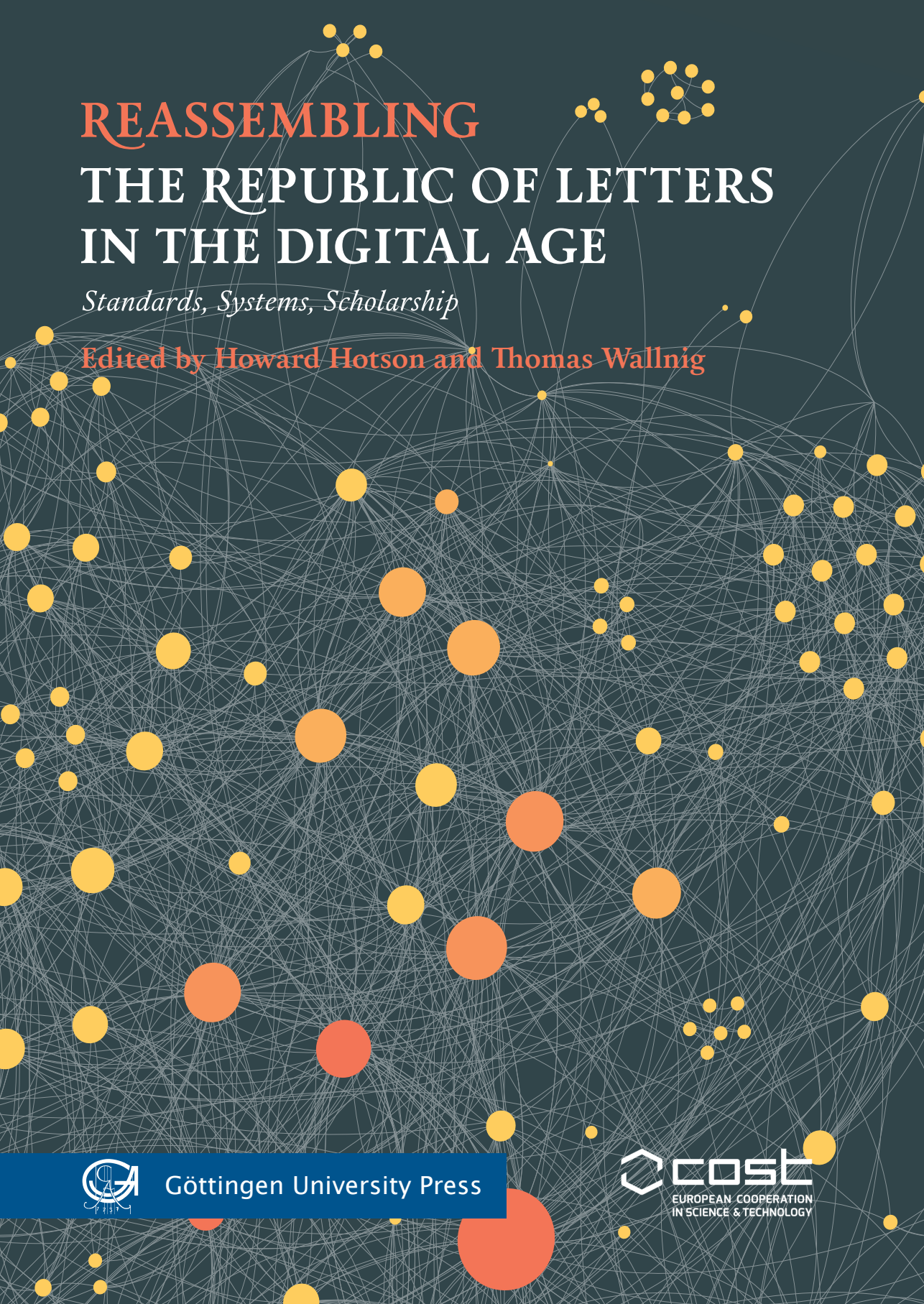
REASSEMBLING THE REPUBLIC OF LETTERS IN THE DIGITAL AGE

Standards, Systems, Scholarship

Edited by Howard Hotson and Thomas Wallnig

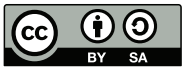


Göttingen University Press

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

Howard Hotson/Thomas Wallnig (eds.)

Reassembling the Republic of Letters in the Digital Age



This article/publication is based upon work from COST Action IS 1310 supported by COST (European Cooperation in Science and Technology).

COST (European Cooperation in Science and Technology) is a funding agency for research and innovation networks. Our Actions help connect research initiatives across Europe and enable scientists to grow their ideas by sharing them with their peers. This boosts their research, career and innovation.

www.cost.eu

Funded by the Horizon 2020 Framework Programme of the European Union

Published by Göttingen University Press 2019

Reassembling the Republic of Letters in the Digital Age

Standards, Systems, Scholarship

Edited by Howard Hotson and Thomas Wallnig

Göttingen University Press
2019

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de

Address of the Editors
Prof. Howard Hotson
E-Mail: howard.hotson@history.ox.ac.uk
https://www.history.ox.ac.uk/people/professor-howard-hotson

PD Dr. Thomas Wallnig
E-Mail: thomas.wallnig@univie.ac.at
https://ifg.univie.ac.at/ueber-uns/mitarbeiterinnen/projektmitarbeiterinnen/thomas-wallnig/

Typesetting and layout: Christoph Kudella
Cover design: Tommaso Elli and Beatrice Gobbo
DensityDesign Research Lab, Milan
Cover picture: Network graph

https://www.univerlag.uni-goettingen.de/
ISBN: 978-3-86395-403-1
DOI: https://doi.org/10.17875/gup2019-1146

Table of Contents

I Reassembling the Republic of Letters

I.1 Introduction

Howard Hotson and Thomas Wallnig

1 The Challenge: Reassembling the Republic of Letters

Between 1450 and 1650 a communications revolution transformed the capacity of learned Europeans to engage with one another. In the previous period, universities had been established and endowed with legal rights and privileges designed to help spread learning across Europe, and in these two centuries their number effectively quadrupled: to the forty-five universities active in Europe in 1450 were added 120 new universities by 1650, as well as nearly 100 immediately sub-university institutions. During the same period, printing presses, the paper they used, and the books they produced grew in number even more rapidly. By 1500 – only 45 years after the first Gutenberg bible – 282 known printers had produced over 28,000 surviving editions originally totalling perhaps 20 million copies; and during the sixteenth century those numbers would increase more than tenfold. The networks created by the book trade were added to monastic, academic, and diplomatic communication channels which were also growing in this period, and scholarly exchange was further stimulated by the Renaissance rediscovery of the familiar letter as literary genre. During the sixteenth century, a dense postal network developed between imperial free cities which allowed ordinary people to exchange letters within the Holy Roman Empire; and to this was added an imperial postal service in 1597. In England, the Royal Mail, created by Henry VIII in 1516, was opened for public use in 1635; and similar arrangements in the Low Countries, France, Spain and elsewhere steadily increased the volume of correspondence in circulation with-

in Europe. During the seventeenth century, newsletters in script and print evolved into newspapers, and learned correspondence into the first learned journals.

The subsequent exchange of steadily increasing quantities of learned letters knit Europe together in unprecedented ways fundamental to the revolutionary intellectual developments of the early modern period. Thematically, these letters meander unpredictably through the entire world of learning, displaying the pursuit of knowledge with an immediacy rarely replicated elsewhere. Geographically they spun a web of direct, personal, and reciprocal communication from one end of Europe to another and beyond it to Asia and America. Socially, correspondence bound together people who had never met or might not care to mix socially: princes and aristocrats, gentlemen and scholars, men and women, diplomats and officer-holders, physicians and apothecaries, clergymen and schoolteachers, students and tutors, printers and booksellers, merchants and travellers, instrument makers, craftsmen, alchemists, and astrologers: all these and many more jostle together in the most representative correspondences of the period. Last but not least, learned letters have a unique fascination of their own, representing the closest we will ever come to eavesdropping on the informal conversations between men and women of learning hundreds of years ago

In idealistic moments, key figures in this new world of learned exchange saw themselves as living the most meaningful parts of their lives in a new kind of imagined community which they called the *orbis eruditorum* ('the world of the learned'), the *sodalitas doctorum* ('the fraternity of the learned'), or the *respublica litteraria*: 'the republic of letters'. This new world was an open society in which bonds and duties were created not by law, custom, or power relations but by mutual services to the cause of learning; a meritocratic society in which status was determined neither by birth nor by wealth but by learning and insight; and a transnational and tolerant community, existing above and beyond the narrower bounds of ethnicity, nationality, profession, and even religious confession, and held together by intellectual aspirations, learned values, and cultural ideals. From the pan-European scholarly renaissance of the Erasmian era in the early sixteenth century, via the period of the new scientific societies of the seventeenth to the heyday of the European Enlightenment in the eighteenth century, the republic of letters helped foster the advancement of learning and the integration of cultural norms while germinating and disseminating the series of intellectual breakthroughs that ushered in the modern world. As a learned ideal (if not always as an actual practice), this imagined community helped formulate, propagate, police, and institutionalize a whole range of values and practices relevant to Europe's idealized self-image today.

Given its obvious relevance, this priceless cultural heritage is strangely marginal to the cultural memory of Europe. Tens of thousands of volumes of early modern printed letters are preserved in Europe's libraries, containing millions of letters, not counting the even larger quantities of manuscript epistles. Although a large body of specialized literature is devoted to the republic of letters, it is scarcely even a standard category for introductory undergraduate study, much less one familiar to the

general public, and hardly features in the popular historical imagination at all. Why has this vitally important community disappeared from Europe's cultural memory? And how can modern scholarship be equipped both to recapture it more fully and to communicate it more vividly to policy makers, young people, and the general public?

The answers to all of these questions are both technological and political. Part of the problem results from the limitations of scribal correspondence as a communication technology. Letters are messages sent between people in different places. They can only perform their communicative function by being dispersed. The very exchange of letters which helped create the *respublica litteraria* therefore dispersed the documentation required to study it. Reassembling the scattered letters, even of a single famous individual, remains an extremely laborious process, sometimes requiring lifelong labours of whole teams of scholars. To this is added the further limitations of print technology more generally. Ink on paper is an excellent medium for deciphering transcribing, translating, and annotating letters and publishing the results, but inadequate as a means of navigating, analysing, and visualizing the immense quantities of complex data needed to gain an overview of this phenomenon.

The constraints imposed by these technologies have been further compounded for centuries by political conditions. In its late seventeenth- and eighteenth-century origins, the *historia litteraria* was consciously transnational in scope; but the professional study of history and vernacular literature came of age in the long nineteenth century under the conditions of nationalism. The greatest archives, libraries, and museums of Europe founded in this era are typically organized along national categories, and so are the biographical dictionaries and historical syllabi ultimately deriving from them. As a consequence, national narratives and cultural heroes are far more firmly institutionalized than their transnational equivalents. Most historical and literary teaching and research are still organized along national lines, to which transnational narratives are marginal.

In recent years, however, the political and technological context has shifted palpably, increasing the urgency of a fresh wave of research on the republic of letters while transforming the means with which it can be pursued. The success of nineteenth-century institutions in fostering national identities now retards the development of the transnational level of identity that Europe today urgently needs. In order to end the fratricidal patterns of the past and to flourish in a world of global commerce and exchange, Europe needs to be integrated economically. But economic integration requires monetary integration: commerce can flow most freely without the fluctuation of dozens of independent currencies. Monetary integration, however, requires political integration: without it, the union cannot be preserved by transfers from those regions that benefit from the common currency to those that are harmed by it. But political integration is impossible, in turn, without cultural integration: Europeans will only share more sovereignty when they have been persuaded that what they have in common is more important than what

divides them. In these new circumstances, stories of genuine transnational integration and achievement take on unprecedented value. The European Union urgently needs to invest in those transnational stories at something like the scale and intensity of the huge investments required to create national identities in the long nineteenth century.

If shifting political conditions have increased the public value of the history of transnational entities like the republic of letters, rapidly changing technology is simultaneously transforming the prospects for a fresh assault on this difficult scholarly topic. Digital technology, in a word, is far better adapted to reassembling data on the republic of letters than anything available previously; and the reasons for this advantage are rooted in the nature of letters themselves. In order to serve their function, manuscript letters needed to include five key pieces of information: the recipient and their address had to be recorded if the letter was to reach its destination; the sender and date of sending had to be recorded for the message to be interpreted properly; and the place of sending had to be included to facilitate a reply. These five data points, plus a sixth for the current location of the letter, provide a simple, stable, and intuitive data model normally sufficient to distinguish one letter from another. Thanks to this model, large quantities of data already exist in several different kinds of catalogues and inventories, and still larger quantities can be generated relatively easily from printed and manuscript letters. Human beings are not very good at analysing data of this kind unaided, and as a consequence most of this data has been lying dormant. But such analysis is precisely what computers do best, and the last two decades have consequently seen the spontaneous proliferation of digital catalogues, inventories, archives, and editions of correspondence, many including some combination of manuscript images, machine-readable transcriptions, annotations, biographical details, and bibliographical records.

In order to bring these independently created and hosted data silos together into a single pool of homogeneous data, attention must now be focused on creating and disseminating shared standards and systems. What is needed is a shared digital platform on which whole communities of scholars can collaborate in piecing back together the innumerable scattered tesserae of the republic of letters into something approaching a coherent mosaic. At the heart of this platform must be a digital toolset capable of facilitating and transforming every stage of the scholarly process: cataloguing individual manuscript letters; reassembling entire correspondences; transcribing, annotating, and analysing texts; visualizing huge collections of complex metadata; and even modelling the thematic structure of entire networks of correspondences. In this way, the digital revolution of our own time can help resolve the scholarly problem resulting from the communications revolution of the early modern period. Likewise, the collective study of the *respublica litteraria* can help fashion a new international scholarly community, safeguarding and propagating what was best in the values and practices of the old one, while overcoming some of its limitations as a predominantly male and primarily elite phenomenon.

Together these processes can transform current understandings of this early form of transnational, knowledge-based civil society, inject powerfully distilled depictions of it into the public domain, and thereby help reinforce European values and shape European identities in the future.

2 The Action: Objectives, Scope, and Structure

Realizing this immense potential can only be achieved in stages. A really adequate conception of the republic of letters and its values cannot be reinjected into European cultural memory until a representative body of readily navigable, analysable, and visualizable material has been assembled collaboratively. Such collaboration requires a shared electronic platform; but such a platform cannot be built or populated before its general specifications have been agreed with a large range of potential partners. The necessary precondition for all of these developments, therefore, is a carefully organized process of negotiation, in which representative committees assemble from across Europe all the diverse kinds of expertise needed to devise the standards, specifications, and arrangements required for the proper functioning of a 'scholarly social machine' for global collaboration in the epistolary field.

For this purpose, most funding agencies are poorly suited. National funding schemes are not set up to address international issues of this kind. Research funding agencies are often ill-suited to long-term experiments with the specification and use of emerging technology. This project has been fortunate in the support it has received from two complementary funding agencies of a rather different mould. A series of four grants from the Andrew W. Mellon Foundation of New York has made possible a decade-long experiment in the design, development, and population of a union catalogue of learned correspondence collaboratively, undertaken by the *Cultures of Knowledge* project in Oxford and resulting in the open-access digital resource, *Early Modern Letters Online* (EMLO). As well as generating a great deal of experience in and reflection on a large set of interrelated challenges, this project has served to confirm the appetite of the international scholarly community to collaborate in new ways made possible by digital technology and identified the core of an international and interdisciplinary network interested in pursuing this objective.[1]

In order to take these discussions to the next level, what was needed was a funding scheme designed to gather an interdisciplinary network geographically coextensive with the early modern republic of letters itself and to sustain for several years the discussions within it needed to sharpen the conceptions, to identify the needs, to define the standards, to devise the technical solutions, and to generate

[1] Howard Hotson, 'Cultures of Knowledge in Transition: *Early Modern Letters Online* as an Experiment in Collaboration, 2009–2019', in Simon Burrows and Glenn Roe, eds., *Digitizing Enlightenment: Digital Humanities and the Transformation of Eighteenth-Century Studies*, Oxford University Studies in the Enlightenment (Liverpool: Liverpool University Press / Voltaire Foundation, forthcoming 2020).

the funding models capable of setting both scholarship and public engagement in this field on a new footing, and of piloting the solution to analogous challenges in cognate fields as well. For this purpose, nothing could be more appropriate than a COST networking grant.

COST is an intergovernmental framework for European Cooperation in Science and Technology. Rather than funding research, resource creation, or IT systems development, COST Actions provide the networking support needed to ensure that nationally funded initiatives add up to something greater than the sum of their individual parts. Between April 2014 and April 2018, COST Action IS 1310, *Reassembling the Republic of Letters, 1500–1800*, was funded to pursue four years of structured discussions devoted to designing (in the words of the project's subtitle) 'a digital framework for multilateral collaboration on Europe's intellectual heritage'. Its objective can be described as 'networking to the fourth power': it aimed to assemble a network to design new networking infrastructure to support a scholarly network studying past networks.

In more prosaic terms, the objectives of this Action are essentially two-fold: technical and historiographical. The technical objective was to plan a state-of-the-art digital system within which to collect a pan-European pool of highly granular data on the republic of letters. The historiographical agenda generated the fresh research questions needed to design the infrastructure and used the emerging technology to experiment with new methods, pose new questions, and answer old ones. The relationship between these two agendas was reciprocal: just as the infrastructure must be designed to address scholarly research questions, the possibilities opened up by the new infrastructure should shape the historiographical agenda in turn. But although the ultimate goals are historiographical, technical issues nevertheless initially take precedence: the primary objective is to devise technical means to facilitate new forms of scholarship which can only be glimpsed today.

These hybrid aims determined the interdisciplinary breadth and geographical scope of the Action and also helped define the range of activities and modes of participation which it employed. To begin with, a community capable of undertaking these discussions must of necessity be highly interdisciplinary. To fulfil its task, the Action needed to attract experts from many different areas of expertise: the histories of science, philosophy, ideas, and media as well as neo-Latin and modern philology were needed for studying the many different dimensions of the republic of letters per se. Library and archival science contributed expertise on cataloguing and preserving letters in script and print. Information technology was crucial to the core task of designing the standards, tools, and infrastructure necessary to these scholarly communities. Data interaction design, visualization, and communication are emerging fields focusing on helping communities design new interfaces and processes for exploring big data. Intellectual property law and knowledge management have important contributions to make as well.

No less important was to obtain participation from as many European countries as possible. By the second year, the Action had recruited nearly 100 Manage-

ment Committee members and substitute members from thirty-two of the thirty-five countries that participate in COST, as well as participants from the United States and Canada. A similar number of affiliates from even further afield also participated informally in the network, many of these being early career scholars who actively participated in many different forms of scholarly exchange.

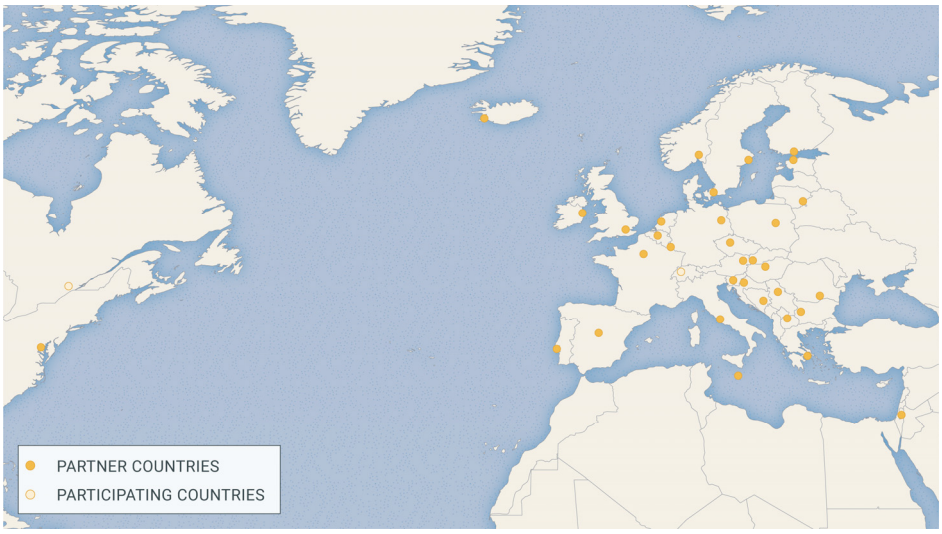


Figure 1: Partner and participating countries in COST Action IS 1310 (visualization by Tommaso Elli)

Such participation took many different forms. Ultimate decision-making authority rested with a Management Committee, composed of up to four members from each participating country. Most Management Committee Members were also members of the six Working Groups, which also included a small number of additional members specially recruited to add necessary expertise. Affiliates to the Action were also welcome to participate in many different ways, especially the early career scholars via Training Schools and STSMs (discussed further below). The chairs of the six Working Groups constituted the core of the Action's Steering Group, which was supplemented by the STSM coordinator and the coordinator of the forthcoming conference. The Action as a whole, the Management Committee, and the Steering Group were led by a chair and vice-chair, supported by a grant manager and project administrator (Dobrochna Futro), webmasters (Giorgio Uboldi and Tommaso Elli), and a website editor (Sue Hemmens).

At the core of the Action were over sixty events and exchanges. The Action was punctuated by three large conferences. The first, coordinated by Elizabeth Williamson, met in St Anne's College, Oxford, in March 2015. The second was convened by Anna Skolimowska in the Institute for Interdisciplinary Studies 'Artes

Liberales' of the University of Warsaw in June 2016. The third, coordinated by Jean-Paul De Lucca, met in the Valletta Campus of the University of Malta in January and February 2018. Training schools were convened at the beginning and end of this series in the University of Oxford (March 2015) and the Estonian Academy of Sciences in Tallinn (March 2018): the first involved partners from Cambridge, The Hague, Lancaster, London, Milan, and Stanford; the latter was led by Miranda Lewis and hosted by Kristi Viiding. Two other mid-scale meetings were held in Como in April 2016 and May 2017: hosted by Paulo Ciuccarelli in partnership with Charles van den Heuvel and supported by Tommaso Elli, these events inducted members of the Action into the 'design sprint' process for generating new kinds of data interaction design in partnership with members of his DensityDesign Research Lab at the Politecnico di Milano (see further ch. IV.1). In addition, a dozen Management Committee and Working Group meetings were hosted in Brussels, Prague, Vienna, Lancaster, Zagreb, Dublin, The Hague, Lisbon, Chester, Wolfenbüttel, Budapest, and Oxford.

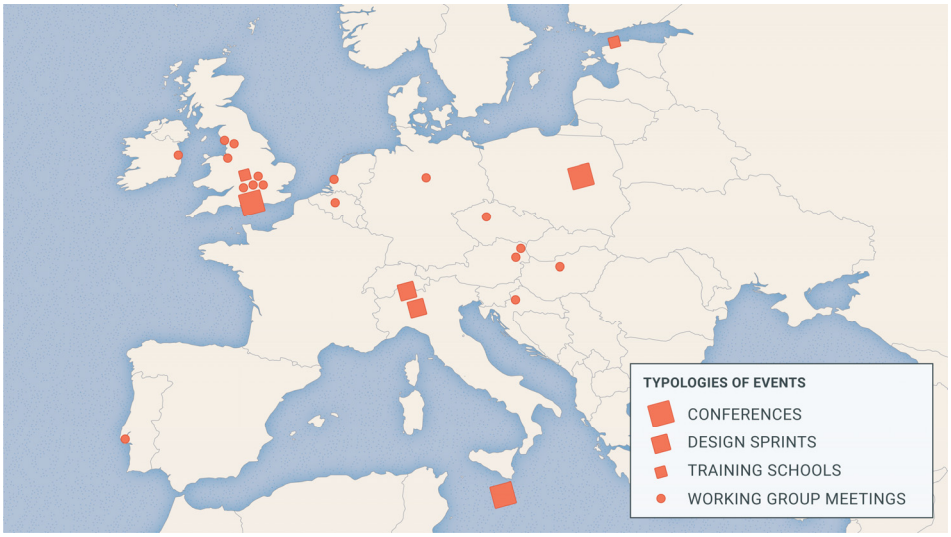


Figure 2: Distribution of events organized by COST Action IS 1310 (visualization by Tommaso Elli)

Among the most fruitful components of the Action was the programme of so-called 'Short-term Scientific Missions' (STSMs), ably coordinated by Vanda Anastácio of the University of Lisbon. This programme funded exchange visits of early-stage researchers from one country to partner institutions in other countries within the network in order to exchange expertise. Over forty such exchanges, ranging between one and three weeks, were funded within the Action, and their

results loom large, particularly in section IV of this volume. The following graphic gives an impression of the pattern of this multilateral exchange.

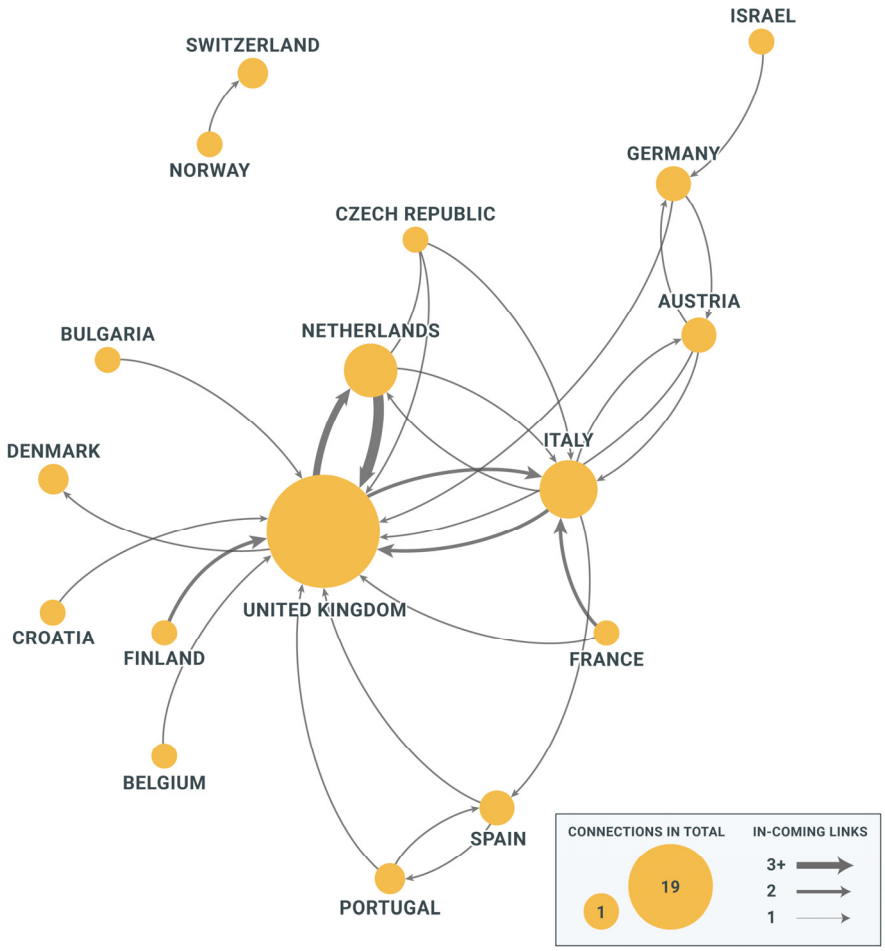


Figure 3: Short-term Scientific Missions undertaken within COST Action IS 1310 (visualization by Tommaso Elli)

The core structure of the Action derived from the intuitive data model for letters discussed above. Four Working Groups dealt, respectively, with the temporal and spatial (WG 1), biographical and social (WG 2), and textual and topical (WG 3) dimensions of letters, as well as the formal and genre conventions of letters, their material properties and archival histories (WG 4). The fifth working group studied the technical challenge of exchanging data between participating projects and insti-

tutions, while the sixth addressed the challenges of analysing and exploring complex data visually.

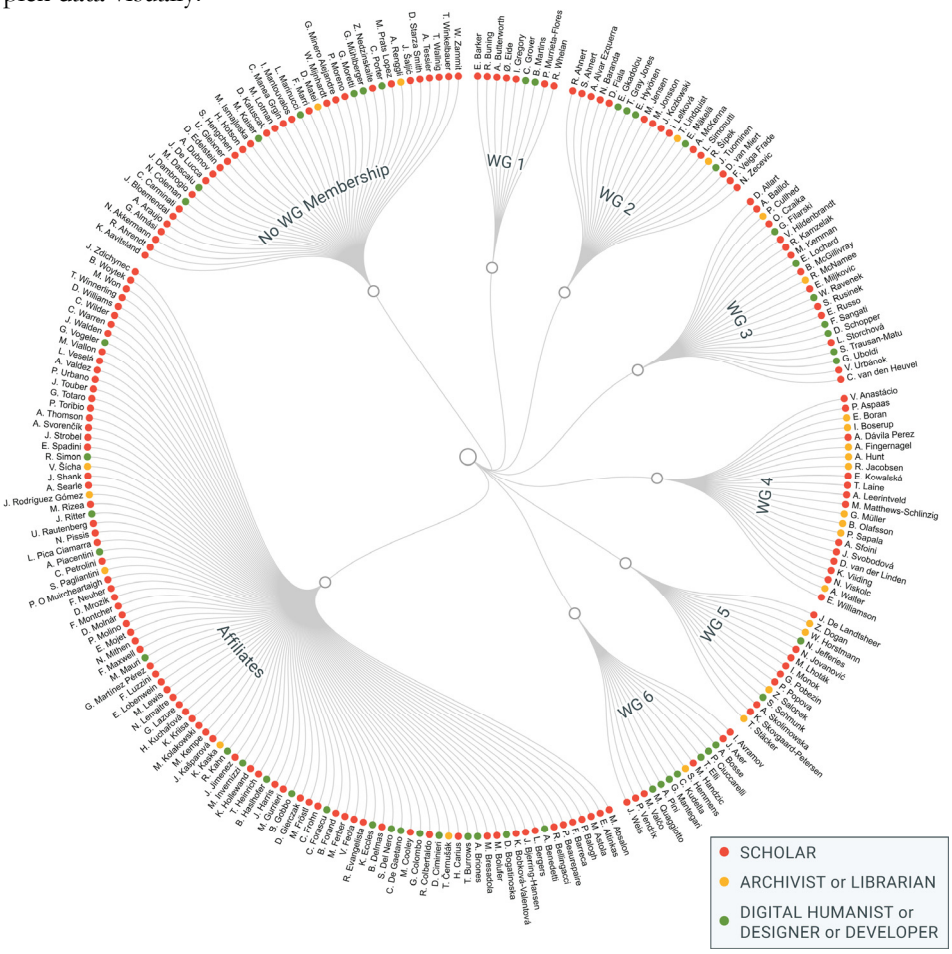


Figure 4: COST Action IS 1310 participants organized by Working Group and colour-coded by specialism (visualization by Tommaso Elli)

WG 1: Space and Time are among the most fundamental dimensions of exchange networks, and to date among the least well understood. Digital technology offers unprecedented assistance in grasping these dimensions, but, to realize this potential, we need standards for presenting spatial and temporal information (chs. II.2–3). WG 1 also hosted a dialogue between historians, geographers, and IT experts exploring how fresh scholarly questions can be answered through the application of geographical information systems, network analysis, Natural Language Processing, and visualization strategies (ch. IV.2). WG 1 was led by Ian Gregory, professor of Digital Humanities in the Department of History at Lancaster University,

and a specialist in the application of geographical information systems to humanistic materials.

WG2: People and Networks addressed the most complicated entities involved in the exchange of learned letters. The basic technical task was to lay the foundations for a prosopographical data model (ch. II.4). The corresponding scholarly and design objective was to experiment with tools and methods for visualizing and analysing data sets – from individual people to multidimensional networks (chs. IV.4–5). WG 2 was led by Eero Hyvönen, director of the Helsinki Centre for Digital Humanities (HELDIG) at the University of Helsinki and research director of the Semantic Computing Research Group (SeCo) at Aalto University. Expertise on quantitative network analysis was provided by Ruth and Sebastian Ahnert.

WG 3: Text and Topics considered digital means of engaging with letter texts themselves. One focus was on tools to aid transcription, annotation, and collaborative editing (ch. III.3). Another was the variety of means of modelling topics and mining texts (chs. II.5, III.4, IV.6). WG 3 also discussed how the tools developed in all the WGs could be integrated into a user-friendly 'Virtual Research Environment' (ch. IV.7). Leadership of WG 3 was provided by Charles van den Heuvel, professor of Digital Method in Historical Disciplines at the University of Amsterdam and head of research of the group on the 'History of Science and Scholarship' at the Huygens Institute for the History of the Netherlands.

WG 4: Documents and Collections pursued two related objectives. One was to contribute to the refinement of a shared data model which includes common definitions of the physical features of the letter, its formal features and genres, and its modes of dissemination and preservation (ch. II.1). The second was to survey the principle sources of epistolary data and to consider the tools and processes needed to collect it (ch. III.1). WG 4 was led by the scholar-librarian Elizabethanne Boran, who edited the correspondence of the most learned Irishman of his day, James Ussher (1581–1656), archbishop of Armagh, while presiding over the Edward Worth Library in Dublin.

WG 5: Data Exchange and Strategic Planning focused the core of its attention on designing a digital system for exchanging metadata within a distributed network of nodes and hubs (ch. III.5). Members of this working group were responsible for many of the more technical chapters in the book, including those considering how to model the republic of letters as a whole (ch. I.3), the events within it (ch. II.6), the letters which helped knit it together (ch. II.7), and the means of reconciling large quantities of data on those letters (ch. III.2). WG 5 was led by Thomas Stäcker, at the time deputy director of the Herzog August Bibliothek, Wolfenbüttel, responsible for the library's programme of digitization and the Wolfenbüttel Digital Library, and now director of the Universitätsbibliothek in Darmstadt. Arno Bosse, digital project manager at the *Cultures of Knowledge* project in Oxford, helped coordinate much of the work of this group and took a lead role in writing up some of its key results.

WG 6: Visualization and Communication explored the application of data interaction design to the challenges involved in navigating huge quantities of data and in communicating both the scholarly and the technical interest of the Action to diverse audiences. Crucial to this process was a pair of 'design sprints' held in Como in the Spring of 2016 and 2017, which brought the scholarly and technical components of the Action together with experts in data interaction design. Some of the principles involved are described in chapter IV.1, and the results of this work feature prominently throughout section IV. WG 6 was led by Paolo Ciuccarelli, scientific director of DensityDesign Research Lab at the Politecnico di Milano, where his students and colleagues created the Action website, contributed indispensable expertise to the design sprints, and have helped develop many of the visualizations published in this volume. Tommaso Elli merits particular mention due to his sustained engagement with all of these activities throughout the duration of the Action, including the polishing of numerous figures in the final stages of the production process.

3 The Book: Structure and Distinguishing Features

The origin of this book in four years of interdisciplinary discussions structured by this framework is not difficult to perceive.

Section I introduces the Action itself (ch. I.1), the republic of letters as an historical phenomenon (ch. I.2), and the task of modelling such an entity as a technical challenge (ch. I.3). These introductions are designed to ensure that readers from diverse disciplines have the background necessary for understanding the scholarly and technical discussions which follow.

Section II is devoted to developing standard means of describing the various dimensions of epistolary data. This involves the basic definition of letters themselves and the distinction of their various states and genres (ch. II.1); the establishment of standardized means for collecting data on the spatial (ch. II.2) and temporal (ch. II.3) dimensions of epistolary metadata, on the people who exchange letters and are mentioned in them (ch. II.4), and on the topics discussed in them (ch. II.5). The meta-category of events also requires attention (ch. II.6), since it is central both to the prosopographical data model and the process-based approach to modelling correspondence. The concluding chapter introduces a new data model for letters which synthesises many of the conclusions of this section (ch. II.7).

Section III then considers the processes and tools needed to handle unprecedented quantities of data of these kinds. First, various kinds of epistolary metadata must be distinguished from one another, various sources of each kind identified, and various processes envisaged for assembling that data for further work (ch. III.1). Then tools will be needed for reconciling and normalizing large quantities of data arriving from a wide range of sources (ch. III.2). For dealing with textual data, tools for transcription and editing require attention (ch. III.3) as well as those for

modelling the topical content of the texts themselves (ch. III.4). The greatest challenge, and the culmination of this chapter, consists of outlining the kind of distributed infrastructure which would allow individual institutions to retain control of their own materials while allowing users to assemble and interact with data from across a distributed system (ch. III.5).

Section IV provides a preliminary glimpse of the scholarly results which might emerge from applying new tools and methods to unprecedented quantities of material. First, a preliminary introduction to the emerging field of data interaction design is required (ch. IV.1). Then thumbnail case studies explore some of the innovative options for dealing with the geographical (ch. IV.2), chronological (ch. IV.3), and prosopographical (ch. IV.4) dimensions of the republic of letters and of the analysis of networks central to it (ch. IV.5). The prospects for mining huge quantities of textual data are also discussed (ch. IV.6) before a concluding chapter considers how all of these tools and methods might be combined in a 'Virtual Research Environment' (ch. IV.7).

Section V begins by providing (1) a concluding synopsis of the recommendations emerging from the volume and (2) a sketch of the prospects for radically multilateral collaboration created by such shared infrastructure. It then shows how infrastructure for reassembling scattered correspondence might be incrementally expanded (3) first to gather data on the republic of letters in the round and then (4) to handle other forms of intellectual, cultural, and economic exchange in other times and places, before returning to consider (5) the social and political implications of the project, including its potential impact on European identities.

Several other unusual features of this book require comment before concluding. Perhaps the most obvious is the wide range of means adopted for acknowledging the diverse forms of collaboration which have contributed to this experimental undertaking. One aim of the volume is to help bridge the divide between the scholarly and technical communities which must collaborate in this field. Synthesising expertise from radically different domains requires collaboration. Collaboration requires co-authorship. Co-authorship comes in many different kinds and degrees, and this volume has attempted to accommodate many of them. A few contributions are single-authored or written equally by two or three people. In other cases, one person has written up results obtained through collaboration with multiple colleagues. Most inextricably collaborative are those cases that have emerged through collaboration between people with very different kinds of expertise in STSMs or design sprints. When many such collaborative case studies are bundled into a single chapter, this can result in a chapter authored by a dozen people. Fully acknowledging the collaborative origins of many of these chapters represents a challenge to the forms of authorship dominant in the humanities and embedded in the research assessment regimes which have proliferated across Europe in recent decades.

Along with co-authorship, a second characteristic of the chapters that follow is an emphasis on clarity. This too is rooted in the decision to expound technical and scholarly material between two covers. The future of this field requires enhanced communication and mutual comprehension across these large disciplinary divides. The most basic precondition of such collaboration is a kind of mutual respect too often lacking from departmentalized academic culture, in which the complementary strengths and limitations that colleagues with very different areas of expertise bring to a project are properly acknowledged and respected. To bridge this divide, all contributors to the volume have agreed to write in language that all members of the community can comprehend, and to provide sufficient background knowledge for their contributions to be comprehensible to non-specialists.

This emphasis on clarity and comprehensibility has a third consequence for the nature of the volume. This book is obviously not the place to look for detailed archival, philological, or historical scholarship, for fully rounded and nuanced case studies, or for grand syntheses: it is far too early for the latter, and there is no space for the former. Likewise, this volume is not the place to look for detailed, technical specifications of digital infrastructure which could be immediately implemented: such language would be incomprehensible to many readers and would quickly be rendered obsolete by the rapid pace of technical development. Instead, the focus of this volume is on the clear exposition of general problems and of the general principles structuring possible solutions to them. The aim of this volume is not so much to provide definitive solutions to individual problems as to outline a framework of interrelated challenges that need to be confronted together if the potential of this field is to be realized. Even after many of the technical solutions sketched out here are rendered obsolete by the rapid pace of technological progress, the overall framework may continue to provide a point of reference for coordinating work across such a large and complex field.

These emphases on collaboration, co-authorship, clarity, and a complex network of interrelated challenges relate to a fourth characteristic of this volume: the attempt to design infrastructure from the bottom up. Consortia already exist devoted to developing high-level, generic environments, including DARIAH, CLARIN, Europeana, and TEI. Scholars in humanities disciplines are often slow to engage with such generic tools and environments, especially when built from the 'top down' without broad-based input from the scholarly community. The problem is that these high-level infrastructures will be adopted by individual scholarly communities only if they are carefully tailored to their needs; this process of tailoring is laborious, painstaking, and time-consuming; it can only be undertaken in active collaboration with the scholarly communities whom the infrastructure is designed to serve; and such collaboration will only be forthcoming if a fully engaged scholarly community has 'taken ownership' of the process of designing the tools and populating the infrastructure itself. In other words, the digital transformation of humanistic scholarship will take place at an intermediate level of general-

ity, in which scholarly communities, working 'from the ground up' meet digital infrastructure providers working 'from the top down'.

Viewed from this perspective, this volume summarizes an experiment in pursuing the former and perhaps the less common of these complementary approaches. It began with a scholarly community eager to devise tools and techniques for its own needs and working methods; and it then built outwards to address more general problems and upwards to more generic instruments applicable to a wider variety of uses. The proposals emerging from this process outline an intermediate-level environment connecting individual scholars, projects, and institutions to very high-level digital infrastructure. One of the most important contributions of the project may be to exemplify a process through which a very broad international and interdisciplinary scholarly community becomes deeply involved in designing integrated transnational infrastructure, filling it with content, curating that content by means of its collective expertise, and thereby inhabiting that infrastructure as a routine part of cutting-edge scholarly practice. Since, by common consent, the social and cultural challenges of adapting and adopting digital infrastructure are at least as daunting as the technical challenges of building it, this is no insignificant achievement.

I.2 What Was the Republic of Letters?

Dirk van Miert, Howard Hotson, and Thomas Wallnig

1 Definitions and Distinctions

When European historians describe the intellectual movements of the early modern period, they frequently refer to the Renaissance of the fifteenth and sixteenth centuries, the scientific revolution of the seventeenth, and the Enlightenment of the eighteenth. The first two of these terms, however, are not 'actor's categories'. Contemporaries did not use them to describe the intellectual enterprises which bound them together: they are terms devised centuries later by historians and retrospectively imposed on the period from outside. While *lumière* was a favourite metaphor for mid-eighteenth-century contemporaries to describe their age, it was subsequent historians who applied it retroactively as the defining characteristic of the entire century and more.

When early modern intellectuals before c. 1750 referred to the enterprise which they shared, they sometimes used the Latin phrases *respublica litteraria* or *respublica litterarum*. From the later seventeenth century onward, these terms made their way into a variety of European vernaculars, most influentially in the French term *la république des lettres*, but also in English as 'the republic of letters', or 'commonwealth of learning' and in German as *die Gelehrtenrepublik*. What precisely do these terms mean? What were the implications of thinking of the learned community of Europe as a 'republic of letters'? Does this term describe the manner in which those who used it actually behaved, or merely a set of ideals or aspirations?

This chapter endeavours to provide preliminary answers to some of these questions by proceeding in three stages. First, we unpack the implications of these

Latin terms *res publica* (1.1) and *litteraria* (1.2) in a programmatic and theoretical fashion. Second, we consider the vexed question of how far the realities of the republic of letters deviated from these ideals (1.3) and the multiple dimensions of that deviation, including time, space, discipline and language (1.4). Finally (in section 2), we sketch out a rough framework for thinking about how these dimensions vary with time and place. Exactly what contemporaries meant by these terms is open to interpretations that vary from one scholar to the other, and much of that variety is caused by the different periods on which modern interpreters focus or the changing contexts in which early modern people employed the phrase. None of these brief discussions can have any claims to definitiveness. Instead they are offered here merely to ensure that all readers of this volume, irrespective of area of academic specialism, have access to a set of basic conceptions of what the term 'republic of letters' might be thought to signify.

1.1 *Res publica*

The obvious starting point is to look at the root meaning of these Latin terms. Literally, *res publica* means 'the public thing', something created and held in common by a large number of people. To describe the world of learning as a *res publica* in this most basic sense is to emphasize that learning is a good common to all nations, regions, and confessions and perhaps to suggest that it has been created and sustained collectively as well.

But what kind of *res* is a *res publica*? The standard answer would be that this 'public thing' is a political entity, a state or commonwealth.[1] To suggest that the learned community is a republic in this sense implies that it is an independent political entity, an autonomous, sovereign, self-governing authority, a law unto itself. The *respublica litteraria* therefore implicitly exists alongside the empires and monarchies and republics which exercised political authority in early modern Europe, but it is independent of these polities because it is not subject to their authority (*regnum*). Much the same might be said of the relationship between the *respublica litteraria* and ecclesiastical authority (*sacerdotium*). The term *respublica litteraria* bears some affinity to the contemporary notion of a *respublica Christiana*, but the nature of that affinity is unclear.

What the term does seem to imply, however, is that the laws of the *respublica litteraria* are not established by kings or parliaments, nor by popes, councils or churches: they derive from the citizens of the republic of letters themselves. Better

[1] A vast literature exists on ancient, early modern and contemporary notions of republicanism, and the precise meaning of the term is contested. See for instance Knud Haakonssen, 'Republicanism', in Robert E. Goodin and Philip Pettit, eds., *A Companion to Contemporary Political Philosophy* (Oxford: Blackwell, 1995); Martin van Gelderen and Quentin Skinner, eds., *Republicanism: A Shared European Heritage*, vol. 1: *Republicanism and Constitutionalism in Early Modern Europe*; vol. 2: *The Value of Republicanism in Early Modern Europe* (Cambridge: Cambridge University Press, 2002), and, for an influential modern restatement, Philip Pettit, *Republicanism: A Theory of Freedom and Government* (Oxford: Oxford University Press, 1997).

still, many of them are laws of nature, which learned men and women have learned to recognize and to canonize from within the fabric of reality itself. Princes do not dictate the axioms of geometry or the principles of natural theology. Churches do not legislate grammar, rhetoric, logic, history, or optics. It is up to the citizens of the republic of letters to debate these rules for themselves and also to enforce them: those who refuse to obey the most fundamental of these laws lose the rights of citizenship and are ostracized if not banished from the learned republic.

If the *respublica litteraria* is not ruled by princes or parliaments, by popes or preachers, then this implies that it is a self-governing entity; and this brings us to another association of the original term. In Roman history, the republic is that phase of development during which the Roman people had freed themselves of subjection to a king and had not yet been subjected to an emperor (traditionally dated 509–27 BC), which was also the period during which Rome extended its rule throughout the Mediterranean world.[2] Those wealthy, landowning families who had overthrown kings and tyrants shared regal powers among themselves through election to a multiplicity of public offices – political, military, and religious. Later, both the ability to vote in these elections and to hold some of these offices was extended to the most reputable men from wealthy plebeian families, particularly those who had distinguished themselves through service to the *res publica*. In order to maintain the *res publica*, great emphasis was placed on civic virtue, active participation, and the rule of law.

In similar fashion, the *respublica litteraria* might regard itself a self-governing community, free of arbitrary rule, in which authority was shared in proportion to the individual's intellectual merit and service to the cause of learning as a whole; and foremost among these services was that of helping to overthrow unjustified intellectual tyranny and to (re)gain intellectual liberty. In the sixteenth century in particular, these republican ideals were reinforced by the tendency of many humanists to take the Roman senator Cicero as their stylistic model. The humanists implicitly reference the Roman Republic by employing its discourse, praising each other as consuls, triumvirs, and princes, or even (and not in the pejorative sense) as dictators. Those who seek to try to silence others are styled as tyrants. No great scholar was called a 'king of the republic of letters'. Indeed, although access to at least modest wealth remained a virtual precondition of citizenship and social status remained an asset, the republic of letters surpassed the Roman Republic in embracing merit and personal accomplishment rather than birth and family history as qualifications for full citizenship.

[2] Harriet I. Flower, ed., *The Cambridge Companion to the Roman Republic* (Cambridge: Cambridge University Press, 2004).

1.2 *Litteraria*

So much, then, for the implications of the first term in the phrase. What is implied by the second term? One slightly misleading answer is suggested by the standard English translation of this term as 'the republic of letters'. The specific understanding of the republic of letters in terms of an epistolary network is bound to be enhanced, unconsciously or not, by the fact that in the dominant languages – Latin, French, and English – the *respublica litterarum* has a felicitous second subliminal connotation that enriches the association with communality and communication: 'letters' meaning also 'epistles'.[3] Epistolary communication was a central, pervasive, and characteristic feature of the *respublica litteraria*. In their letters, learned men (and some women) shared information about work-in-progress and recently published books; they gossiped about colleagues and recommended students; they reflected on the politics of universities, princes, and the Church; and they reported on family matters and personal health. Letters were meant to be answered: reciprocity was a vital principle, and the letter-writers honoured the cult of communication. So to foreground correspondence in our conception of the republic of letters is not entirely unwarranted; but to define the *respublica litteraria* primarily as a republic of epistolary communication would be a grave mistake.

Fortunately, for German, Spanish, and Dutch scholars, the notion of a republic of *Briefe* or of *cartas* sounds very peculiar. For this reason, Germans in the early modern period adopted the very different translation of *respublica litteraria* as *Gelehrtenrepublik*, *Republik der Gelehrten*, or even *gelehrte Republik*, that is, the 'republic of the learned'. Likewise in Latin, alternative designations abounded in the early modern period, including *orbis eruditorum* ('the world of the learned'), the *sodalitas doctorum* ('the fraternity of the learned'), *mundus litterarius* ('the learned world'), *omnis litteratorum cohors* ('the cohort of all men of letters'). The English designation 'the commonwealth of learning', seems to capture both the idea that this was a social world of learned people, as well as a wider 'world' that included not only people but also institutions and infrastructure. Indeed, some early modern scholars thought of the republic of letters as the assembly of learned institutions such as universities and societies.[4]

[3] Constance M. Furey, *Erasmus, Contarini, and the Religious Republic of Letters* (Cambridge: Cambridge University Press, 2006); Hanan Yoran, *Between Utopia and Dystopia. Erasmus, Thomas More, and the Humanist Republic of Letters* (Plymouth: Lexington Books, 2010).

[4] Hans Bots has recently styled the republic of letters as the 'intellectual world of Europe', and described it not only anthropologically as a community of people bent on the exchange of knowledge, but also as a 'world' that included practices such as epistolary traditions, institutions such as universities and societies, commercial stakeholders such as the book printers and traders, and a reflective discourse embodied in the medium of the journal. Hans Bots, *De republiek der letteren. De Europese intellectuele wereld 1500–1760* (Nijmegen: Vantilt, 2018).

So if 'letters' does not primarily mean 'epistles', neither does it imply merely literary pursuits in the narrow sense of *belles-lettres*: instead, the term embraced the whole of learning, including the natural and social sciences, mathematics, history, and the learned disciplines of medicine and civil law. But here too, a number of limitations must be acknowledged.

For one, even before the Reformation, sacred letters overlapped only partially with the *litterae* defining this learned republic. The ultimate sources of authority for institutionalized, academic theology lay outside the *respublica litteraria*. The legislators in matters theological were churchmen, not mere scholars. After the Reformation, however, the European *res publica* claimed the right to dispute some theological questions framed in terms of antiquarianism, law, or philosophy. In such cases, an irenic trans-confessional exchange was possible, as is shown by the case of Leibniz's church reunion projects,[5] in which astronomy played an important role.

Second, as the term also implies, the 'republic of letters' was a commonwealth, not merely of words, but of written and later printed words. Within this commonwealth, forms of learning without a literary pedigree were marginalized. In other words, the learning which defined this learned republic was 'liberal' in the classical sense of embracing the liberal arts and marginalizing if not quite completely excluding the vulgar, mechanical, or manual ones. This sidelining of artisanal forms of knowledge, based on craft skill, passed down in practice-based learning from master to apprentice, and typically involving manual labour, was explicitly advocated even by many who (like Francis Bacon) acknowledged their importance and celebrated their fertility. Technology and engineering, chemistry and alchemy, the visual arts of painting and sculpture were at a disadvantage, even if these boundaries were neither static nor unbreachable. Rubens earned his citizenship on the basis of his literary accomplishments and learning, not the brilliant painting for which he is now famous. Spinoza was a citizen by virtue of his philosophical writings, not his expertise as a lens grinder. The exception proves the rule: the class of artisans most warmly welcomed within the republic of letters were the ones who dealt in letters: that is, the printers, publishers, and booksellers.

1.3 Ideal, Reality, and Practice

From the term *respublica litteraria*, therefore, can be easily extracted an implicit conception of the realm of learning as a self-governing commonwealth of learning, independent of states and churches, bound together by shared notions of intellectual virtue and mutual service to the common good, exercised by a prosperous but meritocratic elite, reasserting the liberty of thought as paramount throughout all domains of secular learning, and extending its beneficent rule over the whole world

[5] Wenchao Li, Hans Poser, and Hartmut Rudolph, eds., *Leibniz und die Ökumene* (Stuttgart: Franz Steiner Verlag, 2013).

of the arts and sciences. From this idealized self-conception, a number of questions inevitably arise. Etymology, to begin with, is an insecure foundation for an empirical discipline such as history. Did self-proclaimed citizens of the republic of letters actually think of themselves in these terms? If so, did these conceptions spring fully formed from the head of Erasmus, or did they evolve more gradually before reaching mature form midway through the early modern period? And even for those who fully embraced this shared conception, to what extent did real citizens of the republic of letters attain this ideal? If we must concede that the reality often fell short of the ideal, did the idea of the republic of letters nevertheless exercise an agency of its own, with the power to shape individual and collective behaviour at least to some extent? These are among the most interesting questions underlying historical scholarship in this field.

Given that it was bound together primarily by values, attitudes, aspirations, and practices rather than formal institutions or legal obligations, one useful starting point is to describe this world of scholars and scientists as an 'imagined community'. To do so is to adopt and repurpose a phrase developed by Benedict Anderson to explain the formation of 'nations' in the nineteenth century.[6] However intuitively apposite the phrase may be, it must be remembered that Anderson's 'imagined community' of the nation, in distinction to the elite republic of letters, was solidly cemented through the use of a national language. This conception of the 'nation' could only arise *after* the fall of Latin as a *lingua franca*, the universal language of learning from the Roman Empire to the Renaissance and beyond, accessible to all quarters of western and central Europe on an equal basis, and the acknowledged repository of most of its learning and even more of its ongoing discourse. To a large degree, the *respublica litteraria* was also an imagined community built on a shared language, which remained its second language until the decline of the republic of letters at the end of the Ancien Régime.

To this abstract and imagined community must of course also be added the more concrete conception of the republic of letters as a social network, or perhaps better a network consisting of many different kinds of networks. In this sense, the republic of letters is understood as all those involved in the pursuit of liberal learning: professors, secretaries, courtiers, physicians, lawyers, or 'virtuosi' rich enough to support themselves. In such a model, the scholars are nodes, and the letters they addressed to one another constitute one species of edges, that can be qualified in terms of weight (number of letters) and direction (sender–recipient). Such an epistolary network is a proxy for a social network which can be further mapped by identifying other sorts of ties: oral communication (recorded in diaries, for instance), friendship (documented by occasional verses and inscriptions in *alba amicorum*), teacher–student relations (evidenced for example by disputations), client–

[6] Benedict Anderson, *Imagined Communities. Reflections on the Origin and Spread of Nationalism* (London: Verso, 1983). It is used to characterize the republic of letters in, for example, Robert Mayhew, 'British Geography's Republic of Letters: Mapping an Imagined Community 1600–1800', *Journal of the History of Ideas* 65 (2004): 251–76, see https://doi.org/10.1353/jhi.2004.0029.

patron relations (immortalised in dedications), and concrete social relations such as ties of kinship, profession, religion, origin, or shared identities in terms of sex and social position.

If, therefore, the republic of letters is thought of as such a social network, then it must not be regarded as an isolated or monolithic one. Scholarly and scientific networks were not self-contained; they interacted with broader parts of society, as well as with institutions of political, religious, and economic power. The disputes within these institutions were not confined to scientists and scholars arguing over a specific world view (or paradigm, if we want to use an iconic term),[7] but very often those conflicts overlapped: Galileo's trial before the Roman inquisition, in which a scientific debate became a religious and social one,[8] may be the most prominent example. The republic of letters was formed by several networks, overlapping with each other, that were part of a broader context of networks within early modern society.

Needless to say, the social reality of the republic of letters often departed from the abstract intellectual ideal. No ordinary political or ecclesiastical institution ever perfectly exemplifies its ideals in practice, and the republic of letters was no exception, not least because its concrete social reality was deeply intertwining with the political, confessional, and social institutions of its age. Realpolitik, confessional strife, and even social conventions obstructed the free exchange of communication, and scholars and scientists embodied ideas that followed the priorities or their own agendas, whether or not these were dictated by contexts that set limits to the circulation of knowledge.[9] In fact, scholars often vilified each other in pamphlet wars and vicious polemics, battling against charlatanism, plagiarism, vanity, and arrogance, but sometimes in the process descending into prying, spying, mudslinging, or outright lying – ironically excused, of course, by the unimpeachable objective of defending 'the' truth. Yet scholars should beware the temptation to construct a disjuncture between the republic of letters as a discursive ideal and as a quarrelsome and unedifying reality. For the ideal was also, on a day-to-day basis, to greater or lesser extents, exercised by early modern people and hence brought into practice. Idealism and pragmatism could converge in strategic attempts to lay hands on information that was in the possession of the confessional or political

[7] The well-known concept of 'paradigm shift' was indeed developed by Thomas S. Kuhn, a physicist looking at the history of early modern science, who wanted to understand the social dynamics behind the refusal of one explanatory system and the acceptance of another. The transformation of cosmology – from geocentric to heliocentric – plays a pivotal role in his study: Thomas S. Kuhn, *The Structure of Scientific Revolutions* (Chicago: Chicago University Press, 1962); id., *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought* (Cambridge, MA: Harvard University Press, 1957).

[8] For the social dimension, see the famous study of Mario Biagioli, *Gallileo, Courtier: The Practice of Science in the Culture of Absolutism* (Chicago: University of Chicago Press, 1993).

[9] Nick Hardy has most recently warned against describing the republic of letters as striving to objectivity, irenicism, or ecumenism through reasoned debate or even self-regulating polemics. Nicholas Hardy, *Criticism and Confession. The Bible in the Seventeenth-Century Republic of Letters* (Oxford: Oxford University Press, 2017). Such optimistic views are espoused by Peter N. Miller, *Peiresc's Europe: Learning and Virtue in the Seventeenth Century* (New Haven and London: Yale University Press, 2000).

'other'.[10] After all, for the exchange of texts, objects, and ideas (epitomized in the expression *commercium litterarium*, often used by early modern scholars in connotation with the term *respublica litteraria*), it was more effective to seek common ground than to stress differences.[11] Modesty, friendliness, openness, constancy, patience, forgiveness, and industry were frequently upheld in practice as well as theory as the moral codes of the republic of letters. Early modern scholars were seldom naively fooled by high-minded protestations of mutually shared purposes, but the fact that early modern scholars usually employ the phrase 'republic of letters' when they praise the services of their interlocutors to some kind of common good was performative as well as discursive: for it meant not only a description of merits, but also a prescription to heed codes of conduct that were not laid down (at least not in the earlier history of its manifestations) but socially constructed and transmitted.

1.4 Dimensions: Time, Space, Discipline, and Language

If the republic of letters was not isolated from or uncontaminated by the concrete institutions of its day, neither was it undifferentiated. On the contrary, it varied with time, space, and discipline in ways that scholars trapped in narrow scholarly specialisms sometimes fail to appreciate.

Some of the confusion about the long history of the republic of letters, for instance, is due to the fact that research is carried out by different communities: early modernists tend to be separated uncomfortably into several distinct communities studying Renaissance humanism, the phenomenon of the Enlightenment, and the century in between; and it cannot even be taken for granted that historians and philologists of that specific time share the same conceptual framework. Moreover, the earlier letter-writing culture of the high Middle Ages and the later one, after the transitional period around 1800, also shares more features with the early modern republic of letters than the division in academic sub-chronologies suggests.

Another myopic affliction is the orientation of much scholarship on the republic of letters westwards. The study of the republic of letters is dominated by historians of England, the Dutch Republic, France, and increasingly Germany.[12] But all

[10] Hans Bots, *Republiek der letteren. Ideaal en werkelijkheid* (Amsterdam: APA-Holland University Press, 1977); Lorraine Daston, 'The Ideal and Reality of the Republic of Letters in the Enlightenment', *Science in Context* 4 (1991): 367–86, see https://doi.org/10.1017/S0269889700001010.

[11] Hans Bots and Françoise Waquet, eds., *Commercium litterarium. La Communication dans la république des lettres. Forms of Communications in the Republic of Letters, 1600–1750* (Amsterdam and Maarssen: APA-Holland University Press, 1994); Hans Bots, *De republiek der letteren. De Europese intellectuele wereld, 1500–1760* (Nijmegen: Vantilt, 2018), 36–8.

[12] For instance, Hans Bots, *De republiek der letteren. De Europese intellectuele wereld, 1500–1760* (Nijmegen: Vantilt, 2018); Hans Bots and Francoise Waquet, *La République des lettres* (Paris: Belin, 1997); Marc Fumaroli, *La République des lettres* (Paris: Gallimard, 2015); Martin Mulsow, *Die unanständige Gelehrtenrepublik. Wissen, Libertinage und Kommunikation in der frühen Neuzeit* (Stuttgart and Weimar: J. B. Metzler, 2007); Herbert Jaumann, *Die europäische Gelehrtenrepublik im Zeitalter des Konfessionalismus* (Wiesbaden: Harrassowitz, 2001); Conrad Wiedmann and Sebastian Neumeister, eds., *Res publica litteraria: Die Institutionen der Gelehrsamkeit in der frühen Neuzeit* (Wiesbaden: Harrassowitz, 1987); most American authors also focus on this western European history: Anne Goldgar, *Impolite Learning. Conduct and*

the wonderful research into scholarly networks that is conducted today in Portugal, Spain, and even Italy, not to mention eastern European and Scandinavian territories as well as other parts of the world, needs to be integrated into a new picture.[13] Rather than attempting to come to a universally applicable and generally acceptable definition of what the republic of letters entailed, it is wiser to acknowledge that early modern scholars themselves failed to agree, as much as modern historians do. In other words, we need to accept the multiplicity of meanings of the republic of letters and its varying reaches across time and space.

Such multiplicity should prevent us from ascribing certain programmes to a reified and idealized republic of letters, and from using those ideals as criteria for modern scholars to decide who was in and out of the republic of letters. Admittedly, there is a tendency to characterize the republic of letters as a tolerant community bent on neutrality in religious and political issues and epistemic humility when it comes to truth claims. And while the equivalence between the republic of letters and tolerant or even radical thought was promoted by some of the latter's proponents, especially in the eighteenth century, others continued to use the term in a highly generic way, and without any revolutionary or even 'enlightened' aspiration. In fact, few of those scholars ever cared to make clear what they meant by the term, in particular before the eighteenth century.

A further barrier to synthetic understanding is the disciplinary separation of this huge field into linguistic domains, with Neolatinists, Romanists, and Italianists each focusing on discrete languages in a world that itself was multilingual. Fortunately, manifestos advocating for new *longue durée* accounts, geographical inclusiveness, and translation studies have made an impact, not to mention the growing attention to the agency of female letter-writers and learned women. Although studies of the black, Muslim, Jewish, or queer republic of letters are probably far away, the study of the republic of letters has ceased to be the exclusive domain of histo-

Community in the Republic of Letters, 1680–1750 (New Haven and London: Yale University Press, 1995); Dena Goodman, *The Republic of Letters. A Cultural History of the French Enlightenment* (Ithaca and London: Cornell University Press, 1994); April Shelford, *Transforming the Republic of Letters: Pierre-Daniel Huet and European Intellectual Life 1650–1720* (Rochester, NY: University of Rochester Press, 2007); Carol Pal, *Republic of Women. Rethinking the Republic of Letters in the Seventeenth Century* (Cambridge: Cambridge University Press, 2012).

[13] For instance Gábor Almási, *The Uses of Humanism: Johannes Sambucus (1531–1584), Andreas Dudith (1533–1589), and the Republic of Letters in East Central Europe* (Leiden: Brill, 2009); Maria Berbeara and Karl Enenkel, eds., *Portuguese Humanism and the Republic of Letters* (Leiden and Boston: Brill, 2012); Caroline Winterer, 'Where is America in the Republic of Letters?', *Modern Intellectual History* 9:3 (2012): 597–623, see https://doi.org/10.1017/S1479244312000212. Beyond the West: Min Chŏng, *18세기 한중 지식인의 예공화국* [The Republic of Letters of Korean and Chinese Intellectuals in the Eighteenth Century] (Kyŏnggi-do P'aju-si: Munhak Tongne, 2014); Muhsin J. al-Musawi, *The Medieval Islamic Republic of Letters: Arabic Knowledge Construction* (Notre Dame, IN: University of Notre Dame Press, 2015).

rians of scholarship, poetry, linguistics, and religion and is attracting new sociologically minded historians.[14]

Indeed, what can be easily shared across disciplines is the data referring to the republic of letters. Its long history, its broad geography, its complex stratification, and its multilingual media will be amendable to systematic study once technical systems and scholarly cultures are adapted to allow the exchange of unprecedented numbers of records, texts, and associated data. This is the reason that the subject has also piqued the interest of digital humanists, as this volume demonstrates. For in order to come to a chronologically, geographically, socially, and linguistically inclusive big picture of the republic of letters, one which also reflects its broader demographic, economic, environmental, and geopolitical context, we need to pool existing knowledge about the republic of letters, identify the gaps, and set the agenda for filling these *Forschungsdesiderate*. Such an approach relies on large data sets and requires digital instruments and environments.[15]

Here is where academic variety poses challenges. Different disciplines use historical data in different ways. They draw on different types of primary sources and cultivate different methodological traditions. This variety is replicated in the digital world. Network analysis works with quantitative methods rooted in the social sciences. Digital philology works with automated ways of morpho-syntactic tagging and a quantitative version of stylistic analysis. On the other hand, some of the more literary or historical approaches will focus on the semantics of individual words or texts, often in a qualitative way. Not only should digital humanists be aware of their disciplinary traditions; they should also be prepared to translate them into the digital domain, and to see them transform and converge. This urgent need for standardization, transformation, and convergence is the *raison d'être* of this volume. Organized in terms of the different dimensions of that data, this volume sketches the digital conditions to make such a cooperation possible.

Before we embark on an analysis of the problems and take stock of the solutions developed so far, however, we will provide a chronological overview of the history of the republic of letters. This chronological outline, admittedly generalizing in its attempt to identity a series of intellectual turning points, introduces us to some important figures, topics, and tendencies. Together, the story provides a frame for positioning the digital work for which this volume sets the agenda.

[14] Rens Bod, *A New History of the Humanities. The Search for Principles and Patterns from Antiquity to the Present* (Oxford: Oxford University Press, 2013); Peter Burke, *What Is the History of Knowledge?* (Cambridge: Polity Press, 2016).

[15] A rough estimate is that between 1 and 2 million scholarly letters survive from the early modern period scattered over hundreds of libraries and archives in and outside Europe; see III.1.

2 A Chronological Survey

This chronological survey is structured by means of a conventional periodization into 'centuries'.[16] Although it attributes to each century a particular label, these characterizations are by no means exclusive. Perhaps they were not even dominant in the periods described. Yet they do indicate new developments that attracted attention, while established modes of research were not abandoned.

2.1 The Fifteenth Century: Revitalizing Ancient Literature

Our first recorded use of the expression *respublica litteraria* dates from the early fifteenth century. In 1417, the Italian humanist Francesco Barbaro wrote a long letter to his colleague Poggio Bracciolini, praising him for his many discoveries of manuscripts with new texts of ancient Roman authors. The carefully crafted letter, obviously meant for a larger public than just the recipient, bestows on Poggio the early modern equivalent of a lifetime achievement award for 'bringing to this Republic of Letters the largest number of aids and equipments': he wanted Poggio's 'immortal merits to be placed in the light of Europe'.[17]

After that one incidental blip on the historians' radar, the term remains hitherto unrecorded for most of the fifteenth century. But during the course of that century, Italian humanists frequently idealized learned communality by such terms as the society of the learned (*societas litteratorum*), the erudite world (*orbis eruditus*), or the fellowship of letters (*sodalitas litteraria*).[18] Other equivalents were the 'learned world' and 'world of the learned' (*orbis litterarius, orbis litteratorum*). The rector of the Sorbonne in 1470 used the term 'coetus doctorum hominum'.[19] In book titles we come across 'the learned state' (c. 1462: 'politia literaria').[20] The Frisian scholar Rudolph Agricola, writing from Heidelberg, speaks in a letter from 1584 to Reuchlin about serving the 'literaria respublica patriae nostrae'.[21] Harking back to Christine de Pisan's *Cité des dames* of 1405, the Italian humanist Laura Cereta in 1488 described the historical record of learned women as a 'muliebris respublica'.[22] A

[16] Parts of the following text have been published in: Dirk van Miert, 'What Was the Republic of Letters? A Brief Introduction to a Long History (1417–2008)', *Groniek* 205:4 (2014): 69–87, see https://ugp.rug.nl/groniek/article/view/27601/25014.

[17] Marc Fumaroli, 'The Republic of Letters', *Diogenes* 143 (1988): 129–54, see https://doi.org/10.1177/039219218803614307. [Franceso Barbaro], *Fr. Barbari et aliorum ad ipsum epistolae* (Brescia: Rizzardi, 1743), 5.

[18] Fritz Schalk, 'Von Erasmus' *Res publica literaria* zur Gelehrtenrepublik der Aufklärung', in Fritz Schalk, *Studien zur französischen Aufklärung* (Frankfurt am Main: Vittorio Klostermann, 1977), 143–63.

[19] Hans Bots and Françoise Waquet, *La repubblica delle lettere* (Bologna: Mulino, 2005), 19.

[20] Angelo Camillo Decembrio, *De politia litteraria*, c. 1462. (Augsburg: Steyner, 1540; Basle: Hervagius, 1562).

[21] Agricola (Heidelberg) to Johann Reuchlin, 9 November 1584: Rudolph Agricola, *Letters*, ed. and tr. with notes by Adrie van der Laan and Fokke Akkerman (Assen: Van Gorcum, 2002), 230, line 3 (letter no. 41).

[22] Laura Cereta, *Epistolae iam primum e MS in lucem productae*, ed. Jacobus Philippus Tommasinus (Padua: Sebastianus Sardus, 1640), 195. The notion was taken up by Carol Pal as a label for the female

year later, in 1489, the expression *respublica litteraria* appears for the first time in a book title.[23] In 1491, an editor of Bonaventura's work (Nuremberg: Koburger) speaks of people who never contribute to the *litteraria respublica*. The humanist Conrad Celtis mentioned the expression in an oration of 1492, and in 1498 the cartographer Johannes Stabius called Ingolstad a 'litteraria respublica'.[24]

A particular popular ideal in the fifteenth century was the learned conversation between a handful of friends gathered round a dinner table in some villa in the countryside – a setting reminiscent of the ancient Greek philosophical symposium.[25] If friends could not meet, the letter acted as a medium for long-distance communication. Already in Antiquity, correspondence was conceptualized as a dialogue between absent friends.[26] The expression 'republic of letters' reappeared in 1494, in a seminal work by Desiderius Erasmus against ignorance of intellectual culture, the *Anti-barbari*, but the book was only published in 1520.[27] The restless traveller Erasmus, who lived by the motto 'I wish to be a citizen of the world', was the exemplary spider in a pan-European epistolary web of learning,[28] and his contemporaries frequently employed the terms *respublica litteraria* and *respublica litterarum*.[29]

After Martin Luther's break with Rome in the decades after 1517, the term took on renewed significance, because it could act as an alternative to the idea of the pan-European *respublica Christiana*, which now lay in tatters. Latin remained as the *lingua franca* of this somewhat stripped-down ideal of unified *secular* learning in Europe, although Italian was often employed in the Italian peninsula, and similarly French in the kingdom of France and Spanish on the Iberian peninsula. In the

community within the seventeenth-century republic of letters: Carol Pal, *The Republic of Women. Rethinking the Republic of Letters in the Seventeenth Century* (Cambridge: Cambridge University Press, 2012).

[23] [Aelius Donatus], *Aelij Donati Grammatici, Pro impetrando ad rempub. litterariam aditu: novitijs adolescentibus grammatices rudimenta quam aptissime dedicate* (Venice: per Theodorum de Ragazonibus, tercio kalendas mensis decembris (29 November) 1489 – no original copy located, but according to WorldCat, Wroclaw University Library holds a Xerox copy); also: Geneva: per Johannem de Stalle Allemannum, 15 May 1493 (University Library Basle, UBH DE VI 24:5). Bots and Waquet, *La repubblica delle lettere*, 12–3, refer only to an edition of Venice 1491, but give no bibliographical details.

[24] Bots and Waquet, *La repubblica delle lettere*, 13.

[25] Schalk, 'Von Erasmus' *Res publica literaria* zur Gelehrtenrepublik der Aufklärung'.

[26] For a brief overview of the theory of letter writing, see Dirk van Miert, 'Letters, and Epistolography', in Anthony Grafton, Glenn Most, and Salvatore Settis, eds., *The Classical Tradition* (Cambridge, MA: Harvard University Press, 2010), 520–3.

[27] *Antibarbarorum liber unus* (Basle: J. Frobenius, 1520), 41: 'Alii enim literariam Rempu[blicam] tanquam funditus deletam cupiunt; alii imperium non quidem prorsus extinguere, sed arctioribus finibus includere moliuntur. Postremi ita Remp[ublicam] salvam esse volunt, ut afflictissimam velint, quippe in qua ipsi tyrannidem occupent.'

[28] Lisa Jardine, *Erasmus, Man of Letters: The Construction of Charisma in Print* (Princeton, NJ: Princeton University Press, 1993). See also the work of Christoph Kudella (e.g. 'The Correspondence Network of Erasmus of Rotterdam. A Data-driven Exploration', Unpublished PhD Thesis, University College Cork, 2017).

[29] Constance M. Furey, *Erasmus, Contarini, and the Religious Republic of Letters* (Cambridge: Cambridge University Press, 2006); Hanan Yoran, *Between Utopia and Dystopia. Erasmus, Thomas More, and the Humanist Republic of Letters* (Plymouth: Lexington Books, 2010).

fifteenth century, sociable Italian humanists formed local or regional pockets with a European outreach. They largely engaged with classical philology and ancient history, with Greek as the new fashion, in particular after the conquest of Constantinople by the Ottomans in 1453. This event led to the appearance in Italy of Greek scholars who brought knowledge and manuscripts. When the Italian Renaissance started to spread beyond the Alps in the second half of the century, the learned networks became wider and more interconnected.

2.2 The Sixteenth Century: The Turn to Christian History

In the sixteenth century the citizens of the republic of letters turned to the Bible and the important Christian authors of late Antiquity, known as the Church Fathers (Augustine, Ambrose, Jerome, Basil, among others). In their study of the interplay of Greek, Roman, and Christian Antiquity, humanists relied not only on texts, but increasingly on material culture: the remnants of Antiquity which might anachronistically be called archaeological objects: coins, statues, and monuments, often showing images and inscriptions. The study of religious practice in historical and cultural perspectives also started to pave the way for a view that saw Christianity in its relation to other religions, and thus as one religion among others.

However, the exchange of knowledge was not limited to such traditional themes and objects. Increasingly, scholars communicated about the results of medical experiments and sent each other recipes. While some turned to the skies to see if what they observed confirmed what ancient authors had written about the stars, planets, and comets, others bowed their noses to the ground in search of plants and flowers, in an attempt to match them with the herbs described by Greek and Roman authors. It was through intense cooperation and exchange of letters, in and beyond Europe, that it dawned upon the citizens of the republic of letters that the ancients had not been omniscient. Apart from letters, the speed of information exchange and the comparison of data was significantly accelerated by the development of the printing press. In 1439, Johannes Gutenberg had first used movable types in Mainz, and after 1500 the printing press started to conquer Europe.

Also due to the increasing amount of information circulating,[30] from the second half of the sixteenth century onwards, specialized networks started to develop, focusing, for example, on astronomy or botany, antiquarianism or theology. Most of the erudites were both scholar and scientist, theologian and philosopher, but many felt they had to take a more special interest in one of these subjects.

The wars of religion, which scourged France, the Low Countries, and the Holy Roman Empire (i.e. 'Germany') in the half-centuries before and after 1600, sparked off endemic controversies over the history of the early Church, the authority of the Church Fathers, and the good and the bad parts of ancient philoso-

[30] Ann M. Blair, *Too Much to Know. Managing Scholarly Information before the Modern Age* (New Haven and London: Yale University Press, 2010).

phies, new confessions and future politics: the more or less distant past became a yardstick for the present. This intense bickering created a buzz, which drew ever more people into the realm of the republic of letters and made them participate in both the cult of communication and the practice of polemic. Wars were not only fought on the battlefields, but also in the public and semi-public fora of intellectual exchange: letters, pamphlets, book, speeches, and other kinds of academic texts that would allow scholars to take stances, or create polemical distance. Not only knowledge in the fields of polemical theology or historical jurisprudence was required; also the boundaries between the sciences, the arts, and the technical disciplines were blurred: Leonardo da Vinci offering his services as a military engineer to Ludovico il Moro, duke of Milan, is an example for this.

However, political and religious difference brought scholars and scientists into conflict not only with one other but also with legal and ecclesiastical authorities. While in the late Middle Ages heresy was still considered to be a theological matter, the growing complexity and interconnection of fields of knowledge in the sixteenth century led to an extension of censorial measures far into the spheres of natural philosophy and history. The 'index of forbidden books', established by the Roman curia in 1559, is perhaps the most prominent example, but should not disguise the fact that censorship was also exercised by secular authorities, and not only in Catholic parts of the world.

While, thus, on the one hand scholars with unorthodox ideas had to fear persecution, and possibly also face the stake (like the philosopher Giordano Bruno, executed in Rome in 1600), on the other hand the pan-European dimension of the republic of letters was not forgotten: scholars still communicated across religious, political, and ideological boundaries. In order to gain access to much coveted information, it was wise to ignore such hot-button issues and pretend that all learned people were working towards the same goal of universal peace and justice – and by so doing, many in fact advanced precisely that ideal.[31]

2.3 The Seventeenth Century: The Rise of Natural Science

Significant portions of the sixteenth and the early seventeenth centuries saw Europe engaged in brutal wars with religious underpinnings. Among the most important episodes we may evoke the Edict of Nantes (1598), granting rights to the French Calvinists (Huguenots) after several decades of ferocious civil war, before being revoked in 1685, which forced Huguenots into exile again; the Eighty Years' War (1568–1648), during which the mainly Protestant Netherlands successfully fought for their independence from the Catholic Spanish crown; and the Thirty Years' War (1618–48), leading to a confessionally divided Holy Roman Empire

[31] Jeanine De Landtsheer and Henk Nellen, eds., *Between Scylla and Charybdis. Learned Letter Writers Navigating the Reefs of Religious and Political Controversy in Early Modern Europe* (Leiden and Boston: Brill, 2011).

and, in the mid-term, to a French supremacy in Europe (especially after the end of the war with Spain in 1659). In the Baltic region, Sweden, Russia, and Prussia emerged, while the previous Ottoman expansion in the Balkans came to a stop, yielding to a Counter-Reformation conquest of large parts of Hungary. Overseas empires were held by Spain, Portugal, France, England, and the Netherlands, and they often drew on the infrastructure of either private enterprises (e.g. the East India Company) or religious orders (e.g. the Jesuits).

These circumstances had repercussions on the republic of letters not only in that they framed the lives, movements, and, eventually, deaths of its personnel. They are also responsible for some of the data peculiarities we will be dealing with in the following chapters: the significance of choosing one calendar system over another (ch. II.3), the political meaning of place names in different languages (ch. II.2), or the use of pseudonyms in order to evade censorship (ch. II.4).

Around the republic of letters itself, however, not only the political landscape changed, but also the framework of institutions and media, and with them changed the learned and scientific priorities. Until halfway through the seventeenth century, scholars had relied almost exclusively on letters and books for news about the world of learning. This changed after the arrival, in 1665, of the first learned journals. The journals appeared in Paris and London and were published by newly established academies. These academies (among others, the Académie des Sciences and the Royal Society of London) were loosely modelled on the fifteenth- and sixteenth-century Italian societies of humanists and learned courtiers, but were now sponsored or supported by bearers of political power, who hoped to profit from the results of the rapidly transforming and increasing research into the natural world.

In terms of learned content, the natural sciences now emancipated themselves from the influence of Aristotelian thought, in which natural history, natural philosophy, and physics had been neatly organized into a metaphysical framework. Openly criticizing previous traditions, but in fact heavily indebted to them, Francis Bacon (d. 1626) and René Descartes (d. 1650) inspired new generations of both radicals thinkers and pious observers of God's creation to reread the Bible and the Book of Nature in the light of new philosophical and historical frameworks.[32] Mathematics and the cosmos, cometology and meteorology, motion and matter, atomism and the vacuum – such themes occupied many of the greatest minds of the century. Yet other brilliant intellectuals, such as the earlier mentioned Calvinist scholar Joseph Scaliger, the Church of Ireland archbishop James Ussher, and the Catholic theologian Dénis Petau, continued to deepen the insight into human and biblical history by developing comparative chronologies, while Jesuit scholars such as Athanasius Kircher studied Asian history and Egyptian antiquities (along with

[32] Eric Jorink, *Reading the Book of Nature in the Dutch Golden Age, 1575–1715* (Leiden and Boston: Brill, 2010).

magnetism):[33] the world deepened chronologically, expanded geographically, and grew more complicated mathematically, while becoming conceivable for the first time in Western history without the idea of the biblical God, thanks to the Jewish-Dutch philosopher Baruch de Spinoza.

Whatever the field of knowledge may have been, however, there had always been some implicit consensus about the duty of scholars to communicate and respect the authorial rights of one's colleagues, as appears from the many polemics arising from the infringement of such codes. Yet, from the late seventeenth century onwards, the republicans of letters consciously reflected on the ideals of tolerance, and on the codes of conduct, which should structure the republic of learning.[34] At the same time, the medium of print periodicals (with textual genres directly inspired by correspondence) reached growing parts of the increasingly literate public, so that the overspill of learned ideals into the sphere of political theory may be one of the major reasons for the perceived intersection between the republic of letters and the Enlightenment.

2.4 The Eighteenth Century: The Philosophical Republic of Letters

After the classical turn of the fifteenth century, the ecclesiastical and biblical turn of the sixteenth century, and the natural scientific turn of the seventeenth century, the republic of letters experienced a philosophical turn in the eighteenth century, when obsessive letter-writers such as Gottfried Wilhelm Leibniz (d. 1716) were relieved by no less maniacal correspondents such as Voltaire (d. 1778).

For many, the French enlightened republic of letters was too dismissive of the study of the text and remnants of Antiquity to count as truly learned, but the new generation in fact built upon the accomplishments of the humanists, scholars, and scientists of the Renaissance. Denis Diderot and Jean le Rond d'Alembert in their famous *Encyclopédie* (1751) ransacked not only recent predecessors such as the 'critical' histories of Pierre Bayle (1697) and Jacob Brucker (d. 1744), but also the seventeenth-century journals and the sixteenth-century commentaries on pagan and Christian texts from Antiquity and the early Middle Ages. However, they did not dwell on them for mere historical interest; instead, they critically assessed their thinking against the philosophical systems of their day.

Philosophers were not the only citizens of the republic of letters. Take for example the famous botanist Carl Linnaeus (d. 1778). He corresponded with about

[33] Anthony Grafton, 'A Sketch Map of a Lost Continent: The Republic of Letters', *Republics of Letters* 1:1 (2008). See http://arcade.stanford.edu/rofl/sketch-map-lost-continent-republic-letters, accessed 20/03/2019.

[34] Dena Goodman, *The Republic of Letters. A Cultural History of the French Enlightenment* (Ithaca and London: Cornell University Press, 1994); Anne Goldgar, *Impolite Learning. Conduct and Community in the Republic of Letters, 1680–1750* (New Haven and London: Yale University Press, 1995); Martin Mulsow, *Die unanständige Gelehrtenrepublik. Wissen, Libertinage und Kommunikation in der frühen Neuzeit* (Stuttgart and Weimar: J. B. Metzler, 2007); Sari Kivistö, *The Vices of Learning. Morality and Knowing at Early Modern Universities* (Leiden and Boston: Brill, 2014).

200 people within Sweden and twice as many from other countries, in Europe but also Asia and Africa. About 3,000 of the letters, which he received from 660 people, have survived. On top of that, there are about as many letters that he wrote himself and sent off.[35] The statesman and natural philosopher Benjamin Franklin (d. 1790) was also an avid letter-writer: a constant flow of letters from and to him crossed the Atlantic. His surviving correspondence numbers some 15,000 letters. In the letters he addressed politics and electricity, naturally, but they also show that Franklin was interested in subjects ranging from meteorology to morality. And if we look at the greatest philosopher of the century, Immanuel Kant (d. 1804), the creator of modern ethics and metaphysics, we observe that he, too, exercised the praxis typical of a citizen of the republic of letters. He kept up an extensive and lively correspondence, which was inscribed in the learned sociability characteristic of the republic of letters. He composed poems for visiting scholars and funeral poetry for deceased colleagues, and he took great pleasure in learned conversation at the dinner table. He received many students who called on him with letters of recommendation, often from scholars who merely used those students as an excuse to address Kant. However, whereas the philosopher Johann Gottlieb Fichte (d. 1814) still spoke of a *literary republic of the learned*, Kant never once mentioned the concept of the republic of letters in his writings. Is this symptomatic for the diminishing currency of the metaphor?[36]

On the level of European politics, the apparent decline of the republic of letters may be related to the rise of political entities – states – that were in need of expert knowledge in order to legitimize their authority and to exert power on a calculable basis. All fields of knowledge were involved, and universities as well as academies now consolidated the findings of past centuries, transforming them into curricula, or applying them to the agenda of state building (e.g. by creating land registers).[37] These institutions were often framed in a national paradigm, and referenced a specific state: knowledge became 'French' or 'English', and no longer related to the transnational commonwealth of learning.

To be sure, national stereotypes were not new in the republic of letters, but during the eighteenth century, knowledge became one of the fuels of nation-based statehood. Latin as a *lingua franca* was abandoned definitively, and although the republic of letters after 1800 persisted in the form of a 'republic of *belles-lettres*' and,

[35] For the edition of Linnaeus's correspondence, see http://linnaeus.c18.net/Doc/lbio.php, accessed 20/03/2019.

[36] Dirk van Miert, 'Immanuel Kant and the Republic of Letters', paper read at Oxford, *Annual Meeting of the British Society for Eighteenth-Century Studies*, 8 January 2015; Kasper Risbjerg Eskildsen, 'How Germany Left the Republic of Letters', *Journal of the History of Ideas* 65:3 (2004): 421–32, see https://doi.org/10.1353/jhi.2005.0004.

[37] William Clark, *Academic Charisma and the Origins of the Research University* (Chicago and London: The University of Chicago Press, 2006).

eventually, the 'scientific community' of our day, its prevailing reference became the respective academic culture of a state, a nation, or an empire.[38]

[38] For the survival of the republic of letters in the modern age see Peter Burke, 'The Republic of Letters as a Communication System', *Media History* 18:3–4 (2012): 395–407, see https://doi.org/10.1080/13688804.2012.721956.

I.3 How Do We Model the Republic of Letters?

Christoph Kudella

With contributions from Neil Jefferies

1 Data in the Humanities

The term 'data' is not traditionally used by humanities scholars to describe the objects of their studies. Yet, with the rise of computational methods and their employment for new lines of enquiry in the humanities over the last few decades, humanities scholars are now engaging with 'data' at an exponentially increasing rate.[1] Fortunately, this new mode of research – generally associated with the term 'digital humanities' or cognate terms – exhibits a comparatively high level of epistemological self-reflection, as Christine Borgman has shown.[2] This brief introduction to data, and to the challenges of modelling the data pertaining to the republic of letters, is intended primarily to ease scholars in the humanities into the more detailed discussions of data models that follow, especially in section II of this book.

One of the broadest and yet most concise definitions of data in the humanities was devised by Christof Schöch: 'Data in the humanities could be considered a digital, selectively constructed, machine-actionable abstraction representing some

[1] See Christof Schöch, 'Big? Smart? Clean? Messy? Data in the Humanities', *Journal of Digital Humanities* 2:3 (2013), see http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/, accessed 20/03/2019.

[2] See Christine Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet* (Cambridge; London: MIT Press, 2010).

aspects of a given object of humanistic inquiry'.[3] As this definition makes clear, data cannot be identified with documentation or even with information: it is an 'abstraction' derived from documentation and reformulated in a manner which can be processed by computers. Moreover, as this definition also suggests, one of the peculiarities of humanistic data is the consciousness that it has been 'selectively constructed'.[4] Data, in other words, is not merely latent in the historical documentation but has been constructed by a series of decisions made by the researcher who has extracted it from the documentation. What humanistic data 'looks' like, that is, its structure and content, is shaped by the researcher's choice of source material, data model, and technology, by decisions regarding the use of external resources, and much more. These decisions in turn are determined by the kind of analyses that the researcher intends to conduct; and they likewise influence the analyses which can subsequently be carried out based on the data, and therefore also their reusability and interoperability.

Since data in the humanities are present in every step of the digital research process and also determine that process in many respects, it is of the utmost importance to model data very carefully at the outset. Data models define unambiguously and explicitly the shape in which the data is created and stored and enable information systems to process queries. Explicit data models are also required by third parties needing to understand the structure of a data set prior to re-use as a condition of interoperability. Approaches to data modelling can be distinguished from one another in a variety of ways. Perhaps the most important distinction with regard to the subject of this book is that between curation-driven and research-driven data modelling.

Curation-driven data modelling is the approach taken by libraries, archives, and museums. Its principal purpose is to catalogue holdings according to the standards of the respective institutional community. Research-driven data modelling organizes data with a view to answering a particular research question. Put simply, the contrast between these two approaches is that of standardization and simplicity versus granularity and nuance. The contrasts of these two approaches often prevent data reuse: data created for cataloguing purposes may lack many of the features necessary for answering research question, while data collected using a research-driven approach can sometimes not easily be integrated into the catalogues of libraries and archives, both because essential information is lacking and because data is captured in non-standard ways.

In addition to these different objectives and styles of data generation are differences in the methods of information organization used and still deeper differences in information technology. These underlying differences cannot be dealt

[3] See Schöch, 'Big? Smart? Clean? Messy?'.

[4] See Johanna Drucker, 'Humanities Approaches to Graphical Display', *Digital Humanities Quarterly* 5:5 (2011), see http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html, accessed 20/03/2019.

with exhaustively in this brief introduction, but examples of them will be provided at suitable points in the following discussion.

2 Modelling the Republic of Letters

As already pointed out in chapters I.1–2, the republic of letters was more than just a network of correspondences. This imagined community was bound together by networks of verbal discourse (in universities, academies, coffee houses, and salons), scribal communication (via letters and other manuscripts), printed publication (of journal articles as well as books), and the exchange of physical objects (such as natural historical specimens and cultural artefacts), as well as many different kinds of scholarly travel. Given the interaction of so many different kinds of entities, research in this field clearly cannot begin simply by adopting a single, pre-existing comprehensive data model covering this whole field. Existing models for letters, for instance, handle very different information in very different ways from prosopographical data records describing the life of a person in a structured fashion; and even when standards already exist for these different domains, they cannot easily be reconciled. The only way forward is therefore to devise a set of interoperable data models, drawing on existing standards where they are available and creating new ones where necessary.

Given the complexity of the republic of letters as a whole, it makes sense to focus initially on the more limited task of the modelling of letters. The exchange of letters is one of the most characteristic activities of the *respublica litteraria*, one of the best documented, and also one of the most complex. As a consequence, devising a data model for letters raises many of the fundamental problems pertaining both to other aspects of the republic of letters and to modelling historical data and the process of research data creation in the humanities in general.

Before modelling can be carried out, the underlying material must first be localized and the phenomena it documents analysed. With regard to the letters, the problem of discoverability therefore arises first. As noted in chapter I.2, an estimated 1–2 million early modern letters are scattered throughout public and private libraries and archives in Europe and beyond. Many remain uncatalogued. Others have been 'catalogued' in very different ways: to note only the most salient point (further elaborated in ch. III.1, sect. 3.1), librarians and archivists use quite different standards and instruments for describing their collections. At the national level, some countries have attempted to assemble existing catalogue data into analogue or digital union catalogues, such as *Kalliope*[5] (see further ch. III.1, sect. 3.2), which enhance discoverability but compound the problem of data heterogeneity. As suggested above, still others have been inventoried by projects with unusual, research-driven data models designed to answer different questions.

[5] See http://kalliope-verbund.info/en/index.html, accessed 20/03/2019.

Even when standardized data is easily discoverable, however, they often do not meet all the requirements of researchers. To begin with, collection-level catalogue records may not contain information about individual letters (see further in ch. III.1, sect. 3.3). Even when letters are catalogued individually, even the most basic metadata – recording sender and recipient as well as the places and dates of sending and receipt – are often unrecorded, missing, or unreliable; nor is sufficient distinction always made between the many forms in which a letter can be preserved, such as drafts, scribal copies, and altered versions as well as letters actually sent (see further ch. II.7). One of the principal modes of preserving early modern correspondence is in printed letter collections or 'epistolaries' as well as in modern scholarly editions (on which see ch. III.1, sect. 2). Here too, comprehensive indexes of these sources are lacking, and when relevant printed collections are found no concordances typically exist to reconcile material preserved in archival holdings on the one hand, and printed letter collections and editions on the other.

Once sufficient quantities of material have been collected, the essential requirements for a data model for letters can be derived from an analysis of the phenomena one encounters within them. At this point, there are several approaches to modelling this information, as well as various technologies for storing and processing the corresponding data.[6] Three common approaches are presented here: the relational data model, the Text Encoding Initiative (TEI), and the Resource Description Framework (RDF).

The first modelling approach presented here is the relational model. In this approach, entities, their attributes, and the relationships between these entities are identified and represented. In a correspondence, three entities can be clearly distinguished from each other: 'letters', 'correspondents', and 'locations'. Between these entities, four distinct relationships exist, distinguished by their roles. One 'correspondent' can be the sender or recipient of many 'letters'. Likewise, one 'location' can be the origin or destination of many 'letters'. Thus, 'correspondents' and 'locations' have a one-to-many relationship to 'letters'. The relationships between these entities alone are not sufficient to identify a letter within a corpus or several corpora, since the same sequence of relationships might occur more than once. By adding additional attributes like a date specification to the 'letter' entity, a high – albeit not always conclusive – degree of identification at the level of the individual letter can be achieved by referring to the entities, their relationships, and attributes. This may sound abstract at first, but becomes understandable when one looks at how data collection, storage, and retrieval based on this model can work in an implementation of this model. In a Relational Database Management System (RDMS) such as MySQL, the corresponding data for the letters would be stored in at least three different tables: for letters, correspondents, and locations. Each row in the letters table would reference the sender and recipient of a letter by referencing

[6] For a broader introduction see Seth van Hooland and Ruben Verborgh, *Linked Data for Libraries: How to Clean, Link and Publish Your Metadata* (Chicago: Neal-Schuman, 2014).

entities in the correspondents table. In the same fashion, the location of dispatch or reception of a letter would be recorded in each row of the letters table by referencing the locations table (see fig. 1). This approach essentially corresponds to the tabular recording of letter metadata and is in principle not dissimilar to the structured lists of letters found in many editions of correspondence. In comparison to a purely tabular listing, the advantage of a relational model lies in the use of unique keys, which ensure unique, unambiguous referencing of entities.

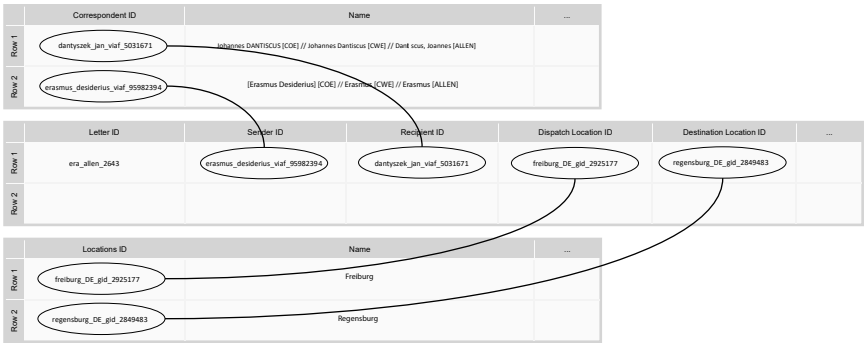


Figure 1: A potential (reduced) table structure of a relational database of correspondence metadata

The second modelling approach presented here originates from the field of digital editing, which is also dealt with more extensively in chapter III.3. Most digital editions today are based on XML[7] and follow a set of mark-up rules, the TEI-Guidelines,[8] developed since the late 1980s. The most striking difference between a database of correspondence metadata, as previously presented, and a digital edition, is of course that the latter also represents the entire letter content. What they have in common is the information which distinguishes an individual letter (often called 'metadata'):[9] the metadata which make up the bulk of a database must also be recorded in a structured manner in digital editions of correspondence. Methodologically, there is also a big difference: while in relational databases, the data is stored in linked tables, in TEI-XML-based editions, data is stored in a tree-like structure. The information about sender, recipient, location and date of dispatch are stored in the so-called TEI header of each letter in certain nested elements. The first element, <correspAction>, 'contains a structured description of the place, the

[7] The Extensible Markup Language (XML) is a markup language for encoding hierarchically structured data in a both human-readable and machine-readable format according to a defined set of rules.

[8] See http://www.tei-c.org/guidelines/, accessed 20/03/2019.

[9] In general, metadata provides information about other data. In the case of the data model for letters employed in this volume, the core metadata are the names of the sender and recipient, the places of sending and receipt, the date of sending, and the location of the manuscript of printed letter. Together these both describe key features of the letter and distinguish it from other letters. The data model is discussed in detail in chapter II.7.

name of a person/organization and the date related to the sending/receiving of a message or any other action related to the correspondence'.[10] The second element, <corrispContext>, 'provides references to preceding or following correspondence related to this piece of correspondence',[11] thereby providing a structural map of the relation of the epistolary item to two other items.

While the previous approaches are focused on the materiality and properties of individual letters, the primary aim of sharing data is to understand letters in relation to other letters and the people, places, and events that locate them in a broader historical and intellectual context. This opens the door to a large set of potential entities and inter-relationships that can be cumbersome to represent in the relational and hierarchical models already mentioned, and which are likely to expand over time as a result of scholarly activity. Of particular interest is the need to express differing and sometimes contradictory assertions that reflect the current state of knowledge based on evolving evidence.

These observations lead us towards graph representations of information, where any two entities can share a number of relationships which may be qualified in some way, limiting them in terms of time, location, or even evidentiary source, for example. This is a much more flexible approach that suits the more narrative, unstructured nature of humanistic information rather better than more rigidly designed hierarchies or tables. The Resource Description Framework (RDF),[12] along with related tools and protocols from the Semantic Web,[13] is the most commonly used graph-orientated framework, with the additional benefit that it is readily extensible and explicitly designed with networked information sharing in mind.

The basic construct in RDF is the 'triple' (see fig. 2), consisting of a subject and object (termed 'nodes'), and a predicate, a directional relationship between the two (termed a 'directed arc', or 'edge').

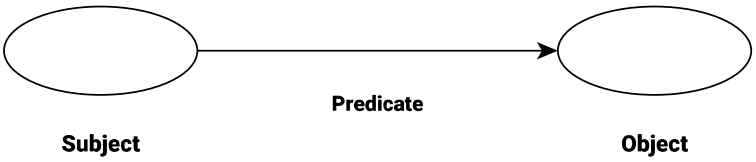


Figure 2: Schematic of an RDF triple

[10] See http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-corrispDesc.html, accessed 20/03/2019.

[11] Ibid.

[12] See https://www.w3.org/RDF/, accessed 20/03/2019.

[13] See https://www.w3.org/standards/semanticweb/, accessed 20/03/2019.

For example, a letter can be related to one person as the sender and to another person as the recipient. As more nodes and edges are added, an overall, network-like, 'graph' of objects and relationships emerges. By selecting a single node, such as a person, we can readily access all the related nodes – seeing their relationships to other people and institutions, letters sent and received, and mentions in other documents and letters.

Both nodes and arcs are (ideally) not just textual descriptions but persistent Uniform Resource Identifiers (URIs) for distinct online 'resources' that have attributes that define them and/or provide further information (also using RDF). A key aspect of RDF graphs are the ontologies that provide typologies and vocabularies for describing resources which are defined using either the RDF-Schema (RDFS)[14] or the Web Ontology Language (OWL).[15] A key strength of this approach is flexibility, in that the basic triple structure remains constant, and can be used as the basis for code development, while the expansion and extension is through additional ontology terms. In particular, in chapter II.6 we identify the existing Provenance Ontology (PROV-O)[16] as the basis for our event model, and in chapter II.7 it can be seen that the letter model is clearly graph-like, and hints at the existence of a potentially expansive set of roles (and consequent ontology) involved in the broader context of letters.

In practice, RDF, as defined here, is an abstract construct which can be serialized in a number of well-defined formats, such as JSON-LD[17] or RDF/XML,[18] depending on the use to which the data will be put and the preferences of the coder. This also allows a great deal of flexibility when converting legacy data to RDF since the most convenient format can be chosen. Once data is available in one format, code is readily available to translate between these formats as required.

3 Modelling People

In order to reconstruct the republic of letters as a network of correspondences, one of the most fundamental requirements is to identify the sender and recipient of the individual letters uniquely and unambiguously. While this may at first sound trivial, the difficulty of this task in fact presents a major obstacle to progress in this field.

The names of persons that one encounters in early modern manuscript and printed letters are usually not standardized: each sender denotes the recipient as they see fit. Given that different people share the same name, the same person can use different names and titles, and that learned exchange also abounds in pseudo-

[14] See https://www.w3.org/2001/sw/wiki/RDFS, accessed 20/03/2019.
[15] See https://www.w3.org/OWL/, accessed 20/03/2019.
[16] See https://www.w3.org/TR/prov-o/, accessed 20/03/2019.
[17] See https://json-ld.org/, accessed 20/03/2019.
[18] See https://www.w3.org/TR/rdf-syntax-grammar/, accessed 20/03/2019.

nyms, pen names, and learned named forms, the nature of the difficulty begins to become apparent. A data model for letter metadata must enable such uncertainties and ambiguities to be recorded and, when possible, resolved. Doing so serves two purposes: first, distinguishing the letter metadata created by the researcher from the historical documentation from which it derives makes explicit the process of inference undertaken by the researcher; and second, a granular recording of uncertainty and ambiguity makes it possible subsequently to filter out uncertain or ambiguous data according to one's own research purpose.

Reconciling the multiple ways of referring to the same person requires an unambiguous point of reference. In modern printed editions of correspondence, this disambiguation is usually provided by a footnote in which brief biographical information about the person is presented along with an assertion that a correspondent referred to in a variety of ways is one and the same person. Such discursive footnotes, however, are not readily processed computationally. The system of authority files that has been developed over decades in library and information science, providing unique alphanumeric identifiers for persons and other entities, supplies solutions for this problem.[19] Associating a range of different personal names and titles with such an identifier renders research data of different provenance (such as several letter corpora) interoperable with each other. In concrete terms, this means that an information system is able to identify a person as being identical by means of an external ID, irrespective of the particular specifications of a person's name in different sets of research data.

In the previously mentioned relational database, a corresponding unique authority-file ID could be recorded in a column in the table 'correspondents' for each entry. In a TEI-based capture of letter metadata, the same ID could be listed in the @ref attribute in the respective sub-elements of the <correspDesc> element, i.e. <persName>. When using RDF, the sender or receiver nodes would have one or more additional relationships linking them to various name authority identifiers. One key advantage of the RDF approach over TEI is that identifiers could be added to person nodes after the event without requiring additional editing of any letter records.

Of equal importance is the requirement of being able to record inferences and uncertainty regarding the sender or recipient of a letter. It is precisely this aspect that constitutes the granularity of research-driven data models that is often lacking in curation-driven data modelling and capture. In a relational database, such information could be recorded in corresponding columns in the table 'letters', for example by using Boolean operators that denote if the sender or recipient has been inferred by the person capturing the data or a previous editor, and whether this identification is uncertain. For recording such information in TEI, the situation is

[19] E.g. on a national and/or language-based level such as the Integrated Authority File (GND, https://www.dnb.de/EN/Standardisierung/GND/gnd_node.html), oriented on subject-domains such as CERL thesaurus (https://data.cerl.org/thesaurus/_search), or as virtual aggregation such as the Virtual International Authority File (VIAF, https://viaf.org/); all accessed 20/03/2019.

currently still inconsistent: a qualification of the information by the @cert attribute is basically feasible, but it is not yet clear for which sub-elements of the <correspDesc> element this should be allowed and how this information should be processed.[20]

RDF allows great freedom to qualify and expand on assertions by the simple mechanism of creating a new node in the graph that represents the assertion itself – a process called reification.[21] In practice, as described in chapter II.6, we find that the vast majority of the assertions we are interested in qualifying in this way are events. Thus instead of saying that someone is the sender of a letter we can say that they were involved in a 'sending event' and the outcome of the event was a sent letter. We can now add attributes to the sending event such as a date, a location and, for example, an indication of the evidence for these assertions such as the details of the letter enclosure. Most crucially, for disputed evidence, we can also record who made the assertion – so a letter may have multiple contradictory 'sending events' based on differing evidence cited by different scholars.

Developing unique identifiers is of little use, however, unless we can uniquely identify individuals, and this raises another key problem: namely, what kind of information is required to identify individuals uniquely. In contemporary data management systems, the core information required usually consists of the name, date, and place of birth as well as date and place of death when these are known. For historical individuals, such data is often lacking, and in any case it is not normally associated with every reference to the person; so disambiguation often requires a far richer data set, including not only the vita of the person concerned but also information about acquaintances, professions, spheres of activity, and the like. While the former core information is usually listed in the authority files, the latter are often missing. To make matters worse, authority file entries usually do not contain any provenance information for the individual statements given. The latter is, of course, also a weak spot of many biographical handbooks on a granular level. In both cases, it complicates the assessment of the reliability of the data, which in practice often raises the problem of whether the person appearing in a corpus of correspondence is actually the individual described in an authority file entry or not.

The only adequate solution to these problems is to record information on correspondents in a part of a more elaborate prosopographical data model. In the case of a relational database, this data could be recorded in columns in the table 'correspondents'; in the case of a TEI-based data capture, by means of separate register entries. Such an approach that utilizes authority file IDs where available, but adds the identifying information itself, has several advantages. On the one hand, all data can potentially be enriched with data provenance, which not only facilitates the evaluation of reliability but is also highly relevant for data re-use scenarios. On the

[20] See https://digiversity.net/2015/perspectives-of-the-further-development-of-the-correspondence-metadata-interchange-format-cmif/, accessed 20/03/2019.

[21] See https://www.w3.org/DesignIssues/Reify.html, accessed 20/03/2019.

other hand, this information is essential for the identification of persons for whom no authority file IDs are available across several corpora.

Using RDF, this is much simpler since we already have nodes representing people, places, organizations, and events created as a consequence of letter description. All that is needed is to select or define, if necessary, suitable prosopographical and genealogical ontologies and we can readily start to assert relationships between people and organizations. As these relationships are necessarily bounded in terms of time, space, and normally some organizational context, they can be represented by events. The reification approach mentioned earlier allows us to make reference to the evidentiary sources, which would be represented by additional nodes.

4 Modelling Time

Another problem arises with regard to the dating of letters. The fact that many letters are either missing a date or are only partially or vaguely dated, poses quite a challenge from an information science point of view. In addition, there is also the problem that a dating can be uncertain or ambiguous. Several requirements for a data model can be derived from this. An adequate data model must ensure that it is possible to differentiate between information that was provided by the data creator or editor and the information present in the source material itself. It also must provide means to record which aspects of the dating are uncertain or ambiguous. In concrete terms, this means that this information must be represented in a very granular fashion (that is to say, the data model for a date must be broken down into many different elements indicating the date as given, the editor's interpretation of the date as given, the calendar used, and issues of uncertainty or ambiguity). These issues are all discussed in detail in chapter II.3. It is because this granularity is often missing in curation-driven models that they are often not suitable for reuse by researchers.

To complicate matters further, not all letters can be dated exactly to a specific day, and this necessitates the use of time intervals indicating the earliest and the latest possible dates of dispatch. This means that an adequate date model must be able to deal both with discrete and continuous time. This, in turn, makes the processing of this data considerably more difficult: for queries that refer to specific periods of time, both letters dispatched on a specific day and those sent within a longer period of time must be considered together.

In a relational database, such granular information could be recorded in the 'letters' table by using a series of columns: one column that records the information as it represents itself in the source material, one column that represents the date in a machine-readable format according to the year-month-day ISO 8601-(2019 Part 1) specification, and additional columns that use Boolean operators to mark inference, uncertainty, and/or ambiguity. In order to record letters that can only be dated to a specific time frame, the computable date column would have to

appear twice, in order to enable the recording of the start and end of the respective time-span, accompanied by a new column that uses a Boolean operator to signify whether the letter has a time range or not.

In the case of TEI-based data capture, the TEI-Guidelines specify generic ways of dealing with time intervals, but fail to encompass all the cases one might encounter when dealing with correspondence metadata. As in the case of the encoding of the sender and recipient of a letter in TEI, the further development of the <corresp Desc> element as well as the Correspondence Metadata Interchange Format (CMIF)[22] is expected to cover those issues.

As has already been discussed in the preceding sections, in RDF, an event-based approach arises logically from the need to deal with evidentiary uncertainty in general, and no special mechanisms need to be invoked to deal with time issues specifically.

In all cases, a new standard will allow many of these problematic issues to be captured and expressed in an extended date format: this standard, due to be released soon, is ISO 8601-2019 Part 2, which is based on the Extended Date/Time Format (EDTF).[23] As part of the urgently needed implementation of this standard, transformation routines will be needed from this extended standard back to the simpler year-month-day ISO-8601-1 format, to ensure the processability of this data (in derivative form) in older systems.

5 Modelling Space

Identifying the places of sending and receipt of a letter likewise requires the capacity to record multiple names for the same place as well as uncertainties and ambiguities in granular form. An additional problem of great importance and complexity, however, is that of territoriality: the researcher frequently needs to analyse geographical data at more than one level. For instance, as well as being able to pinpoint the precise location from which a letter is sent on the map, one may need to know the hierarchy of superordinate political or ecclesiastical entities of which it was a part: after 1466, for instance Gdańsk was an autonomous city in the province of Royal Prussia, which formed part of the Kingdom of Poland, which in turn was part of the Polish-Lithuanian Commonwealth, until the rest of Polish Prussia was annexed by the Kingdom of Prussia in 1772 followed by Gdańsk itself in 1793.

The absence of such hierarchies of political and ecclesiastical entities from modern gazetteers severely limits their utility for historical purposes. Working within the territoriality of the twenty-first century generates anachronistic analyses. For instance, a calculation of the number of letters sent by Erasmus to places with-

[22] See https://correspsearch.net/index.xql?id=participate_cmi-format&l=en, accessed 20/03/2019.

[23] See https://www.loc.gov/standards/datetime/edtf.html, accessed 20/03/2019.

in the Holy Roman Empire will be completely different from the same calculation based on the boundaries of Germany today. Such difficulties are further compounded when one looks at the individual territories *within* a complex entity like the Holy Roman Empire, where political boundaries can be astonishingly complex, and dynastic relationships and confessional identities change frequently. Both the data and the framework for such an historical gazetteer are currently lacking. Projects such as the *World Historical Gazetteer Project*[24] and *EM Places*[25] are currently developing the necessary frameworks, but the challenge of providing a comprehensive data set will not easily be resolved. These issues are discussed in more detail in chapter II.2.

6 An Event-based Model of the Republic of Letters

Having considered the modelling of letter records and the entities of which they are composed – people, places, and dates – it is worthwhile returning in conclusion to the problem of modelling the republic of letters as a whole. Although correspondence networks provide an excellent point of departure for modelling this community, there was far more to the republic of letters than letters. The commonwealth of learning was knit together by many other forms of learned exchange as well, from learned travel (documented by travel diaries and *alba amicorum*), to the publication and collecting of books (documented in publishers' lists and library catalogues), to the activities documented in the archives of universities and other learned institutions. The structured collection of such information is the task of prosopography.

A direct but basic connection between the data models for correspondence and prosopography is provided by authority files, which enables the basic linking of letter senders or recipients with prosopographical information. A far more robust link, however, is to model biographies as event streams and to regard the sending or receipt of a letter as just one of the many events to be included in such a stream. This handling of correspondence metadata as a species of prosopographical data has reciprocal benefits to both data models. For instance, on the one hand, the letter metadata can provide prosopographical data on a person's geographical location at the date of sending; on the other, prosopographical data can be used to infer the location of a sender or recipient at a specific time if that is not indicated by the letter metadata itself.

Even more fundamentally, the concepts, tool, and infrastructure devised for event-based letter modelling of this kind can be adapted for event-based modelling of all the other transactions and modes of exchange which constituted the republic of letters and even more general patterns of intellectual and cultural exchange.

[24] See http://whgazetteer.org/, accessed 20/03/2019.

[25] See https://github.com/culturesofknowledge/emplaces, accessed 20/03/2019.

These opportunities are most fully elaborated in the conclusion to this volume (ch. V). In order to provide a foundation for this enterprise, chapter II.6 in this volume outlines recommendations for an event-based data model for the republic of letters that has the potential to integrate many different aspects of this field of research into one model.

II Standards: Dimensions of Data

II.1 Letters

Elizabethanne Boran, Marie Isabel Matthews-Schlinzig, and Signed, Sealed, and Undelivered (Rebekah Ahrendt, Nadine Akkerman, Jana Dambrogio, Daniel Starza Smith, and David van der Linden)

With contributions from Antonio Dávila Pérez, Christoph Kudella, and Roberta Colbertaldo

1 What Is a Letter?

Elizabethanne Boran with contributions from Antonio Dávila Pérez

'A letter consists of written communication typically addressed to one or more named recipients, and identifies the sender and conveys a message'.[1] This minimalist definition, though it provides a useful starting point, does not reflect the complexity of the letter and, more generally, letter writing in the early modern period.[2]

[1] Terttu Nevalainen, 'Introduction', in Terttu Nevalainen and Sanna-Kaisa Tanskanen, eds., *Letter Writing* (Amsterdam: John Benjamins, 2007), 1.

[2] Nor, for that matter, does the definition in the *Oxford English Dictionary*, which states that a letter is 'a written, typed or printed communication, sent in an envelope by post or messenger': quoted by Michael Sinding, 'Letterier: Categories, Genres, and Epistolarity', in Marie Isabel Matthews-Schlinzig and Caroline Socha, eds., *Was ist ein Brief? Aufsätze zu epistolarer Theorie und Kultur/What Is a Letter? Essays on Epistolary Theory and Culture* (Würzburg: Königshausen & Neumann, 2018), 21–37; at 22. Sinding notes the difficulties inherent with this definition and reminds the reader that 'Letters may be written by almost anyone, to almost anyone, using almost any materials, and delivered in almost any

For, as Nevalainen rightly notes, 'letter writing has always been situated activity' and therefore, in order to understand the concept of a letter in the early modern period, we must first look at early modern epistolary theory to see how contemporaries understood the term.[3]

'To expect all letters to conform to a single type, or to teach that they should, as I notice even learned men sometimes do, is in my view at least to impose a narrow and inflexible definition on what is by nature diverse and capable of almost infinite variation'.[4] So wrote Erasmus of Rotterdam in his *Opus de conscribendis epistolis* (Basle: Johann Froben, 1522), a text on letter writing which did much to shape understanding of epistolary theory in the early modern period. When pushed to produce a concise definition, Erasmus provided two formulations, each based on ancient and medieval commentators: a letter is 'a mutual conversation between absent friends', and 'a letter is a conversation between two absent persons'.[5]

The first formulation owed much to ancient definitions: for example, the Ancient Greek rhetorician Isocrates' letters were invariably addressed to known associates and written in a familiar style.[6] The even more influential Demetrius of Phalerum (c. 350–c. 280 BC), looked back to the definition of Artemon, the editor of Aristotle's correspondence, who stated that a letter was a conversation halved.[7] The idea of a letter as a continuation of a conversation by other means (in this case written), was further popularized by Cicero (106–43 BC), who spoke of 'Amicorum colloquia absentium'.[8] Thus in ancient epistolary theory there were four essential elements in the definition of a letter: it was (1) a written conversation; (2) written in a familiar style; (3) between people who were known to each other; (4) who were absent from each other. Absence was the necessary pre-condition for any letter.[9]

---

way' (22). See also Armando Petrucci, *Scrivere lettere. Una storia plurimillenaria* (Rome and Bari: Laterza, 2008).

[3] Nevalainen, 'Introduction', 1.

[4] Erasmus, 'De conscribendis epistolis', translated and annotated by Charles Fantazzi, in Jesse Kelley Sowards, ed., *Collected Works of Erasmus*, vol. 25: *Literary and Educational Writings*; vol. 3: *De conscribendis epistolis, Formula, De civilitate* (Toronto: University of Toronto Press, 1985), 12.

[5] Ibid., 20; and 'Conficiendarum epistolarum formula', translated and annotated by Charles Fantazzi, in ibid., 258.

[6] Robert G. Sullivan, 'Classical Epistolary Theory and the Letters of Isocrates', in Carol Poster and Linda C. Mitchell, eds., *Letter-writing Manuals and Instruction from Antiquity to the Present* (Columbia: University of South Carolina Press, 2007), 8.

[7] Carol Poster, 'A Conversation Halved: Epistolary Theory in Greco-Roman Antiquity', in Poster and Mitchell, eds., *Letter-writing Manuals*, 23.

[8] Gabriella Zarri, 'Sixteenth-Century Letters: Typologies and Examples from the Monastic Circuits', in Regina Schulte and Xenia von Tippelskirch, eds., *Reading, Interpreting and Historicising: Letters as Historical Sources* (European University Institute, Fiesole, 2004), 40 see http://hdl.handle.net/1814/2600.

[9] As Constable notes, this element of 'absence' allowed a widening of the concept of a letter, for it could be understood both spatially and temporarily: Giles Constable, *Letters and Letter-collections* (Turnhout: Brepols, 1976), 14.

Medieval commentators of the *ars dictaminis*, such as the early thirteenth-century author Guido Faba, added more elements to the definition of a letter. In his *Summa dictaminis*, written c. 1228–9, which Camargo notes was 'probably the single most influential treatise of the *ars dictaminis*', he included the following definition of a letter:

> *An epistle is a booklet sent to one or several absent persons, and it is called 'epistle' from epi, which is 'beyond', and stola or stolon, which is 'sending', because it makes the sender's desire clear 'beyond' a messenger's capacity to expound it. For on account of the mind's forgetfulness and the multiplicity of affairs and the distances of journeys, many things would be omitted, which an epistle represents like a mirror.*
>
> *The epistle was invented for two reasons. The first was so that the secrets of friends might be concealed through it, whence it is named from epistolo, that is, 'I conceal'. The second reason was so that it might express better than a messenger what is sent.*[10]

Faba's definition thus added four additional elements to that of the Greek and Roman commentators: (5) a letter could be a 'booklet'; (6) it was something that was 'sent'; (7) it provided secrecy; (8) it allowed for greater accuracy. Later thirteenth-century commentators, such as Conrad of Mure, added a fifth: (9) to enable conversation between persons unable to communicate directly with one another.[11] This moved the definition beyond a communication between friends to one between potential strangers, a definition later mirrored by Erasmus's second formulation, that a letter might be defined as 'a conversation between two absent persons'.

Erasmus's *Opus de conscribendis epistolis* spawned a host of imitators. As the sixteenth century progressed, writers not only sought to give advice on writing letters in Latin but increasingly identified the vernacular market as a growth area.[12] In general, such treatises on letter writing focused less on definition of letters and more on the division of letters into various genres. However, one area relevant to definition, which both medieval and early modern epistolary theory considered, was the question of the parts of a letter. Medieval treatises on letter writing agreed that there should be five parts to a letter:

1. the *salutatio*: the address, usually including the name of the author and the name of the addressee;

[10] Martin Camargo, 'Where's the Brief? The *ars dictaminis* and the Reading/Writing between the Lines', in Carol Poster and Richard Utz, eds., *The Late Medieval Epistle* (Evanston, IL: Northwestern University Press, 1996), 2.

[11] Camargo, 'Where's the Brief?', 2.

[12] See, for example, Angel Day's best-seller, *The English Secretorie* (London: Robert Walde-grave, 1586), which was reprinted in 1592, 1595, 1599, 1607, 1614, 1621, and 1635.

2. the *exordium*: a short statement or epigram outlining the qualities of both author and addressee and the circumstances of writing the letter;
3. the *narratio*: the main subject of the letter;
4. the *petitio*: the request;
5. the *conclusio*: the subscription, which might also include a brief summary of the argument.[13]

Of these the most important were the address and farewell, for it was these which identified the document as a letter.[14] Additional medieval requirements – that a letter be brief and that it should keep to a single subject were largely ignored in the early modern period. However, as Burton notes, the fact that the five parts were modelled on the six rhetorical parts of Ciceronian orations ensured that they were retained in one form or another by humanist scholars in their reform of the *ars dictaminis*.[15] The *salutatio* and farewell continued to be given pride of place: Erasmus spends some part of his *Opus de conscribendis epistolis* on the formula for greetings and farewells which should be employed.

The *salutatio* provides us with a useful starting point in our search for the essential characteristics of an early modern letter. If we concentrate less on the function of a letter and the circumstances of its creation, and more on its form, we may isolate the following essential elements whose combination enable us to recognize a text as a letter:

- A letter text includes the name of the sender/s.
- A letter text includes the name of the intended recipient/s.
- A letter text includes the place name of the letter's origin.
- A letter text includes the place name of the letter's destination.
- A letter text includes the date the letter was written.

True, there are examples where the address of the recipient may be missing, or the date of the letter might have been lost (or, in the case of a partial draft, not included), but in the main, the combination of these five elements indicate that a text is a letter. Ultimately, then, a letter is a written communication (in manuscript and/or

---

13 Rita Costa Gomes, 'Letters and Letter-writing in Fifteenth-Century Portugal', in Schulte and Von Tippelskirch, eds., *Reading, Interpreting and Historicising*, 24.

14 Constable, *Letters and Letter-collections*, 17. As Constable notes, not all letters had a subscription, for writers of secret letters necessarily wished to conceal their identity (18). The salutation and valediction were considered vital areas of a letter by ancient writers also: Sullivan, 'Classical Epistolary Theory and the Letters of Isocrates', in Poster and Mitchell, eds., *Letter-writing Manuals*, 7–20, esp. 9.

15 Gideon Burton, 'From *Ars dictaminis* to *Ars conscribendi epistolis*', in Poster and Mitchell, eds., *Letter-writing Manuals*, 92–3.

print form), between two different entities (sender/s and recipient/s), which is located in space and time by the addition of addresses and a date.

2 The Letter as Text

Elizabethanne Boran with contributions from Antonio Dávila Pérez

At its most basic level, a letter is both text and material object and any definition of a letter must take into account the interaction of both in order to understand it fully.[16] Any definition of a letter must consider the fact that early modern letters often survived in more than one state. The arrival of printing further complicated what was already a very complicated process. In the early modern period a letter might survive in the following multiple states:

1. as a draft – this might exist in various stages of composition;
2. a holograph letter (sometimes called the 'original'), written by a scribe and signed by the author;
3. a holograph letter, written wholly by the author;
4. a copy of the letter, written by a scribe;
5. a copy of the letter, written by the author;
6. a copy of the letter, written by the recipient/s;
7. a printed copy of the letter;
8. a revision of the letter;
9. a summary/abstract of the letter in manuscript form;
10. a summary of the letter in print form;
11. a translation of the letter by either the recipient or someone in her/his circle.

The correspondence of the seventeenth-century scholar James Ussher (1581–1656), archbishop of Armagh, exhibits many of these states.[17] Ussher's letter to David Rothe, the Roman Catholic bishop of Ossory, exists only as an undated partial draft.[18] When Ussher became archbishop in 1625 much of his correspond-

[16] On this point see James Daybell, *The Material Letter in Early Modern England. Manuscript Letters and the Culture and Practices of Letter-writing, 1512–1635* (Basingstoke: Palgrave Macmillan, 2012).

[17] Elizabethanne Boran, ed., *The Correspondence of James Ussher 1600–1656*, 3 vols. (Dublin: Irish Manuscript Commission, 2015).

[18] Ibid., vol. 3, 1143. An example of a letter that was substantially revised during its composition is Gerardus Joannes Vossius's letter to Ussher of 1/11 January 1632 from Amsterdam, now in the

ence was drawn up by scribes and simply signed by him.[19] Nonetheless, many of his holograph letters survive, sometimes in manuscript volumes devoted to letters.[20] Copies written by others were also kept in miscellaneous volumes. Some letters were evidently more privileged than others: for example, Ussher kept copies of his outgoing letters to many of his new continental correspondents, presumably because he was afraid that his initial letters to them might be lost in the post – and also to keep track of his correspondence with them which, unlike that of correspondence closer to home, might take more time to elicit a reply.[21] Equally, some of Ussher's letters were reprinted by contemporary scholars keen to link themselves to him.[22] Indeed, many exist only in print form, primarily because his seventeenth-century editor, Richard Parr, on printing an edition of Ussher's life and letters in 1686, evidently decided to destroy the holographs, which Ussher (unlike so many other authors), had thoughtfully archived for posterity.[23] Some later editors silently revised the text (though in Ussher's case such textual changes were minor).[24] Some of Ussher's letters to William Laud, archbishop of Canterbury, were kept as manuscript abstracts and are now in the Bodleian Library, Oxford.[25] One letter exists only in translation: Ussher's letter to James Frey, 17 September 1635, was translated into German, presumably by Frey (or by a member of his circle), who was professor of Greek at the University of Basle, so that he might send on the import of the original to a different, vernacular-speaking readership.[26] To these eleven states we might, to further complicate matters, add a ghostly twelfth – the all too frequent missing letter, whose existence may be deduced but which is no longer extant.[27]

Bodleian Library, Oxford, MS Lett. 83, 67r–68v. An initial manuscript draft of this, indicating many revisions, exists at Bodleian Rawl. MS Lett. 84, 58r–59r. The letter was printed by Paul Colomiès, *Gerardi Joan. Vossii et clarorum virorum ad eum epistolae* (London: Sam Smith, 1690), 186–8, and is now in Boran, *Correspondence*, vol. 2, 566–73.

[19] Ibid., vol. 2, 504–5: Ussher to the Lords Justices Adam Loftus and Richard Boyle, 3 April 1630, Drogheda. Other examples of this practice may be found in Museum Plantin-Moretus, which holds some letters signed by Chrisophe Plantin (c. 1520–1589), but written by his son-in-law, Jan Moretus (1543–1610).

[20] See, for example, Bodleian Library, Rawl. MS Lett. 89, which is a volume of letters, evidently collected by Ussher himself.

[21] Boran, *Correspondence*, vol. 2, 474–6: the manuscript draft of Ussher's letter to Louis de Dieu, 1 October 1629, now in the Bodleian Library, Rawlinson MS. C. 850, 61r–v, was also available in a manuscript copy in Bodleian Library, Tanner MS 461, 56r–v.

[22] Ibid., vol. 3, 1171: Ussher's letter to Nicholas Bernard, c. 1656, was reprinted by Bernard in his *The Judgement of the Late Arch-bishop of Armagh, and Primate of Ireland*, 2nd edn. (London: John Crook, 1658), 110–13.

[23] Boran, *Correspondence*, vol. 1, xxxvii.

[24] Ibid., vol. 1, 56, fn 2, gives an example of this type of minor change in earlier print editions of a letter from Samuel Ward to Ussher, 6 July and 1 August, 1608.

[25] Ibid., vol. 2, 505–6 which is taken from Bodleian Library, Sancroft MS 18: Ussher's letter to William Laud, 5 April 1630.

[26] Ibid., vol. 2, 671–3.

[27] Ussher, writing to Samuel Ward on 15 March 1630, mentions a letter from Ward dated 11 January, which is now lost: ibid., 500.

As these examples demonstrate, letter states are chiefly categorised by using the following criteria:

The *contents* of the letter (nos. 1, 8, 9, and 10). Here the existence of a holograph letter greatly aids identification of other states, such as drafts, revision, and abstracts in other formats, but, as Ussher's letter to Rothe demonstrates, it is also possible to identify drafts in the absence of holographs. In this case the subject matter indicates that the letter is part of a known ongoing communication process. As Vossius's letter to Ussher of 1/11 January 1632 shows, another useful indicator of draft status is the frequency of deletions in the exemplar.

The *handwriting* of the letter and signature (nos. 2, 3, 4, 5, and 6). To distinguish whether an item was written/copied by its author, a scribe, or a recipient, it is necessary to compare the handwriting of the letter with exemplars from both sender and recipient in order to clarify correct identification.

The *format* of the letter (no. 7). Early modern letters exist in both manuscript and/or printed form. In the early modern period it was relatively common to destroy manuscript holographs once they had been published in print.

The *language* of the letter (no. 11).

3 The Letter as Object

Signed, Sealed, and Undelivered (Rebekah Ahrendt, Nadine Akkerman, Jana Dambrogio, Daniel Starza Smith, and David van der Linden)

Letters do not simply bear the words of authors to their recipients, they can also be interpreted as carefully crafted composites of substrate and writing substance. The reading of a letter begins long before it is opened, as its material features communicate a series of silent cultural assumptions. In the last decade and a half, scholars have increasingly turned their focus to the material features of letters, particularly in the early modern period.[28] Digital resources for the study of letter collections have also begun to factor materiality into their research remits, raising

[28] See most notably Sara Jayne Steen, 'Reading Beyond the Words: Material Letters and the Process of the Interpretation', *Quidditas* 22 (2001): 55–69; Alan Stewart and Heather Wolfe, *Letterwriting in Renaissance England* (Washington, DC: The Folger Shakespeare Library, 2004); Colette Sirat, *Writing as Handwork: A History of Handwriting in Mediterranean and Western Culture*, ed. Lenn Schramm (Turnhout: Brepols, 2006); Alan Stewart, *Shakespeare's Letters* (Oxford: Oxford University Press, 2008); James Daybell, 'Material Meanings and the Social Signs of Manuscript Letters in Early Modern England', *Literature Compass* 6:3 (2009): 649–67, see https://doi.org/10.1111/j.1741-4113.2009.00629.x; Daybell, *The Material Letter in Early Modern England*; Harry Newman, '"A seale of Virgin waxe at hand/Without impression there doeth stand": Hymenal Seals in English Renaissance Literature', in James Daybell and Andrew Gordon, eds., 'New Directions in the Study of Early Modern Correspondence', special issue of *Lives and Letters* 4:1 (2012): 94–113; Heather Wolfe, '"Neatly Sealed, with Silk, and Spanish Wax or Otherwise": The Practice of Letter-Locking with Silk Floss in Early Modern England', in Susan P. Cerasano and Steven W. May, eds., *In the Prayse of Writing* (London: British Library, 2012), 169–89. Before this more recent interest, Pierre Chaplais, *English Royal Documents: King John–Henry IV, 1199–1461* (Oxford: Clarendon Press, 1971), pursued similar questions.

new possibilities for archival access and data-driven analysis of epistolary materiality.[29] This section summarizes the essential concerns of the study of letters as objects, and explains how two current interrelated projects – *Signed, Sealed, and Undelivered* (*brienne.org*) and *Letterlocking* (*letterlocking.org*) – are seeking to theorize them in new ways and implement tools for their further study.

For much of history, a letter could not simply be rushed off: even a short, informal note would require some degree of planning and preparation. As objects, letters should be considered in relation to a series of other objects on a letter-writer's desk, which might include an inkwell, standish (inkstand), candle, feather quill (carved with a pen-knife) or pen, seal matrix, dust-box, blotter, scissors, whetstones, wax jack, and rulers.

The letter proper begins with a substrate: paper, papyrus, and parchment are most familiar to us now, but in other traditions clay, wax tablets, bark, etc. were also used. Scholars of material letters ask how thick this substrate is, whether (if it is paper) it is hand-made, if its chain and wire lines (also known as laid lines) are visible, and if its watermark enables us to identify its source. All these details enable us to understand the document's make-up. We also need to ask if it has survived largely intact or whether damage (such as mould and ink corrosion) or interventions over the years might have destroyed or altered some of its material evidence. Before writing, the substrate may need to be trimmed for neatness, prepared to ensure it did not absorb too much ink, then folded into a suitable shape for writing, often a bifolium; a crease running parallel to one edge can serve as a writing margin.

Early modern letter-writers usually made their own ink, and the quality of ink can drastically affect a letter's afterlife – too much acidity, and it will eventually eat through the paper. Invisible inks can more subtly alter the physical state of the paper, if made visible by the recipient or interceptor: chemicals need a reagent such as water, which might leave the paper crinkled; fluids such as milk or the juice of citrus fruit need the heat of a flame to oxidize, which might leave the paper scorched. Many letters may survive with hidden writing which has never been made visible.

Once written, letters' contents have their own materiality, for example, where signatures or marginalia are placed, if cross-writing is employed, or how much blank space is left around the writing.[30] Letter-writers may employ cryptology – which can take the form of cryptography (ciphers) or steganography (codes, riddles, invisible inks) – to disguise their message, or to embed a hidden communica-

[29] Daniel Starza Smith, 'The Material Features of Early Modern Letters: A Reader's Guide', in Alison Wiggins, ed., *Bess of Hardwick's Letters: The Complete Correspondence c. 1550–1608* (2013). See http://www.bessofhardwick.org/background.jsp?id=143, accessed 20/03/2019.

[30] Jonathan Gibson, 'Significant Space in Manuscript Letters', *The Seventeenth Century* 12:1 (1997): 1–9, see https://doi.org/10.1080/0268117X.1997.10555420; Anna Bryson, *From Courtesy to Civility: Changing Codes of Conduct in Early Modern England* (Oxford: Oxford University Press, 1998); Giora Sternberg, 'Epistolary Ceremonial: Corresponding Status at the Time of Louis XIV', *Past and Present* 204:1 (2009): 33–88, see https://doi.org/10.1093/pastj/gtp018.

tion within the overt one. These, too, have their own material conventions and histories. Other material features sometimes found on historical correspondence include postal marks, seals, ribbons, and sealing wax. The borders and edges of a writing substrate can sometimes be decorated: letters announcing a death, or written during a period of mourning, might be edged in black; gilt edging is common; blue- and green-edged letters also exist. Edging is a feature which can easily be overlooked on a digital surrogate.

Early modern letters were composed in an age before the mass-produced gummed envelope had been invented. This usually meant that, after writing, the writing surface itself had to be folded up to become its own sending device. This process is called 'letterlocking', 'the act of manipulating and securing an epistolary writing substrate (such as papyrus, parchment, or paper) to function as its own envelope.[31] Letterlocking is a subcategory of a 10,000-year information security tradition, pertaining to epistolary materials, and its study encompasses the materially engineered security and privacy of letters, both as a technology and a historically evolving tradition. Letterlocking demonstrates that letters were for centuries folded and otherwise manipulated to become their own envelopes, and that this process has a rich history. Archival letters can today seem like flat, fossilized, two-dimensional artefacts, but letterlocking reminds us that they were once dynamic, three-dimensional objects which travelled through space and worked as engineered objects, often including sophisticated anti-tamper mechanisms.

The material features of letters – in particular the letterlocking aspects – have hitherto rarely been captured in epistolary databases, which have largely focused on standard content-related metadata such as date, place, author/sender, and keywords. One notable exception is *Bess of Hardwick's Letters*, for which Wiggins et al. recorded a number of material features in the metadata, making this 234-item corpus searchable by fifteen standards, including 'Letters with seals', 'Letters with significant space', 'Endorsements', 'Subscriptions', and 'Sewn'.[32] These metadata capture standards usefully group letters that exhibit common physical features, enabling them to be studied and compared more easily.

The main repository which records materiality among its metadata is the epistolary union catalogue EMLO, which enables contributors to note postage marks, endorsements, enclosures (both letters with enclosures and letters that are enclosures), seals, paper type, paper size, and handling instructions.[33] Although EMLO

[31] Jana Dambrogio, 'Historic Letterlocking: The Art and Security of Letter Writing', *Book Arts/Arts du Livre Canada* 5:2 (2014): 21–3. Letterlocking videos illustrate how these letters were once folded and secured shut, and other resources including vector diagrams and a monograph are in preparation for publication. See the 'Letterlocking' channels on YouTube (https://www.youtube.com/c/Letterlocking) and Vimeo (https://vimeo.com/letterlocking), both accessed 20/03/2019. The field of letterlocking was initially developed by Dambrogio, and first introduced at the annual conference of the American Institute for Conservation and Historic and Artistic Works (AIC), Minneapolis, MN, 2005.

[32] See https://www.bessofhardwick.org/filter.jsp?filter=1, accessed 20/03/2019.

[33] See http://emlo.bodleian.ox.ac.uk/advanced, accessed 20/03/2019.

already allows for a rich variety of data to be captured, the incorporation into EMLO of the Brienne Collection, an extremely well-preserved archive of 2,600 undelivered letters sent from all around late seventeenth-century Europe,[34] has prompted the development of new metadata standards to record different kinds of evidence, both material and ephemeral.[35]

From a material standpoint, the Brienne Collection presents a series of opportunities and challenges related to the state in which the letters have been preserved: all are archived in their folded state, and some 600 of them have never been opened (even by their original addressee). These features have inspired two new metadata fields capturing (1) whether letters are still unopened and (2) if they have been stored folded. The *Signed, Sealed, and Undelivered* project is also developing metadata standards which more overtly define and capture evidence of letterlocking. In particular we seek to record letterlocking *formats* and *categories*. Formats refer to the shape the letter takes when folded into a packet (e.g. 3 = triangle, 4 = quadrilateral, 5 = pentagon). Categories are distinguished by the number and combination of steps required to make a packet, including (for example) folds, slits, and locks. The *Unlocking History* research team, led by Dambrogio and Smith, is working to refine format and category information into metadata standards that can be globally adopted.[36] The material features of the letterlocking data – and thus, by extension, of the letters – open up new and exciting avenues for scientific analysis, allowing scholars to relate letters' content to their material features, and to explore technological trends and innovation across centuries, borders, and cultures.

From a more immaterial perspective, but still pertaining to letters as objects, the Brienne letters challenge commonly accepted notions about the nature of correspondence routes which may necessitate further revision of EMLO's metadata fields. The letter as an object is the product not just of one 'author', but of an entire system. EMLO, like most correspondence databases, had operated on the assumption that the *origin* of a letter is one fixed geographical location where it was physically written, while the *destination* is the location of the addressee. But what if there are multiple hands writing from various locations to a destination that is never achieved? Many of the Brienne letters came to The Hague, after all, not according to the will of their senders but by accident or omission: a good number of the letters now in 'La Haye en Hollande' were intended for 'La Haye en Touraine', a small village in France.

---

[34] Rebekah Ahrendt and David van der Linden, 'The Postmasters' Piggy Bank: Experiencing the Accidental Archive', *French Historical Studies* 40:2 (2017): 189–213, see https://doi.org/10.1215/00161071-3761583.

[35] Rebekah Ahrendt, Nadine Akkerman, Jana Dambrogio, Daniel Starza Smith, and David van der Linden, eds., 'The Brienne Collection', in *Early Modern Letters Online*, Cultures of Knowledge, http://emlo-portal.bodleian.ox.ac.uk/collections/?catalogue=brienne-collection, accessed 20/03/2019.

[36] Two forthcoming publications co-edited by Dambrogio and Smith will set out the terms of letterlocking in more detail: a monograph, *Letterlocking*, and the *Dictionary of Letterlocking* (DoLL).

Furthermore, The Hague was not the only place where these letters stopped along their originally intended routes; many bear the marks of other post offices along the way, sometimes including the dates on which they were there. In a sense, *all* of the letters in the Brienne Collection arrived at the 'wrong' destination, whether due to incomplete or indecipherable addresses or to the absence, death, or non-acceptance of their addressees. The reasons for non-delivery were carefully recorded by the post office in The Hague on nearly every letter; notes such as 'in England', 'departed', 'refused', attest to the many hands and voices implicated in the traffic of correspondence, which deserve themselves to be acknowledged and recorded. Thus, *Signed, Sealed, and Undelivered* and EMLO have developed metadata fields to record the address as intended, the route taken by the letter across time and space (including special handling instructions), and the reasons why it was never delivered.

4 The Letter as Genre – Early Modern Letter Genres (1500–1800): Definitions, Conceptualizations, Metadata

Marie Isabel Matthews-Schlinzig

Descriptive metadata records on letters should ideally also include information on epistolary subgenre/s. In order to identify an appropriate way of recording this information, it is necessary to reflect on four aspects: (1) how epistolary subgenres are conceptualized; (2) how they were defined in the past; (3) how they are studied in the present; and (4) how the current state of knowledge on letter genres can best be integrated into data sets.

4.1 Conceptualizing Letter Genres

Letter genres (i.e. epistolary subgenres)[37] are commonly conceptualized in relation to a particularly distinct feature: this includes (1) a function, theme, or purpose (e.g. farewell letter, love letter, blackmail letter); (2) institutional or social contexts (e.g. chancery letter, children's letter); (3) a particular linguistic quality, style, and/or textual form (e.g. gallant letter, epistolary treatise); or (4) prominent material qualities (e.g. illustrated letter). Some letter genres are characterized by a combination of such defining features – think, for instance, of the use of black wax to seal letters of mourning.

Letter genres emerge from and develop through epistolary cultures and practices, and in particular due to specific social, educational, economic, and institu-

[37] The terms 'epistolary subgenres' and 'letter genres' are used here interchangeably to designate different types of primarily non-fictional letter writing, including published correspondence.

tional needs.[38] Hence, many subgenre designations (e.g. condolence letter) traditionally reflect specific types of social and cultural interaction. This social – as well as cultural – 'grounding' (Charles Bazerman) of letters accounts for the fact that the existence, lifespan, age, characteristics, and names of epistolary subgenres can vary across geographical space and time, i.e. between cultures and languages, and within societies. Furthermore, developments in printing and publishing, as well as (modern) scholarly interests have played key roles in subgenre creation and change.

The multiple factors contributing to the development of epistolary subgenres are not the only reason why it is not always straightforward to delineate letter genres clearly or assign an individual piece of correspondence to a specific subgenre. Such difficulties are exacerbated further by basic characteristics of the epistolary form as well as of letter cultures as such: key factors include the polymorphism and malleability of letters, the epistolary form's inherent poetic qualities, and the often complicated transmission history of letters.

Letters, like 'texts' more generally, 'may', as Michael Sinding points out, '"coincide with" multiple genres in various degrees'.[39] Genre mixes – both in a specific as well as a broader sense – are common: one letter can be associated with several epistolary subgenres. Furthermore, just as letter writing in general can integrate several ways of recording and representing information as well as adapt other genres of writing (e.g. poetry, calculations, drawings), so epistolary subgenres can encompass miscellaneous types of text in several styles and material forms. The subgenre of scholars' letters, for instance, comprises such different forms as familiar correspondence, epistolary treatises, and a range of paratexts, including epistles dedicatory.[40]

Ultimately, this multiplicity of forms relates back to the question raised earlier in this chapter: what actually constitutes a letter and how – if at all – can one differentiate it from other types of writing? While its perhaps most defining characteristic nowadays, the 'epistolary bracket' (i.e. the presence of an address and a salutation plus a signature and, potentially, a leave-taking formula) persists in genres of writing that have emerged from the epistolary form (e.g. legal documents such as letters patent),[41] it is not always fully present in letters, let alone a defining feature of some epistolary subgenres (e.g. billets, published letter series).

---

[38] See Charles Bazerman, 'Letters and the Social Grounding of Differentiated Genres', in David Bartin and Nigel Hall, eds., *Letter Writing as a Social Practice* (Amsterdam: John Benjamins, 1999), 15–29, at 17–20.

[39] Sinding, 'Letterier: Categories, Genres, and Epistolarity', in Matthews-Schlinzig and Socha, eds., *Was ist ein Brief?*, 21–37, at 25.

[40] Thomas Wallnig, 'Gelehrtenbriefe', in Marie Isabel Matthews-Schlinzig, Jörg Schuster, and Jochen Strobel, eds., *Handbuch Brief: Von der Frühen Neuzeit bis zur Gegenwart*, 2 vols. (Berlin and Boston: De Gruyter, publication planned for 2020).

[41] See Bazerman, 'Letters', 20. On the subject of other genres (as well as 'hybrid sub-genres') emerging from letter writing in the early modern period, see also Gabriella Del Lungo Camiciotti, 'Letters and Letter Writing in Early Modern Culture: An Introduction', *Journal of Early Modern Studies* 3 (2014): 17–35, at 25–8. See https://doi.org/10.13128/JEMS-2279-7149-14163.

The fundamentally rhetorical nature of letters adds a further layer of complexity to the picture: while some epistolary subgenres are primarily non-fictional (e.g. papal bull/brief), others (e.g. letter poems) tend to encompass a large proportion of fictional or semi-fictional texts. Although this chapter centres on non-fictional writing, it should nevertheless be noted that letter writing takes place on a continuum of literary and factual writing activities.[42] Not least because of this, fiction can play a key role in the emergence and/or development of letter genres; after all, fiction does not only respond to, and document, but can also modify or set new standards for contemporary genre conventions (e.g. through epistolary novels).

Last but not least, a number of processes that take place during the transmission of a letter (e.g. publication) can also influence its association with one or more epistolary subgenre/s. The transformations that letters can undergo from being first drafted to – potentially – being circulated in print (e.g. fictionalization), can be manifold.

From these observations it is evident that letter genres are fundamentally as dynamic as the social and cultural processes they both manifest and shape. If we wish to identify epistolary subgenres and assign them to individual letters we must therefore be cognizant of two interrelated factors: firstly, the culture-historical contexts for the classification as well as use of letter genres in the past and, secondly, our own theoretical premises.

4.2 Early Modern Letter Classes

As Gabriella Del Lungo Camiciotti has remarked, '[i]n the early modern period […] correspondence acquired the characteristics and uses that were to become typical of the genre in the whole modern period'.[43] That this is also true of many epistolary subgenres can be gleaned from a range of sources (e.g. manuscript letters, letter collections, collections of model letters, letter manuals, and treatises on the epistolary form). However, while many a letter genre, especially from the later early modern period, might seem familiar to twenty-first-century readers, the subgenre's historical conceptualization can deviate considerably from present understanding. A comparative study of source material both from different parts of the period and from different language contexts will reveal significant variation both in classification methods and definitions of letter genres; there are also discrepancies between letter theory and letter writing practices.

Already before the sixteenth century, letter writing manuals had included definitions of a number of letter 'genres': e.g. 'Francesco Negro's 1492 Modus epistolandi […] list[ed] some twenty kinds of letters (varying from the epistola expurga-

[42] For useful, if diverging, theoretical reflections on the aesthetic, imaginative, or fictional potentials of the epistolary form, see e.g. Robert Vellusig, 'Die Poesie des Briefes: Eine literaturanthropologische Skizze', and Inka Kording, 'Epistolarisches: Die achtfache Relationalität in Briefen', in *Was ist ein Brief?*, 57–75 and 77–89. See also section 4.3 of this chapter on Claudio Guillén.

[43] Camiciotti, 'Letters', 18.

tiva to the epistola domestica)'.[44] Initially, classical oration alongside classical letter writing served as the key frameworks for describing and categorizing epistolary subgenres; this was especially pertinent to letters written in Latin.[45] New thoughts on classifying epistolary subgenres evolved gradually.

Two different versions of Erasmus's *Opus de conscribendis epistolis* (1522) illustrate such a process. In an early draft (c. 1499–1500), Erasmus 'classifie[d] subcategories of letters under the categories of the oration (deliberative, demonstrative, and judicial)'; letters he could not describe in relation to these categories he called 'extraordinary'.[46] The 1522 version of the *Opus* refined this model. Not only did Erasmus differentiate between 'mixed' and 'unmixed' letters, i.e. those that cover one or many topics, before subdividing the latter again in relation to the rhetorical categories mentioned. He further replaced the term 'extraordinary' with 'familiar', and added, almost as an afterthought, a fifth category of letters that dispute, investigate, and teach.[47]

This and other developments of letter genre classifications were in part due to changing conceptualizations and uses of the epistolary genre: from the Renaissance onward, letter theoreticians and practitioners began to consider the letter as a more flexible form than it had been to their medieval predecessors and contemporaries who continued to use the *ars dictaminis*.[48] Other causes driving epistolary subgenre development, which cannot be discussed in detail here, include increases in the variety of social groups using letters (e.g. a growing amount of non-official correspondence and increasing numbers of middle-class, as well as, later, lower-class letter-writers and -readers), and more frequent use of vernacular languages in letter writing.

A look at classifications of epistolary subgenres from both the earlier and later parts of the period indicates, also, that the classical rhetorical tradition continued to influence subgenre concepts arising in the seventeenth and eighteenth centuries. In *Empire of Letters* (2005), Eve Tavor Bannet briefly contrasts letter 'classes' present in eighteenth-century English-language manuals with those found in some sixteenth-century manuals.[49] In the latter, she writes:

---

[44] Burton, 'From *Ars dictaminis* to *Ars conscribendi epistolis*', in Poster and Mitchell, eds., *Letter-writing Manuals*, 88–101, at 89.

[45] See Eve Tavor Bannet, *Empire of Letters: Letter Manuals and Transatlantic Correspondence, 1680–1820* (Cambridge: Cambridge University Press, 2005), 55. See also Lawrence D. Green, '*Dictamen* in England, 1500–1700', in Poster and Mitchell, eds., *Letter-writing Manuals*, 102–26.

[46] Judith Rice Henderson, 'Humanism and the Humanities: Erasmus's *Opus de conscribendis epistolis* in Sixteenth-Century Schools', in Poster and Mitchell, eds., *Letter-writing Manuals*, 141–77, at 146–7.

[47] Ibid., 147.

[48] See Burton, 'From *Ars dictaminis* to *Ars conscribendi epistolis*'. See also Jane Couchman and Ann Crabb, eds., *Women's Letters across Europe, 1400–1700: Form and Persuasion* (Aldershot: Ashgate, 2005), 7–8.

[49] Bannet, *Empire*, 55. Specifically, she refers to Justus Lipsius's *Principles of Letter-writing* (*Epistolica institutio*, 1591) and Day's *The English Secretary* (1599). Compare e.g. for an introduction to epistolary subgenres in German-language letter manuals in the seventeenth and early eighteenth centuries,

> *Stylistically, letters were grave, learned or jocular; functionally, they were 'exhortatorie, accusatorie, commendatorie, excusatorie, congratulatorie, resonsorie, consolatorie, jocatorie, nunciatore, criminatorie, lamentatorie, mandatorie, dehortatorie, objurgatorie, petitorie, comminatorie, expostulatorie, ematorie, conciliatorie, laudatorie, or intercessorie'.*[50]

> [In the eighteenth century] *recognized classes of familiar letter* […] *included: letters of business; letters of advice; letters of praise or commendation; letters of recommendation; letters of remonstrance; letters of exhortation; letters of entreaty, request or petition; letters of counsel; letters of complaint; letters of reproof; letters of excuse; letters of thanks; letters of invitation; letters of congratulation; letters of consolation, comfort, or condolence; letters of visit; letters of compliment; letters proffering assistance; letters of merriment or raillery; and letters mixing two or more of the above.*[51]

Irrespective of potential similarities and differences between older and newer letter genres (which in many cases are still awaiting analysis), one of the more significant links between them is that their names reflect 'speech-act functions', i.e. they 'name possible verbal actions'.[52] Other, more recent forms of classification, found for instance in German-language letter manuals, identify letter subgenres in relation to the letter authors' 'intent', or a particular social group they belong to (e.g. merchants, 'gentlewomen', chancery officials, students, and soldiers).[53]

As Bannet reminds us, conceptualizations of subgenres emerging in the later part of the period should also be approached with care, since some '[e]ighteenth-century letter classes […] contained types of letter that we do not now think of together'.[54] 'Letters of Business', for instance, 'dealt with court, political, administrative or government business, as well as with commercial affairs. […] [T]hey could include letters of advice, counsel, remonstrance, command, request, recommendation, offering assistance, complaint, reproach and excuse'.[55] The hierarchical relationship between letter genres observable in this last example is another feature that is subject to historical change and cultural variation (as already indicated, e.g. by Erasmus's categories of 'mixed' and 'unmixed letters', etc.).[56] At the same time,

Carmen Furger, *Briefsteller: Das Medium 'Brief' im 17. und frühen 18. Jahrhundert* (Cologne: Böhlau, 2010), especially 135–46.

[50] Bannet, *Empire*, 55, note 3.

[51] Ibid., 55–6.

[52] Sinding, 'Letterier', 33.

[53] Furger, *Briefsteller*, 141.

[54] Bannet, *Empire*, 57.

[55] Ibid., 57–8.

[56] For further examples see e.g. Furger, *Briefsteller*, 136–8.

Bannet's description highlights that letter genre mixes could be an integral part of historical conceptualizations of epistolary subgenres.

Finally, when studying historical subgenre concepts, we should remind ourselves that historical letter theory does not represent a complete record of epistolary subgenres in use in the early modern period. Letter treatises and manuals mostly had a clear agenda: they were used to codify practices that were already present, establish best practice guidelines, and/or to instigate change. Manuals also tended to have educational purposes, were influenced by changing ideological concerns, and often addressed a particular readership. As a consequence, some letter genres that were commonly used and even present in other types of publication either did not find their way into letter treatises and manuals, or did so only comparatively late.

An example of this is the epistolary subgenre of last letters written before death. The English language offers a number of names for letters in this subgenre, including 'farewell letter', 'farewell note', 'last/final letter' as well as 'last note', 'death-bed letter', 'suicide note', 'death note', and 'suicide letter'.[57] This – in comparison to other European languages – very long list of designations suggests a particularly differentiated cultural presence of the subgenre. Hence, in the eighteenth century, at least some examples of such letters found their way into English-language letter manuals.[58]

This was, however, not the case in other languages and cultures, such as, for instance, in German: in fictional and non-fictional epistolary practice, the subgenre was well known and designated mostly either with *letzter Brief* ('last letter') or *Abschiedsbrief* ('farewell letter'). Yet, letter manuals did not include 'last letters', and defined *Abschiedsbrief* (or *Abschiedsschreiben*), only as polite notes which it was customary to write 'to good friends and patrons' when one parted from the place where the latter lived.[59] Given such gaps in historical letter theory, the widest possible range of relevant sources and manifestations of letters have to be taken into account in order to reconstruct historical knowledge about a letter genre.

From our present, scholarly point of view, both the lack of a historical theorization of specific epistolary subgenres, as well as the variety of different conceptualizations and designations that have developed over time/across cultures for what can essentially be described as one letter genre (e.g. letters consolatorie and letters of consolation), make it necessary to formulate new, 'historical-critical' definitions of epistolary subgenres. Although such definitions should reflect historical concepts and practice, they will, at least in part, also be constructs that are retroactively applied.

---

[57] Marie Isabel Schlinzig, *Abschiedsbriefe in Literatur und Kultur des 18. Jahrhunderts* (Berlin and Boston: De Gruyter, 2012), 8.

[58] Ibid., 102.

[59] Ibid., 7. Contemporary dictionaries tended to define *Abschiedsbriefe* as legal documents or certificates of service issued by employers (6).

4.3 Modern Concepts

In the more recent past, the number of research projects and studies focusing on individual letter genres has been increasing.[60] Nevertheless, more comparative and interdisciplinary research on the theory, history, and relationships of epistolary subgenres is needed. There is, for instance, still no consolidated typology of letter genres that are frequent in the early modern period. Ideally, such a catalogue would also include definitions and examples, and be flexible enough to serve as a basis for work on correspondence across different disciplines and cultures. As a step toward such a typology, the end of this section offers a (not yet comprehensive) list of epistolary subgenres that makes use of 'historical-critical' letter genre designations.

Conceptually, this list (and indeed this contribution as a whole) is informed by work currently under way for the interdisciplinary project *Handbuch Brief* ('Letters – A Handbook').[61] In order to throw into greater relief the relative merits and limitations of the handbook and thematize alternative approaches, it will be discussed here alongside two other, pertinent pieces of scholarship that define models for describing epistolary 'subtypes', genres, and/or subgenres. Both have a specific historical focus and are based on different disciplinary points of view; in what follows, they are used to highlight aspects of letter genres that current 'historical-critical' conceptualizations do not necessarily capture.

In his article 'Letters: A New Approach to Text Typology', linguist Alexander T. Bergs defines letters as a '"surface" or "super" text type' which, he suggests, 'can be subdivided into […] socio-pragmatic subtypes on the basis of socio-psychological and pragmatic dimensions and factors, including speech act and accommodation theory'.[62] In defining his subtypes, Bergs reflects on (1) a particular set of primary source material, the late Middle English Paston letters; (2) the social positions of letter-writer and addressee; and, adapting primarily Karl Bühler, (3) the function that language fulfils in each case.[63] Accordingly, Bergs distinguishes five subtypes:

---

[60] This includes – with regard to letter genres present in the early modern period – the anonymous letter, bridal letter, diplomat's letter, humanist's letter, last letter written before death, letter of friendship, letter poem/heroid, love letter, maternal letter, open letter, physician's letter, scholar's letter, secret letter, woman's letter, as well as the child's letter. The literature on early modern epistolary genres is large and continuously growing. Due to the limited length of this contribution, an extended bibliography of relevant publications cannot be included here. See, for selected references, 'Sources for Early Modern Letters' at The Warburg Institute: https://warburg.sas.ac.uk/research/completed-research-projects/scaliger/sources-early-modern-letters, accessed 20/03/2019; a listing of online resources on the *Cultures of Knowledge* website http://www.culturesofknowledge.org/?page_id=2319, accessed 20/03/2019, and the 'Bibliographie der Briefforschung': http://www.textkritik.de/briefkasten/forschungsbibl_a_f.htm, accessed 20/03/2019.

[61] See note 40.

[62] Alexander T. Bergs, 'A New Approach to Text Typology', in Nevalainen and Tanskanen, eds., *Letter Writing*, 27–46, at 27.

[63] Ibid., 29 and 33 respectively.

- report (= descriptive, neutral)
- request (= appellative, socially inferior to superior)
- orders (= appellative, socially superior to inferior)
- counsel (= descriptive)
- phatic (= phatic-descriptive-expressive).[64]

From his analysis of the Paston letters, Bergs concludes that '[i]n late Middle English, these socio-pragmatic subtypes do not necessarily differ in their overall structure and make-up, though they do show interesting and significant differences in their use of formulaic language, speech acts, and their functional elaboration of certain linguistic variables'.[65] '[S]ocio-pragmatic subtypes […]', he also acknowledges, 'may show a great deal of overlap and should thus be treated as non-discrete constructs'.[66] It is not only in this respect that Bergs's subtypes relate to historical letter genres as discussed above: as he himself points out, historical letter theorists such as Erasmus also adopted a classification of letters 'based on functional properties'.[67] Indeed, some of Bergs's subtypes appear very close to epistolary subgenres as defined above (see 4.1): e.g. his 'request' is similar to 'letters of petition', and Bergs himself identifies the 'Petition [i.e. 'request'] as belonging to the text type "letter"'.[68]

As Bergs acknowledges, one of the main limitations of his list of subtypes is that it is not comprehensive. Yet, the subtypes' simplicity and flexibility, which is partly due to their 'reflect[ing] universal language functions', as well as their inherent focus on social relationships also makes them a potentially useful tool.[69] In this context, Bergs seems to suggest that his typology could be transferable to other periods, languages, and cultures, and that Bühler's model could, for instance, be used to identify clusters of epistolary subgenres that share specific language functions.[70]

The typology of letter writing Claudio Guillén developed in his seminal *Notes toward the Study of the Renaissance Letter* also foregrounds certain formal linguistic and literary qualities as well as publication contexts of the epistolary form.[71] When surveying the field of the 'Renaissance epistle', Guillén writes, 'it becomes neces-

64 Ibid., see, for example, 34, table 1.

65 Ibid., 42.

66 Ibid., 37.

67 Ibid., 43, note 3.

68 Ibid., 33–4.

69 Ibid.

70 Ibid., 34, and 43, note 3 respectively.

71 Claudio Guillén, 'Notes toward the Study of the Renaissance Letter', in Barbara Kiefer Lewalski, ed., *Renaissance Genres: Essays on Theory, History, and Interpretation* (Cambridge, MA and London: Harvard University Press, 1986), 70–101.

sary to distinguish at least seven kinds of writing, each of which can be seen as following its own career and rhythm of development'.[72] His typology includes:

- the neo-Latin prose letter
- the prose letter in the vulgar tongue
- the neo-Latin verse epistle
- the verse epistle in the vernacular tongue
- the tradition of the theory of the letter
- practical manuals for letter writing
- letters inserted within other genres.

In relation to our previous discussion, this list could be understood as a typology of epistolary subgenres that is alternative to the classifications introduced so far (e.g. in Erasmus, or by Bannet). However, it can also be seen as a typology of epistolary genres that, while subordinate to the 'letter as genre' in general, are superordinate to epistolary subgenres as defined earlier (see 4.1). Two features support the second understanding: firstly, Guillén, in his discussion of 'the prose letter in the vulgar tongue', mentions the 'Latin letter' as well as the 'Humanist and philological letters' as kinds (or, in our terminology 'subgenres') of prose letters. Secondly, many of the letter genres mentioned earlier (e.g. scholars' letters) could be associated with several or all of Guillén's categories. In this situation, it is also helpful to remind oneself that, as Guillén says: 'A piece of writing can be a hybrid; and to the question of its generic definition the answer need not be […] either yes or no'.[73]

His typology has many merits: it is broad yet includes specific foci on language and form that allow us to capture both similarities between epistolary subgenres and their manifold incarnations in manuscript and print. The inbuilt attention to the trans- or, as Guillén terms it, 'supranational' existence of epistolary forms as well as to the close interrelationship of epistolarity and literature, is a particularly attractive feature.[74] However, although in theory this typology could be extended to include later material, it would not be finely grained enough to capture individual letter genres and their development across the whole of the early modern period in great detail.

The encyclopedic *Handbuch Brief: Von der Frühen Neuzeit bis zur Gegenwart* takes a more generalized approach: it aims at presenting an overview of current scholarly

[72] Ibid., 71.
[73] Ibid., 82.
[74] Ibid., 91.

knowledge on the 'real historical letter' in a systematized form.[75] Detailed information on epistolary subgenres can be found in two (out of four) sections of the book: Section III consists exclusively of entries on different letter genres, and Section IV (which focuses on the history of letter writing and letter cultures) includes discussions of further epistolary subgenres in the context of period-specific phenomena. Of particular relevance to early modernists are entries in Section III on billet; bridal letter; business letter; émigré's letter; illustrated letter/artist's letter; missionary's letter; last letter written before death (including the deathbed letter, suicide note, etc.); letters from exile; letter of mourning/consolation (including condolence letter); letter poem/heroid; letter to the editor; literary correspondence; love letter/erotic letter; open letter; patient's letter; petition; philosophical letter; prison letter; scholar's letter; travel correspondence; threatening letter/ransom note.

In Section IV, there are entries which focus on the period between the late fifteenth and the early nineteenth centuries and which thematize administrative and courtly correspondence; artists' correspondence; authors' correspondence; diplomats' correspondence; émigrés' correspondence; gallant letter; humanists' correspondence; letter theory; monastic correspondence; musicians' correspondence; philosophical letter; physician's letter; royal (and other noble) correspondence; scholar's letter; scientific correspondence; society letter; woman's letter.

Some of the letter genre designations used in the handbook rely on conceptualizations from the late seventeenth and eighteenth centuries, others reflect more recent (scholarly) concerns; entries in the previous section of this chapter describe letter genres from a long-term and 'historical-critical' point of view.[76] The handbook does not aim at presenting a comprehensive typology of letter genres; this is precluded by mainly three factors: (1) gaps in the current research landscape, and the facts that (2) entries originate from a range of disciplines (e.g. history, literary studies, linguistics) and (3) authors are encouraged to include innovative perspectives.

Although the handbook project will be an important step towards a comprehensive 'historical-critical' typology of early modern letter genres, the latter remains a project of the future. It should, *in addition* to the epistolary subgenres already listed in relation to the *Handbuch Brief: Von der Frühen Neuzeit bis zur Gegenwart* (and mindful of the classifications assembled by Bannet), include the following: biblical letter; didactic letter; circular letter; chancery letter; child's letter; classical letter; council letter; defamatory letter; diaspora letter; ecclesiastical letter; emperor's letter/edict; epistle (including epistle dedicatory); epistolary treatise; letter of advice; letter of complaint; letter of compliment; letter of congratulation; letter of excuse; letter of exhortation; letter of intelligence/news (i.e. newsletter); letter of introduc-

[75] The *Handbuch Brief* project comprises four volumes: one each on antiquity and the medieval period, and two on the time from the early modern period until the present day; the present discussion focuses on the latter.

[76] See 4.2.

tion/recommendation; letter of invitation; letter of merriment or raillery; letter of praise or commendation; letter of remonstrance/reproof; letter of thanks; letter of visit; letters close; letters of credit; letters patent; letters proffering assistance; letter to the reader; moral letter; missionary letter; papal bull/brief; polemical/satirical letter; report; royal letter (including e.g. *lettres de cachet*); secret letter.

4.4 Letter Genres in Metadata

The previous discussion of three scholarly approaches to classifying letter writing and epistolary subgenres or types has reminded us not only of the differences between disciplinary approaches to the epistolary form. The relative merits as well as limitations of these classifications also illustrate that capturing – in one neat model – all the many parameters which we can refer to in conceptualizing letter genres is difficult. Consequently, the inclusion of information on letter genres in descriptive metadata is also a challenge.

Yet, not least in view of the evident lack of scholarship on letter genres (including, for instance, in the field of genre theory), a systematic collection of information on epistolary subgenres in digital projects (e.g. databases, digital editions of correspondence, archives, etc.) appears particularly necessary and would unlock great research potential.[77] Including data on letter genres in digital projects would, to name just one example, facilitate locating source material (which is particularly valuable in cases of epistolary subgenres for which only few examples have survived). Altogether, it is fair to assume that the use of digital tools to collect as well as analyse information on letter genres would lift research in this field – both nationally and internationally as well as within and across disciplines – to new levels.

Mindful of these chances but also the challenges mentioned, I would like to suggest that information on letter genres should be tagged not in relation to a typology that is specific to a discipline, language, or culture, but instead with the help of a limited number of keywords. Transcending fixed letter genre definitions, these keywords would reflect the categories most commonly used both in historical and present-day conceptualizations: i.e. (1) speech acts and social actions (e.g. 'advice'); (2) major themes (e.g. 'love'); (3) social and/or institutional identity of author or context of letter (e.g. 'merchant', 'church'); (4) key linguistic, stylistic, formal features (e.g. 'expressive', 'gallant', 'verse'); and (5) miscellaneous aspects (e.g. 'illustrated'). Ideally, keywords such as these should be derived from as wide a range of letter material as possible and consolidated over time.

[77] For a most valuable contribution to letter (genre) theory more generally see Sinding, 'Letterier'; his essay is also instrumental to the 'keyword' solution suggested here.

Using these five dimensions, the large and varied genre of 'scholars' letters' (a more recent relative of a key early modern letter class – the 'learned letter') could, for instance, be tagged and recognized with keywords such as those below:[78]

1. 'inform', 'recommend', …
2. 'patronage', 'philosophy', 'science', …
3. 'university', 'courtly service', 'emigré', …
4. 'familial', …
5. 'containing bibliography/scientific objects', 'illustrated', …

Various types of scholars' letters – such as the 'epistle dedicatory' (or the 'letter to the reader', etc.) – could be differentiated with the help of other, sometimes additional keywords: e.g. (1) 'dedicate', (4) 'verse', 'epistle'. By the same token, more detailed information on publication contexts or text types could also be recorded, e.g. in dimension (5) 'poetry collection', 'spiritual'. Specific combinations of keywords would then allow users of databases to look for scholars' letters in general as well as, for instance, for epistles dedicatory only, or, even more specifically, for epistles dedicatory published in a poetry collection and/or a piece of spiritual literature.

To allow comparability across different databases, the sets of keywords to be used for recording genre information should be standardized and publicized to users of the database. The keywords should always be chosen with an eye to optimizing flexibility, and expandability. Of course, including this information would require additional labour and potentially create considerable sets of data. However, given the potential benefits of making this data available, which include a better understanding of letter genres and their interrelatedness with various types of texts, it seems definitely worth the effort. All this, of course, must be preceded by a consideration of how we model letters as metadata, which is the subject of chapter II.7 of this book.

[78] I would like expressly thank Howard Hotson and Thomas Wallnig for their input on this paragraph.

II.2 Place

Arno Bosse

1 Challenges

Historical places pose many difficulties for scholars not well resolved by existing digital resources.[1] Places can operate on a huge range of scales. At the micro scale within the city, a place may reference specific quarters, streets, buildings, and even individual rooms. At the macro scale, places such as villages, towns, and cities are themselves situated within larger units such as a county, a country, an empire or multiple monarchy, a continent, and a hemisphere. The accurate historical analysis of large sets of documents requires the confident assignment of records to place entities at various scales and the nesting of these places within one another. The ways in which these historical levels are nested, however, is not fixed. At every level – from the empire to the study – places appear and disappear continuously. Over time, those places that endure can pass from one dynasty, ruler, country, county, municipality, or religious confession to another. Accurate historical analysis requires the capacity to capture all these changes, at all of these different levels. Without this, we will be able to search for documents referencing individual cities (e.g. Oxford and Leiden) but not between larger polities (e.g. England and Hol-

[1] I would like to thank my colleagues Howard Hotson, Miranda Lewis, and Matthew Wilcoxson for their help in defining the goals and shaping many of the formative ideas described in this chapter. In particular, I would also like to acknowledge the contributions made by our development partners at the KNAW Humanities Cluster, and by our external consultants: Graham Klyne in creating the semantic data model for *EM Places*, and Glauco Mantegari, who designed its user interface. A full list of contributors to *EM Places* can be found on its repository page in *GitHub*. See https://github.com/culturesofknowledge/emplaces, accessed 20/03/2019.

land), still less between Castile and the Catholic regions of the Holy Roman Empire. To compound these difficulties, the names or toponyms given to individual places change over time, as do the shape, size, and location of the polities enclosing them.

A single example illustrates the practical difficulties involved.

Toponyms. A good example is the city known today by the Polish name of Wrocław, but previously known also as Breslau (in German), Vratislav (in Czech), and Vratislavia (in Latin). Further widely used variants include its name in Hungarian (Boroszló), Hebrew (ורוצלב or Vrotsláv), Yiddish (ברעסלוי or Bresloi), and Silesian German (Brassel). Breslau also gave its name to a duchy which in turn was part of the Duchy of Silesia (the Latin and English term: Śląsk in Polish, Schlesien in German, Schläsing in Silesian German, Slezsko in Czech, etc.).

Polities. From 1469 to 1490, Breslau was subject to the king of Hungary. From 1490 to 1526, it was incorporated (along with Bohemia, Moravia, and Lusatia) into the Kingdom of Bohemia, which formed part of the Holy Roman Empire. In 1526, the crown of Bohemia (and with it Silesia) passed to the Austrian Habsburgs, who ruled it for over two centuries with the brief exception of the period of the Bohemian Revolt (1618–20) and the occupation by Saxon and Swedish troops during the ensuing Thirty Years War (1618–48). After further occupation during the War of the Austrian Succession, Breslau was formally ceded with most of Silesia to the Kingdom of Prussia in 1742, which survived the collapse of the Holy Roman Empire in 1806. With the unification of Germany in 1871 it became part of the second German Reich. After the First World War, it became the capital of the Prussian Province of Lower Silesia during the Weimar Republic. During the Potsdam Conference after the Second World War, the Soviets insisted on transferring most of Silesia to Poland, where it remains today.

Locations. While the current geographical coordinates of the city of Wrocław are easily found in a gazetteer and are, broadly speaking, as valid today as they were in the sixteenth century, the same cannot be said for the geographic boundaries of the polities that enclosed it throughout its history. If we consider only Wrocław's political-administrative polities between 1500 and 1800, these were the County of Wrocław, the Duchy of Wrocław, the War and Domain Chamber of Wrocław, the Governorate of the Duchy of Silesia, the Province of Silesia, the Bohemian Crown, the Habsburg Monarchy, the Kingdom of Prussia, and the Holy Roman Empire. Very little reliable historical data exists on the geographic extents of these boundaries and even this data is limited to a small number of specific points in time. Accurate, temporally continuous spatial data for historical political-administrative, ecclesiastical, or other types of polity boundaries is rarely available anywhere in Europe before the second half of the nineteenth century.

To this first, broadly shared set of challenges,[2] we must now add a second set specific to working with historical places in correspondence. The three most important kinds of place metadata in correspondence are the origin of a letter, its destination, and the places mentioned in the body of the text. Because the place of writing or origin was typically noted in the body of the letter, we possess a great number of such references. The difficulty, however, is that in many cases, these references are indirect, incomplete, or simply idiosyncratic. Consider, for example, the following examples of places of origin recorded in the Tudor Domestic State Papers (1509–1603)[3]: 'From the Tower', 'On leaving Werthona', 'My lodging in Petty France', 'From his Majesty's Fort by Plymouth', 'At my gouty lodging', 'At Mr. Pettifer's over against the Red Cow', 'At sea'.[4] The decision to include actual (i.e. as written) or inferred places of this kind in an historical gazetteer or instead record the places as part of the work metadata is seldom straightforward. Letter destinations present a somewhat different challenge as very little data on them (i.e. the original address as written) is still available to us. In the rare instances where letters have survived unopened (as, for example, in the Dutch Brienne collection)[5] they share some of the same kinds of imprecision encountered for origins (e.g. 'Monsieur Vachon loge chez Monsieur pointy mestre Chirurgien a lenseigne de la Ville dorange proche le plain a la haye').[6] Finally, places mentioned in the body of a letter raise greater problems still, since here we now also need to decide whether to accommodate and disambiguate the vastly richer forms of indirect and figurative place references available in extended prose.

Two further place-related challenges of particular relevance to correspondence have to do with different kinds of information *associated* with a place, and with the ways in which two or more places are *connected* to each other. For example, it is very useful to know which calendar (e.g. Julian or Gregorian) was in predominant use in an historical region, as this information can later be drawn on to help infer how to assign a calendar to a letter correctly (these issues are covered in depth in chapter II.3, 'Time'). But recording the dates of transition from one calendar to another is

[2] For a recent review of shared requirements for historical geo-gazetteers, see Lex Berman, Ruth Mostern, and Humphrey Southall, 'On Historical Gazetteers', *International Journal of Humanities and Arts Computing* 5 (2011): 127–45, see https://doi.org/10.3366/ijhac.2011.0028. For a mapping of these scholarly requirements to Semantic Web technologies, see Karl Grossner, Krzysztof Janowicz, and Carsten Keßler, 'Place, Period, and Setting for Linked Data Gazetteers', in Lex Berman, Ruth Mostern, and Humphrey Southall, eds., *Placing Names: Enriching and Integrating Gazetteers* (Bloomington and Indianapolis: Indiana University Press, 2016), 80–96.

[3] See https://www.gale.com/intl/c/state-papers-online-part-i, accessed 20/03/2019.

[4] By my rough estimate, such place types make-up a good 10–15 per cent of all letter origins in the Tudor Domestic State Papers data set.

[5] *Early Modern Letters Online*, *Cultures of Knowledge*, 'Brienne Collection', http://emlo-portal.bodleian.ox.ac.uk/collections/?catalogue=brienne-collection, accessed 20/03/2019.

[6] 'Monsieur Vachon lodges with Monsieur Pointy, master surgeon at the sign of the Orange City near the plain at the Hague'. Planche, Marie (fl. 1690) to Vachon (fl. 1690) in: 'Brienne Collection', *Early Modern Letters Online*, *Cultures of Knowledge*, http://emlo.bodleian.ox.ac.uk/profile/work/bc606b51-ff87-4f15-8dde-655610519e28, accessed 20/03/2019.

contingent on the existence of adequately dated, hierarchically nested polities with which these data can be associated. In contrast to this, when recording data on postal networks, we want to be able to record the attributes of the connections between places such as distance, travel time, postage fees, and itineraries, and then draw on these data to calculate, for example, the cost of maintaining a regular correspondence of a specified size between London and Paris.

The primary requirements for a resource capable of fully addressing these challenges, therefore, are, firstly, to prepare a data model capable of capturing historical changes in toponyms and hierarchical polities in a manner that can also accommodate incomplete, uncertain, and approximate data, and, secondly, to implement that model in a Linked Open Data database which allows early modern researchers both to draw on current geospatial information and to enrich and add to this with correspondence-specific elements, thereby contributing new, well-attributed open access data sets for reuse and further refinement.

Modern, large-scale gazetteers such as *GeoNames*, the *Getty Thesaurus of Geographic Names* (TGN), and *WikiData* are invaluable resources for disambiguating, locating, and reconciling current, extant places. But none fulfils the requirements for a geographic resource for early modern correspondence.[7] Most crucially, none is currently able to provide a consistent chronology of the changing contexts that a given place has occupied throughout its history.[8] Funded by a grant from the Andrew W. Mellon Foundation, the *Cultures of Knowledge* project at the University of Oxford[9] has been working closely with the KNAW Humanities Cluster[10] in pursuit of a solution to this problem.

EM Places (Early Modern Places)[11] is a collaboratively curated, historical gazetteer for the sixteenth to eighteenth centuries, and represents the first of what will eventually become three Linked Open Data resources[12] also comprising '*EM*

[7] For a recent review of requirements for a historical geo-gazetteer, see Berman, Mostern, and Southall, 'On Historical Gazetteers'. For a mapping of these requirements to Linked Open Data features, see Karl Grossner, Krzysztof Janowicz, and Carsten Keßler, 'Place, Period, and Setting for Linked Data Gazetteers'.

[8] *Das geschichtliche Ortsverzeichnis* (http://gov.genealogy.net) gazetteer offers a rich, open-access data set with historical administrative-political boundaries. However, its pre-nineteenth-century data is very sparse, and coverage is limited to Europe. A commercial product, *EurAtlas: History and Geography of Europe* (https://euratlas.com) offers GIS shape files with polities and boundaries from the early medieval period to the present day, but these data cannot be shared publicly, and are only available at 100-year intervals. For some European countries, this gap is partially filled by digital resources such as the Danish *DigDak* historical gazetteer (http://www.digdag.dk), the Polish *Atlas Fontium* research project (http://atlasfontium.pl), or *A Vision of Britain Through Time* (http://www.visionofbritain.org.uk); all accessed 20/03/2019.

[9] See http://www.culturesofknowledge.org, all accessed 20/03/2019.

[10] See https://huc.knaw.nl, all accessed 20/03/2019.

[11] See https://github.com/culturesofknowledge/emplaces, all accessed 20/03/2019.

[12] See http://www.culturesofknowledge.org/?p=8455, accessed 20/03/2019.

People' and *EM Dates*[13] built on a shared humanities infrastructure platform developed at the Humanities Cluster[14] of the KNAW in Amsterdam.

EM Places is being designed to meet four goals, selected in consultation with members of the COST Action *Reassembling the Republic of Letters* as well as further meetings and discussions with early modern and geospatial researchers.[15]

The first is to provide *a resource for identifying early modern places* by means of their historical name variants as well as their current names. This is necessary to ensure that searches for historical name variants can readily be resolved to their current names and locations. To this end, *EM Places* will combine current and alternative place names with a current administrative hierarchy and location data from a small number of reference gazetteers and then add further place name attestations provided by contributors from primary sources. Users will be able to browse and search for places on multiple criteria, refine their results over facets, and export their search results. To facilitate the semi-automatic cleaning and disambiguation of bulk metadata, *EM Places* will function as a reconciliation service for *OpenRefine*.[16]

The second is to provide *a means to contribute richer historical contexts to places*. *EM Places* will provide means for capturing basic historical data on political-administrative and ecclesiastical hierarchies (in the future, potentially also military and judicial hierarchies) and where available, georeferenced historical maps and related resources and bibliographies. It will also offer means to record the dates of transition between official calendars (e.g. from the Julian to the Gregorian) in an historical region for reuse in *EM Dates* (see ch. II.3 for a discussion of *EM Dates*). Custom attributes describing 'connections' between places (e.g. the time required for mail to pass between two stations on a named postal route) will also be supported.

The third is to *source and credit fully all contributions to the gazetteer*, whether by individual researchers or project teams. Regular contributors with registered accounts on *EM Places* will be able to submit new data or suggest revisions to existing data using either a web interface or a bulk upload facility. More experienced users with editorial privileges will have the means to review and approve these contributions. Users will be able to see whether data in the gazetteer originated from a reference gazetteer such as *GeoNames* or was added by an individual contributor. Contributors to *EM Places* will be able to call up a listing of their own additions and revisions to the gazetteer.

---

13 See https://github.com/culturesofknowledge/emdates, accessed 20/03/2019.

14 See https://huc.knaw.nl, accessed 20/03/2019.

15 The 'Space and Time' working group of the COST Action IS1310 *Reassembling the Republic of Letters* (see http://www.republicofletters.net/index.php/working-groups/space-and-time/, accessed 20/03/2019) and further presentations and discussions at conferences and workshops in the United Kingdom, Italy, and the United-States.

16 See https://github.com/OpenRefine/OpenRefine/wiki/Reconciliation, accessed 20/03/2019.

The fourth and final goal is to *make the EM Places source code and data sets easily accessible and reusable by others*. To this end, the source code for *EM Places*, based on the *Timbuctoo* infrastructure[17] developed by the KNAW Humanities Cluster, will be shared under open source and made available for reuse in virtual containers. The data in *EM Places* will be shared under open-access licences and distributed over multiple channels: as user-initiated exports of individual records from the application itself, on popular open repositories such as *GitHub*, and via the *EM Places* GraphQL API. The intent is to prepare the gazetteer in as transparent and collaborative manner as possible to allow it to become a useful resource for the early modern research community and an active participant in the *Pelagios Commons* network.[18] In support of this, in addition to tabular and RDF formats, *EM Places* will export its data in the new Linked Places[19] Geo-JSON gazetteer interconnection format for reuse by the *World Historical Gazetteer*.[20]

EM People is currently in development at Oxford and Amsterdam and is expected to be released as an initial pilot in Autumn 2019. The next section presents an overview of the key concepts underlying the *EM Places* data model and the proposed workflows for adding and revising data to the resource. This is followed by a section outlining the major functionality of the gazetteer based on a review of a mock-up of the interface for a single place record.

2 Data Model and Workflow

2.1 Data Model: Places, Names, and Locations

EM Places is deeply indebted to the *Pleiades* gazetteer of ancient places for its key concepts of place, location, and name.[21] As in *Pleiades*, a 'place' in *EM Places* is a geographical and historical *context* for an entity 'constructed by human experience'. A place, therefore, is an abstract entity and does not in itself require either a spatial or a temporal reference. In this way, the notion of 'place' is able to accommodate, where required, unnamed, unlocated, and no longer extant places. To quote from the *Pleiades* documentation, 'Places may be no larger than a family dwelling or as big as an empire, be temporally enduring or fleeting. They may expand, contract, and evolve over time'.[22]

A 'location' in *EM Places* is a spatial reference, connecting a place to a geographical context. A location, when known, will typically be represented as a lati-

[17] See https://github.com/HuygensING/timbuctoo, accessed 20/03/2019.

[18] See http://commons.pelagios.org, accessed 20/03/2019.

[19] See https://github.com/LinkedPasts/linked-places, accessed 20/03/2019.

[20] See http://whgazetteer.org, accessed 20/03/2019.

[21] See https://pleiades.stoa.org/help/conceptual-overview, accessed 20/03/2019.

[22] Pleiades, 'Technical Introduction to Places', see https://pleiades.stoa.org/help/technical-intro-places, accessed 20/03/2019.

tude/longitude pair, or else as a polygon region, constructed from multiple geo-coordinates, though other representations are also possible. Locations are always accompanied by time-spans. Many place records in *EM Places*, in particular those of historical polities, will have unknown, uncertain, and approximate locations which themselves have changed over time (consider for example, the unsettled boundaries of principalities in central Europe during the Thirty Years' War).

Finally, a 'name' is a textual reference belonging to a place. In *Pleiades*, as in *EM Places*, a place may have one or several names, or even no name at all. A name, however, can only be of one place. Identical names referring to different places are treated in *EM Places* as separate entities. Just like locations, names are always accompanied by their time-spans.

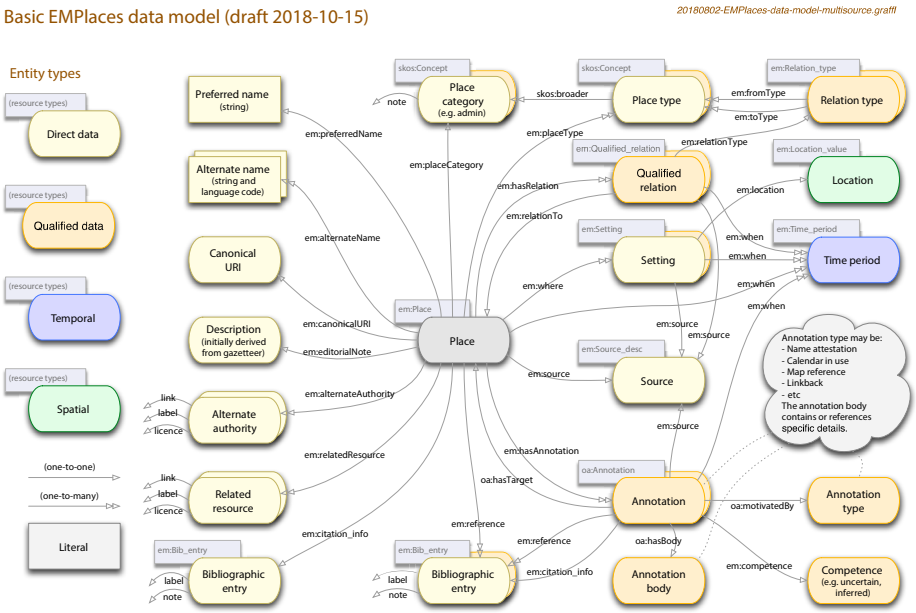


Figure 1: *EM Places* data model overview

The overview entity-relationship diagram reproduced in figure 1 illustrates how the *EM Places* data model incorporates and builds on the concepts of place, location, name, and time period from the *Pleiades* gazetteer. Two aspects of the *EM Places* model are worth highlighting very briefly here. The model has been designed to capture historical time periods either by reference to (exact or approximate) time-spans or by reference to a named period in an external period (e.g. *PeriodO*)[23] or

[23] See http://perio.do, accessed 20/03/2019, accessed 20/03/2019.

date (e.g. *GODOT*)[24] gazetteer. All scholarly assertions in *EM Places* can be qualified. In the current draft of the data model, the values available for this are: 'Definitive', 'Inferred', 'Assumed', 'Uncertain', and 'Approximate'. Further documentation with details on the data model and its sub-components can be found in the 'Models' section of the *EM Places GitHub* repository.[25]

2.2 Data Workflows

From a workflow perspective, *EM Places* place records can be divided into two classes: (1) a required, minimal set of *core metadata* usually ingested from a group of reference gazetteers; and (2) an optional, larger set of predominantly historical, *supplementary metadata* provided by contributing scholars and projects. From a data modelling or data semantic perspective, there is no difference between core and supplementary metadata. Drawing this distinction is nonetheless helpful for explaining the organization of data in the gazetteer, and its appearance in the user interface.

Core metadata. When a new place record is first added to *EM Places*, we attempt to locate that place (if available, via its authority ID) in *GeoNames*, our primary reference gazetteer. If the place is found, we add to *EM Places* a set of core metadata from the reference gazetteer, supplemented by a small number of additional authorities (currently, the *Getty TGN* and *WikiData* – more authorities may be added in the future). Slightly simplified here for clarity, the data retrieved from these sources consists of:

- the preferred (i.e. default) name for this from *GeoNames*
- a merged list of alternative and variant place names from *GeoNames*, *Getty TGN*, and *WikiData*
- location (i.e. geo-coordinate) data from *GeoNames*
- the current administrative-political hierarchy (polity) and place type from *GeoNames*
- a list of matching place authority IDs from *WikiData.*

The information we ingest into core metadata from these sources is assumed to be current and authoritative. For this reason, and to simplify subsequent updates, it is imported into *EM Places* without modification and will normally not be revised or edited again. Drawing on this data, we are then able to represent its location on an *OpenLayers*[26] map, generate a unique and persistent Linked Data URI for the record, and provide means to cite the record with data on its original creator, subse-

[24] See https://godot.date, accessed 20/03/2019, accessed 20/03/2019.

[25] See https://github.com/culturesofknowledge/emplaces/tree/master/models, accessed 20/03/2019.

[26] See https://openlayers.org , accessed 20/03/2019.

quent contributor(s), and the open-access licence(s). In a normal workflow, this process will be scripted and applied semi-automatically to a large number of incoming places from a contributing project.[27]

If, however, the place we wish to add cannot be found in *GeoNames*, a new record is added manually by an *EM Places* editor. This could be necessary when, for example, an early modern village, or a building, cannot be found in our reference gazetteer. Historical entities no longer extant, such as principalities, will always need to be entered manually since these are not recorded in *GeoNames*. Our expectation is that for the initial release of *EM Places* in 2019 we will seed the gazetteer with c. 3,500 core metadata records drawn from previously disambiguated and reconciled place records in *Early Modern Letters Online* (EMLO). Based on an analysis of places of sending and receipt in EMLO, we are confident that this number of records will already cover the vast majority of places recorded in EMLO's catalogue.

Supplementary metadata. All other metadata fields in *EM Places* may be characterized as supplementary metadata which can be optionally appended to a basic, core metadata place record. Supplementary metadata can be contributed to *EM Places* by individual users, research projects, and partner institutions. This type of metadata will include:

- attestations of additional toponyms, or variant attestations of existing toponyms
- links to geo-referenced historical maps
- a short, textual description of a place
- data on historical political-administrative, ecclesiastical (and in future releases, potentially also judicial and military) hierarchies
- date(s) of transition between calendars in a region (e.g. from the Julian to the Gregorian)
- data on connections or associations between places (e.g. the distance between stations on a postal route)
- links to related historical resources
- bibliographical references.

A key prerequisite for the future success of *EM Places* will be our ability to attract large numbers of individual contributions of supplementary metadata from researchers. This is a further reason for first seeding the gazetteer with a large number of basic place records. A scholar wishing to contribute supplementary metadata will be more inclined to do so if the place in question was already found in the gazetteer and does not first need to be created from scratch. Moreover, the principal motivation for doing so will be to fill gaps in the existing data which an indi-

[27] See https://github.com/culturesofknowledge/emplaces/tree/develop/src/geonamesdataexport, accessed 20/03/2019.

vidual scholar or project needs to map or analyse a new data set; but adding such material will only be feasible and attractive if the gaps to be filled are relatively small. We expect that in most circumstances, researchers will contribute a small number of records, so that these can be made via the *EM Places* webform. However, the gazetteer will also have a facility for ingest bulk metadata contributions in tabular CSV/Excel and in RDF format. Both features are derived from functionality already present in the underlying *Timbuctoo* infrastructure.

EM Places will also be able to leverage *Timbuctoo* for user and group permissions (roles) for different aspects of its workflow. The expertise to evaluate contributions of detailed historical data cannot be expected to reside at just one project or institution. Nor would their staff be able to cope with the flood of potential contributions, revisions, queries, and comments from across the relevant academic and research communities. To address this, *EM Places* has prepared a hierarchical permissions model for its workflow. In essence, this distinguishes among public users (who have read-only access), contributors with user accounts (who also obtain the ability to suggest the addition of new place records and revisions to existing supplementary metadata), and editors (who are able to create new place records and approve contributions). Our plan is first to provide editorial access to a small number of trusted partners in our scholarly communities, and later to extend this group by granting the same privileges to our most reliable and regular contributing users. This is a variation of the arrangements already applied and tested in EMLO for epistolary metadata in general.

3 Features and User Interface

This section reviews many of the key features of *EM Places* with the aid of an interface mock-up of the detail view for a sample place record. The place selected for this example, Opole, is a town in Poland with authority records in *GeoNames*, the *Getty TGN*, and *WikiData*. A full-size, high-resolution mock-up of the record detail view for Opole with sample data is reproduced at the end of this chapter and is available as a downloadable PDF file in the 'images' sub-directory of the *EM Places GitHub* repository.[28]

In the mock-up, all interactive text elements are shown underlined. For example, in the 'Citation' section, clicking on the 'MLA' link will copy an MLA formatted version of a citation to the current place into the computer's clipboard.

[28] See https://github.com/culturesofknowledge/emplaces/blob/master/images/current_display.pdf, accessed 20/03/2019.

3.1 Core Metadata: Preferred and Alternative Names, Administrative Hierarchy, Location, Citation, Persistent URI, Authorities

As noted above, we expect that *EM Places* will initially be seeded with a large number of basic records consisting only of merged core data from multiple reference gazetteers. It will take time for these records to be fleshed out with supplementary metadata. In practice, it is likely that for many years the majority of records in the gazetteer will consist only of core data. Besides this, not all users are interested in the same kind of data. We believe that most users will want to draw on *EM Places* as a way to disambiguate and reconcile their own place metadata. For these reasons, all the data fields with core metadata are shown in the top left-hand corner of the screen, in the most prominent position on the interface. If no further metadata is available, no further information on a record will be shown beside (see below) the record's provenance data and links to the gazetteer's export features.

Alongside the larger, highlighted *Preferred Name* of the place, we show a merged list of all unique *Alternative Place Names* found in our reference gazetteers. While useful for disambiguation purposes, these may not include all of the known historical, early modern toponyms. As we will see, metadata on further name variants can be added and viewed in the 'Name Attestations' section.

The *Administrative Hierarchy* section illustrates the partitive (i.e. child–parent) relationships pertaining to a place. In our sample record, the town of Opole is the seat of a third-order administrative polity (called Opole), which is part of a second-order polity (also called Opole), which in turn is part of the Opole Voivodeship. The hierarchy then concludes at the country level, in Poland. Gazetteers organize and model their polities in different ways. For example, in the *Getty TGN* entry for Opole,[29] there is only one polity (Opolskie) between the town of Opole and Poland. In addition, the *Getty TGN* hierarchy does not terminate in Poland but in 'Europe' and 'World'. Poland itself is not defined as a 'Country' in the *Getty TGN* but as a 'Nation'. These and many other differences are the result of the gazetteers being based on two conceptually different systems of modelling geographic data.[30]

To avoid having to arbitrate and resolve these conflicts manually across many thousands of place records, we have opted to build our Administrative Hierarchy exclusively from a (significantly larger data set) maintained by *GeoNames*. Unfortunately, due to the great number and variety of polities *GeoNames* needs to aggregate, it makes no attempt to label their different levels. Instead, these are shown with their generic *GeoNames* feature codes (e.g. ADM1, PPL, PPLA).[31] In the

[29] See http://www.getty.edu/vow/TGNFullDisplay?find=Opole&place=&nation=Poland&subjectid=7007751&english=Y, accessed 20/03/2019.

[30] For a detailed discussion of these differences, see Janowicz Krzysztof and Carsten Keßler, 'The Role of Ontology in Improving Gazetteer Interaction', *International Journal of Geographical Information Science* 22:10 (2008): 1129–57, see https://doi.org/10.1080/13658810701851461.

[31] See http://www.geonames.org/export/codes.html, accessed 20/03/2019.

smaller, manually curated *Getty TGN* gazetteers, the polities are labelled (e.g. district, county, province…).

Location data will consist in most instances of a single decimal latitude/longitude coordinate pair. Modern geo-coordinates are drawn up using the WGS84 geodetic standard with Greenwich as prime meridian. However, in the early modern period (and well into the nineteenth century) multiple, competing prime meridians were in active use.[32] Among the most important of these were the El Hiero (Ferro Islands), Cádiz, and Paris meridians. To raise awareness of their presence, and to help disambiguate variant latitude and longitude references in historical maps and other primary documents, users will have the option of viewing the location coordinates on the basis of several early modern prime meridian systems as well as in degrees, minutes, and seconds.

A *Persistent URI* for each place record in *EM Places* will be generated automatically by the underlying *Timbuctoo* infrastructure for reuse by others to reconcile their places against *EM Places*, and, crucially, as the basis for sharing our records as Linked Open Data. In the interface mock-up, this is currently represented as an ARK persistent identifier.[33] Drawing on this persistent identifier, we plan to offer a simple means to reference an *EM Places* record in several standard bibliographic *Citation* formats.

Finally, the *Authorities* section includes a list of all matching reference IDs found for the place in our three reference gazetteers. More precisely, the list is made up of a subset of relevant authorities for our users. We expect that, in time, and in response to feedback from our users, this list will grow to encompass additional authoritative gazetteer sources.

3.2 Core and Supplementary Metadata: Map and Description

If location data is available for a place, it will by default be shown on an *OpenLayers map*. If, in the case of a polity, additional polygon data is available, then the polity will be shown instead as a region. If supplementary, historical georeferenced data is also available, the modern representation may be accompanied by a limited number of historical cartographic representations from a source such as the *David Rumsey Map Collection*.[34] The different maps will be identified, in the first instance, by date of publication, with further metadata recorded in their respective provenance fields. A future release of *EM Places* may additionally draw on an open-source IIIF viewer to display historical maps published to the International Image Interoperability Framework (IIIF) standard.[35]

[32] See https://en.wikipedia.org/wiki/Prime_meridian, accessed 20/03/2019.
[33] See http://n2t.net/e/ark_ids.html, accessed 20/03/2019.
[34] See https://www.davidrumsey.com, accessed 20/03/2019.
[35] See https://iiif.io, accessed 20/03/2019.

To help further disambiguate places from another, and to offer more context for the Elasticsearch[36] search engine used by *Timbuctoo*, each place record will be seeded (where available) with a short text *Description* extracted automatically from the first paragraph of its entry in *Wikipedia*.

3.3 Supplementary Metadata: Name Attestations, Calendars, and Associated Places

Name Attestations are documented instances of variants of the preferred and alternative names in core data. For example, a reference to Siena, Italy, recorded in manuscript as 'Ciena' not already listed as an alternative toponym in our reference gazetteers can be recorded here, along with its language, date, authority, and source. *EM Places* editors are not obliged to merge or select among conflicting attestations provided to us by our contributors. Such differences can be recorded in the provenance metadata for the attestation and evaluated independently by the users of the gazetteer.

The *Calendars* element is a simple visualization of the predominant calendars (Julian, Gregorian) known or inferred to be in use by the political-administrative authority of a place between 1500 and 1800. This data will at first be automatically pre-populated via a hierarchy of inherited calendars. In the absence of specific information for a region, we will assume that it transitioned from the Julian to the Gregorian calendar on the date set out in Gregory XIII's papal bull (4 October 1582). However, if any parent region in a polity hierarchy transitioned to the Gregorian calendar at a later date, then all child place entities under it (e.g. subordinate regions, or individual towns and cities) are assumed also to have inherited this later date and the default 1582 transition date is overridden. In practice, we will not attempt to record variances from regional calendars used by authors at smaller scales such as individual towns. With few exceptions, the date of transition for an inhabited place will be inherited (i.e. inferred) from its superordinate place entity. In the mock-up, we see that the transition from the 'new style' (i.e. New Year on January 1) Julian calendar took place on 29 January 1584 and that this date is inferred for the town of Opole (and thus from the Duchy of Opele, unless it too inherited the date from its superordinate entity). For a discussion of the issues accompanying calendrical transitions and the reuse of the data recorded in *EM Places* by the *EM Dates* calendrical conversion service, please see chapter II.3, 'Time'.

The *Connections* section is designed to link to additional data on the relationship between two or more places. To accommodate this, we are considering providing a summary list of such relationships showing the name of the related place, place type, and association type with an optional link to an external resource offering additional data on the association. Alternatively (as shown in the current mock-up)

[36] See https://www.elastic.co/products/elasticsearch, accessed 20/03/2019.

this information would be included in the detail view of the relevant place (here, 'St Adalbert Church'). However, the data model, features, and appearance of 'Associated Places' in *EM Places* have not been finalized and remain under active discussion.

3.4 Supplementary Metadata: Historical Hierarchies

In the initial release of *EM Places*, the *Historical Hierarchies* section will display the administrative-political and ecclesiastical hierarchies for an historical place.[37] In the sample data set for Opole, we can see, for example, that the Silesian town of Opole was subordinate to the Duchy of Opole from 1281 until 1521. Further superordinate polities are shown above it, terminating with the Holy Roman Empire. The sub-menu of distinct hierarchies shown in the interface is created automatically from the contributed source data. By reviewing each partitive parent–child relationship we find the last shared date in the hierarchy in which the relationships began, and likewise, the earliest shared date in which it ended, and so we derive a period in which the relationship was true for all entities. In the example shown in the mock-up, none of the partitive relationships began later than 1442 and none ended earlier than 1521. Therefore, the period during which all the entities in the hierarchy stood in this relationship was 1442–1521. After 1521, an element of this set of relationships changed, requiring the construction of a new hierarchy (1521–6). This process is repeated until the termination date of the last valid hierarchy falls outside the scope of *EM Places* (1500–1800).

Note that each underlined entity in the hierarchy 'graph' is a fully-fledged place in the *EM Places* gazetteer. Clicking on it is designed to take the user to its own detail view page. Unlike places imported from our reference gazetteers, historical entities have no 'current' administrative hierarchy. In the current draft interface mock-ups, this section of core data is replaced (where available) by a reference to the historical hierarchy section.

We anticipate that political-administrative (and to a lesser extent, due to a potential multiplicity of denominations, ecclesiastical) data will be easier to find and more useful to most researchers than data on judicial and military hierarchies. In recognition of this, the initial release will focus on administrative and ecclesiastical polities.

[37] The conceptual design and user interface for Historical Hierarchies in *EM Places* was inspired by the *Das geschichtliche Ortsverzeichnis* (http://gov.genealogy.net/) gazetteer. See, for example, the graph representing the administrative, ecclesiastical, military, and judicial hierarchies for Ballum, Germany. (http://gov.genealogy.net/item/show/BALLUMJO45IB), all accessed 20/03/2019.

3.5 Supplementary Metadata: Related Resources, Bibliography, and Feedback

Related Resources provide a means for listing additional, predominantly digital resources related to a place which cannot be usefully represented in a traditional bibliography. A good example of a related resource would be a link to a live search carried out on external databases. For example, a link to a search of all letters in *Early Modern Letters Online* sent to or from a place, or to its mentions in full-text resource such as the ARTFL *Encyclopédie*.[38]

The *Bibliography* section is intended for listing traditional, offline scholarly publications and references. In the initial release of *EM Places*, this will be a simple formatted text list. However, in the future, we would like contributors creating the list to be able to verify easily if an entry had previously been added to a different place record, and to select this, so that as many entries as possible can be added in an efficient and consistent manner. Contributors should also have a means to add a bibliographic entry to a class of records (e.g. to have an entry appear in all places such as towns which are subordinate to the current place in a political-administrative hierarchy).

3.6 Supplementary Metadata: Creator, Contributors, Licence, and Export

Our intent is to distinguish and separately credit the (single) creator of a place record, and the potentially multiple contributors of (in most instances, supplementary) metadata to that record. Although *Timbuctoo* provides the technical means to record and display as a contributor each user who makes any change to a data field, this would give equal credit to someone contributing original data, and someone making a trivial change to the data. Moreover, not all new contributions to the gazetteer are equal (e.g. adding a new, fully sourced name attestation versus adding a new bibliographic entry). The decision as to which types of edits should count as contributions will need to be determined and implemented under an editorial policy.

EM Places is designed to be able to incorporate data from multiple sources, not all of which will be shared under the same open-access *Licence*. Thus, for example, data from *GeoNames* is licensed under a public-domain, CC0[39] licence, while supplementary data contributed by a researcher will need to be shared under a licence that requires attribution, such as CC-BY.[40] Which, and how many, of these licences to display here for users who wish to export data from *EM Places* will likewise need to be settled under an editorial policy.

[38] See http://encyclopedie.uchicago.edu, accessed 20/03/2019.

[39] See https://creativecommons.org/publicdomain/zero/1.0/, accessed 20/03/2019.

[40] See https://creativecommons.org/licenses/by/4.0/, accessed 20/03/2019.

Users will have the opportunity to *Export* the open-access data in *EM Places* on-demand for reuse. We will provide support for sharing this data in tabular formats (CSV, Excel), as Linked Open Data (Turtle/RDF), and in the temporal GeoJSON format employed by the recently defined 'Linked Places'[41] gazetteer interchange format used by the in-development *World Historical Gazetteer*.[42] In the context of a single record shown in the Opole mock-up, an export will only include the data for one place record. In the *EM Places* Search and Browse interface, users will additionally have the opportunity to export bulk records matching all or a selection of their search criteria.

4 Current State and Future Plans

Much work remains to be done in Oxford and Amsterdam before the anticipated launch of the pilot version of *EM Places* in 2019. The project's live status can be tracked on *GitHub*. Our focus in Spring and Summer 2019 will be on creating sample data to test the data model, and in completing the specification for the browse and search functionality and the webform for editors and contributors. Parallel to this, work will begin on completing the *EM Places* APIs (including the planned integration with *EM Dates*), customizing the import and export functionality, and implementing user and group permissions for contributor and editor workflows.

It should perhaps be emphasized in conclusion that, although *EM Places* has been devised in the first instance to handle geographical entities derived from letter texts and metadata, our intention from the outset has been to create a resource useful for much wider application in the early modern period (and, with modification, for earlier and later periods as well). Designing, implementing, and above all populating such a resource is not feasible for individual projects even for relatively limited domains. The more research communities share this infrastructure, the more likely it is to be populated with a comprehensive and reliable data set. Some of the cognates field to which this resource could be applied, and from which it could also be populated, are suggested in the conclusion to this volume (section V).

[41] See https://github.com/LinkedPasts/linked-places, accessed 20/03/2019.

[42] See http://whgazetteer.org, accessed 20/03/2019.

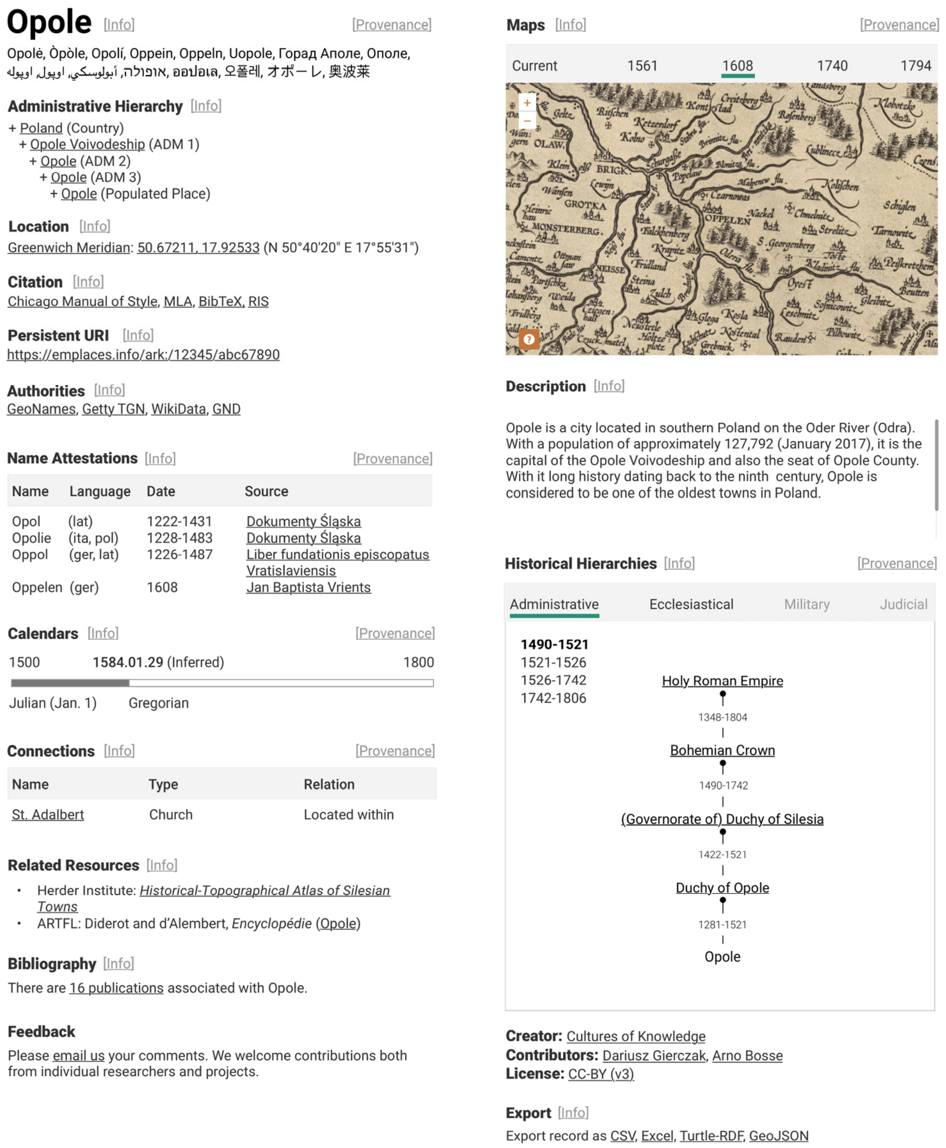


Figure 2: Mockup of *EM Places* record detail

II.3 Time

Miranda Lewis, Arno Bosse, Howard Hotson, Thomas Wallnig, and Dirk van Miert[1]

The heyday of the republic of letters during the sixteenth, seventeenth, and eighteenth centuries witnessed a profound and protracted upheaval in European timekeeping. The centrepiece of this upheaval was the transition between two calendars (1.1): the Julian calendar – devised under the Roman statesman Julius Caesar and in use throughout the Christian world during the Middle Ages – and the modern Gregorian calendar – devised under the Roman pontiff Gregory XIII and propagated in October 1582, but only adopted in Protestant Europe by fits and starts thereafter, in some cases as late as the mid-eighteenth century. Simultaneously, a number of factors further complicated this calendrical change (1.2). For one thing, the Julian calendar was interpreted and expressed in several different ways. At the same time, alternative chronologies were employed based on the Christian ecclesiastical calendar, and on papal or on regnal years. Still further calendars were in use among non-Christian communities in various times and places. Within such a complicated landscape, inferring calendar usage is a difficult problem in its own right (1.3). Even more intractable is the problem of how to handle dates that are incomplete, uncertain, or lacking altogether (1.4).

The simultaneous use of so many different ways of expressing dates has caused confusion to contemporaries and headaches for historians. These problems are aggravated for anyone attempting to assemble a union catalogue of correspond-

[1] With thanks to Anna Skolimowska for her helpful examples regarding the dates used in the correspondence of Ioannes Dantiscus, and to Jeannine de Landtsheer for her comments.

ence embracing the whole of Europe during this period of tumultuous chronological transition. An individual correspondence, if brief in duration and local in scope, poses relatively few problems of this kind, and those that do occur can be resolved on a case-by-case basis in free-text annotations. However, a union catalogue encompassing early modern Europe in its entirety needs to confront the problem of the simultaneous use of multiple timekeeping systems in all its complexity. More specifically, a digital catalogue must overcome three interrelated problems if the data emerging from it are to be precise enough to satisfy scholars and unambiguous enough to be analysed and visualized computationally. First, rigorous means must be developed for reconciling all these forms of timekeeping to a baseline standard (2.1). Secondly, computational methods must be devised to determine when individual territories transitioned from one calendar to the next (2.2). Thirdly, in order to pre-process large new sets of incoming raw data, methods are needed to assign calendars to individual letters in a provisional fashion, pending closer editorial scrutiny (2.3). Fourthly, means must be developed for expressing, analysing, and visualizing the chronology of letters that remain uncertain or incomplete (2.4). The first section of this chapter addresses each of these four problems in turn, and the second section proposes solutions to them.

1 Problems

1.1 The Simultaneous Use of Julian and Gregorian Calendars

Miranda Lewis and Dirk van Miert

Prior to the mid-sixteenth century, Europeans shared a calendar sanctioned by ancient pedigree and long usage. The Julian calendar was the result of a reform of the Roman calendrical system instituted by Julius Caesar shortly after his conquest of Egypt. On the advice of the astronomer Sosigenes of Alexandria, Julius's reform prescribed a common year of 365 days divided into twelve months. Since ancient astronomers reckoned the actual solar year to be 365 and one-quarter days, the calendar was synchronized with the sun by adding one extra day every fourth year, known then as an *annus bissextilis* and now as a 'leap year'.[2]

Although an improvement on previous reckonings, this ancient measurement of the year was still slightly inaccurate: rather than 365 days and 6 hours, the average tropical year is in fact 365 days, 5 hours, 48 minutes, and 45 seconds. Whilst imperceptible initially, this difference of 11 minutes and 15 seconds accumulated gradually over time. Each 128 years, the Julian year moved out of step with the

[2] The solar year is the time that the sun takes to return to the same position in the cycle of seasons, for example, from vernal equinox to vernal equinox, or from winter or summer solstice to its equivalent.

solar year by one additional day.[3] As centuries passed, the annual cycle of spring and autumn equinoxes and the summer and winter solstices began to shift across the calendar, wreaking havoc with the seasons and the ecclesiastical calendar. By the sixteenth century, the vernal equinox was occurring around 11 March rather than 21 March, Easter was sliding towards summer, and Christmas towards spring.[4]

The displacement of liturgical festivals within the established calendar was of particular concern to the Catholic Church. During the pontificates of Pope Paul III (d. 1549) and his successors, leading Italian astronomers considered the problem, among them Aloysius Lilius (c. 1510–1576). Recommendations for a solution continued under Pope Gregory XIII (d. 1585), led by the German Jesuit astronomer and mathematician Christopher Clavius (1538–1612). Clavius's work produced a more precise measurement of the length of the average year as $365\,{}^{97}/_{400}$ or 365.2425 mean solar days; and on this basis he advised that the calendar could be kept synchronized with the seasons simply by skipping three Julian leap days in every 400 years.[5] In addition, to resynchronize the solstices and equinoxes with the seasons, ten days needed to be removed from the year in which the new calendar was instituted. The result of this work is the most widely employed civil calendar in use today, and is known as 'Gregorian' after its patron. Announced on 24 February 1582 in the papal bull *Inter gravissimas*, the transition from the Julian to the Gregorian was mandated to take place in October 1582, with Thursday, 4 October (Julian) to be followed immediately by Friday, 15 October (Gregorian).[6]

Even within the Catholic world, the adoption of the new calendar was not instantaneous. While a number of Catholic countries – including some Italian states,

---

[3] To make matters worse, the lunar calendar, used for calculating the date of Easter, was even more inaccurate. Bonnie Blackburn and Leofranc Holford-Strevens, *The Oxford Companion to the Year. An Exploration of Calendar Customs and Time-reckoning* (Oxford: Oxford University Press, 1999), 682, explain: 'whereas after 76 years the solar calendar was about three-quarters of an hour behind the sun, the lunar calendar was nearly 6 hours behind the moon'.

[4] Martin Luther noted that, in 1538, Easter should not have been celebrated on 21 April but rather five weeks earlier on 17 March. He considered reform to be the concern of the Christian princes, who should act in a united fashion to prevent confusion arising both in everyday events, such as traditional markets, and secular business. See Blackburn and Holford-Strevens, *The Oxford Companion to the Year*, 632.

[5] See August Ziggelaar, 'The Papal Bull of 1582 Promulgating a Reform of the Calendar', in George V. Coyne, Michael A. Hoskin, and Olaf Pedersen, eds., *Gregorian Reform of the Calendar. Proceedings of the Vatican Conference to Commemorate Its 400th Anniversary, 1582–1982* (Vatican, 1983), 201–39, and P. Kenneth Seidelmann, ed., *Explanatory Supplement to the Astronomical Almanac* (Sausalito, CA: University Science Books, 1992). The approximation of $365\,{}^{97}/_{400}$ is achieved by having ninety-seven leap years every 400 years. The papal bull stipulated that a year should become a leap year if its number were divisible by four or by 400 but not by 100 unless it may also be divided by 400. The position of the extra day in the leap year was moved from the day before 25 February to the day after 28 February. In addition, new rules for the calculation of Easter were adopted.

[6] The bull was displayed on the doors of St Peter's on 1 March 1582. For the full text, see Christoph Clavius, *Romani calendarii a Gregorio XIII. P. M. restituti explicatio S. D. N. Clementis VIII. P. M. iussu edita. Accessit confutatio eorum, qui calendarium aliter instaurandum esse contenderunt* (Rome: apud Aloysium Zannettum, 1603), 13–5.

Spain, Portugal, and Catholic parts of Poland – followed the papal instructions immediately, removing ten days from October 1582, France, delayed by objections in the Paris *parlement*, removed the ten days between 9 December and 20 December 1582. However, with only months to make arrangements, and with calendars for 1583 already compiled and printed, not all countries made the change as requested. Many, especially those at a distance from Rome, were slow to introduce reform. In the Germanic lands, this new calendar was adopted at varying points in the 1580s, with Catholic states leading the way between 1583 and 1585. Prague arranged adjustments in January 1584, as did the Catholic cantons of Switzerland. Some states of mixed confession took longer: Transylvania removed the days 15–24 December 1590. In addition, a number of Catholic countries or states chose not to follow Rome's directive to the letter: Tuscany adopted the calendar but did not move the start of the year to 1 January until 1700.[7]

Inevitably, in the religious climate of the sixteenth century, Protestant countries and principalities refused to heed a mandate from the pope. A few accepted reform, notably the Duchy of Prussia, where 22 August was followed by 2 September 1612.[8] In the Low Countries, broadly speaking, the Catholic provinces switched calendars in 1582 or 1583. The northern provinces were split: in Holland and Zeeland, 1 January 1583 Julian was followed by 12 January 1583 Gregorian; but the other Protestant provinces chose to employ the Julian until 1700, and Friesland retained the Julian calendar until 1701. To complicate matters further, Groningen adopted the Gregorian calendar in 1583, reinstated the Julian calendar in 1594, and finally returned to the Gregorian reckoning at the time of its adoption by much of Protestant Europe around 1700.[9] In the Swiss Confederation, the Catholic cantons changed in 1583, 1584, or 1597; but the Protestant cantons retained the Julian calendar by and large until 1700. England did not adopt the Gregorian calendar until 1752.[10]

In short, the attempt to impose a new calendar on a confessionally polarized Europe fragmented the continent's timekeeping for 170 years. For well over a century after the papal bull, most Protestant and Catholic communities operated in

[7] Blackburn and Holford-Strevens, *The Oxford Companion to the Year*, 785. Florentine and Pisan styles began the year on 25 March. Florence counted years from the 25 March following the Nativity; Pisa counted years from the 25 March preceding Christ's birth, which resulted in Pisa being one year ahead.

[8] See: Owen Gingerich, 'The Civil Reception of the Gregorian Calendar', in George V. Coyne, Michael A. Hoskin, and Olaf Pedersen, eds., *Gregorian Reform of the Calendar*, 266; Blackburn and Holford-Strevens, *The Oxford Companion to the Year*, 685; Friderich Karl Ginzel, *Handbuch der mathematischen und technischen Chronologie* (Leipzig: Hinrichs, 1914; fasc. Repr. Munich, 2015), vol. 3, ch. XIV, 266–79, esp. 271; Paul Botley and Dirk van Miert, eds., *The Correspondence of Joseph Justus Scaliger,* 8 vols. (Geneva: Droz, 2012), vol. 1, lvi–lvii; and Roger Kuin, ed., *The Correspondence of Sir Philip Sidney*, 2 vols. (Oxford: Oxford University Press, 2012), lxv–lxvi, and note. See also Christopher Robert Cheney, ed., *A Handbook of Dates for Students of British History* (Cambridge: Cambridge University Press, 2000), 236–41.

[9] See Blackburn and Holford-Strevens, *The Oxford Companion to the Year*, 384.

[10] For this overview, see Walter E. van Wijk, *De Gregoriaansche kalender: een technisch-tijdrekenkundige studie* (Maastricht: Stols, 1932), 56–65.

different time zones, divided not by hours but by ten days, which by the eighteenth century grew to eleven. Expanding the geographical scope to include the Greek and Russian Orthodox worlds extends calendrical discord into the early twentieth century: Russia converted from Julian to Gregorian in 1918,[11] and the Greek Orthodox Church adopted the revised Julian calendar in 1923.[12]

Understandably, there was resistance to change on the part of many Christians who complained that, whilst church feast days remained on the correct numerical date, following removal of the ten days, the feast would be celebrated on the incorrect day.[13] This, in turn, affected traditional fairs and markets.[14] Those who communicated across confessional boundaries faced additional problems: a letter dispatched from Paris on 5 January 1650 might be reckoned by its recipient in London to have been written on 26 December 1649. Confronted by these problems, some writers clarified their dates by marking them as 'new style' or 'n.s.' (Gregorian) or 'old style' or 'o.s.' (Julian). Others removed ambiguity by double dating letters and providing the date in both calendars, for example as "26 December/5 January'. But few adopted such conventions consistently. In consequence, letters might appear to have been received before they were written; and people moving across international borders appear to time-travel: William of Orange left the United Provinces, which used the Gregorian calendar, on 11 November 1688 but arrived in England, which used the Julian, on 5 November.

[11] Russia continued to employ the Julian calendar and, after 28 February 1800, its difference from the Gregorian calendar grew to twelve days. This continued to cause problems into the early twentieth century: In 1908, for instance, the Imperial team arrived in London twelve days too late for the opening of the Olympic Games. See Edward G. Richards, *Mapping Time: The Calendar and Its History* (Oxford: Oxford University Press, 1999), 247. The Gregorian calendar was implemented in Russia on 14 February 1918 with the elimination of the Julian dates of 1–13 February 1918 by order of a Sovnarkom decree signed by Vladimir Lenin on 24 January 1918 (Julian). Subsequent plans called for the abolition of the Christian week altogether and its replacement with a ten-day week, in the manner of the French Revolutionary calendar, but these plans were not realized (Richards, *Mapping Time*, 277).

[12] The Greek Orthodox Church adopted the Revised Julian calendar in 1923. See Blackburn and Holford-Strevens, *The Oxford Companion to the Year*, 687.

[13] See Heribert M. Nobis, 'The Reaction of the Astronomers to the Gregorian Calendar', in George V. Coyne, Michael A. Hoskin, and Olaf Pedersen, eds., *Gregorian Reform of the Calendar*, 243–54; Michael A. Hoskin, 'The Reception of the Calendar by Other Churches', in George V. Coyne, Michael A. Hoskin, and Olaf Pedersen, eds., *Gregorian Reform of the Calendar*, 255–64; Gingerich, 'The Civil Reception of the Gregorian Calendar', 266ff; and Blackburn and Holford-Strevens, *The Oxford Companion to the Year*, 684.

[14] Blackburn and Holford-Strevens, *The Oxford Companion to the Year*, 684–5. For the fairs and festival days affected in England, see also Robert Poole, "'Give Us Our Eleven Days!": Calendar Reform in Eighteenth-Century England', *Past & Present* 149:1 (November 1995): 95–139, esp. 121–9. See https://doi.org/10.1093/past/149.1.95.

1.2 Additional Calendars and Dating Conventions

Miranda Lewis and Dirk van Miert

The disparity between Julian and Gregorian calendars was not the only source of confusion. Further complexity was created by varying conventions within the Julian calendar itself, especially regarding the day on which a new year commenced. The Roman calendar began on 1 January, but in medieval Europe the Church moved the start of the year to 25 March, the Feast of the Annunciation also known as Lady Day, perhaps in an attempt to differentiate itself from the calendar of its pagan predecessors. Thus 24 March 1581 would be followed by 25 March 1582.

By today's reckoning (proleptic Gregorian), the year has always started on 1 January.[15] At the time, however, depending on where you lived, it might not. In Venice, the new year began on 1 March, and this practice was retained in official documents until the Republic fell to French forces under Napoleon in 1797.[16] In England, Florence, Naples, Pisa, or Scotland, the year began on 25 March. Elsewhere, the new year fell on Easter Day, or 24 or 29 September (respectively the equinox or Michaelmas, the Feast of Michael and All Angels). In certain German and Italian states, 25 December (Christmas Day) was when the year turned. Within France alone, the year began on one of these four different days depending on the region. From the sixteenth century onwards, however, some countries – although by no means all – moved the day the year changed to 1 January; others, including England, continued to use 25 March. By the time the Gregorian calendar was introduced in 1582, a handful of countries (both Catholic and Protestant) had made this change already: for example, France had begun to count the year from 1 January in 1564. Simultaneously, other countries persevered with a different start of the year: England did not begin the year on 1 January until it made the switch to the Gregorian system in 1752. Scotland, by contrast, switched the start of the year in 1600 but retained – with England – the use of the Julian calendar. Thus, a man in Paris might date his letter 30 January 1650 and upon receipt this might be recorded in London as having been dated 20 January 1649, or in Edinburgh as 20 January 1650.[17]

---

[15] Working back in this manner is called proleptic Gregorian; see Richards, *Mapping Time*, 251; and *Data Elements and Interchange Formats – Information Interchange – Representation of Dates and Times ISO 8601*, 3rd edn. (2004), 8.

[16] Leofranc Holford-Strevens, *The History of Time: A Very Short Introduction* (Oxford: Oxford University Press, 2005), 128.

[17] Ironically, the bull issued by Gregory is itself a casualty of these confusions. It was signed and dated: 'Datum Tusculi, anno Incarnationis dominicæ MDLXXXI, sexto Kalendas Martii, pontificatus nostri anno X'. Converting from the Roman calendar, 'MDLXXXI, sexto Kalendas Martii' is 24 February 1581 in the Julian calendar, using the year start of 25 March; it is 24 February 1582 when using the Julian calendar using the 1 January year; and in the proleptic Gregorian calendar the date is 6 March 1582.

These complications were compounded still further by classically trained scholars keen to show off their erudition. Many of the citizens of the republic of letters continued to date their letters by the Roman calendar. To some, no doubt, using the Roman systems was merely a consequence of a desire to write in consistently classical Latin; but the practice was not confined to humanist epistles or even to those written in Latin, and many instances survive of letters that use the Roman calendar having been written in the vernacular.[18] To navigate the chronological labyrinth of early modern epistolography, the scholar must also understand how to convert Roman dates into Julian or Gregorian.

The Roman calendar differed from the form of Julian adopted in the Middle Ages in a number of important respects. By the time of Julius Caesar, it was divided into twelve months, beginning with January (Janus, to whom this month was dedicated, was the god of transitions), and the months were subdivided further into an arrangement based originally on the phases of the moon.[19] The kalends, the nones, and the ides made up the three distinctive sections: kalends set the first day of a month (it had once marked the new moon); ides (which had been based on the full moon) fell in the middle of the month, namely the 13th or the 15th, depending on the month; nones occurred nine days before the ides, but it is important to remember that Roman practice was to count inclusively and thus nones (which means 'nine') is actually eight days before the ides.[20]

Within each section of the month, days in the Roman calendar were named by counting backwards (once again, counting inclusively). Those between *Kalendae* and *Nonae* were called: 'the day before *Nonae*'; 'the 3rd day before *Nonae*' (with no '2nd day before *Nonae*' because *Nonae* itself was the first day, and thus 'the 2nd day before' and 'the day before' were one and the same); 'the 4th day before *Nonae*'; and 'the 5th day before *Nonae*'. Days between *Nonae* and *Idus* were numbered similarly to 'before *Idus*', while days after *Idus* were counted inclusively to before the *Kalendae* of the following month. In leap years, the *dies bissextus*, or extra day, was inserted before 'VI Kal. Mar'.[21]

Given the complexities across Europe associated with the date selected as the start of the year, it is unsurprising that, in the Roman calendar, scholars of the early modern period today encounter problems associated with the dates in the latter

[18] See, for example, the letter from William Stukeley to Edward Wilson, 25 July 1725, written in English with the date recorded as '8 Calend. Aug. 1725' (Bodleian Library, University of Oxford: MS Don. d. 90 p. 132 [no. 44]).

[19] January and February as months had been added on 500 years previous to Julius Caesar's reform; until then there had only been ten months in each year. Under the Julian reforms, while the names of the months were retained, the number of days in most of them were changed, and an additional day was inserted every four years. Additionally, the year 46 BC was adjusted and the year contained 445 days. See Richards, *Mapping Time*, 214.

[20] Inclusive counting reckons the day from which you are counting as day one (rather than taking the day before as day one).

[21] For a more detailed explanation, see Richards, *Mapping Time*, 210–11, and Cheney, ed., *A Handbook of Dates*, 145.

part of December. When a letter is dated in the Roman calendar 'VIII Kal. Jan. 1642', this could be 25 December 1642, or it could be 25 December 1643, depending upon which Julian year start was being employed. While some early modern individuals believed that the year given in the Roman kalends of January was the one in which they were writing, others maintained it specified the following year (that is, the new year in which the kalends of January fell and from which they were counting backwards, and assuming they were in a country that used 1 January as the start of the year).[22]

These three distinctions – between Julian and Gregorian calendars; between the various beginnings of the new year; and between Roman and Julian/Gregorian dating styles – are the most common chronological complications to confront the student of early modern learned letters. But they are not the only ones. Within Christian Europe, dates are sometimes expressed with reference to liturgical calendars,[23] saints' days,[24] or papal or regnal years.[25] Outside the Christian community, specialists encounter dates expressed in the Islamic[26] or Jewish calendars.[27] There were dating systems such as the French Republican calendar imposed by political regimes,[28] or others, for example the Masonic calendar, employed solely by members of a specific organization.[29] These are just a few instances of the dating systems in use in the period.

[22] For specific examples and further clarification, see Botley and van Miert, eds., *The Correspondence of Joseph Justus Scaliger*, vol. 1, lviii. But early modern writers themselves were not consistent, and Botley and van Miert cite the example of letters written and dated by Scaliger on the same day and caution scholars to 'flag as uncertain the year of every Roman date between the 14th and the 30th of December (that is, from XIX Kal. Jan. to III Kal. Jan.) unless it can be dated securely by other evidence'.

[23] See Blackburn and Holford-Strevens, *The Oxford Companion to the Year*, 757–9.

[24] For example, the letter from Ioannes Secundus to Ioannes Dantiscus of 21 February [1532] is dated 'pridie Petri Vincti' (see *Corpus of Ioannes Dantiscus' Texts & Correspondence*, http://dantiscus.al.uw.edu.pl/?f=letterSummary&letter=756, accessed 20/03/2019). When calculating a date from a saint's day, a scholarly eye is often required. Which St John, for example, is being referenced: St John the Evangelist or St John the Baptist? Identification often depends upon the region in which the letter originates.

[25] Regnal years (calculated from an accession date) are not always as straightforward as might be imagined. For example, there is problem with Charles II, who counted his reign from 30 January 1649 (or 30 January 1648 by the Julian English calendar), when his father was executed, while others count it from the Restoration of the monarchy in May 1660.

[26] See Richards, *Mapping Time*, 232–5; and Blackburn and Holford-Strevens, *The Oxford Companion to the Year*, 731–5.

[27] See Richards, *Mapping Time*, 220–30; and Blackburn and Holford-Strevens, *The Oxford Companion to the Year*, 722–30.

[28] See Blackburn and Holford-Strevens, *The Oxford Companion to the Year*, 742–5.

[29] Although a number of Masonic dating systems exist, that of the *Anno Lucis* is used most widely in documentation. *Anno Lucis* is based on the chronology calculated by Archbishop James Ussher (1581–1656) that dates the creation to 23 October 4004 BC and involves adding 4004 to the year.

1.3 The Difficulty of Inferring Calendar Usage

Miranda Lewis and Dirk van Miert

The problems arising from the simultaneous use of multiple calendars are not confined to the technical difficulty of reconciling them with one another. As we shall see, reconciliation of this kind is a purely technical issue, to which definitive solutions exist. More intractable is a secondary difficulty, arising from the fact that letter-writers normally fail to indicate which calendar they are using.

In the absence of this information, previously it has been left to the scholar to deduce which calendar is employed in any particular letter. In relatively localized exchanges – between people in the same country sharing confession and homeland with one another – this is often unproblematic: the calendar employed can be safely assumed to be the one in official use at the time and place in which the letter was written.

When corresponding across national and confessional boundaries, however, letter-writers may decide, for a number of reasons, *not* to adopt the calendrical norms of the place from which they write. Travellers writing home, for instance, might adopt the conventions of their home country rather than those of the country from which they write. Diplomats or merchants might decide to adopt the calendar of their home country even when writing to countrymen who are also abroad. Early modern individuals clearly associated the two calendars with the two main confessional blocks: a letter from one Pieter Corneleszn in Alkmaar to the ministers of the Dutch Church in London is marked with the date 'desen 11 May 1586 stilo papali'.[30] Learned writers might adopt the calendars used by their addressee's confession as a courtesy. A vivid example of the resulting difficulties can be found in the correspondence of the early seventeenth century's foremost chronologer, J. J. Scaliger.

On 13 August 1602, Scaliger wrote two letters to two of his regular correspondents. Both resided in Augsburg. Both letters may be assumed to have been consigned to the same courier. The letter to David Hoeschelius, who was a Protestant, Scaliger dated 'III Non. Augusti Iuliani'. The letter to Marcus Welser, who was a Catholic, he dated 'Idib. Augusti', adopting the Gregorian calendar. Augsburg was an imperial free city inhabited by Catholics and Protestants alike; and Scaliger evidently expected these two confessional communities to use different calendars, and dated his letters accordingly. The ease with which he slipped between these two calendars should serve as a warning to editors to tread very carefully in this area.[31]

[30] See Jan H. Hessels, ed., *Epistulae et tractatus cum reformationis tum ecclesiae Londino-Batavae historiam illustrantes: Ecclesiae Londino-Batavae archivum*, vol. 3, pt. 1 (Cambridge, 1897), 819, letter 1016.

[31] Botley and van Miert, eds., *The Correspondence of Joseph Justus Scaliger*, vol. 1, lvii.

In circumstances such as these, inferring calendar use is a speculative problem, and different scholars have employed very different methods of addressing it. One logical possibility is to convert all the dates of their letter headers into the Julian calendar.[32] The opposite procedure is to convert everything to Gregorian.[33] A third approach is consistently agnostic: simply to record the date as marked and not attempt to guess which calendar was used.[34] Within the domain staked out by these three logical options, a number of hybrid solutions exist. A fourth practice is to adjust to Gregorian whenever possible, while leaving dates as marked in doubtful cases.[35] A fifth alternative is to give both calendar dates in the letter header whenever possible and to infer from the place of sending which calendar has been used in any specific instance, again leaving the date as marked wherever there is doubt.[36]

This diversity of practice raises a third set of issues for anyone attempting to assemble a union catalogue from multiple scholarly editions and inventories. If a union catalogue is to be created by merging all of these inventories and many more, a common standard must be developed. Moreover, if the data in such a union catalogue is to be subjected to automated analysis and visualization, agnosticism regarding calendar use must be minimized. The reason for this is that large-scale digital analysis and visualization accommodates less ambiguity than traditional textual scholarship. In print editions, delicate issues of judgement can be handled in detailed prose annotations, which then inform the reading of relatively small numbers of letters. When doubts persist, the safe option is simply to mark the calendar, and therefore the date, as unknown. But when analyzing and visualizing large corpora of letters computationally, free-text discussion cannot be reliably parsed, and sidestepping uncertainty causes more problems than it solves. In most cases, there are grounds for regarding the use of one calendar as more probable than another; and the analytical distortion caused by incorrectly inferring calendar use in some cases is far smaller than that caused by setting aside all letter records for which the calendar cannot be identified with certainty.

For all of these reasons, the compilation of a union catalogue of correspondence requires a consistent method for making clear and defensible inferences regarding calendar use from the available data. Such a method should mirror the inferences made by scholars themselves. It should also be capable of distinguishing between more and less certain inferences, and should record the security of its

[32] As, for example, in A. Rupert Hall and Marie B. Hall, eds., *The Correspondence of Henry Oldenburg*, 13 vols. (Madison, WI: University of Wisconsin Press; London: Mansel; London: Taylor & Francis, 1965–86).

[33] See, for example, in G. Anton C. van der Lem and Cornelis S. M. Rademaker, *Inventory of the Correspondence of Gerardus Joannes Vossius (1577–1649)* (Assen and Maastricht: Van Gorcum, 1993).

[34] See the example of Roger Kuin, ed., *The Correspondence of Sir Philip Sidney*, 2 vols. (Oxford: Oxford University Press, 2012).

[35] As has been done in the Scaliger edition, see Botley and van Miert, eds., *The Correspondence of Joseph Justus Scaliger*.

[36] See the dating policy adopted in Noel Malcolm, ed., *The Correspondence of Thomas Hobbes*, 2 vols., The Clarendon Edition of the Works of Thomas Hobbes (Oxford: Oxford University Press, 1994).

inferences in a manner readable by both humans and machines. Although capable of being manually overruled by scholarly editors, such a method must also be capable of computational treatment for two reasons. On the one hand, some degree of automation will be necessary in order to standardize retrospectively tens of thousands of records already accumulated. On the other, distributed infrastructure will make automation even more necessary. The capacity to pool the results of multiple catalogues, inventories, digital archives, and libraries of printed and manuscript correspondence will generate letter records for curation on a vast scale; and these will only be processed efficiently if preliminary standardization is undertaken computationally, pending further refinement by scholars.

1.4 Incomplete and Uncertain Dates

Thomas Wallnig, Arno Bosse, and Miranda Lewis

The problems that arise when contemporaries fail to indicate which calendar they are using are a subset of the far larger problem of incomplete dates. When a letter includes a complete date but fails to specify the calendar used, then the data is incomplete, and systems need to be developed to complete it. Once the missing data is supplied (that is, once the calendar is tentatively identified), a separate tool can establish automatically the unique, unambiguous, and precise relationship between a date as given and a date in the proleptic (modern) Gregorian calendar.

Analogous but far more intractable difficulties arise when the date provided is itself incomplete, whether or not a calendar is specified. For instance, in some letters, a day and month is specified but not a year. In others, reference is made only to a day of the week, without day, month, or year. On other occasions, a feast day and year are provided, without mentioning day or month. Written notes that are part of a local communication system, within a town or a court for example, may only record an hour or a part of the day. Further chronological difficulties arise for letters that were not, in fact, written on one particular day. For example, a precise date might be recorded for the drafting of a letter, but not for the production of the final copy as sent. In other cases, letters could be composed over several days, thus requiring a 'date range' rather than a single date.

Still more awkward are the numerous letters in which no date of any kind is provided, whether because the writer failed to include it, or because it was left out by a copyist, or because part (or indeed all) of the letter has been lost. In such cases, dating must be inferred speculatively with varying degrees of certainty in many different ways: from the date of receipt; from the date at which a response was written; from the place of a specific letter in a sequence; from references to the letter or its content in other documents; or from dateable details of the content of the letter itself.

These problems and many more bedevil the ascription of precise and certain dates to innumerable early modern learned letters. Since uncertainty arises from many different sources, many different patterns of inference are needed to date undated letters and many different degrees of uncertainty arise from these inferences. In a traditional hard-copy edition or inventory, these inferences can be articulated at length one at a time. In order to analyse and visualize large quantities of correspondence metadata, however, standard means are required for representing both uncertain and incomplete dates and for the inferences involved in attempting to clarify or complete them.

2 Solutions

Arno Bosse and Howard Hotson

The complexities of competing calendrical systems encumber the proper dating of letters by the lone scholar. The problems posed by these complexities are greater still when a union catalogue is being compiled from multiple sources and huge numbers of letter records need to be processed. Before proposing solutions to these problems, we first need to divide these complexities into manageable parts.

In essence, three different problems are entangled together. The first problem is that of converting dates from one known calendar to another. For this procedure, reliable solutions are available (2.1). The second problem is to determine which calendars are in use in individual places at specific times. For this purpose, new resources are needed, such as the gazetteer proposed below, which indicates when specific places transitioned from one calendar to another (2.2). The third and hardest problem is to develop procedures for provisionally inferring the calendar employed in any individual letter and for assigning a degree of certainty to the inference (2.3). Each of these three components will now be discussed in turn.

2.1 Converting between Calendars

Arno Bosse

Individual tools for converting dates from one known calendar to another already exist. For converting between Julian and Gregorian dates, resources are readily available. These include tools both for developers seeking to create their own applications and for researchers, editors, and other users employing free conversion utilities on the web. Other tools can handle the Roman and Latin versions of the

Julian calendar, including those using Roman numerals.[37] Individual tools likewise exist for calculating the date of the Jewish Passover, Easter Sunday, and the other dates of the ecclesiastical calendar dependent on them,[38] and for handling the Hebrew, Islamic, French Revolutionary, and other calendars as well.[39] For developers, online resources such as *World of Science*,[40], and print references such as Reingold and Dershowitz's *Calendrical Calculations*,[41] explain the historical and mathematical context and provide the algorithms for converting dates between a wide range of calendrical systems. Open-source software libraries and applications implementing these conversion routines can readily be found on *GitHub* and other public-code repositories.

There is thus no lack of individual software solutions for converting dates between calendars. What is currently lacking is a well-maintained, centrally accessible resource providing facilities for three sets of uses: (1) a webform for converting dates while individual letter records are input by scholars; (2) a facility, based on spreadsheets, for converting calendrical data when large numbers of letter records are ingested by a catalogue data editor; and (3) the well-documented API and data exchange format needed by developers. As a starting point, such a resource would need to be capable of dealing with all widely used early modern calendar dates, that is, Julian, with year starting 25 March; Julian, with year starting 1 January; Gregorian; Roman; Hebrew; and Ottoman. Subsequent extension could include more complicated conversions, for instance, from the liturgical calendar.

2.2 Tracking Calendar Usage in Specific Places

Arno Bosse

As well as these tools for converting between calendars, resources are needed for determining which calendars were in use in specific times and places.

The background functionality for such a tool will be provided by *EM Places*, the historical gazetteer for the early modern period currently under development at *Cultures of Knowledge*.[42] As discussed in more detail in the previous chapter, *EM Places* will offer a database of political-administrative place entities or 'polities' containing data on when these polities existed historically and where they were situated at any given time within hierarchies of larger and smaller polities. Every specific

[37] See http://www.csgnetwork.com/julianmanycalconv.html; http://www.softhawkway.com/rcalc.htm, and the electronical version of the classic Grotefend, http://bilder.manuscripta-mediaevalia.de/gaeste/grotefend/grotefend.htm, both accessed 20/03/2019.

[38] See https://www.staff.science.uu.nl/~gent0113/easter/easter_text2a.htm, accessed 20/03/2019.

[39] See http://www.fourmilab.ch/documents/calendar/, accessed 20/03/2019.

[40] See Edward M. Reingold and Nachum Dershowitz, *Calendrical Calculations. The Ultimate Edition*. 4th edn. (Cambridge: Cambridge University Press, 2018).

[41] See https://doi.org/10.1017/9781107415058.

[42] See https://github.com/culturesofknowledge/emplaces, accessed 20/03/2019.

place recorded in the gazetteer – for example, a city from which a letter may have been sent – will be situated within one of these political-administrative hierarchies.

In order to track calendar usage in individual times and places, *EM Places* will also record the dates at which individual polities transitioned from one calendar to another (specifically, between the Julian calendar with different dates for the start of the year, and from Julian to Gregorian). In this way, when metadata are inputted or uploaded indicating that a letter was sent on a particular date from a particular place, *EM Dates* will be able to draw on *EM Places* to determine which calendar was in use in that place at that time.

As noted in the previous chapter, it is not expected that *EM Places* will be populated with a comprehensive data set for all European places at the outset. For instance, the difficult task of generating a consensus on the shifting hierarchical relationships between geographical entities will only be accomplished gradually over time. Likewise, it will not be possible initially to supply comprehensive data on dates of calendrical transition for all regions of Europe. Instead, a first stage of work will prioritize large polities where political authority was well consolidated by the late sixteenth century (such as France and England) and regions containing towns and cities from which the largest numbers of letters were sent or received. In this way, data on the main centres of intellectual communication can be provided within a manageable initial data set. No less importantly, a structure will have been provided within which the scholarly community can pool its expertise to enhance this data to better serve its needs. In other words, this aspect of *EM Dates* and *EM Places* is not a ready made service, furnished with a comprehensive data set at launch: instead, like EMLO, it offers a framework within which the scholarly community can collaborate in creating the resources it needs. In the meantime, an incomplete set of calendrical data represents an improvement on the current situation, in which chronological data typically fails even to identify the calendars used, much less to reconcile dates in different calendars with one another.

A related challenge will be to decide how to handle places for which no reliable information on calendrical usage is readily available. In such cases, two general axioms will be applied in the absence of historical information to the contrary. The first is that, within polities in which political and ecclesiastical authority is well consolidated, subordinate polities transition between calendars at the time dictated by superordinate authorities. For instance, Munich, Ingolstadt, and all the other towns and cities of Bavaria transitioned from Julian to Gregorian at the same time as the duchy that governed them; and the same transition was affected in the Upper Palatinate when it was annexed to Bavaria during the Thirty Years' War. The second axiom is that, in the absence of any other information, it will be assumed that all Roman Catholic territories transition from Julian to Gregorian on 15 October 1582, and that all non-Roman Catholic territories do not. Although this is of course a gross oversimplification, it will still provide a more adequate starting point than making no claim at all. When data reveals that a specific historical polity (e.g. the 'Duchy of Opole') transitioned at a later date (in this case, January 1584), the

first axiom implies that all the polities *below* it (e.g. the town of Opole) also transitioned at that date, while the second axiom implies that assumptions about calendrical usage in the polities *above* it in the hierarchy remain unchanged. As *EM Places* is populated with an increasingly comprehensive data set on calendrical usage, any anomalies initially arising from these assumptions will gradually disappear.

2.3 Inferring Calendar Usage in Individual Letters

Howard Hotson

The most difficult of these problems is that of identifying which of the two homologous calendars and their various interpretations was used in any individual letter. Detecting the use of the Roman, ecclesiastical, Jewish, or Islamic calendars is easy, since each records dates in a different fashion. However, Julian (irrespective of which day marks the start of the year) and Gregorian calendars all record their dates in the same way – that is, in terms of years ('anno Domini') and sequentially numbered days of the month – making it impossible to determine at first glance which is being used. Unless the letter explicitly indicates which calendar is used, which calendar was used must be inferred from the contextual data available, hence the need for a tool for provisionally inferring calendar use from contextual data, pending further scholarly scrutiny.

The purpose of such a tool is four-fold. First, it should aim to replicate the most simple and straightforward inferences made by scholars. Second, when the data is inadequate for confident scholarly assertion, it should supply a reproducible 'best guess'. Third, since not all of these inferences will be equally robust, the tool should also indicate the confidence of the inference in a manner legible to both computers and human users. Fourth, the tool should also be capable of provisionally assigning calendars to letters automatically when ingesting or retrospectively standardizing large quantities of unedited letter records.

How then to devise a tool capable of performing these functions? The starting point is the acknowledgement that an inference depends on the data available. When little data are available, the inference may be simple but insecure. When more data is available, the inference may be more complicated, but also more certain. The following discussion first treats the most basic pattern of inference employed by scholars and then considers how to render it computational.

In the simplest cases, the only data available for automatic analysis are the places and dates contained in the letter records themselves. When only date and place of sending are known, the inference is simple but insecure: the calendar employed is most probably the one standard at that place and time at which it is written, but the confidence level of this inference is relatively low. When the place of both sending and receipt are also known, a more complicated inference is required. If both places use the same calendar, the use of that calendar can be inferred with a

higher degree of confidence. If the two places use different calendars, however, the calendar in use at the place of sending remains the more probable inference, but confidence in the inference is lower than in either of these two other cases.

A more sophisticated tool might also analyse data contained in the person records of each of the two correspondents. A basic person record contains dates and places of birth and death. Places of birth and death can help determine whether the countries from and to which a letter is written are the long-term places of residence of the two correspondents, and doing so is useful for inferring calendar usage. This possibility would produce six separate geographical reference points: the places of birth and death of the two correspondents and their places at the time of correspondence. If all six of these data place both correspondents in regions using the same calendar, then the probability that they are using that calendar is very high. If, on the other hand, these six geographical data points are mixed, their relative weight must be measured. Consider, for instance, a letter written from Utrecht to Venice in 1620. Judging from the places of sending and receipt alone, we might conclude (with a low level of confidence) that the letter uses the Julian calendar. However, if person records indicate that both sender and recipient were born and died in Venice, then the greater likelihood is that they are using the calendar in official use in their home city, that is, the Gregorian.

This discussion suggests a basic mechanism for inferring probable calendar usage and measuring the confidence of the inference. In a somewhat simplified version of this basic model, only four factors are in play (aside from the date of sending): these are the places of sending and receipt, and what might loosely be called the *patria* (or homeland) of the sender and recipient.[43] These four factors need to be weighted differently. The active party (the letter-writer) is more important in choosing which calendar to use than the passive party (the recipient); so the location and *patria* of the sender must count more than those of the recipient. The location (of writer and recipient) might also be treated as more important than their homeland (since the most convenient thing for any contemporary is to follow the timekeeping practices of one's locality).[44]

Precisely how these four factors would be weighted requires careful study. The first step in such a study would be to mock up a set of metrics and test the results against scholarly intuition and eventually empirical data. To begin this process:

[43] 'Nationalities', in turn, are indicated by places of birth and death, but these can be treated as unitary to simplify the initial construction of the model.

[44] In determining 'nationality', place of birth counts more than place of death.

- Location of sender (the most important datum) is assigned 4 points.
- Location of recipient (less important than location of sender) is assigned 3 points.
- Nationality of sender (the next most important factor) is assigned 2 points.
- Nationality of recipient (the least important of these four factors) is assigned 1 point.

The results of this exercise are displayed in figure 1.[45] The rows of the table indicate all the possible combinations of the four factors on which the calculation is based. The first factor (location of sender) is weighted 4, the second (location of recipient) 3, the third (nationality of sender) 2, and the fourth (nationality of recipient) 1. In columns 1 through 4, the letters 'A' and 'B' stand for the two calendars in contention. Column 5 then sums up the points registered in columns 1 through 4 in favour of calendar A. Column 6 then expresses this confidence level with colour rather than a number: colour is deployed here to avoid the impression that this is a precise calculation of probability, and to provide a code which can be used in visualizations.

The maximum confidence level (bright green) is reserved for row 40, where all four columns indicate that both correspondents are located in and indigenous to a *patria* using the same calendar. The minimum confidence level (bright red) is reserved for the two instances (rows 20 and 34) in which equal points are recorded for both calendars: for instance, row 20 describes a correspondent of unknown origin writing from a region using calendar A to a correspondent who is both located in and indigenous to a region which has adopted calendar B.[46] The two instances in which a negative number is returned in column 5 indicate that the preponderance of data suggests that calendar B might be used, despite the location of the letter-writer in a region using calendar A. Between these two extremes, the moderate levels of confidence (3–6) can be produced by many different combinations of factors. Row 4, for instance, totals 4 points because the only datum available is the location of the sender; and row 37 likewise registers the same moderate level of confidence because this is an instance of a letter written from the home country to a countryman travelling abroad.

[45] This chart simplifies the calculus in two important respects. First, it assumes that there are only two calendars at play in the calculation, whereas sometimes there may be three: old and new-style Julian as well as Gregorian. Second, it assumes that places of birth and death indicate the same 'nationality', whereas in fact these may differ. Removing these possibilities reduces the number of permutations dramatically, allowing the method employed here to be illustrated and studied. Once the basic principles are understood, this additional level of complexity could easily be entered into a computational version of this model.

[46] This is a very useful outcome, since it means that in all but two of these thirty-seven cases there is a balance of probability in favour of one calendar or the others, which helps to render the data analysable and visualizable computationally.

		1	2	3	4	5	6
	Data	**Location**		**Patria (homeland)**		**Confidence level**	
		of sender (weight: 4)	**of recipient** (weight: 3)	**of sender** (weight: 2)	**of recipient** (weight: 1)	**sum**	**colour code**
1	**A**				A	1	
2				A		2	
3			A			3	
4		A				4	
6	**AB**			A	B	1	
7			A	B		1	
8		A	B			1	
9			A		B	2	
10		A		B		2	
11		A			B	3	
12	**AA**			A	A	3	
13			A		A	4	
14			A	A		5	
15		A			A	5	
16		A		A		6	
17		A	A			7	
19	**ABB**	A	B	B		-1	
20		A	B		B	0	
21		A		B	B	1	
22	**AAB**	A	B		A	2	
23		A		B	A	3	
24		A	B	A		3	
25		A		A	B	5	
26		A	A	B		5	
27		A	A		B	6	
28	**AAA**		A	A	A	6	
29		A		A	A	7	
30		A	A		A	8	
31		A	A	A		9	
33	**ABBB**	A	B	B	B	-2	
34	**AABB**	A	B	B	A	0	
35		A	B	A	B	2	
36		A	A	B	B	4	
37	**AAAB**	A	B	A	A	4	
38		A	A	B	A	6	
39		A	A	A	B	7	
40	**AAAA**	A	A	A	A	10	

Figure 1: Table for inferring probable calendar usage

The purpose is not to calculate the mathematical probability of one calendar being used instead of another. Rather, the aim is to determine two things in a standardized fashion. The first is to infer which of two calendars is more likely to have been used in any given letter (based on the evidence available): this is necessary in order to analyse and visualize large data sets computationally. The second is to determine how much confidence should be attached to the previous inference: this is necessary to ensure that human users as well as computers do not treat all of these inferences as equally secure.

Despite its simplicity, this basic model appears to provide a surprisingly effective solution to this problem. Yet it need not be regarded as a finished product: it could be further developed in a variety of ways. In the first place, a formula for disaggregating *patria* into places of birth and death is needed: these might be weighted the same or differently. Second, confessional differences might be added to the model, to help resolve the problem of multi-confessional polities such as Augsburg, where more than one calendar is in use simultaneously. Third, where even richer prosopographical data is available, the relative social standing of the two correspondents at the specified point in time might be taken into account as well: in multi-confessional polities like the Holy Roman Empire, a lower-status sender might adopt the calendar preferred by the higher-status recipient as a sign of respect, irrespective of questions of origin and confession. Fourth, the weightings attached to individual factors could be adjusted, if this produced more satisfactory calculations.[47] Introducing weights which are not integers could be easily accommodated due to the fact that colour coding is the main means of communicating the confidence level, rather than precise numerical values.

In any case, a fundamental principle of this system is that scholarly judgements can always override computationally generated inferences, provided that scholars are also willing to assign a confidence level to their judgements. The primary purpose of such an automated system would be to assign calendars *provisionally* to unedited data ingested in bulk. Whenever carefully curated data is ingested – for instance, from meticulously compiled inventories and editions – the scholarly judgement of the editors must conclusively override the computationally generated inferences. To be more specific, expert users will be able to override these automatic inferences, for instance, for individual letters (e.g. when the calendar used is explicitly stated), or for all the letters from a specific person (when their usage is consistent). On the other hand, in the not-too-distant future, distributed infrastructure will create the capacity to pool the results of multiple catalogues, inventories, and digital archives and libraries of printed and manuscript correspondence. At that point, it will be necessary to complement scholarly curation of individual data records with automatic data curation on a very large scale.

[47] One might argue, for instance, that these initial weightings undervalue nationality relative to location. For instance, if two Spanish ambassadors correspond with one another from postings in London and St Petersburg, they may use the Gregorian calendar despite the continued use of Julian in the countries from which they write.

Finally, it should also be stressed that empirically derived rules could also be deployed to override these calculations systematically in specific cases. If, for instance, it is discovered that ambassadors from a given country are required to report home and to write to one another using the calendar of the home country, this can be implemented as a general rule. Likewise, if research determines that the religious minority in a given country normally adopts the calendar used by their co-religionists abroad, this might be introduced as a special rule which overrides the normal calculus of this system.

2.4 Handling Incomplete and Uncertain Dates

Thomas Wallnig, Arno Bosse, and Miranda Lewis

Given the great complexity of the inferences involved, there is no immediate prospect of automated inference providing precise dating for undated or partially dated letters. What a union resource does require, however, is a standard means of representing the various forms of unknown, uncertain, and incomplete dates in a manner which is accurate, consistent, and computationally tractable. Without such a standard, there is no prospect of including the huge numbers of imprecisely dated letters in automated analysis or visualization, or of undertaking semi-automated bulk conversions of calendrical data.

Among the available standards that might be adopted in response to this problem, two merit brief consideration here. Chapter 13.1.2 of the TEI guidelines provides a very basic set of tags to describe periods of time and *termini ante quem* as well as *post quem*.[48] Since it is already incorporated within the TEI guidelines, this system can facilitate broad interoperability (e.g. in the context of <corresp Desc>, for which see chapter II.7); but its basic form is generic and does not address the variety of problems outlined above. In a similarly generic way, but with extended features, the conceptual reference model CIDOC CRM offers categories for the description of some kinds of temporal uncertainty under the heading 'E2 – temporal entity'.[49] Most of them – like, for example, 'time span', 'occurs during', 'overlaps with', 'meets in time', 'happens during', 'ongoing throughout', 'at some time within', 'minimum/maximum duration' – reflect CIDOC CRM's mission to relate cultural heritage artefacts to descriptive categories of complex semantic value (such as 'the Bronze age'). In that sense, a reuse of this model for correspondence metadata is not inconceivable, but would require adaptation.

[48] See http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html, accessed 20/03/2019.

[49] See http://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_5.0.4.pdf, accessed 20/03/2019.

Something similar can be said about means for measuring the closeness and overlap of query and annotation intervals,[50] or visualizing the overlap of time-spans while taking into account uncertain beginnings and endings.[51] These too have been framed primarily as solutions to a somewhat different problem: namely, the difficulty of assigning data to broad historical periods, rather than the uncertainty arising from the lack of precisely dating to year, month, and day. For example, an 'axial age', one of the test settings of *topotime*, overlays the lifespans of roughly a dozen ancient religious leaders and philosophers; uncertainties regarding their exact dates of birth and death are visualized by means of geometrical figures that become more acute the shorter the documented period of activity is. This solution provides a good overview of the simultaneous activity of ten to fifteen individuals with often considerable biographic uncertainty, but the visual concept will not be adequate for the visualization of uncertain dates of tens of thousands of letters.

The ISO standard for representing date and time digitally, ISO 8601:2004,[52] provides a means of avoiding ambiguities between different conventions for representing dates. A simple example is the conflict between the American month-day-year convention and European preference for day-month-year. In this case, ISO specifies the use of YYYY-MM-DD. ISO is currently silent on the problem of representing incomplete, uncertain, and unknown date data, but a new draft revision, ISO 8601-2 (due for adoption in 2019)[53] is intended to deal with this deficiency. It describes precise means for representing uncertain or approximate dates, and cases in which portions of dates are unspecified, as well as time intervals with uncertain, unknown, or open start and/or end dates. Adopting this standard is important, since tools such as *EM Dates* will get far more traction if contributors do not first have to convert their incomplete, uncertain, and unknown date metadata into a format that other software can understand. Bulk conversions of many dates can also be exported by *EM Dates* into the new standard so that other applications can automatically recognize uncertain or approximate conversions.

The employment of methods and tools ultimately depends on the purpose: if the goal is that of pointing the user of a union catalogue to the uncertainty of a letter date, basic tagging and intuitive colour coding may suffice. If, however, research questions can be formulated that relate to the different types of uncertainty outlined above, then it will be necessary to create more sophisticated semantic systems.

---

[50] See Tomi Kaupinnen et al., 'Determining relevance of imprecise temporal intervals for cultural heritage information retrieval', *International Journal of Human-Computer Studies* 68:9 (2010): 549–60, see https://doi.org/10.1016/j.ijhcs.2010.03.002. This and the following resource have been recommended by Bruno Martins.

[51] See http://dh.stanford.edu/topotime/docs/TemporalGeometry.pdf; http://dh.stanford.edu/topotime/; https://github.com/kgeographer/topotime, both accessed 20/03/2019. See also http://perio.do.

[52] See https://www.iso.org/standard/40874.html, accessed 20/03/2019.

[53] See https://www.iso.org/standard/70908.html, accessed 20/03/2019.

II.4 People

Howard Hotson, Thomas Wallnig, Jouni Tuominen, Eetu Mäkelä, and Eero Hyvönen

1 Preamble

Modelling time (II.3) is complicated, above all for the period in which Europe transitioned from a variety of Julian calendrical conventions to the modern, Gregorian one. But all of these calendars were based on mathematical models; so the reconciliation of calendars is ultimately a technical problem of the kind that computers handle easily, and many of the problems involved in it are amenable to definitive solutions.

Modelling space (II.2) is more complicated still, and not all problems raised by it can be solved in the foreseeable future. For instance, although it is possible to envisage a system that locates points within multiple, parallel, evolving spatial hierarchies, we cannot yet foresee a user-friendly technology or data set that allows scholars to trace the constantly shifting boundaries of those sometimes extremely complicated geographical entities.

Modelling people (II.4) is even more complicated than modelling space. People can be neither modelled with mathematical precision (like time) nor fixed with mathematical precision (like places). They move about constantly in time and space, interacting with one another and gaining and losing attributes constantly.

In light of this extreme complexity, there is no prospect of adumbrating a comprehensive prosopographical data model in this short space or at this stage of project development. Instead, this chapter sketches out the principles to be employed in modelling names (in section 2) and people (section 3), discusses a few

classes of events and people with which the development of a robust data model might begin (part 3.7, which is continued in chapter IV.4), and concludes with a brief account of some of the issues involved in implementing these principles technically in the context of the CIDOC CRM conceptual reference model (section 4).

2 Modelling Names

2.1 The Problem of Ambiguity

Same name, different people. Early modern European proper names are combinations of finite quantities of given names, surnames, and titles.[1] This naturally gives rise to ambiguity, when multiple people share the same or similar names.

For instance, the Thesaurus compiled by the Consortium of European Research Libraries (CERL)[2] includes roughly 150 different people by the name of 'Johannes Fabricius' (the Latin translation of 'John Smith'), all of them from the sixteenth, seventeenth, and eighteenth centuries.[3]

This problem is compounded by the tendency of nuclear and extended families to use the same given and surnames both simultaneously and throughout successive generations. In this case, not only names but also often geographical locations, professions, and blood relations are confusingly similar.

An example familiar to historians of colonial New England is the name 'John Winthrop', which could refer to the elder man of that name (1587/8–1649), the founding governor of the Massachusetts Bay Colony, his son, John Winthrop the Younger (1606–1676), a colonial governor of Connecticut, the latter's eldest son, Fitz-John Winthrop (1639–1707), another colonial governor of Connecticut, and the elder's great-great-grandson John Winthrop (1714–1779), a professor at Harvard College.

Different names, same person. The resulting problems of disambiguation are multiplied by the tendency to refer to the same person in many different ways, and to spell the same name in multiple ways, potentially in each of several different languages.

For instance, within the CERL Thesaurus, Hugo Grotius's name is spelt more than fifty different ways on the title pages of books in European libraries.[4] Within

[1] For guidance through the complexities of this topic, see Carole Hough, ed., *The Oxford Handbook of Names and Naming* (Oxford: Oxford University Press, 2016), esp. pt. III: 'Anthroponomastics'.

[2] See https://data.cerl.org/thesaurus/_search, and https://www.cerl.org/, both accessed 20/03/2019.

[3] During the course of almost exactly one year, between 24/11/2017 and 22/11/2018, the number rose from 114 to 150 with some few non-pertinent names, see https://thesaurus.cerl.org/cgi-bin/search.pl?query=johannes%20fabricius&type=p&set=full, accessed 20/03/2019.

[4] See http://thesaurus.cerl.org/record/cnp01440072, accessed 20/03/2019; the figure one year earlier was twenty-eight.

the 8,034 letters to and from Hugo Grotius digitized within the Dutch *ePistolarium* project, the Swedish general Lennart Torstensson (1603–1651) is referred to 486 times under thirty-nine different name forms.[5]

This problem is exacerbated by the practice of referring to individuals by their titles: the same person can have several different names and multiple titles successively, and the same title can pass sequentially from one person to another.

Noble and other titles. Churchmen are often referred to by their titles, which change as they work their way up the ecclesiastical hierarchy and are passed from person to person over time. Noble titles could also change in rapid succession: Francis Bacon's early continental readers often referred to him as 'Verulamius', in reference to the noble title which appears on the title page of the *Instauratio magna*: he was created Lord Verulam in 1618 but became 1st Viscount St Alban only three years later, in 1621.

Toponymics. The medieval tradition of associating a person with a place – famous in the cases of Thomas Aquinas and Leonardo da Vinci – was continued by a founding figure of the republic of letters, Erasmus of Rotterdam. The young man originally known as 'Johannes Amosus Nivnizensis' (i.e. from Nivnice in Moravia) later became famous under the Latinized surname 'Comenius' (derived from another nearby town: Komňa).[6]

Noble toponymics are a special case. Noblemen were often proud to be known by the denomination of their estates. Charles-Louis de Secondat, for instance, published his works under a title derived from his dominions: baron de La Brède de Montesquieu.

Pseudonyms were adopted for a variety of reasons. One species is aliases adopted to hide the identity of the writer. A prominent example is the notorious *Tractatus theologico-politicus*, which Baruch de Spinoza's publisher released under the names of several recently deceased scholars. So intractable were the problems caused by this practice that Vincentius Placcius published the first bibliographical guide to deciphering pseudonyms already abundant in the republic of letters in 1708: the *Theatrum anonymorum et psuedonymorum*.[7]

Pen names. François Marie Arouet de Voltaire is said to have used 178 different pen names throughout his literary career. The most common of these is that of 'Voltaire' itself, which he invented for himself. As he wrote to Rousseau in 1719, 'I

[5] See http://ckcc.huygens.knaw.nl/?page_id=677 , accessed 20/03/2019.

[6] Gottfried Zedler and Hans Sommer, eds., *Die Matrikel der Hohen Schule und des Paedagogiums zu Herborn* (Wiesbaden: Harrassowitz, 1908), 56, no. 1472 (30 March 1611), mistranscribed 'Nivmizensis'. See Gustav Toepke, *Die Matrikel der Universität Heidelberg von 1386 bis 1662* (Heidelberg: Selbstverlag, 1884–93), ii. 265, no. 74. Erasmus was illegitimate, and Comenius was an orphan, whose precise place of birth is disputed.

[7] Later landmarks include Antoine-Alexandre Barbier, *Dictionnaire des ouvrages anonymes et pseudoynmes* (1806–8). Modern guides include Adrian Room, *Dictionary of Pseudonyms: 13,000 Assumed Names and Their Origins*. 5th rev. edn. (Jefferson, NC: McFarland & Co., 2010).

was so unhappy under the name of Arouet that I have taken another'.[8] Arouet was not a noble name, and may have been uncomfortably redolent of *à rouer* ('to be beaten up') and *roué* (a debauchee). As a child, he had apparently been known as 'le petit volontaire' ('the determined little thing'), and Voltaire is an anagram of the 'Arouet l[e] j[eune]', which can also be formed by reversing the syllables of his family's home in Airvault, in the region of Poitou.

Learned name forms. As long as Latin remained the universal language of learning in Europe, vernacular surnames had Latinate forms as well. Some were formed merely by adding a Latin suffix to a vernacular name. Others translated the root word into Latin, like the example of 'Fabricius' given above. More elevated still was to use Greek: at the suggestion of his great-uncle, Johann Reuchlin, Philipp Schwartzerdt ('black earth') adopted the more elegant Greek surname 'Melanchthon' (Μελάγχθων). Another Protestant theologian, Johann Hussgen / Heussgen / Huszgen first etymologized his name as Hausschein ('house-shine') and then Graecicized it to Oecolampadius, from *οἰκο-* ('house') and λαμπάδ- ('torch', 'lamp').

Religious names. An additional difficulty is added by religious communities, in which people entering change their names. While names in the *saeculum* often reflect the – dynastic or regional – agenda of the parents, the religious name usually tells something about the specific monastery or the order and its veneration for specific figures: Matthias Leopold and Franz Philipp Pez echoed a Counter-Reformation Habsburg name spectrum before becoming Bernhard and Hieronymus in the context of Melk Abbey.

Given this range of options, the choice of name to use will normally reflect the specific communicative situation, such as the purpose, privacy, confidentiality, and formality of the communication, as well as the familiarity of the persons communicating. This opens a variety of research questions regarding the statistics of name use and their relation to other dimensions of learned epistolary practice. For present purposes, however, they represent one of the basic challenges to be overcome by a prosopographical system adequate for use on the republic of letters.

2.2 Removing Ambiguity: Unique Identifiers and Authority Files

Authority files provide an invaluable aid in resolving ambiguities of this kind. Traditional authority files typically chose one name form and spelling (e.g. for a person, place, or institution) as the standard, main heading of an alphabetical card file and then collected variant spellings, forms, and titles under that standard heading. Digital authority files now typically substitute a unique identifier for the standard name form. These unique identifiers allow the reconciliation of multiple name forms

[8] Voltaire to Jean Baptiste Rousseau, c. 1 March 1719. *Electronic Enlightenment*, ed. Robert McNamee et al. Vers. 2.1 (University of Oxford, 2010); see http://www.e-enlightenment.com/item/voltfrVF0850079_1key001cor, accessed 20/03/2019.

without the need to designate one or another as standard. They also represent these names in a machine-readable format which allows for interoperability with other catalogue resources.

Existing authority files have mostly arisen from cataloguing initiatives, normally originating in a library context. Some of these cataloguing initiatives are national in origin but have taken on an international dimension as well: a prime example is the Gemeinsame Normdatei (GND, also known as the Integrated Authority File and Universal Authority File),[9] which handles personal names, subject headings, corporate bodies and other objects. It is used for documentation in libraries and increasingly also by archives and museums. The GND is managed by the Deutsche Nationalbibliothek in cooperation with library networks in German-speaking Europe and elsewhere. Others originate in international consortia, such as the CERL Thesaurus mentioned above, which was created from the references to people and places in the consortium's catalogues of printed books. Still others are private initiatives, such as the Virtual International Authority File (VIAF),[10] which typically contain large numbers of records still awaiting disambiguation. Additional authority files can be harvested from digitized copies of national biographical dictionaries.

The most obvious limitations of existing authority files arise from their origin. Authority files derived from national biographical dictionaries are typically national in scope. Given the international scope of the republic of letters, an international authority file is needed. Authority files arising from library catalogues are typically restricted to authors and printers of printed books. These lists overlap with catalogues of writers of learned letters; but far more people wrote surviving letters from far more places than are found in catalogues of books. In fact, the vast majority of people listed in a catalogue of correspondence – often 80 per cent or more – are not recorded in existing authority files. As a consequence of these two limitations, scarcely one in five writers of early modern learned letters is found in existing authority files.

Towards a personal authority file for the republic of letters. Solving this problem is relatively straightforward: the community of scholars interested in the republic of letters needs to collaborate in creating and curating a shared authority file which will provide unique identifiers for all the individuals encountered in their sources. The basic source of data for this authority file will originate primarily from catalogues of learned letters as well as books. *Early Modern Letters Online* provides a point of departure for such an authority file in its c. 25,000 curated biographical entries. These data are currently being used to create 'Early Modern People', an editable, linked open data resource for bibliographic and prosopographical information on early modern people shared under CC0, and accessed via an API and a web interface. The obvious step for enhancing such a list would be to assemble a union list

[9] See http://www.dnb.de/EN/Standardisierung/GND/gnd_node.html, accessed 20/03/2019.

[10] See https://viaf.org/, accessed 20/03/2019. Additional authority files exist, for example, in the context of *WikiData*, the Library of Congress, the Bibliothèque nationale de France, UK National Archives, or the *Oxford Dictionary of National Biography*.

of additional authority files which can be used to identify individuals missing from existing top-level resources like CERL and VIAF. For Great Britain, for instance, in addition to the standard national biography, these would include resources such as the *Clergy of the Church of England Database*, *History of Parliament Online*, and the online registers of students in Oxford and Cambridge.[11]

A far more serious limitation of existing authority files is that they lack references to the evidence on the basis of which disambiguation has been conducted, an account of the logic of inferring disambiguation from that evidence, and acknowledgement of the scholars who have located the evidence and made the inferences. In order to meet these requirements, scholarship requires something more than a longer list of unique identifiers. It needs a description scheme for learned people that can provide a framework for producing prosopographical data better suited to its needs.

3 Modelling People

In order to develop a robust and appropriate model for citizens of the republic of letters, a host of questions must be answered:

What do we mean by prosopography? Why do we need it? How must it be structured to meet these needs? How should they be referenced? With what sort of authorities should it be populated? How can such a model balance durability and flexibility, ease of use and comprehensiveness? Which events are central to such a model and which peripheral? With which categories of people can the development of a robust prosopographical data model best begin? Answering all of these questions will not provide a prosopographical data model for the republic of letters, but it does provide a set of requisites which such a data model must meet.

3.1 Definition

What do we mean by 'digital prosopography'? In essence, we mean biographical data in structured and machine-readable form.

It is the structured nature of prosopographical data that differentiates it from other collections of biographical information. Instead of constructing rounded biographies of individuals as prose narratives, each structured by the unique features of an individual life, prosopography assembles data on a finite set of shared features of the lives of multiple individuals typically belonging to a well-defined group. Sometimes the features documented and the structure of the documenta-

[11] The *Clergy of the Church of England Database* 1540–1835 (CCEd), see http://theclergydatabase.org.uk/; *History of Parliament* Online (HoPO), see https://www.historyofparliamentonline.org/; Joseph Foster, ed., *Alumni Oxonienses 1500–1714* (Oxford: Parker, 1891), at *British History Online*, see http://www.british-history.ac.uk/alumni-oxon/1500-1714; John A. Venn, *Alumni Cantabrigienses* (London: Cambridge University Press, 1922–54); searchable version at https://search.ancestry.co.uk/search/db.aspx?dbid=3997; all accessed 20/03/2019.

tion are pre-established by the structure and content of statistical or administrative records. Alternatively, the structure of the prosopography may be determined by the scholarly purposes of the individual or group assembling it. Sometimes the aim is to generate robust, statistic comparisons, or to write a collective biography of a social group. In other instances, the objective is merely to provide a framework within which scarce and scattered fragments of documentation can be pieced together.[12] For this reason, the prosopographical impulse is best established for relatively distant groups about which documentation is fragmentary, such as the Roman Republic and the Byzantine Empire.

Originally compiled and published in a variety of print media, many of the senior prosopographical projects have transitioned to digital media, often pioneering the application of computing technology to humanistic scholarship in the process.[13] The enormous advantages of digital media for prosopographical work are essentially two-fold. On the one hand, the accessibility and malleability of digital media are ideally suited to resources produced collaboratively by many people in many places which evolve continuously over time. On the other hand, the structured nature of prosopographical data can be rendered machine-readable more readily than prose biographies, allowing computational methods to be applied to the analysis and visualization of large bodies of complex data. It is therefore justified to speak of 'digital prosopography' when the basic biographical material comes in the form of machine-readable data, and the processing and analysis of the data is carried out by computational means.

3.2 Purpose

Disambiguation requires data. Authority files provided with unique identifiers are a necessary precondition for large-scale disambiguation, but not a sufficient condition. The question of whether two similar names refer to one and the same person cannot be resolved with reference to a list of names. It requires, in addition, biographical data on individual people with those names: their dates of birth and death, location histories, and other associations and attributes. The more we know about two historical John Smiths, the more confident we can be about attributing new information to one of them rather than the other: if one died in 1710, we know that he cannot have written a letter dated 1714. For this reason, disambiguation requires access to large bodies of detailed biographical data.

Data for most people is extremely fragmentary. Existing authority files typically lack the references needed to check facts and disambiguate people. For the majority of correspondents not treated in readily available reference works, such facts need to be pieced together from scattered sources. Indeed, even for better attested individ-

[12] See for instance Katharine S. B. Keats-Kohan, ed., *Prosopography. Approaches and Applications. A Handbook* (Oxford: P&G, 2007).

[13] See the *Digital Prosopography of the Roman Empire* (http://romanrepublic.ac.uk/) and the *Prosopography of the Byzantine Empire* (http://www.pbe.kcl.ac.uk/); both accessed 20/03/2019.

uals, the data available in biographical dictionaries is often inadequate to resolve questions of detail, such as the location of an individual at a particular time. For all of these purposes, the scholarly community needs to be able to assemble and interrogate large quantities of highly granular data. The distinctive purpose of the prosopographical system proposed here is not to provide complete data sets or polished prose biographies of individuals absent from standard biographical reference works: it is, rather, to assemble incrementally in usable form scraps of information on lesser-known individuals as they come to the attention of the contributing community of scholars. The key question is how to structure such a data repository so that it can accommodate the kind of data needed for these and related purposes.

Understanding networks requires more than epistolary data. Although letters are a uniquely informative record of personal contact between individuals, they were not the only form of contact. The exchange of letters typically rests on pre-existing networks of direct social exchange, so intellectual networks can only be fully understood with reference to data documenting non-epistolary as well as epistolary contact. In order to reconstruct those social networks, we need more than the epistolary data: we also need contact histories, that is to say, collections of scraps of information documenting the direct, personal encounters between individuals in close physical proximity to one another. Genealogical data are also needed, along with means of capturing confessional, academic, professional, and other associations. Developing standard data models for all aspects of intellectual exchange is therefore a precondition for a more adequate data-driven exploration of the republic of letters.

Data is also needed for a host of other scholarly purposes. Needless to say, many other scholarly purposes are also served by compiling such biographical data. For instance, in order to understand the content of letters, rich biographical information on the correspondents is often required. For this reason, well-edited correspondences generally contain a biographical register of correspondents. Recreating such registers for each, separate edition involves enormous duplication of effort and expense which could be radically reduced by an online repository for information of this kind. To take another example, in order merely to track an individual's itinerary in time and space, a continuous location history is needed. Unless the individual in question has compiled a detailed travel diary, such an itinerary can only be patched together from details recorded in a variety of often very fragmentary sources. A third case is the need for standardized data on such crucial questions as confessional identity, political affiliation, professional activities, institutional memberships, and so on, in order to analyse and understand communities. If such data is to be employed analytically and comparatively, the categories structuring it need to be precise and robust. Confronted by challenges such as these, the problem of structuring prosopographical data rapidly grows in complexity.

3.3 Structure

Attributes versus events. Many prosopographical data models are based on static attributes, such as confession, social status, or profession. Although convenient as a shorthand, these static attributes are inappropriate for an historical data model because most of these attributes change over time. For instance, strictly speaking, a religious confession is not typically acquired at birth but during a ritual (such as baptism) performed some time afterwards. In any case, confessional affiliation can subsequently be changed by religious conversion, never more so than during the era that saw the expulsion of the Jews and Moors from the Iberian peninsula after 1492, the Protestant Reformation, the Ottoman conquest of Hungary, the eventual reconversion of most of the Czech lands, Poland, and Hungary to Catholicism, and the Revocation of the Edict of Nantes in 1685.

Likewise, although an individual may obtain a social status at birth, that status can change repeatedly during their lifetime. Even more obviously, profession is an attribute acquired relatively late in life: Newton was not born as Lucasian professor or master of the Mint. Professions are not exclusive: one can practise (or be qualified to practise) several different professions simultaneously. Nor is professional qualification a precondition for deep involvement in most areas of intellectual interest: the typical citizen of the republic of letters in Restoration England was an amateur: the gentleman virtuoso. In consequence, treating profession as a unitary attribute is extremely problematic: Leibniz was – simultaneously or successively – lawyer, court councillor, mathematician, mining engineer, librarian, court historiographer, journal editor, and president of the Berlin Academy of Sciences, to name only a few of his primary spheres of activity.

Strictly speaking, the only exceptions are the attributes acquired at birth: sex, parentage, and (via parents) other living family relations and ancestors. These core data (on which see section 3.6 below) should be handled as part of the 'birth' event. All attributes acquired after birth should be regarded as events.

Modelled prosopographically in this fashion, a life is an 'event stream', beginning with birth, ending with death, and including an indefinite number of documented events in between, typically relating to other persons at specific places and times. The more granular and complex the data becomes, the more time-consuming entering it will be. This raises the key practical question of how to balance the need for intuitive and efficient inputting of data with the need for rigour and detail (see section 3.5 below).

3.4 References

Another serious limitation of existing authority files (and many prosopographical data models) is that they are based on unreferenced assertions. Again, the origin of these authority files in library catalogues is one explanation for this defect: in order to check the reliability of a traditional authority file, one needs merely to return to

the library catalogue from which it has been derived and thence to the books themselves.

For scholarly purposes, however, mere assertion is inadequate. Scholars need to know the basis on which assertions are being made. In a digital resource, the basis of an assertion – whether in primary sources, secondary literature, reference works, or further data – should ideally be linked directly from the assertion itself. Moreover, because interpreting the sources and providing the documentation requires scholarly work, a prosopographically enhanced authority file must also credit the authorship of contributions, accommodating references, source documentation, biographical data, and scholarly credit for each iota of data assembled.

A solution to this prolem is provided by the so-called factoid approach to data modelling, developed and implemented in several research projects in King's College, London. A factoid is a 'spot in a source that says something about a person or persons'.[14] In other words, the factoid model identifies a relationship between a given source and a given person, and also attributes the identification of that relationship to a specific author working at a specific point in time. For instance, as noted above, several different places have been tentatively identified as the birthplace of Comenius. Rather than simply choosing one or mentioning all three, the factoid model documents each of these claims with reference to one or more specific texts and (where possible) attributes each to a specific author and time as well as embedding all these identifications in the data. Several large projects of digital prosopography have been successfully working with this model,[15] and software solutions are already available for implementing this approach.[16] One of the central consequences of adopting the factoid model is the 'transformation of texts into databases', that is, the reprocessing of biographical narratives into collections of source references.[17] Given the complexity of the documentary heritage of the republic of letters and the centrality of practices of self-fashioning to it, the next component of this project must be to review the sources that can inform new and old-style prosopographies of the commonwealth of learning.

3.5 Sources

Secondary sources. Ultimately, all assertions of biographical fact must be based on primary sources; but anchoring everything in primary sources would impose unsupportable burdens on the community populating this resource, above all in its

[14] See https://factoid-dighum.kcl.ac.uk/what-is-factoid-prosopography-all-about/, accessed 20/03/2019.

[15] See https://factoid-dighum.kcl.ac.uk/factoid-prosopographies-at-cchddh-kcl/, accessed 20/03/2019.

[16] For instance, see http://pdr.bbaw.de/software/ae/, accessed 20/03/2019.

[17] John Bradley and Harold Short, 'Texts into Databases. The Evolving Field of New-style Prosopography', *Literary and Linguistic Computing* 20:suppl. 1 (2005): 3–24, see https://doi.org/10.1093/llc/fqi022.

initial stages. The most convenient and manageable starting point for gathering details on obscure but not entirely unknown individuals will therefore often be via reference works and other secondary literature – including (in the case of the *respublica litteraria*) the extensive genre of *historia litteraria* generated from the seventeenth century onwards[18] and (more generally) the wealth of biographical reference works that proliferated in the nineteenth century. Harvesting this material could potentially be accelerated by Natural Language Processing and Named Entity Recognition techniques, which are currently being used to extract biographically relevant data from the full texts of Dutch biographical dictionaries already available in machine-readable form.[19]

Primary sources. Secondary works are useful points of departure because they contain generalizations based on larger data sets. Those generalizations are useful as points of departure, adequate for some purposes, and permanently valuable for those who want only the summary of the data. For the specialist, however, a far more detailed picture is required, which must ultimately be drawn directly from the primary sources. For instance, for some purposes it might be adequate to record merely that in 1663–6 the young Martin Lister (1639–1712) travelled to Montpellier to obtain some of the medical learning for which he was later famous. But the travel journal he kept at the time allows a far more detailed account of this formative journey, including the route of his outbound and inbound travels, the staging posts and places he visited along the way, the people he met, the letters he sent and received (many no longer extant), and the books he read – not to mention his natural historical expeditions, the dissections he performed, and the salons, gardens, and libraries he visited.[20] An adequate prosopographical system will need to allow the inputting first of a summary record of Lister's voyage, and then of such rich biographical detail, while retaining the summary record as a means of navigating the broad outline of Lister's biography.

As noted above, perhaps the most distinctive purpose of this system would be to provide a structured repository for stray references to poorly documented people. If twelve different people inputted twelve different events related to an itinerant intellectual in twelve different archives in multiple cities and countries, the result could be the emergence of a coherent profile of a significant figure who might otherwise have remained virtually unknown.

---

[18] *Historia litteraria*, in broad terms, means 'history of learning'. During the seventeenth and eighteenth centuries, several (bio-)bibliographies and journals were published under that heading. The goal was to provide a structured overview of the development of a specific knowledge discipline, and of extant knowledge in general. For a useful introduction, see Frank Grunert and Friedrich Vollhardt, eds., *Historia literaria. Neuordnungen des Wissens im 17. und 18. Jahrhundert* (Berlin: Akademie-Verlag, 2007).

[19] http://www.biographynet.nl, which draws material from the Dutch Biography Portal: http://www.biografischportaal.nl/en/; a similar procedure is being adopted to the *Österreichisches Biographisches Lexikon* within the framework APIS: https://www.oeaw.ac.at/acdh/projects/apis/; see also https://sites.google.com/view/bd2019, all accessed 20/03/2019.

[20] Anna Marie Roos not only reproduces the text of this journal but richly illustrates and annotates its content in 'Every Man's Companion: Or, An Useful Pocket-Book. The Travel Journal of Dr Martin Lister (1639–1712)', http://lister.history.ox.ac.uk/index.html, accessed 20/03/2019.

At the other end of the spectrum, the system might also be populated by ingesting institutional or other records documenting tens of thousands of people in a structured manner. For instance, uploading the matriculation register of Leiden University, from 1575 to 1811, would not only furnish a data set highly relevant to the broader subject of intellectual traffic and exchange: it would also create tens of thousands of event records which could subsequently be incorporated into person records when matches with, for example, letter records became apparent. Needless to say, to process large data sets of this kind, the use of the reconciliation tools outlined in chapter III.2 will be needed.

3.6 Core versus Supplementary Data

Combining ease of use with sophistication. The approach to modelling people outlined above raises an obvious practical difficulty. A prosopographical spreadsheet of the traditional kind simply assigns certain attributes to people without providing a basis for the assertion. The system advocated here (1) treats those attributes as the products of events at particular places and times; (2) recognizes that the knowledge about those events mostly depends on written documentation; (3) references that documentation as evidence for the events; and (4) attributes the provision of evidence to specific scholarly contributors. The flexibility, accuracy, and reliability of the data modelled in this way is greatly increased, but the cost in additional time and effort of creating that improved data is potentially very high. This cost also needs to be minimized in order to maximize scholarly contribution to the system. In other words, what is needed is a balance between opposite features: we need a system that is easy to learn and to use but also comprehensive and sophisticated, capable of covering many cases in a detailed and rigorous fashion.

Durability and flexibility are two other opposing characteristics that need to be balanced. On the one hand, a data model must be robust and user-tested before it is opened up for general use, so that it does not require fundamental overhaul after launch. Few scholars will be tempted to spend lengthy periods of precious time inputting data into a model or system that may need comprehensive restructuring or may even be rendered altogether obsolete. Yet, on the other hand, it is highly unlikely that any such prosopographical model will be perfect at launch and not require further refinements, emendations, or extensions. Indeed, the only way to test and perfect a system is through dialogue with a large and diverse community of users. So the prosopographical system must combine solid fundamentals with the capacity for adaptation to different uses and evolution in response to changing needs.

Core versus supplementary data. A solution to these interrelated problems is suggested by a basic feature of the EMLO data model: namely, its distinction between core and supplementary metadata. A basic letter record ideally records the six core metadata fields that collectively distinguish one letter from another, namely: sender, place, and date of sending; intended recipient; place of receipt; and reference.

Supplementary data and metadata take many forms, including further information on states, versions, material, and textual features. Similarly, a place record (described above in chapter II.2) consists of core fields and supplementary fields.

This foundational distinction between core and supplementary data fields has several advantages. To begin with, this distinction helps balance ease with comprehensiveness: there is no need for a new user to master the complete data model (for instance, to worry about how to identify watermarks or letter-locking techniques) before beginning to use it. Second, it also helps begin to address the high cost of event-based factoid modelling: a contributor is obliged only to contribute core data (when available) in order to put a letter, place, or person on the system. No EMLO letter record is regarded as 'incomplete' merely because it has not included watermarks or letter-locking. By parity of logic, no person record would be regarded as incomplete merely because every documented event in that person's life had not been grounded in primary sources. A third advantage is to address the need for both durability and flexibility: work on supplying core data can begin first, with supplementary metadata fields added later, without obliging earlier contributors to return and redo their work. This is even more important in the case of prosopography, since it allows a robust person model to be developed carefully, incrementally, one step at a time, as it were, in modular fashion. First, models can be introduced for those attributes that are common to everyone; then those events that are central to the republic of letters can be added; and afterward additional events can be introduced once the expanded data model is carefully devised and tested.

Summary versus detailed metadata. A similar distinction is needed to solve an analogous problem. For the majority of correspondents, documentation is scanty; and one of the primary purposes of this prosopographical system is to capture in structured form every iota of data found anywhere on such little-known figures. But other figures suffer from the opposite problem: they left behind so much documentation that teams of scholars have spent decades merely attempting to survey it. Leibniz is a paradigmatic case. His archive of tens of thousands of documents was placed under seal immediately after his death. A bare chronicle of his life contains well over 2,000 entries, and if his over 800 works, articles, and reviews were added as events of their own, along with the 20,000 surviving letters he exchanged with others, the broad outline of his life's work will become lost in a mass of fine detail.[21] For such extensively documented individuals, we need a means of distinguishing between events of greater and lesser significance, perhaps on a sliding scale. One means of doing so would be to distinguish major categories of events,

[21] Kurt Müller and Gisela Krönert, *Leben und Werk von Gottfried Wilhelm Leibniz: Eine Chronik* (Frankfurt am Main: Vittorio Klostermann, 1969). His c. 50,000 working papers are itemized in Eduard Bodemann, *Die Leibniz-Handschriften der Königlichen öffentlichen Bibliothek zu Hannover* (Hanover, 1889; facs. repr. Hildesheim: Georg Olms, 1966); his published works are listed in Émile Ravier, *Bibliographie des œuvres de Leibniz* (Paris, 1937; facs. repr. Hildesheim: Georg Olms, 1966). The basic online resource for the correspondence is now the *Personen- und Korrespondenz-Datenbank der Leibniz-Edition*, see https://leibniz.uni-goettingen.de, accessed 20/03/2019.

such as changes of location, profession, confession, or social status, as well as major interventions in intellectual life, such as major publications, honours, or first and last letters exchanged with major correspondents.

3.7 Selection: Which Events to Model?

Core person data is used in establishing identities today. The most basic set includes current and past name forms, and date and place of birth. An extended set includes names of parents and their dates and places of birth and marriage. In the case of historical persons, date and place of death are added. The virtue of this basic data is that it is common to all people without exception: everyone is born at a specific time and place to two biological parents; and everyone dies at a specific time and place. As in the case of core letter records, not all of these data are necessarily known in every instance; but records can usually identify an individual or letter unambiguously even if one or more of these data are unavailable.

Supplementary genealogical data potentially takes many forms. One of the most fundamental, simplest to model, and useful for establishing networks is an extension of the core biological data set itself to include brothers and sisters (including half-siblings, adoption, etc.), grandparents and other ancestors, marriages and children, and extended family members by blood and marriage. Needless to say, these supplementary data *cannot* be handled in the same manner as core data: otherwise the regress would theoretically be infinite. A data model is nevertheless needed for those instances in which such supplementary biological data is readily available.

Supplementary data on the 'respublica litteraria'. People are complicated and infinitely various. Modelling every aspect of their behaviour is an infinite task. For this reason, it would be unwise to attempt at the outset to devise a data model for every aspect of the life of every person who has ever lived on this planet, or even for every early modern European. A better strategy is to expand the supplementary aspects of the model incrementally, and to begin with those activities central to the 'republic of letters' or, better still, the 'commonwealth of learning'. In other words, the specialized prosopographical data model needed by this community should focus in the first instance on the kinds of events typically central to the life of an early modern intellectual. This general strategy could be developed in several complementary ways.

Events that establish learned relationships. Since the commonwealth of learning is a system of communication and exchange, one approach would focus on modelling the acts of communication and exchange which constitute that commonwealth. Abstractly considered, these include the various media and modes of exchange. The media of communication include the exchange of spoken, written, and printed words as well as the non-verbal exchange of images and material objects. The modes of exchanging those media include oral exchange through learned institutions, scribal exchange through postal systems, exchange of printed materials through the book market, and so on. Viewed from this perspective, the attempts to

model the exchange of correspondence represent just one of the many communicative events which bound together the republic of letters. In order to capture and analyse data on the republic of letters more generally, data models are needed for the other communicative events as well.

Events documented in huge quantities. A complementary approach would begin with those forms of intellectual exchange for which the most voluminous and systematic documentation survives. Again, this approach sees letters as just one of the kinds of documentation of learned exchange which survives in large quantities. University matriculation registers, for example, document another kind of learned exchange even more systematically than letters. *Alba amicorum* – the friendship books popular with humanists and others in the sixteenth and seventeenth centuries – provide further documentation of the encounters of named individuals as they moved about in time and space. Bibliographical data documents the production of printed books, while the printed book sales catalogues which proliferated from the mid-seventeenth century onwards provide one record among others of book collecting.

Events documented in institutional archives. A third approach, again complementary to the others, is to attempt to model all the transactions systematically documented in the archives of learned institutions. Particularly appropriate for this kind of attention are those institutions, founded in the medieval period, which were replicated across Europe on a fairly standard organizational blueprint which varied relatively little from the beginning to the end of the early modern period. Two examples – universities and learned religious orders – are discussed further in chapter IV.4. Literary academies and scientific societies emerged *during* this period and documented their activities more and less systematically in ways sufficiently similar to one another that they might also be reduced to similar models.

3.8 Conclusion

In the first part of this chapter, an outline was given of the problems regarding the complex relationship between biographical data and learned personae. In the second part, unique identifiers and authority files were presented as a means of disambiguating person records. The third and longest part was dedicated to outlining the prerequisites of a description framework that could provide the basis for a prosopographical data model appropriate to the scholarly crowdsourcing of data on citizens of the republic of letters.

Such a description framework still falls far short of a data model. Translation from the former to the latter would have to begin with a consideration of existing standards in the field of biographical data modelling, including factoid-oriented models[22] and those which are part of the broader TEI framework;[23] historically

[22] See https://factoid-dighum.kcl.ac.uk/fpo-factoid-prosopography-ontology/, accessed 20/03/2019.
[23] See http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html, accessed 20/03/2019.

grounded models,[24] and those which are not.[25] An advantage and limitation at the same time consists in the fact that the discussion is global in scope and nature, and therefore does not always properly address the complexity of historical sources.[26]

One of the most prominent standards for the structured description of cultural artefacts is CIDOC CRM. CIDOC is the International Committee for Documentation convened by the International Council of Museums. CIDOC CRM is the 'Conceptual Reference Model' devised by the committee to provide 'definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation'.[27] Since cultural heritage is created by people, the CIDOC CRM includes the basis of a prosopographical data model. The concluding part of the chapter describes an attempt at rendering prosopographical data from EMLO as Linked Open Data, by way of a domain-specific extension of CIDOC CRM.

4 Implementation: Representing 'Early Modern People' as Linked Data

Jouni Tuominen, Eetu Mäkelä, and Eero Hyvönen

As noted in chapter I.3, Linked Data is a promising approach for aggregating heterogeneous, distributed data on the republic of letters on a pan-European level. The core of the Linked Data paradigm is the use of unique identifiers for entities, such as people and places, in order to assemble data on those entities scattered throughout the Internet. Linked Data facilitates the publication of ontologies covering such entities for shared authority control.

As a pilot Linked Data publication, the *Early Modern Letters Online* (EMLO) database has been converted into a graph-based RDF data model.[28] In this section we shall introduce the idea of using the early modern person collection of EMLO as a Linked Data service with open APIs and user interfaces. We will outline a Linked Data modelling approach for biographical information of early modern people.

In addition to purely epistolary data, EMLO contains prosopographical information related to the people in the database, modelled as events and social rela-

[24] See http://ontologies.symogih.org/index.html, accessed 20/03/2019.

[25] See https://duraspace.org/vivo/about/, accessed 20/03/2019.

[26] See https://zenodo.org/record/1041978#.XA-mazGNy72, accessed 20/03/2019.

[27] See ICOM: https://icom.museum/en/; CIDOC: http://network.icom.museum/cidoc/; and CIDOC CRM: http://www.cidoc-crm.org/, all accessed 20/03/2019.

[28] This idea and system is presented in more detail in Jouni Tuominen, Eetu Mäkelä, Eero Hyvönen, Arno Bosse, Miranda Lewis, and Howard Hotson: 'Reassembling the Republic of Letters - A Linked Data Approach', in Eetu Mäkelä, Mikko Tolonen, and Jouni Tuominen, eds., *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference*, 76–88, see http://ceur-ws.org/Vol-2084/paper6.pdf, accessed 20/03/2019.

tionships. Events cover activities that the people have participated in during their lives, such as birth and death, ecclesiastical and educational activities, creation of works, and changes of location and residence. The event metadata includes the event name, type, participants and their roles, time span, location, and source information. The prosopographical data was converted into RDF format using CIDOC CRM for the event-based modelling and W3C's PROV model for representing the roles of participants in the events. The event-based modelling is discussed in more detail in chapter II.6.

After the RDF conversion, the EMLO data was published in a SPARQL endpoint, a technology standardized by W3C. This approach allows the data to be queried in a distributed manner and utilized as a semantic glue to integrate different databases. For efficient use of the shared ontologies, a user interface component, Federated SPARQL Search Widget,[29] can be integrated, for instance, into letter cataloguing systems. Using such an approach, different data providers already receive strong identifiers for people and places as part of the data input process, with no need to reconcile the data later.

As a continuation of the effort of modelling early modern people, CIDOC CRM was extended by introducing role-centric modelling, known as the Bio CRM model.[30] Bio CRM provides the general data model for biographical data sets. The individual data sets concerning different cultures or time periods, or collected by different researchers, may introduce extensions for defining additional event and role types.

The core design principles of the Bio CRM data model are:

- The model is a domain-specific extension of CIDOC CRM, making it applicable not only to biographical data but also to other cultural heritage data as well.
- The model makes a distinction between enduring unary roles of actors, their enduring binary relationships, and perdurable events, where the participants can take different roles modelled as a role concept hierarchy.
- The model can be used as a basis for semantic data validation and enrichment by reasoning.
- The enriched data conforming to Bio CRM is targeted to be used by SPARQL queries in flexible ways, using a hierarchy of the roles in which participants can be involved in events.

[29] See https://github.com/SemanticComputing/federated-sparql-search-widget, accessed 20/03/2019.

[30] For more details of the model, see Jouni Tuominen, Eero Hyvönen, and Petri Leskinen, 'Bio CRM: A Data Model for Representing Biographical Data for Prosopographical Research', in Antske Fokkens, Serge ter Braake, Ronald Sluijter et al., eds., *Proceedings of the Second Conference on Biographical Data in a Digital World 2017*, 59–66, see http://ceur-ws.org/Vol-2119/paper10.pdf, accessed 20/03/2019.

Bio CRM makes a clear distinction between a person's attributes, relations between people, and events in which people participate in different roles:

- *Attributes* are properties of persons that are assumed to characterize them independently of time and space. For example, place and time of birth can be modelled as attributes.
- *Relations* are established between people and are assumed to characterize the people independently of time and space. For example, father-of is such a relation. Relations can, however, have time and space as qualifiers, e.g. student-of. For example, Ferdinand Bol (1616–1680) was a student of Rembrandt in 1630–41, starting his own studio in 1641, but can be characterized as a student of Rembrandt in general. If a distinction between these two senses is needed, his years in Rembrandt's studio as a student can be represented as an event.
- *Events* take place in time and space and involve participants in different roles, expressing the ways in which persons participate in events – for example, an officiant in a certain baptism event.

In Bio CRM, a person is represented as an instance of bioc:Person, a subclass of cidoc:E21 Person. This instance-of relationship is persistent and never changes during the life of the person. In order to identify a person, the person is associated with core data: appellations (i.e., names and identifiers in other data repositories), birth time and place, and death time and place, using CIDOC CRM. A person's birth and death are represented as a Birth/Death event, which can be qualified with time and place. Birth can also incorporate information about the mother and father.

In addition to core data, a person can also have other attributes and relationships, and participate in events. Having a role (e.g. 'teacher') may be temporary or something inherent to that person at all times, even if it is possible also to specify when exactly the role was present (e.g. a professorship). Being a teacher by virtue of one's education is different from saying that the person happened to participate in a particular teaching event (such as giving a lecture) or occupying a particular institutional post (such as that of extraordinary professor in a university).

One main advantage of using roles is that they allow the number of properties to be minimized, because different properties for different roles are not needed. Instead, different role classes are used. Such a model is simpler to query using SPARQL and provides the user with a natural hierarchy of role concepts.

II.5 Topics

Howard Hotson and Eero Hyvönen

Standards and tools are needed for navigating the chronological, geographical, and social dimensions of correspondence networks (chs. II.2–4 above). These needs are obvious and undeniable; and although each of these three dimensions is more challenging than the previous one, work on providing each is well underway. Scholars also need tools for navigating the conceptual or topical dimension of letter corpora; but providing a conceptual gazetteer might at first sight seem impossible, and no such tools are currently in the making, at least not for the early modern period. This chapter first explores the desirability of such a tool by developing the analogy of a conceptual gazetteer with a geographical one (section 1). It then explores the feasibility of such a tool by surveying approaches developed for generating similar tools for navigating modern knowledge (section 2). Finally, it provides an example of how a historically structured topical gazetteer might be constructed (section 3).

1 Desideratum

1.1 A Historically Structured Geographical Gazetteer

As already established in chapter II.2, a historically structured geographical gazetteer is needed in order to navigate efficiently throughout the physical geography of the republic of letters. Such a gazetteer has three functions: disambiguation, location, and contextualization.

Disambiguation minimizes ambiguity by referring to a single unique identifier all the different ways of naming a given place in different languages, periods, and traditions.

Location situates a physical place precisely on a set of global coordinates, further distinguishing it from any other place with a similar name or location.

Contextualization is needed to arrange the places thus named and located in multiple, parallel, evolving hierarchies. Hierarchies are needed because each place may be both subdivided into smaller places and situated within larger ones. Multiple, parallel hierarchies are needed because a place may be situated within multiple, parallel systems of organization – political, ecclesiastical, military, judicial, and so on – which overlap with one another without sharing boundaries. Evolving hierarchies are needed because both the official name of a place and its situation within these hierarchies may change over time.

The first two of these functions are reasonably well served by existing gazetteers, while the third requires new data and functionality.

1.2 A Historically Structured Conceptual Gazetteer

Although intellectual historians are interested in the spatial relationships between intellectuals, they are even more interested in the conceptual relationships between ideas. They therefore need aids for navigating abstract conceptual geography even more than they need aids for navigating concrete spatial geography. The question therefore arises whether it is possible to envisage a kind of historically structured conceptual gazetteer, in loose analogy to the functions of a geographical gazetteer outlined above.

Disambiguation. Creating efficient means of navigating the conceptual space of a correspondence requires addressing the problem of ambiguity. Early modern learned correspondence employs many languages. In each language there are often multiple more or less synonymous terms which might be used as keywords or index entries for a particular concept. If no order is imposed on the choice of keywords, their utility is greatly reduced: if ten different indexers use ten different words to refer to one and the same thing, they will not collectively create efficient means of navigating their shared domain. In analogy to the unique identifiers for variously named places, intellectual historians ideally need standard terms for designating concepts or topics encountered within correspondence corpora. In a traditional hard-copy edition of a correspondence, this problem is addressed in a pragmatic manner in devising a topical index. In a collaboratively populated union catalogue of correspondence, the analogous solution would be to develop standard vocabularies of keywords for mapping the conceptual dimension of letters in a systematic way. However, since the 'linguistic turn', intellectual historians have become acutely sensitive to the ways in which many key terms and concepts have shifted meaning in the past, and this shift in meaning is now one of the main ob-

jects which they study.[1] Historians therefore need navigational aids more subtle than a single set of keywords: they need semantic nets capable of capturing the multiple meanings of words at any one time and the ways in which the meaning of those words change over time.

Location. The second function of a geographical gazetteer is to pinpoint places in space. For this function, there is no obvious analogy in a conceptual gazetteer. Ideas and topics do not exist in a two- or three-dimensional space, and therefore cannot be pinpointed with reference to a simple set of mathematical coordinates. As a consequence, the only way in which we can 'map' the relationship of ideas and disciplines with one another is via the semantic nets mentioned above and the hierarchical relationships discussed below.

Contextualization. Not unlike spaces, concepts are often arranged in hierachies: a basic conception can be broken down into its component parts, or clustered together with others to make fields or disciplines. Conceptual hierarchies can therefore serve to situate individual terms within knowledge organization structures. But these hierarchies must be multiple and evolving, because competition between intellectual systems is a central feature of early modern intellectual history. This basic problem requires more extended discussion below (1.3).

1.3 The Key Requirement: Mapping between Multiple, Parallel, Evolving Hierarchies

An example may help clarify this point. Before Copernicus, the realm above the moon was regarded as different in substance from the terrestrial realm; planets were regarded as wandering stars, the study of the movement of the planets was therefore regarded as part of astronomy, and astronomy was typically situated as one of the four mathematical disciplines of the medieval quadrivium, which in turn made up an intermediate part of the undergraduate university curriculum. After Newton, the study of the movement of the planets came to be regarded as part of celestial mechanics, which was a branch of physics and a part of natural philosophy, which formed part of the more advanced, philosophical level of the undergraduate curriculum. Between Copernicus and Newton, the location of the subject of planetary motion within the circle of the disciplines was in flux. No single disciplinary hierarchy is capable of capturing the shifting location and position of this topic (the movement of the planets) within this disciplinary matrix (astronomy, cosmology, physics, natural philosophy, and indeed natural science), especially

[1] See for instance John E. Toews, 'Intellectual History after the Linguistic Turn: The Autonomy of Meaning and the Irreducibility of Experience', *The American Historical Review* 92:4 (1987): 879–907, see https://doi.org/10.1086/ahr/92.4.879; Quentin Skinner, *Visions of Politics*, vol. 1: *Regarding Method* (Cambridge: Cambridge University Press, 2002); Reinhart Koselleck, 'Introduction and Prefaces to the *Geschichtliche Grundbegriffe*', *Contributions to the History of Concepts* 6:1 (2011): 1–37, see https://doi.org/10.3167/choc.2011.060102.

when the very nature and content of the disciplines are constantly being renegotiated.

Moreover, during the eighteenth and ninteenth centuries, reorganization of the arts and sciences was undertaken on a large scale. The modern natural sciences emerged from 'natural philosophy'; the modern social sciences emerged from 'moral philosophy'; and in consequence 'philosophy' itself collapsed from an all-embracing discipline, pursuing knowledge for its own sake throughout all the areas of speculative thought, to a far more compact discipline dealing with a smaller number of fundamental problems such as existence, knowledge, values, mind, and language. Given the radical divergence of these two conceptions, it is futile to attempt to navigate the space of ancient, medieval, or early modern philosophy equipped solely with the modern conception of the discipline; still less can the modern categories help us understand how the older conception mutated into the modern one. To navigate this domain efficiently, we need to be able to map the shifting location of topics and disciplines within multiple, parallel, evolving hierarchies.

This kind of radical redefinition was not unique to 'philosophy'. Other major disciplinary categories – such as 'art', 'science', 'history', and 'religion' – underwent equally dramatic reconfigurations in this period.[2] One might even go further to suggest that this process of disciplinary redefinition was the defining feature of the early modern period of intellectual history. Early modern intellectal history begins when ancient and medieval categories are called into question and ends when those categories assume more or less their modern configuration. In any case, these shifts explain why fully modern knowledge organization structures are anachronistic and of limited utility when attempting to understand the early modern period. The republic of letters was the community within which that transformation took place. As a consequence, it is impossible to navigate the conceptual landscape of the republic of letters efficiently without the multiple, parallel, evolving conceptual hierarchies capable of mapping between the various stages of this transformation.

How then to set about devising such a navigational tool? In pursuit of this question, the following section reviews approaches to mapping the organization of knowledge today and notes their roots in analogous early modern literatures. We then review a particularly fertile genre of early modern intellectual production which gave rise to a specific example of how early modern materials might be redeployed to create the kind of navigational aid we need.

---

[2] Władysław Tatarkiewicz, 'Classification of the Arts', in Philip P. Wiener, ed., *The Dictionary of the History of Ideas*, 5 vols. (New York: Scribner's Sons, 1973–4); Katharine Park and Lorraine Daston, eds., *Early Modern Science*, in David C. Lindberg and Ronald L. Numbers, eds., *The Cambridge History of Science*, vol. 3 (Cambridge: Cambridge University Press, 2006); Gianna Pomata and Nancy G. Siraisi, eds., *Historia: Empiricism and Erudition in Early Modern Europe* (Cambridge, MA: The MIT Press, 2005); Peter Harrison, *The Territories of Science and Religion* (Chicago: University of Chicago Press, 2017).

2 Modern Knowledge Organization: Parallel Literatures and Early Modern Roots

The flood of new information being made accessible by digital technology has raised in particularly acute form the problem of organizing topics of knowledge and representing them in navigable structures. Solutions to this problem are currently being sought in several cognate disciplines, including (1) philosophy and ontology, (2) linguistics, (3) terminology, (4) library and information sciences, and (5) computer science. This sections considers each of these approaches in turn.

Early modern intellectuals, seeking to manage the flood of new information unleashed in the early age of print, faced analogous problems;[3] and many of the strategies currently being pursued are digitally supercharged versions of solutions devised in the sixteenth and seventeenth centuries. The parallels between the old solutions and the new raise a hitherto unexploited possibility: namely, the possibility of using new, digital techniques for redeploying these older solutions in order to allow scholars of the early modern period to navigate their early modern data by means of early modern terms, categories, and hierarchies.

2.1 Philosophy/Ontology

The original meaning of the term 'ontology' comes from philosophy. Ontology in this sense is the branch of philosophy that studies 'the most general features of what there is, and how the things there are relate to each other in the metaphysically most general ways'.[4] This notion originates from the metaphysics of Aristotle (384–322 BC). In his ontology, Aristotle identified ten categories as a flat list, such as Substance, Quality, Quantity, Relation, Position, etc.[5] He also introduced the first formal logic system based on propositions and domain-independent rules, syllogisms. Porphyry of Tyre (234–c. 305), a Greek Neoplatonist, arranged ontological categories into a dichotomous hierarchy of sub-/supertypes, thus introducing the idea of a semantic net. For example, body can be animate or inanimate, animate can be rational or irrational, and so on. Medieval scholastics and logicians, such as Peter of Spain, developed ontological ideas further and reduced them to graphic form (fig. 1). During the sixteenth and seventeenth centuries, the Ramist tradition pursued the project of reducing the entire field of knowledge to a net of definitions and divisions represented by continuous series of dichotomous tables; and it was from this tradition (as we shall see more below) that the term 'ontology'

[3] Ann Blair, *Too Much to Know: Managing Scholarly Information before the Modern Age* (New Haven: Yale University Press, 2011).

[4] Thomas Hofweber, 'Logic and Ontology', in Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Summer 2018 edn.), see https://plato.stanford.edu/archives/sum2018/entries/logic-ontology/, accessed 20/03/2019.

[5] Paul Studtmann, 'Aristotle's Categories', in Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Fall 2017 edn.), see https://plato.stanford.edu/archives/fall2017/entries/aristotle-categories/, accessed 20/03/2019.

emerged. Today, ontological research in philosophy aims to develop formal models of foundational categories and logic behind everything.[6]

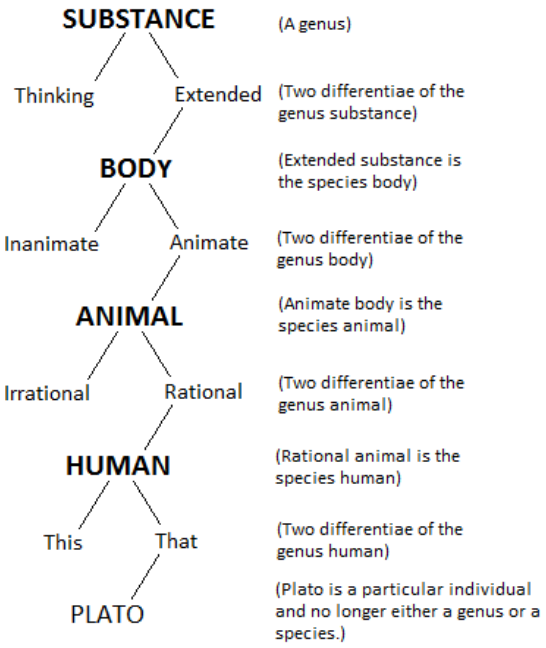


Figure 1: Subcategories of substance in graphical form according to Peter of Spain (1329)[7]

In other disciplines, the term 'ontology' carries different meanings; and further confusion is created by the overlap of the term 'ontology' with 'vocabulary', which also has its own established connotations. According to W3C,[8] vocabularies define the concepts and relationships (also referred to as 'terms') used to describe and represent an area of concern. This means that there is no clear distinction between 'ontologies' and 'vocabularies' in the context of the Semantic Web, even though the term 'ontology' is usually preferred for more complex and formal knowledge organization systems. Furthermore, both terms are used to refer to different kinds of knowledge organization systems, such as large domain-specific gazetteers, classifications, and controlled vocabularies of up to millions of concepts, or small do-

[6] John Sowa, *Knowledge Representation. Logical, Philosophical, and Computational Foundations* (Pacific Grove, CA: Brooks/Cole, 2000).

[7] Source: 'Porphyrian Tree', Wikipedia (Licence CC BY-SA 3.0).

[8] See http://www.w3.org/standards/semanticweb/ontology, accessed 20/03/2019.

main-independent metadata schemas of tens of concepts. The following paragraphs briefly outline perspectives on ontology within other scientific disciplines.[9]

2.2 Lexicography and Linguistics/Words

In linguistics, the idea of ontologies is closely related to the idea of semantic dictionaries, where words are organized not alphabetically but according to their meaning. The design, compilation, use, and evaluation of dictionaries is studied in the field of lexicography.

An early example of a semantically organized dictionary is Roget's *Thesaurus*,[10] a work that has been continuously in press since its first publication in 1852. Within it, the world is divided into 1,000 categories organized into a three-level hierarchy. The most widely used (psycho-)linguistic vocabulary/ontology on the Semantic-Web today is WordNet.[11] In a manner roughly comparable to Roget's work, word meanings in WordNet are organized into cognitive synonym sets called synsets that refer to concepts. Homonymous word forms with different meanings belong to several synsets. The synsets based on substantives, verbs, adjectives, and adverbs are organized into hierarchical subnets of their own. WordNet is part of the Linked Open Data Cloud[12] with many mappings to related data sets. The English WordNet has been translated at least partly into many other languages.[13]

The semantically organized dictionaries created in the early modern period could provide a starting point for developing similar tools for organizing historical materials in appropriate structures. One example is illustrated in section 3.5.

2.3 Terminology/Terms

Terminology is the study of terms and their use.[14] The major problem addressed in this discipline is the confusion created by the use of words with multiple meanings in everyday life and in different professional contexts. Major sources of ambiguity in determining the references for a word include synonymy, where several words have the same meaning (e.g. 'buy' and 'purchase'); polysemy, where a single word has different but related meanings (e.g. 'head' of an arrow vs. 'head' of a man), and

---

[9] Based on Eero Hyvönen, *Linked Data: Publishing and Using Cultural Heritage Linked Data on the Semantic Web* (Palo Alto, CA: Morgan & Claypool, 2012).

[10] Barbara Ann Kipfer, ed., *Roget's International Thesaurus*, 7th edn. (New York: Harper Collins Publishers, 2011).

[11] Christiane Fellbaum, ed., *WordNet. An Electronic Lexical Database* (Cambridge, MA: The MIT Press, 2001).

[12] See http://linkeddata.org/ for the international Linked Data movement and https://lod-cloud.net/ for the Linked Open Data cloud of openly available, mutually interlinked data sets; both accessed 20/03/2019.

[13] See http://www.globalwordnet.org/, accessed 20/03/2019.

[14] Heidi Suonuuti, *A Guide to Terminology* (Helsinki: Finnish Centre for Technical Terminology, 2001).

homonymy, where a single word has different unrelated meanings (e.g. 'bank' as a business and as a river bank).

Terminological analysis is a standardized methodology (by ISO 704:2009) for clarifying language use by creating specific terminologies for different communities with precise, harmonized definitions. Concepts are analysed and defined by organizing them into a concept system that is represented as a graph using four major relation types: equivalence, genericity, partitivity, and associativity. Terminological definitions are typically normative, that is, they are recommendations for a particular interpretation and use of terms in a domain. The analysis may be multilingual.

Terminological analysis, designed to eliminate ambiguity by means of normative definitions, is by no means confined to the contemporary world. The project of creating an artificial language free of ambiguity and mapping onto a hierarchical structure for organizing all of knowledge was a central project of the seventeenth century, intimately related to the advent of modern science. The vast body of source material generated by this enterprise, and the voluminous historical literature analysing it,[15] provide resources for mapping the structures of knowledge which proliferated in this period. None of these multiple systems created in the seventeenth century became sufficiently normative to constitute an ontology shared by a significant portion of the early modern intellectual world. But they nevertheless were central to the process of restructuring learning outlined above, and eventually produced results of enduring value, such as Linnaeus's system of biological classification.

2.4 Information and Library Science

Thesaurus[16] construction is also studied in Information Science, Library Science, and Information Technology.[17] Here the goal is to create controlled vocabularies for indexing or tagging purposes and for use in information retrieval systems.[18] A thesaurus of this kind consists of terms and semantic associations between them. The main categories of them are equivalence, hierarchy, and association. However

[15] See for instance the pioneering work by Paolo Rossi, *Clavis universalis: Arti della memoria e logica combinatoria da Lullo a Leibniz* (Milan: Ricciardi, 1960; rev. edn. Bologna: Il Mulino, 1983, 2006); trans. by Stephan Clucas as *Logic and the Art of Memory: The Quest for a Universal Language* (London: Continuum, 2000); James Knowlson, *Universal Language Schemes in England and France, 1600–1800* (Toronto: University of Toronto Press, 1975); Mary M. Slaughter, *Universal Languages and Scientific Taxonomy in the Seventeenth Century* (Cambridge: Cambridge University Press, 1982); Jaap Maat, *Philosophical Languages in the Seventeenth Century: Dalgarno, Wilkins, Leibniz* (Dordrecht and Boston: Springer, 2004); Rhodri Lewis, *Language, Mind and Nature: Artificial Languages in England from Bacon to Locke* (Cambridge: Cambridge University Press, 2007).

[16] Douglas J. Foskett. 'Thesaurus', in Allen Kent, Harold Lancour, Jay E. Daily, eds., *Encyclopaedia of Library and Information Science*, vol. 30 (New York: Marcel Dekker, 1980), 416–62.

[17] Jean Aitchison, Alan Gilchrist, and David Bawden, *Thesaurus Construction and Use: A Practical Manual* (London: Europa Publications, 2000).

[18] Lots of standards have been created for thesaurus construction, including the widely used monolingual standard ISO 2788, its British equivalent, BS 5723, and the US Standard ANSI/NISO Z39.190-1993.

(as in terminology), there are many more refined relations that can be employed for partial equivalence, class-instance relationship, refined hierarchical relationships, or refined associative relationships. When cataloguing a book, a cataloguer selects terms from a controlled vocabulary to supply the metadata element values describing the book's content in a harmonized way. During information retrieval, the same terms can be used for constructing queries. Similarly constructed thesauri based on historical principles could be developed for cataloguing early modern material.

A related tool widely used in libraries is the classification system. The idea of classification systems is to create a system of categories into which material can be assigned.[19] A typical example is the Dewey Decimal Classification, used for storing books in particular positions on the shelves. The focus is different form thesauri and ontologies which concentrate on describing concepts.

As Europe was flooded with printed books in the sixteenth and seventeenth centuries, manuals appeared on the principles for selecting and organizing a choice library which pioneered developments fundamental to modern information and library science. The burgeoning libraries of bibliophilic rulers, patricians, and universities were designed to provide conspectus of the whole of knowledge. Since books are physical objects, this genre confronted the problem of arranging the whole of knowledge not merely conceptually but also spatially. In doing so, this genre articulated principles which might also be redeployed in developing a conceptual gazetteer or ontology for the period in which they were written.[20]

In many cases, using only one thesaurus of classification is not feasible for classifying data, but a set of orthogonal thesauri or classifications schemes is needed, as suggested in facet analysis,[21] a classification scheme introduced in information sciences by S. R. Ranganathan in the 1930s. For example, consider books published in different languages, on different topics, and in different countries. If they were classified using only one hierarchy, this would contain massive amounts of categories for virtually all combinations of languages, topics, and countries. Using facets, one can simply classify the books by triples (for language, topic, country) without enumerating these as categories in a large hierarchy. The facets themselves can be organized in a natural way as smaller hierarchies. The corresponding idea of faceted search[22] and browsing[23] was invented and developed independently by

[19] Daniel Joudney, Arlene Taylor, and David Miller, *Introduction to Cataloging and Classification. Library and Information Science Text Series*, 11th edn. (Santa Barbara, CA: Libraries Unlimited, 2015).

[20] See for instance Helmut Zedelmaier, *Bibliotheca universalis und Bibliotheca selecta. Das Problem der Ordnung des gelehrten Wissens in der frühen Neuzeit* (Cologne, Weimar, and Vienna: Böhlau, 1992); Paul Nelles, 'Books, Libraries and Learning from Bacon to the Enlightenment', in Giles Mandelbrote and Kim Manley, eds., *The Cambridge History of Libraries in Britain and Ireland*, vol. 2 (Cambridge: Cambridge University Press, 2006), 23–35; Eric Garberson, 'Libraries, Memory and the Space of Knowledge', *Journal of the History of Collections* 18 (2006): 105–36, see https://doi.org/10.1093/jhc/fhl011.

[21] Amanda Maple, *Faceted Access: A Review of the Literature. Technical report, Working Group on Faceted Access to Music* (Middleton: Music Library Association, 1995).

[22] Daniel Tunkelang, *Faceted Search* (Palo Alto, CA: Morgan & Claypool, 2009).

several research groups, and is also called 'view-based search'[24] and 'dynamic taxonomies'.[25] The principles deployed in developing sets of orthagonal thesauri and classification systems could be adapted for the purpose of providing historians with 'multiple, parallel hierarchies' of terms and categories, with which to organize their material.

2.5 Computer Science

In computer science, the term 'ontology' (with a lower-case letter) is used to refer to a formal data structure that can be processed using algorithms.[26] An ontology is a formal, explicit specification of a shared conceptualization. The key words in the definition are: (1) formal, i.e. an ontology has well-defined syntax and semantics; (2) explicit, i.e. an ontology can be represented and processed algorithmically; (3) shared, i.e. an ontology is agreed upon in a community and facilitates communication between its member agents; (4) conceptualization, i.e. an ontology presents a model of the real world.

Research on computational ontologies includes theoretical aspects of foundational ontologies, such as the Basic Formal Ontology,[27] and more practical works in the field of Semantic Web and Linked Data.[28] Semantic Web ontologies[29] are today widely used in digital humanities. Ontology models here are founded on description logics,[30] so the loop back to Aristotle's ontology and idea of reasoning syllogisms has been closed. Description logics facilitate the definition of terms, called 'classes', using other terms and their properties. A class consists of individuals that share the properties of the class and its superclasses. For example, Plato

[23] Marti Hearst, Jennifer English, Rashmi Sinha, Kirsten Swearingen, and Ping Yee, 'Finding the Flow in Web Site Search', *Communications of the ACM* 45:9 (September 2002): 42–9, see https://doi.org/10.1145/567498.567525.

[24] Steven Pollitt, *The Key Role of Classification and Indexing in View-based Searching.* Technical report, University of Huddersfield, UK, 1998.

[25] Giovanni Maria Sacco, 'Dynamic Taxonomies: Guided Interactive Diagnostic Assistance', in Nilmini Wickramasinghe, ed., *Encyclopedia of Healthcare Information Systems* (Hershey, PA: Idea Group Publishing, 2005). The idea was integrated with Semantic Web ontologies in Eero Hyvönen, Samppa Saarela, and Kim Viljanen, 'Application of Ontology Techniques to View-based Semantic Search and Browsing', in *The Semantic Web: Research and Applications. Proceedings of the First European Semantic Web Symposium (ESWS 2004)* (Berlin and Heidelberg: Springer Verlag, 2004), 92–106, see https://doi.org/10.1007/978-3-540-25956-5_7.

[26] Nicola Guarino, Daniel Oberle, and Steffen Staab, 'What Is an Ontology?', in Steffen Staab and Rudi Studer, eds., *Handbook on Ontologies*, 2nd edn. (Berlin and Heidelberg: Springer Verlag, 2009), 1–17, see https://doi.org/10.1007/978-3-540-92673-3_0.

[27] Robert Arp, Barry Smith, and Andrew D. Spear, *Building Ontologies with Basic Formal Ontology* (Cambridge, MA: MIT Press, 2015).

[28] Pascal Hitzler, Markus Krötzsch, and Sebastian Rudolph, *Foundations of Semantic Web Technologies* (Boca Raton, FL: CRC Press, 2009).

[29] Dean Allemang and Jim Hendler, *Semantic Web for the Working Ontologist. Effective Modeling in RDFS and OWL* (Amsterdam: Morgan-Kaufman Publishers, 2008).

[30] Franz Baader, Ian Horrocks, Carsten Lutz, and Uli Sattler, *An Introduction to Description Logic* (Cambridge: Cambridge University Press, 2017).

could be defined as an individual of the class Man that is a subclass of the class Person. Furthermore, analogously to Aristotle's syllogisms, logic rules are used based on modern logic systems; here different subsets of standard predicate logic, that can be processed efficiently by theorem-proving algorithms. The underlying logical data describing the individuals can be seen in a dualistic way as a semantic net whose connections can be enriched with new connections by applying logical rules to the data, i.e. by reasoning. The most central Semantic Web vocabularies, i.e. ontology standards[31] in use are Resource Description Framework (Schema) RDF(S),[32] Simple Knowledge Organization System (SKOS),[33] and Web Ontology Language (OWL).[34]

Knowledge organization systems on the Semantic Web can be deployed as ontology services, based on Linked Data publishing principles,[35] that can be used not only by human users but also by machines.[36] Such services can be shared by their user community and be integrated with legacy systems, such as cataloguing systems in museums and archives, in order to ease the use of ontologies and to foster semantic interoperability when indexing data in a distributed environment. Examples of ontology services online include, e.g. the Bioportal of biomedical ontologies[37] and the Finnish national ontology service Onki/Finto[38] of vocabularies/ontologies for different application domains.

3 Early Modern Knowledge Organization

As the foregoing examples suggest, the project of constructing what we have described as a kind of conceptual gazetteer was energetically pursued in the sixteenth and seventeenth centuries. One particular approach to this project merits closer consideration. Its basic constituents were twofold. One (3.1) was the virtually universal tendency to regard all possible subjects of discussion as 'places' – that is 'topics' or 'loci'. The second (3.2) was the only slightly less widespread tendency to

[31] See https://www.w3.org/standards/semanticweb/ontology, accessed 20/03/2019.

[32] RDF(S) provides the foundational graph-based data model for Linked Data on the Semantic Web with hierarchy systems for representing classes, instances, and their properties, as well as basic constraints for using the properties.

[33] SKOS is an extendable vocabulary developed for representing various knowledge organization systems, such as keyword thesauri, geographical gazetteers, authority files, etc., in terms of RDF(S) and OWL.

[34] OWL is a versatile high-end description logic system for representing complex ontological models that can be used for many reasoning tasks for enriching data.

[35] Tom Heath and Christian Bizer, *Linked Data: Evolving the Web into a Global Data Space*, 1st edn (Palo Alto, CA: Morgan & Claypool, 2011).

[36] Mathieu d'Aquin and Natalia Noy, 'Where to Publish and Find Ontologies? A Survey of Ontology Libraries', *Journal of Web Semantics* 11 (March 2012): 96–111, see https://doi.org/10.1016/j.websem.2011.08.005.

[37] See https://bioportal.bioontology.org/, accessed 20/03/2019.

[38] *Finto* (http://finto.fi) is the deployed version of the national Finnish ontology service prototype ONKI (http://onki.fi); both accessed 20/03/2019.

distribute those 'places' spatially on the printed page by means of logically defined and diagrammatically represented conceptual hierarchies. It was out of this latter branch of the tradition that the term 'ontology' first emerged (3.3). Carried to its logical conclusion (3.4), this tradition produced works which sought to map all the 'places' in the entire domain of human discourse within a single comprehensive conceptual and pedagogical hierarchy (3.5), which could provide an obvious starting point for the modern project of developing a conceptual gazetteer for the early modern period.

3.1 Topics

The defining characteristic of this tradition was to conceive the entire domain of human discourse as a set of 'topics' or 'places'. This is hardly surprising, since etymologically the two words mean the same thing. The English word 'topic' derives from the Ancient Greek word *topos* (τὸπος, plural *topoi*) which means 'place'. In Greek dialectical theory and rhetorical practice, a *topos* was a 'place' in which to find words or subject matter for discourse or argument on a particular theme. The Latin equivalent of *topos* was *locus* (plural *loci*), which also means 'place'. The supreme Latin orator, Cicero, echoed the Greek conception by defining *loci* as 'the seats of an argument' ('argumenti sedes'), that is, places where material for discourse can be found (Cicero, *Topica*, 2, 8). So the idea that a 'topic' is a 'place' is as old as the classical tradition, and built deep into modern languages.

It was only in the sixteenth century, however, that this idea inadvertently provided the basis from which what might be regarded as conceptual gazetteers developed. The 'Renaissance' was devoted to the 'rebirth' of aspects of classical civilization. The central vehicle for this rebirth was the revival of classical languages, Greek and above all Latin, and the ideas and values communicated by them. In order to master classical literature, the culture contained in it, and the challenge of reading, writing, and speaking in classical Latin, scholars and educationalists, especially in northern Europe, developed techniques for processing the fruits of one's reading in the classics via books of *loci communes* or 'commonplaces'.[39]

The pioneer of Northern Renaissance humanism, the Frisian Rudolf Agricola (c. 1444–1485), advocated the use of two types of 'places' to his students: generic *sedes argumentorum* (such as definition, genre, type, etc.), relevant to all subjects for discussion; and more specific, subordinate terms (called *capita* or 'headings') under

[39] An excellent brief introduction is provided by Andreas Fuchs and Thorsten Fuchs, 'Loci communes', in *Brill's New Pauly*, Classical Tradition volumes, ed. Manfred Landfester, English edn. by Francis G. Gentry. See https://doi.org/10.1163/1574-9347_bnp_e1501750, accessed 20/03/2019. For a fuller treatment, see Ann Moss, *Printed Commonplace-Books and the Structuring of Renaissance Thought* (Oxford: Oxford University Press, 1996). Even larger in scale, but less apposite for present purposes is the three-volume series Joop Koopmans and Nils H. Petersenn, eds., *Commonplace Culture in Western Europe in the Early Modern Period,* 3 vols. (Leuven: Peeters, 2010–11). See also Wilhelm Schmidt-Biggemann, *Topica universalis: Eine Modellgeschichte humanistischer und barocker Wissenschaft* (Hamburg: Meiner, 1983).

which ancient knowledge could be systematically compiled. A generation later, the 'prince of the humanists', Erasmus of Rotterdam (c. 1469–1536), was advocating the use of 'places' as a means of mastering ancient language and literature: the student should first prepare a list of topical *loci* and then digest the fruits of his reading into these pre-prepared places to create a book of 'commonplaces' collecting together passages relating to the same theme or topic, which could then be drawn upon to provide *copia rerum et verborum*, an abundance of matter and of words with which to express it. By the time of Philip Melanchthon (1497–1560), the leading educational theorist of the Lutheran Reformation, the term *loci communes* no longer referred to the aphorisms and adages collected under these headings but to the *capita* or *tituli* under which such literary gleanings were collected. The leading Lutheran educationalist of the next generation, Johann Sturm (1507–1589), enthusiastically promoted a 'grand scheme for digesting the whole universe of Latin language into "domicilia" ['little houses'], "cellae" ['store rooms'], "sedes" ['seats'], or "receptacula" ['receptacles'], lodged within a "natural order" of "places" constituted by "res divinae", "res naturales", and "res humanae" [divine, natural, and human things]'.[40]

3.2 Hierarchies

As these collections grew, their headings multiplied, and the need arose for some means of reducing these proliferating headings to some kind of order. By this time, the emphasis of these collections had moved from words to things: as well as collecting classical words and phrases, in order to speak and write Latin like Cicero, educators like Melanchthon and Sturm used commonplace books to collect material as the basis for teaching all the disciplines of the academic curriculum. As a consequence, the *loci* became identified with the chief subjects of philosophical enquiry, and the *capita* became the 'chapters' treating the constituent parts of all the disciplines of the university curriculum. 'Do not think commonplaces are to be invented casually or arbitrarily', Melanchthon warned: 'they are derived from the deep structures of nature, they are the sets and patterns to which all things correspond'. The best way of organizing a commonplace book, in consequence, was 'to follow the divisions of the intellectual disciplines'.[41]

In the oral cultures of Antiquity and the Middle Ages, such structures of 'places' had been stored in the memory by arranging them spatially in imaginary places such as 'memory palaces'.[42] During the age of print, the spatial representation of these places was transferred to the printed page in diagrammatic representation of

[40] Moss, *Printed Commonplace Books*, 150.

[41] Moss, *Printed Commonplace Books*, 121, 129.

[42] The classic study is Frances Yates, *The Art of Memory* (London: Routledge & Kegan Paul, 1966, frequently reprinted); Jonathan D. Spence, *The Memory Palace of Matteo Ricci* (London: Faber, 1985); more recently: William N. West, *Theatres and Encyclopaedias in Early Modern Europe* (Cambridge: Cambridge University Press, 2002).

structures of knowledge.[43] Here too, a modern analogy immediately suggests itself. In modern computer science, ontologies and Linked Data are commonly presented in diagrammatic form.[44] The reason is obvious: clear diagrams, outlining networks of distinctions visually, are often more readily comprehended than complex sets of logical formulae or dense passages of prose. The advantages of visualization for readily apprehending the logical relationships between things was systematically exploited in the early modern period.[45]

The most persistent and widespread attempt to render the structure of learning visible was pursued within the pedagogical and philosophical tradition deriving from the Parisian humanist Petrus Ramus (1515–1572).[46] One ideal of this tradition was to structure the presentation of any academic discipline by means of dichotomies, that is, binary distinctions of concepts into pairs of contradictory terms. In an extreme case, the entire text of an elementary treatise was reduced in this manner to a continuous set of such bifurcating tables. More commonly, this technique was used to provide diagrammatic tables of contents, displaying at a glance not only the sequence of parts but also their logical relationships to one another.

The next stage was for scholars to begin publishing voluminous commonplace books which collected passages from canonical authors under topical headings. A former student of Ramus in Paris, Theodor Zwinger (1533–1588), devised a comprehensive set of bifurcating tables to create a unified structure within which to collect *loci classici* on any subject. Published in 1565 under a title redolent of the earlier tradition of topical memory – *Theatrum vitae humanae*, that is *The Theatre of Human Life* – this work had grown by 1604 through four subsequent editions into a work of five folio volumes of 4,376 pages.[47]

More characteristic still was the notion that the structure revealed in this way was the structure of the discipline itself, rather than merely one exposition of it. When combined with the assumptions underlying the commonplace tradition outlined above, the results were sets of dichtomous tables which purported to show precisely how each *locus*, place, or topic was situated within the structure of the discipline as a whole. When the whole 'circle of the disciplines' was brought to-

[43] Walter J. Ong, *Ramus, Method and the Decay of Dialogue* (Cambridge, MA: Harvard University Press, 1958, repr. Chicago, IL: University of Chicago Press, 2004).

[44] Aba-Sah Dadzie and Mathew Rowe, 'Approaches to Visualising Linked Data: A Survey', *Semantic Web* 2:2 (April 2011): 89–124, see https://doi.org/10.3233/SW-2011-0037.

[45] For a splendid overview, see Steffen Siegel, *Tabula: Figuren der Ordnung um 1600* (Berlin: Akademie Verlag, 2009).

[46] The foundational work is Ong, *Ramus, Method and the Decay of Dialogue*. On the influence, see Mrdechai Feingold, Joseph S. Freedman, and Wolfgang Rother, eds., *The Influence of Petrus Ramus: Studies in Sixteenth and Seventeenth Century Philosophy and Sciences* (Basle: Schwabe, 2001); Howard Hotson, *Commonplace Learning: Ramism and Its German Ramifications, 1543–1630* (Oxford: Oxford University Press, 2007); Steven John Reid and Emma Wilson, eds., *Ramus, Pedagogy and the Liberal Arts: Ramism in Britain and the Wider World* (Aldershot: Ashgate, 2011).

[47] Helmut Zedelmaier, 'Navigieren im Text-Universum: Theodor Zwingers *Theatrum vitae humanae*', *Metaphorik* 14 (2008): 113–35. See https://www.metaphorik.de/sites/www.metaphorik.de/files/journal-pdf/14_2008_zedelmaier.pdf, accessed 20/03/2019.

gether to constitute an 'encyclopaedia', individual *loci* were thereby situated within the entire universe of learning. The culminating work of this post-Ramist encyclopaedic tradition – the *Encyclopaedia* published by Johann Heinrich Alsted in 1630 – therefore provides a useful example of the manner in which such a work could be used to create a conceptual gazetteer.

3.3 Ontology

Before concluding with a look at Alsted's work, it is worth noting that the term 'ontology' emerged from the midst of this tradition. Until recently, the earliest known use of the Latin term *ontologia* was in a marginal note within a major work: the philospohical dictionary published in 1613 by the leading German Calvinist philosopher at the University of Marburg, Rodolph Goclenius the Elder (1547–1628).[48] An earlier usage has more recently been found in a far more obscure source: a mini-encyclopaedia which Jacob Lorhard (1561–1609) composed for use in the gymnasium in St Gallen, where he was rector.[49]

Lorhard and Goclenius were bound by many links which indentify the context in which ontology emerged. Both were educationalists in the German-speaking branch of the Reformed (or loosely 'Calvinist') tradition. Both were interested in developing encyclopaedic approaches to teaching the entire university curriculum of their day. Both were strongly influenced by the Ramist tradition.[50] Lorhard's book reduced the eight disciplines of his school's curriculum – Latin and Greek grammar, logic, rhetoric, astronomy, ethics, physics, and 'metaphysics or ontology' – to a continuous set of Ramist tables.[51] Both drew inspiration from other main works of this tradition.[52] The two men were almost certainly acquainted with one

[48] Goclenius, *Lexicon philosophicum quo tanquam clave philosophiae fores aperiuntur*, 1613 (repr.: Hildesheim: Georg Olms, 1980), 16: the term is written in Greek in a marginal note to the entry on 'abstractio materiae'.

[49] Lorhard's ontology is the focus of Peter Øhrstøm, Henrik Schärfe, and Sara L. Uckelman, 'Historical and Conceptual Foundations of Diagrammatical Ontology', in Uta Priss, ed., *Conceptual Structures: Knowledge Architectures for Smart Applications: 15th International Conference on Conceptual Structures, ICCS 2007, Sheffield, UK, July 22–7, 2007* (Berlin: Springer, 2007), 374–86, see https://doi.org/10.1007/978-3-540-73681-3_28. The usage was first noted in Joseph S. Freedman, *Deutsche Schulphilosophie im Reformationszeitalter (1500–1650)*, 2nd edn. (Münster: MAKS, 1985). For the broader tradition, see Piotr Jaroszyński, *Metaphysics or Ontology?*, trans. Hugh McDonald (Leiden: Brill, 2018), ch. 8: 'The Founders of Ontology: From Lorhard to Clauberg', see https://doi.org/10.1163/9789004359871_011.

[50] On the relevant branch of this tradition, see Howard Hotson, *Commonplace Learning.*

[51] Jacob Lorhard (1561–1609), *Ogdoas Scholastica, continens diagraphen typicam artium: grammatices latinae, grammatices graeca, logices, rhetorices, astronomices, ethices, physices, metaphysices, seu ontologiae. Ex praestantium huius temporis virorum lucubrationibus, pro doctrinae & virtutum studiosa juventute confecta* (St Gallen: Straub, 1606). The work was posthumously republished under a more typical and descriptive title: *Theatrum philosophicum: in quo artium ac disciplinarum philosophicarum plerarumq[ue] omnium […] praecepta, in perpetuis schematismis ac typis, tanquam in speculo, cognoscenda obiiciuntur* (Basle: Waldkirch, 1613).

[52] Lorhard's treatment of 'metaphysics or ontology' reduced to a set of Ramist tables the most innovative and influential metaphysical work produced in the German post-Ramist tradition, Clemens Timpler's *Metaphysicae systema methodicum*, first published in Steinfurt in 1604 and reprinted eight times in a number of German Reformed printing centres. Marco Lamanna, 'Correspondences between

another: in 1607, the year in which his encyclopaedia was published, Lorhard briefly assumed a theological professorship in Marburg, where Goclenius was the mainstay of the philosphical faculty.

For our purposes, these links are the more important because Goclenius demonstrably and Lorhard almost certainly influenced a student in Marburg who represents the culmination of the entire post-Ramist encyclopaedic tradition of Reformed Germany: Johann Heinrich Alsted (1588–1638).[53] Alsted also adopted the term *ontologia* and disseminated it in the work which brought the German Reformed post-Ramist encyclopaedic tradition to its culmination.[54]

3.4 Alsted's *Encyclopaedia* (1630) as Conceptual Gazetteer

Disciplines. In size, Alsted's masterpiece was one of the most voluminous works of an immense publishing tradition: its 5,000 folio columns of densely printed eight-point type contain well over 3 million words. In scope, it was the most comprehensive work produced by the international Ramist tradition. As its title indicates, the encyclopaedia is divided into seven tomes. These seven tomes are then subdivided into thirty-five books in the manner indicated in figure 1.[55] Most of these books treat a single discipline. At the core of the work (tomes II–V) are the disciplines relating to an expanded university curriculum: six 'philological' disciplines propaedeutic to the philosophical course (expanding on the medieval *trivium*); ten branches of theoretical philosophy (including the medieval *quadrivium* of mathematical disciplines as well as metaphysics and natural philosophy); four branches of practical philosophy; and the three higher faculties of the university curriculum (theology, law, and medicine). To these Alsted added four more innovative *praecognita* which lay the philosophical and pedagogical foundation of the entire work in the first tome, a pioneering treatment of the mechanical arts in the sixth, and in the seventh the thirty-seven *Farragines disciplinarum* or 'mixtures of disciplines' which escaped the topical framework of the body of the work. So, in essence, this work is structured around the standard medieval and early modern university curriculum,

Timpler's Work and that of Lorhard', see https://www.ontology.co/essays/correspondences-timpler-lorhard.pdf, accessed 20/03/2019, and his monograph, *La nascita dell'ontologia nella metafisica di Rudolph Göckel (1547–1628)* (Hildesheim: Georg Olms, 2013).

53 Alsted studied in Marburg in 1606–7. He acknowledged Goclenius as his philosophical teacher, who contributed an epigram to his *Encyclopaedia*. Howard Hotson, *Johann Heinrich Alsted, 1588–1638: Between Renaissance, Reformation and Universal Reform* (Oxford: Oxford University Press, 2000), 8, 12; ibid., *Commonplace Learning*, 27, 230–1. Lorhard's *Lysis duorum sophismatum pro omnipraesentia carnis Christi in Eius Persona* (Marburg, 1607) is dedicated to Hermann Vultejus and Gregor Schönfeld. See Marco Lamanna, 'Birth of a New Science: The History of Ontology from Suárez to Kant': recorded in https://www.ontology.co/history.htm, accessed 20/03/2019. Schönfeld was a professor of theology in Marburg who taught Alsted in these years; Vultejus was professor of law, chancellor at University of Marburg, and Alsted's blood relative.

54 Alsted, *Cursus philosophici encyclopaedia libris XXVII* (Herborn: Corvinus, 1620), vol. 1, 149. *Encyclopaedia, septem tomis distincta* (Herborn, 1630; facs. repr. Stuttgart: Frommann-Holzboog, 1989), 573 (citing Goclenius).

55 Derived from Alsted, *Encyclopaedia*, 1, 3, 1867.

expanded almost to bursting point by the new intellectual developments of the sixteenth centuries, and published just before the new philosophies of the seventeenth century throw this traditional organization into confusion.

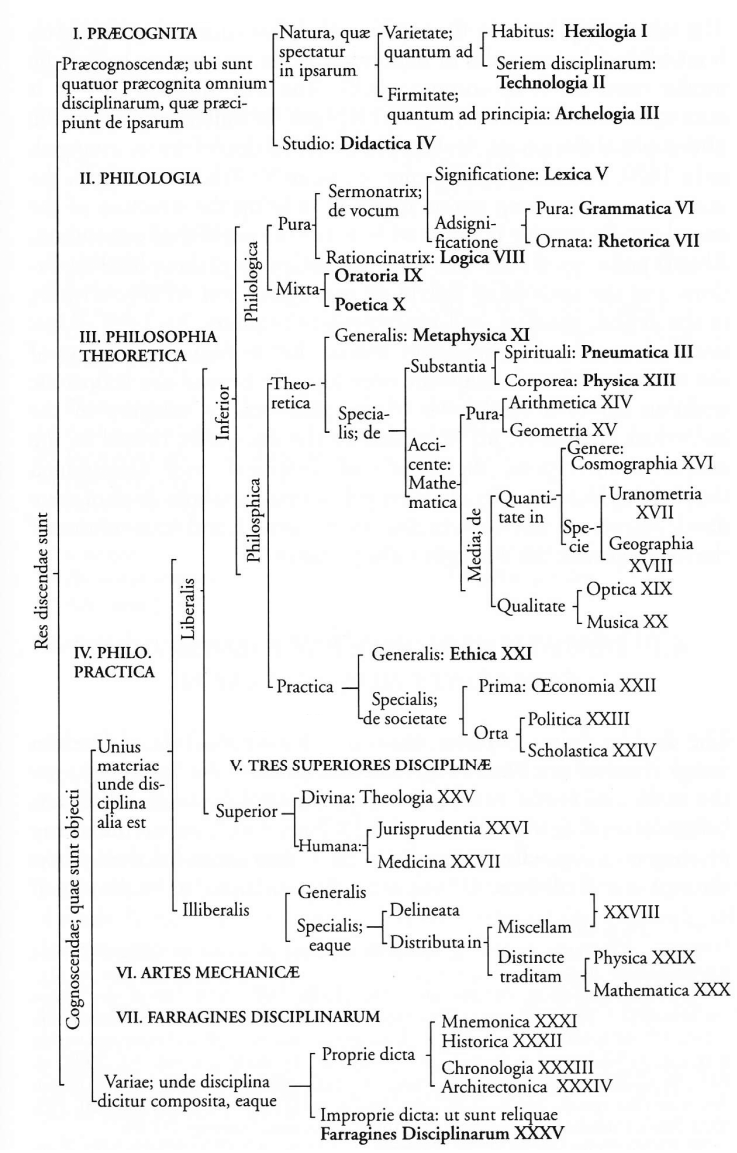


Figure 2: The structure of Alsted's *Encyclopaedia* (1630)[56]

[56] Derived from Alsted, *Encyclopaedia*, 1, 3, 1867. First published in Hotson, *Commonplace Learning*, 253.

Topics. The thirty-five books of Alsted's *Encyclopaedia* are subdivided into 115 parts; and these 115 parts are further divided into over 1,000 chapters. Each of these chapters is regarded as a 'topic' or 'commonplace', listed in the *index locorum communium* at the back of the work. The precise location of each of these 'places' within the logical structure of the work is indicated by the further set of bifurcating tables which are placed at the conclusion of each of the 115 parts. Figure 3 reproduces one of these tables: the table depicting the first part of physics – one of the largest books in the *Encyclopaedia*, which is broken down into eight parts.[57] The general part of physics is then further dvidied, by a series of mostly dichotomous distinctions, into fifteen chapters (indicated by Arabic numerals at the end of each branch of the table). When combined with the overall table reproduced in figure 2, the complete set of these subsidiary tables situate each of the 1,000 'places' (i.e. *loci*, *topoi*, topics, or chapters) within the structure of the *Encyclopaedia* as a whole.

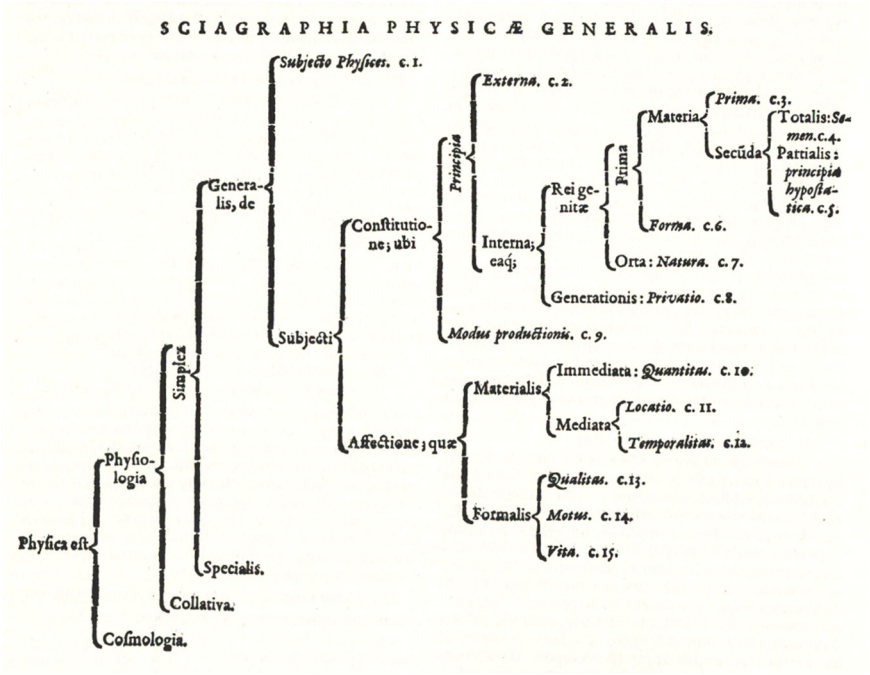
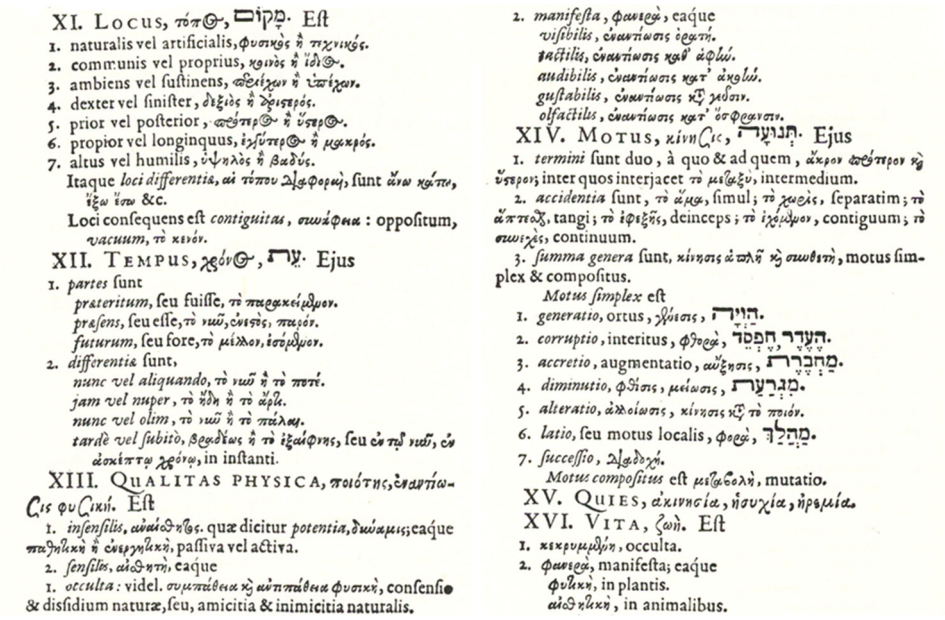


Figure 3: Table outlining 'general physics', from Alsted's *Encyclopaedia* (1630)

[57] Alsted, *Encyclopaedia*, 'Sciagraphia physicae generalis', 688. The roughly 1,000 chapters are handily tabulated in a concluding 'Index librorum, capitum, et locorum communium', *Encyclopaedia*, fols. hhhhh2v–6v, here 'Loci physici pars I', fol. hhhhh 4r.

Words. Each of these chapters is then further broken down into a thematically organized list of the technical terms employed within them. For this purpose, Alsted's *Encyclopaedia* also contains a tri-lingual lexicon or 'nomenclator' for each of the thirty-five books of the *Encyclopaedia*, providing terminology in all three of the ancient sacred languages: Latin, Greek, and (where appropriate) Hebrew. An extract from the nomenclator for the first part of physics is provided in figure 3.[58] The large Roman numerals in each heading correspond to the chapters of the first part of physics enumerated in figure 2. The Arabic numerals subordinate to each heading list the terms or pairs of terms associated with that topic. In some cases, these terms are further divided and subdivided. In the case of chapter 14, for instance, the general term *motus* ('motion') is divided into three categories (1. *termini*, 2. *accidentia*, and 3. *summa genera*); the third of these is divided in two (*motus simplex* and *compositus*), and the first of these subdivisions is further distinguished into seven different kinds (from *generatio* to *successio*).


XI. Locus, τόπος, מָקוֹם. Eſt
1. naturalis vel artificialis, φυσικὸς ἢ τεχνικός.
2. communis vel proprius, κοινὸς ἢ ἴδιος.
3. ambiens vel ſuſtinens, περιέχων ἢ ὑπέχον.
4. dexter vel ſiniſter, δεξιὸς ἢ ἀριστερός.
5. prior vel poſterior, πρότερος ἢ ὕστερος.
6. propior vel longinquus, ἐγγύτερος ἢ μακρός.
7. altus vel humilis, ὑψηλὸς ἢ βαθύς.
Itaque *loci differentia*, αἱ τόπου διαφοραὶ, ſunt ἄνω κάτω, ἔξω ἔσω &c.
Loci conſequens eſt *contiguitas*, συνάφεια : oppoſitum, *vacuum*, τὸ κενόν.

XII. Tempus, χρόνος, עֵת. Ejus
1. *partes* ſunt
praeteritum, ſeu fuiſſe, τὸ παρακείμενον.
praeſens, ſeu eſſe, τὸ νῦν, ἐνεργείᾳ, παιόν.
futurum, ſeu fore, τὸ μέλλον, ἐσόμενον.
2. *differentia* ſunt,
nunc vel aliquando, τὸ νῦν ἢ τὸ ποτέ.
jam vel nuper, τὸ ἤδη ἢ τὸ ἄρτι.
nunc vel olim, τὸ νῦν ἢ τὸ πάλαι.
tardè vel ſubitò, βραδέως ἢ τὸ ἐξαίφνης, ſeu ἐν τῷ νῦν, ἐν ἀσκέπτῳ χρόνῳ, in inſtanti.

XIII. Qualitas physica, ποιότης, ἐνεργείαι φυσικαί. Eſt
1. *inſenſilis*, ἀναίσθητος. quæ dicitur *potentia*, δύναμις; eaque παθητικὴ ἢ ἐνεργητικὴ, paſſiva vel activa.
2. *ſenſilis*, αἰσθητὴ, eaque
1. *occulta* : videl. συμπάθεια καὶ ἀντιπάθεια φυσικὴ, conſenſio & diſſidium naturæ, ſeu, amicitia & inimicitia naturalis.
2. *manifeſta*, φανερὰ, eaque
viſibilis, ἐνεργείαι ὀρατή.
tactilis, ἐνεργείαι καθ' ἁφήν.
audibilis, ἐνεργείαι κατ' ἀκοήν.
guſtabilis, ἐνεργείαι καὶ χύσιν.
olfactilis, ἐνεργείαι κατ' ὄσφρησιν.

XIV. Motus, κίνησις, תְּנוּעָה. Ejus
1. *termini* ſunt duo, à quo & ad quem, ἄκρον ἀφ' ὅτου καὶ ὕστερον; inter quos interjacet τὸ μεταξύ, intermedium.
2. *accidentia* ſunt, τὸ ἅμα, ſimul; τὸ χωρὶς, ſeparatim; τὸ ἅπτεσθαι, tangi; τὸ ἐφεξῆς, deinceps; τὸ ἐχόμενον, contiguum; τὸ συνεχὲς, continuum.
3. *ſumma genera* ſunt, κίνησις ἀπλῆ καὶ συνθετὴ, motus ſimplex & compoſitus.
Motus ſimplex eſt
1. *generatio*, ortus, γένεσις, הֶוְיֵה.
2. *corruptio*, interitus, φθορὰ, חֶפֶד הֶעְדֵּר.
3. *accretio*, augmentatio, αὔξησις, מַחְבֶּרֶת.
4. *diminutio*, φθίσις, μείωσις, מִגְרַעַת.
5. *alteratio*, ἀλλοίωσις, κίνησις κατὰ τὸ ποιόν.
6. *latio*, ſeu motus localis, φορὰ, מַהֲלָךְ.
7. *ſucceſſio*, διαδοχή.
Motus compoſitus eſt μεταβολὴ, mutatio.

XV. Quies, ἀκινησία, ἡσυχία, ἠρεμία.

XVI. Vita, ζωὴ. Eſt
1. κεκρυμμένη, occulta.
2. φανερὰ, manifeſta; eaque
φυτικὴ, in plantis.
αἰσθητικὴ, in animalibus.

Figure 4: The 'nomenclator' of general physics, chapters 11–16, from Alsted's *Encyclopaedia* (1630)

It is worthwhile enumerating the levels of this hierarchy. At the bottom of the hierarchy (level 1) we have tens of thousands of individual terms (such as *generatio* and *successio*), some of which are grouped in subordinate clusters (level 2, in this case *motus simplex*) and superior groupings (level 3, here *summa genera*) within the

[58] Alsted, *Encyclopaedia*, 'Nomenclator physicae', 242.

individual *loci,* topics or chapters (level 4, here *motus*). All told, twenty terms are provided for discussing the topic of 'motion'. Motion is one of fifteen 'places' within *Physica generalis* (level 5), which is one of the eight parts of physics (level 6). Physics, in turn, is one of the ten disciplines of theoretical philosophy (level 7), which together make up one of the seven tomes of the *Encyclopaedia*. A glance at figure 2 reveals, however, a further eight levels between the discipline of physics and the encyclopaedia as a whole.[59] So Alsted's *Encyclopaedia* was erected on a knowledge hierarchy created by mostly binary distinctions and consisting of at least fifteen levels.

Assembling the complete set of these Ramist tables would demonstrate the relationship of every one of tens of thousands of tri-lingual terms within the 1,000 'commonplaces' or chapters of the 115 parts of the thirty-five books of the seven tomes of the work as a whole. Together, they offer a standardized vocabulary, organized according to the main topics in each discipline, rearranged in hierarchical order as subdivisions of the entire 'encyclopaedia' or 'circle of the disciplines' as a whole. Like the semantically organized vocabularies created in modern lexicography, linguistics, and terminology, the order of the words in these nomenclators is not alphabetical. Instead, the order is 'topical', deriving from the order of the subject matter in the body of the treatise, that is from the order of the *loci communes* themselves. Moreover, like the ontologies created to organize data on the Semantic Web, this structure is hierarchical: each term in Alsted's nomenclator is subordinated to more general terms which define the basic topics, and these topics are situated by means of a net of definitions and distinctions within the overall edifice of the encyclopaedia.

Perhaps most significant of all, there is nothing fanciful in describing this entire structure as a conceptual gazetteer. The root metaphor of the entire work is to regard each 'topic' as a *locus* or 'place', with a fixed location within a comprehensive structure. The fundamental purpose of the work is to allow students to navigate throughout the whole *orbis doctrinae* or 'world of learning', not via two-dimensional geographical coordinates, but with reference to a formal hierarchy of categories structured by a continuous net of dichotomous logical distinctions.

It is tempting to go one step further and suggest that this structure fulfils the conditions of a formal ontology outlined above. It is (1) formal (manifesting a well-defined syntax and semantics); (2) explicit (capable of being represented in a set of dichotomous tables); (3) shared (both derived from and providing the basis of

[59] Physics (which treats of body) and pneumatics (which treats of spirit or soul) make up the substantial part of theoretical philosophy (level 8). Together with the mathematical sciences (which study one of the accidents of matter, namely extension), these constitute the special part of theoretical philosophy (level 9), to which must be added metaphysics (here regarded as the general part) to make theoretical philosophy as a whole (level 10). Theoretical and practical are the two parts of philosophy (level 11) which, with philology, make up the inferior part (level 12) of the liberal disciplines (level 13). The liberal and mechanical arts constitute the two main kinds of disciplines (level 14), which are the main species of things to be known (level 15), to which must be added Alsted's novel cluster of praecognita to make up the entire encyclopaedia (level 16).

formal educational provision in a network of schools and universities); and (4) regarded as a conceptualization which presents a model of the real world.

3.5 The Need for Multiple, Parallel Hierarchies

Of these four conditions, it is the third which is the most doubtful. Alsted's *Encyclopaedia* occupies a central place in the history of its genre. The last of a long line of topically organized, pedagogically orientated, Latin encyclopaedias, it is the culmination of a long medieval and Renaissance tradition. But that is not to say that this work represented the philosophical or pedagogical consensus of the sixteenth or seventeenth century. On the contrary, by the time the work was published in 1630, that consensus was fast disintigrating.

Alsted was a member of the generation midway between Copernicus and Newton. Deeply conscious of the limitations of the received world view, and eager to replace it with something better, he and his generation were not yet in possession of the materials with which to fashion a comprehensive, authoritative, durable alternative on firm foundations. So he organized his encyclopaedia for the most part along conventional lines, but sought to correct some of the defects of traditional learning at the level of individual disciplines, part of disciplines, topics and terms. The result was a compromise which pleased neither the more traditional educationalists of his day nor the more adventurous innovators. For generations, figures of the stature of Comenius (1592–1670) and Leibniz (1646–1716) proposed to replace his encyclopaedia with something more up to date. But all failed, and in consequence of their failure his work remains one of the most valuable gazetteers for the conceptual geography of this transitional period of European intellectual history.[60]

The point is that, just as no single modern ontology can capture the world view of the early modern republic of letters, no early modern ontology – however finely ramified or influential – represents more than one alternative available to early modern intellectuals. At the very least, the structure developed by Alsted would need to be complemented by others. For starters, a more conventional encyclopaedic scheme is needed, such as that of the highly successful *Margarita philosophica* first published by Gregor Reisch in 1503. For the later period, the obvious choice would be the greatest book of the eighteenth century, the famous *Encyclopédie* created by d'Alembert and Diderot as a kind of summa of the learning of the Enlightenment. These works also contained tables outlining the structure of knowledge, and if further examples are needed they can be found in this period in great profusion. The key step would then be to map the place of individual words, terms, concepts, and topics from one of these structures to another to provide a

[60] This process is the subject of Howard Hotson, *The Reformation of Common Learning: Post-Ramist Method and the Reception of the New Philosophy, 1618–1670* (Oxford: Oxford University Press, forthcoming).

tool capable of sorting, faceting, and querying data by means of multiple competing ontologies.[61]

The effort required would not be trivial: this exercise would represent a major research project in its own right. But its difficulty should also not be overstated. In creating a standard modern ontology, a huge amount of conceptual work is required to define terms, calibrate distinctions, and determine relationships. A huge amount of promotion and negotiation is then required to gain the acceptance of the user community. In the case of these early modern 'ontologies', the conceptual work has already been done by the traditions that they encapsulate; and precisely insofar as they are traditional their acceptance by the early modern user community is also a *fait accompli*. While implementing them would be an ambitious enterprise, it would also be far easier than devising and gaining acceptance for a new ontology.

The result would be significant in its own right: a tool for exploring the evolution of knowledge organization systems in a high level of detail. But the chief return on the investment would be practical: the creation of a far more adequate basis for creating keywords, for relating them hierarchically, searching, sorting, faceting, browsing, and analysing at any level of generality from the individual terms at the bottom of the hierarchy via topics, subdisciplines, disciplines, and clusters of disciplines to the entire encyclopaedia. Although far too laborious to construct for any single research project in this field, such tools could play an invaluable role over the long term as part of a distributed infrastructure shared by interlinked research teams throughout the world (on which see chapter III.5 below).

[61] A similar proposal is outlined in Stefan Heßbrüggen-Walter, 'What People Said: The Theoretical Foundations of a Minimal Doxographical Ontology and Its Use in the History of Philosophy', in Constanze Baum and Thomas Stäcker, eds., *Grenzen und Möglichkeiten der Digital Humanities* [special issue of the *Zeitschrift für digitale Geisteswissenschaften* 1] (Wolfenbüttel: Forschungsverbund Marbach Weimar Wolfenbüttel, 2015), see https://doi.org/10.17175/sb001_001.

II.6 Events

Neil Jefferies with Gertjan Filarski and Thomas Stäcker

1 Introduction

Looking over the previous sections, it becomes evident that almost all aspects in the domain of interest to the COST action have strongly time-dependent elements:

Spatially, as discussed in II.2, the geopolitical landscape of Europe during the period is particularly fluid as a consequence of conflict, dynastic developments, and political manoeuvering.[1] Understanding the broader political, ecclesiastical, and other locational contexts in which a letter is written and read can thus be crucially dependent on dates.

Temporally, as discussed in II.3, calendars were subject to wide variation in use, depending on date and location. Awareness of these changes is therefore essential for mapping of dates-as-marked to a consistent (modern) calendar in order to sequence events and establish causal relationships.

Letters, as distinct from other literary artefacts, are defined by the acts of sending and receipt and as a result letter corpora are necessarily time-ordered. This is described in more detail in chapter II.1.

People, as discussed in II.4, are naturally the most dynamic entities involved in these networks, requiring potentially innumerable events to model accurately.

[1] 'Interactive World History Atlas since 3000 BC', *GeaCron*, see http://geacron.com/home-en/, accessed 20/03/2019.

1.1 Events as Distinct Entities

As such, within all the entities (people, places, letters, calendars, etc.) that are represented in our data model, there is the need to have a consistent expression of events to capture this temporality. Events, however, establish relationships between multiple entities rather than being an attribute belonging to any one, so an independent construction makes sense. Specifically, an event locates a participant at a particular time and place along with a reference to an object such as a letter that is either involved in the event (e.g. being written), or, if not involved, provides the evidence for the event (e.g. mentioning it later).

Indeed, it can be argued that time provides the key structuring and descriptive element of the collections that we seek to aggregate. As historians, we are interested in the construction of narratives which both contextualize, and permit a better understanding of, the underlying evidential materials. This chapter considers a number of existing general approaches to events in the light of the characteristics of the material discussed in earlier chapters and presents some specific pragmatic recommendations.

2 Modelling Considerations

In terms of constructing a suitable event model there are a few factors that need a little expansion to provide some more focus to the exercise.

2.1 Scope of Event Model

Primarily, the model is concerned with events that either change an entity, change its context, or, in the case of documents, are depicted within their content. For certain events central to the republic of letters – such as sending or receiving a letter or publishing a book – detailed, standardized ontologies are needed to ensure interoperability between corpora. However, other events, such as those depicted in letters, can vary so widely that there is little to be gained by such a granular treatment, and broader classification accompanied by textual descriptions may prove sufficient. In some cases, it may be possible and prudent to reuse existing ontologies from other domains.

2.2 Capturing Provenance and Context

When attempting to construct a distributed digital representation of the republic of letters, the need to capture and represent provenance emerges at a number of levels. Most obviously, it is useful to consider the onward journey of letters from the initial recipient (whether intended or not) into subsequent collections and archives. In this case, it is evident that a model that is already based on the notion of a letter being transmitted from one party to another can be readily extended to cover ac-

cession and aggregation events involving further parties that bring the letter into new and different contexts.

Including these types of event in the model, along with related entities such as collectors and collections, provides a mechanism for capturing additional contextual information that is useful for understanding the nature of the current distributed corpus. Examples could include material concerning the authenticity of individual letters, or documentation that might provide an indication of the completeness or selectiveness of a particular collection. To generalize, in terms of an event-based model, provenance can be viewed as the series of contexts in which an entity exists, the events that describe the transitions between them, and the evidence for those events.

2.3 Digital Provenance and Sources

Besides the traditional archival notions of provenance concerning physical artefacts, a distributed system also needs to address provenance in the digital sense, which differs from the physical case in several significant ways:

- Software can transform digital objects in significant ways without necessarily affecting their intellectual content. For example, changing an image file format from TIFF to JPEG2000 can result in every single byte of a file being changed, even though the underlying image that it represents is identical to the original. As such, when dealing with digital surrogates such as images and transcriptions arising from the digitization process, it is useful to capture both human- and machine-mediated events. The idea of provenance thus needs to be extended to cover both the context and state of a digital object.
- Digital objects can be replicated without loss between collections. Once that has occurred, subsequent actions, such as the types of transformation mentioned above, can cause the two objects to diverge in content. This divergence can also occur in the form of differing metadata developing over time as a result of scholarly or library activity. In order to understand correctly the derivation of the content of an object, it is thus necessary to know the source of the original copy, the time that the copy was made, and any subsequent changes.
- Digital objects can be disaggregated and the parts reassembled in different ways. Indeed, the ability to split metadata files into individual assertions that can be combined and compared with similar assertions from other sources is essential for interoperating between the different systems that comprise the distributed corpus of the republic of letters. In order to allow information to be fed back to the original sources it is therefore sometimes necessary to have source and change information at the level of individual assertions.

The net result is that, in the digital world, we have the ability and need to capture provenance information at a much more granular level. However, this increased level of detail must be managed carefully since the increased data volumes can lead to inefficiencies for machines and confusion for human users.

2.4 Graph Representations

While events contextualize letters, letters themselves document many events in turn. In a similar manner, as alluded to earlier, the representations of time and place cannot be entirely disentangled. Such interrelationships are not effectively described if we adopt a hierarchical library catalogue model that is centred on letters. A more suitable approach would be a graph-like representation[2] with letters, agents, events, and locations as nodes. While each of these could be represented as tables within a relational database (as is the case with EMLO[3] currently) experience has shown that the number and type of relationships between entities increases rapidly as the corpus grows to include new material and scholarship. Accommodating this evolution within a relational database becomes cumbersome and the resulting queries hard to understand or maintain. True graph representations such as RDF (Resource Description Framework[4]) and the related tooling provide a much more flexible and scalable approach for this type of material – treating each entity as a distinct node with the ability to have arbitrary typed relationships with any other node.

2.5 Notions of Agency

The model for an event that is emerging is characterized by the concept of agency: the entity or entities that actively instigate an event, and secondary entities that influence or delegate their involvement in the event. While people and organizations are the primary agents of interest at the outset, the emergence of digital tools and automation means that software will emerge as a key agent when considering the provenance of metadata.

Closely related to agency is the concept of an agent's role, both in the context of a particular event and also within a broader organizational context (formal or otherwise), which can provide additional subtlety to event descriptions. For instance, when a letter is sent, we can distinguish the author of a letter from a signatory organization and express the relationship of the author to that organization. Without roles, authorship, signing, and sending would have to be represented as

[2] 'Graph Data Modeling 101', *Cambridge Intelligence*, see https://cambridge-intelligence.com/graph-data-modeling-101/, accessed 20/03/2019.

[3] 'Early Modern Letters Online : Home', see http://emlo.bodleian.ox.ac.uk/, accessed 20/03/2019.

[4] 'RDF - Semantic Web Standards', see https://www.w3.org/RDF/, accessed 20/03/2019.

separate events, greatly adding complexity, and constructing a formal ontology that adequately captures the meaning of organizational membership roles is unlikely.

2.6 Bounded Entities

Adopting an event-based model brings to the foreground two key characteristics of entities that are often only partially addressed in other systems. First and foremost, the fact that almost all entities are bounded in terms of time and space: they have a beginning, a finite lifespan, and an end and exist at particular locations during that time. This is evident for people, but less so for places and letters, for example. This also applies even to abstract entities such as classifications: as discussed in chapter II.5, major disciplinary categories like 'philosophy' and 'art' change shape profoundly in the course of the early modern period.

This perspective also makes clear that many of the attributes of an entity are the result of agency and thus apply only for a portion of the total existence of the entity. Recognizing the existence of these bounds, making them explicit and using them in combination with relatively simple inference logic can be useful both for checking the sanity of assertions but also for aiding identification and deduplication efforts by helping identify probable matches based on shared bounds.

2.7 Periods

While events are a useful construct, they are not sufficient to represent completely bounded entities and attributes. *Periods* define spatio-temporal intervals for which an assertion is valid. These assertions may be as fundamental as existence (a person's lifespan) or as simple as a label ('The Middle Ages'). Periods can, but do not need to, include a spatial element based on the observation that many historical 'periods' need some form of spatial qualification: for example, the Second World War began and ended at different times in different countries.

Periods can, in many ways, be regarded as analogous to the geopolitical entities described in chapter II.2 in that they describe geographically dependent time intervals, as opposed to time-dependent geographic boundaries. To continue this analogy, periods can include other periods and events in a manner that can be useful for validation and inference: for example, a person's presence at various locations necessarily takes place during their lifetime (except for final interment). As with places, a period may be part of multiple such hierarchies: the Bohemian Revolt (1618–20), for instance, is generally regarded as the first, brief phase of the Thirty Years' War (1618–48). Periods can be broadly classified according to the way that they are defined:

- *Explicit periods* are defined by bounding events. While there may be disagreement about the precise timing and location of the bounding events, the fact that they bound the period is not debatable. Examples would include the reign of a monarch bounded by their coronation and death, or someone's

presence at particular location bounded by their arrival and departure. Explicit periods can thus be derived from the existence of their bounding events.

- *Implicit periods* are less clearly defined, and may even have several disputed definitions so that a level of uncertainty is itself a defining characteristic. Instead, assertions can be made that events or periods fall within a named implicit period, and the period is consequently defined by the aggregation of these assertions and thus depends on a somewhat subjective assessment of the validity of these assertions. In general, such periods end up acquiring a 'core', defined by assertions about which there is little argument, and a 'fuzzy' boundary of varying size which is subject to more interpretation. A good example of this would be the time-span covering the 'republic of letters'. This term is generally referred to as an historical phenomenon of finite historical duration, that is to say, a phenomenon which did not always exist, which came into existence at some time, and then ceased to exist at another. But since the 'republic of letters' is (as suggested in chapter I.2) an 'imagined community', it is naturally not easy to say precisely when it began and ended. It should, however, be possible to arrive by consensus at an earliest date prior to which the 'republic' definitely does not exist and a subsequent date by which it definitely does exist. Together these dates would define an interval covering its 'coming into existence' which can be used to frame a debate. A similar process applied to its ending would yield a latest date by which it is definitely in existence and an earliest date by which it has definitely ceased to exist.

3 Existing Approaches Embodying Events

The relevance of an events-based model is not unique to the use case presented here. Rather, it can be seen as an emergent characteristic that begins to appear in many scenarios as a result not just of improved data collection and analytical tooling, but also of an increasing emphasis on workflows and reproducibility that introduces such concepts into contemporary research activity. In order to avoid unnecessary reinvention, it is prudent to look at existing standards in the library and digital humanities space to identify opportunities for reuse.

3.1 CIDOC CRM (Conceptual Reference Model)

CIDOC CRM[5] hails from the museum/cultural heritage domain (but not, historically, libraries) and at the outset emphasized the importance of context and its representation in terms of events. However, as a reference model, it is a rather abstract and unwieldy framework in practice, with most implementations taking

[5] 'Home | CIDOC CRM', see http://www.cidoc-crm.org/, accessed 20/03/2019.

shortcuts or only including subsets of the standard. Experience gained during the ongoing OXLOD[6] project suggests that almost any data model can be mapped to CIDOC CRM, but at the expense of increased complexity and somewhat reduced transparency. As such, it can be useful for unifying diverse corpora but is less suitable for operational information storage.

3.2 Schema.org

Schema.org[7] is an open standard originated and still largely supported by leading search engines for contextualizing entities on the web and thereby improving their discoverability. OCLC (the Online Computer Library Center[8]) is a strong advocate and has enhanced WorldCat[9] to support schema.org. It is implemented as additional semantic markup within regular HTML and consequently imposes no requirements as far as underlying data storage formats, although RDF would, of course, be the simplest to implement. Schema.org includes representations of people, places, and documents. Although the initial version did not include events, they were added in subsequent revisions. Thus, much of our data can easily be exposed via schema.org online, and this practice should be recommended. This is explored in greater detail in chapter III.5.

3.3 BIBFRAME

BIBFRAME[10] is an emerging linked-data library cataloguing standard led by the Library of Congress, with a view to replacing the various MARC[11] standard cataloguing variants in current use. Events were introduced in version 2.0 which, among other changes, represents a significant shift away from a basically literal translation of MARC into RDF that characterized the initial iteration. At this stage, events are somewhat limited in scope to those depicted in works and those that embody works such as performances. A complex formal model including agency is not yet fully developed. However, BIBFRAME is a component of the ongoing Linked Data for Libraries[12] initiative looking at the potential for linked-data use in

6 'Oxford Linked Open Data Pilot | Gardens, Libraries & Museums', see https://www.glam.ox.ac.uk/oxford-linked-open-data-pilot, accessed 20/03/2019.

7 'Home - Schema.Org', see http://schema.org/, accessed 20/03/2019.

8 'OCLC Home', see https://www.oclc.org/en/home.html, accessed 20/03/2019.

9 'OCLC Adds Linked Data to WorldCat.Org', see https://www.oclc.org/en/news/releases/2012/201238.html, accessed 20/03/2019.

10 'BIBFRAME Model, Vocabulary, Guidelines, Examples, Analyses (BIBFRAME – Bibliographic Framework Initiative, Library of Congress)', see https://www.loc.gov/bibframe/docs/index.html, accessed 20/03/2019.

11 'MARC 21 to BIBFRAME 2.0 Conversion Specifications (BIBFRAME - Bibliographic Framework Initiative, Library of Congress)', see https://www.loc.gov/bibframe/mtbf/, accessed 20/03/2019.

12 'Home' *LD4L*, see https://www.ld4l.org/, accessed 20/03/2019.

libraries more broadly and more event-centric approaches are emerging in related projects.

4 Specific Semantic Event Models

In addition to the event models embedded in the broader frameworks mentioned above, there are a couple of essentially stand-alone models that can be reused in larger frameworks that bear further examination.

4.1 Simple Event Model

The Simple Event Model[13] (SEM) was developed[14] by the Natural Language Processing (NLP) community partly in response to the perceived complexity of approaches such as CIDOC CRM on one hand, and overly simple or domain-specific models on the other. The aim was to be rich enough to model all the types of events that can be extracted from texts via linguistic analysis, while being simple enough to be usefully implementable. A number of projects and initiatives have subsequently used the model successfully, but the model itself did not progress to become a formal standard, although it can be considered a de facto one. Although we are considering a much broader range of events, and for uses other than just analytics, the SEM nevertheless aligns quite well with many of our use cases. A major simplifying aspect of SEM compared to CIDOC CRM is the concept of roles that indicate how agents participate in an event – which is complex and inconsistent to encode in CIDOC CRM.[15]

4.2 W3C PROV-O (Provenance Ontology)

W3C PROV-O[16] is a defined standard[17] which uses slightly different terminology to SEM but which is largely congruent and compatible with it. As a consequence, the adoption of SEM does not preclude a subsequent move towards PROV-O.

Although provenance might seem to be a relatively narrow class of events, in practice, almost any event can be encoded using the same basic pattern (shown in Figure 1), and almost all events are described precisely because they do provide

---

13 'The Simple Event Model', see http://semanticweb.cs.vu.nl/2009/11/sem/, accessed 20/03/2019.

14 Willem Robert van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber, 'Design and Use of the Simple Event Model (SEM)', *Web Semantics: Science, Services and Agents on the World Wide Web* 9:2 (1 July 2011): 128–36, see https://doi.org/10.1016/j.websem.2011.03.003.

15 'How to model Roles in the CIDOC-CRM RDF encoding', see www.cidoc-crm.org/sites/default/files/Roles.pdf, accessed 20/03/2019.

16 Luc Moreau, Paul Groth, James Cheney, Timothy Lebo, and Simon Miles, 'The Rationale of PROV', *Web Semantics: Science, Services and Agents on the World Wide Web* 35 (1 December 2015): 235–57, see https://doi.org/10.1016/j.websem.2015.04.001.

17 'PROV-O: The PROV Ontology', see https://www.w3.org/TR/prov-o/, accessed 20/03/2019.

context and/or provenance to the entities to which they relate. In addition to roles, PROV-O does introduce two other useful concepts: delegation, allowing one agent to act on behalf of another; and influences, which capture 'secondary agents' that are not necessarily direct participants in an event. These are very useful constructs for effectively representing humanistic narratives in a formal data model.

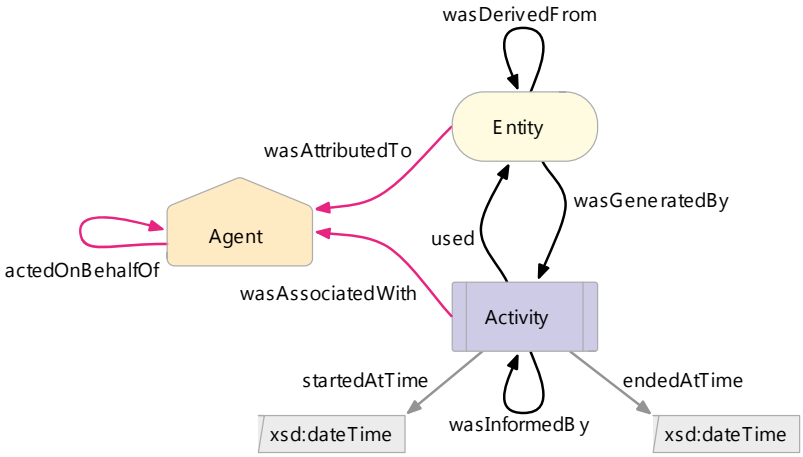


Figure 1: Basic Elements of the PROV-O Ontology[18]

5 Expressing Uncertainty

One key requirement for any system that aims to represent accurately data and metadata derived from historical sources is the need to capture and express uncertainty. This is needed so that scholars can effectively evaluate these inferences. At this point, it is useful to distinguish uncertainty, where there is unclear or conflicting evidence, and vagueness, where there is clear evidence but it has an intrinsic lack of specificity. For example, the date of sending of an undated letter is uncertain but that of a letter inscribed 'Easter 1768' is vague.

An initial step towards enabling effective evaluation is to capture the provenance of each assertion, in terms of who made it, when, and on what basis (for example, a link to an evidentiary source). The basis of the assertion is the crucial factor that distinguishes vagueness from uncertainty. By qualifying information in this way, differences of scholarly opinion with multiple contradictory assertions can be accommodated without causing confusion to other users of the material. While it is possible to use PROV-O, as described in the previous section, the Web

[18] 'PROV-O: The PROV Ontology', see https://www.w3.org/TR/prov-o/, accessed 20/03/2019.

Annotation Data Model[19] provides a simpler solution which can be mapped programmatically to PROV-O, or other annotation formats such as NanoPub,[20] as required.

Some systems attempt to compute or assign numerical measures of certainty to assertions, but this is only meaningful when data sources have well-defined statistical characterizations, which is simply not the case for a broad selection of historical sources. The general application of such overall measures thus, at best, adds relatively little value and, at worst, can be a distraction and a potential source of erroneous deductions. Such measures, essential for the use of many analytical and visualization tools, *can* be computed more reliably across a selection of material chosen for greater consistency, probably along with certain simplifying assumptions on the part of the research (for an example, see ch. II.3, section 2.3). It is therefore better to make sources and uncertainty explicit and allow the scholar to make these informed decisions. Assertions based on vague statements are, by definition, certain unless they attempt greater precision than the vagueness allows.

5.1 Spatial Uncertainty

The earlier chapter on modelling space (II.2) already discussed mechanisms for capturing the temporal variation of places within multiple broader geopolitical hierarchies. As this already handles the definition of regions (as opposed to points on a map), spatial uncertainty can be readily accommodated by referencing appropriate regions when indicating locations. More nuanced variations – for example, locations with the same name – can be expressed using multiple qualified assertions, as described in the previous section.

5.2 Temporal Uncertainty

Temporal uncertainty is rather more difficult to accommodate within existing models. Most models allow us to specify start and end dates/times for an event. A basic, and straightforward, extension of this is to allow the start and end points to be intervals in their own right, defined by:

lower bound: earliest date/time that the event *could* have started/ended

upper bound: earliest date/time by which the event *must* have started/ended.

Both of these assertions may well need to be qualified in some manner because they will probably be determined by external events or sources.

However, as discussed in the chapter on modelling time (II.3), the nature of the materials of the early modern period presents additional types of temporal uncertainty that elude such simplistic modelling. For example, it is possible to

[19] 'Web Annotation Data Model', see https://www.w3.org/TR/annotation-model/, accessed 20/03/2019.

[20] 'Nanopub.Org', see http://nanopub.org/wordpress/, accessed 20/03/2019.

know the date and month of a letter but be uncertain about the year as a result of calendrical changes. Fortunately, Extended Date Time Format (EDTF[21]) modifies the ISO 8601–2004[22] standard time format in common use to include such uncertainty in individual components of a date-time reference along with a broader range of temporal references such as seasons, quarters, etc. At the time of writing, the ISO8601-2018 update has confirmed the inclusion of EDTF functionality, but is not yet released. Along with the use of bounding intervals detailed above, it is probable that the majority of temporal issues can be successfully captured.

6 Implications of an Event-based Approach

The logic of building an historical data model around events as discrete entities is clear and compelling; but this event-based approach also requires a shift in mindset and results in increased complexity; so an adequate justification of this approach requires a concluding discussion of its potential benefits for users. These apply to both how data is captured and used but also, subsequently, how the corpus is enriched and developed through ongoing scholarship.

6.1 Narrative-based Discovery

Modelling events as distinct entities both simplifies and improves the discovery and navigation of a corpus through several mechanisms.

Viewpoint Agnosticism: Any reasonable database system will allow users to search by the content of metadata records: finding people and places by name, and letters based on textual content. However, browse and faceted views based on people and places become much richer as a result of indexing events rather than letters through the inclusion of important contextual information such as biographical and geopolitical events in the search results. Such views therefore display all the information that is known about a person or place rather than just the relevant letter records. For example, selecting a person could display biographical events including itinerary details and relationships to various organizations alongside their letters – providing immediate contextualizing information. In doing so, it transforms a catalogue viewpoint into a knowledge system.

Temporal Indexing: As a result of a single consistent representation of events, time becomes a viable dimension for discovery. In practice, this is likely be combined with a person and place reference and displayed using a timeline or calendar form to give a discovery mechanism that aligns much more closely with the historical narratives that potentially drive such queries.

---

[21] 'Extended Date Time Format (Library of Congress)', see https://www.loc.gov/standards/datetime/, accessed 20/03/2019.

[22] 'ISO 8601 Date and Time Format', see https://www.iso.org/iso-8601-date-and-time-format.html, accessed 20/03/2019.

Narrative-based Discovery: The addition of temporal parameters (and the use of inference) permits the construction of much richer and more natural queries. These allow the discovery of entities through their relevance to a broader stated narrative: 'Where was Comenius in 1612 and who was he likely to have encountered as a consequence?'

Simplified Data Model: It is possible to produce similar search results by querying traditional person, place, and letter records individually and merging the results. However, if such records aim to capture the same richness as event streams, they will need to have multiple date and place fields within each record (e.g. birth date and place, matriculation date and place, death date and place, etc.). As new events are added, database table structures and the related queries will need to be updated, which can quickly become cumbersome and error-prone. An event-based model eliminates this requirement.

Visualization and Analytics: In many cases, visualization and analytics combine letters with one or more dimensions of people, place, or time. These are greatly facilitated by the simplified querying that an event-based approach allows.

6.2 Enhanced Discourse

Adopting an event-based model along with mechanisms for qualifying assertions permits a much more holistic approach to constructing a knowledge system for the republic of letters (or, indeed, in general). The model recognizes that much of the meaning of a letter or other artefact depends on the broader intellectual and geo-political context in which it arises and exists, and therefore aims to capture and express this fact.

A ready extension of this paradigm is to realize that the current evolving intellectual discourse is merely an extension of this extant provenance, is thus equally relevant to the understanding of these materials, and can be captured and discovered using much the same mechanisms. Disputed interpretations can be represented in the same way as an uncertain date: with multiple events/annotations qualified by links to relevant evidence or publications. Building awareness of this essential continuity into current models allows the construction of systems that should remain relevant and useful in the longer term.

II.7 Letter Model

Neil Jefferies, Howard Hotson, Christoph Kudella, and Miranda Lewis with Thomas Stäcker, Gertjan Filarski, and Thomas Wallnig

Chapter II.1 discusses in some detail the complexity of defining precisely what a letter is. Bearing that in mind, this chapter is not aiming to construct a comprehensive data model for letters and/or related communication channels. The focus will, instead, be on defining the elements of a data model that are necessary to answer three fundamental questions:

- What is it that distinguishes a letter from other literary forms? This establishes the scope for content of type 'letter' in our systems, and consequently the complexity of the rest of the data model.
- What are the essential characteristics of one letter that distinguish it from another letter? This is a crucial distinction that defines the approach to the systematic deduplication and reconciliation of records when multiple, distributed resources are brought together.
- How do we relate the various versions of a letter to one another? This establishes the basis for linking different versions of a letter together for discovery and comparison purposes.
    a) First, we will consider how versions are dealt with in existing bibliographic models.
    b) Second, we will present a model for the case of letters bearing in mind both existing bibliographic practice and the analysis presented in preceding chapters.

1 What Is a Letter?

The definition of a letter is designed to be as inclusive as possible of the variations discussed in previous chapters and therefore focuses on two essential criteria distilled from preceding analysis:

- *A letter is a physical object that is intended by one party to convey a message to another party primarily in textual form*. This distinguishes it from other goods that may be sent between people but can include cases where the letter is not the primary delivered item (e.g. when a message is enclosed along with other goods). As physical artefacts, surviving letters are generally identified in the data model by references to catalogue records, and those that do not survive by references in other extant sources.
- *The intent of the creator of the letter is for a version of it to be transmitted to another, specified, party or parties*. While the act of transmission is the key event that defines a letter, it is essential that this refers to any version so that drafts and literary copies are covered even though they are not sent themselves. The empirical evidence of this intent is typically found in features of the letter genre (identified in ch. II.1) which are either:
    a) necessary for the letter to be transmitted to the intended recipient (typically, the name and address of a specific individual);
    b) necessary in order for the recipient to understand the origins of the message and respond to it (i.e. the identification of the sender, of the date of sending, and of the place to which a response can be sent); and
    c) stylistic conventions which indicate to the recipient that the document is a letter (such as opening and closing salutations).

Note that this definition does not require specific recipients to be identified: open letters (with no specific addressee) and letters of recommendation (addressed 'To whom it may concern') are not excluded from this definition. Similarly, the mode of transmission is not proscribed, so that printed letters of dedication still fall within the definition of a letter. Open letters, letters of recommendation, and dedicatory epistles can nevertheless be distinguished in subordinate features of the data model, since for many purposes researchers may wish to treat them separately.

2 Distinguishing Letters

An important consideration in the construction of a data model for letters is that letters, and the evidentiary records of letters (e.g. mentions), can be extant in multiple forms which do not agree perfectly with one another; and individual scholars can interpret this surviving evidence in different ways, particularly in speculatively dating letters or identifying their senders, recipients, or places of sending and receipt. In the absence of originals, or with imperfect copies, the disposition of any

particular copy or record with respect to a particular letter identity is not necessarily clear-cut. Consequently, each record potentially represents an additional interpretation of the basic metadata about underlying truth: the original letter. This metadata is simply a formalization of the typical functional features of the letter genre, as mentioned in the previous section and chapter II.1.

2.1 Matching Copies to a Sent Letter

Which drafts and copies pertain to a particular 'letter actually sent'? A convenient guide in most cases is the basic letter metadata that is likely to be present, at least in part, in most forms of record. This suffices for most cases in which such metadata is available, because it is rare for the same person to write more than one letter to another person on the same day, although drafts may of course be dated earlier. More complete certainty can be provided by considering textual details, including incipit and explicit (the opening and closing words of the text, normally not including the salutations), abstracts, and transcription in whole or in part, if available. However, texts can be identical in the case of high-quality photographic reproduction, significantly altered in the case of early drafts, and substantially re-written in the case of edited letters.

Two special cases, where these assumptions break down, require more careful consideration:

- The first case occurs with letters in which the metadata are the same and the text is different. Such cases can arise, for instance, in the midst of a battle or major political upheaval, when a diplomat or military commander dispatches more than one letter to a superior on the same day. Since the essential purpose of a letter is to convey a message, characterized primarily by the text, these should therefore be identified as distinct letters. In other words, letters are ultimately distinguished by their message, not by their metadata; and letters with identical core metadata can be distinguished by persistent identifiers as well as by supplementary metadata.
- The opposite case comprises letters in which the text is the same but the metadata are different. Such cases arise, for instance, when the same correspondent writes several textually identical letters to the same addressee and sends them either on the same day or on successive days via different routes or couriers, in order to increase the chances of at least one copy reaching its destination.[1] Once again, since the purpose of a letter is to convey a message, determined principally by its text, these documents must be regarded as cop-

[1] Several instances are preserved in the National Archives in London. For instance, John Doddington (1628–1673), the English resident in Venice, sent multiple copies of letters from Turin on 17, 26, and 30 April, and 11, 18, 21, 24, and 29 May 1670 to the English secretary of state, Sir Joseph Williamson. See Alexandre Tessier, 'The Correspondence of John Doddington', *Early Modern Letters Online*, see http://emlo-portal.bodleian.ox.ac.uk/collections/?catalogue=John-dodington, accessed 20/03/2019.

ies of the same letter, *provided* that the texts are essentially identical. In principle, we can assume that there is a first Sent Letter, and that subsequent documents can be designated Sent Copies of the same, even if, in practice, it may not always be possible to determine the sequence in which the copies were sent.

2.2 Sent Letter not Available

When we lack the copy of the letter actually sent, we normally take existing drafts or copies in manuscript as evidence that a letter was actually sent and as close approximations of the text of the letter which was sent.[2] Some attributes of the Sent Letter can thus be inferred from these sources. Printed copies of letters also constitute such evidence, although in this case more care needs to be taken, since there is a greater likelihood (depending on the period, author, and type of letter) that such copies have been edited substantially before publication in print. Mentions of letters (e.g. in other letters) provide further evidence of otherwise missing letters, sometimes provide indication of their contents, and furnish conclusive evidence that they were both sent and received.

2.3 Material Properties

Generally, only physical artefacts can have material properties, although some physical attributes of recently missing objects may be gleaned from surviving photographic evidence, for example. A detailed data model needs to be capable of capturing all of these, but not all of them are equally useful to scholarship or likely in fact to be provided, so it makes sense to concentrate on those that can inform the identification or interpretation of letters. However, there are already extensive standards for describing material properties in use in the archives and manuscripts disciplines so there is no need for letter-specific developments.

- Material properties typically recorded for sent letters include the quality of the paper, the disposition of the text on the page, the presence of gilding, ribbons, seals, and the manner in which the letter has been folded – which are relevant to the manner in which the letter should be read and interpreted.
- Other copies of letters may also have useful evidential material properties – such as the reuse of paper from a previous draft or letter – which can help determine when drafts, copies, and edited versions were made.

[2] Evidence occasionally emerges indicating that a letter extant in draft was never polished up and sent, but this evidence is sufficiently rare not to undermine confidence in the assumption. At least as likely is the possibility that letters sent were never successfully delivered to the intended recipient, because they were lost in transit, for instance, or because the recipient had moved without a forwarding address.

3 Existing Bibliographic Models

The capture of correspondence metadata is, in many respects, a cataloguing activity. Although many steps in this process require background knowledge in both the humanities and information technology, metadata, as such, are primarily the area of expertise of library and information science (LIS). Consequently, it is pertinent to review the conceptual models developed by LIS specialists. In doing so, we must be mindful that the core-use cases for bibliographic records in library systems are resource discovery and stock management, rather than the enabling of scholarly analysis and discourse. This is a crucial distinction: for present purposes, we are more concerned with accurately representing knowledge about letters we believe to have existed based on a variety of evidentiary sources, rather than describing the physical content of defined collections which may include multiple copies of the same print edition of a given letter.

3.1 Functional Requirements for Bibliographic Records (FRBR)[3]

FRBR (1997) is a conceptual model of the structure of library information resources developed over the span of two decades. Its main benefit over traditional catalogue records (i.e. flat, field-based files) is that it employs an Entity-Relationship model[4] amenable to efficient implementation using a relational database system. It is, however, still rooted in requirements for managing modern print collections.

The FRBR model conceptualizes these entities in three groups, of which the 'products of intellectual or artistic endeavour' is of potential relevance to the modelling of letters. This is shown in figure 1 below.

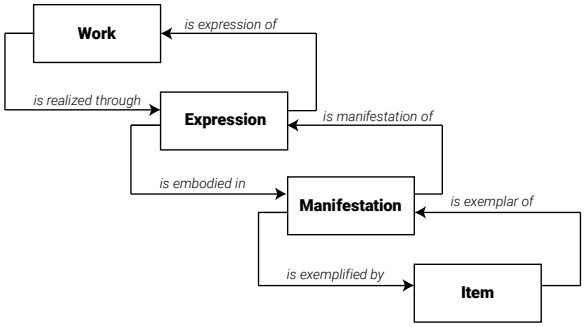


Figure 1: The entities of the FRBR model

[3] 'IFLA – Functional Requirements for Bibliographic Records', see https://www.ifla.org/publications/functional-requirements-for-bibliographic-records, accessed 20/03/2019.

[4] Peter Pin-Shan Chen, 'The Entity-Relationship Model – Toward a Unified View of Data', *ACM transactions on database systems: TODS* 1:1 (March 1976): 9–36, see https://doi.org/10.1145/320434.320440.

The entities in this group have strictly defined relationships to one another:

- *Work* is defined as a 'distinct intellectual or artistic creation'.
- *Expression* is defined as the 'intellectual or artistic realization of a work'. A *work* can be realized in many *expressions*, but an *expression* can only be the realization of one *work*.
- *Manifestation* is defined as the 'physical embodiment of an expression of a work'. An *expression* can be embodied in many *manifestations*, and a *manifestation* can embody many *expressions*, which potentially results in many-to-many relationships.
- *Item* is defined as 'a single exemplar of a manifestation', and is the only entity that has, potentially, a physical, information bearing, existence. A *manifestation* can be exemplified by one or many *items*, but an *item* can only exemplify one *manifestation*.

Applying the model to our use case, a letter itself, in the abstract sense of an intellectual creation, should be framed as a *work* entity. This is realized in one or more *expressions*, e.g. drafts, the autograph letter, as well as possible contemporary or subsequent copies. On the basis of these *expressions*, contemporary and/or modern editions may have been created, which constitute *manifestations*. These editions in turn are being held as *items* in one or many libraries worldwide (see fig. 2).

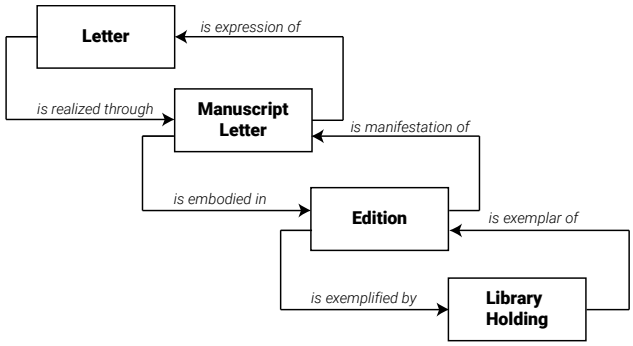


Figure 2: A letter in the FRBR entity model

This model is inappropriate as the basis for modelling letters for one very simple reason: if we regard the *item* as the only representation of a physical artefact in the model, then drafts, autograph letters, and contemporary copies should also be considered as *items*. This makes the *edition* level effectively redundant and the distinction between *expression*, *edition*, and *item* somewhat unclear. This shortcoming in the model also applies to many other manuscripts and unique materials.

Furthermore, the relationship between *expression* and *manifestation* raises further difficulties when we consider that the model must be populated on the basis of

extant evidence, and that our aim is to document our knowledge of the *existence* of letters rather than just the physical *items* that survive:

- The surviving *item* may not be an exemplar of the primary *expression* of the *work* (i.e. of the autograph letter). For instance, when the only surviving copy of a letter is found in the archive of the sender, the extant document is unlikely to be the autograph letter: it may be a draft retained or a copy made prior to sending. Alternatively, recipients may retain the autograph or a copy of it in a letter book; and a copy may also have been made of any of these various states by a third party.
- Many other modes of preservation are conceivable, but with regard to humanistic correspondences the following scenarios, in particular, must be considered regarding the historical tradition:
  a) Recipients copying letters written by eminent authors for further circulation and subsequent copying, each one of which is one step further removed from the original.
  b) Later attempts by senders to collect their letters with the aim of publishing them. For example, Erasmus asked his correspondents to create copies of his original letters from the period when he had not yet begun to keep copies.

3. Letters known to have existed by references in other letters and documentation but for which no physical *item* exists.
4. With regard to editions published in the early modern period, in the vast majority of cases, we do not know which *expression(s)* of a letter formed the basis for a particular contemporary edition.[5]
5. Letters may have been 'revised' before they were published in an early modern edition, as has been attested for many humanistic correspondence corpora. This practice effectively creates a new *expression* derived from an unknown *item*.
6. Finally, we must consider that some early modern editions are not based on any *expression(s)* of the letter itself but on a previous edition. These may be fitted into FRBR only if we accept that the *expression–manifestation* relationship is more conceptual than a representation of reality.

Consequently, although the FRBR definitions of *works* and *items* can be applied to the historical and contemporary treatments of letter materials, the definitions of *expression* and *manifestation* with their constrained relationships within the model hierarchy prove problematic.

[5] Indeed, the manuscript/letter sent to the printer was almost always discarded after use.

3.2 FRBRoo ('FRBR, object oriented')[6]

FRBRoo (2015) is a development of the FRBR model that transforms it from a hierarchical entity-relationship model to a more object-orientated formulation, and aligns it with the more event-oriented approach of CIDOC CRM.[7] However, attempting to retain the basic entity definitions of FRBR, which are not a priori process-oriented at the same level of detail as CIDOC CRM, leads to significant additional complexity. For example, many distinct classes of 'Work' are introduced, reflecting the different processes involved in creating them, but with a consequence that the top-level abstraction of creative intent becomes unclear. Along with the other practical complexities of the CRM identified in chapter II.6, this has led to somewhat limited adoption, except through programmatic generation from simpler sources.

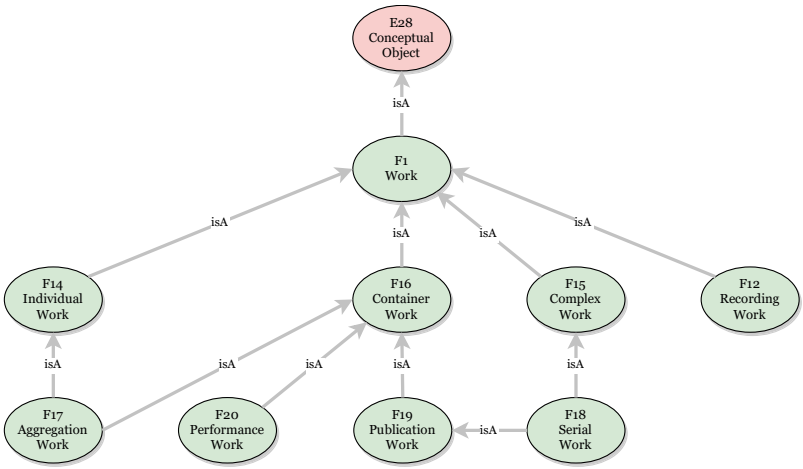


Figure 3: The proliferation of works in FRBRoo[8]

[6] 'IFLA – Definition of FRBRoo: A Conceptual Model for Bibliographic Information in Object-oriented Formalism', see https://www.ifla.org/publications/node/11240, accessed 20/03/2019.

[7] 'Home | CIDOC CRM', see http://www.cidoc-crm.org/, accessed 20/03/2019.

[8] Hans-Georg Becker, 'FRBR, Serials and CRM – Linked Open Data @hbz – Hbz Wiki', https://wiki1.hbz-nrw.de/display/SEM/2012/01/03/FRBR%2C+Serials+and+CRM, accessed 20/03/2019.

3.3 BIBFRAME[9]

BIBFRAME was developed partly in response to these limitations. As the foregoing discussion indicates, the print-orientated mindset behind FRBR does not lend itself to describing adequately the broad range of materials that libraries actually hold (including not only digital content but also special collections and archives). Rather than generating a more expansive model, BIBFRAME offers a simpler model, grounded in item description, with looser definitions that could be adapted to a wider range of content and will not be too onerous to populate and use. BIBFRAME prioritizes flexibility and ease of use by replacing the strict, defined hierarchies of FRBRoo with the ability to tag and to add relationships.

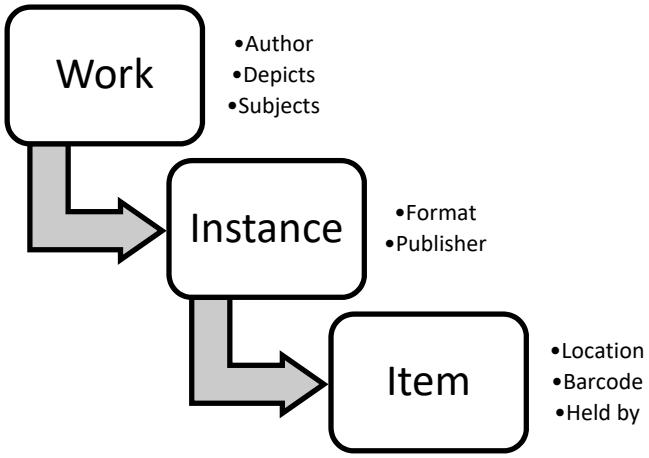


Figure 4: BIBFRAME 2.0 model overview

In BIBFRAME, *work* and *item* have broadly similar definitions to FRBR; but, between them, only a single intermediate level, *instance*, is provided, that groups together identical items (such as copies of the same edition of a book). BIBFRAME explicitly does not attempt to embody, in the core model, any information about the process by which artefacts are created or the relationships between them, and thus avoids many of the issues discussed in sections 3.1 and 3.2. As it is a Linked-Data model, additional relationships can readily be added to accommodate this information if required.

[9] 'BIBFRAME Model, Vocabulary, Guidelines, Examples, Analyses (BIBFRAME – Bibliographic Framework Initiative, Library of Congress)', see https://www.loc.gov/bibframe/docs/index.html, accessed 20/03/2019.

4 A New, Process-based Conceptual Model for Letters

The model proposed here attempts to combine the relative simplicity of BIBFRAME[10] but introduces some of the process orientation of the FRBRoo approach. A process focus is essential since letters are defined by the act of sending and by letter features that facilitate that act. In particular, stages (and the related originals) map quite closely to the letter types identified and agreed upon by a diverse group of scholars at the start of the *Cultures of Knowledge*[11] project.

The following diagram illustrates how the model is very simple for the common case of a single well-preserved sent letter in a collection (highlighted in yellow), yet can readily expand to include the additional complexities alluded to earlier.

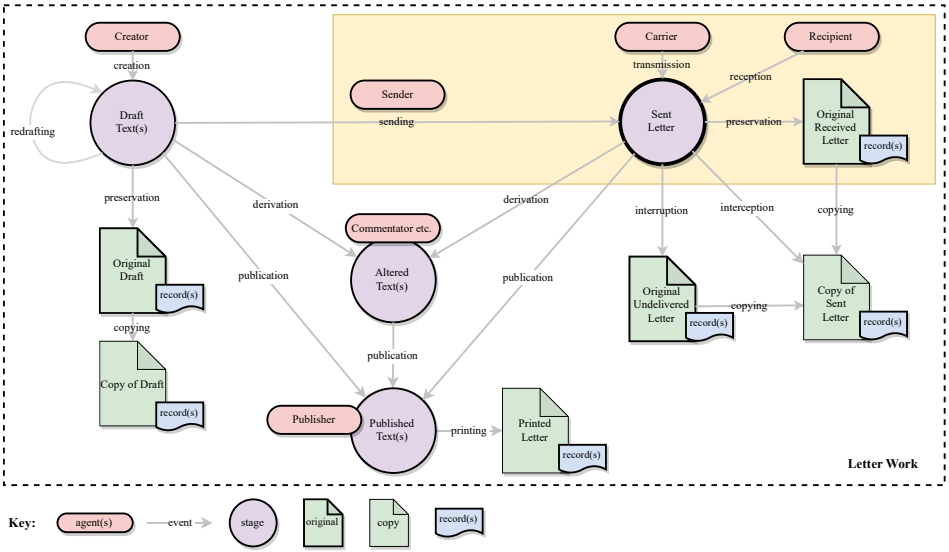


Figure 5: A process-oriented letter model

4.1 Key Components of the Model

The key elements of the model are the stages, originals, copies, and records. These are related by events modelled as described in chapter II.6 but, for clarity, only the essential events that relate to these elements and some of the more critical agents are shown. For example, all the copying events should have an associated agent.

[10] Like BIBFRAME, the model is also, of necessity, graph-based and amenable to representation using RDF and subsequently exposed as Linked-Open-Data.

[11] 'Cultures of Knowledge: Networking the Republic of Letters, 1550–1750', see http://www.culturesofknowledge.org/, accessed 20/03/2019.

- *Letter Work*: Conceptual essence of a resource, defined in this case as a single occurrence of the intent of the letter creator to transmit a message, characterized by the existence – actual or inferred – of a physical sent letter. A Letter Work, most importantly, provides a top-level persistent identifier for subsequent citation and reconciliation with new letter information. Other than that, its metadata is that of the underlying Sent Letter stage which is in turn drawn from either a surviving original or, by inference on the basis of evidence provided, by drafts, copies, and mentions. This is broadly equivalent to Work in BIBFRAME and FRBR terms.
- *Stages*: A letter passes through a number of distinct abstract stages during its existence, with transitions between states as a result of activity by various agents (recorded as events as discussed in ch. II.1). A stage is only included for a Letter Work if it corresponds to one or more extant evidentiary artefacts: a surviving original, physical or digital copies thereof, or references in other sources. The fact that stages are abstract allows the attachment of events to a letter in a consistent and meaningful way in the absence of surviving originals. However, the attributes of a stage should be drawn from the originals, in as much as they are known. The primary stages defined by the model are:
  a) *Draft Text*: The draft stage reflects the process of the creation of the letter and its preparation for sending and necessarily ends when a letter is actually sent. It may involve a number of different agents such as the creator of the letter, a scribe if the letter is dictated, and even a distinct sender, but usually these are the same person. Extant original drafts are, obviously, not sent and typically not intended to be sent. The text of a draft *need not* depart significantly from the text of a letter actually sent; but draft texts are part of the process of creating the letter sent while the copy of a finished letter taken prior to sending is not (it relates to the Sent Letter stage). A draft text *can* depart significantly from the text of the letter sent, particularly when a letter passes through several drafts before sending. Multiple drafts are regarded as pertaining to the same letter sent based on evidence that they are intended to convey essentially the same message to the same recipient in the same general period of time. The multiple textual variants generated as part of the process of composing a letter *prior* to sending must be differentiated from the alterations to the text of copies of a letter made *after* sending, for instance in preparing them for print publication (see point c. below).
  b) *Sent Letter*: There can only be a single Sent Letter state, which can be either *Undelivered* (for a letter that does not reach its intended recipient for any reason) or *Received*. This singular state and the associated original thus define the main identifying attributes of the *Letter Work* uniquely. *Intercepted* letters, more common in diplomatic correspondence, can result in additional copies of the sent letter being made en route, after

which the original may continue on its journey and be *Received*, or it could simply remain *Undelivered* (in which case, *Interception* is a special case of *Interruption*). In the anomalous case, mentioned in section 2.1, of multiple identical letters being sent from the same sender to the same recipient by different routes, the earliest letter sent is considered to be the Sent Letter, with subsequent dispatches being designated as Sent Copies. In cases where not all letters are delivered, or the sending order is not clear, the Sent Letter should be considered to be the first delivered, representing the earliest successful transmission of a message.

c) *Altered Text*: An altered text can be derived from artefacts of any other stage either singly or in combination. Unlike an ordinary copy, it must have textual content that differs materially from the source material as a result of a scholarly or editorial process, by the original author or others. Such changes are typically made prior to publication in print, either because the author wants to polish the style or enrich the substance of a letter, or because author or editor wish to remove from the letter material that is confidential, personal, or otherwise not deemed suitable for publication. Consequently, multiple different Altered Text stages may exist, corresponding to the actions of different agents on different sources. If the content is not changed significantly, it should be considered a copy of an earlier original rather than a new stage. Altered texts can also include annotations, amended or abridged manuscript copies, and scholarly editions.

d) *Published Text*: Not all altered texts are necessarily published, and they may be published by different publishers in different forms (such as physical and digital editions). Equally, letters and drafts may be published directly without alteration. A dedicatory letter may be published without it actually being sent, in which case, the Letter Work would be defined by the Published Letter stage rather than the Sent Letter. In practice, it may be possible to assert that a specific published or altered text is derived from specific originals or copies (which would be represented by events linking those sources to the Published text stage).

3. *Originals and Copies*: Copies refer to the different physical or digital artefacts that embody the content of a letter in any particular stage. A physical original[12] naturally has primacy over other copies in terms of defining the attributes of the overarching state. Additional copies should not differ significantly from the original – or they should be considered part of a derived Altered

[12] Autographs are orthogonal to the classification of originals and copies. An autograph is merely a document written by its author. Thus a draft, a letter sent, *or a copy* can be an autograph (the latter, for instance, a copy of a finished letter made by the author prior to sending). Equally, an original draft need not be an autograph (it could be written by a scribe from dictation), and the same applies theoretically to a letter sent, although ordinarily at least the signature would be written by the sender's own hand.

Text stage instead. As a consequence, an original and its copies basically reference a common text which maps readily to the idea of a version in formalist literary theory[13] and also aligns well with modern computational approaches. However, this is not necessarily true for other metadata (e.g. early modern printed copies often lack basic metadata, such as the places to which letters are sent). Copies represent a direct embodiment of the letter content, and can refer to manuscript copies, print editions, or digital images or texts, but not, for example, references in other letters which would be classified as Records. Modern print or digital editions with multiple identical instances would generally be represented as constituting a single distinct copy. Copies exist for a variety of different *purposes*:

a) Authors can copy their own letters prior to sending in order to retain a record of the complete correspondence.
b) Others can copy third-party correspondence for personal use, for deposit in a repository, or for circulation.
c) Copying also takes place in the act of publication and in preservation in photographic and digital media.

4. *Records*: Records capture interpretation(s) of the evidence contained in the documents themselves: i.e. they involve deciphering the names, places, and dates that provide the core metadata of a letter record. Two scholars, looking at precisely the same manuscript, can interpret the ink on paper in different ways, even when dealing with core metadata; and records can diverge further still when dealing with supplementary metadata. Thus, multiple records can exist for any particular document. The integrity of all records assembled in a union catalogue must be maintained in order to ensure that the provenance of the assertions about the underlying document in the catalogue can be ascertained.
   a) The *author* of a letter can retain a mere listing of letters sent, or abstracts and partial transcriptions of them.
   b) The *recipient* of a letter can record the news conveyed by letters in a diary or commonplace book, can mention previous letters in responding to them or in passing on news to others, and can quote from letters received in print publications.
   c) *Third parties* – including historians writing much later – can mention letters that they have read or seen, and construct inventories of extant collections for various purposes. Such notices pertain to all the various states and copies of a letter, including drafts, letters actually sent, edited versions, and copies of them, including those in printed and digital editions.

[13] Ivor Armstrong Richards, *Practical Criticism: A Study of Literary Judgment* (London: Kegan Paul Trench Trubner and Company Limited, 1930), see http://archive.org/details/practicalcritici030142mbp, accessed 20/03/2019.

d) *Scholars* can subsequently subsume under one entry in a modern inventory or edition records of multiple states and copies of the same letter.
e) *Librarians* will construct catalogues of letters in their collection. Such catalogues may also contain records of other artefacts, such as diaries, that themselves contain records of letters.

In some cases, when the original copies have been lost, these records may be the only extant evidence that exists. Within the data model, a record should contain, at least, the core letter attributes that can be gleaned from the record, and the source of the record.

4.2 Agents and Events in the Process-based Letter Model

This core conceptual model now needs to be related to the contextual framework of people, places, and time. Since, as discussed in chapter II.6, events are the primary mechanism for expressing these relationships, so it is necessary to define a small number of core events that are fundamental to the concept 'letter'. As described above, the essence of a letter is that it is intended as an act of communication. Consequently, it would be prudent to consider the field of Communication Theory[14] to see if suitable models already exist that can be adapted for letters.

Shannon and Weaver's foundational model[15] in the field identifies the *source*, *receiver*, and *transmission medium* as key elements in a communication. These three elements map quite readily to *sender*, *recipient*, and *carrier* in our conceptualization of a letter. However, their model neglects the *content* of the communication, which is fundamental to our model although, admittedly, this aspect is not pertinent to Shannon and Weaver's subsequent analysis. A later development, Berlo's SMCR[16] Model of Communications,[17] addresses this shortcoming by distinguishing the Message (i.e. the content) and the Channel (i.e. the mode of transmission). From the standpoint of an event terminology, Berlo's Source, Channel, and Receiver can all be regarded effectively as agents[18] that act on the message. These equivalences can be mapped as follows:

[14] 'Communication Theory', *Wikipedia*, 1 September 2017, see https://en.wikipedia.org/w/index.php?title=Communication_theory&oldid=798321176, accessed 20/03/2019.

[15] Claude Elwood Shannon and Warren Weaver, *The Mathematical Theory of Communication* (Urbana: University of Illinois Press, 1971).

[16] Source-Message-Channel-Receiver.

[17] David K. Berlo, *The Process of Communication* (New York, NY: Holt, Rinehart, & Winston, 1960).

[18] See discussion of Agency in II.6.

Table 1: Equivalences between SMCR Model and the Letter Model

SMCR Model	Letter Model	Key Event
Source	Creator / Sender (Agents)	Creation / Sending
Message	Letter (Entity)	
Channel	Carrier (Agent)	Transmission
Receiver	Recipient (Agent)	Reception

Relating these events to the conceptual model, it is evident that they must, by definition, apply to the Sent Letter, characterized by a surviving or inferred original.

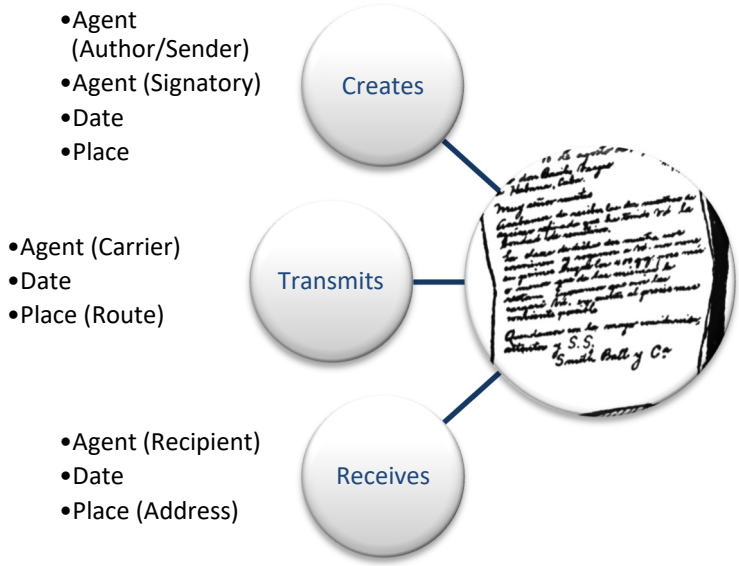


Figure 6: Core predicates of an event-based letter model

This model effectively captures the essential distinguishing characteristics that make a particular letter unique: sender, date, location, destination, recipient, and textual content. However, the nature of surviving letters dictates that a number of assumptions often have to be made in order to populate the model:

- In general, it is assumed that the recipient is the addressee and that the letter was successfully delivered to the address specified. There may be external evidence of this fact, such as mentions in subsequent letters or the presence of the letter in a recipient's collection, but this is not guaranteed.
- Without additional evidence, it is not possible to specify the date that a letter is received, although it is obviously bounded by the sent date.
- We generally have relatively little information about the transmission event, although more research into postal routes[19] may enable reasonable inferences to be made (see further ch. IV.2 below). Undelivered collections such as Simon de Brienne's trunk[20] are an interesting exception where the interrupted transmission event is a key feature of the collection.
- *Forwarding or redirection* of letters requires that creation and sending are treated as distinct events, since they happen at different times and are sometimes carried out by different agents. If the letter (either the original or a second, enclosed, letter) is sent on unaltered, then the first recipient becomes the sender for the next leg of the letter's transmission to a second recipient and address, effectively making the forwarder part of the transmission process. If the first recipient modifies the letter content before forwarding it, then they are authors of a new, derived, version of the letter in the form of an *Altered Text* as well as senders. Thus a single original can relate to multiple distinct Sent Letter stages.
- *Held* letters, where the recipient of a letter holds it, unread, to be picked up by the intended reader is treated in a similar way. The holder of the letter becomes part of the transmission process and *holding* should be a valid mode of transmission.
- *Publication* in print, considered a valid transmission channel, with publisher/printers and distributors as the key agents involved. A letter could be released directly to a publisher, or sent to one party in the expectation that it will be printed for consumption by a wider audience. As discussed in the next section, this is a simplification, since print dissemination typically involves a number of agents with different roles (compositors, printers, publishers, distributors, booksellers, etc.). Print dissemination may be represented by more than one event if sufficiently detailed information is available.

[19] 'Postal Networks: Early Modern News Networks', see https://earlymodernnewsnetworks.wordpress.com/tag/postal-networks/, accessed 20/03/2019.

[20] 'A Postal Treasure Trove', Signed, Sealed, & Undelivered, see http://brienne.org/unlockedbriennearchive/, accessed 20/03/2019. See also ch. II.1

4.3 Other Key Events

There are many other event types that could be related to letters beyond the core model presented above. However, we would like to focus on three particular types that are particularly relevant to the study of the republic of letters through surviving letters:

- *Mentioning:* As discussed above, the mention of one letter in another letter effectively makes the source letter a record for the mentioned letter. In fact, this mention can sometimes be the only evidence for letters that have not been preserved. In practice, this record could include a 'mentioning' event that relates the two letters at the creation time and place of the source letter. An analogous mechanism can also be used to capture mentions of other entities of interest such as people, organizations, places, and events in the broader contextual model.
- *Aggregation:* A key aspect of the preservation and study of letters is the practice of creating collections, which, importantly, results in the creation of letter *records*. This act can be considered as an event that affects the material and/or intellectual context of the letters involved and should therefore be recorded, if possible, along with the agents and motivations responsible. Examples include:
    a) *Accession:* A key part of the material provenance of letters is their movement into, and between, personal and institutional collections in various locations, especially where physical and digital instances may exist. As the documentary basis expands to include a broader range of sources and records, understanding the criteria and circumstances of collection creation is likely to become increasingly important.
    b) *Compilation:* Creating published editions is an editorial process which draws together specific versions of letters, and possibly generates new versions. As a text-oriented action, it may not always be clear which *items* constitute the source material, so a *version* reference may be more appropriate. Some editions are based entirely on earlier editions and should probably be considered as derivative works of the previous *edition*, with only a secondary relationship to the original letter entities through the derivation relationship.
    c) *Selection:* For the purposes of analysis or exposition, a scholar may select a superset or subset of material in existing collections and editions. Capturing this process provides a ready mechanism for linking to the subsequent scholarly outputs, which helps contextualize and thereby enriches the items in the selection.

3. *Publication:* Publication is a potentially complex event, conflating a number of distinct events (typesetting, printing, distribution, etc.) into a simplifying umbrella event. It is possible to tease out a little of this detail by allowing multi-

ple roles (typesetter, printer, etc.) if necessary. As more information about the printing and book trades[21] becomes available, these relationships may become more amenable to exploration.

5 Relationship of the Model to TEI <corresp Desc>[22]

The minimal model proposed herein maps readily to many aspects of the <corresp Desc> structure whilst being flexible enough to accommodate a broader range of possible activities that are either covered by other elements within TEI, or not at all since they lie outside the realm of the text itself. Events, obviously, correspond closely with the <corresp Action> elements, while the materiality of the letter ('Letter as Object') is captured by Originals and Copies in the Conceptual model, while the content ('Letter as Text') maps to the corresponding Stage.

The notion of a threaded discussion, with subsequent letters written in response to prior letters, lies outside the scope of the model presented here since, in the absence of explicit *mentions*, this is a much more subjective interpretation. In practice, these relationships should be expressed as annotations[23] with appropriate attribution and citation of evidence. Where explicit mentions exist, these relationships can be derived programmatically via simple inference logic.

6 Conclusion

Having analysed the shortcomings of existing bibliographic models, we present a new model that captures the complexity of the letter as an object of study. In particular, it focuses on the processes of letter creation, transmission, and publication which define the variants of the form, and the multiplicity of surviving artefacts and evidence that may result.

By adopting a graph approach to constructing the model, which only includes elements when they are necessary, the model is very simple for the common case of a single surviving sent letter, but readily expands to accommodate the existence of multiple drafts, complex transmission routes, published dedicatory letters, and other complexities. At the same time, it is still possible to map the basic components of the model to elements within the existing <corresp Desc> approach.

---

21 'London Book Trades – Lbtwiki', see http://lbt.bodleian.ox.ac.uk/mediawiki/index.php/Main_Page, and 'British Book Trade Index', see http://bbti.bodleian.ox.ac.uk/, both accessed 20/03/2019.

22 Peter Stadler, Marcel Illetschko, and Sabine Seifert, 'Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing <corresp Desc>', *Journal of the Text Encoding Initiative* 9 (24 September 2016), see https://doi.org/10.4000/jtei.1433.

23 'Web Annotation Data Model', see https://www.w3.org/TR/annotation-model/, accessed 20/03/2019.

The abstract stages and linking events drawn from Communication Theory mean that those elements of the model can be flexibly applied to other modes of communication. For example, the stages and many of the events presented are perfectly applicable to email if we replace the physical Sent Letter with a digital Sent Email as the element that defines the overarching Work. Consequently, the model permits the representation and scholarly analysis of multi-modal intellectual exchange when the need arises.

III Systems, Methods, and Tools

III.1 Assembling Metadata

Dirk van Miert and Elizabethanne Boran

With contributions from Gábor Almási, Ivan Boserup, Clizia Carminati, Per Cullhed, Antonio Dávila Pérez, Vittoria Feola, Andreas Fingernagel, Ad Leerintveld, Gerhard Müller, Alexa Renggli, Patryk Sapala, Justine Walden, and Axel E. Walter

1 Introduction

The most basic precondition for collaborative scholarship on learned correspondence is to assemble relevant data scattered in innumerable places across and beyond Europe. This complex task is rendered more difficult by the fact that the sources of such data can be distinguished in several different ways. One such distinction relates to the *media* in which letters are preserved: some letters are preserved in print, others in manuscript. Another distinction relates to the state of the catalogue *records*. Some collections of correspondence are thoroughly catalogued at the item level: in such cases, individual letters are listed separately, whether those letters are published in a printed volume or preserved in manuscript. In other cases, the only available catalogue records describe collections of letters rather than individual items: on the one hand, basic bibliographical data relate to whole collections of printed letters; on the other, collection-level descriptions relate to entire folders or boxes of manuscript letters. The limiting case is collections of manuscript materials which include letters but have yet to be catalogued at all.

This chapter will explore some of the systems and processes needed to assemble epistolary metadata in all of these forms on a large scale. Since published letter collections are better catalogued, more accessible, and easier to work with than manuscript collections, these printed materials are handled first. Since the first stage in the process of collecting a census of printed letters is to collect bibliographical data on printed letter collections, the assembling of collection-level descriptions naturally precedes the extraction from them of item-level records. Prior to either stage, however, a brief consideration of the history and nature of printed letter collections is in order. A similar order of exposition will then be followed in dealing with the more troublesome problem of manuscript letter collections. The question of how individual letter records can be reduced to the same format and reconciled with other data will be postponed to the next chapter.

2 Letters in Print

2.1 Printed Letter Collections: The History and Hazards of a Textual Genre

Publishing one's own letters was already a customary procedure in Ancient Rome: Cicero and Pliny the Younger carefully selected and prepared some of their letters for public dissemination. Ever since Cicero's *Epistolae ad Atticum* were rediscovered by Petrarch (1304–1374) in 1345, humanists sought to emulate their ancient forebears by collecting their own letters in volumes of correspondence. The discovery stimulated Petrarch to make a selection of his own letters and to publish them after a process of thorough re-editing. The humanist Pietro Bembo (1470–1547) did the same thing a generation later, and the Renaissance movement as a whole spread such activity across Europe.[1] The arrival of the printing press ensured that this tributary of manuscript correspondence would be transformed into an ocean of readily available letter models throughout Europe. The proliferation of letter collections in Renaissance Europe undoubtedly held implications both for early modern letter writing and for how collections were created. As Cecil H. Clough observes, 'A letter collection was seen by the humanist of the Renaissance as a literary work in its own right'.[2] This reminds us that early modern letter collections should be treated warily: whether preserved in manuscript or in print, a 'letter collection' may have undergone a variety of different kinds of editorial intervention in the early modern period.

[1] Cecil H. Clough, 'The Cult of Antiquity: Letters and Letter Collections,' in Cecil H. Clough, ed., *Cultural Aspects of the Italian Renaissance. Essays in Honour of Paul Oskar Kristeller* (New York: Zambelli, 1976), 33–67.

[2] Ibid., 35.

In the first place, the principal correspondent himself, in collecting his own correspondence for posterity with an eye to publication, was often tempted to weed out material he deemed unsuitable for inclusion. A prime example is the policy of the learned archbishop of Armagh, James Ussher (1581–1656), who excluded from his archive letters from family members, projecting instead an emphasis on his official role.[3] A further phase of selection is evident in Ussher's seventeenth-century editor, Richard Parr (1616/17–1691), who included in his work on Ussher's life and letters not only letters by and to Ussher, but also letters which were, at best, only tangentially connected with him.[4] Further varieties abound. Some editors, early modern and modern, included epistles dedicatory, while others did not.[5] Equally, fictitious letters might be included alongside genuine epistles, since 'letter collections' in manuscript and printed form did not necessarily imply a collection of correspondence actually sent, but often harked back to the *ars dictaminis*, the medieval treatises on the art of letter writing, by including letters written as models for emulation rather than for sending. Printed letter collections (known as 'epistolaries') also collected a variety of material: edited collections were not always devoted to one scholar but sometimes included an anthology of letters and other sources by many different hands.[6] The ready market for printed epistolaries ensured that scholars increasingly kept manuscript archives of their correspondence. As a result, the archive of early modern correspondence available in both manuscript and printed form is vast and is not limited to Latin sources. Vernacular letter books became increasingly common and were used not only by scholars but also by ambassadors, merchants, and, increasingly, anyone who could write. In short, whether dealing with manuscript collections or printed letter collections, scholars must be alert to a range of possibilities: are the documents that have been preserved undoctored autographs, silently censored manuscript copies, letters never actually dispatched, or perhaps even purely literary compositions never intended for sending?

More surprising is another form of exclusion which reflects the origin of the epistolary as a Renaissance literary genre: most letter collections published before the late seventeenth century contain only the letters written by the principal correspondent, without the answers that person might have received; and this exclusion is maintained irrespective of whether these collections were published by the author during his or her lifetime, or posthumously published by relatives or students, acting either on instructions from the primary author or of their own accord. From the latter half of the seventeenth century onwards, however, a development is

[3] Elizabethanne Boran, ed., *The Correspondence of James Ussher 1600–1656*, 3 vols. (Dublin: Irish Manuscript Commission, 2015).

[4] Richard Parr, *The Life of the Most Reverend Father in God, James Ussher* (London: Nathanael Ranew, 1686).

[5] Michael Hunter, Antonio Clericuzio, and Lawrence M. Principe, eds., *The Correspondence of Robert Boyle*, 6 vols. (London: Pickering and Chatto, 2001), i, xxxiv.

[6] For an early example, see *Epistulae diversorum philosophorum* (Venice: Aldus Manutius, 1499).

noticeable away from the literary criterion, in which a collection of letters represents the epistolary compositions of a single pen, towards the historical criterion, in which an epistolary preserves the discussions undertaken at a distance between one learned individual and their contemporaries, in which even minor scholars were deemed worthy of a place.

The transition appears to have begun in 1670, when Johann Andreas Bosius (1626–1674) published the letters that Thomas Reinesius (1587–1667) had exchanged with Christian Daum (1612–1687) arranged chronologically, to make their conversation as easy to follow as possible. Three more traditional collections of letters, exclusively by Reinesius, had recently appeared, and Bosius felt the need to explain in his preface the still unusual editorial choice of publishing both sides of the conversation:

> *I have included the letters by Daum because otherwise the letters of Reinesius cannot be sufficiently understood, and because I am aware that great men have deplored the fact that the same thing has not been done in the letters of Scaliger, Casaubon and other famous men. I will do the same for other letters, if I am allowed to publish more.*[7]

A rather different experiment was conducted in Petrus Burmannus's edition of the correspondence of Marquard Gudius (1635–1689) and Claude Sarrau (d. 1651), published in 1697. As Burmannus explained, 'I have first given the letters of Gudius himself to friends and acquaintances, and then I added the ones which friends wrote to him'.[8] This may represent a compromise between the traditional letter collection, emphasizing the literary productions of one author, and the emerging practice of documenting entire epistolary conversations. From an historical perspective, the disadvantages of this method of organization seem obvious: Burmannus invited his reader first to read all of Gudius's letters to others and then to move back in time again to start with letters that others wrote addressed to Gudius. Anyone wanting to read the epistolary conversation in chronological order was forced continuously to flip back and forth. Closer inspection reveals, however, that the collection was so incomplete that there were hardly any letters responding to one another anyway. But when others organized more complete correspondences in this way, its disadvantages became apparent: in the edition of 'the letters of Gerardus Joannes Vossius and of other illustrious men to him' published in 1690,

---

[7] *Thomae Reinesii Epistolae […] ad cl[arissimum] v[irum] Christianum Daumium: In quibus De variis scriptoribus disseritur, loca obscura multa […] Accedunt alia ejusdem, et ipsius Daumii epistolae ad Reinesium*, ed. Joannes Andreas Bosius (Jena: Gothofredus Schulzen, 1670), sig. A4v: 'Adjeci Damianas, quod satis alias intelligi Reinesianae non possent, quodque non ignorabam, magnos viros doluisse, quod idem Scaligeri, Casauboni, aliorumque clariss. virorum epistolis factum non esset. Idemque et aliis, si plures edere licuerit, praestabo'.

[8] *Marquardi Gudii et doctorum virorum ad eum epistolae […] et Claudii Sarravii […] epistolae*, ed. Petrus Burmannus (Utrecht: Franciscus Halma and Gulielmus van de Water, 1697), sig. **v: 'Praemisimus ipsius Gudii ad Viros, quibuscum ipsi amicitia et usus intercessit, Epistolas, quibus subjunximus, quas ejus amici ad illum dederunt'.

the epistolary dialogues could only be reconstructed by leafing back and forth between Vossius's letters in the first part of the volume and those addressed to him in the second.[9]

As the *historia litteraria* displaced Renaissance epistolography as the main motivation for publishing letter collections, the preference for publishing epistolary conversations in chronological order was finally consolidated. A watershed can be found in 1708, when the advice of the polymath Daniel Morhof (1639–1691) was posthumously printed in his much used *Polyhistor*. Morhof dedicated a paragraph to 'Ordering letters chronologically' and wrote: 'But this I would prefer with authors of letters, to have the answers joined, so that we can judge everything better. I would also mention the letter dates. For good reasons, Thomasius desires both of these in the preface which precedes his edition of the letters of Boxhorn'.[10] Theodor Janssonius ab Almeloveen (1657–1712), the industrious editor and life-writer, referred to both Bosius and Morhof in the preface to his monumental third edition of the correspondence of Isaac Casaubon (1559–1614):

> *Because the famous Andreas Bosius taught me that learned men greatly deplored that there were no answers added to the letters of Scaliger and Casaubon, since without these they could not be sufficiently understood, I have at the top of each letter in the margin written the number of the letter of Scaliger, or Baudius or Lipsius or others to which Casaubon responds. At the top I have given the number of the letter which responds to it. But if these responses have not yet been published, I have included them.*[11]

This paper trail was continued by Adamus Henricus Lackmannus (1694–1754) in 1728: in the preface to his rather miscellaneous edition of the *Letters to Lossius* and other people's letters to various others, he agreed that it is useful to add the answers, citing the prefaces of Bosius and Burmannus in support and referring to

---

[9] *Gerardi Joannis Vossii et clarorum virorum ad eum epistolae collectore Paolo Colomesio Ecclesiae Anglicanae presbytero. Opus omnibus philologiae et ecclesiasticae antiquitatis studiosis utilissimum,* ed. Paulus Colomesius (London: Samuel Smith, 1690). The book was republished in 1691 and 1693, with different page ranges.

[10] Daniel Morhof, *Polyhistor, sive de notitia auctorum et rerum commentarii, quibus praeterea varia ad omnes disciplinas consilia et subsidia proponuntur*, vol. 1 (Lübeck: Petrus Böckmannus, 1688), bk. 1, ch. 23 ('De epistolarum scriptoribus', p. 275: 'Illud tamen ego velim, in epistolarum scriptoribus semper responsorias adjungi; ita rectius de omnibus judicaremus. Velim et tempore epistolarum sollicite adnotari. Quae duo non sine causa desiderat in Epistoliographis Thomasius praefatione illa, quam Boxhornii epistolis a se recusis praemisit'.).

[11] Isaac Casaubon, *Epistolae, insertis ad easdem responsionibus, quotquot hactenus reperiri potuerunt, secundum seriem temporis accurate digestae*, ed. Theodor J. ab Almeloveen (Rotterdam: Caspar Fritsch and Michael Böhm, 1709), sig. **r: 'Deinde quum Vir clarissimus, Andreas Bosius [note: Praefat. praem. Daumii et Reinesis Epistol. Vide Morhof. Liter., bk. 1, ch. 23. §. 5.], me docuisset viros doctissimos vehementer dolere, Scaligeri Casaubonique Epistolis non additas Responsorias, quod sine iis satis intelligi nequeant; Epistolarum principiis in ora adscripsi numerum Epistolarum Scaligeri, Baudii, Lipsii aliorumve, ad quas respondet Casaubonus; in calce vero, quo loco ad Casaubonianas illorum virorum Responsoriae reperiantur; insertis tamen iis Epistolis, quae hactenus lucem non viderunt'. Almeloveen referred to Johannes Möller's 1708 edition of Morhof's *Polyhistor*.

precedents such as Almeloveen's edition of Casaubon's letters.[12] By the end of the 1720s, it had become normal practice to document the epistolary dialogue as fully as possible. The scholarly reader needed to be able to reconstruct the contents of the discussions and learn not only about the ideas developed by a single author or the style in which they were expressed, but about his web of communication and the benefit he derived from communication with others. This coincided with a growing awareness of the context of learning, and a shift from a focus on the history of heroic scholarship to the history of scholarship in more general terms. Perhaps this points to a growing historical self-awareness in the republic of letters itself.

Yet other pitfalls remained hidden in early modern printed epistolaries, ready to trip up the unwary scholar. Modern critical editions – such as those of Justus Lipsius (1547–1606), Joseph Justus Scaliger (1540–1609), and Casaubon – are based wherever possible on autograph letters; and painstaking, word-for-word collation with the early modern printed collections has revealed the many ways in which early modern editors silently tweaked texts, censored passages, or ignored entire letters, for stylistic, personal, or political reasons. The posthumous edition of the correspondence of Casaubon, published in 1638, for instance, silently omits not only references to his family life but also to social pleasantries exchanged with his correspondents, apparently in an attempt to construct Casaubon's posthumous identity as a more masculine and resolute hero. Some of his scathing remarks on Catholic enemies were toned down, producing the impression of a more composed and less passionately involved scholar.[13] Unfortunately, returning to the autograph letter is not always possible, and not merely because the original letter has been lost accidentally. The sad fact is that autographs or apographs were often destroyed after the printed editions became available, evidently in order to make it impossible to 'get behind' the edited letters.

[12] *Epistolae diversi argumenti, maximam partem a variis ad Lucam Lossium & post eum a Duraeo, Langwedelio, Boeclero, Portnero, Berneggero, Freinshemio aliisque ad alios exaratae*, ed. Adamus Henricus Lackmannus (Hamburg: widow of Theodorus Christophorus Felginerus, 1728), sig. [8]v: 'Immo et ex re et emolumento Rei litterariae est, Epistolis adjungere responsiones. Joannes Andreas Bosius, edens Thomae Reinesii, Medici ac polyhistoris excellentissimi, ad v[irum] c[larissimum] Christianum Daumium Epistolas, in praefat. ita: *Adjeci Daumianas, quod satis alias intelligi Reinesianae non possent, quodque non ignorabam, magnos viros doluisse, quod idem Scaligeri, Casauboni, aliorumque clarissimorum Virorum epistolis factum non esset.* Eundem fere in sensum Petrus Burmannus in limine praestantissimi operis Epistol. quod, adplaudentibus musis, *Ultrajecti* 1697. lucem vidit: *praemisimus ipsius Gudii ad viros, quibus cum ipsi amicitia et usus intercessit, Epistolas, quibus subjunximus, quas ejus amici ad illum dederunt.* Add. Gerardi Joannis Vossii et Clarorum Virorum ad eum Epistolae. Joannis Keppleri, item, Pauli Sarpii, Isaaci Casauboni et aliorum. Quis enim omnes recenseat?'

[13] Paul Botley and Maté Vince, eds., *The Correspondence of Isaac Casaubon in England*, vol. 1 (Geneva: Droz, 2018), 65–6.

2.2 Assembling Collection-level Descriptions: Towards a Bibliography of Early Modern Printed Letter Collections

Despite these pitfalls and limitations, printed letter collections provide an attractive starting point for assembling the huge quantities of data needed to form a data-driven impression of the republic of letters as a whole. Their advantages as a point of departure are several. Many letters are only preserved in print. Printed letter collections are more accessible than manuscripts, since they typically survive in multiple copies. Printed texts are easier to read than handwritten ones, opening up the possibility of experimenting with crowdsourced metadata and automatically generated machine-readable text. Printed collections already benefit from the work of their editors in assembling related material in one place. Bibliographical records provide ready-made collection-level descriptions of printed letter collections. Large numbers of these records can be identified relatively easily via meta-catalogues such as *WorldCat* or the *Karlsruher Virtueller Katalog* as well as national bibliographies of early modern books, such as short title catalogues VD 16, VD 17, EEBO, and ECCO, in addition to chronologically more comprehensive catalogues such as *Gallica* or geographically more inclusive ones such as *Europeana*.[14] Moreover, titles can be automatically exported from these catalogues to bibliographical reference databases such as *Zotero*, *Endnote*, or *Refworks*, facilitating the first stage of data collection considerably.

Best of all, four substantial bibliographies of letter collections already exist, providing abundant material with which to begin. The oldest of these is the bibliography of epistolaries published by Arenhold in 1746, which lists 816 titles, organized by country of publication, which cover the entire European space.[15] More recent bibliographies of printed letter collections have been national in scope. A second major resource is the bibliography embedded within Monika Estermann's four-volume inventory of printed letters to and from German authors of the seventeenth century, which lists 567 epistolaries.[16] Whereas these epistolographies were printed between 1600 and 1750, a second set of four volumes was added to this series by Thomas Bürger, who gives a bibliography of approximately 1,066 works printed between 1751 and 1980 that contain published letters.[17] A third

[14] See www.worldcat.org; https://kvk.bibliothek.kit.edu; www.ustc.ac.uk; http://estc.bl.uk/; https://www.kb.nl/en/organisation/research-expertise/for-libraries/short-title-catalogue-netherlands-stcn; https://opacplus.bib-bvb.de/; http://www.vd17.de/; https://eebo.chadwyck.com/home; https://www.gale.com/intl/primary-sources/eighteenth-century-collections-online; https://gallica.bnf.fr; www.europeana.eu; all accessed 20/03/2019.

[15] Silvester Johannes Arenhold, *Conspectus bibliothecae universalis historico-literario-criticae epistolarum: Typis expressarum et m[anu]s[crip]tarum, illustrium omnis aevi et eruditissimorum auctorum* (Hanover: Haeredes Foersteriani, 1746).

[16] Monika Estermann, *Verzeichnis der gedruckten Briefe deutscher Autoren des 17. Jahrhunderts. Teil 1: Drucke zwischen 1600–1750*, 4 vols. (Wiesbaden: Harrassowitz, 1992–3).

[17] Thomas Bürger, *Verzeichnis der gedruckten Briefe deutscher Autoren des 17. Jahrhunderts. Teil 2: Drucke zwischen 1751 und 1980*, 4 vols. (Wiesbaden: Harrassowitz, 2002). In addition to bibliographies of epistolaries, these works contain item-level records of c. 110,000 letters printed between 1600 and 1980.

important source is the unpublished, typewritten bibliography of the 492 epistolaries consulted in compiling the so-called 'Apparatus Molhuysen', an index card file of 39,000 item-level descriptions of letters written to or from early modern scholars residing in the Dutch Republic.[18] A fourth major achievement is the three-volume bibliography of Italian correspondents, compiled by Corrado Viola, which lists printed letter collections of 3,827 letter-writers, including many modern editions, some including just a few letters.[19]

In order to capitalize on these advantages, a new resource known as EROL (*Epistolaries of the Republic of Letters*) was created in 2016–17 by a series of postgraduate students working under Dirk van Miert's supervision in Utrecht who came to Oxford on COST-funded STSMs to begin work on a comprehensive bibliography of letter collections printed in the early modern period. In February 2016, Lara Bergers began work on this project by importing all the titles in the bibliographies of Arenhold, Estermann, and Molhuysen into a *Zotero* database.[20] In 2017, this database was expanded with the help of three more STSMs. Mandi Astola managed to import titles automatically from a PDF of the first two volumes of Viola's bibliographies of Italian epistolographies: the first added 4,575 titles and the second another 2,840.[21] Justine Walden had used an STSM to assemble a huge quantity of additional collection-level and item-level data on learned correspondence in Italy, thanks to yet another STSM.[22] With Walden's permission, Astola added another 4,712 titles from Walden's bibliography to EROL. Another STSM allowed Celine Frohn to add another 200 titles of English epistolaries to EROL: because English materials were absent from the existing bibliographies, she extracted titles from the *English Short Title Catalogue* with the help of keyword searches.[23]

Thanks to this collaborative effort, as of January 2019 EROL contains records of 14,160 printed works that include at least one letter, but which typically contain dozens or even hundreds. Nevertheless, EROL remains a work in progress, with much still to be done. The title descriptions in EROL are not standardized. Since they were automatically imported from various repositories with different standards, the titles in EROL are not uniformly structured and are sometimes incom-

---

[18] 'Lijst van geëxcerpeerde boeken voor hs. Ltk. 1643 (apparaat Molhuysen)'. The list, typewritten, with manual additions, is kept in a single copy in the Special Collections department of Leiden University (shelf-mark DOUSA 80 1604).

[19] Corrado Viola, *Epistolari italiani del settecento. Repertorio bibliografico* (Verona: Fiorini, 2004); id., *Epistolari italiani del settecento. Primo supplemento* (Verona: Fiorini, 2008); and id., with Valentina Gallo, eds., *Epistolari italiani del settecento. Repertorio bibliografico. Secondo supplemento* (Verona: QuiEdit, 2015).

[20] Mojet's STSM report, see http://www.republicofletters.net/index.php/emma-mojet-a-database-of-early-modern-epistolaries-by-arenhold-estermann-and-molhuysen/, accessed 20/03/2019.

[21] Astola's STSM report, see http://www.republicofletters.net/wp-content/uploads/2017/05/GP3_Astola_Scientific-Report-revised.pdf, accessed 20/03/2019. Viola's *Secondo supplemento* (2015) is now ready for inclusion as well.

[22] Walden's STSM report, see http://www.republicofletters.net/index.php/justine-walden-the-wealth-of-early-modern-italian-letters/ , accessed 20/03/2019.

[23] Frohn's STSM report, see http://www.republicofletters.net/wp-content/uploads/2017/05/GP3_Frohn_STSM-Report-Frohn.pdf, accessed 20/03/2019.

plete: title-metadata is sometimes missing and sometimes appears in the wrong fields. This does not prevent the use of EROL as a finding aid for identifying epistolaries, but more data cleansing must take place before reliable analysis can be undertaken. In order to pilot such analysis, however, Riccardo Bellingacci used a fifth STSM, supplemented by the first design sprint in Como in the spring of 2016, to experiment with the visualization of EROL data (discussed further in chapter IV.3 below).

Cleaning up of EROL is on the agenda of a project funded through the European Research Council's (ERC) Consolidator programme: 'Sharing Knowledge in Learned and Literary Networks (SKILLNET): the Republic of Letters as a Pan-European Knowledge Society'. Part of SKILLNET's objective is to gain insight into the size, spread, and structure of the republic of letters. Before EROL can do so, however, it will have to include more countries and present a more representative list, geographically speaking. Major deficits are epistolaries printed in France, on the Iberian peninsula, in the Nordic and Baltic regions, and in eastern Europe. Yet, as soon as EROL in its current state has been cleaned up, it will be made available on http://www.skillnet.nl.

2.3 Extracting Item-level Descriptions: Towards a Union Catalogue of Early Modern Learned Correspondence in Print

Assembling collection-level descriptions of printed letters – i.e. of epistolaries – is not an end in itself: its ultimate purpose is to prepare for the extraction of large quantities of individual letter records. One ready-made source of abundant metadata of this kind is the *Corpus Epistolicum Recentioris Aevi* (CERA), which contains digital facsimiles of ninety epistolaries, totalling 55,000 pages of text, published between 1520 and 1770 in Germany and neighbouring countries. Each high-quality page image has been scanned with optical character recognition to provide a rough transcription which was then manually corrected to produce machine-readable XML or HTML files. A treasure trove with potential application for Natural Language Processing, CERA currently lacks item-level metadata, and the number of letters it comprises has not been established.[24]

The most ambitious strategy for extracting item-level descriptions from printed epistolaries, however, is crowdsourcing. This strategy aims to exploit two of the principal advantages of printed over manuscript collections: the fact that huge quantities of them have already been scanned and published online, and the fact that printed texts are far more legible to non-specialists than handwritten ones. To pursue this strategy, a crowdsourcing project known as CEMROL (*Collecting Epistolary Metadata of the Republic of Letters*) was built by the Humanities Lab of Utrecht

[24] In this case, a core component of the necessary metadata could be harvested from Estermann's *Verzeichnis der gedruckten Briefe deutscher Autoren des 17. Jahrhunderts*, discussed in sect. 2.2. above.

University and launched in December 2018 by the SKILLNET project.[25] Drawing in many ways on the technical, logistic, and intellectual knowledge gathered during the COST Action, SKILLNET and CEMROL aim to develop mutually beneficial exchanges with EMLO and other projects that have sprung off the COST Action. The item-level descriptions produced by CEMROL will ultimately be integrated into EMLO as a major contribution towards assembling a catalogue increasingly capable of documenting the full geographical scope and chronological development of the republic of letters.

The challenges in CEMROL are both technical and social, and – as usual with exploratory ventures of this kind – the social challenges outweigh the technical ones. The tasks offered to the public are two-fold: first, to draw boxes around epistolary metadata on the page, and second, to transcribe the text in those boxes: for the time being, the interpretation of these transcriptions is left to experts. The workflow of CEMROL is improving as more people use it and provide feedback. Issues in transcriptions are tackled in brief tutorials, with videos explaining how to manage the system. One obvious but challenging area for development involves semi-automating aspects of the workflow. At present, CEMROL gives the crowd the opportunity to standardize proper names through a drop-down menu containing the names in the authority file of EMLO, but crowd members are likely to make mistakes. Another experiment is with the automated translation of Roman dates into modern dates (dd/mm/yyyy). One problem in this process is that non-experts cannot be expected to indicate whether they think dates are Julian or Gregorian if the style is not indicated. Perhaps semi-automated processes such as those described in chapter II.3 can be implemented instead. The same applies to the standardization of names. Deduplication is another challenge: the mechanisms described in chapter III.2 can be employed to identify likely duplicates and merge them automatically on command, speeding up the process of assembling metadata that reflect actual numbers of letters. Ultimately, the SKILLNET team members are responsible for cleaning up data, and the question of how labour-intensive this quality control is going to be will have to be answered in the course of 2019. A very different, social challenge for CEMROL is to build up a crowd: some projects have found that a gaming component helps to incentivize contributions, by awarding points for every letter marked or transcribed, and developing hierarchies of contributors on that basis; or by closely integrating the most productive members of the crowd with the project through continuous outreach and quick responses.

CEMROL has some obvious advantages over other crowdsourcing projects: apart from the relatively good legibility of type, people can opt for their source language of choice. Moreover, SKILLNET is prioritizing certain editions over others, but crowd members are invited to suggest sources of their own preference. Eventually, the CEMROL environment may also be used to harvest metadata from manuscript sources, although this would require a more extensive instruction and

[25] See https://cemrol.hum.uu.nl/, accessed 20/03/2019.

is expected to draw a much smaller crowd due to the difficulty of the handwriting. Perhaps if the crowd only mark information, that second task of transcription might be automated with the help of *Transkribus* or other software that can be trained to decipher handwritten texts. Another scenario would be to use manuscripts for controlled crowds, such as library cataloguing staff or students taking a course in palaeography. This brings us, finally, to the issue of handwritten letters

3 Letters in Manuscript

3.1 Manuscript Letter Collections: The History and Hazards of an Archival Category

Collecting correspondence metadata – whether at the item or collection level – is considerably more difficult for manuscript letters than for printed ones. The most obvious difficulty is that script is more difficult to read than print, especially when one considers the transnational and multilingual scope of the republic of letters. In addition, collections of printed correspondence come in well-catalogued units reproduced in multiple copies and often distributed throughout many different repositories. Manuscript letters, by contrast, are typically unique, are often uncatalogued, are scattered all across Europe by the very act of sending, and have very often subsequently been incorporated into many different types of holdings and preserved by means of often unpredictable and contingent processes.[26] A brief survey of the vicissitudes of archival collections of manuscript letters is therefore the necessary starting point of a discussion of how to assemble catalogue records of them.

Collections of handwritten letters may consist of autographs or holographs (letters handwritten by their authors), or of apographs (copies made by someone other than the author of the letter). Libraries typically do not assemble letters into a single epistolary category, but organize them instead according to their origins. These origins may be in the *Nachlass* or working papers of an individual, in the archive of a family or institution of which the individual was a part, or indeed in a corpus of correspondence assembled by a collector. As suggested by the recent 'archival turn', insight into the original context in which items were assembled

[26] The correspondence of the great Spanish orientalist Benito Arias Montano (1527–1598) provides a nice example of both of these problems. His surviving letters are preserved in large quantity in the Swedish Royal Library in Stockholm, the Museum Plantin-Moretus in Antwerp, and the Archivo General de Simancas, with smaller numbers scattered from Warsaw to Chicago; and even that subset preserved in Antwerp poses serious palaeographic challenges, on which see Antonio Dávila Pérez, 'Crítica textual en los borradores latinos conservados en el Museo Plantin-Moretus de Amberes', in María Teresa Muñoz García de Iturrospe and Leticia Carrasco Reija, eds., *Miscellanea Latina* (Madrid: Sociedad de Estudios Latinos, 2015), 509–20.

yields information about the status the item had for the collecting person or institute.[27]

Working papers. Throughout their lives, citizens of the republic of letters typically assembled their incoming and outgoing correspondence within larger collections of working papers. As they reached the end of their working lives, many consciously reviewed, ordered, and in some cases purged their letters with a view to archival preservation or posthumous publication (on which see further chapter IV.3).

Family archives. Entirely private individual archives from the early modern period that are not part of a larger archival body only survive under special circumstances (for example, if hidden and forgotten, removed as a result of legal sequestration, or auctioned off). If we possess a scholar's *Nachlass*, it usually means that he has bequeathed it to an enduring group which could care for it after his death. In rare cases normally involving individuals of elevated social status, this might be a family archive. In England, for instance, the foundation of the State Papers Office in 1578 established a trend and many of the most important noble families followed suit.[28] Family archives of the great houses have therefore also survived, sometimes after dispersal in different collections.

Institutional archives. A more common route to survival lay with bequeathing one's letters to a learned institution with which the scholar had some kind of connection. Here the options are numerous. The letters of Christian Daum (1612–1687) are preserved primarily in the school in Zwickau in which he taught.[29] Those of mathematician and cryptographer John Wallis (1616–1703) are preserved in large numbers in the archives of the University of Oxford, over which he presided.[30] A large proportion of the surviving manuscript letters of the Polish astronomer Johannes Hevelius (1611–1687) was obtained after his death by the Observatoire in Paris.[31] The letters sent to several former court librarians in Vienna, among them Sebastian Tengnagel (1563–1636), are kept in the manuscript collection of

[27] See most recently Liesbeth Corens, Kate Peters, and Alexandra Walsham, eds., *Archives & Information in the Early Modern World* (Oxford: Oxford University for the British Academy, 2018). See also Filippo de Vivo, Andrea Guidi, and Alessandro Silvestri, eds., *Archivi e archivisti in Italia tra Medioevo ed età moderna* (Rome: Viella, 2015), and, less recent but still topical, the volume by Michael Hunter, ed., *Archives of the Scientific Revolution: The Formation and Exchange of Ideas in Seventeenth-Century Europe* (Woodbridge: Boydell Press, 1998).

[28] James Daybell, *The Material Letter in Early Modern England. Manuscript Letters and the Culture and Practices of Letter-Writing, 1512–1635* (Basingstoke: Palgrave Macmillan, 2012), 223.

[29] Lutz Mahnke, *Epistolae ad Daumium: Katalog der Briefe an den Zwickauer Rektor Christian Daum (1612–1687)* (Wiesbaden: Harrassowitz, 2003).

[30] Philip Beeley and Christoph J. Scriba, 'The Correspondence of John Wallis' in *Early Modern Letters Online*, Cultures of Knowledge, see http://emlo-portal.bodleian.ox.ac.uk/collections/?catalogue=john-wallis, accessed 20/03/2019.

[31] 'Inventaire détaillé de la correspondance de Johannes Hevelius', Bibliothèque de l'Observatoire, C1/1–16: see https://alidade.obspm.fr, accessed 20/03/2019.

the Austrian National Library.[32] The correspondence of Bernard (1683–1735) and Hieronymus Pez (1685–1762) is preserved primarily in the Benedictine monastery at Melk in Lower Austria, in which they lived out their learned lives.[33] The enormous *Nachlass* of the great German philosopher and mathematician Gottfried Wilhelm Leibniz (1646–1716) was sealed on his death and preserved in the ducal library in Hanover which now bears his name.[34] No less important are the archives of correspondence assembled by learned institutions themselves, such as the over 4,300 early letters in the archive of the Royal Society of London, the catalogue of which is currently being prepared for publication on EMLO.[35] People have often fulfilled a specific role in the institution which preserves their papers – as a chancellor, a bishop, an ambassador, a professor, a secretary, or a host of other roles – and their archives may *also* contain records arising not just from private correspondence, but from the fulfilment of an official duty; yet the distinctive trait of citizens of the republic of letters is precisely that their papers reach *beyond* those official roles, addressing questions and reaching out to people beyond the remit of their official job description.

Collectors. Very different from all of the foregoing arrangements is the case of letters gathered by early modern collectors, often scholars themselves. When searching for the letters of scholars who left no intact *Nachlass*, these collections provide an obvious starting point. The largest letter collections of this kind are often named after their collectors. Prominent examples include Hamburg's Uffenbach-Wolf collection, a collection of thousands of autograph letters assembled by the Frankfurt book collector Zacharias Conrad von Uffenbach (1683–1734) on his many travels and expanded after his death by the polyhistor Johann Christoph Wolf (1683–1739); Erlangen's Trew collection, consisting of hundreds of letters brought together by the Nuremberg physician Christoph Jacob Trew (1695–1769); the Danish Royal Library's Thott collection, which formed part of the library of the Danish Count Otto Thott (1703–1785); the collection of 38,000 letters accumulated by and now named after the Swedish physician Erik Waller (1875–1955), now in the Uppsala University Library; the many thousands of letters gathered by Pierre (1582–1651) and Jacques (1591–1656) Dupuy, now in the Bibliothèque Nationale de France; the British Library's Burney collection, which formed part of the 13,000 items of the library of the classical scholar Charles Burney junior (1757–1817); and Leiden University Library's Papenbroeck collection, deriving from the

[32] See https://geschichtsforschung.univie.ac.at/forschung/oorpl/, accessed 20/03/2019, as well as chapter III.3. Other correspondences kept at the Österreichische Nationalbibliothek include those by Hugo Blotius (1533–1608) and Peter Lambeck (1628–1680).

[33] See http://www.oapen.org/search?identifier=445402, accessed 20/03/2019 (vol. 1); http://www.oapen.org/search?identifier=576952, accessed 20/03/2019 (vol. 2); as well as https://unidam.univie.ac.at/nachlass/195, accessed 20/03/2019 (Pez papers).

[34] See, most recently, Howard Hotson, 'Leibniz's Network', in Maria Rosa Antognazza, ed., *The Oxford Handbook of Leibniz* (Oxford: Oxford University Press, 2018), 563–90.

[35] The Royal Society, GB 117: early letters from correspondents in natural philosophy sent to the Royal Society and its fellows (1613, 1642, 1651–1740).

collector Gerard van Papenbroek (1673–1743).[36] Fortunately, these collections all have printed catalogues. Some of these, such as the Waller collection, have turned digital as online databases.[37] The records of others were integrated into digital meta-catalogues (e.g. the letters in Erlangen's Trew collection are recorded in *Kalliope* and those of the Thott collection are in the Royal Library's *Brevbase* (see below, under 'Denmark' in 3.2). Another way of making these descriptions of collections available is by putting them on the Internet as searchable PDFs, such as the Bibliothèque Nationale de France did with Dorez's inventory of the Dupuy collection.[38] Libraries also hold letters that are not part of particular collections: for instance, Leiden University Library shelves numerous letters under the class mark 'BPL', which stands for 'Bibliotheca Publica Latina'.

In order to understand how best to use such collections, it is often vitally important to understand the objectives of their collectors. A small example of how the history of an archive is shaped by a collector's agendas is provided by MS 983 in the Utrecht University Library. This manuscript, in the shape of a bound book holding copies of some 200 letters, was drawn up by the Utrecht antiquarian Arnoldus Buchelius (1565–1641) in the first half of the seventeenth century. A substantial number of these were originally written by or to Buchelius himself, but there are many others that were exchanged between other people. Buchelius often chose not to copy out the entire letter, but limited himself to providing metadata and a brief outline. What were Buchelius's criteria of selection? Did he merely copy out everything that he could cast eyes on? Or is there a particular strategy involved, which has led to a particular set of letters that gives modern historians insight into the status of representativity of the information held in this collection? Inspired by the COST Action, Dirk van Miert and five research master students at Utrecht University (Jan Fongers, Anne Haak, Erell Smith, Tirreg Verburg, and Chantal van der Zanden), in January 2018 entered the metadata of all letters mentioned in this manuscript into a *NodeGoat* area.

Mapping these records (fig. 1) shows that almost all of these letters were written either to or from Utrecht and that three-quarters of the letters in the collection were addressed to people living in Utrecht. This suggests that Buchelius collected much material in Utrecht itself and that on his travels he copied out those letters that were written by or to his fellow citizens.

---

[36] Nilüfer Krüger, *Supellex epistolica Uffenbachii et Wolfiorum*, 2 vols. (Hamburg: Hauswedell, 1978); Eleonore Schmidt-Herrling, *Die Briefsammlung des Nürnberger Arztes Christoph Jacob Trew, 1695–1769, in der Universitätsbibliothek Erlangen* (Erlangen: Universitätsbibliothek, 1940); Leon Dorez [vol. 3: Suzanne Solente], *Bibliothèque nationale. Catalogue de la collection Dupuy*, 3 vols (Paris: Leroux, 1899-1928). The Thott collection is integrated into the Danish Royal Library's online letter catalogue *Brevbase:* see under 3.2. In 3.2. the Waller collection is referenced under the heading of Sweden.

[37] See http://www.ub.uu.se/finding-your-way-in-the-collections/selections-of-special-items-and-collections/waller-collections/waller-manuscript-collection/, accessed 20/03/2019.

[38] Available online as a PDF: see http://visualiseur.bnf.fr/Visualiseur?Destination=BnF&O=NUMM–209160, accessed 20/03/2019.

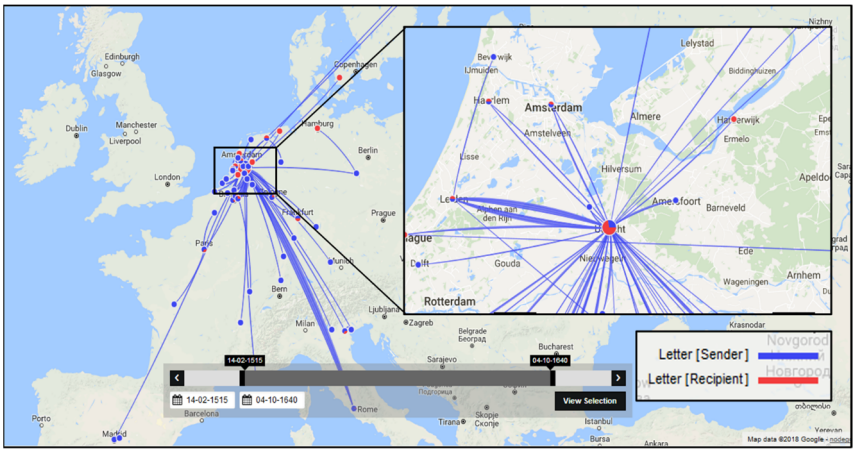


Figure 1: *NodeGoat* geographical visualization of the metadata in Utrecht University Library, MS 983

A social network visualization (fig. 2) shows that Buchelius himself was within three degrees of separation from the most prolific letter-writers in the collection: although there are numerous isolated interactions between other individuals, only a cluster of them – surrounding Theodorus and Lambertus Canterus – generated a significant number of letters. Looking at the wider context of Buchelius's work, it can be no coincidence that Buchelius authored a *Traiecti Batavorum descriptio* and a notebook with annotations on Utrecht families. In short, this particular collection was part of a personal archive assembled to serve a particular interest in local history, and the preservation of these particular letters was meant to provide grounds for asserting Utrecht's importance as a significant node within the republic of letters.

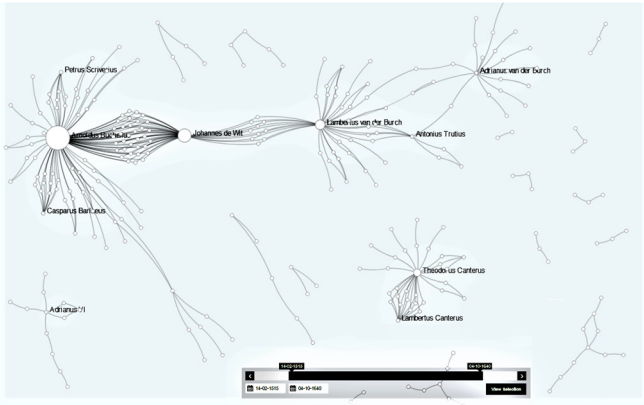


Figure 2: *NodeGoat* social network visualization of the metadata in Utrecht University Library, MS 983

Another good example is the catalogue of the collection of Bartolomeo Gamba (1766–1841), a sub-keeper at the Biblioteca Marciana in Venice who amassed some 4,000 manuscript letters to and from eminent early modern Italian scholars, many documenting relations between the University of Padua and Viennese court physicians. Material on this once coherent collection has now been dispersed between the Österreichische Nationalbibliothek in Vienna and the Biblioteca Museo Civico in Bassano del Grappa (which own the early modern correspondence) as well as the Marciana Libray in Venice (which holds Gamba's own correspondence about his collection). In order to reassemble virtually Gamba's collection of early modern letters, Vittoria Feola has put together a team of collaborators in all of these institutions and funding from the University of Padua (DiSSGeA), the Medical University of Vienna, the Gerda Henkel Stiftung, with supplements from CofK and COST Action IS1310.[39]

Archives and libraries. Through all these labyrinthine routes, many letters have found their way into omnibus public libraries and archives; and here further distinctions must be made. Archives, to simplify slightly, keep records arising from administrative activities, while libraries store printed books. Consequently, libraries and archives describe their holdings in different ways.

A library catalogue typically focuses on the bibliographical entity, usually the book and its bibliographical metadata: author, title, publisher, place, and year. Letters are bound together in volumes and catalogued in many different configurations: whole volumes can be described in a generic way (for example, by reference to the main recipient, perhaps a range of dates, or with additional references to places and principal correspondents), or individual letters can be itemized separately. Letters can also be part of manuscripts that are *not* letter collections. This often happens when letters become part of the daily working apparatus of scholars, and are thus organized together with other materials in thematic dossiers.

Archive inventories structure their directories hierarchically, proceeding from the general description of a holding (e.g. the papers of a given person, or the output of a given chancery over a period of time) to its individual parts (such as correspondence, working papers, and works in manuscript, or in the products of various bureaus within a chancery). In such an archival scenario, one should not be surprised to find inventory entries that point to 'correspondence', without providing information on quantity, people involved, content, or even the time-span. The notorious description category 'miscellanea' may point to the most interesting and unknown materials, but it also entails a lot of painful manual cataloguing work.

[39] More on this project can be found on EMLO (http://emlo-portal.bodleian.ox.ac.uk/collections/?catalogue=bartolomeo-gamba) and the CofK blog (http://www.culturesofknowledge.org/?p=5810) as well as the STSM report on the COST website (http://www.republicofletters.net/wp-content/uploads/2015/11/feola_stsm.pdf), all accessed 20/03/2019.

3.2 Sources of Epistolary Metadata on Manuscript Letters: Challenges and Opportunities

With contributions from Ivan Boserup, Clizia Carminati, Per Cullhed, Andreas Fingernagel, Antonio Dávila Pérez, Ad Leerintveld, Gerhard Müller, Alexa Renggli, Patryk Sapala, Justine Walden, and Axel E. Walter

Given the complexity of the archival history of early modern letter collections in Europe, it is not surprising that the modes of cataloguing these materials and of publishing those catalogues online also varies greatly from one institution and country to another. The amount of data available online is now enormous, but finding it is not easy and dealing with it once found is more difficult still, given the great variety of cataloguing styles and standards. In order to provide a preliminary overview of existing resources, Working Group IV organized a series of presentations at the first Action conference in Oxford in March 2015. These presentations, updated with reference to further discussions, provide the basis of this section. This brief guide to available resources does not aim to be comprehensive: it merely seeks to highlight some of the differing approaches taken to date and some of the challenges and opportunities facing scholars eager to track down manuscript letters. It begins with a survey of the high-level resources available: international resources, union catalogues at the national level, and catalogues of national institutions. It concludes with a survey of some printed inventories of early modern learned correspondence, to which might be added the constantly proliferating digital resources being created by individual institutions.

International resources

The senior project under this international heading is Paul Oskar Kristeller's magisterial *Iter Italicum*, published in six volumes between 1963 and 1997. The title of the work is misleading: from the third volume onwards this repertory expanded to include humanist manuscripts in Austria, Great Britain, Greece, Hungary, Israel, Japan, Liechtenstein, Luxembourg, Malta, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, Romania, South Africa, Spain, and Yugoslavia. As Kristeller noted in 1989, the difficulty of compiling the work was compounded by the ever-changing library landscape. By the time of publication, some private libraries had disappeared while others had newly appeared. In others repositories, the original shelf-mark system had been superseded by newer ones. Where institutions had already provided detailed catalogues, Kristeller was content to draw attention to these, rather than replicate their findings. Given the enormous scope of the enterprise he had to remain content with collection-level descriptions, rather than itemizing letters individually. Despite these inevitable limitations, Kristeller's *Iter Italicum* remains a vital printed resource for letter collections in Italy and elsewhere. The

use of these unwieldy volumes has recently been facilitated by the production of a digitized version available on CD and online via subscription.[40]

More recently, several digital projects have begun creating union catalogues and archives of early modern learned correspondence which are international in scope. The oldest major resource of this kind is the *Electronic Enlightenment* (EE). Uniquely among major digital resources in this field, EE provides online access to a huge collection of early modern learned letters previously published by many different presses in hard-copy editions which are still under copyright, and for this reason the full data set is only available on subscription. As of the autumn of 2017, EE included 77,251 letters and other documents as well as over 10,000 biographical entries linking people across Europe, the Americas, and Asia from the early seventeenth to the mid-nineteenth centuries. The origins of the project in the Voltaire Foundation in Oxford help explain its particular focus on the French Enlightenment. Now transferred to the Bodleian Library, EE is joining forces with the *Oxford Text Archive*, and discussions are underway to expand collaboration with other Oxford resources in the field.[41]

Early Modern Letters Online (EMLO) is a decade-old experiment in the collaborative population of a union catalogue of the republic of letters, based in Oxford, funded by the Mellon Foundation via the *Cultures of Knowledge* (CofK) project directed by Howard Hotson.[42] At the core of the project are individual user accounts which allow contributors to collaborate with the project's digital editor, Miranda Lewis, and her team of 'digital fellows' in the curation of catalogues to a high and uniform standard prior to publication. Each catalogue is published with a separate page describing the chief correspondent, the scope of the correspondence, and the collaborators, institutions, and funding bodies involved in its creation. Individual records are linked to authority files for people and places, and to further resources on and off EMLO, often including abstracts, letter texts, and digital images of manuscripts and early modern printed books. From the outset, the project has set out to serve the international community working on early modern learned epistolary exchange: in the first 100 catalogues published on EMLO, the thirty-nine Brit-

[40] *Iter Italicum: A Finding List of Uncatalogued or Incompletely Catalogued Humanistic Manuscripts of the Renaissance in Italian and Other Libraries*, compiled by Paul Oskar Kristeller (London: The Warburg Institute; Leiden: E. J. Brill, 1963–92). CD ROM: consultant ed. Luciano Floridi (Leiden: Brill, 1995). Online version accessible via the Iter: Gateway to the Middle Ages and Renaissance (https://www.itergateway.org/resources/iter-italicum) and Brill (https://brill.com/view/serial/KRIS), both accessed 20/03/2019.

[41] *Electronic Enlightenment*: see http://www.e-enlightenment.com/, accessed 20/03/2019. On its origins, see Nicholas Cronk and Glenn Roe, 'Electronic Enlightenment', in Simon Burrows and Glenn Roe, eds., *Digitizing Enlightenment: Digital Humanities and the Transformation of Eighteenth-Century Studies*, Oxford University Studies in the Enlightenment (Liverpool: Liverpool University Press / Voltaire Foundation, forthcoming 2020).

[42] EMLO: http://emlo.bodleian.ox.ac.uk/home. CofK: http://www.culturesofknowledge.org/, both accessed 20/03/2019. On its origins, see Howard Hotson, 'Cultures of Knowledge in Transition: Early Modern Letters Online as an Experiment in Collaboration, 2009–2018', in Burrows and Roe, eds., *Digitizing Enlightenment* (forthcoming).

ish correspondences were significantly outnumbered by sixty-one continental ones. In the course of the COST Action, EMLO was increasingly adopted by the international community as a shared resource and a natural home for metadata on early modern learned correspondence.

Another valuable international resource for the field is the online listing of *Sources for Early Modern Letters* first developed by the Warburg Institute during the project editing Scaliger's correspondence and now transferred to Utrecht University and maintained by the SKILLNET project.[43] While this naturally concentrates on individual projects, it also provides a brief listing of printed sources available elsewhere.

National resources

Austria. The Austrian National Library (in common with many other national libraries) contains not only individual manuscript letters bound with other material but also bundles of letters (*commercium litterarum*) bound together in volumes solely devoted to correspondence. These bundles contain a variety of materials, such as an individual scholar's attempt to construct a personal archive, an institution's gathering of correspondence to and from an individual scholar, or an institution's collection of letters to and from a variety of correspondents. Andreas Fingernagel drew attention to the fact that in the second half of the nineteenth century a project was started to collect the autographs and seal impressions in the Austrian national collection. This project, now online, includes c. 300,000 autographs, and enables scholars to search both for individual letters and letter collections.[44]

Croatia. CroALA (*Croatiae Auctores Latini Collectio Electronica*) presently includes 449 documents with a date range between 976 and 1984.[45]

Denmark. The Danish Royal Library's letter catalogue originated in an alphabetical card catalogue but is now available online under the title *Brevbase.* As Ivan Boserup demonstrated, it includes item-level descriptions of over 200,000 letters, dating from the seventeenth and eighteenth centuries, which are in the manuscript collection of the Royal Library. Such institutional catalogues now function in practice as a 'meta-catalogue' – an umbrella-search engine of the multiplicity of separate collections within which letters have been acquired and preserved by the library.

Estonia. Kristi Viiding is the principal investigator of a project on the correspondence of the well-known Livonian humanist David Hilchen (1561–1610). His correspondence of c. 800 letters is in the process of being edited and the metadata

[43] See https://warburg.sas.ac.uk/research/completed-research-projects/scaliger/sources-early-modern-letters and https://skillnet.nl/sources/, both accessed 20/03/2019.

[44] The search screen for the Austrian National Library can be found at https://search.onb.ac.at/primo-explore/search?sortby=rank&vid=ONB&lang=de_DE, accessed 20/03/2019.

[45] See http://www.ffzg.unizg.hr/klafil/croala/, accessed 20/03/2019.

has already been included in EMLO.[46] This type of initiative is representative of a multitude of individual editorial projects across Europe. Here the focus is less on the institutional holdings and more on the individual scholar.

Finland. An important resource based in Finland is *The Corpora of Early English Correspondence* (CEEC400). Created by a partnership between the Academy of Finland and the University of Helsinki, this project has amassed a large body of letters in English in order to test how methods created by sociolinguists studying present-day languages could be applied to historical data. The CEEC family of corpora currently covers 400 years from 1400 to 1800, and is being united into a structure whole consisting of over 5 million words.[47]

France. The catalogue of the huge Collection Dupuy, which includes thousands of letters, was printed at the end of the nineteenth century by Leon Dorez in two volumes and is now online.[48] It represents just the tip of the iceberg of early modern correspondence available in the Bibliothèque Nationale de France. The online *Catalogue Collectif de France* (CCFr) allows the user to search simultaneously not only the general catalogue of the Bibliothèque Nationale but also the catalogues of digitized municipal library collections.[49] Despite these innovations, locating early modern correspondence at item level in French libraries remains a challenge, not least due to the disparate nature of the municipal collections.

Germany. An effort to integrate the manuscript catalogues of hundreds of libraries and archives online is the *Kalliope Verbundskatalog* (*Kalliope* Union Catalogue), which started from a collection of 1.2 million card files. As Gerhard Müller noted, *Kalliope* is a constantly growing entity which now holds almost 2,240,000 million records of correspondence (134,000 of which are from before 1800), from over 19,100 collections held in 950 institutions. It is, therefore, an invaluable point of departure for scholars interested in tracking correspondence in early modern Germany. Likewise, scholars interested in letter metadata from the German context can draw on *corresp.Search*, a web application created by Stefan Dumont at the Berlin-Brandenburgische Akademie der Wissenschaften, which is closely connected with the creation and curation of the respective TEI elements (in particular, <corresp>; see I.3 and III.5). Currently *corresp.Search* offers metadata for roughly 47,000 letters (between 1510 and 1991), including 10,390 before 1800, and either provides bibliographical information to retrieve the respective printed edition, or links to a potential digital edition.[50]

---

[46] See http://emlo-portal.bodleian.ox.ac.uk/collections/?catalogue=david-hilchen, accessed 20/03/2019.

[47] Project home page: http://www.helsinki.fi/varieng/domains/CEEC.html. Further information: http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/index.html; both accessed 20/03/2019.

[48] Solente, *Catalogue* [= Dorez, *Catalogue*, vol. 3], a PDF is available at http://visualiseur.bnf.fr/Visualiseur? Destination=BnF&O=NUMM–209160, accessed 20/03/2019.

[49] See https://ccfr.bnf.fr, accessed 20/03/2019.

[50] https://correspsearch.net/index.xql?l=en, accessed 20/03/2019.

Hungary. The Manuscripts Department of the National Széchényi Library (OSZK) includes a collection of some 30,000 letters, catalogued card-files, dating from the mid-sixteenth century to the present day (with the vast majority relating to the nineteenth and twentieth centuries). For the early modern period, Gábor Almási has counted some 870 letters (copies and autographs) in the holdings of the OSZK.[51]

Ireland. While there are numerous print publications of individual correspondences, few online resources are specifically devoted to early modern correspondence held in Irish repositories. One of the most up-to-date is emerging from the project *The Reception and Circulation of Early Modern Women's Writings, 1550–1700* (RECIRC); but, as the name suggests, this project does not limit itself to Irish writers alone or solely to correspondence.[52] The online 'Sources: A National Library of Ireland Database for Irish Research' is the initial entry point for scholars interested in tracking material about Irish correspondents;[53] but given the colonial and religious history of Ireland, archives and libraries in Ireland, the United Kingdom, and continental Europe hold important material as well. As elsewhere, item-level description in catalogues of manuscripts is not always available and many letters may only be found by painstakingly trawling through manuscripts.

Italy. A pioneering survey of Italian materials is contained in Kristeller's *Iter Italicum*, which is treated above since it is actually international in both origin and scope. More recently, a variety of digital resources for Italian correspondences have begun to proliferate. As Justine Walden noted, the bulk of early modern Italian letters in manuscript are scattered throughout many different libraries, but others can be found in the Archivi di Stato (state archives). Researchers can search the holdings of all 103 state archives through a single interface;[54] but the difficulty of tracking down correspondence to or from a specific individual in this way is compounded by the tendency of these catalogues to identify holdings by the name of their principal collector, rather than by subject or the names of correspondents.

Some of these problems are now being overcome by a new generation of collaboratively populated online resources. A prime example is the online catalogue and archive of Italian literary correspondences in the early modern period, known as *Architet*.[55] As Clizia Carminati explained, *Architet* concentrates on letters by and to Italian writers, and letters that relate to Italian literature and culture. The database not only provides the names of the sender(s), recipient(s), date, place of the sender(s) and of the recipient(s), but also all names and books quoted in the letter, things of relevance, the incipit and the source (i.e. the place where the original manuscript letter is now kept or the reference edition, if published). As Carminati

[51] These can be found in OSZK Quart. Lat. 783, Quart Lat 998. Fol. Lat. 1394, Fol. Lat. 1647. Fol. Lat. 1661, Fol. Lat. 1673. Quart. Lat. 1621. Fol. Germ 594 and 'Levelestár'.

[52] See http://recirc.nuigalway.ie/, accessed 20/03/2019.

[53] See http://sources.nli.ie/, accessed 20/03/2019.

[54] See http://siusa.archivi.beniculturali.it/cgi-bin/pagina.pl, accessed 20/03/2019.

[55] See http://www.archilet.it, accessed 20/03/2019.

noted, each letter is not merely 'filed'; it is studied, so that the content may be understood by anyone visiting the website. Moreover, catalogue records can be enhanced with JPG images of letter texts when available or can link out to images of letters on other open access websites, such as Google Books. The database is openly accessible and constantly increased in collaboration with a large community of contributors. The objective of the project is to place authors and texts into a context, and to document literary choices and cultural relationships, thereby revealing new perspectives on early modern history and the history of literature, ideas, religious thought, and art.

Lithuania. Axel E. Walter, in his presentation on the collection and cataloguing policies of different institutions in Lithuania, highlighted another cataloguing problem: in many cases more information is available about individual letters, whether in print or manuscript, than there is concerning whole letter collections. Searching is made more difficult as different cataloguers preferred different standards. A union catalogue comprehending different collections is therefore unavailable.

Netherlands. Another major resource is the Dutch online union catalogue known as the *Catalogus Epistularum Neerlandicarum* (CEN).[56] As Ad Leerintveld explained, CEN contains approximately 2 million records of letter (many post-1800) in the collections of the Royal Library at The Hague; four university libraries (Amsterdam, Leiden, Groningen, and Utrecht); and several other significant collections (including the Stadsarchief en Athenaeumbibliotheek in Deventer, the Tresoar Library in Leeuwarden, the Zeeuwse Bibliotheek in Middleburg, and the Letterkundig Museum and Museum Meermanno in The Hague). CEN thus provides an essential starting point for the study of early modern scholarship in the Netherlands. However, as Dirk van Miert subsequently observed, CEN also has significant drawbacks. First, this union catalogue is not comprehensive: although the holdings of several major university libraries are listed in CEN, many archives are not, including the Nationaal Archief in The Hague. Second, CEN is not yet freely available online: only a library that contributes to CEN can grant access to it, and most scholars working outside the Netherlands have no access to this valuable resource. A deeper problem is that a union catalogue of this kind is only as good as the records contributed to it by partner institutions, which vary in quality. For instance, if several letters are preserved written by the same sender to the same recipient and classified under the same shelf-mark, some of the catalogues assembled by CEN aggregate all of these letters into a single record. As a consequence, although an online search currently identifies slightly over 50,000 hits for the period up to 1800, the actual number of letters is probably two or even three times as large. It should also be noted that the CEN does not limit itself to Dutch letters but instead gives details of letters held in Dutch collections. CEN aims to overcome the first two of these problems in the future by including more partner li-

[56] The URL is: picarta.pica.nl/DB=3.23, accessed 20/03/2019.

braries and archives and by offering open access to this valuable catalogue worldwide.

Poland. As Patryk Sapala explained, the vicissitudes affecting the survival of letter collections in national and private libraries are nowhere more painfully visible than in the case of Poland. The large-scale dispersal of Polish collections began during the second partition of Poland (1795), when several large collections were taken to the Imperial Public Library in St Petersburg. During the latter part of the nineteenth century, private collectors sought to fill the gap by preserving correspondence at libraries such as the Ossolinski Library (Zakład Narodowy im. Osslińskich) in Lviv; but this effort was piecemeal and led to a proliferation of individual catalogues. To make matters worse, the material that had been repatriated by the Russians was later burnt by the Nazis during the Second World War and further dispersal and rearrangement took place during the Communist era. In 2003, a fundamental finding aid appeared in the survey of surviving manuscript collections in Poland, edited by Danuta Kamolowa and Teresa Sieniatecka.[57]

Portugal. 'Post Scriptum' is an online listing of private letters written in Portugal and Spain during the early modern period.[58] It is designed to enhance the linguistic study of a range of private communications in the Iberian peninsula. Although the scope of this digital archive is broader than learned correspondence, 'Post Scriptum' provides an interesting starting point for scholars interested in the republic of letters in early modern Portugal and merits further development.

Spain. While Kristeller's *Iter Italicum* includes a survey of Renaissance material in Spanish libraries, in many respects it has been superseded by the proliferation of online resources. Antonio Dávila Pérez drew attention to a number of these that are particularly useful for tracking early modern learned correspondence in Spanish libraries and archives: these include *Hispana*, which lists 617 digital collections;[59] *Pares*, a portal containing digitized manuscripts from the most important Spanish archives;[60] and the *Catálogo colectivo del patrimonio bibliográfico español*, which covers books printed in Spain from the fifteenth century onwards.[61] The latter catalogue is complemented by the *Biblioteca virtual del patrimonio bibliográfico español,* a digital library which includes manuscripts as well as printed books.[62] These are important resources, but Dávila Pérez also stressed the continuing need for in-depth archival research, since the catalogues of manuscripts on which these online resources are based often include vague, imprecise, inaccurate or incomplete information. Out-

[57] Danuta Kamolowa and Teresa Sieniatecka, eds., *Zbiory rękopisów w bibliotekach i muzeach w Polsce* [Manuscript collections in Poland] (Warsaw: Biblioteka Narodowa, 2003; reissued 2014); Tomasz Makowski and Patryk Sapala. *Rękopisy w zbiorach kościelnych* [Manuscripts in church collections] (Warsaw: Biblioteka Narodowa, 2014).

[58] See http://www.clul.ulisboa.pt/en/10-research/662-p-s-post-scriptum, accessed 20/03/2019.

[59] See http://hispana.mcu.es, accessed 20/03/2019.

[60] See http://pares.mcu.es/ParesBusquedas20/catalogo/search, accessed 20/03/2019.

[61] See http://catalogos.mecd.es/CCPB/ccpbopac/noserver.htm?dir=/CCPB/ccpbopac, accessed 20/03/2019.

[62] See http://bvpb.mcu.es/es/inicio/inicio.do, accessed 20/03/2019.

side Spain, the *Spanish Republic of Letters* (SRL) project is working to create correspondence catalogues needed to put early modern Spanish intellectuals on the emerging map of the European republic of letters.[63] Coordinated in the University of Windsor, Canada, by Guy Lazure, Cal Murgu, and Dave Johnston, SRL currently contains fifty correspondence catalogues containing 3,559 letters.

Sweden. Per Cullhed of the University Library of Uppsala highlighted two valuable union catalogues recently developed in Sweden. *Opac Libris* brings together records from several relevant initiatives, including the Waller collection (containing 38,000 manuscripts, mainly letters, on the history of science and medicine), the catalogue of c. 5,000 letters to and from the great Swedish botanists Carl Linnaeus (1707–1778), and an *Alba-Amicorum* project started in 2015.[64] The ALVIN portal is a digital repository for archives, images, books, manuscripts, maps, objects, sound, video, musical material, and software.[65] Though ALVIN's remit is much broader than early modern correspondence, it represents an interesting innovation in the provision of manuscript correspondence online. Rather than waiting until all its letters have been properly catalogued, ALVIN allows digital images of letters to be published online first, and metadata to be added later, potentially through scholarly crowdsourcing. This arrangement established an interesting precedent, which should be studied by other institutions with large collections of uncatalogued correspondence.

Switzerland. The platform *e-manuscripta* provides free access to digitized manuscript material from Swiss libraries and archives.[66] This impressive portal was developed and financed cooperatively by three major Swiss libraries: the Zentralbibliothek Zürich, the Universitätsbibliothek Basel, and the ETH-Bibliothek. Launched in 2013, it continues to be managed collaboratively by these three institutions in conjunction with the Swiss National Library. Many other institutions have also made their stock available, thereby expanding the range of material hosted on the portal. As Alexa Renggli noted, the range of manuscripts included is very broad, including music, maps, drawings, and photographs; yet correspondence of individuals and institutions is prominent: of the more than 75,800 items currently available, over 33,000 are letters and over 20,000 of these letters are dated before 1800. Best of all, each record is accompanied by a high-resolution image of the manuscript, provided with a permanent link ensuring long-term access; and these images can not only be studied online and embedded in other digital resources via the IIIF protocol but also downloaded for study as PDF files. A major innovation is the installation of a transcription tool, and future crowdsourcing projects are also envisaged.

[63] See http://cdigs.uwindsor.ca/srl/, accessed 20/03/2019.

[64] See http://libris.kb.se/, accessed 20/03/2019.

[65] See http://www.alvin-portal.org/alvin/home.jsf?dswid=–4620, accessed 20/03/2019.

[66] See https://www.e-manuscripta.ch/, accessed 20/03/2019.

United Kingdom. The richest single source of edited texts of early modern English correspondence is *Oxford Scholarly Editions Online* (OSEO), which provides online access to nearly thirty editions previously published in hard copy by the Oxford University Press (OUP).[67] Navigating these machine-readable edited texts is facilitated by pre-structured metadata, which the OUP have begun passing to *Cultures of Knowledge* for extraction, curation, enhancement, and publication on EMLO. Published catalogues arising from this material include Philip Sidney, Lady Anne Conway, Thomas Hobbes, Elisabeth Stuart, and Elias Ashmole. Future acquisitions include John Locke, Samuel Pepys, Joseph Addison, James Boswell, Samuel Johnson, Adam Smith, Jonathan Swift, and Lady Mary Wortley Montagu.

Early Modern Letters Online (discussed above under international resources) provides open access data on over 100 early modern correspondences, British and continental. In addition, it is also generating catalogues on some of the richest deposits of learned letters in the UK. Its founding catalogue was a digitized and curated version of the 'Index of Literary Correspondence', an index-card file of 48,668 letters found within 487 volumes of early modern manuscript correspondence in the Bodleian Library.[68] Currently in preparation is a curated catalogue of c. 4,300 early letters (before 1740) in the archive of Britain's premier scientific society, the Royal Society of London.[69] Work has also begun on preparing a catalogue of over 300,000 letters in the English State Papers for the Tudor and Stuart periods: the catalogue of Tudor letters has already benefited from extensive work by Ruth and Sebastian Ahnert in the context of the AHRC-funded project *Tudor Networks of Power*, while the remaining work will be conducted under a second AHRC grant for a project entitled *Networking Archives*.[70]

The Helsinki-based *Corpora of Early English Correspondence* is discussed above under Finland.

United States of America. The most relevant union resource for the republic of letters from this quarter is Founders Online undertaken by the US National Archives, which provides access to full-text versions of the correspondence and other writings of six major shapers of the United States: George Washington (1732–1799), Benjamin Franklin (1706–1790), John Adams (1735–1826), Thomas Jefferson (1743–1826), Alexander Hamilton (1757–1804), and James Madison (1751–1836).[71] Over 181,000 fully annotated and searchable documents are included from the authoritative *Founding Fathers Papers* projects, ranging from 1706 to 1836.

Another American initiative with a long-term impact on the field is the *Mapping the Republic of Letters* project at Stanford University. Working with Stanford's Hu-

[67] See http://www.oxfordscholarlyeditions.com/, accessed 20/03/2019.

[68] See http://emlo-portal.bodleian.ox.ac.uk/collections/?catalogue=bodleian-card-catalogue, accessed 20/03/2019.

[69] The Royal Society, GB 117: early letters from correspondents in natural philosophy sent to the Royal Society and its fellows (1613, 1642, 1651–1740).

[70] See https://networkingarchives.org/, accessed 20/03/2019.

[71] See https://founders.archives.gov/, accessed 20/03/2019.

manities + Design Lab, this project has innovated above all in the development of new tools for analysing and visualizing digital data (most famously in the case of *Palladio*); but a range of research projects on the long eighteenth century is also generating significant quantities of high-quality data, some of which has already been published on EMLO.[72]

Published inventories of individual scholars

In preparation for modern editions of ego-correspondences, several inventories have been compiled and published in print which could yield coherent collections of high-quality digital metadata. In 1968 appeared the ground-breaking *Inventaire* of the correspondence of Lipsius, compiled by Gerlo and Vervliet. This catalogue is still a vade mecum for the editors of Lipsius's correspondence, although the inventory has been corrected and supplemented extensively, to the point of being superseded by the volumes of Lipsius's correspondence that are now in print.[73] Meanwhile, the school of Paul Dibon, Hans Bots, and Eugenie Bots-Estourgie, carried forward by the Amsterdam Institute for Neo-Latin and Neo-Philology, published extensive inventories of such letter-writers as André Rivet (1971),[74] Johannes Fredericus Gronovius (1974),[75] Caspar Barlaeus (1978),[76] and later on Gerard Vossius (1993),[77] Theodorus Janssonius van Almeloveen (1997),[78] and Hadrianus Junius (2010),[79] listing all the different manifestations of each single letter. The latest offspring was the long-awaited inventory of the correspondence of Petrus Scriverius, started in the early 1980s by the institute's director, Pierre Tuyman, and published in 2018.[80] Similar projects include the inventories of the correspondenc-

[72] See http://republicofletters.stanford.edu/; http://hdlab.stanford.edu/; and http://emlo-portal.bodleian.ox.ac.uk/collections/?catalogue=athanasius-kircher, all accessed 20/03/2019.

[73] Aloïs Gerlo and Hendrik D.L. Vervliet, *Inventaire de la correspondance de Juste Lipse 1564–1606* (Antwerp: Éditions scientifiques Erasme, 1968). A digital catalogue based on these resources is well under way in EMLO.

[74] Paul Dibon, *Inventaire de la correspondance d'André Rivet (1595–1650)* (The Hague: Nijhoff, 1971).

[75] Paul Dibon, Hans Bots, and Eugénie Bots-Estourgie, *Inventaire de la correspondance de Johannes Fredericus Gronovius (1631–1671)* (The Hague: Martinus Nijhoff, 1974).

[76] Koert van der Horst, *Inventaire de la correspondance de Caspar Barlaeus (1602–1648)* (Assen: Van Gorcum, 1978).

[77] G. Anton C. van der Lem and Cornelis S. M. Rademaker, *Inventory of the Correspondence of Gerardus Joannes Vossius (1577–1649)* (Assen: Van Gorcum, 1993); online version at http://emlo-portal.bodleian.ox.ac.uk/collections/?catalogue=gerardus-joannes-vossius, accessed 20/03/2019

[78] Saskia Stegeman, 'Patronage en Dienstverlening: het netwerk van Theodorus Janssonius van Almeloveen (1657–1712) in de republiek der letteren', Doctoral Dissertation, Radboud Universiteit, Nijmegen, 1997; ead., *Patronage and Services: The Network of Theodorus Janssonius van Almeloveen (1657–1712)* (Amsterdam: APA-Holland University Press, 2005), 538–72.

[79] Chris Heesakkers and Dirk van Miert, 'An Inventory of the Correspondence of Hadrianus Junius (1511–1575)', *Lias. Journal of Early Modern Intellectual Culture and Its Sources* 37:2 (2010): 109–268, see https://doi.org/10.2143/LIAS.37.2.2115446.

[80] Michiel Roscam Abbing and Pierre Tuynman, *Petrus Scriverius Harlemensis (1576–1660). A Key to the Correspondence, Contacts and Works of an Independent Humanist* (Leiden: Foleor Publishers, 2018).

es of Jean Bouhier (1975),[81] Pasquier Quesnel (1989),[82] Jean Henri Samuel Formey (2003),[83] d'Alembert (2009),[84] and the gigantic inventory in six volumes of the correspondence of Jean-Alphonse Turrettini (2009).[85] For Germany, a recent inventory is the one of Johann Valentin Andreae's correspondence (2018).[86] Italy also has many such inventories, such as the ones listing the correspondence of the Manutius family (1957),[87] of Paolo Ruffini (1997),[88] and of Cassiano dal Pozzo (1991).[89] In Spain, there was a tendency during the last decades to skip the process of publishing inventories and proceed directly to editing letters, although a provisional catalogue has been printed for the large project of Benito Arias Montano's letters (2002).[90] The adjective 'provisional' anticipates that such inventories become outdated once critical editions are underway or finished. Thus, the editors of the correspondence of Joseph Scaliger compiled and updated their inventory during their work and published it only on EMLO once the edition had been completed and printed.[91]

Smaller inventories appeared in the meantime in journals and in the appendixes of dissertations and monographs all across Europe: noteworthy examples include Charles B. Schmitt's inventory of Jacques Daléchamps's correspondence,[92] Axel E. Walter's inventory of Georg Michael Lingelsheim's correspondence (2004),[93] Peter Korteweg's census of the letters to and from Johannes Drusius

[81] Françoise Weil, *Jean Bouhier et sa correspondance, [vol.] 1: Inventaire* (Paris: Université Paris-Sorbonne, 1975).

[82] Joseph A. G. Tans and Henri Schmitz Du Moulin, with Hans Bots, H. Buycks, and Cornelius P. Voorvelt, *La Correspondance de Pasquier Quesnel: inventaire et index analytique. [vol. 1], Inventaire* (Brussels and Louvain: Nauwelaerts and Bureau de la R.H.E., Bibliothèque de l'Université, 1989).

[83] Jens Häseler and Rolf Geissler, *La Correspondance de Jean Henri Samuel Formey (1711–1797): inventaire alphabétique […] avec la bibliographie des écrits de Jean Henri Samuel Formey* (Paris: Honoré Champion, 2003).

[84] Irène Passeron et al., *Jean le Rond d'Alembert – Oeuvres complètes; Série V: Correspondance générale. Vol. 1: Inventaire analytique de la correspondance, 1741–1783* (Paris: CNRS, 2009).

[85] Maria-Cristina Pitassi, with Laurence Vial-Bergon, Pierre-Olivier Lechot, and Eric-Olivier Lochard, *Inventaire critique de la correspondance de Jean-Alphonse Turrettini*, 6 vols. (Paris: Honoré Champion, 2009).

[86] Stefania Salvadori, *Inventar des Briefwechsels von Johann Valentin Andreae (1586–1654)* (Wiesbaden: Harrassowitz, 2018).

[87] Ester Pastorello, *L'epistolario manuziano: inventario cronologico-analitico, 1483–1596* (Florence: Leo S. Olschki, 1957).

[88] Francesco Barbieri and Franca Cattelani Degani, *Catalogo della corrispondenza di Paolo Ruffini* (Pisa: ETS, 1997).

[89] Anna Nicolò, *Il carteggio di Cassiano dal Pozzo: Catalogo* (Florence: Leo S. Olschki, 1991).

[90] Antonio Dávila Perez, 'El epistolario de Benito Arias Montano, catálogo provisional', *De Gulden Passer* 80 (2002): 63–129.

[91] See http://emlo-portal.bodleian.ox.ac.uk/collections/?catalogue=joseph-justus-scaliger, accessed 20/03/2019.

[92] Charles B. Schmitt, 'The Correspondence of Jacques Daléchamps (1513–1588)' *Viator* 8 (1977): 399–434, and 409–34, see https://doi.org/10.1484/J.VIATOR.2.301574.

[93] Appendix to Axel E. Walter, *Späthumanismus und Konfessionspolitik: Die europäische Gelehrtenrepublik um 1600 im Spiegel der Korrespondenzen Georg Michael Lingelsheims* (Tübingen: Max Niemeyer Verlag, 2004), 478–545.

(2006),[94] Leo van Santen's listing of Ludwig Crocius's letters (2014),[95] and Martin Mulsow's overview of Christoph August Heumann's correspondence (2017).[96] Modern editions are typically accompanied by appendixes listing the metadata of the letters published; such appendices are structured inventories of correspondences and may be found in such works as Kemke's edition of the correspondence of Patrick Young (1898),[97] or Sophie van Romburgh's edition of Franciscus Junius the Younger (2004).[98] Digitizing such lists is a relatively simple task, although a laborious one. Such tables can be OCR'd and automatically turned into electronic tables, but they require much manual correction and curation.

3.3 A Framework for Assembling Catalogue-level Descriptions

The resources itemized in the previous section represent only the principal points of departure for exploring and assembling a comprehensive union catalogue of learned correspondence in early modern Europe; yet the quantity of data potentially available even within these resources is dauntingly large. The following chapter (III.2) describes the development of semi-automated tools and systems to facilitate and accelerate the transformation of this often rough-and-ready data into high-quality normalized catalogue metadata. Yet even with the assistance of such tools, alternative approaches will be needed if we wish to assemble a comprehensive pool of relevant data in an efficient and properly prioritized fashion.

One obvious approach is to develop means of collecting collection-level descriptions of correspondence archives as a first step towards providing item-level descriptions. Providing collection-level descriptions is obviously a more difficult challenge for manuscript letters than for printed ones. The early modern printed letter collections discussed in section 2 can easily be described by well-established bibliographical standards; huge quantities of data of this kind have already been assembled in existing bibliographies; and the project of assembling a comprehensive set of such collection-level descriptions is now well under way in the EROL project described in section 2.2.

Analogous attempts to assemble collection-level descriptions of manuscript correspondence do not enjoy these advantages, but neither must they begin *de novo*.

---

[94] Peter Korteweg, *De Nieuwtestamentische commentaren van Johannes Drusius (1550–1616)* (Melissant: s.n., 2006).

[95] Leo van Santen, *Bremen als Brennpunkt reformierter Irenik. Eine soziologische Darstellung anhand der Biographie des Theologen Ludwig Crocius* (Leiden and Boston: Brill, 2014).

[96] Martin Mulsow, 'Der Verbesserer. Heumanns Poecile im Kontext seiner Korrespondenz mit der Gelehrtenrepublik. Mit einem Inventar von Heumanns Briefwechsel,' in Martin Mulsow, Kasper Risbjerg Eskildsen, and Helmut Zedelmaier, eds., *Christoph August Heumann (1681–1764). Gelehrte Praxis zwischen christlichem Humanismus und Aufklärung* (Stuttgart: Franz Steiner Verlag, 2017), 39–70.

[97] Johannes Kemke, *Patricius Junius (Patrick Young), Bibliothekar der Könige Jacob I. und Carl I. von England: Mittheilungen aus seinem Briefwechsel* (Leipzig: M. Spirgatis, 1898).

[98] Sophie van Romburgh, *'For my worthy freind Mr Franciscus Junius': An Edition of the Correspondence of Francis Junius F.F. (1591–1677)* (Leiden and Boston: Brill, 2004).

At a COST Action meeting held in The Hague in 2016, members of Working Group 4 addressed these issues and advocated the use of existing international standards for describing correspondence collections as far as possible. Two leading international standards, ISAD (G) and DACs, were singled out for special consideration.[99] ISAD (G) is the acronym for the 'General International Standard Archival Description', a comprehensive document adopted by the International Council on Archives in 1999. DACs refers to 'Describing Archives: A Content Standard', which was published by the Society of American Archivists in 2013. In addition, EAD (Encoded Archival Description) was advocated for the 'Reference Code' since EAD is the standard recognized by the Library of Congress for encoding archival collections in a manner that both reflects the hierarchical nature of archival description and which is compatible with SGML (Standard Generalized Markup Language) and XML (Extensible Markup Language).[100] After studying these models, the following key elements were agreed upon for any future online union catalogue capable of accommodating collection-level descriptions of correspondence:

1. Reference Code: EAD (Encoded Archival Description) to be used.
2. Name and Location of Depository.
3. Title: the name of the collection in existing finding aids.
4. Dates of Creation (of record).
5. Name of Creator (of record): author of file to be named.
6. Extent: including number of letters, whenever possible.
7. Level of Description: the only practicable course would be to ask the archive/library to use their own original system, since this would normally conform to ISAD (G) norms.
8. Scope and Content: to be included as far as possible.
9. Access: this field should include the URL of institutions (vital for small repositories) and note any access restrictions.
10. Languages and Scripts: ISO standards 639–2 and 639–3 to be used.
11. Administrative and Biographical History: provenance and collection description history to be noted, and existing finding aids listed.

---

[99] For ISAD(G), see the *General International Standard Archival Description*. 2nd edn. Adopted by the Committee on Descriptive Standards, Stockholm, Sweden, 19–22 September 1999 (Ottawa, 2000), see https://web.archive.org/web/20111027061153/http://www.icacds.org.uk/eng/ISAD%28G%29.pdf. For DACS: 'Describing Archives: A Content Standard', 2nd edn., see https://www2.archivists.org/standards/DACS; both accessed 20/03/2019.

[100] See https://www.loc.gov/ead/, accessed 20/03/2019.

12. Access points: including URLs of institution's access points if available and links to cognate institutions.

In addition to these standard fields, a number of additional fields would increase the utility of such a resource:

13. EULO unique identifier for collections.
14. Reference note: to include both printed and manuscript collections (e.g. link to printed edition of Grotius; links to articles on collections; URL links to same.

As in the analogous case of crowdsourcing item-level records (see the discussion of CEMROL in section 2.3 above), the chief difficulty in rolling out a programme of assembling collection-level descriptions of this kind would be that of recruiting participants in a huge range of libraries and archives across and beyond Europe. The technical precondition for such a campaign would be an online interface prompting archivists, librarians, and other scholars to enter properly formatted material under all of these headings. The best means of recruiting contributors would presumably be to act via international organizations such as the Consortium of European Research Libraries (CERL) and the International Council on Archives. Again, such a campaign would be most likely to succeed if coordinated with a broader campaign to assemble item-level metadata on a large scale, either as an end in itself or as a precondition for a systematic campaign of digitization. But how exactly can this be done and what are the necessary prerequisites? For answers to these questions see the infrastructure outlined in chapter III.5.

III.2 Reconciling Metadata

Eero Hyvönen, Ruth Ahnert, Sebastian E. Ahnert, Jouni Tuominen, Eetu Mäkelä, Miranda Lewis, and Gertjan Filarski

1 Introduction

The strategies and methods outlined in the previous chapter (III.1) are intended to increase the rate at which letter records are made available for inclusion in a union catalogue. The future development of distributed infrastructure (discussed in chapter III.5) could accelerate this rate further still. But as section II established in detail, letter records are problematic as all three of their main components – dates (II.3), places (II.2), and people (II.4) – can be expressed in multiple different ways. Add to this the possibility that the same letter could exist in many different states (II.1), and we are confronted with a fundamental problem: namely that of reconciling hundreds of thousands of letter and component records.

As the rate at which letter records are made available increases, tools will be needed to help accelerate the rate at which they can be processed for publication within a homogeneous union catalogue. Yet this huge volume of letters will also need to be processed to a higher degree of precision than was necessary in the analogue age. Human readers, flipping through traditional indexes card by card, can easily process minor variants in the form of names, dates, places, and sources; and the pace of scholarly work previously allowed time to puzzle over more difficult problems one at a time. Computers, analysing and visualizing thousands of records almost instantaneously will require clean, unambiguous data in order to produce reliable results. We therefore need tools capable of producing very high-quality data very quickly and in very large quantities.

Because data curation involves several different processes, several different tools will be required. Work on creating such tools has already been undertaken by several of the partners of the COST Action. A team in Cambridge and Queen Mary University of London has developed a tool to speed the reconciliation of the tens of thousands of letter-writers and recipients encountered in a single archive of over 100,000 letters (section 3). A team in Oxford and Helsinki has developed a tool for reconciling the people and places encountered in similar quantities of letter records originating from scores of smaller correspondence inventories (section 4). A team in Amsterdam, meanwhile, is leading the development of plans for more fully automated data reconciliation tools capable of meeting the most demanding of scholarly requirements (section 5). Before introducing these initiatives, this chapter sketches out some of the basic principles and approaches involved in the process of data reconciliation (section 2).

2 Data Reconciliation

Eero Hyvönen

The process of making heterogeneous data sets interoperable with each other is known as data reconciliation. Data reconciliation must proceed at two different levels – syntactic and semantic – because data sets can be heterogeneous in two different ways. In linguistics, syntax relates to the structure of a sentence and semantics to the meaning of the individual words and phrases within it. Similarly in computer science, syntax refers to the manner in which data is structured, and semantics refers to the significance of the individual elements within those structures and their relations. One set of tools is needed for reconciling data sets which differ syntactically (i.e. are structured differently). A different set of tools is needed for reconciling data which differ semantically (i.e. which express the same meanings in different ways).

Syntactic interoperability and data cleaning. On the syntactic level, the same data can be structured in many different ways. For example, an inventory of correspondence consisting of precisely the same data (e.g. names of sender and recipient, place of sending and receipt, date of sending, etc.) can be presented in tabular formats such as CSV (Comma Separated Values), or as JSON (JavaScript Object Notation) objects,[1] or as RDF graphs (Resource Description Framework).[2] In addition, data values, such as dates, person names, and numeric values may be represented in different forms: the same Gregorian date, for instance, can be represented as *yyyy-mm-dd*, as *dd-mm-yyyy*, and in many other configurations as well. In order to be made interoperable, these differing data structures and representations need

[1] See https://www.json.org/, accessed 20/03/2019.

[2] See https://www.w3.org/RDF/, accessed 20/03/2019.

to be transformed into a common format: for instance, when publishing Linked Data, everything needs to be transformed into RDF.

A task closely related to syntactic reconciliation is data cleaning, that is, the removal from the data of typing errors and irregular formatting. Syntactic transformations are in many cases technically fairly straightforward to do but may be tedious. There are many tools available to facilitate data cleaning and transformations, such as *OpenRefine*[3] and *Karma*.[4] Others have been specifically adapted for the purpose of preparing spreadsheets for ingesting into EMLO.

Semantic interoperability. More formidable are the challenges encountered in data reconciliation at the semantic level. Reconciliation at this level is needed due to the fact (established in detail in chapters II.2–5) that data can mean different things even when expressed in similar or identical ways. For example, the date 1 January 1600 means three different things depending on whether the calendar used is Julian with the year beginning on 1 January, or Julian with the year beginning on a different date (e.g. 1 March, or 25 March), or Gregorian. Similarly, the name 'Neustadt' could refer to over twenty different places in Germany alone and a dozen more elsewhere, not to mention to innumerable people with that surname. On the other hand, completely different representations can also refer to the same entity: 'Leiden', 'Leyden', and 'Lugdunum Batavorum' all refer to the same place in the sixteenth century (although the Latin originally referred to a slightly different place), and the term 'archbishop of Armagh' also referred, during a specific interval, to James Ussher. Ambiguity of this kind prevents the accurate processing of data digitally. When searching and browsing, errors and semantic confusion in the aggregated data result in low precision (i.e. a search returns results for all the 'Neustadts' rather than the specific one of interest to the researcher); and incomplete aggregated data likewise results in low recall (i.e. search results return only those records in which James Ussher is mentioned by name).

Further difficulties arise on a higher structural level, where incompatible ways of representing knowledge frequently occur: for instance, if the creator of a text is indicated as an 'author' in one data set and a 'writer' in another, then 'writers' are not found when looking for 'authors', or 'authors' when looking for 'writers', both of which lower recall. In many cases, the semantic content in one data set cannot be represented by using the knowledge structures of another data set. In such cases, a more fundamental underlying knowledge representation scheme is needed for representing both data sets and their relations in common terms. To represent the different expressions and manifestations of the Bible in print or in other media formats, for example, a deeper notion of the underlying idea of the Bible as an immaterial work and its physical representations is required. The same applies to the individual letter, not to mention people and places.

[3] See http://openrefine.org/, accessed 20/03/2019. See below section 4.

[4] See http://usc-isi-i2.github.io/karma/, accessed 20/03/2019.

Major approaches. This section considers data reconciliation from the perspective of integrating cultural heritage linked data contents that are represented using different kinds of metadata models.[5] In this context, two major approaches are in use for reconciling (meta)data. First, within the Dublin Core framework,[6] different document-based schemas can be harmonized by using the Dumb-Down Principle. The idea is to map metadata elements onto each other within a hierarchy. For example, if the 'author' and the 'writer' of a text are represented as sub-properties of a more general property 'creator', the machine can understand the relationship between the different kinds of creators. Similar hierarchies can be established for other classes of concepts. This allows queries to be expanded by moving up the hierarchy: searching for 'creator' returns more results than either 'writer' or 'author'. Alternatively, a fundamental underlying ontology[7] describing the domain of discourse can be modelled and different metadata schemes transformed into it. The best-known examples of this approach are the CIDOC Conceptual Reference Model[8] for cultural heritage museum data, and the Functional Requirements for Bibliographic Records (FRBR) (and the related conceptual models FRAD and FRSAD) for library data,[9] that were recently consolidated as the IFLA Library Reference Model.[10] CIDOC CRM and the FR-models are being combined into a Conceptual Model for Bibliographic Information in Object-Oriented Formalism (FRBRoo).[11] In contrast to the document-centric approach of Dublin Core, the conceptual reference models above are event-centric in nature: within them, the world of discourse is described in terms of events and their related constituents – especially participants, place, and time – in a manner highly suitable to modelling exchange within the republic of letters as well as the underlying biographical and prosopographical data.[12]

---

[5] Eero Hyvönen, *Publishing and Using Cultural Heritage Linked Data on the Semantic Web* (Palo Alto, CA: Morgan & Claypool, 2012).

[6] See http://dublincore.org/, accessed 20/03/2019.

[7] The term 'ontology' here refers to ontologies as formal, shared, structured models of data used in computer science for representing knowledge, not to ontology as a branch of philosophy studying the nature of being, existence, and reality. See Steffen Staab and Rudi Studer, eds., *Handbook on Ontologies*, 2nd edn. (Berlin: Springer Verlag, 2009).

[8] CIDOC CRM. ICOM/CIDOC Documentation Standards Group/CIDOC CRM Special Interest Group, *Definition of the CIDOC Conceptual Reference Model*, Version 6.2.3. Christian-Emil Ore, Martin Doerr, Patrick Le Bœuf, S. Stephen Stead (eds.), 2018. See http://www.cidoc-crm.org/Version/version-6.2.3, accessed 20/03/2019.

[9] See FRBR. IFLA Study Group on the Functional Requirements for Bibliographic Records, *Functional Requirements for Bibliographic Records*, Revised 2009. See https://www.ifla.org/publications/functional-requirements-for-bibliographic-records, accessed 20/03/2019.

[10] See https://www.ifla.org/publications/node/11412, accessed 20/03/2019.

[11] IFLA Working Group on FRBR/CRM Dialogue, Definition of FRBRoo, *A Conceptual Model for Bibliographic Information in Object-Oriented Formalism*, Version 2.4. Chryssoula Bekiari, Martin Doerr, Patrick Le Bœuf, and Pat Riva (eds.), 2016. See https://www.ifla.org/publications/node/11240, accessed 20/03/2019.

[12] See Jouni Tuominen, Eero Hyvönen, and Petri Leskinen, 'Bio CRM: A Data Model for Representing Biographical Data for Prosopographical Research', in Antske Fokkens, Serge ter Braake, Ronald

Semantic disambiguation and entity linking. A recurring basic problem in data reconciliation is how to map literal, i.e. textual, expressions in (meta)data, such as names of persons and places, onto their corresponding unique meanings, defined in a reference registry (domain ontology). Consider the case of reconciling epistolary data for the EMLO database. Whenever metadata comes from two different collections, say A and B, the 'authors' and 'recipients', and the 'origins' and 'destinations' of the letters in A and B are literal names of persons and places, possibly written in different languages, with different orthographies and perhaps even different transliteration practices. In order to reconcile the data in A with the data in B (and, by extension, with the data in all the other inventories brought together in EMLO), the best practice is not simply to designate one of these name forms as standard: instead, these names have to be mapped to Unique Resource Identifiers (URIs) for each of the people and places represented in the underlying common EMLO ontologies of people and places. This process is called Named Entity Linking (NEL),[13] and ensures that letters to or from the same person or place can be accurately aggregated and identified.[14]

A key challenge in NEL is semantic disambiguation, that is, the task of selecting the correct referent from multiple possible choices. For instance, identical or nearly identical names can be used for the same entity (e.g. Johann Strauss, a father, and Johann Strauss, his son). Alternatively, the same name may be used for different kinds of entitles (such as persons and places, in the case of 'Neustadt' mentioned above). In order to disambiguate data of this kind, reference must be made to contextual data. For instance, if a letter from Johann Strauss is dated after the death of the father, we can reliably infer that it must have been written by the son. In principle, it is therefore possible to devise an inference engine which could ascribe letters in cases such as this when sufficiently rich prosopographical data is available.

However, especially when dealing with historical materials, we often lack the contextual data needed to disambiguate such cases with certainty or even to identify a referent candidate tentatively. This commonly leads to two equally undeniable outcomes. One results from creating a separate resource and minting a unique URI for each unreconciled entity: in this case, the register ontology will become populated with multiple instances of individuals that actually refer to the same entity. When additional contextual data becomes available, further disambiguation may be possible, and this requires a method of merging these records into a master record.

Sluijter et al., eds., *Proceedings of the Second Conference on Biographical Data in a Digital World 2017*, 59–66, see http://ceur-ws.org/Vol-2119/paper10.pdf.

13 Delip Rao, Paul McNamee, and Mark Dredze, 'Entity Linking: Finding Extracted Entities in a Knowledge Base', in Thierry Poibeau, ed., *Multi-source, Multilingual Information Extraction and Summarization, Theory and Applications of Natural Language Processing* (Berlin: Springer Verlag, 2013), 93–115, see https://doi.org/10.1007/978-3-642-28569-1_5.

14 Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran, 'Evaluating Entity Linking with Wikipedia', *Artificial Intelligence* 194 (2013): 130–50, see https://doi.org/10.1016/j.artint.2012.04.005.

However, the linking of named entities is an uncertain process whenever the contextual data is inadequate; and even greater problems emerge when named entities are linked on the basis of questionable data or dubious inferences from it. What is needed in order to avoid these opposing difficulties is a method for dealing with uncertain links between named entities: we need methods for identifying possible duplicates, for flagging these for further investigation, and for merging records into one master record which also registers incomplete, uncertain, or fuzzy knowledge.

Depending on the desired level of precision and recall in NEL, fully automatic semantic disambiguation and linking process may not be possible. Fortunately, it is often possible to identify problematic instances and to develop semi-automatic tools where difficult cases can be resolved with the help of a human expert. In the following sections, two semi-automatic tools are described, both developed for use with epistolary data. Section 3 describes the *Disambiguation Engine* for identifying references to the same persons in metadata deriving from the English State Papers in the Tudor period. Section 4 describes *Recon*, a tool for disambiguating, linking, and de-duplicating EMLO metadata and registries deriving from multiple sources. Finally, in section 5, the future development of inference-based data linking is discussed in a more general setting.

3 Disambiguating People in a Single Large Data Set: The *Disambiguation Engine*

Ruth Ahnert and Sebastian E. Ahnert

An excellent example of the need for disambiguation tools is provided by the AHRC-funded research project *Tudor Networks of Power*. The objective of this project was to study the networks documented by 132,747 letters in the central archive for English history between 1509 and 1603, the Tudor State Papers, held at the National Archives. In the process of digitizing this material to create *State Papers Online* (SPO), Gale Cengage had created XML metadata that provided a basis from which to begin. In its raw state, this metadata contains over 37,000 different values in the name fields of senders and recipients. But in many cases, more than one of these values refer to a single person, due to spelling variants, name changes through marriage, or the acquisition of titles or formal roles, for instance. In these cases, the labels need to be de-duplicated, meaning that two or more labels are collapsed onto one. In many other cases, the same label might refer to different people in different manuscripts, either because different individuals occupied the same office over time, or simply because some names were common to more than one individual. In this case the labels need to be disambiguated, meaning that a label needs to be split into two or more existing or new labels.

In order to speed this work of de-duplication and disambiguation, the project team developed the *Disambiguation Engine*. In order to disambiguate metadata relia-

bly, scholars must consider it alongside contextual data. The *Disambiguation Engine* accelerates this process by allowing scholars to move quickly and easily between three very large layers of data all included in State Papers Online: (1) the basic metadata fields (including names of sender and recipient, date and place of sending, and document identifier/manuscript reference); (2) the supplementary information, such as letter abstracts, added by the editors of the Calendar of State Papers, which were begun in the nineteenth century; and (3) the digital scans of the original manuscript letters themselves. Before loading the data into the *Disambiguation Engine*, each name label is assigned a unique URI (Uniform Resource Identifier), in this case a number. The tool then allows the user to record their disambiguation and de-duplication decisions by mapping these URIs to other existing URIs, or to new URIs. The principal function of the *Engine*, however, is to present to the scholar as efficiently as possible the information from all three levels of State Papers Online relevant to disambiguating and de-duplicating these labels.

In the on-screen presentation of this information (shown in fig. 1), the basic decision-making process reads from left to right. The main left-hand panel lists all the 37,000 name labels in alphabetical order, preceded by their automatically generated URI (the number). When any name label on this list is highlighted (like the 'Duke of Montpensier' in this case, labelled 1 in the figure), an editable box appears (labelled 2) where a number can be entered to map that label to another URI. A name label can be mapped to multiple IDs, separated by a comma, normally where there are multiple senders or recipients; in this case, however, the URIs are separated by semicolons to distinguish the two separate holders designated by that title across the manuscripts where that name label appears. Here the sequence of IDs mirrors the sequence of manuscripts with this name label, which appear when you highlight his name. This sequence appears in the box labelled 3. When a manuscript reference in box 3 is clicked, the full details of that letter appear in the panel in the top-right corner (labelled 4), including details of the letter recipient, place of writing, and date sent. The word 'Images', which appears in the second line of text in box 4, allows the user to link through to a scan of the original letter.

The tool also allows different forms of navigation and search, enabled by the boxes labelled 5 and 6. The navigation box (labelled 6) allows the user to jump to any ID in the main left-hand box, and it also toggles the whole list of names, which can be viewed in box 5. The entire list view is useful to search for name variants or any other keywords in the name labels that might aid the decision process. Moreover, by clicking on the components of the name and title in the left-hand box (such as, for example, 'Duke', 'Montpensier', etc.), all other instances of that component can be found; these will appear in box 5, where you can use the hyperlinked name-list to jump to those other instances.

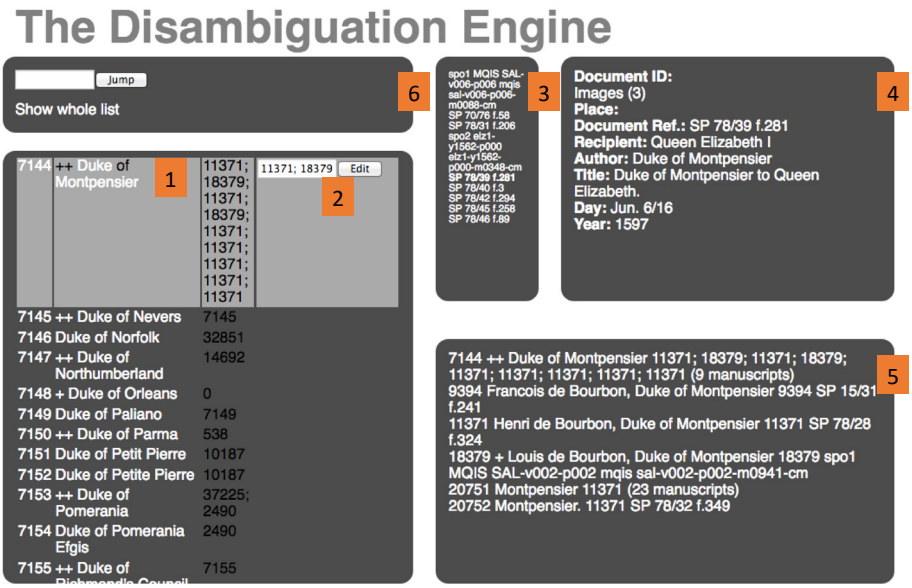


Figure 1: The interface of the *Disambiguation Engine*, used in the *Tudor Networks of Power* project

In many cases the metadata alone might be enough to provide a positive identification: for example, the highlighted label in figure 1, 'Duke of Montpensier', can be disambiguated using the dates of the letters, cross-referenced with historical lists of title-holders. In cases for which no date is provided or the date provided is insufficiently precise to allow disambiguation, it is often necessary to examine the description of the letter provided by the Calendar entry, or even to examine the text of the letter or the signature on the digital scan of the manuscript. This is often the most definitive way to make a positive identification in ambiguous situations. One example is the case of James Fitzgerald, earl of Desmond (URI 13081). Over the entire period of the Tudor State Papers, four separate individuals with the first name James held the title of earl of Desmond, and in addition the title was challenged on the earldom's second creation (when it was held by James Fitzgerald, the 1st earl of Desmond) by James FitzThomas FitzGerald, the Súgán earl of Desmond. Confusingly the 1st earl and the Súgán earl were both imprisoned in the Tower of London around the same time, and both petitioned the government for mercy, thereby creating some difficulties in telling them apart. In this context their signatures are often the only way to distinguish them. When in doubt, our decision on the whole was to over-disambiguate: a person who could not be positively identified as another person in our data set would retain their URI, or, in the case of multiple letters associated with one name, we would assign them their own URI. Given that the aim of this project was a network analysis of the correspondence,

this leads to an under-connected rather than an over-connected network, which is preferable to erroneously exaggerating a person's significance.

In the process of disambiguating and de-duplicating records from the Tudor State Papers, over 37,000 original URIs, based on the metadata name labels, were mapped to a set of 20,656 final URIs, of which over 1,000 were newly generated. Such work, even aided by such a tool, takes a significant time-commitment. The process described above took one of the team nine months of working on the task full-time, plus almost another nine months of undertaking the work alongside a standard UK academic job (teaching plus administration). Leaving adequate time for these data processing steps is vital, as any quantitative analysis results will be highly dependent on the accurate and comprehensive disambiguation and de-duplication of the underlying data. Moreover, the complexity of some of the decisions, including the need to compare handwriting and signatures to distinguish some historical persons, shows that caution should be used when employing automated disambiguation and de-duplication processes. However, once this work has been undertaken you can begin analysis of the data with confidence. You can read some of the outcomes of the *Tudor Networks of Power* project in chapter IV.5.

4 Reconciling Data from Multiple Large Data Sets: *Recon* and *Mare*

Jouni Tuominen, Eetu Mäkelä, Miranda Lewis, and Eero Hyvönen

The *Cultures of Knowledge* project has collaborated with Aalto University in Finland to create another reconciliation tool, known as *Recon*, for use in scenarios where data is complex and accuracy is of paramount importance. Such a scenario appears when data from many different sources is prepared for publication on *Early Modern Letters Online* (EMLO).[15] EMLO is a union catalogue of early modern learned correspondence being populated collaboratively by a network of contributors, including scholars working on a specific collection or edition of correspondence, librarians, archivists, and publishers. The metadata contributed by these partners can be input using either a customized spreadsheet or via EMLO's own input web form (known as *EMLO-Collect*). In order for data originating from many different sources to be rendered interoperable, personal names and place names in the incoming metadata either have to be matched to the existing person and place records in the EMLO database or new person and place records (and URI identifiers) need to be created for them. It is to assist this matching process when the metada-

[15] See http://emlo.bodleian.ox.ac.uk/, accessed 20/03/2019.

ta is ingested from a spreadsheet that two semi-automatic tools, *Mare* and *Recon*, were developed.[16]

Mare is a simple tool to extract aggregate columns from tabular data. The user interface of *Mare* is depicted in figure 2. The tool is used in the EMLO spreadsheet workflow to collect all unique personal names (labelled 1 in the figure) and place names from a correspondence data set with contextualizing information, such as the years of activity (labelled 2) based on the dates of the letters that involve particular people or places. When processing personal names, also the places that are involved in the letters (labelled 3) can be used as a contextualizing information, and vice versa. The tool functions in an automatic way after the reconciliation rules have been specified by the user.

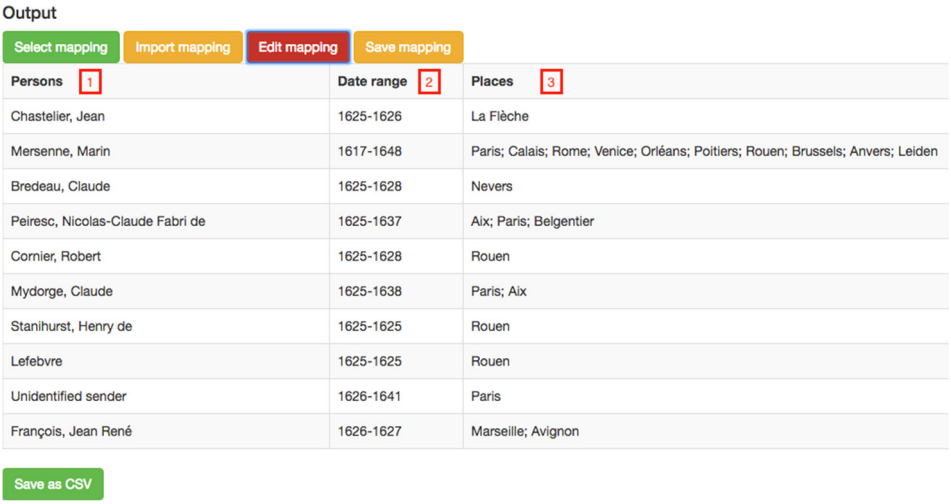


Figure 2: A sample output of the *Mare* tool, used for preprocessing personal names of an ingested spreadsheet. The boxes represent: (1) a list of distinct personal names of the authors and recipients in the spreadsheet; (2) the activity years of the authors and recipients, based on the dates of the letters; (3) the places the authors and recipients have been active in, based on the origins and destinations of the letters

Recon is designed for digital humanities scenarios where accuracy is of paramount importance and resources are sometimes, but not always, available for manual verification of candidates for matching. This means that: (1) the matching cannot be done entirely automatically; (2) the tool has to return as many potential matches as possible for the user to consult and consider a 'match'; and (3) the user has to be supported in the manual verification process with the provision of contextual

[16] Both tools are available as Open Source via *GitHub* at http://github.com/jiemakel/mare (*Mare*) and http://github.com/jiemakel/recon (*Recon*), accessed 20/03/2019.

information concerning the match candidates. Compared to reconciliation tools such as *Silk*[17] and *OpenRefine,*[18] *Recon* focuses on the manual checking of the match candidates, with a simple, fast, and intuitive workflow, using a browser-based user interface.

The *Recon* user interface is depicted in figure 3. The tool reads the personal and place names on a contributor's spreadsheet (produced by *Mare*), possibly supplemented with contextual information when it is available (labelled 1 in this figure). Working through the data rows (labelled 2), *Recon* sends SPARQL queries to a triplestore containing the authority to match against, here all the people and place records from the EMLO database.[19] For each person or place in the spreadsheet, a list of potential candidate matches from EMLO is offered to the user (labelled 3). The matches are prioritized according to the string similarity of the name, enhanced by any other criteria contained in the SPARQL query used. For example, the date of the letter in the contributor's database can be compared with the birth and death dates of the EMLO person records to rule out candidates not alive when the letter was written. The user has the option to specify whether or not there is a match, or to leave a query indicating uncertainty or requesting further investigation. When the spreadsheet has been processed, *Recon* re-exports to the contributor the original data supplemented with the EMLO IDs of the matched people and places. Where no matches have been identified, new people and place records are created either manually in the EMLO editing interface (known as *EMLO-Edit*) or by using a bulk upload script, and their IDs are inserted into the spreadsheet either manually or by running the list a second time through *Recon*. After the data set has been checked by the contributor and adjustments made where required, it can be ingested into EMLO using the complete list of people and place IDs.

In addition to assisting in the reconciliation of incoming metadata with EMLO, *Recon* can be used to reconcile data already within EMLO. The population of EMLO with data from multiple catalogues has led to the gradual proliferation of multiple records of the same letter. For example, a letter exchanged between Martin Lister and Edward Lhwyd preserved in the Bodleian Library may be documented by three separate letter records: one in the calendar of Lister's correspondence contributed by Anna Marie Roos; the second in the calendar of Lhwyd's correspondence contributed by a team led by Brynley F. Roberts; and the third from the card catalogue of the manuscript correspondence in the Bodleian.

---

[17] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov, 'Discovering and Maintaining Links on the Web of Data', in Abraham Bernstein, David R. Karger, Tom Heath, Lee Feigenbaum, Diana Maynard, Enrico Motta, and Krishnaprasad Thirunarayan, eds., *Proceedings of the 8th International Semantic Web Conference (ISWC 2009)* (Berlin: Springer Verlag, 2009): 650–65. https://doi.org/10.1007/978-3-642-04930-9_41.

[18] Ruben Verborgh and Max De Wilde, *Using OpenRefine* (Birmingham: Packt Publishing, 2013).

[19] Jouni Tuominen, Eetu Mäkelä, Eero Hyvönen, Arno Bosse, Miranda Lewis, and Howard Hotson, 'Reassembling the Republic of Letters - A Linked Data Approach', in Eetu Mäkelä, Mikko Tolonen, and Jouni Tuominen, eds., *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference*, 76–88, see http://ceur-ws.org/Vol-2084/paper6.pdf.

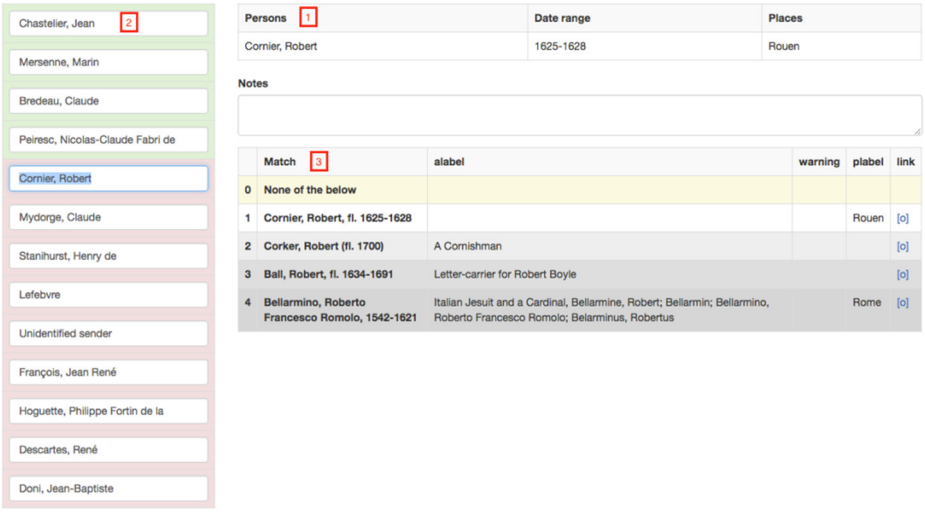


Figure 3: The user interface of *Recon*, used in matching personal names of an ingested spreadsheet to the existing person records in EMLO. The boxes represent: (1) the personal name that is currently being matched, with contextual information, including a date range and places of origin and destination of correspondence; (2) the personal names of the data rows in the spreadsheet, where green indicates a match that the user has specified, and red a personal name that has not yet been processed by the user; (3) the ranked list of potential candidate matches with contextual information, possible warnings (e.g. the person was not alive when a letter was written), and link to the EMLO record of the person

In order to locate multiple records of the same letter, *Recon* is configured to run SPARQL queries across the EMLO data set to identify letters with the same sender and recipient and other similar metadata fields, in particular dates, places of origin and destination, repository and shelf-mark information, or printed edition details. The tool then ranks potential duplicate matches for a given letter by taking into account the proximity of the dates, string similarities of textual metadata fields, etc. EMLO editors can then determine which match, if any, is reliable; and if it is decided that the same letter has been recorded more than once by different contributors, a bridge link is inserted, preserving the two scholars' independently created metadata representations while indicating that they refer to the same letter.

Whilst working with *Recon*, EMLO's editors have reported that the process of calling up records to identify matches provides the opportunity to consider people, place, and letter records in different combinations, and to view metadata 'from different angles' to those encountered elsewhere in their workflows. In consequence, errors are spotted and corrected, or potential matches of records other than those required for the current search are observed, and these records can be cleaned and augmented in tandem on an ongoing basis.

5 An Automated Approach to Data Quality: *CommonPlace*

Gertjan Filarski

As indicated by the foregoing discussion, within the humanities the view prevails that high-quality data sets can only be delivered by a process of meticulous manual selection, disambiguation, and curation conducted by trained specialists. The tools discussed in sections 3 and 4 are therefore only semi-automated: their function is to present scholars as efficiently as possible with the data they need in order to disambiguate and reconcile data swiftly and accurately. However, it has also been indicated in passing that the accumulation of homogeneous contextual data on a system provides a basis on which carefully devised programs can begin to make reliable inferences in a fully automated fashion. As increasingly large sets of raw data become increasingly frequently available, innovative work on fully automated processing of this kind will be needed. With this problem in mind, the COST Action has also generated a new proposal for more advanced (semi-)automation of this kind in a project known as *CommonPlace*.[20]

CommonPlace will implement a multi-part strategy to assure data quality across the entire knowledge graph. First, and above all, we believe in empowering our users in all editorial decisions. Based on the extensive provenance trails created for all revisions in the knowledge graph, we will allow users to decide what to accept and what to reject. A user can give precedence to certain sources, override data with an alternative interpretation and in- or exclude assertions or rejections made by others. This level of intervention relies on the expertise of the user and is intended primarily for professional users – scholars, data journalists, teachers, etc. – who are qualified to prepare and curate data selections.

A second level of quality control, complementing the first, is directed at lay users who will not be able to participate in the fine-grained curation of a data set. For these users, *CommonPlace* will implement a ranking system that will by default serve the highest ranked data to the user. The ranking system is an extension (advocated by Hyvönen, et al.)[21] of the five-star model proposed by Tim Berners-Lee. In this extended ranking, we look not just at the technical quality and availability of the data but also test a source's quality by measuring its granularity in relation to similar data in the graph. Greater weight can also be granted to those values that are in agreement with other data sets.

Finally, we will implement a dual strategy of complementary automated and semi-automated data quality management tools. Automated algorithms will constantly scan the graph for certain patterns. The tools will be designed to be flexible

[20] See https://commonplace.eu/, accessed 20/03/2019.

[21] Eero Hyvönen, Jouni Tuominen, Miika Alonen, and Eetu Mäkelä, 'Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Data Sets', in Valentina Presutti, ed., *The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers* (Cham: Springer Verlag, 2014), 226–30, see https://doi.org/10.1007/978-3-319-11955-7_24.

enough to support both explicit rule-based and inferential machine-learning approaches. Some of the detected patterns will result in a data object being flagged for editorial review. These types of checks will become more complex with the growth of the knowledge graph. Depending on the defined policy settings of the hub, an editor with sufficient permissions will be able to review and correct all flagged data objects. This environment will be optimized for the bulk correction or rejection of certain well-defined classes of automated findings. All editorial changes will add a metadata annotation to the knowledge graph with provenance data – in the case of a (semi-)automated correction, which version of which algorithm made the addition with which level of certainty at which moment in time. Such provenance trails allow other users to alter their view on a data selection in order to explicitly include or exclude (semi-)automated corrections, or to accept them only when these were made above a minimum threshold of certainty. *CommonPlace* intends to use these automated tools especially for certain kinds of secure inferences. Link inferences are a good example: a tool can reliably infer that Person A and Person B are siblings from the fact that they have parental links to the same mother and father. A logical second example is mutually exclusive assertions: if the knowledge graph indicates that John Smith died in Trieste and that John Smith died in Prague, these two name labels cannot refer to the same person.

While the possibilities outlined in this concluding section remain entirely *in potentia*, enough has been said to indicate that there are abundant opportunities for ongoing development in this field.

III.3 Transcribing and Editing Text

Charles van den Heuvel, Montserrat Prats López, Thomas Wallnig, Chiara Petrolini, Elena Spadini, and Elizabeth R. Williamson

With contributions from Günter Mühlberger

1 Introduction

Charles van den Heuvel

Transcribing texts written in antiquated handwriting and old forms of vernacular and ancient languages is a demanding enterprise. For no class of texts is this challenge more acute than when transcribing learned correspondence, which combines documents in many different hands, sometimes from several different countries, and sometimes written in several different languages. For centuries, the difficult task of transcribing these materials has been vested in scholarly experts, and the same applies still more to the very different challenge of editing and annotating such texts for publication.

Today, this scholarly monopoly is being challenged from at least two different directions. On the one hand, the challenge of reassembling a comprehensive collection of documentation of the republic of letters, within the resource constraints affecting humanistic scholarship generally, requires the exploration of more efficient and cost-effective means of generating large bodies of reliable transcription. On the other hand, the demands of annotating correspondences which can range

geographically and thematically across the entire *orbis litterarum* can also benefit from large-scale collaboration on shared digital platforms.

Needless to say, digital technology can contribute in many different ways to meeting these challenges. Since the task of assembling and curating epistolary metadata is at the heart of the present volume, there is no space here to explore all the different potential aspects of this challenge in detail. Instead, this chapter focuses on four key parts of a broader field of potential developments. In section 2, Montserrat Prats López discusses the advances of the crowd-sourcing model for transcribing letters without losing quality in the production of letter editions. The tension between 'broader public' and 'scientific community' addressed here is reminiscent of those at the core of the republic of letters itself. Thomas Wallnig and Chiara Petrolini then discuss an experiment with the even more direct application of digital technology to transcription: namely, the pioneering of automatic handwriting recognition using the software *Transkribus*. In section 4, some of the tools and methods for transforming such transcriptions into digital scholarly editions are discussed by Elena Spadini. Finally, Elizabeth Williamson surveys some of the issues that arise when collaborative projects curate larger digital text corpora with long and complicated provenance trails of their own.

2 Crowd Science in the Humanities: Transcription and Quality Control

Montserrat Prats López

2.1 Introduction

The idea that people can accomplish more in the domain of science by joining forces than they could alone – first prominently articulated by Francis Bacon in his *Instauratio magna* of 1620 – has driven many academic endeavours in history. The 1714 Longitude Prize and the creation of the first *Oxford English Dictionary* are only two of the most cited examples of involving the broader public in such scientific efforts. Calling on the general public to help in solving organizational, societal, and scientific problems is not really new, but recent technological developments have contributed to the expansion of such collaborative practice.

Crowd science[1] is the term often used to refer to this collaborative practice as 'an emerging organizational mode of doing science'[2] driven by the use of the Internet. The increasing access to the Internet and the declining costs of digital technology facilitate reaching and coordinating larger numbers of distributed people

[1] The word 'science' is used here to refer to all fields of academic research.

[2] Chiara Franzoni and Henry Sauermann, 'Crowd Science: The Organization of Scientific Research in Open Collaborative Projects', *Research Policy* 43:1 (2014): 1–20, at 7. See https://doi.org/10.1016/j.respol.2013.07.005.

and integrating their contributions to crowd science projects at a relatively low cost.[3]

2.2 Benefits of Crowd Science

Crowd science projects can potentially provide time and resource efficiency to a research project by engaging large numbers of people to collect, process, or analyse (textual) data, perform a task, or solve a problem.[4] Resource efficiency in particular has become an important benefit for public universities and research institutes as the latest economic recession has increased the pressure to find new ways of funding their research projects[5] or of allowing them to perform research that could not have been done without citizens' contributions.[6] Time and resource efficiency can be achieved by engaging large numbers of people who are willing to contribute their knowledge and skills without financial compensation, and to invest their free time in transcribing and annotating the vast number of pages offered through crowd science projects.

Other advantages of involving the public in the research practice through crowd science include the access to a broader pool of knowledge and science outreach. That is, science can benefit from the contributions of individuals with specialized knowledge, and from the increased ability to spot inaccuracies and verify results, as more diverse eyes and brains focus on the same task or build upon each other's contributions.[7] Moreover, crowd science can also be seen as a means of public outreach, improving the public's understanding of and involvement in science.[8]

---

[3] Allan Afuah and Christopher L. Tucci, 'Crowdsourcing as a Solution to Distant Search', *Academy of Management Review* 37:3 (2012): 355–75, see https://doi.org/10.5465/amr.2010.0146.

[4] Daren C. Brabham, 'Crowdsourcing: A Model for Leveraging Online Communities', in Aaron Alan Delwiche and Jennifer Jacobs Henderson, eds., *The Participatory Cultures Handbook* (New York: Routledge, 2013), 120–9; Franzoni and Sauermann, 'Crowd Science'.

[5] Thomas Estermann and Anna-Lena Claeys-Kulik, *Financially Sustainable Universities – Full Costing: Progress and Practice* (Brussels: European University Association, 2013), see https://eua.eu/resources/publications/387:financially-sustainable-universities-towards-full-costing-progress-and-practice.html, accessed 20/03/2019.

[6] It is important to note that crowd science does not substitute the work of academics. Relying only on the crowd to supply input for research may have consequences for quality and put at risk the jobs of expert academics. The need for crowd contributions versus ethical issues for the academic labour market are considered in Hauke Riesch and Clive Potter, 'Citizen Science as Seen by Scientists: Methodological, Epistemological and Ethical Dimensions', *Public Understanding of Science* 23:1 (2014): 107–20, at 118. See https://doi.org/10.1177/0963662513497324.

[7] Franzoni and Sauermann, 'Crowd Science'; Lars Bo Jeppesen and Karim R. Lakhani, 'Marginality and Problem-solving Effectiveness in Broadcast Search', *Organization Science* 21:5 (2010): 1016–33, see https://doi.org/10.1287/orsc.1090.0491; Mark N. Wexler, 'Reconfiguring the Sociology of the Crowd: Exploring Crowdsourcing', *International Journal of Sociology and Social Policy* 31:1–2 (2011): 6–20, see https://doi.org/10.1108/01443331111104779.

[8] Jonathan Silvertown, 'A New Dawn for Citizen Science', *Trends in Ecology & Evolution* 24:9 (2009): 467–71, see https://doi.org/10.1016/j.tree.2009.03.017.

2.3 Crowd Science in the Humanities

The number of crowd science projects has grown in recent years,[9] not only in the natural sciences[10] but also in the humanities.[11] An increasing number of crowd science projects in the humanities are being used as research-supporting initiatives to make literary and historical manuscripts digitally accessible to both the research community and the wider non-academic public. Before manuscripts are edited and made available online, they need to be transcribed and annotated. These are typically the tasks being outsourced to the general public through crowd science projects.

Crowd transcription projects can be developed in a number of different ways. One is to join a general crowd science platform, such as *Zooniverse*,[12] and to benefit from standardized functionality and technical support.[13] A second option is to build a specialized transcription environment, such as *eLaborate*.[14] A third possibility is to combine ready-made technologies with custom-made tools, such as the *Transcription Desk*[15] developed for the *Transcribe Bentham* project.[16]

The *Zooniverse* platform supports many crowd science projects covering various fields of research. Though it started by crowdsourcing the classification of images, the platform is in constant development, adding new functionality to support other types of tasks, such as the transcription of lengthy manuscripts in projects like *Shakespeare's World*.[17]

In contrast, *eLaborate* – developed by the Huygens Institute for the History of the Netherlands – is a collaborative online environment exclusively designed to support the digital transcription and editing of manuscripts. The source code of

[9] Franzoni and Sauermann, 'Crowd Science', 1; Andrea Wiggins and Kevin Crowston, 'Surveying the Citizen Science Landscape', *First Monday* 20:1–5 (January 2015), see https://doi.org/10.5210/fm.v20i1.5520.

[10] Andrea Wiggins and Kevin Crowston, 'From Conservation to Crowdsourcing: A Typology of Citizen Science', in Ralph H. Sprague, Jr., ed., *Proceedings of the 44th Annual Hawaii International Conference on System Sciences, 4–7 January 2011, Koloa, Kauai, Hawaii, USA* (Piscataway, N.J.: IEEE, 2011), see https://doi.org/10.1109/HICSS.2011.207.

[11] Kaja Scheliga, Sascha Friesike, Cornelius Puschmann, and Benedikt Fecher, 'Setting up Crowd Science Projects', *Public Understanding of Science* 27:5 (2016): 515–34, see https://doi.org/10.1177/0963662516678514.

[12] See https://www.zooniverse.org, accessed 20/03/2019.

[13] Nathan R. Prestopnik and Kevin Crowston, 'Citizen Science System Assemblages: Understanding the Technologies That Support Crowdsourced Science', in *Proceedings of the 2012 iConference, 7–10 February 2012, Toronto, ON, Canada* (New York, NY: ACM, 2012), 168–76, see https://doi.org/10.1145/2132176.2132198.

[14] See http://elaborate.huygens.knaw.nl, accessed 20/03/2019.

[15] See http://transcribe-bentham.ucl.ac.uk/td/Transcribe_Bentham, accessed 20/03/2019.

[16] Tim Causer, Justin Tonra, and Valerie Wallace, 'Transcription Maximized; Expense Minimized? Crowdsourcing and Editing *The Collected Works of Jeremy Bentham*', *Literary and Linguistic Computing* 27:2 (2012): 119–37, see https://doi.org/10.1093/llc/fqs004.

[17] See https://www.shakespearesworld.org/, accessed 20/03/2019.

this tool is available on *Github* for anyone who wishes to adopt, modify, or develop it further.[18]

The use of open source technologies and their customization is the strategy adopted by the well-known *Transcribe Bentham* project. The project started in 2010 with the aim to complete the *Collected Works of Jeremy Bentham* by crowdsourcing the transcription of hitherto untranscribed manuscripts.[19] Volunteers transcribe online by means of the *Transcription Desk*, which was custom-build especially for this project.[20]

2.4 Managing Quality Concerns

Although crowd science has many advantages for the academic community as well as for society in general, the quality of citizens' contributions remains a point of concern among scholars.[21] Researchers reject inaccurate or unverified information, because the quality of their research results and their professional legitimacy depends on doing so. Concerns about quality arise from the uncertainty associated with crowd science and from a perceived knowledge boundary between professional scientists and the general public.

Uncertainty emerges because participation in crowd science is voluntary and open to anyone, hence not involving employment contracts.[22] As a consequence, it is not known in advance who will participate.[23] That is, it is a priori not certain how many people will join a project, what type of knowledge they will bring, how much time they will spend, or how much effort they will contribute to the crowdsourced task.

The perception of a knowledge boundary between scientists and the public is deeply grounded. Fully qualified professional researchers have spent years studying

---

[18] See http://elaborate.huygens.knaw.nl/?page_id=107, accessed 20/03/2019.

[19] Causer, Tonra, and Wallace, 'Transcription Maximized; Expense Minimized?', 120.

[20] Ibid., 121–3.

[21] Johan Oomen and Lora Aroyo, 'Crowdsourcing in the Cultural Heritage Domain: Opportunities and Challenges', in *Proceedings of the 5th International Conference on Communities and Technologies, 29 June–2 July 2011, Brisbane, QLD, Australia* (New York, NY: ACM, 2011), 138–49 , see https://doi.org/10.1145/2103354.2103373; Riesch and Potter, 'Citizen Science as Seen by Scientists'; S. Andrew Sheppard, Andrea Wiggins, and Loren Terveen, 'Capturing Quality: Retaining Provenance for Curated Volunteer Monitoring Data', in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing , 15–19 February 2014, Baltimore, MD, USA* (New York, NY: ACM, 2014), 1234–45, see https://doi.org/10.1145/2531602.2531689; Andrea Wiggins, Greg Newman, Robert D. Stevenson, and Kevin Crowston, 'Mechanisms for Data Quality and Validation in Citizen Science', in *Proceedings of the 7th IEEE International Conference on e-Science Workshops (eScienceW 2011), 5–8 December 2011, Stockholm, Sweden* (Piscataway, NJ : IEEE, 2011), 14–9, see https://doi.org/10.1109/eScienceW.2011.27.

[22] Henri Simula, 'The Rise and Fall of Crowdsourcing?', in *Proceedings of the 46th Annual Hawaii International Conference on System Sciences, 7–10 January 2013, Wailea, Maui, HI, USA* (Piscataway, NJ : IEEE, 2013), 2783–91, see https://doi.org/10.1109/HICSS.2013.537; Mark N. Wexler, 'Reconfiguring the Sociology of the Crowd: Exploring Crowdsourcing'.

[23] Franzoni and Sauermann, 'Crowd Science'.

specific topics[24] and mastering established scientific practices by means of rigorous doctoral training and peer review to guarantee the quality of research outcomes. Moreover, the transcription of historical manuscripts is a knowledge-intensive and time-consuming task. It involves reading and interpreting hard-to-decipher handwritten texts –problematic because of variations in length, difficulty of the language, readability of the handwriting, and condition of the paper – and accurately transcribing them.

However, research shows that uncertainty can be managed, and knowledge boundaries can be partially crossed. First, the crowd of volunteers is not as anonymous and massive as the term 'crowd science' can lead us to think. It is often argued that, similar to what occurs in online communities in general,[25] the participation in crowd science follows the Pareto principle, which asserts that only a few people contribute most of the work.[26] For example, although more than 3,000 people had registered to participate in the *Transcribe Bentham* project between October 2012 and June 2014, only eleven of these people transcribed 100 folios or more.[27] Second, some participants in crowd science are more skilled than we may think,[28] in other words, the 'long tail' of expertise includes a 'head' of rather skilled people[29] with knowledge at the margins[30] of specific academic fields.

Finally, recent research has shown that, in order to foster high-quality outputs, the professionals leading crowd science projects invest in knowledge management practices and support the learning process of citizen participants.[31] These knowledge management and quality-assuring practices, including the types of technologies used, differ among projects. For instance, to ensure quality, some project leaders recruit participants by targeting specific communities or knowledgeable individuals, share their knowledge through manuals and face-to-face meetings, and assess the quality of contributions themselves as expert academics. In other crowd science projects, project leaders recruit through open calls, make use of online guidelines, forums, and feedback through e-mails, and select experts among the

[24] Steve Miller, 'Public Understanding of Science at the Crossroads', *Public Understanding of Science* 10:1 (2001): 115–20.

[25] Samer Faraj, Sirkka L. Jarvenpaa, and Ann Majchrzak, 'Knowledge Collaboration in Online Communities', *Organization Science* 22:5 (2011): 1224–39, see https://doi.org/10.1287/orsc.1100.0614.

[26] Franzoni and Sauermann, 'Crowd Science'; Trevor Owens, 'Digital Cultural Heritage and the Crowd', *Curator: The Museum Journal* 56:1 (2013): 121–30, see https://doi.org/10.1111/cura.12012.

[27] Based on the research by Montserrat Prats López, 'Managing Citizen Science in the Humanities: The Challenges of Ensuring Quality', Doctoral Dissertation, Vrije Universiteit Amsterdam, 2017: ch. 3. See http://hdl.handle.net/1871/55271.

[28] Daren C. Brabham, 'The Myth of Amateur Crowds', *Information, Communication & Society* 15:3 (2012): 394–410, see https://doi.org/10.1080/1369118X.2011.641991.

[29] Alpheus Bingham and Dwayne Spradlin, *The Long Tail of Expertise* (London: Pearson Education, 2011).

[30] Lars Bo Jeppesen and Karim R. Lakhani, 'Marginality and Problem-solving Effectiveness in Broadcast Search'.

[31] Based on research by Prats López, 'Managing Citizen Science in the Humanities', chs. 2 and 3.

crowd to review and correct the contributions of other crowd members.[32] Altogether, this indicates that quality concerns, although understandable, are a challenge which is dealt with by managing knowledge flows and controlling the quality of outcomes.

2.5 Consequences of Managing Knowledge and Quality in Crowd Science

Managing knowledge and quality in crowd science seems to go against the open nature of this phenomenon. First, participation in crowd science is said to be open,[33] yet, to ensure quality outcomes, project leaders often target specific communities of people, which limits open participation and the knowledge diversity that goes with it. This is a trade-off that organizations engaging in crowd science projects need to take into account.

Second, despite the usual open access availability of the final outcomes of crowd science,[34] research shows that the intermediate results of crowd science projects are often only made available to the public after crowd contributions have been assessed and corrected, normally by the responsible scientific experts.[35] This seems to imply again a trade-off between openness and quality control. However, quality is usually a priority for crowd science project leaders,[36] and openness is not seen as a dual concept, hence, differing, non-mutually exclusive degrees of openness can be applied[37] at each stage of a crowd science research project.

Moreover, the reluctance to publish intermediate results can also be due to a tension between the open nature of crowd science and the publishing paradigm associated with traditional science. Researchers usually only publish finished work because it is related to their legitimacy and recognition in the scientific community,[38] while crowd science projects allow opening up intermediate outcomes through digital (research) products or online databases that can be expanded and modified over time. This may seem a trade-off between openness and scientific legitimacy, but the emergence and growth of the digital humanities field indicate a potential shift away from the traditional publishing paradigm.

To sum up, if research organizations are to benefit from the potential time and resource efficiency offered by a novel way of organizing research like crowd sci-

---

32 Ibid., ch. 2.

33 Franzoni and Sauermann, 'Crowd Science'.

34 Ibid.

35 Prats López, 'Managing Citizen Science in the Humanities'.

36 Ensuring the quality of contributions from the crowd is a priority for project leaders. This is explicitly discussed in ibid.; and in Riesch and Potter, 'Citizen Science as Seen by Scientists', 116.

37 Angus Whyte and Graham Pryor, 'Open Science in Practice: Researcher Perspectives and Participation', *The International Journal of Digital Curation* 6:1 (2011): 199–213, at 206–7. See https://doi.org/10.2218/ijdc.v6i1.182.

38 Roberta Lamb and Elizabeth Davidson, 'Information and Communication Technology Challenges to Scientific Professional Identity', *The Information Society* 21:1 (2005): 1–24, see https://doi.org/10.1080/01972240590895883.

ence, they need to make sure that researchers become familiar with crowd science's benefits and challenges, and learn how these are dealt with by fellow researchers.

3 Handwritten Text Recognition: *Transkribus* and Learned Correspondence

Thomas Wallnig and Chiara Petrolini

With contributions from Günter Mühlberger

3.1 Handwritten Text Recognition and *Transkribus*

Crowdsourcing uses a digital platform to coordinate the work of a large community of people on a single and relatively simple large task, such as transcribing large bodies of handwritten text. More recently, a more direct application of digital technology to the task of manuscript transcription has been pioneered as well: namely, Handwritten Text Recognition (HTR).

For some time, Optical Character Recognition (OCR)[39] has been applied to the digitization of printed text. First applied to a relatively small range of modern typefaces, OCR has gradually been extended backward to deal with early modern printed pages as well. This technology has been feasible because printing is a mechanical method of reproducing text. In theory, every piece of type cast in the same matrix is identical to every other; and every printed letter produced by such a piece of type is identical to every other. In practice, of course, the imperfection of each stage of this production process inevitably creates variability; but the range of variability is sufficiently narrow to allow even quite basic forms of OCR to produce high-quality digitizations of printed texts.

Applying the same process to handwritten text is far more challenging because of the far greater range of variability in every stage of the writing process. Different periods, cultures, and languages, writing with different implements, fashion the letters of the alphabet in different ways; and so do different writers, and indeed even the same writer in different circumstances. As a consequence, progress in the challenging task of automatic transcription of handwritten text has had to await the arrival of OCR fortified by sophisticated forms of artificial intelligence, namely by the application of a neural network approach to the field of HTR.[40]

---

[39] For a description of OCR practices, see Fotos Jannidis, Hubertus Kohle, and Malte Rehbein, eds., *Digital Humanities. Eine Einführung* (Stuttgart: Metzler, 2017), 193–8 (including bibliography in English). A large corpus of works by Neo-Latin authors processed has been made accessible by way of OCR techniques in the '*Corpus Automatum Multiplex Electorum Neolatinitatis Auctorum*'; see https://www2.uni-mannheim.de/mateo/camenahtdocs/cera_e.html, accessed 20/03/2019.

[40] Gundram Leifert, Tobias Strauß, Tobias Grüning, and Roger Labahn, *Cells in Multidimensional Recurrent Neural Networks*. Eprint arXiv (2014), see https://arxiv.org/abs/1412.2620v2, accessed 20/03/2019.

In the fast-moving field of HTR, *Transkribus* is at the cutting edge of international development.[41] *Transkribus* is a service platform for the transcription, automated recognition, and searching of historical documents. The platform is part of *READ* (Recognition and Enrichment of Archival Documents), an e-Infrastructure project funded by the European Commission and developed by a team around Günter Mühlberger at the University of Innsbruck. It combines two sets of functionalities: the juxtaposition and matching of scans of (handwritten) text with transcriptions (made within the system or imported into it); and the annotation and digital editing of the emerging texts.

Transkribus uses machine learning (neural networks) to train 'models' irrespective of different languages and styles of handwriting. In an iterative process of interaction between the scholar (who creates the transcription) and the software (that 'learns' from the subsequent corrections), a so-called 'model' is created that feeds the specific documentation back into the overall program. The training procedure itself consists of the upload of images, the recognition of baselines and regions of the text on the image, and the manual or automated line-by-line transcription of its content. Once a sufficient quantity of manually transcribed text has been generated by the user (typically about fifty pages), the training process can be started: by pressing a button, the scholar enables the program to 'learn' – that is, to recognize and extract patterns automatically – from what has been fed into the system regarding the scholarly validated relationship between the scanned image and the text.

The trained model can then be applied to additional scans of the same original source, for which no transcription has yet been made. The program then automatically generates a transcription, also for the pages that have already been transcribed. Checking the 'human' transcription against the one generated by the program results in the percentage of characters that have been rendered by the machine in exactly the same way as by the scholar. This indicator of accuracy is called Character Error Rate (CER).[42] With sufficient training data, CER below 5 per cent is achieved in many cases. Once the automated transcription is in place, it can be annotated by means of a set of predefined and customizable Text Encoding Initiative (TEI)-compatible tags.

Core modules of *Transkribus* were developed by research groups from the universities of Valencia and Rostock. The consortium has been drawing on sources, among others, within the FP7 and H2020 framework programs of the European

[41] For a general introduction to the scope and functionalities of the program, see https://transkribus.eu/wiki/index.php/Main_Page. For the broader context of the related research infrastructure, see http://transcriptorium.eu/; https://read.transkribus.eu/; both accessed 20/03/2019.

[42] However, results of CER have to be considered with caution, in that 'characters' that have been rendered in the wrong way by the program may include signs of punctuation, or spots or stains on the original that do not belong to the text at all.

Union. This means that the platform is in constant development, with version 1.5.0 being the current release (in late 2018).

The field is developing very quickly, and the status reflected in the following paragraphs should thus be read with due caution. The presented case study dates back to 2015–16, and it is determined by the state of knowledge of the respective participants, as well as by the then actual capabilities of the tool itself.

3.2 Testing *Transkribus* at the University of Vienna, 2016

In 2015, a pair of workshops was organized at the University of Vienna (UV) as part of a longer series devoted to testing the utility of *Transkribus* for academics from a range of different fields. The samples of text employed for testing were drawn from the following sources: the registers of Pope Innocent III (Latin, early thirteenth century); the 'Magnum legendarium Austriacum' (Latin, c. 1200); Byzantine prayer books (Greek, mainly twelfth to fourteenth centuries); diaries kept in the UV's collection of women's estates (Sammlung Frauennachlässe; mainly German, nineteenth to twentieth centuries) as well as in its collection of life accounts (Dokumentation lebensgeschichtlicher Aufzeichnungen; mainly German, nineteenth to twentieth centuries); and likewise from a diary by Hans Posse, national socialist art collection envoy (d. 1942); eventually, the sample also included materials from the correspondence and papers of the brothers Pez OSB (d. 1735 and 1762, respectively).[43]

During the first workshop, the basic handling of *Transkribus* was explained on the basis of sample scans brought along by the participants: uploading the scans, generating transcriptions (by copy–paste from extant documents or manually, but in any case line by line), working with annotations, and eventually creating a validation file as the basis for an assessment of the CER. The second workshop (two months later) was dedicated to the evaluation of work done with sample data. User feedback pointed to the usability of the interface (e.g. distinguishing between expert and non-expert user interfaces for easier navigation) and the legal status of uploaded material: all uploaded files can be deleted and will not be shared without the consent of the holder. The majority of the participants also successfully managed to define text regions and baselines within the scans. In most cases, the manual definition seemed preferable to the automated process. An element of particular interest was the rearrangement of the reading order – for instance, in the case of insertions or marginalia – and the improvement of the respective functionalities regarding the line sequence.

A fundamental point is that *Transkribus* primarily requires (from the scholar) and generates (as a program) diplomatic transcriptions, not edited text. This explains the above-mentioned integration of HTR and digital editing functionalities, which allow for normalizing annotation of an exactly reproduced sequence of

[43] See https://www.univie.ac.at/monastische_aufklaerung/en/startseite.html, accessed 20/03/2019.

characters. This becomes particularly evident in the case of highly abbreviated Latin texts. The machine can be told to match an image with the exact sequence of characters as well as with a predefined set of abbreviation graphemes, but this has to be done manually.

Once in place, the text – matched to the image – can be annotated with a view to a critical digital edition. The program can express specific textual configurations like strikethroughs, corrections, changing ink colours, or initials. Furthermore, the text can also be annotated semantically, e.g. in terms of person and place names. No problems arose from the exporting functionalities that allow a variety of export options for the text, images, and tags – e.g. as PDF, MS Word, or MS Excel files.

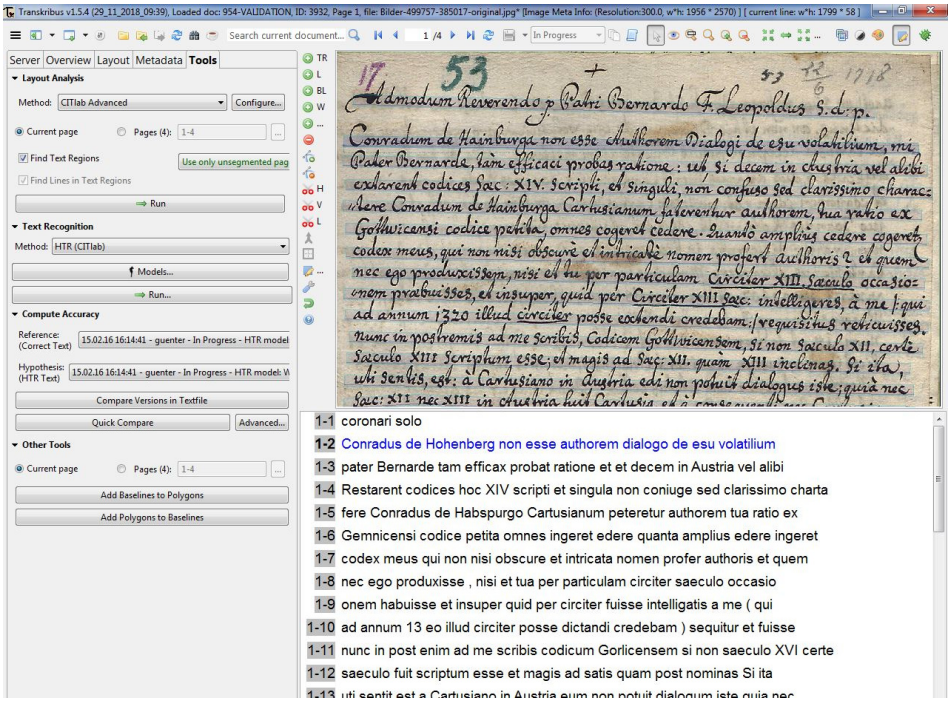


Figure 1: Validation file for a letter from Leopold Wydemann OCart to Bernhard Pez, 1718

With a sufficiently large sample (typically of about thirty pages) and the respective scans in place, it was possible to initiate the training process of the machine. This process was carried out by the Innsbruck team. The key to the process is a sequence of algorithms which first provide a variety of reading suggestions for each sequence of characters, and then calculate the probability of each suggestion (see fig. 2). In older versions of *Transkribus*, the production of sufficiently accurate results depended heavily on the availability of background vocabularies for individual

languages, against which the reading variants could be checked. But with the implementation of new modules, and particularly neural network technologies, in the recent versions of *Transkribus*, the recognition is done solely on visual features, without reading them against the options recorded in the dictionaries. In fact, once the character error rates produced by HTR fall below 5 per cent, the supplementary use of dictionaries does not usually improve the accuracy further.

Figure 1 and 2 give an impression of some of the steps in the process. In the first image (see Fig. 1), the scan of a letter is juxtaposed to its first automated transcription (2016), generated on the basis of a model that had taken into account c. fifty manually transcribed pages of the same handwriting. At that time, the validation file featured a word error rate of 55 per cent, and a character error rate of 22 per cent (including punctuation). In 2017, tests with the same material, after additional training of the model, produced results of c. 4–5 per cent CER.

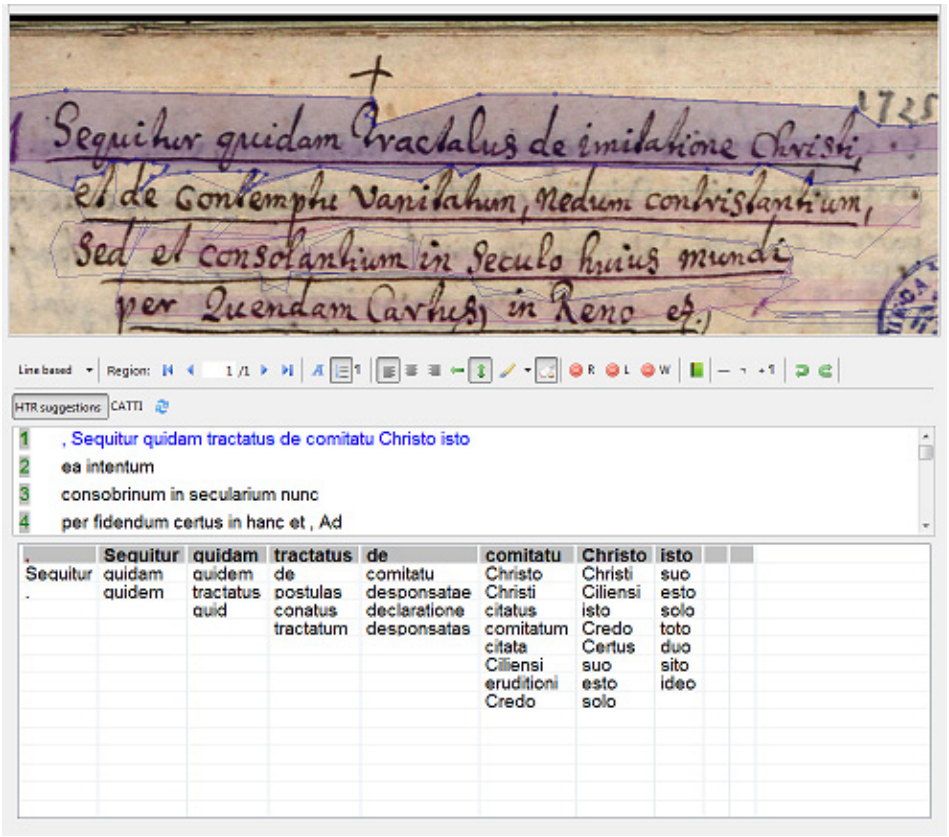


Figure 2: Automated transcription and suggestions, copy of a treatise (by Leopold Wydemann) in the Pez papers (feature available in previous versions of *Transkribus*)

Figure 2 illustrates the way in which *Transkribus* offers a list of alternative readings for each word.

The Vienna workshops have provided the participants with an outline of the workflows, the approximate resource requirements, and the expected outcomes. On this basis, one may distinguish between at least five possible scenarios:

• *Use of Transkribus as a manual transcription and digital edition tool.* Transcribing itself does not become faster, but the digital nature of the transcription and digital edition make it more easily interoperable (with other similar sources, repositories, and databases).

• *Use of Transkribus as a tool for automatic generation of large quantities of draft transcriptions.* Editorial experience shows that any draft transcription, even if done manually, has to be collated by a scholar. If there is no funding to have persons generate such draft transcriptions – and if the machine outcome is sufficiently satisfactory – it can make sense to begin (larger) editorial projects by having the machine create a first draft transcription.

• *Use of Transkribus as a tool for automatic generation of large quantities of draft transcriptions and apply keyword search.* The probability algorithms described above make it possible to search for keywords within all 'probable' readings of a text (with customizable degrees of probability). With due respect for the current experimental nature of this feature, this functionality could become attractive in making large numbers of scans (e.g. from archival or library contexts) automatically searchable. For example, a pilot project is currently underway on the early modern records of the Dutch Republic's Estates General.[44]

• *Use of Transkribus as a tool for automatic generation of large quantities of draft transcriptions and selective annotation.* If work is being done on smaller corpora, it is also conceivable to annotate manually (and correct paleographically) only the names of persons and places (see below). This would, e.g. in the case of correspondence corpora, reconcile the training effect for the machine with the soundness of the scholar-generated indices. If desirable, one could select individual draft transcriptions and transform them into digital editions.

• *Use of Transkribus as an e-learning tool.* This functionality is currently under development and will be available for both teaching and citizen science applications, for generating transcriptions, and for practising reading skills. It is not inconceivable that it might also be used for matters of quality control.

3.3 Applying *Transkribus*: The Tengnagel Correspondence Project, 2018

As stated above, unpublished corpora of learned correspondence represent a particularly difficult challenge for HTR. They are usually not particularly vast (in comparison to official records, for instance), they contain handwriting of many differ-

[44] See https://www.huygens.knaw.nl/workshop-automated-handwritten-text-recognition-transkribus-and-the-resolutions-of-the-dutch-states-general/, accessed 20/03/2019.

ent people often in non-standardized forms, and they often contain only small samples of text by one and the same hand, which are sometimes written in haste. In order to test the assumption that selective annotation might be the best way to make use of *Transkribus* in such a setting, a Vienna-based project, funded by the Austrian Science Fund until 2020, has developed a respective test scenario.

The project focuses on the figure of Sebastian Tengnagel – librarian at the Imperial Library in Vienna from 1608 to his death in 1636 – his correspondence, and the collection of oriental manuscripts at the court library.[45] Four volumes of Tengnagel's letters, held at the Österreichische Nationalbibliothek (Cod. 9737$^{q\text{-}t}$), have been digitized and uploaded onto the central Transcription and Recognition Platform (TRP) server. First experiments have been carried out with Cod. 9737^{t}. This volume consists of 232 letters (making a total amount of 714 digitized pages as JPEG images) written between 1609 and 1635 by forty different correspondents who used not only different languages but also different characters: within a single page by Elias Opala (Cod. 9737^{t}, 320r), for example, we can find six different alphabets.

Once uploaded, the letters are being processed in the way described in the previous section, and then selectively annotated. In this way, a close link between the transcribed text and the materiality of the transcribed manuscript is being maintained, and the complex and sometimes creative structure of the early modern document can easily be preserved: *Transkribus* makes it possible to combine the text and images of the original manuscript, and to make the final image fully searchable. It allows the document to be separated into multiple areas, fully respecting the original structure and making immediately clear, for instance, the intervention of different hands and a number of chronological layers.

Transkribus also uses an inherently multi-user system providing interaction and dialogue among multiple users as well as control over others' actions through versioning. In the case of Tengnagel's letters, this means that different scholars with specific skills can work on the same document in different locations, editing the transcription, helping to decipher difficult passages or words, and – most importantly – transcribing and translating the passages in Arabic, Hebrew, Ottoman, Greek, and other alphabets which often figure in Tengnagel's correspondence (recently the virtual keyboard has been extended to Hebrew and Arabic).

This collaborative approach is essential to the Tengnagel project, and thus the recently developed web interface for simplified transcription is a particularly welcome feature. It makes it very easy to share a single document or a collection of documents even with users less familiar with *Transkribus* by just sharing a link. Without installing the software, they will be able to contribute to the transcription and to add personal comments and annotation of information in Tengnagel's papers. It would also be conceivable to improve the space for individual comments

[45] See http://www.univie.ac.at/oorpl, accessed 20/03/2019.

by way of a wiki so that dialogue about any letter, page, or single sentence could be documented.

Another crucial function of *Transkribus* is tagging and annotation. It is very easy to create new tags (that is, standardized, machine-readable labels to be attached to various parts of the transcribed text). In Tengnagel's case we have created, among others, tags for book titles, for languages such as Arabic, Greek, and Hebrew, for quotations, abbreviations, doubts in transcription, and so on, while tags for person names and place names already exist. *Transkribus* makes it possible to export any single page, group of pages, or entire document, in various formats (including PDF, DOCX, TEI, RTF). The user can therefore choose to export just the images, the transcription, or the tags.

However, what remains most exciting about *Transkribus* is the way in which its cumulative machine learning ability reshapes access to handwritten historical documents. Once a certain number of folios are transcribed, it is possible to train an HTR model and to perform an automated transcription of the texts. Every user benefits from the work of other users because the data are collected centrally, even though it is possible to keep each collection private, without the need to share documents directly. Within the Tengnagel project, a model has been created (using 5,478 transcribed words) that will be improved as the work continues. However, the project focuses more on enriching metadata by selective annotation; it will not go 'beyond' the reading process into digital editing, but rather think about metadata creation, pattern discovery, network analysis, and similar fields. Finally, since late 2017, the so-called Keyword Spotting feature has been included in *Transkribus*. This enables users to search for words directly in the image – and not only in the transcribed full text. This is a much more powerful method, and when recognition rates rise above 20–30 per cent CER, all words will be found with a high degree of confidence.

4 Producing Scholarly Digital Editions

Elena Spadini

In order to engage with the text of the letters of a correspondence, for distant as well as close reading and for producing a digital edition, the text must be available in electronic format. In the previous sections of this chapter, some of the methods for obtaining the text of a work in a suitable medium have been described, such as crowdsourcing transcription and automatic handwriting recognition. In this section, we focus on complementary strategies for producing scholarly editions in a digital framework.

A scholarly edition presents a reliable text and can take different shapes. In the case of a work transmitted by a single witness (that is, a manuscript, print, or born-digital copy), the text will be transcribed and normalized (more or less, from dip-

lomatic to interpretative), producing a documentary edition; this is mostly the case for correspondences. In the case of a work transmitted by various witnesses, a critical text is established, based on one witness or on the stemmatic reconstruction of the archetype, and accompanied by the variants from the other witnesses. A scholarly edition also provides materials useful for understanding the work's form and content, such as notes, introductions, apparatus, glossary, and others.

In the digital realm, a distinction must be made between digitized and digital editions. Digitized editions are the product of the digitization of an existing printed edition to make it available electronically, for instance, in PDF format. A digital edition, on the contrary, includes contents and functionalities (for example, advanced search, on-the-fly collation, interactive visualizations) that would be lost if converted to a printed medium.[46]

In what follows, we focus on different aspects of the creation of a scholarly digital edition. First, we define the term 'editing tools', which is the basic unit of an editorial platform, and provide a comparative study of a particular type of tool for transcription and encoding. Second, we explore the role of standards for encoding and annotation in the development of such tools. Finally, we briefly address the use of distant reading strategies and of Natural Language Processing (NLP) techniques in this context. The three sections pursue tasks that are connected, while potentially independent. Together, they provide insights into the process of creating a scholarly digital edition.

4.1 Editing Tools: A Comparative Analysis of Transcribing and Encoding Tools

An editing tool is here defined as a piece of software used in the creation of a scholarly digital edition: such a tool can potentially be used for transcription, annotation, collation, and, ideally, all the other tasks involved in the process.[47,48] A

[46] See Patrick Sahle, *Digitale Editionsformen, Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels*, vol. 3 (Norderstedt: Books on Demand, 2013), 141–2 and 149; Greta Franzini, 'Catalogue Digital Editions', https://dig-ed-cat.acdh.oeaw.ac.at/, accessed 20/03/2019.

[47] This section presents extracts from the DiXiT report 'Editing Tools. Transcribing and Encoding', including questions concerning the entire workflow necessary for the creation of a scholarly digital edition, see http://hdl.handle.net/20.500.11755/7227e906-2bad-4b8a-9611-1dd351d8bb85. A special issue of the journal *RIDE* (https://ride.i-d-e.de/) dedicated to tools and environments for digital scholarly editing is in preparation (November 2018); all accessed 20/03/2019.

[48] Considering several formalizations of the editing task and the peculiarities of a digital project, we can list: collection of witnesses (doc/image management and metadata), transcription, encoding, named-entity recognition, semantic enrichment, collation, analysis, constitution of the critical (or copy) text, compilation of apparatuses, compilation of indexes, preparation of paratextual material, data visualization. This list leaves out the final step of publication and might not include project-specific tasks. See for example Paul Maas, *Textkritik* (Leipzig: Teubner, 1927); Michael L. West, *Textual Criticism and Editorial Technique* (Stuttgart: Teubner, 1973); Tara Andrews, 'Digital Techniques for Critical Edition', in Valentina Calzolari and Michael E. Stone, eds., *Armenian Philology in the Modern Era: From Manuscript to Digital Text* (Leiden: Brill, 2014), 175–95, see https://doi.org/10.1163/9789004270961_008; Wilhelm Ott, 'Strategies and Tools for Textual Scholarship: The Tübingen System of Text Processing Programs (TUSTEP)', *Literary and Linguistic Computing* 15:1 (2000): 93–108,

number of editing tools can be combined in an editorial platform (also known as 'editing environment', or 'workbench'). Word processors, concordancers, or lemmatizers are programs, that is, digital tools. Dictionaries and glossaries can be computer applications or appear in print: that is, these tools may or may not be digital. 'Extending the toolkit of traditional scholarship'[49] is doubtless one of the aims of digital humanities. The core of each discipline can be found in its toolkit and its applicability: a tailor might not anticipate the next dress to be commissioned by her or his client, but with the right measuring tape, thimble, shears, and other tools it can be made. Within digital media, scholars have created digital equivalents of non-digital media. Furthermore, certain tools can only be available in electronic format. As Andrews points out, the revolutionary aspects of digital scholarly editing are not only related to the new publication media, but mostly consist in what happens, behind the scenes and behind the screen, in the process of editing with the aid of a partially new toolkit for textual scholarship.[50]

The development of ad hoc tools for the creation of a scholarly edition is not new. The history of editing tools remains to be written, but some pioneering projects in the field are well known. *TUSTEP*, for instance, is a toolbox for the scholarly processing of textual data, designed at the Computing Center of the University of Tübingen, first implemented in the 1970s and constantly upgraded until today. *Collate* is a collation tool developed by Peter Robinson in the early 1990s, which has only recently been superseded by its successor, *CollateX*.[51] Theoretical reflection on digital tools for the humanities has been on the increase in past decades, fostered by landmark writings,[52] projects[53] and conferences.[54] On a practical note, a number of resources are available today; two main repositories of humanities applications, not focused only on editing but more generally on research tools for

see https://doi.org/10.1093/llc/15.1.93; Joris van Zundert and Peter Boot, 'The Digital Edition 2.0 and the Digital Library: Services, Not Resources', *Bibliothek und Wissenschaft* 44 (2011): 141–52.

[49] Anne Burdick, Johanna Drucker, Peter Lunenfeld, Todd Presner, and Jeffrey Schnapp, *Digital Humanities* (Cambridge, MA: The MIT Press, 2012), 12.

[50] Tara Andrews, 'The Third Way: Philology and Critical Edition in the Digital Age', *Variants* 10 (2013): 61–76, see https://doi.org/10.1163/9789401209021_006.

[51] Ronald Dekker and Gregor Middell, *CollateX*, 2010, see http://collatex.net/, accessed 20/03/2019.

[52] See for example John Bradley, 'Tools to Augment Scholarly Activity: An Architecture to Support Text Analysis', in Harold Short, Dino Buzzetti, and Giuliano Pancaldi, eds., *Augmenting Comprehension Digital Tools and the History of Ideas* (London: Office for Humanities Communication, 2002), 19–48; John Unsworth, 'Tool-time, or "Haven't We Been Here Already?": Ten Years in Humanities Computing' (Washington, DC, 2003); Willard McCarty, *Humanities Computing* (Basingstoke [England] and New York: Palgrave Macmillan, 2005).

[53] E.g., 'Project Bamboo', see https://wikihub.berkeley.edu/display/pbamboo/Documentation; 'Interedition', see http://www.interedition.eu/, both accessed 20/03/2019.

[54] Recently, *Easy Tools for Difficult Texts*, Cost Action IS1005 'Medioevo europeo' and Huygens ING, The Hague, April 2013; *Research Summit on Collation of Ancient and Medieval Texts*, COST Action IS1005 'Medioevo europeo' (Münster, October 2014); *Scholarship in Software, Software as Scholarship: From Genesis to Peer Review* (University of Bern, January 2015).

scholarly use, are *Digital Research Tools Directory* (DIRT)[55] and *Research Tools for Textual Studies* (TAPOR).[56]

Given the availability of so many tools for scholarly editing, a discussion of them all might quickly become overwhelming. A selection of them has therefore been analysed in the tables below, focusing on encoding and transcribing while also considering collation for one of the environments. This selection follows a strictly empirical criterion of 'user-friendliness'. This selection has been restricted to tools that require minimal computer literacy,[57] and to browser-based or portable applications, for which no installation is needed.[58] The tools that have been analysed are *T-Pen* and *CWRC-Writer*; the environments are *eLaborate*, *TextGrid*, and *Ecdosis*. All them have been tested[59] and compared. These tools and environments differ in so many different ways as to make a systematic comparison difficult. Some of their points of variation are represented in tables 1, 2, and 3 below.

Table 1: Comparative analysis of tools and environments (1)

	Web-based or standalone	**Licence**	**Documentation**
T-PEN	Web-based	Open source ECL-2.0	Users
CWRC	Web-based	Open source	Users
eLaborate 4	Web-based	Open source GNU GPLv3	Users and developers
TextGrid	Portable (connection with the server needed)	Open source Policy available on the website.	Users and developers
Ecdosis	Web-based	Open source	

[55] 'Digital Research Tools', http://dirtdirectory.org/, accessed 20/03/2019.

[56] 'Tapor', http://www.tapor.ca/, accessed 20/03/2019.

[57] Minimal computer literacy includes here basic knowledge of XML, but not of programming; also, the tools should be usable without consulting complex manuals.

[58] A browser application is a computer program which runs online, accessible through a website in the browser. *TextGrid* is the only software not running in the browser that is taken into account here; it is portable, in the sense that it does not require installation, but just to be copied and run.

[59] For an in-depth analysis, see note 47.

Table 1 describes technical features: all the selected tools and environments are web-based or portable, and have been released open source. Regarding the documentation, the solutions adopted vary greatly, and have consequences for ease of use by scholars and developers.

Table 2: Comparative analysis of tools and environments (2)

	Steps of the editing process	**Annotation and markup**	**Import and export**
T-PEN	Metadata management, transcription, encoding, annotation	Annotation. Possible TEI encoding	Import: TXT and XML Export: PDF, XML, plain text, HTML
CWRC	Metadata management, transcription, encoding, semantic enrichment	TEI encoding (*embedded*) and RDF encoding (*standoff*)	Import: plain text
eLaborate 4	Metadata management, transcription, annotation, publication	Annotation	Import: plain text Export: plain text, XML (not available on the normal user interface)
TextGrid	Metadata management, transcription, encoding, collation, lexicon creation, paratext creation, publication (+ Lemmatizer for German texts)	TEI encoding	Import: plain text, XML Export: plain text, XML
Ecdosis	Metadata management, transcription, encoding, paratext creation, event editor, publication	Markdown encoding	Import: plain text, XML Export: plain text, XML

In the first column of table 2, the distinction between tools and environments is evident, with the latter covering a higher number of steps of the editing process. The second column differentiates tools that allow for the encoding of the text from those offering annotation facilities, in combination or in alternative to the encoding. The import and export formats, mostly plain text or XML, detailed in column 3, are also dependent on the encoding choice.

The characteristics explored in table 3 concern the possibility to use these software packages as complete workbenches for the editing process: the image management, the search functionality, and the resources for online collaboration are important facilities common to the majority of the tools and environments under analysis.

Table 3: Comparative analysis of tools and environments (3)

	Search functionalities	**Image management**	**Online collaboration**
T-PEN	No	Advanced Text/Image link (line level)	Yes Leader project and users with different rights
CWRC	No	No	No
eLaborate 4	In metadata: advanced and *friendly* In text: full text search	Advanced Text/image link (page level)	Yes Leader project and users with different rights
TextGrid	Full text search in the project metadata and in TextGrid Repository	Advanced Text/image link (manual)	Yes Leader project and users with different rights
Ecdosis	No	Advanced Text/image link (word level)	Yes Leader project and users with different rights

4.2 Standards for Text Encoding and Annotation

The proliferation of transcribing and encoding tools depends to some degree on the acceptance of the TEI[60] as a standard for the encoding of texts.

Encoding is a fundamental step: a way of putting 'intelligence in the text'[61] – a kind of intelligence that computational tools can process. In practice, encoding in this context refers to the practice of inserting tags into the text, in order to label a portion of it (a letter, a word, a sentence or any combination of these) for the purpose of identifying or adding information. Typical examples include: the identification of a semantic entity, such as a name or a date; the identification of a structural entity, such as a title or a paragraph; or additional information, such as the expansion of an abbreviation or the comparison with other witnesses. The tags, or elements, labelling portions of text in this way constitute the 'markup'.

TEI provides Guidelines for this purpose and the data format it uses is XML.[62] XML-TEI is not the only existing data format for creating Scholarly Digital Editions: other XML languages, markup (such as LaTeX) and markdown syntax, or the code of web pages, HTML, can be used. However, XML-TEI markup, extremely rich and explicitly devoted to text encoding, has become a standard for transcribing and editing literary texts and historical sources, which are normally the objects of scholarly editions. The TEI Guidelines provide around 500 tags for marking up all sorts of phenomena occurring in texts. A quick look at the Guidelines Table of Contents will suffice: the twenty-three chapters are devoted each to a different aspect of text encoding, including verse, performance texts, dictionaries, language corpora, writing modes, linking, and many others.

The use of TEI as a standard plays an important role in the development of tools: a TEI-compliant tool is useful for a large community of practitioners,[63] and this is one of the reasons for the production of increasing numbers of transcribing and encoding tools that use this standard.

Using TEI for encoding correspondence material (letters, postcards, billets, etc.) is equally common. The Guidelines provide specific tags for the markup of the text and for the corresponding metadata (that is information about the document such as sender, receiver, date).

For the text, in particular, the tags for encoding default structures are available (TEI Guidelines, ch. 4: Default Text Structure); some of the tags frequently used

---

60 'Text Encoding Initiative', see http://www.tei-c.org/, accessed 20/03/2019.

61 Susan Hockey, *Electronic Texts in the Humanities. Principles and Practice* (New York: Oxford University Press, 2000), vi.

62 The history of the Text Encoding Initiative is deeply intertwined with that of humanities computing (or digital humanities) and of the XML specification. See 'TEI: History', see http://www.tei-c.org/About/history.xml, accessed 20/03/2019.

63 As the TEI Guidelines are extremely rich and, more importantly, the object to encode, that is – text – is equally variegated, a customization of the TEI is likely to be used in each specific project. This makes it difficult to provide tools that can work out of the box for every single project, without some degree of customization.

for letters are <opener>, <closer>, <dateline>, <salute>, <signed>, <postscript>.[64]

For metadata, the Correspondences Special Interest Group[65] proposed a new section, which was integrated in the TEI Guidelines in Spring 2015. The element carrying the correspondence metadata is <correspDesc> (correspondence description).[66] Together with its subset of tags, it can be used to encode information about the sending, the receipt, the transmission, the redirection, and the forwarding of a message.

Using the Correspondence Description model as an interoperable standard for the encoding of correspondence metadata, the research group TELOTA has promoted the project *correspSearch*, with the aim of indexing letter collections. The database, which can be queried on the project website, already hosts more than 27,000 letters.[67]

Before the addition of the Correspondence Description section to the TEI Guidelines, one of the most successful projects in the realm of XML encoding of correspondence has been the *Digital Archive of Letters in Flanders* (DALF)[68] created by the Centrum voor Teksteditie en Bronnenstudie. The schema designed is an extension of the TEI P4 DTD,[69] and has been adopted with modifications in other projects, as the well-known edition of Van Gogh's letters.[70]

A TEI encoding can also be considered complementary, and not alternative, to other forms of annotations. The Semantic Web and its technical standards, for instance, are increasingly central in structuring and representing cultural heritage data. A mixed use of XML-TEI and XML-RDF has been already implemented in a few projects devoted to correspondences, among which are *Vespasiano da Bisticci. Lettere*[71] and *Burckhardtsource*.[72]

---

[64] The definitions of these elements, together with technical information and examples, are part of the TEI Guidelines. In particular, 'Elements Common to All Divisions', see http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DS.html#DSDTB, accessed 20/03/2019.

[65] Information on Special Interest Groups is available at 'TEI: SIGs', see http://www.tei-c.org/activities/sig/, accessed 20/03/2019.

[66] 'Correspondence Description', see http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD44CD, accessed 20/03/2019.

[67] 'correspSearch', see http://correspsearch.net/index.xql, accessed 20/03/2019.

[68] Centrum voor Teksteditie en Bronnenstudie. 'DALF', see http://ctb.kantl.be/project/dalf/, accessed 20/03/2019.

[69] An XML schema defines which tags, where, and how they should be used in an XML document. The DTD format for schemas has now been replaced by RNG. A TEI customization is expressed through an XML/ODD document. P4 refers to a previous release of the TEI Guidelines, the current one being P5.

[70] Jansen Leo, Hans Luijten, and Nienke Bakker, eds., *Vincent van Gogh: The Letters* (Amsterdam and The Hague: Van Gogh Museum & Huygens ING, 2009), see http://vangoghletters.org, accessed 20/03/2019.

[71] See in particular 'La base di conoscenza', at 'Vespasiano da Bisticci, Lettere', see http://vespasianodabisticciletters.unibo.it/base_conoscenza.html, accessed 20/03/2019. In this case, RDF triples are mainly used to build knowledge in the metadata. From occurrences inside the text, a link is established to the metadata, using TEI; the metadata and the relations between them (i.e. Linked Data) are then stored as triples in external RDF files, connecting to others data sets. An ad

4.3 NLP and Editions as Data

In a scholarly digital edition, the images of the documents, texts, and additional materials (notes, explanations, introductions) are not organized in a book, but digitally: all the information becomes electronic data, which is organized in an information architecture.[73]

Once the scholars engage with data, new forms of analysis become possible. Text in digital form is data, to which NLP techniques or Data Mining algorithms can be applied. Furthermore, data is structured: every individual structure can be statistically inspected and visualized. A closer look at each of these possibilities will clarify the potential of this approach.

NLP is a comprehensive label, referring to the processing and analysis of text and speech by means of computer applications. Among the techniques that are most used in the field of scholarly editing, we can find algorithms for linguistic annotation (such as part-of-speech tagging or stemming) and for Named Entity Recognition (NER). While the former adds information to the word, automatically annexing the corresponding part-of-speech or stem, the latter identifies the proper nouns in a text. NER can also be used for automatically encoding semantic entities such as names of persons and places (for instance, in TEI, see above).

NLP techniques might indeed be useful in the process of creating an edition, just as much as linguistic competences have always been necessary for editing a text. Data Mining, on the other hand, would mostly be used at the end of the process, exploiting the edition as a *corpus* for investigation. Data Mining can either take the form of quantitative analysis leading to data visualization or consist of the application of statistical models, as in the case of topic modelling, stylometry, or sentiment analysis.

A look at the current state of the digital editing field shows that projects seldom embrace the approach just described, taking advantage of the (generally) large amount of data produced in scholarly editing. In most of the cases, the editorial work, the linguistic analysis, and the text analysis remain separate, reproducing the traditional distinction between textual criticism and literary criticism. NLP and Data Mining techniques are not admitted into the editing process; thus their results cannot be integrated in the auxiliary materials provided alongside the text in a scholarly edition.

This state of the art might have multiple explanations. Among these is the fact that for a long time, producing scholarly editions has been considered an ancillary

---

hoc ontology has been built, in order to handle the variety of relations among data specific to the project.

72 'Burckhardt Source Project', see http://burckhardtsource.org/, accessed 20/03/2019. In this edition, an HTML version is generated from the TEI-XML encoding, following common practices. Annotations (triples) are finally created, using the tool *Pundit* that allows annotation of web pages.

73 An information architecture is simply some structure for organizing the data, for instance, the schema for the markup (see above section 4.2), the ensemble of database fields or the ontology for Linked Data.

activity in relation to historical and literary studies (*philologia ancilla historiae*), ratifying their separation of concerns. Most importantly, in the context of digital editing, an edition has not yet come to be regarded as a set of data that can be processed by means of a variety of algorithms and visualized in different ways: in the words of Cummings and colleagues, 'Many editors often think they have "only text", not "data"; but the structures created by the edition contain more data than they think'.[74]

This 'misunderstanding' of the nature of text and data and of what can be done with them appears to be stronger whenever literary materials are involved: it seems less pronounced for the edition and analysis of historical, and more generally cultural, sources. Indeed, it is in the edition of correspondences that some advances in the integration of scholarly editing, NLP, and Data Mining are to be found. The *ePistolarium* project,[75] providing the edition of letters of scholars active in the Netherlands in the seventeenth century, is at the forefront of this process. Using techniques such as topic modelling and keyword analysis in combination with NLP, it offers meaningful ways to engage with the text of correspondences. As such, it serves as an inspiring example for imagining the digital scholarly editions of the future.

5 Digital Provenance of Texts and Additional Resources: EMED

Elizabeth R. Williamson

Reassembling the republic of letters requires large-scale collaboration on previously unedited manuscript sources, but no less important is the repurposing of existing print and digital materials which have already absorbed vast amounts of scholarly labour. More generally still, any collaboratively compiled catalogue, archive, or edition builds on the work of many scholars, sometimes spread over several generations. This section will suggest how much care and critical attention will be required to ensure that due weight is given to the layered histories, agencies, and labour buried within any collaboratively crafted resource of this kind.

Digital editing itself has the potential to be a highly collaborative and highly iterative process: not only can many people work simultaneously or successively on a digital edition, but digital data in standardized formats lends itself to reuse by further groups with divergent purposes at different times. This potential introduces complexities around provenance, attribution, and acknowledgement that are rele-

[74] James Cummings, Martin Hadley, and Howard Noble, 'It Has Moving Parts! Interactive Visualisations in Digital Publications', paper presented at the DiXiT Workshop 'The Educational and Social Impact of Digital Scholarly Editions', Rome, 2017. See http://dixit.uni-koeln.de/programme/materials/#aiucd2017, accessed 20/03/2019.

[75] *ePistolarium*, see http://ckcc.huygens.knaw.nl/epistolarium/, accessed 20/03/2019.

vant both to contributors and end users, and are heightened when an edition makes use of existing scholarship. Accordingly, this section explores the notion of 'collaborative digital provenance' with reference to the Folger Shakespeare Library's resource, *A Digital Anthology of Early Modern English Drama* (EMED).[76] The idea that digital projects have an important provenance or history of use and participation is often elided, or at least not attended to sufficiently, in the race to produce a smooth user experience and glossy finish. Attending to digital provenance is essential in order to credit labour, acknowledge previous decisions and limitations on inherited data, and understand fully the place of the digital in archival history and resource creation broadly writ, especially considering the fact that digital humanities projects often associate themselves with the open source philosophy and involve or build upon legacy projects (both digital and print). EMED is a research collection and digital edition that is overtly concerned with its own complex inheritance, and with its relationship with sister and legacy projects. Further, the editors attempt to use this acknowledgement of complexity as a teaching and learning opportunity by scaffolding their texts with additional resources. EMED is thus introduced as a means to raise issues that are highly pertinent to the collaborative, iterative creation of a textual archive capable of illustrating the republic of letters.

EMED is a resource for exploring early modern English plays performed between 1576 and 1642 by London's professional companies and written primarily by people other than Shakespeare. In order to be included in the project, each play had to survive in at least one edition printed before 1660. This provided a manageable corpus of 403 plays. EMED was published by the Folger Shakespeare Library in 2016–17. The editing and encoding team was Meaghan Brown, Elizabeth Williamson, and Michael Poston. EMED provides extensive copy-specific metadata for all 403 plays, as well as documentary editions of a subset of twenty-nine plays, and additional resources for learning and teaching. The editing team worked with the Roy Rosenzweig Center for History and New Media at George Mason University to create a search and browsing portal for the play metadata using Drupal. Users can easily delve into the standardized, linked data to explore an author's works, a company's repertoire, a publisher's output, and so on. EMED also links outwards to freely available critical editions, including those on Digital Renaissance Editions, as well as to records in the English Short Title Catalogue, and in the *Database of Early English Playbooks* (DEEP).[77] The metadata for the plays was carefully curated from multiple sources, including images and metadata from *Early English Books Online*, Martin Wiggins's *British Drama 1533–1642: A Catalogue*, and the

[76] Meaghan Brown, Michael Poston, and Elizabeth Williamson, *A Digital Anthology of Early Modern English Drama* (Folger Shakespeare Library), see https://emed.folger.edu/, accessed 20/03/ 2019.

[77] *Digital Renaissance Editions*. Internet Shakespeare Editions, University of Victoria, 4 April 2015, see http://digitalrenaissance.uvic.ca/, accessed 20/03/2019; *English Short Title Catalogue* (ESTC), British Library, see http://estc.bl.uk/, accessed 20/03/2019; Alan B. Farmer and Zachary Lesser, DEEP: *Database of Early English Playbooks*, 2007, see http://deep.sas.upenn.edu, accessed 20/03/2019.

DEEP database.[78] DEEP has a broader scope than EMED, as it includes variants, issues, and subsequent editions, and so this is not simply duplication of material. Explicit permission was obtained to reproduce DEEP's information on paratextual materials; such sharing of data enabled EMED to offer more data on the platform, as well as share amendments with DEEP where very occasional corrections were made.

Being able to share and update collaboratively is a major advantage to open digital projects. Equally, being clear about where data comes from is essential to encourage responsible use and reuse of that information. In collecting and refining data for EMED, the editors encountered several examples of how metadata on *Early English Books Online* (EEBO) had become corrupted somewhere along the remediation process (which for EEBO is complex, as explained further below). This meant that the editors at times became detectives, for example identifying a holding repository after spotting a Folger Shakespeare Library ruler on one image in a play supposedly held by the British Library, or identifying a 'franken-edition' that had been silently amalgamated online from different copies of the same edition in different libraries.[79] The risk of error and compounded erroneous assumption is introduced whenever data and files move across media.

This risk is particularly acute when working with inherited texts, as is the case for EMED's Featured Plays. These carefully edited texts are documentary editions representing a single early modern printed work: each text is derived from a transcription of a printed playbook, initially created by the *Early English Books Online Text Creation Partnership* (EEBO-TCP).[80] Yet these base texts (and their associated image sets and metadata) had already undergone a series of remediations that the average EEBO user may be unfamiliar with. In brief, in the 1930s, University Microfilms International (UMI) began working with libraries to photograph rare books to create microfilm copies; these began to be scanned and digitized initially for CD-ROM in the 1990s; and these digitized images then moved online to become the widely used EEBO subscription site, owned by ProQuest-Healy. Founded in 1999, the non-profit Text Creation Partnership (EEBO-TCP), a collaboration between over 150 libraries, has been manually transcribing many thousands of these imaged documents through an outsourced double-keying enterprise, and on 1 January 2015 EEBO-TCP released 25,000 of these texts into the public domain, making them freely available to use and re-purpose.[81]

---

[78] *Early English Books Online*. ProQuest-Healy, see https://eebo.chadwyck.com/home, accessed 20/03/2019; Martin Wiggins and Catherine Richardson, *British Drama, 1533–1642: A Catalogue* (Oxford and New York: Oxford University Press, 2012).

[79] For more detail, see Meaghan Brown, 'Where Is That Book? Tracing Copies Imaged for EEBO', *The Collation* (blog), 17 March 2016, see https://collation.folger.edu/2016/03/where-is-that-book/, accessed 20/03/2019.

[80] A documentary edition is a version of the text that seeks to reproduce the individual witness that is its source, including its particular features.

[81] For further discussion of EEBO's particular history, see Ian Gadd, 'The Use and Misuse of *Early English Books Online*', *Literature Compass* 6:3 (May 2009): 680–92; https://doi.org/10.1111/j.1741-

These EEBO-TCP transcriptions underwent yet another stage of processing before they reached EMED. The *Shakespeare His Contemporaries* (SHC) project took from the TCP release a subset of c. 500 early dramatic texts, adapted the encoding, and used student contributors to fill gaps.[82] They also automated the addition of part-of-speech tagging and regularized spellings using *MorphAdorner*, an NLP toolkit developed by Philip Burns.[83] SHC is more focused on linguistic analysis and bulk correction and so has a different set of priorities to EMED: determining differences in priorities and how they manifest in encoding policies is an essential part of working with inherited texts. EMED edited and proofed these texts to ensure accuracy, which was necessary for both the original copy-specific transcription from EEBO-TCP and the regularized spelling variant provided by MorphAdorner. EMED also added further information from the playbook and adapted and standardized the encoding, in part bringing these texts into alignment with EMED's sister project at the Folger, *Folger Digital Texts*.[84] At heart, EMED aimed to increase the trustworthiness and scholarly authority of the base texts provided by the valuable EEBO-TCP transcription and encoding initiative.

The resulting EMED documentary editions are potentially difficult texts for students to grasp, since they have both a complex print and digital history and a messiness inherent to early printed playbooks. The editors wanted to make this complexity part of the narrative of these texts. Issues of provenance can be all too easily ignored or elided in digital editing projects but were central to EMED's philosophical approach. Not least, it was important to make clear that these are not modern critical editions: one proposal made by the Resources section is that students can build on these early texts to create their own critical edition. In creating a dedicated page on editing, the editors attempted to situate the digital provenance of the text in its existing textual and editorial history, making an argument that this is worthy of scholarly attention as well as providing an introduction to editorial theory. Additionally, the Encoding Path on each play entry page is a bespoke account of this remediation for every play. This articulates the precise pathway that the text went through to go from the library shelf to the EEBO-TCP transcription, to the SHC encoded version, to the EMED documentary edition. Where possible, the editors identified the shelf-mark for the transcribed copy itself, and link directly to the library catalogue record: determining these helped catch the mistakes in the repository attribution mentioned earlier.

4113.2009.00632.x and Bonnie Mak, 'Archaeology of a Digitization', *Journal of the Association for Information Science and Technology* 65:8 (2014): 1515–26, see https://doi.org/10.1102/asi.23061.

82 The SHC corpus was available for download from *Github* and *Google Drive* during EMED's lifetime. For an account of the project, see Martin Muller, 'Shakespeare His Contemporaries', *Scalable Reading* (blog), 22 June 2013, see https://scalablereading.northwestern.edu/?p=262, accessed 20/03/2019.

83 Philip R. Burns, *MorphAdorner v2: A Java Library for the Morphological Adornment of English Language Texts* (Evanston, IL.: Northwestern University), see https://morphadorner.northwestern.edu/morphadorner/download/morphadorner.pdf, accessed 20/03/2019.

84 Barbara Mowat, Paul Werstine, Michael Poston, and Rebecca Niles, 'Shakespeare's Plays', *Folger Digital Texts. Folger Shakespeare Library*, see https://www.folgerdigitaltexts.org/, accessed 20/03/2019.

It is too easy to see the digital text as appearing out of thin air, divorced from its print and digital antecedents. Emphasis on remediation as a continuing process, and one in which the user can participate, is intended to disrupt that appearance. In EMED this is achieved through the Resources section, which provides content on editing practice, lesson plans, and modelled exercises (again, produced collaboratively and attributed as such, in this case by two workshops held during the active lifetime of the project, for college faculty and undergraduate students respectively). To ignore complex digital histories is to remove the multiplicity of agents who construct the text, and to obscure the encoding as well as editorial choices made at every stage of its creation. To direct attention towards these is to put them back.

The aim of this section has been to discuss possible methods for editing on a scale appropriate to the ambitious task of reassembling the republic of letters. It necessarily looks past the model of the lone scholar working for years on print editions and towards the possibilities presented by crowdsourcing, Handwritten Character Recognition, tools for digital editing, and collaborative, iterative working on texts opened up to use and reuse. This section has explored some of the consequences of these changes for how we think about texts and digital provenance, using the example of EMED as a project that edited texts that had been processed by several different groups over many decades. A project as ambitious as reassembling the republic of letters can only work through building on existing editions, texts, and projects in a way that champions collaboration and responsible scholarly reuse. To achieve this, we must acknowledge the multiple agents embedded within our primary materials, digital and otherwise. This must be done not only to credit labour – particularly where such labour, often in the archives and libraries sector and in technical development, has too often gone uncredited – but also in order to recognize the myriad individual decisions that materially affect the text as received, whether editorial, encoding-related, design-related, or technological.[85] The digital edition does not merely provide a new environment into which the print edition is shifted; each edition is a wholly new object of scholarly engagement, contiguous with previous instantiations yet with its own complex narratives around agency and materiality.[86]

[85] For further discussion of problems and biases in the recognition of the scholarly work of archivists see Michelle Caswell, '"The Archive" Is Not an Archive: Acknowledging the Intellectual Contributions of Archival Studies', *Reconstruction* 16:1 (2016), see https://escholarship.org/uc/item/7bn4v1fk, accessed 20/03/2019.

[86] For the digital edition as a different entity to the digitized print edition, see Elena Spadini's comments in this chapter.

III.4 Modelling Texts and Topics

Charles van den Heuvel

1 Introduction: The Need for Topic Modelling

The scientific revolution of the seventeenth century was driven by countless discoveries in the observatory, in the laboratory, at sea, in society at large, and in the library. The amount of information in circulation increased dramatically, giving rise to new knowledge, theories, and world views. The other intellectual movements of the early modern period – the Renaissance, the Reformation, and the Enlightenment amongst them – also increased intellectual stimuli and debates throughout Europe. The 'republic of letters' was one of the names which contemporaries gave to the multiple overlapping networks of communication through which these movements were transmitted; and scholarly correspondence was, in turn, one of the most immediate and characteristic of these modes of communication.

Yet the learned letter is in many respects a very difficult as well as a very rewarding genre for the intellectual historian; and one of the sources of this difficulty is its topical structure – or rather, its lack of topical structure. A formal treatise moves through a series of topics in a carefully predetermined logical order. By the seventeenth century, that order was typically captured, in printed treatises at least, in an introductory table of contents. In addition, the location of passages on a huge range of individual topics was also often pinpointed in a printed index at the back of the volume, directing the reader to the precise pages in which relevant material could be found. These two navigational aids enormously assist the scholar in locating within the treatise material relevant to any particular topic; but neither of these

aids is available in an archive, or indeed an early modern printed collection, of learned letters.

The immediacy of the familiar letter derives in part from the fact that it can leap unexpectedly from one topic to a completely different one, sometimes in mid-paragraph; so the 'table of contents' of an edition of correspondence, even a meticulous modern one, normally identifies the sequence of letters by sender, recipient, and date rather than by topic discussed. Teasing out the topics of the letters in an index is a demanding job which is not undertaken even in many modern editions of correspondence, to say nothing of early modern epistolaries or archival collections. This lack of finding aids helps explain why the hundreds of thousands of letters printed in the early modern period are such an underutilized resource. Buried in those letters are vivid passages on every subject imaginable; but how is the scholar to locate, within this uncharted ocean of miscellaneous material, the texts relevant to any particular topic?

For this as for so many otherwise intractable scholarly difficulties, cutting-edge computing technology appears to offer unprecedented solutions. In this case, the most attractive prospect is to use increasingly sophisticated software for 'topic modelling' to help locate passages on a particular topic within a vast and otherwise undifferentiated text corpus.[1] Topic modelling is a method that assists users in identifying topics based on statistical methods that calculate which words occur more frequently in proximity to other words. The computer generates such strings of neighbouring words and the assumption is that those words together probably have a relationship and express a topic. The user can then label this string of assumed related words as a 'topic'. This technology is not only useful for indexing large quantities of letters, but also assists in identifying recurrent themes and scholarly debates in correspondences and in visualizing them in time and space or in networks.

As an introduction to these topics, the central three sections of the chapter discuss (section 2) the most ambitious project to date in the application of topic modelling to early modern learned correspondence; (section 3) the experiments conducted to test this application; and (section 4) the case studies designed to help teach the use of it. Building on the experience of that project, the final two sections sketch two lines of future development: (section 5) the proposal to embed the *ePistolarium* within a network of other tools and resources, and (section 6) the prospects for refining topic modelling with artificial intelligence and human–computer interaction.

[1] This task is different from but potentially complementary to the one pursued in chapter II.5: that chapter proposed means of recapturing the multiple, competing hierarchies of 'topics', 'places', or *loci communes* in which early modern intellectuals organized their material. This chapter proposes means of locating passages on these and other individual and collective topics within vast bodies of text in individual letters and correspondences which lack a clear internal topical structure of their own.

2 Topic Modelling the Dutch Republic of Letters: The *ePistolarium*

Within the 'information society' *avant la lettre* which flourished in the seventeenth century, a disproportionate role was played by the Dutch Republic. At the peak of their economic prosperity and cultural and intellectual fertility, the United Provinces offered a relatively tolerant safe haven for refugee intellectuals from across Europe, some of Europe's most fertile universities and printing houses, and a global trade network with which to communicate throughout the entire world. Between 2008 and 2013, an attempt to harness advanced computing technology to explore this subject was pursued in the project *The Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic* (CKCC) at the Huygens Institute for the History of the Netherlands in The Hague (subsequently moved to Amsterdam).[2]

From the outset, CKCC pursued an agenda distinct from but complementary to other main initiatives in the field. In Oxford, for instance, *Cultures of Knowledge* focused on creating the culture, community, and technology needed to populate a union catalogue of the republic of letters collaboratively.[3] In Stanford, *Mapping the Republic of Letters* concentrated on creating elegant and user-friendly tools for mapping and visualizing such metadata.[4] Meanwhile, in The Hague, CKCC focused from the outset on metadata coupled with the full texts of the letters themselves. The objective was to harness modern information technology to understand not only the networks through which knowledge was exchanged, but the texts that exchanged that knowledge, and the knowledge they exchanged. The questions pursued were rather generic. How, for instance, did knowledge circulate in the seventeenth-century Dutch Republic? How were elements of knowledge – generated at sea, on the battlefield, in overseas trading outposts, as well as in libraries and laboratories – accessed, debated, and deployed by the learned community? How was this new knowledge processed, disseminated, theorized, and ultimately accepted or, for that matter, rejected? A second, deeper set of questions regarded the new methods necessary for answering the first set of questions in new ways. How, most basically, can the tens of thousands of learned letters edited by previous generations of scholars be assembled and analysed to reveal patterns of knowledge production, circulation, and appropriation in all their chronological, geographical, institutional, and thematic complexity? Still other questions related to how digital technology might also allow the fresh insights gained to be made more accessible to interdisciplinary research in the humanities and potentially also to

[2] The *Circulation of Knowledge Consortium* (CKCC) consisted of the institutes of the Royal Netherlands Academy of Arts and Sciences (Huygens Institute and DANS), Dutch universities (Utrecht University and University of Amsterdam) and the National Library of the Netherlands (KB).

[3] The first CofK website is at https://cofk.history.ox.ac.uk/. The second is at http://www.culturesofknowledge.org/, accessed 20/03/2019.

[4] See http://republicofletters.stanford.edu/. This work has been continued across a broader front by the cognate project at Stanford, *Humanities + Design*, see http://hdlab.stanford.edu/, both accessed 20/03/2019.

non-specialist audiences. To address these and related questions, the central objective of the project was to pilot a new kind of e-infrastructure for early modern correspondences known as the *ePistolarium*.

The first stage in developing the *ePistolarium* was to create a testbed for experimenting with these methods consisting of letters to and from leading intellectuals representative of the Dutch golden age, namely: Caspar Barlaeus, Isaac Beekman, René Descartes, Hugo Grotius, Constantijn and Christiaan Huygens, Antoni van Leeuwenhoek, and Jan Swammerdam. This work created a data set of over 20,000 seventeenth-century learned letters in Latin, French, Dutch, and English: an excellent basis for subsequent experiments. Not only were the correspondences of these scholars multilingual: sometimes even single letters were composed in more than one language. The multilinguality of the correspondences of the republic of letters has a huge impact on the method of topic modelling, which is based on counting words that might belong together statistically. To address this challenge, the topic modelling software needed to be enhanced with Natural Language Processing (NLP) techniques based on interactions between computers and human (natural) languages. These techniques are used to program computers in such a way that they can process and analyse large amounts of natural language data. For the development of the *ePistolarium*, a combination of language identification, spelling normalization, named entity recognition, and keyword analysis was used.[5]

Language identification: Topic modelling is based on the semantic similarity between words and between the texts of the documents. The first, necessary step is therefore to identify the language used in any given passage. Language identification detects in which language texts are written so that for each language separately the occurrence of words that potentially express a topic can be counted.

Spelling normalization: In order to establish the word occurrence, the computer needs to distinguish automatically between similar and dissimilar words. However, since topic modelling is based on detecting and counting words with a probabilistic meaningful relationship, it is crucial to know whether we are dealing with similar words that are spelled differently or with words with a different meaning. This difficulty is compounded for the early modern period, when most European vernaculars had not yet standardized modern spelling. The next step, therefore was to normalize early modern to modern spellings.

Named entity recognition: In order to find predefined categories in text, such as persons or geolocations, the computer can automatically classify information in so called 'named entities'. The example of personal names and place names is of interest because many personal names derive from place names. However, based on grammatical rules (syntax), the computer can distinguish between personal names derived from place names and the names of the places themselves.

[5] For an explanation of the various NLP techniques used see http://ckcc.huygens.knaw.nl/?page_id=13, accessed 20/03/2019. For language identification, the N-grame based cumulative frequency addition algorithm was used; for spelling variation VARD2; for NER, the Aho-Corasicj algorithm; and for keyword analysis WMatrix.

Keyword analysis: Not only can the computer be programmed to identify languages, to normalize spelling, or to recognize automatically and distinguish between classed names; it can also be programmed to analyse keywords in such a way that it can recommend alternative suggestions for search.

As a result of implementing topic modelling with this combination of NLP techniques, the database of the *Circulation of Knowledge* project can be queried by typing in keywords, by selecting given search options (facets) in combination (faceted search), and by similarity search. The latter search method, based on topic modelling, presents letters in the database similar to the letter found after a query. It is also possible to upload a text to find a similar letter in the database. The results of all keyword and faceted search queries can be visualized on geographical maps, in timelines, in correspondence networks, and in co-citation networks to map the people that are mentioned in relation to specific topics and debates (fig. 1).

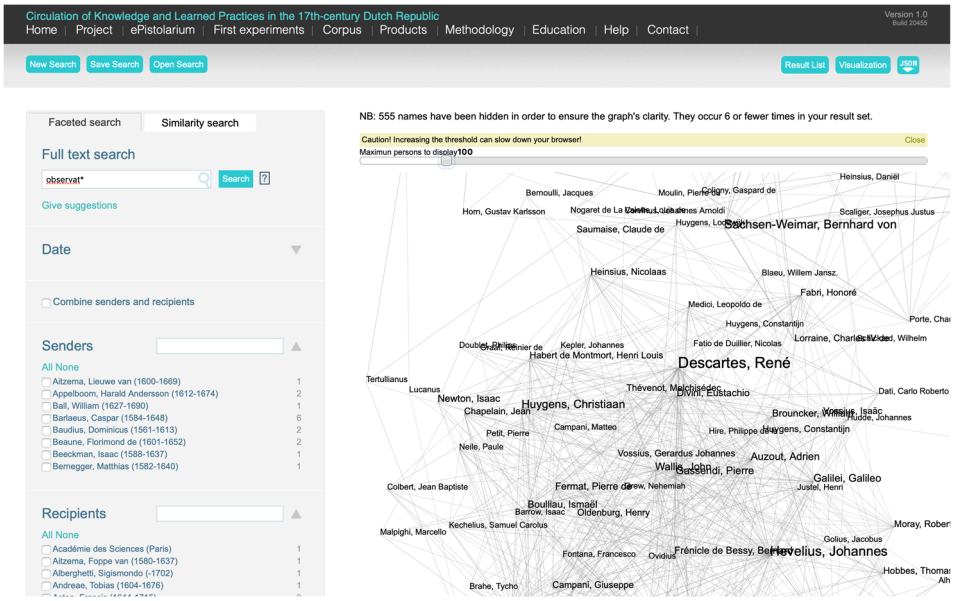


Figure 1: *ePistolarium*: visualization of co-citation network of 100 people of the keyword observat*

During and after the project, that ran from 2008 until 2013, several experiments with the *ePistolarium* tool were set up. First, the open source topic modelling tool *MALLET*, developed at Stanford, was used. Tests by historians of science during the international workshop 'Mathematical Life in the Dutch Republic' held in 2010 in the Lorentz Center in Leiden (Netherlands) revealed a limited success of the *MALLET* tool. Moreover, the historians criticized the lack of control over the

extraction of themes of interest in the database of letters that in their view functioned too much as a 'black box'.[6] For that reason it was decided to compare multiple topic modelling methods. On the basis of this comparative analysis, random indexing was implemented for various reasons.[7] The main expected advantage of this method was that it would scale very well with increasing corpus size. The latter is crucial given the aim to include far more letters in the *ePistolarium* to optimize the functionality and use of the tool.

3 Experiments with the *ePistolarium*

In preparation for the launch of the tool in June 2013, five experiments were undertaken to test its efficacy. In experiments of this kind, failure or partial failure is as significant as success in demonstrating the potential utility of the tool and its limitations. Both the successful and the less successful experiments have been fully documented,[8] and are summarized here to explain both what has been achieved and the further work which is needed in the future.

1. *Searching for simple and concrete topics*. The most basic experiment was undertaken by Floor Haalboom, a student of the history of science at the time, who explored the utility of the *ePistolarium* in locating references to animals in general, to dogs in particular, and most specifically to a very Dutch dairy product: milk. The results of this experiment were to locate rich veins of material from which useful observations could be made. For instance, in the high social circles around Constantijn Huygens, dogs were mentioned as pets and as hunting dogs, while a cluster around Antonie van Leeuwenhoek and Jan Swammerdam referred to dogs primarily as the subjects of scientific experiments. Milk was likewise discussed both as a remedy against illness and as an object of scientific study. More surprising were the figurative descriptions of milk in the correspondence of the mathematician and astronomer Van Nierop, who used milk as a metaphor to describe the chaos in the universe at the moment of its creation. These basic searches confirmed the value of the *ePistolarium* in very readily identifying material relevant to familiar and concrete objects with very well-established vocabularies in all the relevant languages.

2. *Searching for multiple, conceptual topics*. In a second experiment, Dirk van Miert sought more elusive, conceptual terms in order to test a well-defined thesis: namely, whether the 'two cultures hypothesis' of C. P. Snow (1959) could be applied to

[6] At the NIAS-Lorentz workshop *Mathematical Life in the Dutch Republic* organized in Leiden in 2010, c. twenty historians of science participated in the experiment with an early version of the *ePistolarium* tool.

[7] Peter Wittek and Walter Ravenek, 'Supporting the Exploration of a Corpus of 17th Century Scholarly Correspondences by Topic Modeling', in Bente Maegaard, ed., *Proceedings of Supporting Digital Humanities: Answering the Unaskable* (Copenhagen, 2011): http://hb.diva-portal.org/smash/get/diva2:887360/FULLTEXT01, accessed 20/03/2019.

[8] CKCC tab: 'first experiments', see http://ckcc.huygens.knaw.nl/?page_id=1030, accessed 20/03/2019.

the seventeenth century. Snow had argued that, since the Victorian era, intellectual life in the West had gradually split into two cultures: the sciences and humanities. To determine whether this split was visible in the seventeenth century, Van Miert used the co-citation tool that counts the number of people that are mentioned in relation to specific terms. In order to test this hypothesis, Van Miert divided the major figures on the *ePistolarium* into two categories, natural scientists and historico-philologists, and investigated their use of two clusters of terms and their cognates: *observatio* and *historia*.[9] This search resulted in 1,010 hits over the period 1600–1700, with a peak in the period 1660–75. Using the network and co-citation visualization tools of the *ePistolarium*, Van Miert identified two networks discussing the term 'observation': a rather small network around the humanist scholar, diplomat, and jurist Hugo Grotius (whose correspondence is the largest in the *ePistolarium*), and much more frequent use of this term in the network of the mathematician Christiaan Huygens. The names in the latter network most frequently cited in relation to 'observation' are leading natural philosophers and mathematicians of the period, including René Descartes, Johannes Hevelius, Robert Hooke, Robert Boyle, Pierre Gassendi, and Christopher Wren. The search term *historia* yielded the opposite result. The term occurred predominantly at the beginning of the seventeenth century in the correspondence of Grotius and was associated with books and religious matters. When this term was combined with the term 'observation', the co-citation networks revealed a further concentration amongst historico-philological writers, including Claude Saumaise, Gerardus Joannes Vossius, Daniel Heinsius, and Johannes Uytenbogaert. These keyword searches thereby confirmed the traditional historiographical association of the term *observatio* with the natural sciences, and *historia* with more philological and historical work. Significantly, these results also contradicted the hypothesis formulated in Van Miert's recently published work, *Communicating Observations in Early Modern Letters (1500–1675)*,[10] where he argued that the term 'observation' could equally be recognized in letters of an anthropological, historical philological nature and in medical or astronomic correspondences. The crucial question was whether these results revealed a flaw in Van Miert's thesis or a limitation of the data set assembled in the *ePistolarium*. Numerically, the overwhelming majority of the letters in the *ePistolarium* derive from printed editions of the correspondences of well-known scholars; and 90 per cent of them (18,233 of the 20,215 letters) originate in just three sprawling correspondences: those of Hugo Grotius (8,034 letters), Constantijn, and Christiaan Huygens (7,119 and 3,080 letters respectively). These proportions may render this data set

[9] More specifically, the terms *observat** and *observat** were used together with the connotations: *dant*, *dantur*, *dari*, *datis*, *datur*, obseruation, observation, *observatione*, *observationem*, *posita*, *positam*, *positi*, *posito*, *prae*, *prius*, *saltem*, and *semper*.

[10] Dirk van Miert, ed., *Communicating Observations in Early Modern Letters (1500–1675): Epistolography and Epistemology in the Age of the Scientific Revolution* (London: The Warburg Institute; Turin: Nino Aragno Editore, 2013).

unrepresentative of the Dutch republic of letters as a whole. In order to test these conclusions, an even larger data set is needed.

3. *Searching for references to images.* A third experiment, conducted by the historians of science Eric Jorink and Joas van der Schoor, revealed a different set of limitations in the underlying source-base of the *ePistolarium*. Their purpose was to explore the epistemological importance of images to the empirical, scientific cultures of the period. For instance, the autobiographical account of the youth of Constantijn Huygens describes how he was trained by the painter Joris Hoefnagel, and his letters reveal that he surrounded himself with many of the well-known painters, engravers, and architects so prominent in Dutch culture in this period. Could the *ePistolarium* help consolidate our conception of the close links between 'art' and 'science' during the Dutch golden age? In pursuit of answers to this question, Jorink and Van der Schoor tried to map references to images throughout the corpus by searching for polyglot terms for 'image' (such as *afbeelding*, *imago*, *figura*, *figure*, and *vide*) and charting the clusters of authors most inclined to refer to them. The results of this experiment revealed further strengths and limitations of the *ePistolarium* and its underlying data. The strength of the *ePistolarium* tool as an heuristic instrument was revealed in its capacity to locate rich veins of material in a few seconds, which might otherwise be found only by leafing through thousands of pages of texts and indexes. For instance, the term *figura* and its cognates occurred frequently in the network around Descartes. Likewise, *pictura* and its cognates featured in the letters of Grotius, but were restricted mostly to the project of publishing the work of Franciscus Junius, *De pictura veterum* (1597): that is, these references were to a book of images, rather than to the images themselves.

However, Jorink and van der Schoor's investigations also revealed another shortcoming in the data set underlying the *ePistolarium*. The various editors of the volumes that provided its content did not all fully appreciate the importance of images, and, even when they did, the print technology of the time did not allow them to reproduce those images accurately and inexpensively. As a result, some simply skipped the drawings or sketches that correspondents included in their letters (as in the Worp edition of Constantijn Huygens, for example); others substituted these valuable sources with mere schematic representations, or buried facsimiles in the back of volumes (as in the *Oeuvres complètes de Christiaan Huygens* or the Leeuwenhoek correspondence). When these diverse editorial practices are amalgamated in a resource such as the *ePistolarium*, the data set produced is not only incomplete and unreliable in this regard but also highly inconsistent. The limitations encountered by Van Miert, based on the small, selective body of material underlying the *ePistolarium*, is thereby compounded by a second bias resulting from the variable editorial practices of the individual editions themselves.

4. *Co-citation analysis of intellectual networks.* The historian of science Huib Zuidervaart, an expert on early modern optical instruments such as microscopes and telescopes, tested the co-citation feature of the *ePistolarium*. Co-citation occurs when two or more people are mentioned close together in the same text, for in-

stance in one paragraph in relation to a topic. It offers a very useful indicator for establishing the structure of scientific debates. The *ePistolarium* automatically generates co-citations of each paragraph in the letters of its correspondences. To test that functionality, Zuidervaart analysed references to key figures in the correspondence of Christian Huygens involved in the discovery of the ring structure and moons around the planet Saturn. This offered a good test case not only because Christiaan Huygens played a crucial role in debates on this subject, but also because this subject has been studied in detail in eleven different scholarly articles by the expert Van Helden between 1968 and 1996. The automatically generated co-citations could therefore be benchmarked against the results that Van Helden had assembled manually over the years. In other words, this experiment case was designed not only to find new data but also to test the capacity of the *ePistolarium* to reproduce sound results achieved by traditional scholarly methods.

The *ePistolarium* tool returned not only names already discussed in the publications on this topic by the expert Van Helden, but produced additional relevant historical figures. Zuidervaart looked into the scholars that co-cited the word 'Saturn' which occurred in a total of 395 letters in the *ePistolarium* in four periods: 1640–54, 1655–9, 1660–70, and 1671–85. For each period he set the threshold at the ten most mentioned scholars. For the first period the results were disastrous. None of the people that emerged were stakeholders in debates around Saturn. The names that came up could only be connected with *Die Saturni*, that is with Saturday. Fortunately, the analyses with the *ePistolarium* of the successive periods yielded convincing results. For the period 1655–9, eight of the ten people mentioned in the 155 letters were already known from the literature of Van Helden, but to these the *ePistolarium* added two French mathematicians, Fermat and Pascal. A similar result was obtained for the period 1660–70. Once again, eight of the ten scholars cited in a comparable set of 158 letters mentioning Saturn were known, but the *ePistolarium* added Thevenot and Vossius to the discussion. The results for the last period, 1671–85, were even more surprising. In the forty-eight letters with the word 'Saturn', the six figures mentioned most frequently were the usual suspects, but the following six were not earlier mentioned in the selection of articles by Van Helden. Despite the cautionary tale relating to *Die Saturni*, this experiment was successful in both confirming and extending the results of previous research.

5. *Searching for implicit topics*. A still more ambitious experiment conducted by Charles van den Heuvel and Henk Nellen, however, pushed the capacities of the *ePistolarium* to breaking point.[11] Their research question focused on the theme of confidentiality, which recurs regularly in the letters of the Dutch scholar Hugo Grotius, and which is considered an important characteristic of knowledge exchange in the historiography of the republic of letters. Their research question was whether the *ePistolarium* could automatically retrieve material on such a subtle topic

[11] See http://ckcc.huygens.knaw.nl/epistolarium, accessed 20/03/2019, tab: 'first experiments' casus: 'The nature of scholarly communication'.

from other correspondences in the CKCC project.[12] (Some unknown references to confidentiality in the work of Grotius were found, but very few in other correspondences.) Nevertheless, manual checks revealed that confidential information also appeared in the letters of other scholars. In short, this attempt at automatic detection of letters expressing confidentiality was basically unsuccessful.

This result was not surprising. In small data sets, the 'most similar' letter can still have quite different content. A query for words (confidentiality-related or not) based on similarity search in a set of 20,000 letters of which more than a third were written by or to Grotius, almost invariably leads to his correspondence. But apart from the size and composition of the data set, intrinsic features of language, such as implicit language use, can also explain the low recall. While in the Saturn case, words like 'sun', 'moons', 'planets', and 'stars' immediately provide explicit associations with astronomy (or perhaps astrology), the many ways to ask recipients of letters handling specific controversial information in confidence are far more implicit.[13] The different word combinations used by humanistic rhetoricians to express confidentiality are highly variable, as can be suggested by comparing the request 'Interea quas a me tenes litteras tibi habe et Vulcano sacrifica' ('Meanwhile, keep the letters received from me for yourself and offer them to the god of fire') with 'Haec inter nos dicta sunto' ('Consider this information as exchanged between the two of us'). For this reason, it is difficult to determine which strings of words automatically extracted by the computer are representing the same concept or topic. This has implications for the way computer-generated word-strings are tagged as topics. Different levels of abstraction, which often are implicit, have an impact on the quality of the representations of the topics. They affect for instance the quality of visualizations of correspondence network and co-citation networks that respectively represent which authors discuss or are named in relation to specific topics. In short, the very different levels of abstraction encountered in the cases on animals, astronomy, and implicit references to confidentiality, condition the use of text-analytical methods such as topic modelling.

[12] See Henk Nellen, '"In strict confidence": Grotius' Correspondence with His Socinian Friends,' in Toon van Houdt et al., eds., *Self-presentation and Social Identification. The Rhetoric and Pragmatics of Letter Writing in Early Modern Times* (Leuven: University Press, 2002), 227–45; Henk Nellen, 'The Correspondence of Hugo Grotius,' in Christiane Berkvens-Stevelinck, ed., *Les Grands Intermédiaires culturels de la république des lettres. Études des réseaux de correspondances du XVIe au XVIIIe siècles* (Paris: Honoré Champion, 2005), 127–64; and Charles van den Heuvel et al., 'Circles of Confidence in Correspondence. Modeling Confidentiality and Secrecy in Knowledge Exchange Networks of Letters and Drawings in the Early Modern Period', *Nuncius* 31 (2016): 78–106, see https://doi.org/10.1163/18253911-03101002.

[13] Van den Heuvel et al., 'Modeling Confidentiality'.

4 The *ePistolarium* in the Classroom

From the various experiments with the *ePistolarium* it became clear that selection of corpora by historians, selection by editors, and intrinsic features of language all resulted in limitations, sometimes even biases, in the use of topic modelling. To overcome these various flaws, multiple strategies are necessary. Users need to be aware that the interpretation of its generated results should be handled with great care. For that reason it was a great opportunity that two years after the official end of the project in 2013, CKCC was one of the projects selected by CLARIN-NL for the development of an educational module. The primary focus of ongoing work consequently shifted from technical support towards the training of new users of the *ePistolarium*. This CLARIN-Education module aimed at acquainting students with visual analytical methods using the *ePistolarium* tool. Acting on commission of the Huygens Institute, Daan Wegener, an historian and philosopher of science and a teacher of natural sciences, developed three overarching case studies in Dutch and in English. Each case was broken down into sub-cases, assignments, and research questions designed to test the tool and to enable comparisons with similar projects analyzing the republic of letters. These cases can be downloaded from their own tab on the CKCC website.[14]

The first case concerns the reception of Aristotle in the republic of letters. This topic is particularly important with respect to Descartes's presence in the Dutch Republic, and the rapid growth of interest in and outside the Dutch universities in the mechanical philosophy and mathematical descriptions of nature, which implied a rejection of Aristotelian views. The assignments focus on mapping the role of Aristotle in various correspondence networks.

Another case, 'No important news', focuses on the role of news in the republic of letters. In this assignment, students are taught how to refine their queries in order to answer a research question regarding the news. It also addresses the issue of reciprocity in intellectual networks of the republic of letters, a topic of importance also in the earlier case study of confidentiality in the correspondence network of Hugo Grotius.

Aristotle returns in a third case study dealing with the debate between René Descartes and his most important opponent in the Dutch Republic, the Utrecht theologian Gisbertus Voetius. This set of assignments acquaints the user with two facilities of the *ePistolarium* tool designed to help track the circulation of knowledge: namely, the capacity to graph the people in co-citation networks and to visualize the geographical distribution of ideas. At the same time, the user can make comparisons with similar projects such as *Mapping the Republic of Letters* in Stanford and *Cultures of Knowledge* in Oxford.

In a related development, the *ePistolarium* was also used in combination with other tools in another training experiment established by the Brazilian historian of

[14] See http://ckcc.huygens.knaw.nl/?page_id=1456, accessed 20/03/2019.

science Marlon Cesar Alcantara. Alcantara used the networks of Constantijn Huygens to develop a proposal for teaching the history of optical instruments. Students were given the task of creating ego-networks from scholars and artists in the network of Huygens by making connections to relevant documents *outside* the *ePistolarium*. Ego-networks of this kind were created for Francis Bacon, Cornelis Drebbel, René Descartes, Christiaan Huygens, Baruch Spinoza, Rembrandt van Rijn, and Johannes Vermeer. The objective was to develop an interdisciplinary curriculum for teaching students how the advent of optical instruments can be explained by combining networks of scientists, artists, and philosophers with sources about religion and trade. Such experiments with multiple distributed data sets are important to prepare researchers for the use of infrastructures that are needed to handle the big data of the republic of letters.

5 Next Steps: From the *ePistolarium* to a Virtual Research Environment

Although training modules of this kind help introduce users to the possibilities and limitations of the *ePistolarium*, they do not address the deeper problems encountered in using topic modelling for the analysis and visualization of the knowledge circulation and appropriation in the republic of letters.

In order to address these problems, first of all, more data is needed. Since the grant period of the CKCC project ended in 2013, only one substantial collection has been added to the *ePistolarium*: the c. 1600 letters to and from the controversial Huguenot philosopher and polemicist in Rotterdam, Pierre Bayle (discussed further below). In order to overcome the imbalance between very large and very small correspondences, a great many more full-text letters will be needed. For this reason the collaborative efforts, outlined in this volume aiming at 'Reassembling the Republic of Letters' will be crucial to further progress in this field.

In addition to populating the *ePistolarium* with a greater volume of letter texts and improving the technologies for topic modelling per se, several other strategies are under consideration for developing the platform further.

One of these strategies is implicit in one of the subtitles of the *Circulation of Knowledge* project at an early stage: namely, 'A Web-based Collaboratory around Correspondences'. The 'collaboratory' referred to was conceived as a digital environment in which a group of researchers could individually and collectively annotate the letters in the corpus. After the first user tests with the *ePistolarium* described above, it was decided to leave aside the creation of a collaboratory for the moment and to concentrate on the improvement of topic modelling in relation to the search facilities. The justification for this change of plan was simple: before researchers can enrich the texts that take their interest in the *ePistolarium*, they need to be able to find them. If the topic modelling capabilities currently available do not adequately support their queries, researchers will simply abandon the *ePistolarium*

before investing their precious time in enriching its contents. At a later stage of development, returning to these original plans will offer an obvious option for enhancing the environment within which the *ePistolarium* sits.

Ultimately, however, users of a resource such as this will need to be able to work not only with the data and metadata on a stand-alone tool such as *ePistolarium*; they will also need to be able to situate that tool within a broader network of tools and resources devoted to the selection, cleaning, reconciliation, analysis, annotation, and further contextualization of metadata on a large scale, to the automatic transcription crowdsourcing, editing, and annotation of texts, and to deep learning and other experiments on the resulting data, to name but a few. Moreover, as collaborative work on reassembling the republic of letters progresses, these tools and resources will increasingly not be located in a single place but distributed throughout Europe and beyond. In preparation for this kind of contextualization, the *ePistolarium* is being extended with other functionalities and converted into other digital formats to be aligned with the overall infrastructure of the Huygens Institute and the Humanities Cluster of the Royal Netherlands Academy of Arts and Sciences (KNAW), which in turn mirrors the design of the Dutch national Common Lab Research Infrastructure for the Arts and the Humanities, CLARIAH.

A good example of the kind of collaborative exchange that such an infrastructure will facilitate is provided by the correspondence of Pierre Bayle. Between 1999 and 2017, an international team collaborated on a critical edition of Bayle's letters published in hard copy by the Voltaire Foundation in Oxford.[15] More recently, an overlapping team has created an electronic edition of the letters, together with images of the manuscripts, at the Université Jean Monnet Saint-Étienne.[16] No sooner was the edition complete in 2017 than the Bayle correspondence project released its basic metadata for integration into *Early Modern Letters Online* (EMLO), where a catalogue of Bayle's correspondence is embedded within a growing set of related resources, with individual catalogue records linking out to the texts in the electronic edition.[17] The *Cultures of Knowledge* project (responsible for EMLO) then passed its refined metadata to the CKCC project, which ingested it together with the full texts provided by the electronic edition of Bayle's letters, thereby adding a major new data set to the *ePistolarium*. After applying the Natural Language Pro-

[15] *Correspondance de Pierre Bayle*, critical edition established under the direction of Elisabeth Labrousse and Antony McKenna, with the collaboration of Wiep van Bunge, Hubert Bost, Edward James, Annie Leroux, Fabienne Vial-Bonacci, Bruno Roche, and Eric-Olivier Lochard, 15 vols. (Oxford: The Voltaire Foundation, 1999–2017).

[16] *Correspondance de Pierre Bayle*: *Édition électronique*: under the academic direction of Antony McKenna and Fabienne Vial-Bonacci (Institut Claude Longeon, IHPC, CNRS UMR 5037), realized by Pierre Mounier (DSI, Université Jean Monnet Saint-Étienne) on the basis of data of the *Arcane* database created by Eric-Olivier Lochard (Université de Montpellier 3): http://bayle-correspondance.univ-st-etienne.fr, accessed 20/03/2019.

[17] Antony McKenna and L'Édition Électronique de la Correspondance de Pierre Bayle, eds., 'The Correspondence of Pierre Bayle', in *Early Modern Letters Online*, Cultures of Knowledge, http://emlo-portal.bodleian.ox.ac.uk/collections/?catalogue=pierre-bayle, accessed 20/03/2019.

cessing techniques described above, the texts of Bayle's letters can now be analysed and visualized together with all the other letters in the *ePistolarium*.

Further experiments to make the EMLO database interact with the linked data of the CLARIAH Knowledge Graph might lead to further collaboration and data integration between these two and other digital republic of letters projects.[18] Depending on available funding, 'A Web- based Collaboratory around Correspondences' would finally be achieved and complete the *ePistolarium* as a key site for further collaboration in the digitally reassembled republic of letters, involving scholars both individually or collectively, within academic or cultural heritage projects or institutions.

6 Future Prospects: Topic Modelling and Human–Computer Interaction in the VRE of the Republic of Letters

As noted above, the creation and linking of large quantities of epistolary data is fundamental to overcoming the limitations of topic modelling. In order to generate textual material in sufficient quantity, semi-automated means of assembling and reconciling metadata (chs. III.1–2) are needed, along with crowdsourced and automated transcription and collaboratively produced editions (ch. III.3). Conversely, if these methods succeed in generating unprecedented quantities of letter texts, the improvement of techniques for topic modelling these huge bodies of text will grow more urgent.

The difficulty of handling higher levels of abstraction and implicit language, discussed above, will not, however, be resolved merely by increasing the amount of textual material available for modelling. To resolve these problems, solutions are most likely to develop from ongoing experiments with deep learning and the creation of so-called 'deep belief networks' in computer science, and especially in computational linguistics and artificial intelligence. Deep belief networks are created within the context of machine learning. Multiple layers of latent variables are trained via algorithms that attempt to model high-level abstraction in such a way that hidden units can be revealed (deep or hierarchical learning).[19] These methods are also used in the context of NLP and might in the future solve the problem discussed above regarding different levels of abstraction between more and less implicit concepts in topic modelling.

---

[18] The CLARIAH Knowledge graph 'Anansis' is powered by the *Timbuctoo* software of the Digital Infrastructure of the Humanities Cluster of the Royal Netherlands Academy of Arts and Sciences (initially developed at the Huygens ING). Both the *Cultures of Knowledge* project and the *ePistolarium* have an instance in *Timbuctoo* which allows for further linked data interaction between these projects.

[19] Geoffry E. Hinton, Simon Osidero, and The Yee-Whye, 'A Fast Learning Algorithm for Deep Belief Nets', *Neural Computation* 7 (18 July 2006): 1527–54, see https://doi.org/10.1162/neco.2006.18.7.1527.

A further problem relates to changes of topics over time.[20] Whereas in predictive analytics or machine learning 'concept drift' affects the statistical properties of the variables the model tries to predict, it also affects our interpretation of patterns in historical data. The historian of English literature Ted Underwood (2011) addressed this problem eloquently as follows:

Once you create a set of topics, plotting their frequencies is simple enough. But plotting the aggregate frequency of a group of words is not the same thing as 'discovering a trend,' unless the individual words in the group correlate with each other over time. And it's not self-evident that they will.[21]

'Context mining' might allow future systems simultaneously to detect hidden (implicit) concepts and concept drift. Particularly promising in this regard is an experiment pursued in collaboration between Princeton University and the Carnegie-Mellon University. As a supplement to their research project *Modeling the Evolution of Science*, David M. Blei and John D. Lafferty (2006) developed a dynamic model of seventy-five topics, and used it to develop a browser for exploring the evolving structure of hidden topics within the entire content of the journal *Science* over the period 1880–2002.[22] In this browser, a click on one of the top five words from each topic in a given decade leads to a page containing the top 100 words from that topic, links to their distribution in previous and future years, and the articles that exhibit that topic in the highest proportion. These hierarchical layers of words that step by step are contextualized, not only reveal the (partly hidden) content of articles by topics; the browser also shows how a topic has changed over time. Continuation of this kind of tool creation in the future will facilitate the creation and exploration of patterns in machine-readable, dynamic, deep networks of big data.

Despite these promising experiments, scholarly input will continue to be needed in the future for the interpretation and contextualization of these computer-generated patterns. As such, the future use of topic modelling in the digital repub-

[20] For this problem see for instance, Eyal Sagi, Stefan Kaufmann, and Brady Clark, 'Semantic Density Analysis Comparing Word Meaning across Time and Phonetic Space', in *Proceedings of the EACL 2009 Workshop on GEMS: Geometical Models of Natural Language Semantics, 31 March 2009, Athens, Greece* (Stroudsburg, PA: ACM, 2009), 104–11; Rada Mihalcea and Vivi Nastase, 'Word Epoch Disambiguation: Finding How Words Change Over Time', in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 08–14 July 2012, Jeju Island, Korea* (Stroudsburg, PA: ACM, 2009), vol. 2, 259–63; and Yoon Kim et al., 'Temporal Analysis of Language through Neural Language Models', in *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, 26 June 2014, Baltimore, Maryland, USA*, 61–5, see https://doi.org/10.3115/v1/W14-2517.

[21] Ted Underwood, 'Topics Tend to Be Trends. Really: $p < .05$!' *The Stone and the Shell. Using Large Digital Libraries to Advance Literary History* blog, 16 September 2011. See http://tedunderwood.wordpress.com/2011/09/16/topics-are-also-trends, accessed 20/03/2019.

[22] The links http://topics.cs.princeton.edu/Science/ and for the browser http://topics.cs.princeton.edu/Science/browser/ are broken, but references to the experiment are published in David M. Blei and John D. Lafferty, 'Dynamic Topic Models' in *Proceedings of the 23rd international conference on Machine learning, 25–29 June 2006, Pittsburgh, Pennsylvania, USA* (New York, NY: ACM, 2006), 113–20, see https://doi.org/10.1145/1143844.1143859.

lic of letters will continue to require a close human–computer interaction between intellectual historians and artificial intelligence.

III.5 Exchanging Metadata

Arno Bosse, Gertjan Filarski, Howard Hotson, Neil Jefferies, and Thomas Stäcker

1 Introduction

In the Spring of 2016, the University of Oxford,[1] together with technical partners from Germany, the Netherlands, Bulgaria, and the United States[2] and fourteen contributing partners from six further countries,[3] submitted a research infrastructure proposal to the European Commission's Horizon 2020 funding programme.[4] The scheme called for bids, 'to bring together, integrate on a European scale, and open up key national and regional research infrastructures to all European researchers, from both academia and industry, ensuring their optimal use and joint development'. The initial Phase I proposal submitted by the Oxford-led consortium, 'European Letters Online' (EULO), did not receive approval to proceed to the full application stage. Despite this, the discussions and preparatory work on EULO

[1] Faculty of History, University of Oxford; Oxford e-Research Centre, University of Oxford.

[2] Göttingen State and University Library (DE); Huygens ING, Royal Netherlands Academy of Arts and Sciences (NL); Ontotext (BG); Stanford University Digital Library (USA).

[3] Austrian National Library, Vienna (AT); The Czech Academy of Sciences (CZ); State and University Library, Göttingen (DE); Herzog August Bibliothek, Wolfenbüttel (DE); Staatsbibliothek zu Berlin (DE); National Library of Denmark and Copenhagen University Library (DK); Hungarian Academy of Sciences (HU); National Library of Israel, Jerusalem (IL); OCLC Research Europe, Leiden (IO); Consortium of European Research Libraries (IO); Biblioteca Nazionale Marciana, Venice (IT); National Library of the Netherlands, The Hague (NL); British Library, London (UK); Bodleian Library, University of Oxford (UK).

[4] RIA Research and Innovation Action, 'Integrating Activities for Starting Communities', INFRAIA-02-2017, see https://tinyurl.com/ybfmtfkh, accessed 20/03/2019.

coordinated by the COST Action's 'Data Exchange and Strategic Planning' working group,[5] were later drawn on by members of the Action for their own (in part also successful) funding applications.[6]

In essence, the EULO outline proposal consisted of a set of technical standards and workflows for exchanging digitized content on early modern correspondence throughout a distributed network of contributing nodes and integrating hubs. The design anticipated that scholars and editors logging onto a national or thematic service hub would be able to access data assembled from throughout Europe from a tool-rich interface allowing them to speed its standardization and enhancement, to curate it in a protected environment, release it for reuse under conditions suited to their needs, and obtain full acknowledgement for their contributions. In exchange, the contributing nodes would receive all the enhancements of that data provided by the global scholarly community without the need to discard their own records or reinvest heavily in new IT systems: institutions would have the option to install EULO as a free, stand-alone virtual appliance, run it in the cloud as a service, or employ it as a reference implementation to validate their own infrastructure with the programming interfaces and standards in use by the EULO consortium.

In addition, all stakeholders – repositories, publishers, scholars, but also non-expert users – would benefit from sophisticated new tools for navigating, analysing, surveying, and visualizing unprecedented quantities of integrated data, and be able to access related digital resources such as transcriptions, facsimiles, and translations. Our long-term vision for EULO was that, once in regular use in a distributed network, it would, in effect, create an *historical knowledge graph* for the field, capable of transforming our current understanding of the processes of transnational commercial, diplomatic, cultural, and intellectual exchange central to the European experience during the past five centuries and more.

Other contributions to this volume in section II ('Standards: Dimensions of Data') and section III ('Systems, Methods, and Tools') have already outlined many of the essential technical concepts and digital resources required to prepare a research infrastructure for early modern correspondence. The goal of this chapter – drawing in large part on our work for EULO – is to present a high-level overview of the entire research infrastructure and its data workflows. The chapter will offer basic summaries of (1) the challenges the infrastructure needs to overcome; (2) our reasons for proposing a distributed network architecture to address these prob-

[5] Working Group 5: 'Data Exchange and Strategic Planning', see http://www.republicofletters.net/index.php/working-groups/data-exchange-and-strategic-planning/, accessed 20/03/2019.

[6] 'CommonPlace: Linking European Experiences in Search of Shared Identities', submitted in 2017 by the Huygens ING (KNAW) as PI and sixteen partners to the Horizon 2020 EINFRA-21-2017 funding call; 'Early Modern Linked Open Data: A Framework for Aggregating Transnational Cultural Heritage Data and constructing European Identities' submitted in 2017 by the University of Oxford as PI and ten partners to the Horizon 2020 CULT-COOP-09-2017 funding call; 'Networking Archives: Assembling and Analysing a Meta-archive of Correspondence, 1509–1714' submitted in 2017 by the University of Oxford as PI and two partners to the UK AHRC funding scheme.

lems, and (3) how data will flow and be exchanged through the network; and the associated tools and services platform.

In acknowledgement of the ideas reproduced here from the original Horizon 2020 proposal we will continue to refer to this research infrastructure as 'EULO' or 'European Letters Online'.

2 Basic Challenges

To meet the requirements of its predominantly scholarly users, EULO will need to offer coherent and coordinated solutions to two fundamental and closely interrelated challenges.

2.1 Centralized versus Decentralized Functionality

The need for a central pool of homogeneous metadata. The *respublica litteraria* was held together by the exchange of manuscript letters. In order to obtain an adequate picture of the republic of letters, we need to reassemble data and metadata on correspondence scattered across and beyond Europe. Assembling such data in a series of individual silos, each devoted to a single main correspondent, archive, or library, is inherently unsatisfactory for several reasons. Perhaps the most fundamental is the fact that every letter pertains to at least two different correspondences: that of the sender, and that of the recipient. In consequence, a correspondence consisting of letters exchanged between one person and 100 others pertains to 101 different correspondences. In order to create a navigable data set and to avoid wasteful duplication of effort and expense, means need to be found to bring as much of this data together as possible. The place to begin is with epistolary metadata.

The need for decentralized data storage and curation. A monolithic system, in which all data is collected, curated, published, and preserved at a single location, is unworkable both in practice and in principle. *In practice*, the rate at which correspondence data is becoming available has already outstripped the ability of any single project to ingest, standardize, curate, store, and publish this material. Semi-automated metadata standardization can accelerate, to some degree, the rate at which records can be processed; but no central clearing house will be able to process the deluge of data unleashed when major repositories begin to release their catalogue metadata and associated data files at scale under open access. Moreover, the expertise needed to prioritize data creation, to negotiate access to materials, to raise the necessary funding, and to curate to the highest standards the data currently assembled in institutional and national repositories is distributed throughout Europe. *In principle*, major institutional curators of data (such as national libraries and archives but also regional, civic, and other specialist institutions) may need the option to retain authority and control of their original data sets. As metadata catalogue records are supplemented by images of manuscripts, transcriptions, and full online editions,

such national cultural patrimony must clearly reside in the countries, regions, and institutions to which it most closely pertains and where it can best be permanently sustained.

The need for centralized interrogation of distributed data. Although sustainable data storage and curation must be decentralized, many users will still want to be able to access, search, and analyse correspondence data in a centralized manner. Casual users will often want to search all the data available on the system. More specialist users will want to analyse and visualize selected data distributed throughout the system. Contributing users will also want to undertake further research in the context of all the data distributed in the system. For instance, those wishing to standardize, revise, and annotate metadata from one correspondent will need to do so with reference to all the data available to EULO, wherever it may be. New tools and services developed to crowdsource, transcribe, translate, analyse, and visualize this data will need to be exposed to the complete set of data as well as subsets of it. The consolidation of certain types of services and functions is one of the primary objectives of this project, since it is necessary for dealing with the geographically distributed nature of correspondence.

We therefore need systems which allow some processes to be handled centrally, in a consolidated manner, while others continue to be distributed throughout a network of nodes.

2.2 Diversity of Data, Models, Licences, and Software Interfaces

A second, fundamental problem is that the letter data and metadata currently deposited in Europe's libraries, archives, projects, and publishers is not homogeneous. The letters are preserved in different media, catalogued (if at all) to different standards and formats, and consequently also shared and referenced in a variety of different ways. Some metadata (at whatever level of detail) is accompanied by a variety of additional data in different formats, including page images, transcriptions, and/or translations of the letters themselves. As a starting point, the digitized content of the letter (e.g. the image of a manuscript or printed source or its text in machine-readable form) needs to be distinguished from the metadata describing it. Agreement will be needed on the minimal standards that will allow us to assemble or refer to digitized letters or their metadata in a consistent way.

*Metadata models***.** The level of detail in records to be shared differs greatly from one contributor to another. Some catalogues consist only of collection-level descriptions; others offer only the most basic item-level metadata (e.g. names of senders and recipients, dates, and shelfmarks); still others possess curated metadata (including detailed information on dating, locations of senders and recipients, incipits, explicits, abstracts, etc.). Many will not be linked to authority files such as VIAF or *GeoNames* for person and place names. Few, if any, will possess unique authority codes for individual versions of letters and manifestations. The reason for this is that most correspondence catalogues originating in institutional card

catalogues, printed editions, or stand-alone projects were not compiled from the outset with the ability to distinguish clearly among multiple instances of the same letter (e.g. in draft, as sent, as subsequently published etc.).

Licences. Conditions of sharing catalogue metadata and associated data files naturally differ between different content providers. Publishers have a commercial interest in preserving copyright. Research projects may need to embargo material being prepared by scholars as the basis of research publications. National libraries and archives have a duty to preserve national cultural heritage, and a special responsibility for the integrity of the catalogues that render that heritage accessible. One fundamental goal is therefore accommodating the needs of the full range of potential content providers by allowing different kinds of data to be shared and accessed in a variety of ways within a common framework.[7]

Interfaces. Even when data is released for reuse, different institutions use different interfaces (e.g. APIs) on a variety of commercial and open-source platforms to provide programmatic access to their holdings by other applications. The selection of APIs, data standards, formats, and protocols for data sharing must be harmonized to a lowest common denominator to encourage the most efficient and widest possible use.

One of the most basic challenges of this area is therefore to impose at least the minimal level of homogeneity necessary to render this data interoperable.

3 Basic Solutions

The lack of common data models and data quality standards is best addressed by distinguishing a subset of carefully standardized core metadata from other content and resources related to a letter. The separation of centralized and decentralized functionality is achieved by a distributed network of nodes, providing content, and hubs, providing services. The efficient communication and exchange of data between nodes and hubs can be accomplished by drawing on already well-established, lightweight, open data standards, protocols, and interfaces.

3.1 Distinguishing Core Metadata, Supplementary Metadata, and Associated Data

The basic solution to the first problem is to divide a letter's bibliographic *metadata* into two groups: (1) rigorously standardized, core metadata (to become, once curated, 'enriched core metadata') which can be released in the public domain; and (2) a much larger set of supplementary metadata, less prescriptive in standards and

[7] The multi-tier, 'Europeana Publishing Framework' (see https://pro.europeana.eu/post/publishing-framework, accessed 20/03/2019) – which governs and clarifies the relationship between Europeana and its content partners on the principle of 'the more you share, the more you get' – is a potential template for modelling these types of relationships.

(where necessary) more restrictively licensed. As well as reducing the task of standardization to manageable levels, this division allows maximum data sharing and integration (via core metadata) while protecting the intellectual property (within which most of the 'added scholarly value' resides) and institutional resources invested in the collection of supplementary metadata and other, associated digitized resources.

Core metadata on EULO will consist of just seven items: sender and recipient, places of sending and receipt, date of sending, location of the version (i.e. the physical or digitized instance) of a work, and a suitable, persistent identifier. This selection was the result of previous intensive consultations with the relevant scholarly, library, and technical communities and is already in wide use, including as an official TEI-XML extension ('correspDesc')[8] employed, for example, by the Berlin *correspSearch* project.[9] These core metadata fields are the prerequisite for exchanging data across the EULO network and must be shared under a public domain or equivalent licence for unrestricted access (e.g. discovery), revision (e.g. curation), and reuse (e.g. data analysis).

Most nodes will initially only be able to contribute this data in 'raw' form – that is, in whatever (semi-)structured form it currently exists in the contributing node's records. In order to deploy it in the EULO network, the data may need to be further normalized and curated. We anticipate that this curatorial work will ordinarily take place at a regional or thematic hub (although it could, of course, also be carried out by the contributing node in advance of forwarding this data to the hub).

Supplementary metadata above and beyond that described in core metadata such as letter abstracts, language details, or subject keywords, will have been collected by different institutions, projects, and individuals for a great variety of purposes. For this reason, they cannot be normalized to a similarly rigorous, single standard as core metadata but will need to be harmonized on the basis of an optional set of progressively more prescriptive standardization 'profiles'. On this basis, members of the EULO consortium will have the flexibility needed to agree on a set of simple models to describe, for example, the basic descriptions of some physical characteristics of a letter manuscript or settle on a reduced set of essential biographical details of the senders, recipients, and persons mentioned in the texts of letters. Sharing and exchanging supplementary metadata in the EULO network will, therefore, be encouraged and supported by a set of standards but will always remain voluntary.

Associated materials. In distinction to metadata, which benefits most from being organized and discovered centrally, related data files (such as full-text transcriptions or manuscript page images) can continue to be hosted and accessed from the nodes by reference only. This is necessary to accommodate content that can only

[8] See http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-correspDesc.html, accessed 20/03/2019.

[9] See https://correspsearch.net, accessed 20/03/2019. See also chapters I.3 and II.7.

be shared under more restrictive licences. Data files such as high-resolution page images are also much larger than metadata and (given the numbers of records potentially in the EULO network) cannot feasibly be stored and digitally preserved at one central institution or organization. Moreover, associated data in the form of structured full-texts (such as found in digital scholarly editions) is often closely coupled to the digital platforms for which it was prepared and will be of less value to users if shared as stand-alone files, or imported into a different, partially incompatible research or publication environment. In the fullness of time, means will need to be found for pooling freely available, machine-readable textual material for analysis in bulk; but describing these additional arrangements goes beyond the scope of the current conspectus, and will depend to a considerable extent on the arrangements for exchanging metadata outlined here.

3.2 A Distributed Network of Nodes and Hubs

The basic solution to the problem of centralizing some functions and decentralizing others is to design an infrastructure that functions as a distributed network in which nodes undertake the decentralized functions and hubs provide those services that are best organized in a consolidated fashion. In this section, we provide a short overview of the distinction between nodes and hubs and their functions in a distributed network.

The definition of a *node* is simply a server that shares content (i.e. core, supplementary, or associated data) with the EULO network. A server is a *hub* when it provides some additional service (i.e. besides sharing letter data) to other members of the network. For example, a server that collected and indexed metadata from several nodes and made this available on a portal website would function as a hub on the EULO network. A server that shared core and supplementary metadata with the network and also provided data analysis and visualization services would function both as a node (with respect to sharing letter data) and as a hub (with respect to providing analytical services). It is this potential for a multiplicity of hierarchical and non-hierarchical patterns that characterizes a distributed network architecture.

EULO *nodes* constitute a network of distributed content partners. Content providers include libraries, archives, publishers, and research projects with quantities of data and metadata on early modern correspondence. Nodes will assemble (if they are not themselves already data providers) data from one or more content contributors, separate core metadata from other (e.g. associated data) resources, transform, if needed, the 'raw' core metadata into the structured form specified by the EULO standard, and exchange these with one or more hubs providing additional services. In addition, nodes are responsible for the long-term preservation of the original data contributed by their content providers in its original form alongside the enriched and curated metadata returning to them from the hubs.

In practice, we expect that most content providers will prefer to assign the enrichment and curation of their data to a node or a hub. As we have seen, the EULO network supports a hierarchical model, in which scholars, research projects, publishers, and repositories can contribute and exchange data with one or more regional, national, or thematic nodes. For example, a national library (serving as a hub) might team up with an academy of science (serving as a thematic node) to coordinate the prioritization, funding, collection, and enrichment of data from a network of regional and local libraries, archives, and scholarly research projects with holdings in a specific area or discipline.

Hubs serve as aggregation and consolidation points for services provided to the EULO network. A hub, for example, can collect and validate the core metadata released by several contributing nodes, and then provide the staff and the technical facilities to curate and enrich it through a variety of means – automatic, semi-automatic, and manual. In the case, for example, of a thematic hub focused on collecting specific kinds of correspondence collections, this may be accompanied by other responsibilities, including serving as a portal for users to discover and access this data centrally, and for scholars to offer further refinements, comments, and analysis. And as noted above, it is also possible to be a hub and a node, if one is located at an intermediate level in the network hierarchy.

It is very important to recognize that data modified or amended in this fashion at a hub will be offered back again to the originating nodes where it can be optionally consolidated into the original data set (see below, 'Exchanging Data between Nodes and Hubs'). Besides this, other types of hubs (see below, 'Tools and Services Platform') will be able to offer additional services such as transcription or translation by interacting directly with hubs and nodes to the EULO network by means of defined, open interfaces, formats, and protocols.

4 Exchanging Data between Nodes and Hubs

The simplest way to describe the cyclical exchange of data among nodes and hubs in the EULO network is as a series of typical, sequential stages in an integrated workflow. For this reason, we will first describe the flow of data from a contributing node to an integrating hub where this data will be standardized, enriched, refined, and shared, and from where it will be offered back again to the originating node. Other types of hubs may offer different types of services to the network. These are discussed in the section 5, 'Tools and Services Platform'. The key stages of the workflow are illustrated in figure 1.

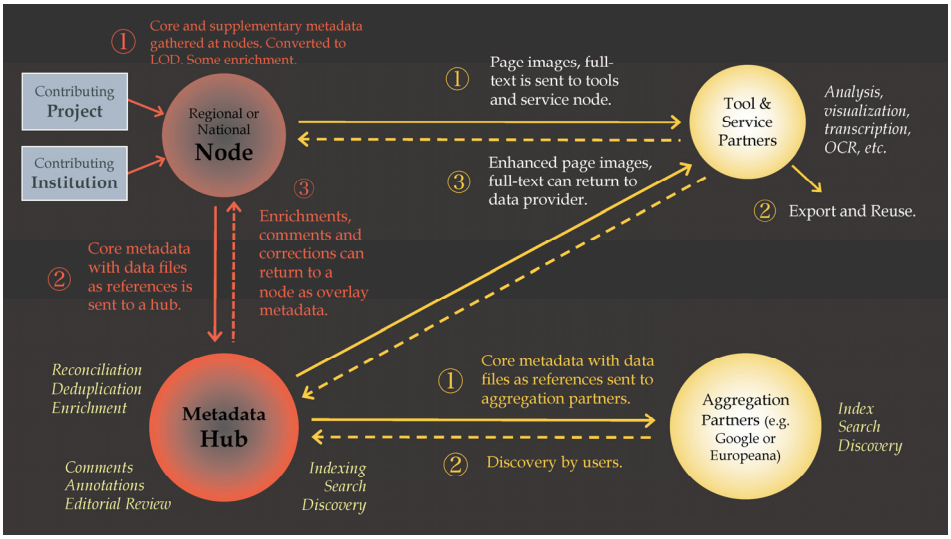


Figure 1: Curatorial and enrichment workflows

4.1 Nodes to Hub: Assembly, Conversion, and Release of Raw Core Metadata

Assembling data and metadata. The first function of a node is to assemble correspondence data and metadata. These resources may represent the holdings of an individual library, archive, publisher, or the product of an individual scholar's research project which will be shared with a thematic or nationally organized hub. Arriving from content providers, it is presumed to be in a 'raw' state which needs processing through several phases to produce the normalized core metadata that provides the main basis for efficient discovery and exchange throughout the system.

Distinguishing core and supplementary metadata. The materials assembled by the node may consist of both core and supplementary metadata as well as letter data in various forms (document images, rough transcriptions, and edited texts). Distinguishing these resources from one another is the second basic task of the node, since they need to be processed differently.

Mapping the core metadata to a structured format. In order to share metadata with the hubs, the nodes must first convert their raw core data into the structured format defined by the EULO core metadata standard. We expect that in the majority of cases, the initial 'on-boarding' data sent to a hub will be in a tabular structure in CSV or Excel formats. If a node already has their data available in a different, well-structured format, such as TEI-XML (e.g. 'CorrespDesc') then this may, at the discretion of the hub, be accepted as well. Nodes will not, at this stage, be expected to provide core metadata as curated Linked Open Data (serialized as RDF/XML,

JSON-LD, Turtle, or similar). This, instead, will be the format in which metadata will be sent back to nodes following its preparation and enrichment at a hub.

Releasing core metadata (and optionally, supplementary metadata) to the hub. For the initial transfer of new data to the hub no particular means need be specified: this could, for example, take place via a simple mechanism such as SFTP or data transfer via HTTP. For subsequent, synchronous exchanges of changes to this initial data set, the EULO standard will specify the appropriate 'push' and 'pull' mechanisms.

4.2 Hub: Curation, Enrichment, and Annotation

At the hub, the raw core metadata collected from the nodes is standardized, curated, and enriched. Enrichment will take several different forms (disambiguation, deduplication, reconciliation, followed by possible further enrichment/annotation) and may be undertaken by several different internal processes and workflows (manual/curatorial, semi-automated, and automated). We are confident that editors standardizing, and enriching bulk ingests of data from EULO nodes and scholars contributing fresh core and supplementary metadata to EULO will benefit from ongoing advances in open-source tools to semi-automate the standardization of metadata and accelerate its enrichment (a first generation of which is described in ch. III.2). The hub is also where scholarly users can discover core metadata from throughout the distributed network and participate in its enhancement and enrichment. All corrections of original data, amendments, and comments by editors and researchers registered at the hub will be recorded as annotations and credited to them with the appropriate technical and scholarly provenance metadata.

Curating and enriching core metadata

The next stage in the process is separately to curate and enrich the places, dates, people, and work components of these letter records in the manner described in section II.

Places. Metadata on places will be disambiguated and reconciled in the manner described in chapter II.2. To ease the identification and location of a place reference in a letter and to associate it with an authoritative, persistent URI, the editors at the hub will soon be able to draw on *EM Places*,[10] a historical geo-gazetteer currently under development by the University of Oxford and the KNAW Humanities Cluster. A EULO hub will be able to compare, over APIs, place references in letter metadata with the place entities recorded in *EM Places*. These are drawn, in the first instance, from current geographical data imported from a small group of reference gazetteers (currently: *GeoNames*, *Getty TGN*, and *WikiData*). However, *EM Places* will also offer a facility for editors and researchers to supplement this

[10] See https://github.com/culturesofknowledge/emplaces, accessed 20/03/2019.

data with additional, fully provenanced historical place name attestations and their historical political-administrative and ecclesiastical hierarchies. *EM Places*, like *EM People*, is being prepared as a collaboratively populated Linked Open Data resource, intended for use by anyone working in the early modern period.

Dates. The challenges accompanying the disambiguation of date metadata and its conversion into a normalized format are described in chapter II.3. To assist with this, the EULO network will draw on *EM Dates*,[11] an early modern date and calendrical resource also under development at the University of Oxford and the KNAW Humanities Cluster. First, a letter's place and date references are compared to a list of places located within historical regions (polities) recorded in the *EM Places* gazetteer. Whenever a matching place is found, *EM Places* will return, if it can, the official calendars in use in that region in the relevant time period. By drawing on this data, *EM Dates* can suggest the appropriate (e.g. Julian to Gregorian) calendar and accompanying date conversion. *EM Dates* can also be used by EULO editors as a resource for semi-automatically parsing and converting into a standardized format, early modern dates written in Roman (or partially Roman) nomenclature (e.g. 'A.D. XVI KAL. MART. 1645'). The date metadata returned to the hub will include provenance data recording how the conversions were made together with ISO 8601-2 codes indicating the certainty and completeness of the resulting dates.[12]

People will need to be disambiguated and reconciled in the manner described in chapter II.4. A particular challenge of disambiguating person references is that the great majority of people mentioned in early modern correspondence are not prominent enough to be found in national biographies such as the *Oxford Dictionary of National Biography*, the German GND, or even in much larger, aggregated sources such as VIAF, or crowdsourced resources, such as *Wikipedia*. To help better identify such persons and assign them with authoritative persistent URIs, plans are underway to create *EM People*, an early modern, Linked Open Data name authority and prosopographical resource.[13] *EM People* will initially be based on the c. 25,000 disambiguated people already recorded in *Early Modern Letters Online*[14] who were either mentioned in, or senders or recipients of letters. Allowing early modern scholars and projects to access freely, download, and subsequently revise and contribute new records on this resource will significantly help editors in the EULO network to identify and disambiguate lesser-known persons and assign them with identifiers.

Letters. The Letter Model outlined in chapter II.7 distinguishes between 'records' of either 'originals' or 'copies' of various 'states' of letters ('draft text', 'sent letter', 'altered text', and 'published text'), all of which are subordinate to a concep-

[11] See https://github.com/culturesofknowledge/emdates, accessed 20/03/2019.
[12] See https://www.iso.org/standard/70908.html, accessed 20/03/2019.
[13] See http://www.culturesofknowledge.org/?p=8455, accessed 20/03/2019.
[14] See http://emlo.bodleian.ox.ac.uk, accessed 20/03/2019.

tual 'letter work' which contains the persistent identifier. The enrichment process for letter metadata will consist largely in identifying and deduplicating the entities of these kinds related to an individual letter work, within the framework provided by the letter model.

Curating and enriching supplementary metadata

Because a hub providing curatorial and enrichment services can, in principle, also be a data contributor, it is possible for it to host and curate its own bespoke supplementary metadata alongside the standardized EULO core metadata. But in practice, because EULO will only suggest levels of standards ('profiles') for supplementary metadata, this data will usually be edited/curated at a node according to the local standards and practices established by the contributing institutions or project. Hubs, therefore, will be able to ingest, store, and index references to supplementary metadata provided by contributing nodes, but will not ordinarily edit it themselves. Lastly, we can also envision that certain types of supplementary metadata (such as biographical or geo-spatial data) could be contributed by external data providers and either linked or merged into the EULO knowledge graph.

4.3 Hub to Node: Return of Enriched Core Metadata to Nodes

All enrichments (which include edits) made at the hub will be treated as annotations on the original data and stored separately according to the W3C Open Annotation model[15] thereby permanently distinguishing it from the original 'raw' data contributed by nodes. The provenance of each enrichment will be tracked and attributed to an individual editor by means of an identification and provenance model. This will allow scholarly work to be attributed to its authors, workflows to be monitored, and quality to be controlled (e.g. by identifying or filtering out all the annotations contributed by an individual user or class of editors). For cross-platform user identification, EULO-hubs will provide support for ORCID identifiers,[16] supplemented by the federated IAA (authentication and authorization) model already widely employed by the European DARIAH[17] and CLARIN[18] research consortia. An opt-in for voluntary registration of further contact details will be provided in order to allow other users to share more details with their colleagues.

[15] See https://www.w3.org/annotation/, accessed 20/03/2019.
[16] See https://orcid.org, accessed 20/03/2019.
[17] See https://www.dariah.eu, accessed 20/03/2019.
[18] See https://www.clarin.eu, accessed 20/03/2019.

Release of enriched metadata by hubs

The normalized and enriched core and supplementary metadata records will be uniquely referenced, indexed, made discoverable, and downloadable as appropriate, over the web. In addition, both hubs and nodes will provide defined APIs to allow, for example, a subset of the metadata curated at a hub to be shared publicly with an external search aggregator such as *Europeana*, *Gallica*, or *Google*, or for related data to be accessed, for example, by an industry partner offering transcription, analysis, or visualization capabilities.

Return of enriched metadata to nodes

Crucially, the enriched records can be pushed back to the originating EULO institutional partner nodes as overlay metadata via the SWORD API protocol.[19] The originating partner can then opt (1) to incorporate the enrichments into its original records; (2) to display the enrichments as annotations while keeping its records unchanged; or (3) to withhold the enrichments, as institutional policy dictates. This means that the contributing institutions have the option of consulting the enriched records returned from the hub but are not obliged to integrate them into their core records. As a result, users benefit from enriched metadata while repositories are not required incrementally to update their records or to replace records generated within the institution with records of complex provenance generated elsewhere. A basic version of this arrangement has been piloted successfully with the incorporation of the hard-copy *Index of Literary Correspondence* in the Bodleian Library within *Early Modern Letters Online*.[20] An image of each index card from this catalogue of early modern intellectual correspondence is accessible from the electronic catalogue record derived from it; and although the electronic records have been corrected, normalized, and enhanced in many respects by a variety of hands, users and the originating institution can always consult and refer to the original file card if they prefer.

User-generated metadata

Not all metadata will arrive in bulk directly from institutional nodes. Scholars working on material outside the nodes will still be able to contribute content into a EULO hub. It will be possible, for example, to create annotations or add related data files (such as digital images and transcriptions of manuscripts, with appropriate permission) to the core metadata. Registered researchers contributing data to the hub in this way will be provided with a personal virtual workspace, within which they can prepare a body of material prior to releasing it. Within this restrict-

[19] See http://swordapp.org, accessed 20/03/2019.

[20] See http://emlo-portal.bodleian.ox.ac.uk/collections/?catalogue=bodleian-card-catalogue, accessed 20/03/2019.

ed space, they can decide with whom in the hub they wish to share these records and to embargo its public release for a period of time during which they can, for example, complete a publication that draws on this data.

5 Tool and Services Platform

Service and tool providers interact with EULO via API services to navigate and access the material. A EULO metadata hub, as the likely entry point for a researcher, will provide a portal for linking to tool services such as visualization and transcription. If the tool requires it, a EULO hub can expose the annotation/enrichment mechanism via a private API key to allow the output of a tool to be routed back to the source repository. This API key mechanism can be expanded, if required, to allow bespoke sharing of content between nodes, the hub, and tool providers without releasing this material publicly. Some of these arrangements are illustrated in figure 2.

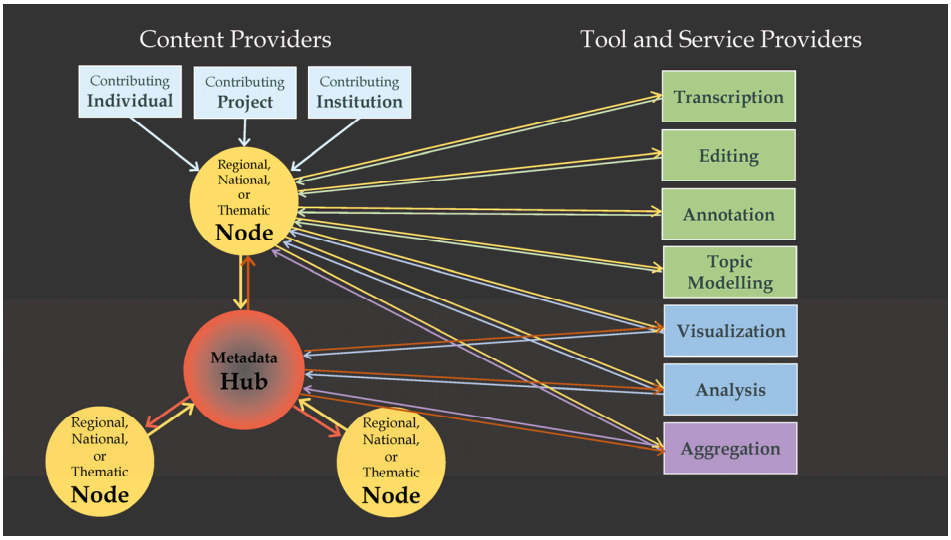


Figure 2: Tools and services workflows

5.1 Nodes to Tool Platforms: Transcription, Editing, Annotation, and Analysis of Texts

No major repository of material in this field will be complete until catalogue metadata is accompanied with digital images of manuscript and printed letters and machine-readable texts. Whenever permissible, these should be released for processing by analytic and exploratory tools appropriate to large digital object collec-

tions – using a EULO hub for discovery and navigation but retrieving texts and images directly from the contributing repositories. In order to facilitate the development of textual content, this toolkit could offer, for example, an interface to transcription and translation tools, with access, quality control, and publication mechanisms aligned with EULO. A transcription tool could leverage hub search services to locate letters of interest with digitized images, retrieve the images directly from the institutional repositories via the International Image Interchange (IIIF)[21] format and then feed the transcriptions back via the enrichment API. This additional content could then be routed back to the originating repository via the enrichment mechanism described above.

5.2 Hubs to Tool Platforms: Browsing, Querying, Visualization, and Analysis of Metadata

Drawing on the data collected in EULO, the web presence will offer modern browsing, querying, and visualization options. These could enable, for example, inferential queries based on semantically structured prosopographical data, and interactive visualizations to explore the contributions and connections among correspondents.

By accessing data in EULO over a public API scholars will be able to employ a variety of analytical tools from related disciplines to further their own research interests and enhance the utility of the aggregated corpus. A visualization tool, for example, can be linked from a search results page, provided that the contributing node supports an entry point that accepts EULO URIs.

5.3 Facilitating Implementation

Reference Models and Guidelines. In order for EULO nodes, hubs, and services to exchange data in a controlled and coordinated manner it will be essential to have clearly defined and documented protocols, APIs, metadata standards, and file formats. Good software development practice for any interoperability initiative dictates that specifications should be accompanied by a reference implementation and a validation test set. This not only provides others with a way to test their own implementation efforts but ensures that the designed specifications are actually implementable and fit for purpose. In addition, future scholarship will be facilitated by guidelines and suggested practices that can be applied to material that is *not* intended for exchange within the current framework, especially when creating new resources or refreshing existing content.

The Semantic Web. The models for exchanging data presented in this document have been described at a high level since our primary concern is to communicate and validate the overall conceptual framework before progressing to detailed tech-

[21] See https://iiif.io, accessed 20/03/2019.

nical specifications. Nevertheless, it should already be evident that the models for almost all of the entities concerned (people, places, events, and letters) are more adequately and easily captured in graph-like representation rather than hierarchies or tables. As such, these models are likely to find their most effective expression using RDF and are intended to be processed and managed using Semantic Web technologies.[22] In the first instance, we expect that these technologies would be implemented at service hubs. Over time, we expect that the technology will trickle down to nodes as their own content technology platforms are revised and upgraded. Although the creation of suitable ontologies (i.e. structured sets of defined classes of entities and their connecting predicates expressed by RDF triples) for EULO will be challenging, we feel strongly that Linked Open Data[23] and its related Semantic Web technologies are the best means available to our community for disseminating and harvesting scholarly data over the web.

[22] See https://en.wikipedia.org/wiki/Semantic_Web, accessed 20/03/2019.

[23] See https://en.wikipedia.org/wiki/Linked_data, accessed 20/03/2019.

IV Scholarship in a Digital Environment

IV.1 Beyond Visualization

Paolo Ciuccarelli and Tommaso Elli

1 Digital Humanities and (Communication) Design United

Paolo Ciuccarelli

This is a general rule of Digital Humanities:
you always need an Italian designer at some point.

– B. Latour, *Rematerializing Humanities Thanks to Digital Traces*, Keynote, DH Lausanne, 2014

On 25 October 1870, Charles Minard, a French civil engineer serving as *inspecteur général* at the École des Ponts et Chaussées in Paris, died at the age of eighty-nine. Besides being an excellent engineer he had also been a pioneer in thematic cartography and statistical graphics: 'The fifty-one *cartes figuratives* that came from his fertile mind and adept hand show a combination of cartographic ingenuity and concern with the graphic portrayal of statistical data that was almost unique during the central portion of the century'.[1] Minard designed one of the most important milestones in the history of the visual representation of data, 'Probably the best statistical graphic ever drawn' according to Edward Tufte.[2]

[1] Arthur H. Robinson, 'The Thematic Maps of Charles Joseph Minard', *Imago mundi* 21 (1967): 95–108, see https://doi.org/10.1080/03085696708592302.

[2] See https://www.edwardtufte.com/tufte/posters, accessed 20/03/2019.

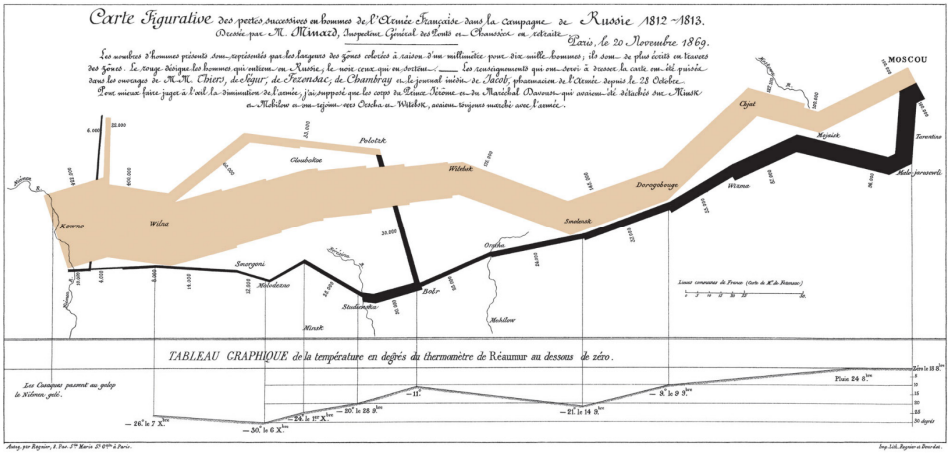


Figure 1: Charles Minard's famous map of Napoleon's disastrous Russian campaign of 1812, with six types of data on a two-dimensional surface: the number of Napoleon's troops, the distance they travelled, the temperature they confronted at each stage, the latitude and longitude, their direction of travel, and their location at specific dates. Lithograph, 62 × 30 cm, published 20 November 1869[3]

In 1871 the *Annales des Ponts et Chausées* published the obituary written by Minard's son-in-law Victorin Chevallier; this text is both a tribute to the scientific career of Minard and a concise yet comprehensive synthesis of what makes the use of visual languages essential in any attempt to derive knowledge from data. Chevallier writes: 'For the dry and complicated columns of statistical data, of which the analysis and the discussion always require a great sustained mental effort, he had substituted images mathematically proportioned, that the first glance takes in and knows without fatigue, and which manifest immediately the natural consequences or the comparisons unforeseen'.[4]

Minard's famous images exemplify, and Chevallier's commentary explicates, a vitally important point for introducing the issue of data visualization into humanistic research practice: visualizations are rapidly becoming ubiquitous in the humanities, not because they are 'trendy', beguiling, or merely decorative, but because they are powerful, effective, and efficient. At the root of the matter is the very natural and universal interaction between our cognitive and sensorial systems. Sight is our 'broadband' sense. Our eyes contain 70 per cent of all sensory receptors.[5] Our brain is a pattern-seeking machine, made especially to process visual signals:

[3] Source: Wikimedia Commons (public domain).

[4] A translation of Minard's obituary by Dawn Finley. Victorin Chevallier, 'Notice nécrologique sur M. Minard, inspecteur général des ponts et chaussées, en retraite', *Annales des Ponts et Chaussées* 2 (ser. 5, no. 15, 1871): 1–22.

[5] Elaine N. Merieb and Katja Hoehn, *Human Anatomy & Physiology*, 7th ed. (San Francisco: Cummings, 2007).

20–30 per cent of the total surface area of the cerebral cortex is largely or exclusively involved in visual processing.[6] Some of these processes are 'low-level' operations performed automatically, sometimes referred to as 'pre-attentive' because they occur without conscious intervention and control: we do not need to 'pay attention' to ensure that they are completed'.[7] Visual perception is also at the heart of human evolution[8] and is intimately connected with the very act of thinking: 'Far from being a mechanical recording of sensory elements, vision proved to be a truly creative apprehension of reality – imaginative, inventive, shrewd, and beautiful'.[9] We think through images which, as Chevallier puts it, 'the first glance takes in and knows without fatigue', because we're all made for that as human beings.

So it's no surprise that, no matter the context and the discipline, anyone working with data ends up using visualization when coping with big and complex sets, in order to improve the process of understanding and making it more efficient. The same is currently happening in the 'digital transformation' of the enquiry processes of humanities scholars: as soon as data – and especially metadata – was made available in unprecedented volumes, the power of visualization in taming them became evident.

How then to harness the capacity of vision to make sense of unprecedented quantities of data in the humanities? An obvious first step is to adopt tools developed mostly in the domain of data analysis to automatize the production of mathematically proportioned images trying to mimic the sophistication of Minard's work. Browsing the web pages of research groups and initiatives in the digital humanities, one frequently encounters laudatory descriptions of the value of tools for visual analysis, such as *Tableau*,[10] for supporting the work of the scholar.[11] What is usually missed or merely implied is the fact that the scholar importing these tools into research practice inevitably and often unconsciously also adopts a scientific, analytical approach to data, information, and knowledge embedded in those tools.

The rapid adoption of these ready-made visual applications was soon followed by a growing awareness of the limitations[12] of introducing tools and methods cre-

[6] Chris I. Baker, 'Visual Processing in the Primate Brain', in Irving Weiner, Randy J. Nelson, and Sheri J. Mizumori, eds., *Handbook of Psychology 3: Behavorial Neuroscience*, 2nd ed. (Hoboken, NJ: Wiley, 2013). See https://doi.org/10.1002/9781118133880.hop203004.

[7] Ian Spence, 'William Playfair and the Psychology of Graphs', in *Proceedings of the American Statistical Association*: *Section on Statistical Graphics* (Alexandria, VA: American Statistical Association, 2006): 2426–36.

[8] William G. V. Balchin, 'Graphicacy', *American Cartographer* 3 (1976): 33–8, see https://doi.org/10.1559/152304076784080221.

[9] Rudolf Arnheim, *Art and Visual Perception: A Psychology of the Creative Eye* (Berkeley, Calif. and London: University of California Press, 1974). See also: Rudolf Arnheim, *Visual Thinking* (London: Faber, 1969).

[10] See https://www.tableau.com/, accessed 20/03/2019.

[11] See as examples of this practice: https://digitalhumanities.berkeley.edu/content-analysis-tableau or http://digitalhumanities.uchicago.edu/node/99 or http://dh101.humanities.ucla.edu/?page_id=163, both accessed 20/03/2019.

[12] 'One of the first discoveries was actually not what we visualized, but what we could not visualize': Dan Edelstein and Paula Findlen, 'Digging into the Enlightenment: Mapping the Republic of Letters'.

ated to foster insights through an analytical approach: encoding data as graphical marks, visual variables, and abstract models dramatically enhances the ease with which numbers can be compared and patterns discovered; but it does not adequately support the interpretative process at the core of humanistic inquiry. Visualization patterns could limit or even mislead the interpretation of scholars, for instance in the case of graphs and networks; visual languages framed in the abstract, rigorous, and quantitative rhetoric of science cope poorly with the complex, multidimensional, and sometimes ill-defined social and historical phenomena of the humanities:[13] the incompleteness, uncertainty, and ambiguity of humanistic data are typically rendered invisible in the pursuit of the ersatz precision and objectivity of scientific, analytical visualizations.

These limitations have become even more apparent in the move to what is sometimes called the 'second wave' of digital humanities:

The first wave of digital humanities work was quantitative, mobilizing the search and retrieval powers of the database, automating corpus linguistics, stacking hypercards into critical arrays. The second wave is qualitative, interpretive, experiential, emotive, generative in character. It harnesses digital toolkits in the service of the Humanities' core methodological strengths: attention to complexity, medium specificity, historical context, analytical depth, critique and interpretation.[14]

To make matters worse, whereas the modes of visualization produced by Minard and other scientists have been devised to visualize structured data produced by statisticians, the digital humanities also need to work with text-heavy, unstructured data which was unavailable when Minard pioneered the field.

This growing acknowledgement of the limited applicability to humanistic material of the analytical approach to data and visualization imported from the sciences naturally led to the search for partnership with other disciplines that could help adapt data visualization to the interpretation and contextual analysis of complex historical data. That is where communication design came into play, because this is precisely what design has been always doing: bridging scientific advancement and human needs by leveraging its nature as an 'interdisciplinary, integrative discipline'[15] placed at 'the intersection of several large fields'.[16] Crafting materials to

(National Endowment for the Humanities: https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HJ-50056-10, accessed 20/03/2019).

13 'Attention was paid on finding effective visual encodings, but for a "generic" idea of flows between cities and persons over time. One of the strongest criticisms to the project involved, in fact, the visual language and the rhetoric adopted in the tool, that, according to Coleman (2010), conveyed a misleading idea of a correspondences network during the Enlightenment as a well-defined and clearly perceivable phenomenon': Giorgio Caviglia, 'The Design of Heuristic Practices. Rethinking Communication Design in the Digital Humanities', PhD Thesis, Politecnico, Milan, 2013.

14 Jeffrey Schnapp and Todd Presner, 'Digital Humanities Manifesto 2.0' [2009]: http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf, accessed 20/03/2019.

15 'The foundation of design theory rests on the fact that design is by nature an interdisciplinary, integrative discipline': Ken Friedman, 'Theory Construction in Designresearch: Criteria, Approaches, and Methods', *Design Studies* 24:6 (2003): 507–22, at 508. See https://doi.org/10.1016/S0142-694X(03)00039-5.

make sense of data through technology, creating interfaces that meet the needs of a specific user in a specific context for a specific purpose is the essence of the design practice. As data, information, and knowledge processes spill from the scientific domain of the analyst into the realm of the humanities, new needs are becoming apparent, since the interpretation performed by digital humanists shows very little resemblance to the analytical approach of a scientist. The greater this difference, the more necessary design becomes in 'translating' data and its processes into this new humanistic domain.

Between 2009 and 2012, a number of pioneering conferences and projects explored the prospects for a closer relationship between humanities and design. In 2009, UCLA's Design Media Arts Department hosted 'the first conference to apply contemporary design theory to emerging issues in the digital humanities' – with the title 'Design Theory + Digital Humanities' – proclaiming that 'learning from communication design, interaction design and industrial design will be vital to 21st century humanistic inquiry'.[17] In 2010, the HyperStudio and Digital Humanities at MIT gathered digital practitioners and humanities scholars together with experts in art and design around 'the past, present, and future of visual epistemology in digital humanities' under the heading 'Humanities + Digital. Visual interpretation'.[18] The importance of this partnership was further articulated by Burdick et al. in 2012, framing design more as an intellectual method and less as a technical activity: 'As Digital Humanities both shapes and interprets this imaginary, its engagement with design as a method of thinking-through-practice is indispensable'. Within the broad area of design a specific role is assigned to communication design: 'Digital humanists have much to learn from communication and media design about […] how to juxtapose and integrate words and images, create hierarchies of reading, forge pathways of understanding, deploy grids and templates to best effect, and develop navigational schemata that guide and produce meaningful interactions'.[19]

The path of the collaboration between the DensityDesign Research Lab and the Stanford Humanities Center (SHC) proves the 'natural' tendency of design and digital humanities to converge. The partnership started as an attempt to overcome the issues that emerged from a first visualization experiment developed by SHC with the Stanford Vis Group: a dashboard-like, quantitative, and analytical visualization tool was developed in the context of the *Mapping the Republic of Letters* initiative, with the goal of supporting the scholarly work of the humanists. If on one side the experiment shed light on the power of visualization as a supporting tool in

16 Facilitated by: 'The nature of design as an integrative discipline places it at the intersection of several large fields': Friedman, 'Theory Construction', 508.

17 'Nowcasting: Design Theory + Digital Humanities', see http://www.dma.ucla.edu/nowcasting/about.html, accessed 20/03/2019.

18 Keynote speakers: Johanna Drucker (UCLA), Lev Manovich (UC San Diego), Ben Shneiderman (University of Maryland), Fernanda Viegas, and Martin Wattemberg (Flowing Media).

19 Anne Burdick, Johanna Drucker, Peter Lunenfeld, Todd Presner, and Jeffrey Schnapp, eds., *Digital Humanities* (Cambridge, MA: MIT Press, 2012), 13.

the analysis of the metadata layer – e.g. to spot patterns at a glance[20] – on the other hand it also brought to the surface the pitfalls of encoding data following the principles of visual analysis, applying visual patterns that rely on the abstraction of mathematics and statistics and do not take interpretation into account. That is what pushed Stanford's humanists towards the more open and agnostic approach of design. In August 2012, the 'Early Modern Time & Networks' event – defined as a 'Design + Humanities workshop' – laid the foundations for a collaboration between the two research centres and, more broadly, for a closer alliance between the two disciplines. The collaboration not only produced a set of custom tools developed following the specific needs of humanistic enquiry; it also led to the consolidatation of the partnership in the form of a research organization under the name 'Humanities + Design Lab'.

This chronicle of steady progress is encouraging; but, as in the first wave of digital humanities, progress in the application of design principles and practices to the digital humanities likewise rapidly throws new limitations into relief. One crucial limitation is highlighted by the tendency to indicate the nature of the new partnership with the symbol '+', which speaks more of juxtaposition than of full integration. Despite the declarations of interest and affinity, designers have rarely been involved in a truly collaborative activity. In many cases, digital humanists or their collaborators in informatics play the role of designer, while fully-fledged 'designers are nowhere to be found'. In place of well-meaning mimicry and the importation of alien competences and methods, a deeper union is needed – as history of design with other disciplines can tell – to unlock the potential and establish a fruitful interdisciplinary collaboration.

It is in this light that the COST Action IS 1310, *Reassembling the Republic of Letters*, provided a unique and timely opportunity to nudge the field forwards from theoretical affinities, hypotheses, and good intentions to the closer integration of actual practice. This was the aim of the Action's Working Group 6: to plunge scholarly colleagues into the midst of three trends that have been shaking up research in information design for some time and are now also sparking discussions in the digital humanities. Of necessity, each of these trends can only be described here in a very succinct manner. Although they merit more detailed treatment, the aim is to say enough to help provoke further discussion.

The first trend begins with growing awareness of the limitations of the 'dashboard', both as a tool and as a metaphor. This metaphor implies that we understand a phenomenon by reducing it to a set of key performance indicators (speed, rpm, altitude, fuel consumption, and the like) which are visualized as analytical

[20] 'While you could have teased that out of the 20 volumes of Voltaire's correspondence', with GIS (geographical information system) mapping technology, you can see it at one glance.': 'Dan Edelstein and the collaborative future of the digital humanities: geeks and poets, unite!', 18 November 2010, by Kristi McGuire, in *The Chicago Blog: Intelligent Commentary, Curated Content, News, Reviews, and All Things Digital* (University of Chicago Press). See https://pressblog.uchicago.edu/2010/11/18/dan-edelstein-and-the-collaborative-future-of-the-digital-humanities-geeks-and-poets-unite.html, accessed 20/03/2019.

patterns organized into a mechanistic or more properly *skeumorphic* interface, that is, an assembly designed to resemble the cockpit of an airplane. Yet, as already established above, this assumption conflicts with the growing appreciation of the complex nature of the phenomena studied in humanistic disciplines. The tools of visual analysis with which digital humanists have mostly been playing are reductive by nature and far from being able to convey the complexity, the multiple dimensions and connections of a social and historical phenomenon. In place of this impatient oversimplification, this first line of enquiry seeks to develop means for weaving the multiple dimensions of a complex phenomenon into a coherent picture, or better into a consistent visual experience: the kind of data experience more appropriate to humanistic disciplines. The focus is less on the efficiency of the visualization of the single indicator and more on the integration of the many dimensions of a complex phenomenon into a coherent visual shape.

This first line of development gives rise directly to a second. Since a single visualization is rarely adequate to convey the richness of a complex issue, multiple views are often needed. Every data set must therefore be transformed into a number of different visualizations, each of which reveals a different pattern and establishes a different perspective. But the exploration of data through sequences of visualizations, like interchanging lenses, leads naturally to narration, to storytelling, to the articulation of a discourse through a series of story points in which data are complemented with contextual information.[21] This could be developed into a more specifically humanistic approach to visualization, in which the structure of the unique pattern of an analytical visualization is increasingly displaced by a narration composed of multiple visualizations. If we start from the idea that every visualization is potentially part of a 'narrative', which produces a message and orientates perception, then visualization begins to be transformed from a technological device into a cultural artefact. The idea of data as something 'constructed' rather than given is something that distinguishes the humanistic from the scientific approach to digital data analysis (see ch. I.3). This idea that visualizations as well as the data underlying them are also constructed brings us to the third ongoing development: namely, the idea that visualization must be seen more as a design process and less as a *product*. Even if the data itself is constructed in the process of being extracted from other underlying historical documentation, the meaning is still not latent in the data: it is constructed in the very act of visualizing it and, still more so, by the process of developing a sequence or kaleidoscopic variety of visual perspectives on the data, which gradually build up a specific interpretation. Visualizations, especially when they are rendered interactive, become design tools in their own right, inter-

[21] Scott Bateman et al., 'Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 10–15 April 2010, Atlanta, Georgia, USA* (New York, NY: ACM 2010), 2573–82, See https://doi.org/10.1145/1753326.1753716.

faces[22] that 'perform typical design activities (i.e. selecting, organizing, manipulate, modeling, representing)'.[23]

This trend in the field of data interaction design is mirrored by an emerging tendency within the digital humanities to prize the scholarly process as much as the scholarly outcomes. While traditional scholarship is typically regarded as a solitary process, taking place within the mind of the lone researcher, the collaborative work of the digital humanities is increasingly seen as an act of construction, a shift from reading to making or 'thinking through making', that once again opens the door for an alliance with design: 'Process is the new god; not product. […] The theory after theory is anchored in MAKING, making in the poetic sense of *poiesis*, but also in the sense of design carried out in action'.[24] From this perspective, it is possible to see the meaning-making, constructive, interpretative process collapsing in the interface[25] that mimics the design-reasoning process, at which point the research process itself can be seen as a design process: 'The question about how design can participate in the digital humanities research, seems to be solved not by bringing design into research processes, but, rather, to see research as a design process'.[26]

Far from being purely theoretical, this line of thinking palpably shaped the design practice institutionalized in the COST Action within a pair of innovative and experimental 'data-design sprints'. The converging interest of the digital humanities and the design community in a constructive approach necessitated that we focused, not merely on the *idea* of sharing the ideas behind these three tendencies in theoretical discussions, but rather on the *practice* of embedding them in our actual collaborative work. It is in this regard above all that the Action sought to move beyond the approach to 'Humanities + Design' which emerged from the MIT conference in 2010. Instead of merely employing existing visualization tools designed for other disciplines and purposes, or merely mimicking design practices without fully mastering them, or simply juxtaposing digital humanities with design, the current challenge is to rethink visualization and design, broadly speaking, as a set of methods *integral to* humanistic scholarship. Better still, the objective is to develop a common approach shared by these two domains, which goes beyond data and analysis to fully integrate a novel, interdisciplinary, visual epistemology in new forms of research organizations and hybrid practices.

---

[22] 'How can we make visualizations function as interfaces, in an iterative process that allows the user to explore and tinker?': John Unsworth, 'New Methods for Humanities Research' (2005), see http://www.people.virginia.edu/~jmu2m/lyman.htm, accessed 20/03/2019.

[23] Caviglia, 'The Design of Heuristic Practices'.

[24] Schnapp and Presner, 'Digital Humanities Manifesto'.

[25] 'Visualizations and interfaces are not conceived as things, but rather, as moments of an interpretative process involving new ways of looking, reasoning and building with and through digital technologies.': Caviglia, 'The Design of Heuristic Practices' (Abstract).

[26] Caviglia, 'The Design of Heuristic Practices'.

In pursuing this goal, we were not breaking entirely new ground. On the contrary, a credible roadmap towards that goal could already be derived from the evolution of the relationship that the DensityDesign Research Lab had established with SHC in pursuit of the *Mapping the Republic of Letters* initiative. The first phase of their shared activities was one of adaptation, with existing tools developed by the lab for other projects bent to try to fit the needs of the humanities scholar. The advantages and disadvantages revealed by this first phase led to a second type of collaboration in which workshops and close collaboration were the keys to developing new, customized tools – such as *Palladio* – by the two disciplines. In the third phase, the convenience of a continuous relationship became so evident that a permanent Humanities + Design research lab was established at Stanford University.

The question arising from this successful collaboration is how similar interaction can be developed on a broader scale. Given that not every institution has the resources of Stanford, how can similarly collaborative co-creation of new modes of visualization and data interaction be fostered on a larger scale and across a broader front? How, indeed, can the union of scholarship and design be embedded in emerging academic practice in a manner which might eventually become the rule rather than the exception? The pursuit of answers to this question resulted in one of the most innovative and successful experiments conducted in the course of COST Action IS 1310. The basis of this experiment was to transform the format of the standard COST working group meetings and training schools into a far more innovative, interdisciplinary version of a hybrid data and design sprint,[27] in which humanists and designers went beyond the mere adding together of their competences: they engaged in an intensive week of physical co-habitation that produced outcomes that could never have been achieved otherwise, while exposing a large, international, and interdisciplinary cross-section of the Action's community to the fertility of collaboration between humanists and designers. The following section describes in detail this experimental mode of collaboration and summarizes some of its results, which loom large in this fourth section of the volume.

[27] 'Not every digital humanist will become a designer, but every good digital humanist has to be able to "read" and appreciate that which design has to offer, to build the shared vocabulary and mutual respect that can lead to fruitful collaborations': Burdick et al., *Digital Humanities*, 13.

2 Design-Sprint Methodology for Reassembling the Republic of Letters

Tommaso Elli

2.1 Design Sprints and Data Sprints

To *sprint*, literally, means 'to run as fast as you can over a short distance'.[28] In the world of software development, the word has been repurposed to denote a delimited time-frame in which to work on specific tasks and to produce specific outcomes for testing. It is one of the components of Agile Software Development and resembles other collaborative digital activities such has hackathons.[29]

The *design sprint* is a variety of this practice which has been widely adopted in the commercial environment in recent years. As developed by Google Ventures,[30] it represents a structured workplan for a heterogeneous group of preselected people directed by a facilitator and unfolding over a five-day period, with each day devoted to a specific task: day one is supposed to *map* or *understand*; day two, to *sketch* or *diverge*; day three, to *decide* or *converge*; day four, to *prototype*; and day five is devoted to *user testing*. This now well-established structure is supported by a variety of online materials designed to maximize efficiency, including checklists to be printed and filled and even presentation templates to guide participants.[31]

On the surface, at least, this model of design sprints is deeply immersed in the world of product development and thus of commercial companies. The objective is normally to design new products, or new features for existing ones, and to arrive swiftly and predictably at a precise plan for follow-up activities. The predefined timetable and time-saving supporting materials prioritize efficiency over flexibility. This represents a first contrast with the objectives of a more scholarly equivalent, which might instead prioritize less concrete outcomes, such as suggesting research questions, learning or developing new methods, questioning underlying assumptions, or increasing understanding of context.

---

[28] *Cambridge Advanced Learner's Dictionary and Thesaurus*, see https://dictionary.cambridge.org/dictionary/english/sprint, accessed 20/03/2019.

[29] 'Agile software development' allows the objectives, requirements, and solutions being developed in a project to evolve through a flexible and collaborative exchange between cross-disciplinary teams of developers and end users. A 'hackathon' brings together computer programmers, sometimes together with graphic designers, project managers, end users and others, to collaborate in software development, generally with a view to creating a functioning product by the end of the event.

[30] Now known as GV, a venture capital company owned by Alphabet Inc., see https://www.gv.com/, accessed 20/03/2019.

[31] An introductory presentation to the design-sprint format is available here: https://www.dropbox.com/s/xm6svbq5ds58xgq/SPRINT%20kickoff%20slides.pdf?dl=0, accessed 20/03/2019. See also: Jake Knapp, John Zeratsky, and Braden Kowitz, *Sprint: How to Solve Big Problems and Test New Ideas in Just Five Days* (New York: Simon & Schuster, 2016); and Wikipedia contributors, '*Design sprint*', in Wikipedia: The Free Encyclopedia: https://en.wikipedia.org/wiki/Design_sprint, accessed 20/03/2019.

A *data sprint* emerges when the idea of 'sprint' bumps into the realm of data. Like a design sprint, a data sprint condenses an intensive collaboration into a short time-frame; but while a design sprint is typically devoted to developing a specific product, data sprints are often far more exploratory, opening up new avenues for exploration rather than closing them down, sketching out new narratives and seeking fresh insights within the data.[32] Unlike the 'results-orientated' teleology of the commercial design sprint, at the beginning of a data sprint (like any other genuine research project) 'no one knows exactly what could or should be reached,' the end users least of all.[33] In place of the pre-established timetable of the commercial design sprint, the data sprint process can unfold in a less structured, iterative manner, meandering across disciplinary boundaries and defining milestones step by step. Because basic solutions to complex problems must be developed in short order, a data sprint typically involves developing 'quick-and-dirty' solutions by the writing and adapting code or by the design of interfaces and data visualizations that are by their nature unfinished.[34] Rather than aiming at the definitive resolution of controversies (which is often not a viable objective in the academic domain), data sprints may merely aim to map 'the cartography of controversies'.[35] Rather than requiring success from the outset, the data sprint invites participants to try-fail-improve their approaches, tools, and methods iteratively.[36]

2.2 Sprints in the (Digital) Humanities

Although the data sprint process outlined above may seem more amenable to academic work than the design-sprint alternative, many aspects of both run counter to long-established assumptions about how humanistic research is best conducted. While most scholars in the humanities are 'lone wolves', researching and writing their books and articles in splendid isolation, the sprint methodology is highly collaborative. While most single-authored academic publications fall squarely within a disciplinary domain, sprints work best when they bring together individuals with very different knowledge bases and skillsets. While traditional humanistic work puzzles over relatively small quantities of extremely complex evidence, a data sprint typically deals with high volumes of abstract representations of such evidence. While scholarship is traditionally 'slow-cooked', with masterworks evolving

[32] Cæcilie Laursen, 'What Is a Data Sprint? An Inquiry into Data Sprints in Practice in Copenhagen', see https://ethos.itu.dk/2017/02/15/caecilie-laursen/, accessed 20/03/2019.

[33] Tommaso Venturini, Anders Munk, and Axel Meunier, 'Data-Sprinting: A Public Approach to Digital Research', in Celia Lury et al., *Routledge Handbook of Interdisciplinary Research Methods* (London: Routledge 2018): 158–163.

[34] Michele Mauri and Paolo Ciuccarelli, 'Designing Diagrams for Social Issues' (full paper). Proceedings of *DRS2016: Design + Research + Society – Future-Focused Thinking*, 3 (2016): 941–56. See: https://doi.org/10.21606/drs.2016.185.

[35] Tommaso Venturini, 'Piccola introduzione alla cartografia delle controversie', *Etnografia e Ricerca Qualitativa* 3 (2008): 369–94.

[36] Venturini, Munk, and Meunier, 'Data-Sprinting'.

slowly over entire scholarly lifetimes, the sprint format accelerates progress wildly to squeeze it into the five-day timeframe. This series of contrasts helps explain why the sprint methodology is still unfamiliar to most researchers in the humanities. In such circumstances, one of the benefits of a five-day experience of full immersion in a cluster of overlapping sprints is to help break through understandable doubts that such an alien process could ultimately benefit one's research.

Admittedly, several aspects of the sprint methodology have been spontaneously emerging within the domain commonly referred to as the 'digital humanities'. The application of digital technology to humanistic research is intrinsically interdisciplinary, consequently collaborative, and typically involved in processing large quantities of data. Moreover, the digital humanities have also inculcated a culture of 'doing and building'.[37] Many of the objects being built are tools, interfaces, and visualizations designed to interrogate data in new ways; and since such objects rarely spring fully formed from the head of an IT systems developer, they typically require an iterative methodology similar in nature to that of a data sprint, if at a more leisurely pace (often, indeed, more like a marathon), involving alternative phases of building and using.

2.3 Data-Design Sprints in COST Action IS 1310

As mentioned in the introduction (ch. I.1), COST does not fund research, resource creation or systems development per se; and although it does fund many familiar categories of networking activities, design sprints are not among them. Since bringing together experts in very different fields from different countries for an extended period of intensive exchange is precisely what the sprint methodology does, it was nevertheless decided to experiment with it, even though this required shoehorning the first sprint rather uncomfortably into the funding framework for a pair of Working Group meetings and the second into the framework of a Training School.

The two meetings shared a common goal: to bring together humanists and designers to collaborate on the case-study-based design of visual interfaces for exploring structured or unstructured data on the republic of letters. Both events brought twenty-five to thirty people together for a five-day period of intensive, exploratory, interdisciplinary collaboration. In both cases, the crucial admixture of expertise in data interaction design was provided by advanced students and associates of the DensityDesign lab, led by Paolo Ciuccarelli in the Politecnico di Milano. In both cases, Como provided a convenient and attractive location to meet. In

[37] '[M]aking a map (with a GIS system, say) is an entirely different experience. DH-ers insist – again and again – that this process of creation yields insights that are difficult to acquire otherwise. […] Building is, for us, a new kind of hermeneutic – one that is quite a bit more radical than taking the traditional methods of humanistic inquiry and applying them to digital objects', see Stephen Ramsay, 'On Building', in Melissa Terras, Julienne Nyhan, and Edward Vanhoutte, eds., *Defining Digital Humanities* (Farnham: Ashgate, 2013), 243–6.

both cases, the assembled company was broken down into smaller Working Groups – seven in the first case, five in the second – who focused their attention on a series of case studies. In the first instance, these case studies were chosen based on applications from Action members and affiliates; in the second they emerged from collective thinking about some of the chief desiderata in the field. These meetings were structurally unusual in two important respects. First, they were residential: all participants spent the full week living in Como, and the 'full immersion' experience proved very effective in generating the kind of focus and commitment necessary for moving from complete unfamiliarity to full engagement with the processes. Second both events staged multiple sprints in parallel within the same building. This provided the opportunity to raise the standard and quicken the pace of each individual group by closing each day's work with a brief flash presentation of the day's achievement toward the goals established on the previous day, which also made explicit what difficulties and perplexities had been encountered. This allowed each group to learn from one another, to benefit both from the presentations of work being conducted in the other groups and from feedback, interventions, and contributions to their own presentations by participants from other groups. The last day featured a longer final presentation of each project, summing up the entire sprint process. This presentation replaced the final user test that characterizes a design sprint with an emphasis on process documentation.

In both events, the design-sprint and data-sprint methodologies were deliberately merged. Like a data sprint, most group work began with data rather than a semi-developed product idea; but, like a design sprint, the process remained orientated towards the development of a prototype. Almost all of the groups opted to mock up or prototype interactive tools: few chose the (only apparently) easier task of developing static or animated visualizations. The humanist scholar who contributed the core data set was typically nominated as the group leader, who also typically contributed the first intuition of the goal of the exercise. Other scholars were distributed among the Working Groups according to their interest in the project objectives, and designers were distributed according to expertise.

The first data-design sprint (framed as a joint meeting of Working Groups 3 and 6) was coordinated by Paolo Ciuccarelli (WG6 leader) and Charles van den Heuvel (WG3 leader). Meeting in Como on 4–8 April 2016, a group of twenty-five people – almost half of them scholars in the humanities and the other half information design researchers, data strategists and developers – were divided into seven Working Groups, each with a separate objective to pursue. Three of the seven projects dealt with the epistolary dimension of the republic of letters, and all three are discussed elsewhere in this section: 'Seeing Echoes' explored means of visualizing text reuse in correspondence (ch. IV.6); 'Visualizing EMLO' scoped out a variety of means of visualizing the temporal, spatial, topical, and linguistic dimensions of correspondence metadata (ch. IV.3); and 'Visualizing Epistolaries' developed means of visualizing the chronological development of collections of letters printed in the early modern period itself (ch. IV.3). Two of the other groups explored

means of visualizing relevant prosopographical data: 'International Lives and National Biographies' investigated a visual browser for data on foreigners and foreign travel from the *Oxford Dictionary of National Biography*; and 'VIA: Virtual Itineraries of Academics' designed an interface for exploring early modern academic travel culture (see ch. IV.4). The final pair focused on bibliographical data: 'Biblio-philus' piloted means of exploring 'the lives of the entire libraries' (when and where books were published, and when, where, and by whom they were acquired); while the final group explored means of visualizing the distribution of copies of a specific book: the first edition of Newton's ground-breaking *Principia mathematica* (1687).[38]

The second data-design sprint (framed as a Visualization Training School) took place at the Chiostrino Artificio in Como on 10–14 July 2017.[39] The organizers identified five different areas where design expertise was urgently needed in helping to create new means of exploring structured and unstructured correspondence data. All were highly productive, and three are documented elsewhere in this section. The first three groups developed partially functional proof-of-concept implementations of tools for exploring 'intersecting correspondences',[40] 'correspondences over itineraries' (see ch. IV.2), and 'visualizations with memory'[41] (see ch. IV.3). The last two groups undertook more conceptual exploration of higher-level interfaces for a 'digital critical editions platform' and an innovative 'virtual research environment' (see ch. IV.7).[42]

2.4 Results and Reflections

Although digital technologies can now greatly assist collaborators working at a distance from one another, the value of intensive, sustained, and direct face-to-face collaboration in this instance was inestimable. This was partly because such a wide variety of disciplinary skillsets had to be applied to each task; but the main reason is that the community had to be taken through a process that was completely unfamiliar to one half of the group and also unusual for the other half. The quick tempo of the week-long sprint, the immersive experience of a residential meeting, and the stimulus provided by all the other groups working in parallel was highly effective in getting all participants to commit fully to the experiment. Some of

[38] A detailed programme of the first meeting, hyperlinked to more detailed reports, can be found at: http://www.republicofletters.net/wp-content/uploads/2017/02/Como-Notes-COST-Action-IS1310-Reassembling-the-Republic-of-Letters.pdf, accessed 20/03/2019.

[39] The detailed programme of the second meeting (formally a Training School) can be found at: http://www.republicofletters.net/wp-content/uploads/2017/12/Como-Training-School-2017.pdf, accessed 20/03/2019.

[40] See an interactive version at: https://iosonosempreio.shortcm.li/intersecting-correspondences, accessed 20/03/2019.

[41] See an interactive version at: https://iosonosempreio.shortcm.li/visualisations-with-memories, accessed 20/03/2019.

[42] The individual projects of the two design sprints are documented in detail: http://iosonosempreio.shortcm.li/como-sprint, accessed 20/03/2019.

these benefits are already well documented in the literature: the particular format of the data sprint prioritizes the processes of learning and exchange rather than just giving a paramount importance to the final outcomes,[43] although the design sprint can create similar effects as well.[44] In this case, however, these benefits were heightened by the fact that the process was entirely new to the scholarly participants and because of the added stimulus of running multiple working groups in parallel.

Another important factor in this regard is that the hybrid 'data-design sprint' methodology explicitly valued not only the artefacts produced at the end of the process but also the experience of sharing the building process itself and, by doing so, learning to understand and to value the languages, assumptions, skills, perspectives, and techniques of the range of specialists from alien domains who formed the team. It is important to emphasize that the benefits were reciprocal if not completely symmetrical. On the one hand, this involved the demonstration to the scholars of the practical knowledge that belongs to information designers. Most obvious in this volume was the manner in which scholars wishing to investigate the use of visualization, were assisted by data visualization experts and gained an insight into the process of visual design and data analysis. Yet the opposite effect is also noteworthy: designers, more accustomed to working with commercial, social scientific, or journalistic data, also had the opportunity to observe and work next to a new set of new 'smart users' with different interests and attitudes, assumptions and objectives, and a rather different relationship to rather different sorts of data. Grasping these disciplinary cultures is a necessary precondition to assisting them in visually conveying their thoughts and advancing their interpretations of historical data, whether in the form of static visualizations, interactive tools, methods, or even plain data sets. More reciprocal exchanges of this kind will be necessary to create communities capable of establishing practices and strategies for designing better tools for the digital humanities.

One of the insights to emerge from the design sprints is that the discussions between scholars, systems developers, and designers should normally begin at the very outset of the research process, when formulating a project and drawing up a research proposal. This is necessary not only for budgeting reasons but also because research results, data models, analytical tools, and modes of visualization are all mutually interdependent: if any of these components is left out of the initial planning and conceptualization process, problems and limitations can emerge at the later stages of the project and cannot be so readily overcome.

This insight has potentially profound implications for creating optimal conditions for innovative work in the digital humanities in the future. Although the appetite among participants for a rematch remains strong, it is unlikely to be satisfied.

---

[43] Laursen, 'What Is a Data Sprint?'.

[44] Imola Unger, 'The Biggest Benefit of a Design Sprint Is Not What You Think It Is' (2017) see https://sprintstories.com/the-biggest-benefit-of-a-design-sprint-is-not-what-you-think-it-is-be807b5e6f71, accessed 20/03/2019.

Replicating the experiment funded by this COST Action will be a rare occurrence: bringing five groups of five people to a five-day residential meeting from many different countries is an expensive enterprise, even when only paying for travel, accommodation, meals, and the venue. Another problem is the fact that properly trained designers are currently a very rare commodity in the digital humanities field, both because very few are trained in working with scholars and their data, and because few institutions have grasped the importance of what they bring to the mix and funded them accordingly. It will be very interesting to see how readily this defect is repaired in the future. Whether, as Bruno Latour famously remarked, every digital humanities project needs an Italian designer, it is probably fair to say that a well-equipped DH centre will have plenty of design expertise close at hand, in order to allow something like the data-design sprint methodology to become firmly institutionalized, and the pace slackened but sustained over much longer distances.

IV.2 Geographies of the Republic of Letters

Ian Gregory, Alexandre Tessier, Vladimír Urbánek, and Ruth Whelan with Claire Grover, Bruno Martins, Yves Moreau, Patricia Murrieta-Flores, and Catherine Porter

1 Introduction

Ian Gregory

Letters are inherently geographical. According to the definition adopted in the second section of this volume (chs. II.1 and II.7), letters are written documents intended to convey a message from one place to another. Without spatial distance between sender and recipient, the problem which letters solve does not even exist. Moreover, the geographies mapped out by letters operate at several other levels too. As well as the places to and from which letters are sent, there are the places discussed within the letters themselves, the itineraries of the correspondents as they move from place to place, and indeed also the routes taken by the letters in their travel from place of sending to place of receipt.

The most common, and most accessible, way that geography is expressed in letters and other texts is through place names. These may be found within metadata, where they usually refer to the place of sending or receipt, or within the texts of the letters, from which they can be extracted using Named Entity Recognition (NER).[1] Place names in metadata tell us about the writer's correspondence net-

[1] Miguel Won, Patricia Murrieta-Flores, and Bruno Martins, 'Democratic NER: Evaluating Name Entity Recognition and Geocoding of Place-names in Historical Documents', in *Frontiers in Digital*

work, while those in full texts tell us how the writers perceived the places that interested them. Once identified, place names in metadata or full text can be allocated to coordinates using a gazetteer, effectively a database table that, at a minimum, provides a coordinate such as a latitude and longitude for each place name.[2] For place names associated with large settlements, a modern gazetteer such as *Geonames* is likely to be adequate; but where there are issues such as change over time, multilingual toponyms, or very local places such as streets or buildings within a particular city, then either a specialized gazetteer needs to be used,[3] or the researcher will have to create one for a particular place or corpus. The configuration of an historical gazetteer suitable for collaborative work on the republic of letters is outlined in more detail in chapter II.2 above. Once a coordinate-based location is available, the place name can be mapped using a geographical information system (GIS) or other form of digital mapping technology.[4]

The process of moving from place names in a corpus to mappable locations is known as georeferencing. In many ways georeferencing can be seen as a technical challenge although the difficulties in identifying where early modern place names refer to can be a significant scholarly undertaking in its own right. Ultimately, the process would be futile unless it led to new forms of analysis and research. Georeferencing opens up a number of new analytic possibilities.[5] The most obvious is that once a corpus has been georeferenced it can be visualized in cartographic form. Contrary to what is often thought, maps rarely answer research questions: instead they describe the locations associated with the corpus and challenge the researcher to explain the patterns shown. In this way, the maps actually create questions for the researcher, such as why something is happening in one place and

Humanities (2018), see https://doi.org/10.3389/fdigh.2018.00002; Rui Santos, Patricia Murrieta-Flores, Pável Calado, and Bruno Martins, 'Toponym Matching through Deep Neural Networks', *International Journal of Geographical Information Science* 32:2 (2018), 324–48, see https://doi.org/10.1080/13658816.2017.1390119.

[2] See: Linda L. Hill, *Georeferencing: Geographic Associations of Information* (Cambridge, MA: MIT Press, 2006); and Humphrey R. Southall, Ruth Mostern, and Merrick L. Berman, 'On Historical Gazetteers', *International Journal of Humanities and Arts Computing* 5:2 (2011): 127–45, see https://doi.org/10.3366/ijhac.2011.0028. See also the essays in Merrick L. Berman, Ruth Mostern, and Humphrey R. Southall, *Placing Names: Enriching and Integrating Gazetteers* (Bloomington: Indiana University Press, 2016).

[3] A good example of a gazetteer created for a specific time and place is the gazetteer associated with the *Map of Early Modern London*, see https://mapoflondon.uvic.ca/gazetteer_about.htm, accessed 20/03/2019. The *Pleiades* gazetteer of the Ancient World provides another example, see: Rainer Simon, Leif Isaksen, Elton Barker, and Pau de Soto Canamares, 'The *Pleiades* Gazetteer and the *Pelagios* Project', in Berman, Mostern, and Southall, eds., *Placing Names*, 97–109; see http://pelagios-project.blogspot.co.uk/p/about-pelagios.html, accessed 20/03/2019.

[4] Ian N. Gregory, Christopher Donaldson, Andrew Hardie, and Paul Rayson, 'Modelling Space and Time in Historical Texts', in Julia Flanders and Fotis Jannidis, eds., *Data Modelling in the Digital Humanities* (London: Routledge, 2019): 133–49.

[5] Ian N. Gregory, Karen Kemp, and Ruth Mostern, 'Geographical Information and Historical Research: Current Progress and Future Directions', *History and Computing* 13:1 (2003): 7–21, see https://doi.org/10.3366/hac.2001.13.1.7.

why it is not happening in another.[6] Maps are also quite a limited form of visualization: they are particularly poor at representing change over time or multiple phenomena simultaneously. To allow researchers to explore these forms of complexity, more innovative forms of visualization are required that will typically be interactive, allowing the researcher to explore more sophisticated spatial and spatio-temporal questions than simple mapping allows. Finally, having locations for places allows the calculation of distances between the places and estimations of the routes between them.

This chapter explores all three of these geographical levels in a series of brief case studies and project proposals. In the first section, Ruth Whelan explores two contemporaneous Huguenot correspondences, and reflects on how the geographies of the places to and from which the letters were sent and the places discussed within the letters can inform our thinking about the spatial dimension of the republic of letters. In the second, Vladimír Urbánek addresses the challenge of highly itinerant correspondents, describing how visualizations might be developed to master the complex patterns mapped out by intellectuals sending and receiving letters as they themselves move from city to city. Finally, Alexandre Tessier proposes a strategy for handling whole systems of postal exchange, which could allow researchers to explore the geographical, chronological, and also monetary dimension of the journey that a letter would have taken to get from its sender to its recipient. The chapter thus demonstrates that we can use digital technologies both to deepen our understanding of the history associated with a corpus, and to develop the technologies to allow us to explore and understand the multiple geographies within the corpus in new ways.[7]

[6] Alan R. H. Baker, *Geography and History: Bridging the Divide* (Cambridge: Cambridge University Press, 2003).

[7] Ian N. Gregory and Alistair Geddes A., 'From Historical GIS to Spatial Humanities: Deepening Scholarship and Broadening Technology', in Ian N. Gregory and Alistair Geddes, eds., *Toward Spatial Humanities: Historical GIS and Spatial History* (Bloomington: Indiana University Press, 2014), ix–xix.

2 Mapping the Republic of Letters: A Spatial Analysis of the Correspondences of the Huguenots Élie Bouhéreau (1643–1719) and Jacob Spon (1647–1685)[8]

Ruth Whelan

The 'Protestant International' is a catchphrase often used to describe the interwoven networks – whether commercial, political, diplomatic, military, clientelist, familial, religious, or intellectual – created in exile throughout Europe by Huguenot refugees who fled France at the time of the Revocation of the Edict of Nantes (1685), when Protestantism was proscribed under the regime of Louis XIV. Both Élie Bouhéreau (in 1686) and Jacob Spon (in 1685) chose exile to avoid forced abjuration, the one going to England (and later Ireland), the other to Switzerland. Both men were scholars, who were regarded by their peers as eminent members of that international yet conceptual early modern space, the republic of letters. This section explores how digital humanities approaches can be used to map their correspondence networks and thereby determine how international the networks created by Huguenots in the republic of letters really were before 1685.

The choice of subjects was slightly serendipitous, as it was determined by finding scholars willing to share their research and data; but a perfect match emerged. Élie Bouhéreau and Jacob Spon were almost exact contemporaries and they came from similar backgrounds. Bouhéreau's father was a pastor from a family established in La Rochelle from the time of the Reformation; his mother was from a merchant family from the Île de Ré and La Rochelle; in 1668 Bouhéreau himself married into one of the most important merchant families in his home town. Spon was the son of a medical doctor from a Protestant family of merchants, originally from Ulm and established in Lyons from 1551. Both men were provincial French Protestant *savants* (learned men) and medical doctors; both consciously inscribed themselves and their work in the conceptual space of the republic of letters. Significantly, they both lived in gateway towns that were, in their diverse ways, important networking hubs and centres of exchange. La Rochelle was one of the most important ports on the west coast and was invested in the triangular trading system between France, West Africa, and the Caribbean or American colonies. Lyons was the gateway to Italy; its transport network via road or river placed it at the centre of a commercial hub, with links to northern Europe, the Mediterranean, the Levant and the Far East. Both men had extensive correspondences in which they communicated, explored, and exchanged learning. And, in both cases, their con-

[8] This project was only possible because Yves Moreau (Université de Lyon Jean Moulin, UMR 5190 LARHRA) generously provided Working Group 1 with a digital copy of his transcriptions of Spon's letters and worked together with us in the session at Lancaster University as we identified, extracted, and analysed the spatial components from the two correspondences.

temporaries recognized each of them as exemplifying, in their different yet similar ways, the values of the republic of letters.[9]

Although early modern doctors rarely travelled outside France, both Bouhéreau and Spon had travelled to Italy and had resided for a time in Paris, participating there in the learned academies and gatherings that so exemplified the collaborative learning of the republic of letters. However, Spon also travelled more extensively for his scholarly pursuits to Germany, Holland, Greece, Switzerland, and the Levant. They were also both authors or aspiring to be so; Spon published the results of his research and travel in his *Recherche des antiquités et curiosités de la ville de Lyon* (1673); and five years later, his *Voyage d'Italie, de Dalmatie, de Grèce et du Levant* (1678, 3 vols.), and the *Miscellanea eruditae antiquitatis* (1685), among other publications. In 1669, Bouhéreau embarked on a translation of Origen's *Contra Celsum (Traité d'Origène contre Celse* (1700) at the instigation of Valentin Conrart, the secretary of the recently founded Académie Française, who also involved his protégé in the revised translation into French of the Psalms used in Reformed worship. Although Bouhéreau's translation of Origen was published some thirty years after he began work on it (his endeavours having been interrupted by family life and the political crisis of the Revocation), he was known from the outset in scholarly and Huguenot circles as its author. In other words, both men had the social capital typical of the educated elite. They had studied the humanities in one of the elite Protestant academies (Bouhéreau in Saumur, Spon in Geneva); they were men of broad education with a cosmopolitan urbanity; they had the income and social standing of doctors of medicine, which they could supplement with their own or their families' commercial interests; they could also use those mercantile and other networks for the transmission of their correspondence.

The extant corpora of letters available for comparison are uneven but cover the years 1661 (Bouhéreau) and 1667 (Spon) to 1685, the year of Spon's death; although letters to and from Bouhéreau are extant after that date, they are mostly concerned with his activities as secretary to Thomas Coxe (1689–92) and Henri de Massue de Ruvigny (1693–7), with whom he travelled extensively in continental Europe. In Bouhéreau's case, metadata for 1,353 letters has been collected by Ruth Whelan, of which 1,112 are held in Dublin (Marsh's Library), 234 in Paris (Bibilothèque du Protestantisme Français), and fifteen in the Archives Départementales de Charente-Maritime. Some 400 letters have been fully transcribed; for comparative purposes 110 of these transcriptions were selected, those sent to Bouhéreau by two of his friends from their student days in Saumur: Jacques Richier de Cerisy, from Normandy, and the marquis Turon de Beyrie from Béarn. Only three of the

[9] Yves Moreau, 'Entre la France, Genève et l'Italie: le réseau de correspondants de Jacob Spon (1647–1685)', in Philippe Martin, ed., *La Correspondance: le mythe de l'individu dévoilé* (Louvain: Presses universitaires, 2014), 115–26; Ruth Whelan, 'West Coast Connections: The Correspondence Network of Élie Bouhéreau of La Rochelle', in Vivienne Larminie, ed., *Huguenot Networks, 1560–1780: The Interactions and Impact of a Protestant Minority in Europe* (New York and Abingdon: Routledge, 2018), 155–71.

extant letters are from Bouhéreau, the others being addressed to him from some 145 correspondents identified to date. As for Spon, 425 letters are extant mostly in archives in Paris (Bibliothèque Nationale de France) and Lyons (Bibliothèque Municipale) and, to a lesser extent, in other European libraries. The metadata has been collected and the letters fully transcribed by Yves Moreau, 135 of them written by Spon (thirteen to correspondents from whom no replies are extant), and 290 addressed to him from eighty-five correspondents.

The two men certainly knew of each other and they had mutual friends, notably, in La Rochelle, the apothecary Élie Seignette (who corresponded with Spon) and the doctor Jean Seignette (who corresponded with Bouhéreau), and who mentioned Spon in a letter to Bouhéreau in February 1685. They had probably met in Paris where they both resided between 1662 and 1664, and where contacts and friendships were established between Protestants attending the Reformed Temple at Charenton just outside the city. They may have met up again when Bouhéreau passed through Lyons on his way to Italy in 1667, and later in La Rochelle in 1683, when Spon travelled in the south and south-west of France with the apothecary Henri Moze, studying mineral waters, Roman antiquities, and other 'curiosities', one of which may have been Bouhéreau's extensive personal library. However, while both men had much in common: their education, profession, religious affiliation, and above all their 'anticomania', that is, their preoccupation with Graeco-Roman Antiquity, their scholarly focus was quite different. Although Bouhéreau was interested in the material culture of Antiquity, and he bought books on the subject after his Italian trip, he was essentially a philologist, textual critic, translator, and bibliophile. As for Spon, he dismissed those studying Graeco-Roman Antiquity just from texts, and not from travelling to the sites where they could conduct their own investigation of its material culture, as *demi-savants* (middling-learned); he was a reader but not a collector of books, which he readily admitted was not his *marotte* (hobby horse). Their similarities and their differences, together with their geographical locations to the east and west of France make them ideal subjects for comparison.

At first sight the major contribution of digital mapping to the study of these two correspondences is its impact: it makes visible and instantly readable what the specialists already know; thus, it is a valuable tool for the communication of research findings to the wider public. Figures 1 and 2, generated from the spatial metadata, serve to illustrate this point.

Élie Bouhéreau resided in La Rochelle, studied in Saumur, and travelled to Paris and Italy: it is hardly surprising then that figure 1, with simple proportional symbols, demonstrates that the largest number of letters addressed to him originate from the west or south-west of France (44 per cent); that Saumur and Normandy, where he studied and where two of his college friends resided, account for 25 per cent of the letters sent to him; or that 26 per cent of the letters were dispatched from Paris, that centre of urbanity and learning, not to mention power, with which all provincial literati desired to have contact.

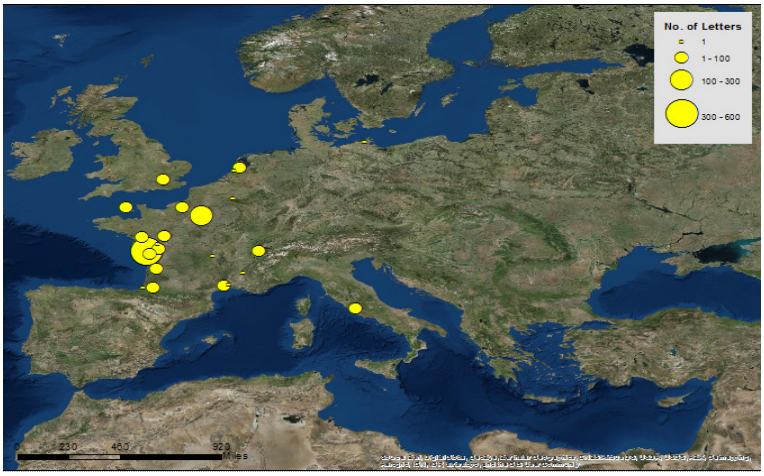


Figure 1: Letters sent to Élie Bouhéreau

Moreover, the European dimension of the correspondence is attested by the expedition of letters to Bouhéreau from London, Holland, Geneva, and Rome. A radial map would show Bouhéreau at the centre of a network that extended well beyond the confines of his provincial origins and place of residence, making and make his cosmopolitism and his standing in the republic of letters instantly visible.

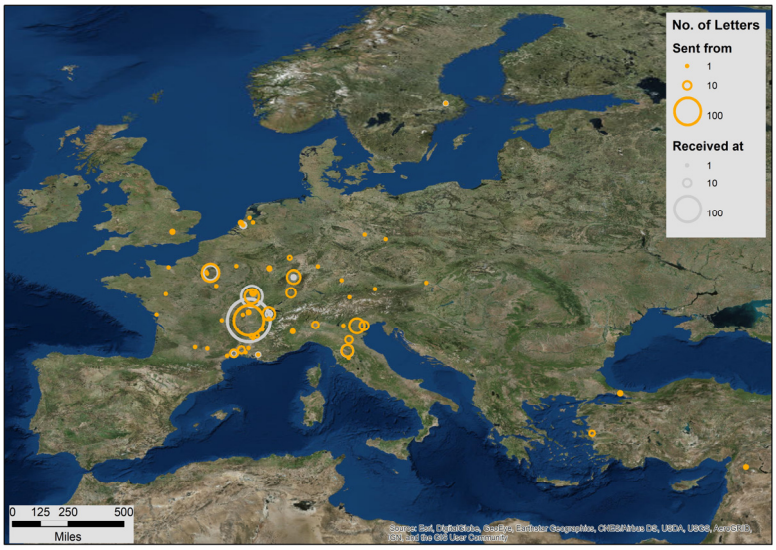


Figure 2: Letters sent to Jacob Spon

Similar conclusions are demonstrated when we consider a spatial representation of Spon's extant correspondence.

Spon resided in Lyons, studied in Geneva, but travelled more widely than Bouhéreau; as a published author, he also had a more prominent standing in the republic of letters, all of which, as might be expected, is illustrated by the mapping of his correspondence network shown in figure. The greatest number of letters were exchanged between Lyons, Dijon, Paris, and Padua; but a radial map would reveal that Spon was at the centre of a network that included England (London), Holland (The Hague, Rotterdam, Utrecht), Germany (Dresden, Leipzig), Switzerland (Basle, Geneva), Italy (Bologna, Florence, Milan, Siena, Turin, Venice, Verona), extending as far north as Uppsala and as far south-east as Constantinople, Smyrna, and Alep. Although fewer letters from Spon's correspondence survive, his network can be shown to be more extensive, in as much as he had contacts in more countries than Bouhéreau, and more international because it was not confined to continental Europe.

However, digital maps turn into an investigative tool that may promote new insights and challenge cherished assumptions when they are used comparatively. At the session in Lancaster, when the two maps above were projected on screen, Yves Moreau observed that he had not realized quite to what extent Spon's correspondence was oriented essentially towards the Mediterranean. And, by comparison with Spon, Bouhéreau's correspondence is clearly revealed to be a network rooted essentially (with some exceptions) in the west, south, and north-west of France. Moreover, radial maps with place names inserted confirm the existence of a spatial hierarchy within the geographical span of the two correspondences: the great urban centres, notably Paris but also London to a much lesser extent, were most important to both men, as were the towns and cities with established universities and active printing houses (in France, Holland, and Italy). These were the places where the most prestigious *savants* resided, where learned exchange occurred at its most intense in formal and informal academies, salons, and gatherings, and where opportunities for self-advancement and promotion of learning might be grasped.

Finally, a computational technique called Named Entity Recognition (NER) was used to identify the place names in all three corpora (letters to and from Spon; letters from Turon de Beyrie and Richier de Cérisy to Bouhéreau); georeferencing assigned geographic coordinates to the place names identified in the corpus through NER, and the results were mapped to produce figure 3. These results produced deeper, even original, insight into the world view embedded in the letters that could not have been reached by other means.[10]

[10] Named Entity Recognition (NER) is a subtask of Natural Language Processing (NLP). It aims to identify, within texts, real-world entities such as names, organizations, and locations. This is carried out through the use of linguistically based techniques taking into account grammar, and of statistical models using machine learning. NER labels sequences of words, for instance, identifying and labelling words that are proper names. A specific form of NER is geoparsing, the process of identifying

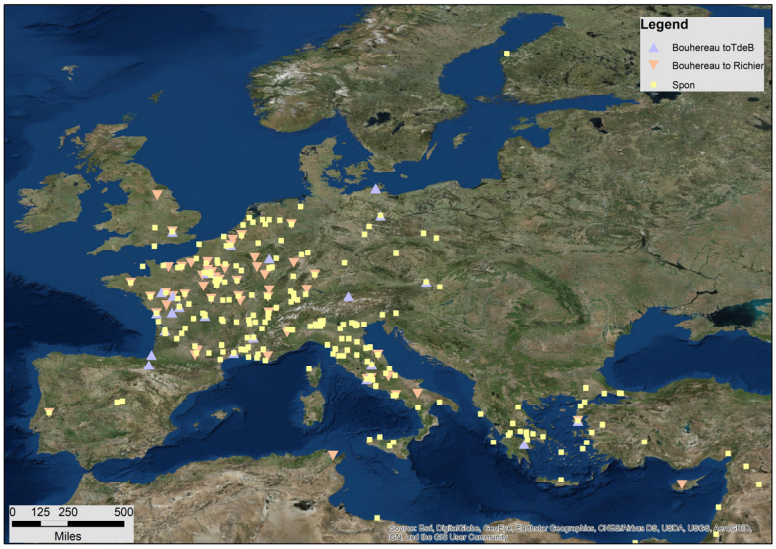


Figure 3: Place names mentioned in the Bouhéreau and Spon correspondences

If Spon's travels to the Levant are omitted, France and Italy emerge as the principal places of importance to both correspondents; Holland, Switzerland, and Germany are also important to one or other of them and sometimes to both; England only registers on their map as London and occasionally Oxford. Although other maps were created from the metadata, text mining, and geographical analysis of these correspondences, they cannot be included for reasons of space. Nonetheless, the three examples above provide a basis for questioning the traditional understanding of the space of the republic of letters by comparing it with the actual geographies of early modern learned men that are defined by the places mentioned in the letters themselves.

The republic of letters is defined as a conceptual space that is transnational, trans-confessional and egalitarian; its objectives are perceived as the pursuit of truth, the generation and communication of knowledge with objectivity and without borders, which involves setting aside what divides and building networks of exchange, dissemination, collaboration, and reciprocal critique; with urbanity and mutual respect regarded as the norm. Élie Bouhéreau and Jacob Spon saw them-

place names in free text and then assigning geographic identifiers to these, such as coordinates. Geoparsing goes beyond simple geocoding in its capacity to disambiguate place names. For this case study, the *Edinburgh Geoparser* was used, an advanced system capable of identifying place names and assigning coordinates to these: see https://www.ltg.ed.ac.uk/software/geoparser/, accessed 20/03/2019. This work was conducted in collaboration with Ian Gregory and Patricia Murrieta-Flores (Lancaster), Claire Grover (Edinburgh), Yves Moreau (Lyons), Catherine Porter (Queen's University, Belfast), and Alex Butterworth (Sussex).

selves, and were perceived by their peers, as eminent members of this pan-European community without borders that was the imagined space of the republic of letters. However, their correspondence reveals that their transnational experience and contacts were not unaffected by national and confessional considerations.

France and home are more significant than 'abroad', Paris is more important than London; their network is confined for the most part to Latinate or Protestant Europe; they are interested in Italy because classical culture functioned as a bond and currency of exchange in the republic of letters; but that interest is not matched by a concomitant interest in Spain, still inextricably connected with the Inquisition in the perception of Protestant Europeans. At the Revocation, both men chose exile, emigrating to destinations that were politically Protestant – Bouhéreau to England and Spon to Switzerland – where they already had established contacts, which illustrates just how devastating the impact of confessional conflict could be on the supposedly neutral space of the republic of letters. Thus, spatial analysis, digital mapping, and comparative study of these two correspondences show the Europe and the republic of letters experienced by two educated French Protestants before 1685, and highlights just how regionally defined (west as opposed to east), limited (France and Italy), urban (cities and university towns), and confessionally impacted that experience really was. In a word, the imagined geography of the republic of letters does not and perhaps cannot map onto the sociopolitical reality of the learned in early modern Europe.

Of course, such conclusions remain tentative, given the limited number of letters extant and the fact that epistolary exchange is but one manifestation of the contact, communication, and collaboration between the learned of the day. However, answers to the questions I put to participants in the Warsaw conference were telling. Is the gap between the ideal and the real that emerges from mapping the Bouhéreau and Spon corpora to be found in other correspondences studied by COST Action IS 1310? Are the actual spaces of the republic of letters different for Latinate and Germanic languages, for north and south, east and west, Protestant and Catholic Europe? Apparently, that can be the case, according to Kristi Viiding (University of Tartu, Estonia). According to her, northern European *savants*, for example in the Baltic, perceived the republic of letters as a pan-European phenomenon, but their experience was more regional and defined by their confession. On the one hand, northern Baltic (Estonian, Livonian, and partly Curonian) Protestants principally communicated with the northern German and, in the seventeenth century, Scandinavian Protestant states, but also with Leiden and Padua (famous for medical studies). On the other, southern Baltic (Lithuanian and Curonian) Catholics were fascinated mostly by Polish and southern German Catholic Universities, but also by Vienna, Paris, and Italy (to which aristocratic families travelled).

That all goes to show that digital mapping can point research in new directions, in this case the need to map the *fragmented reality* of the republic of letters as a corrective to the imagined, nostalgic projection of it. Because individuals and

groups who believed and felt they were a part of a pan-European phenomenon experienced it (and perhaps only wanted to experience it) as a subset of that virtual world. Perhaps, in that respect, they were more like us than we would care to admit.

3 Individual Itineraries

Vladimír Urbánek[11]

3.1 The Problem: Highly Mobile Correspondents

Simple static visualizations are indispensable for getting a quick, overall impression of relatively small and simple data sets; but for exploring larger and more complicated data, more innovative visualizations must be developed. One key example of this imperative is posed by the difficulty of mastering the correspondence of highly itinerant individuals. This section relates an experimental approach to addressing these challenges, developed by one group during one of the five-day design sprints funded by the COST Action in Como.

The career of the Czech educational reformer and pansophist Jan Amos Comenius (1592–1670) provides a clear example of this challenge and an excellent opportunity to begin developing solutions to it. Comenius's surviving correspondence of about 569 letters is not huge, although it is widely scattered both among acknowledged Western centres of scholarship such as Amsterdam, London, and Paris, and in cities and towns such as Leszno, Gdańsk, Elbing, and Brzeg, marginal to the republic of letters, but with lively intellectual traditions, many of them powerfully affected, as Comenius himself was, by the Thirty Years' War and the First Northern War. When visualized in a single static map, however, even this modest data set creates a confusing image (fig. 4). The reason is that Comenius was highly itinerant: throughout his long and tumultuous life, he was repeatedly displaced by the endemic warfare of the period, and moved throughout the same huge swathe of south-eastern, central, northern, and north-western Europe mapped out by his letters, including longer periods of residency in the Bohemian Lands where he was born, the German territories of the Holy Roman Empire where he completed his education, and in Greater Poland, England, Royal Prussia, Hungary, and the Netherlands where he sought refuge. Such is the extreme complexity of his itinerary (indicated in the cartographic portion of fig. 8 below) that superimposing it on the map of his correspondence would render both data sets completely unintelligible.

[11] The work documented in this section was undertaken at a design sprint held in Como, Italy, 10–14 July 2017. The team members were: Vladimír Urbánek (lead); Roberto Evangelista; Magnus Ulrich Ferber; Beatrice Gobbo; and Alex Piacentini. The author gratefully acknowledges their help and especially the technical contribution of Alex Piacentini. Beatrice Gobbo is the author of Figures 5–7, Alex Piacentini of figures 8–9 and of the prototype itself. Beatrice Gobbo and Tommaso Elli have additionally designed figures 5–9. The author also thanks Iva Lelková for providing figure 4.

Figure 4: The geographical distribution of Comenius's sent and received correspondence. The visualization was made by Iva Lelková using *Palladio*

3.2 Parameters of a Solution

This analysis of the problem set the parameters of the solution sought within the remainder of the design sprint. The general objective was to develop a means of visualizing the data defining both the itinerary and the correspondence of Comenius in a manner that allowed the relationship between them to be explored and more perfectly understood. Are changes in geographical scope and prosopographical composition closely related to the person's physical movements or independent of them? More specifically, the aim was to develop a means of discovering whether one drives the other. Does Comenius's relocation from one place to another create new contacts which expand his correspondence network and endure after he has moved on to another place? Or is it the contacts established through correspondence which dictate the pattern of his movements? Or alternatively, does one pattern predominate in some cases, and the other in others? A second objective was to use the case of Comenius to begin developing a tool potentially reusable for other itinerant scholars and their correspondence networks, both in periods of relative peace (such as Erasmus or Leibniz) and in periods of war (as in the case of Comenius's close associate John Dury).

3.3 Matrix-Map Model

The basic problem here is that we have two different data sets – letters moving to and from an individual who is himself frequently on the move – both of which

have a spatial and a temporal dimension. In general terms, the best way to render these four dimensions of the data comprehensible is not to superimpose them on top of one another but to develop two parallel means of visualizing the two data sets, one emphasizing the spatial dimension, the other the temporal one.

With this solution in mind, a 'matrix-map model' was adopted as a means of visualizing the relevant data. As the name implies, this model consists of two basic elements. The more familiar element, on the right, is a map that combines data of two different kinds: the lines represent the journeys undertaken by the principal correspondent, and the circles represent destinations of letters sent by the principal correspondent (in yellow) and origins of letters received by the principal correspondent (in orange). The matrix on the left represents the itinerary of the principal correspondent with two axes: the horizontal axis represents time, and the vertical axis represents a series of places. The places where the principal correspondent actually resided or which he visited during his journeys are not represented by circles (unless he corresponded with somebody in the same city) but with a line.

Both matrix and map are intended to be interactive. The first matrix-map visualization (fig. 5) represents the complete itinerary and correspondence of a figure like Comenius. In the second visualization (fig. 6), the user has chosen to study only a portion of this complete story: the itinerary of the selected portion is represented by the bold line, the later portion by thinner lines, and the correspondence data on the map pertains only to the years under consideration. In the third visualization (fig. 7), the user has chosen only one particular city where a figure like Comenius resided for some time: in this case, only correspondence to or from that city is highlighted on the matrix and the map.

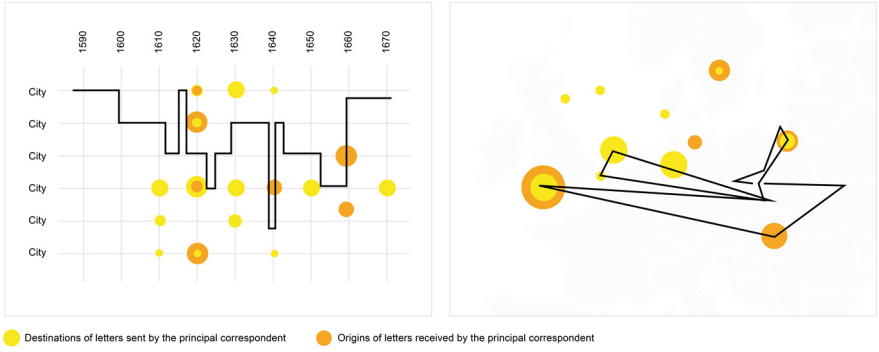


Figure 5: Complete correspondence and itinerary

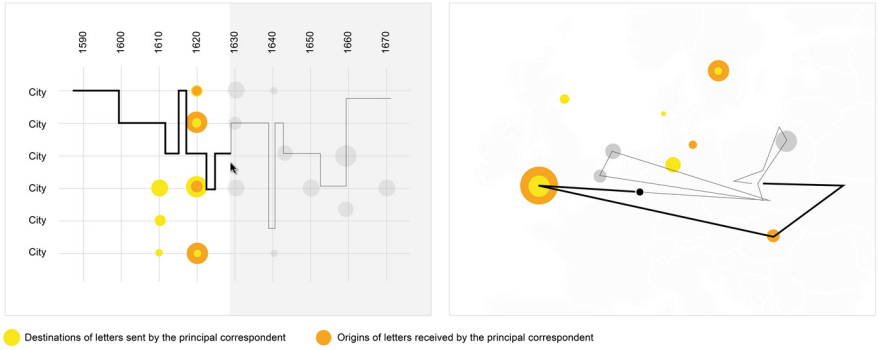


Figure 6: Correspondence and itinerary from defined period

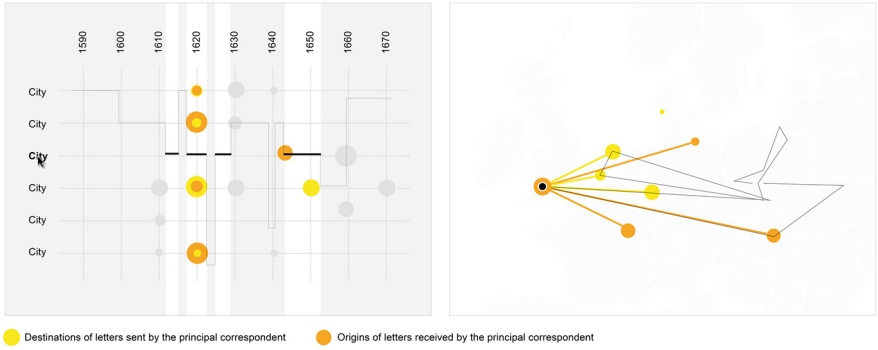


Figure 7: Letters to and from principal correspondent in one city

3.4 Data Preparation

Having defined the problem and formulated a general strategy for addressing it, the next task of the design sprint was to use available data on Comenius to test the suitability of this matrix-map model for exploring the effect of a high degree of mobility on the evolution of a correspondence network. Metadata of Comenius's correspondence, which includes c. 569 letters sent and received, were derived from the Comenius catalogue within *Early Modern Letters Online* (EMLO). Data for Comenius's physical movements were provided by the Department of Comenius Studies and Early Modern Intellectual History at the Institute of Philosophy, Czech Academy of Sciences, in Prague, which had also assembled the correspondence catalogue on EMLO.

Several stages of preparation were needed to adapt the pre-existing data to this new purpose. In order to correct and refine previously incomplete, ambiguous, and unclear dating and placing of the letters, for instance, the latest findings were introduced from ongoing work in Prague on the critical edition of Comenius's correspondence. Since Comenius's location history seemed to lack the structural consistency needed for the visualization it was necessary to simplify and standardize the format of the data and to distinguish between the 'long' periods of residency and 'short' journeys, which would be represented in the graph and map by thick and thin lines respectively. The length of Comenius's itinerary and the breadth of his correspondence created another problem: the over 100 place names on the vertical axis of the matrix posed a considerable challenge not fully resolved even by the end of the project. The group also faced problems related especially to uncertain dates of some letters and unclear chronology of some of Comenius's movements. This was a significant challenge, and in some cases the visual approach combined with detailed comparison of metadata helped the members of the team formulate plausible hypotheses, for instance regarding the chronological order of cities visited. The data of Comenius's location history and his letters were then implemented using various javascript libraries.[12]

3.5 Refined Model

In the refined format developed for exploring this data, individual letters are represented by points – red points for letters sent by Comenius, blue for those sent to him (fig. 8). The times and places of the letters are indicated on the matrix while the map displays his movements and residences and pinpoints destinations of letters sent by Comenius (red circles) and origins of letters received by him (blue circles). Visualizing Comenius's lifelong itinerary and complete correspondence results in a visually overloaded matrix-map. In order to mitigate this effect, and to facilitate the study of individual periods of Comenius's long and turbulent life, his itinerary was divided into seven main stages (labelled A to G in the upper left-hand corner of the interface): clicking on one of these allows the user to visualize data only from this shorter period of time in both the matrix and the accompanying cartographic representation (fig. 9). In this case, the matrix only includes the places relevant to that interval in Comenius's life, considerably enhancing the legibility of the figure as a whole.

[12] Main libraries used were *D3* and *React*; both help connecting data to the elements of web pages. *D3* (https://d3js.org/) is specifically suited for producing dynamic, interactive data visualizations, while *React* (https://reactjs.org/) is used for building user interfaces. Both accessed 20/03/2019.

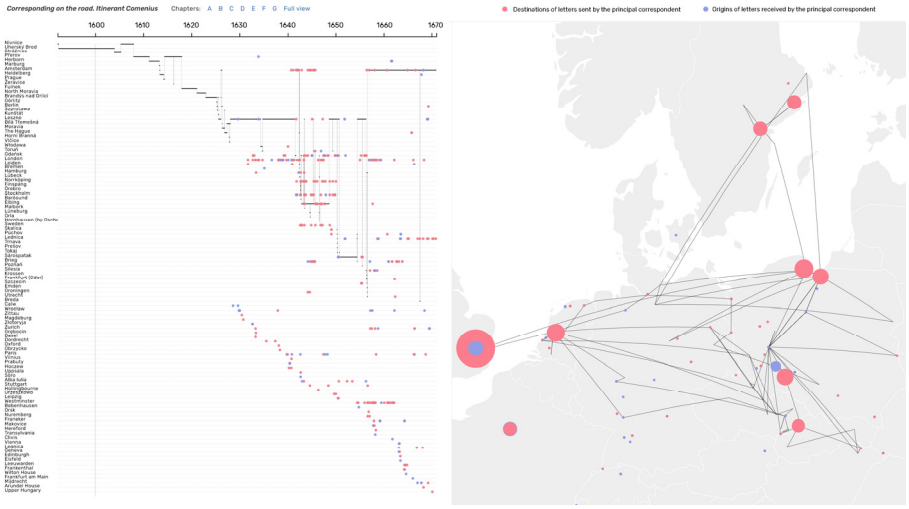


Figure 8: A refined matrix-map prototype, populated with data on Comenius's complete correspondence and itinerary

Figure 9: Visualization of data relating to the period of Comenius's life between his stay in London and the departure from Elbing (September 1641–October 1648)

3.6 Some Basic Findings

One of the most promising features of the refined prototype is that it enables us to represent, in considerable detail, how one person's physical movements are related to changing patterns of his/her correspondence contacts. The results are intelligible at several different levels.

At the most general level, the matrix representing Comenius's entire life (fig. 8) is divided into two roughly equal parts. The top half includes letters sent to and received from places that Comenius visited. In most cases, his epistolary contact with these places seems to have begun at or shortly before his first visit, and in many cases it extended long after his departure from these places as well. The lower half of the matrix includes the letters exchanges with places that Comenius did *not* visit. With the majority of these places, Comenius exchanged only one extant letter, and with only a few of them did he correspond extensively and over a long period of time. This relatively extensive correspondence included not only ecclesiastical letters exchanged with important Reformation centres (like Zurich) and dealing with ecclesiastical agenda but also major intellectual centres (like Paris). This is a very significant finding, since it suggests that our understanding of the correspondence network of an individual such as this is incomplete until it has been analysed alongside the data on their itinerary as well. In other words, in the case of Comenius, personal, face-to-face contacts appear to be very important in generating or intensifying correspondences that are long-lasting. In some cases, however, a personal meeting may lead to a spike in reciprocal contact, after which the frequency of correspondence temporarily decreases, as in the case of Comenius and Hartlib during and after the Moravian's departure from England. Further study is needed to determine whether similar patterns are evident for other correspondents, itinerant and otherwise.

In Comenius's case, there is an active phase in the development of his correspondence contacts with Hartlib and some of his key correspondents, such as Joachim Hübner and Johann Moriaen, which not only includes intensive discussions of Comenius's pansophic project but also prepares his journey to London in 1641 (fig. 9). His itinerary after leaving England reflects previously established correspondence contacts with figures like Moriaen or Hotton in Amsterdam but also an interest in his work from figures with whom he never corresponded (like Descartes) sparked by intermediary figures with whom he did correspond but whom he never met (like Mersenne). At the same time, however, his itinerary in the 1640s reflects development of new correspondence contacts with Sweden, the Netherlands, and the Baltic area (e.g. Gdańsk/Danzig). The prototype can also help formulate questions regarding the frequency of correspondence contacts in certain periods in relationship to a region or city where Comenius was, and its accessibility and connectedness. It is important to remember that we are working only with extant letters and that this is only a part of an actual correspondence exchange. It is striking, therefore, that the correspondence contacts with Hartlib

and figures central to his network were more frequent in the 1630s and early 1640s, even if we know only a small portion of them and even if the correspondence connection between Leszno and London was not as easy as from Gdańsk or Elbing. After moving to Elbing in 1642, the correspondence with Hartlib was less frequent, as demonstrated by the number of extant letters and also by complaints in the letters of John Dury, even though the postal connection between Elbing and London was better.

3.7 Prospects

The matrix-map tool is a useful aid which, with further refinement, could help to generate new research questions and to verify hypotheses. A range of further developments could increase its utility further.

One such refinement could allow the user to drill down from the matrix-map to another view relating more details of the letters and correspondents in specific cities to the cartographic representation of Comenius's movements and letter exchange in the given period. Such a view would facilitate, for example, the detailed investigation of intensive correspondence between Amsterdam and London in the first years after Comenius's arrival in Amsterdam in 1656, and its contrast with his declining epistolographic contacts from his previous period in Poland and Hungary.

Second, the problem of the proliferation of places could be addressed by experimenting with different means of grouping places together in a preliminary matrix view, and then revealing further places as they become relevant. For instance, the initial view might subordinate individual cities to countries or regions to provide an overall picture, and only expose all the detail once the user navigates to a closer look. Alternatively, the system might only display the places relevant to the specific time period which is being investigated.

Third, the user might be given the option of enhancing the timeline on the matrix with important dates or events in European history. This might help grasp the relationship between major events – such as the defeat of the Bohemian Revolt in 1620, the expulsion of Protestants from the Bohemian Lands in 1627–8, the beginning of the English civil wars in 1641–2, or the Peace of Westphalia in 1648 – and major developments in Comenius's itinerary or correspondence network.

No less important would be the opportunity to test the matrix-map model on other data sets, beginning perhaps with other itinerant scholars in the period of the Thirty Years' War whose correspondence and location history intersected with that of Comenius, such as John Dury.

4 Towards an Interactive Exploration of the Postal Network: Systems of Postal Exchange[13]

Alexandre Tessier

Due appreciation of the complexity of postal communications increases still further upon recognition that letters were not transported instantaneously from place to place on a straight line. Letters were physical objects, carried by couriers and ships along established roads and sea lanes. Even after the establishment of formal postal systems, the connections between equidistant places could vary enormously in frequency, cost, and time expenditure. In order to understand the evolution of postal connectivity between various parts of Europe, new systems will be needed that are capable of capturing and displaying data on the time, cost, and routes required to send letters between a multitude of different places and eventually outside Europe.

With this objective in mind, this section explores options for the creation of an interactive map of postal services in the early modern period. The intended tool would provide a Google Maps interface that would allow the user to visualize the itinerary followed by a letter sent within the early modern period from one designated place in Europe to another at a designated time. The section considers models on which to base such a resource before outlining the kinds of data, interfaces, and technical systems required.

Inspiration is provided by three existing web-based tools which explore the duration, cost, and routes of journeys during different historical periods.

The first, ORBIS: *The Stanford Geospatial Network Model of the Roman World*, has been developed since 2011 to analyse and display data gleaned from a host of historical sources on journeys within the Roman Empire at its height, c. 200 CE (fig. 10).[14] The search box to the left is used to enter the parameters of the query. In figure 10, a journey has been selected between two distant points – from Trapezus (Trebizond, on the Black Sea) to Deva (Chester, in north-east England) – in the summer and by any available route; and three successive searches have revealed the most direct route (in purple), the fastest route (in red), and the cheapest route (in green), in the manner familiar from a modern route planner. The strip at the bottom of the screen allows the user to explore these options further with reference to

[13] The work leading to this section was undertaken as part of a Short Term Scientific Mission (STSM) held at the universities of Oxford and Lancaster, 1 November–16 December 2015. I owe many thanks to the following individuals, who helped me at various points during this STSM or in the preparation of this report: Prof. Howard Hotson, Prof. Ian Gregory, Arno Bosse, Miranda Lewis, Dobrochna Futro, Martin Hadley, Prof. Bruno Martins, Dr Stéphane Blond, Prof. Jean-Louis Loubet, Pierre and Antoine Tessier, and Xavier Pilas.

[14] http://orbis.stanford.edu/. On the sources, see 'Building ORBIS: Historical Evidence': http://orbis.stanford.edu/#fn1. For an enthusiastic review from an expert in digital humanities, see Scott Weingart's blog, 'ORBIS: The Next Step in Digital Humanities', 2 May 2012, http://www.scottbot.net/HIAL/?p=15585; all accessed 20/03/2019.

longitude and latitude, duration, distance, and cost via various means of transportation.

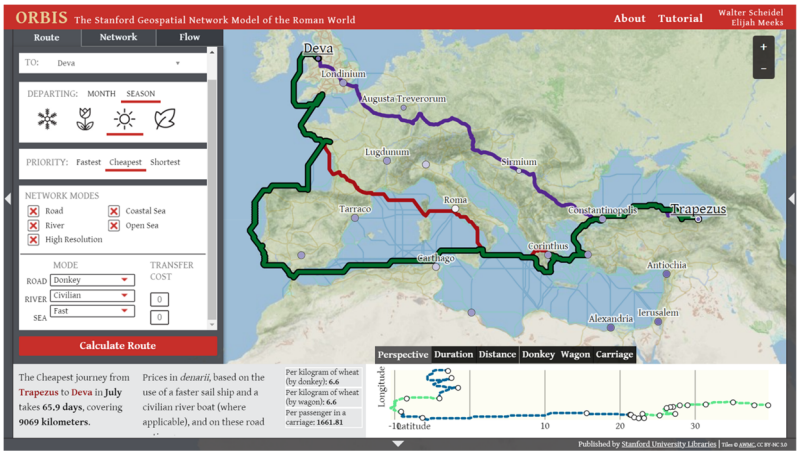


Figure 10: Screenshot of ORBIS showing the most direct (purple), fastest (red) and cheapest (green) routes from Trebizond to Chester

ORBIS also generates synthetic visualizations. Clicking 'Network' (in the top right) and then 'Zones' (in the bottom centre) produces an 'isopleth map' (fig. 11) which uses different shades of green to indicate the time taken to reach various regions from the point of origin. This geography coincides with the *fastest* route between Trebizond and Chester. Clicking 'Flow' (in the top right) superimposes the chosen journeys over what the authors call a 'Minard diagram' (fig. 12), which results from calculating all the most efficient routes to or from the point of origin and aggregating them to display the best routes between whole regions. This geography corresponds to the *cheapest* route selected earlier. The rich exploratory options facilitated by ORBIS provide a major source of inspiration for what could be created for early modern postal communications.

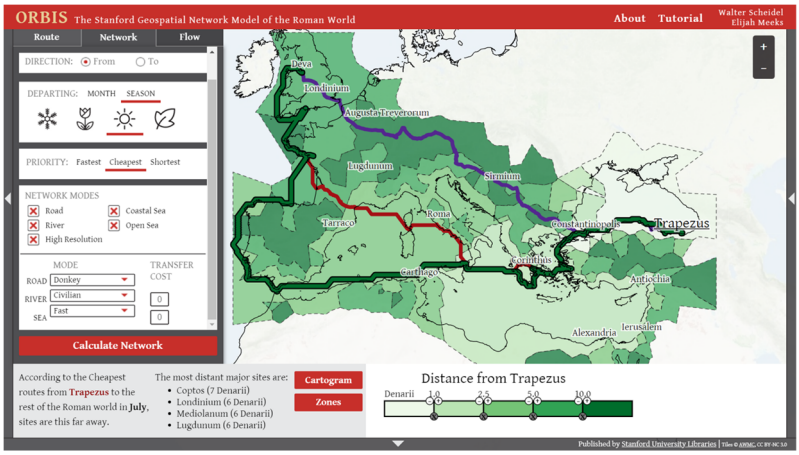


Figure 11: An isopleth map from ORBIS showing journey costs from Trebizond

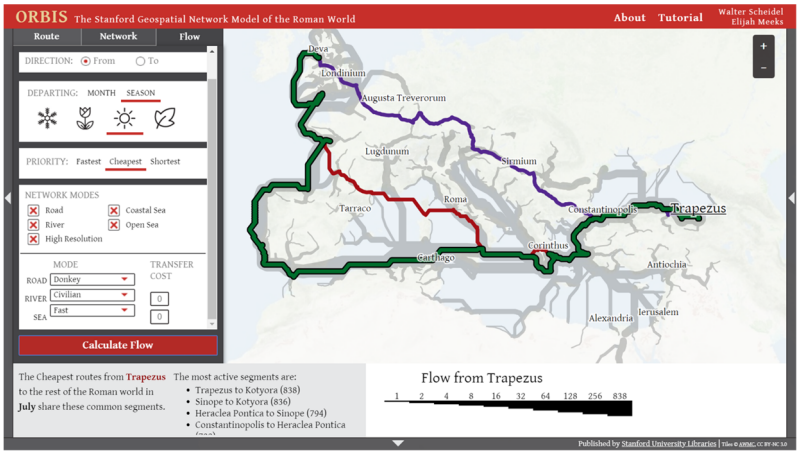


Figure 12: A 'Minard diagram' from ORBIS showing the most efficient routes from Trebizond

The second resource, entitled *OmnesViae* is a web-based tool, also developed since 2011, by René Voorburg. In purpose and achievement, it is similar to ORBIS, and although its scientific ambition is more limited, it offers alternative solutions and new features worthy of close attention. In this case, the underlying data sets derive essentially from a single source: the *Tabula Peutingeriana*, a thirteenth-century copy of an older *itinerarium* or road map of the Roman Empire, possibly depicting its status during the reign of Augustus (27 BC–AD 14). As its name indicates, *Omnes-Viae* provides a cartographic representation of *all* the main overland routes between major centres, on which the most direct (but not necessarily the fastest or

cheapest) road between two selected points is superimposed (fig. 13). In this case, the underlying map layer is provided directly by Google Maps, making it very easy to identify the relationship of ancient sites to modern places. Another fresh feature is the addition, on the left part of the screen, of an iteration of all the points on the map through which the selected route passes: simpler and clearer than the one in ORBIS, it is also far more detailed, as can be seen by scrolling down the list.[15]

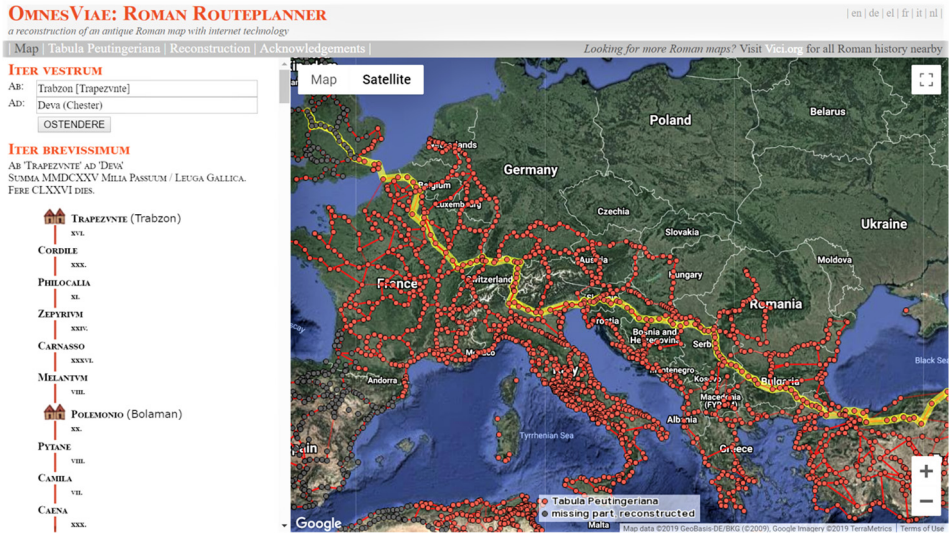


Figure 13: Screenshot from *Omnes Viae* showing its suggested route from Trebizond to Chester

The third website, developed by Cameron Blevins and Jason Heppler, also at Stanford, maps the proliferating network of post offices in the nineteenth-century American West (fig. 14). Apart from the direct focus on postal communication, the particularly relevant feature of this tool is the means of visualizing the *evolution* of a communications network over time. For this purpose, a sliding scale was added to the bottom of the main screen: as well as providing a bar chart representing the number of post offices established in any given year, this scale allows the user to select any portion of the whole period for visualization and study.[16] In this case, the period of rapid growth between 1875 and 1890 has been selected, and the mode of visualization has been chosen that uses colour to distinguish between post offices opened (blue) or closed (orange) during this period, opened *and* closed during these years (red), or opened previously and active throughout these years (purple).

[15] See http://omnesviae.org/fr/, accessed 20/03/2019.

[16] See http://cameronblevins.org/gotp/, accessed 20/03/2019.

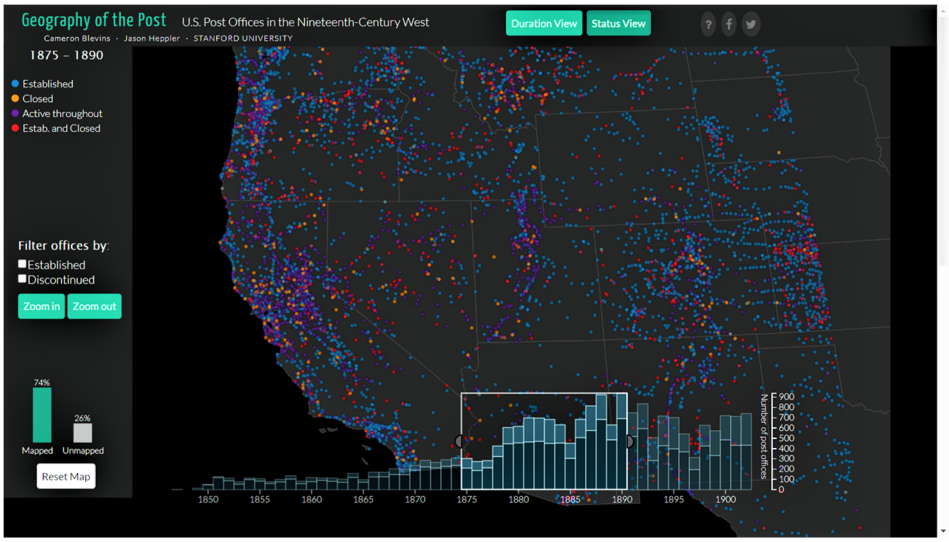


Figure 14: Screenshot from *Geography of the Post* by Blevins and Heppler, showing the evolution of the US Post Offices in the American West between 1875 and 1890

These three existing tools provide encouraging examples of what is possible in this area as well as sources of inspiration for devising something similar for mapping early modern postal communications. Clearly, the global architecture of this analogous system must combine three major components: a graphical interface combining several of the features described above; a gazetteer of early modern Europe containing place names in all relevant forms and their locations; and an additional database, storing data on Europe's evolving postal systems, such as the locations of post offices, together with systems for calculating or estimating the routes between them, the costs incurred, and the formal timetable of postal collection and transportation.

At this preliminary, conceptual stage, it is possible to envisage these three main components with differing degrees of precision. The public online interface can be described with a high degree of accuracy already, and this is the main goal of the brief text which follows. The historical gazetteer coincides in essence with that described in chapter II.2 of this volume, and will not be further described here. The underlying digital architecture remains more speculative, but some of its key features are sketched out below.

It is relatively easy to begin to envisage the basic features of the online, public interface, since several of them would be derived from Google Maps and the three projects described already. Figure 15 provides a preliminary sketch of one possible configuration. In the upper left-hand corner of the screen is a small query panel consisting of three boxes where the user can indicate: (A) the place of origin of a selected letter; (B) the place of destination; and (C) the date of sending (exact or

approximate). Additional options could include information such as the format of the letter (single sheet, double, weight, etc.) relevant to calculating the cost. A virtual button at the bottom of the query panel would allow users to execute a completed request. The large visualization panel in the centre of the screen would then display a topographical map of the probable itinerary followed by the letter from A to B, distinguishing known features of the path from conjectural ones with a visual code such as colour gradient or line thickness. For convenience, users should be allowed to zoom in or out on any zone of this map. When users mouse-over or click on points on or segments of the route, pop-up flags might indicate the next post stage or post office, as well as the day and time when the letter is supposed to have reached this place. Finally, the option of providing the origin and destination by clicking on any places of the global visualization panel should also be considered. This option does not exist in ORBIS, but is offered by Google Maps and greatly improves the flexibility of the whole system

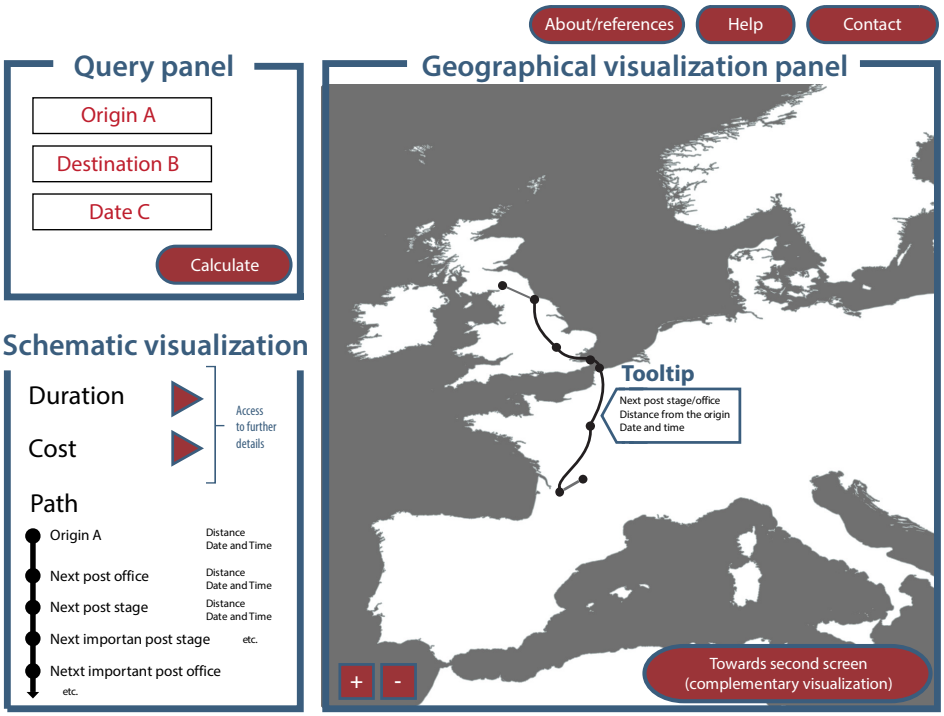


Figure 15: A schematic plan for a public interface

As well as the main geographic visualization panel, a second, smaller panel may be needed to display the results in a more schematic manner. Inspired by the similar feature in *OmnesViae*, the panel in the bottom left of figure 15 reduces the path followed by the letter to a chronological diagram which itemizes the main stages of

the selected postal route together with the distance travelled and the time elapsed since sending. It could also display gross estimates for the delivery date and the final cost. For users desiring further information, small virtual buttons might give access to detailed calculations and propose a bell curve of likely dates of receipt.

Additionally, as in ORBIS, the system should be able to give access to a second screen devoted to the display of synthetic visualizations such as (1) isochronic maps of duration of travel to or from points A or B; (2) isopleth or choropleth maps reflecting the cost of postage from or to A or B; (3) maps of the entire formal European postal network at the time selected; and (4) 'Minard diagrams' depicting the postal connectivity of a given point to all the other points on the map. A further mode of visualization, inspired by *Geography of the Post*, might facilitate exploration of the manner in which postal services evolved on a national or European scale. In addition, chronological diagrams or animations might also provide a notion of the acceleration from one year to another of postal communication between any two points, or of the cost variation in the same time-span, or of the extension of postal networks in Europe. Last but by no means least, the public interface should include a copious explanatory section, in order to provide users with help, but also to defend the scientific validity of the whole model, while facilitating the collection of feedback from users.

The public interface is relatively easy to envisage because it can adopt well-established solutions and because we already have a clear idea of the form of the answers to our questions. By contrast, the inner architecture of the system requires more reflection, since we have to deal with fresh challenges resulting from the specificities of early modern mail services. To begin planning in earnest, it is necessary at the outset to restrict the scope chronologically and geographically to an area of Europe and a period covered by well-documented public postal systems. For this reason, the initial prototype developed to date was limited to France and England at the end of the eighteenth century, a time when sources become abundant. This exercise was further restricted to only four postal routes: London–Berwick and London–Dover in England, and Paris–Calais and Paris–Urrugne (on the Spanish frontier near Bayonne) in France. The goal was to use this restricted network to create an algorithm able to reflect the main characteristics of a mail service that could potentially be expanded to join any place in England to any place in France.

Implementing this simplified system required the development of data sets to represent the post stages and post offices in France and England. Sources from which to derive this data are numerous, including postal maps and guides, schedules for the postal services, and local regulations. Identifying the places mentioned on the old maps was often problematic, sometimes requiring informed choices between several options to create simple, machine-readable data.[17] Information on locations was then structured to create a table including a place of origin (=A), a

[17] For a valuable bibliography of works relating to French postal roads, see Anne Bretagnolle, Timothée Giraud, and Nicolas Verdier, 'Modéliser l'efficacité d'un réseau: le cas de la Poste aux chevaux dans la France pré-industrielle (1632–1833)', *L'Espace Géographique* 39:2 (2010): 117–31.

place of destination (=B), and a date (=C). This top level of data reflects the contents of the query panel in the final public interface. These data are then used to calculate and display the itinerary of the selected letter and the cost and duration of the journey.

In calculating the entire route, several different kinds of inference are required. The itinerary between A and B as a whole potentially joins any two locations in England or France. The first and last stages of the journey link A and B to the two closest post offices, A' and B'. A' and B' were linked, in turn, with the two closest post stations existing on the four postal routes included in the model. The cost of the journey was then relatively easy to calculate, since the attributes of each of the post offices include the cost for sending a letter to or from the capital city. The system must then be instructed to select these figures in the relevant data set and to add the relevant international tax from France to England or from England to France.

The calculation of journey *times* from A to B is more difficult to calculate, requiring a combination of several methods, since we lack data on the journey time between the place of origin and the first post office (AA') and between the last post office and the place of destination (B'B). The simplest expedient is to estimate the journey time in proportion to straight-line distance. By contrast, for any postal town to the capital city of the same country, and between London and Paris, we have prescriptive travel times collected from the sources at our disposal. These must be adjusted to deal with the delays which might occur between journey segments. For any international correspondence between Britain and France, these delays are most likely to occur at three points: in the first postal office of the itinerary, in the central office of the first country, and in the central office of the second country. In addition, it might eventually be possible to adjust the prescriptive times (describing how the postal service was *supposed* to function) with data on how it *actually* functioned gleaned from times of receipt harvested from a small fraction of the potentially huge number of letters from all periods assembled on a collaboratively populated resource such as EMLO.

Within the STSM, a satisfactorily functioning model of this kind was developed and tested using a variety of places in Britain and France. A very straightforward algorithm provided encouraging proof of concept, which can now be developed further. The next steps will naturally be to expand gradually and cautiously, beginning with well-documented areas. By remaining focused on the British and French postal systems at the end of the eighteenth century – immediately prior to the turmoil caused by the French Revolution – an exhaustive model can be achieved for this limited case, including all the postal routes and offices known as active in this time. This could provide a framework for expanding the chronological scope to previous periods and the geographical remit to surrounding countries. Since the documentation available for earlier times and other places is often less complete, expanding the scope of the tool will require the development of means of dealing with progressively greater uncertainties and larger documentary lacunae.

This kind of approach will be needed for the later eighteenth century as well, since even at that time not all letters were sent through formal postal systems. The progress already made in mapping the far patchier data surviving from the Ancient World provides good reason to expect that knowledge and methods can be developed to address these subsequent challenges effectively.

The scholarly benefits of such a system would be evident at multiple different levels. At the micro-level, such a system would provide improved means of determining whether one letter was more probably written before or after another was received – a seemingly trivial consideration, but one sometimes with major consequences for interpreting a correspondence. At the intermediate level, such a system could, for instance, provide a ready means of estimating the expense involved in exchanging a large volume of letters with a wide circle of friends over a long period of time. At the national and international levels, such a system might eventually prove capable of visualizing the shifting patterns of connectivity that helped knit together individual orders, countries, and confessions, as well as the whole of Europe and an expanding world. More generally still, a system populated by postal data alone could both benefit from and contribute to a more general system for tracking modes of transportation of all kinds. Although the availability of reliable data will of course limit the accuracy of such tools, the results already obtained from the far scarcer data available for the Ancient World provide abundant encouragement for further work.

IV.3 Chronologies of the Republic of Letters

Howard Hotson, Dirk van Miert, Alex Butterworth, Glauco Mantegari, Riccardo Bellingacci, Carlo De Gaetano, Christoph Kudella, Michele Mauri, Serena Del Nero, and Azzurra Pini

Introduction

Howard Hotson and Alex Butterworth

Every letter has a chronological as well as a geographical dimension. In fact, every letter is situated, not on one timeline, but on several. The process-based letter model (proposed in chapter II.7) outlines the temporal sequences of drafting, sending, conveying, and receiving a letter, as well as the varieties and vicissitudes of each. Letter records contain further chronological data, on the date of sending and occasionally also of receipt. The characteristics of the networks formed by letters vary over time in ways that traditional network analysis struggles to capture (ch. IV.5). The same applies to the postal and other systems for distributing letters in the first place (ch. IV.2, sect. 4). Collections of letters, in script and print, have histories that need to be mapped and narrated (ch. III.1). The movements of correspondents complicate the histories of their correspondences (ch. IV.2, sect. 3). Last but by no means least, the topics discussed in letters change over time, as do the systems for organizing those topics (ch. II.5), including the very concepts of a 'letter' and the 'republic of letters' themselves (section 4 below).

Given these multiple layers of chronology, time is perhaps the richest field for the innovative visualization of correspondences; but it is also probably the least

well developed. Unlike space, only rather rudimentary means are readily available of visualizing time. The basic convention is to depict chronology on a timeline, an axis reading from left to right, punctuated by precisely quantifiable units, normally years. The centuries-old effort to depict temporal sequence in graphic form has experimented with circles, spirals, and vertical timelines, along with trees and other natural objects, images from biblical prophecy, and more complicated diagrams.[1] More recently, critical reflections on the orientation of timelines and an agenda for their improvement have helped provoke much interesting experimentation around the visualization of complex time.[2] Yet, despite this conceptual work, the options for developing the basic conventions of the timeline have not been systematically explored, non-traditional alternatives remain largely uncharted, and even the recent wave of innovation has yet to inform the tools and visual languages readily accessible for routine scholarly use. Although no full census of the options available can be presented here, the examples that follow, mostly produced by design sprints and STSMs, give some impression of the solutions awaiting development.

1 Correspondence Metadata: Item-level Records

1.1 Histograms

Howard Hotson

The most familiar means of visualizing the chronology of a correspondence is the histogram, which numbers letters on the vertical axis and years on the horizontal one. A useful development of this approach, which has been implemented in *Early Modern Letters Online*, allows the user to compare incoming and outgoing letters in parallel charts, in a single chart, either with outgoing correspondence stacked on top of incoming (as below) or with each series represented by independent bars. Such visualizations provide an almost instantaneous impression of the general 'shape' of a correspondence's chronological development. Only a few years ago, such an impression could only be constructed through a laborious exercise. Although commonplace today, the potential of this simple tool has yet to be systematically exploited. Why, for instance, is Constantijn Huygens's correspondence (figure 1) divided into two quite separate peaks? And why do letters to him preponderate in some periods more than others? Do these changes represent phases in his career, or mere vicissitudes in the survival of his correspondence? How common is this kind of division? And how many other chronological 'shapes' do learned cor-

[1] This history is sumptuously illustrated in Daniel Rosenberg and Anthony Grafton, *Cartographies of Time: A History of the Timeline* (New York: Princeton Architectural Press, 2010).

[2] See the seminal contribution by Stephen Boyd Davis, 'History on the Line: Time as Dimension', *DesignIssues*: 28:4 (Autumn 2012): 4–17, see https://doi.org/10.1162/DESI_a_00171.

respondences take? It is only when we have multiple data sets to compare that such questions even arise.

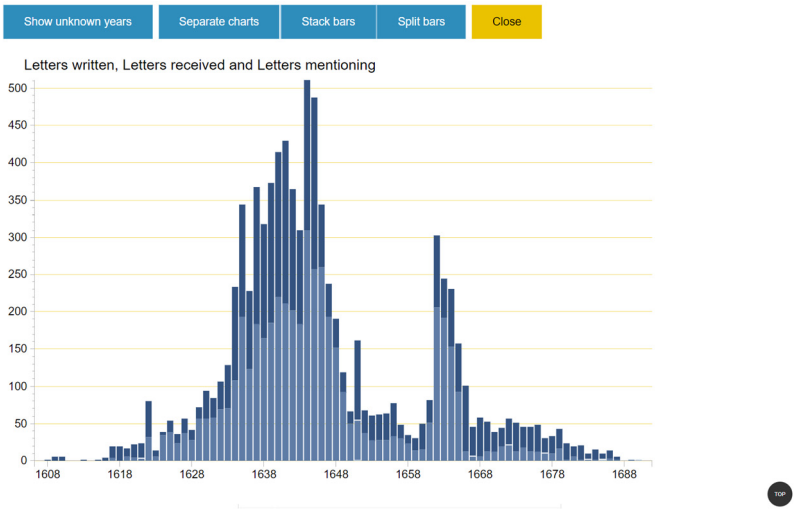


Figure 1: The correspondence of Constantijn Huygens (1596–1687)[3]

Even less has been done to explore the utility of this tried-and-tested mode of visualization on other kinds of readily available data. For instance, did the number of people with whom Huygens exchanged letters vary equally dramatically over time? Even when the total number of incoming and outgoing letters did not vary greatly, was he sometimes exchanging high volumes of letters with a few intimate friends and other times exchanging only a few letters with far larger numbers of people? More interesting results could be generated by graphing the results of textual analysis in a similar way. For instance, with a sufficiently large dataset it should be possible to determine whether patterns of co-citation changed over time. Another set of options could emerge from graphing the results of quantitative network analysis in similar fashion (on which see ch. IV.5). Within a stable network of overlapping correspondences, can changes in the betweenness centrality or eigenvector centrality of a given individual or cluster of individuals be meaningfully graphed in this way? Given the difficulty of capturing change within most forms of network analysis, these are possibilities which merit exploration.

[3] Data derived from Jacob A. Worp, ed., *De briefwisseling van Constantijn Huygens (1608–1687)*, 6 vols., (The Hague: Martinus Nijhoff, 1911–1917) via the *ePistolarium*. Visualization on EMLO: http://emlo.bodleian.ox.ac.uk/profile/person/69f002da-1994-418d-9cc4-d8d544d64121, accessed 20/03/2019.

1.2 Horizontal Bar Charts: Browsing the Chronologies of a Union Catalogue

Carlo De Gaetano, Howard Hotson, Glauco Mantegari, and Azzurra Pini

Histograms – with time represented on the horizontal axis and some other measurable quantity on the vertical one – thus have many underexploited possibilities for future development. Still other options emerge from removing the vertical axis altogether and developing the horizontal timeline in other ways. A few of these options were explored in the first Como data-design sprint by a group charged with investigating new tools for browsing the multidimensional data involving people, space, time, and potentially also topics in the catalogue of early modern learned correspondence, *Early Modern Letters Online* (EMLO).[4]

Lives. EMLO has been assembled partly by bringing together catalogues of the correspondences of individual learned figures. One means of browsing such a union catalogue might begin with a series of horizontal bars running across a chronological grid. The starting and ending points of the bars would represent the birth and death dates of the individuals with their own correspondence catalogue in EMLO.[5] When multiple bars of this kind are stacked on top of one another, the result would be a chart such as that depicted in figure 2. In each case, the name of the correspondent would be provided: clicking upon it would take the user to the EMLO catalogue page, in which more information on the individual, the correspondence, and the state of the data is provided.

Letters. Within the horizontal bars, a vertical line would represent each letter sent or, if the user chooses, letters both sent and received, perhaps distinguished by colour. Since this arrangement would produce only an approximate impression of the distribution of known correspondence throughout an individual's lifetime, clicking on a bar could produce a histogram running along the same axis, more accurately displaying the number of letters sent and received each year.[6]

[4] See http://emlo.bodleian.ox.ac.uk/home, accessed 20/03/2019. Como data-design sprint, 4–8 April 2018, Project 5: 'Visualising EMLO'. Some attention was devoted to non-traditional strategies for visualizing the geographical and linguistic dimensions of the data, but the most fruitful line of development, outlined here, related to the chronological dimension: https://docs.google.com/document/d/1n4BaZ1rSa8LuLn_7ZMIb7ihDyrmzpaNf1yvetuN4oV4/edit, accessed 20/03/2019.

[5] When birth or death dates are unknown, the uncertainty can be represented by shades of grey, with different levels of uncertainty visible by drilling down to a more detailed view. This would require a data model refined in the manner indicated in chapter II.3 to represent different kinds and degrees of uncertainty.

[6] It should be possible to separate and stack the representations of letters sent and received and to zoom to full screen, as on EMLO: e.g. http://emlo.bodleian.ox.ac.uk/profile/person/edb080aa-312f-4e4b-ae3d-0cf4f1eb7a67, accessed 20/03/2019.

Figure 2: Horizontal bar chart for browsing a union catalogue of learned letters

Groups. While the initial view might include all the correspondences in the union catalogue, the user would be able to select clusters of overlapping correspondences for closer study in a variety of ways. After selecting a principal correspondent, one option would be for the system to select automatically all the other figures who exchanged substantial numbers of letters with him or her, whether or not they were provided with separate catalogues of their own within EMLO. These automatically selected correspondences might initially be ranked in terms of the number of letters exchanged with the principle correspondent. The user would then be able to refine this automatic selection in a variety of ways: by delimiting a narrower chronological period, geographical region, or topical field, for instance; or by manually selecting specific correspondents for closer study.

Topics. An additional refinement would colour-code letters by main topic of discussion. Since one letter can discuss multiple subjects, the topics in any given letter would need to be ranked in importance to generate a first overview; but all the topics in the letters could appear sequentially with the right selecting and filtering of topical categories. This option would obviously depend on the availability (1) of a standard system of categorising topics (such as the method outlined in ch. II.5); (2) of an epistolary data model enhanced to accommodate these topics; and (3) of large numbers of epistolary records categorised in this manner, which would also allow users to isolate and analyse correspondence on individual topics. In a system including machine-readable full texts of letters, topic modelling might eventually populate the topical data field automatically, while calculating the number of words devoted to each topic as a way of ranking them in importance.

Conversations. A more easily implemented refinement would allow users to highlight conversations between the correspondents selected. The example below, for

instance, shows how a conversation thread between two individuals might be highlighted as the user hovers over a particular letter within it. Although this particular representation would require metadata indicating which letters responded to which, a simpler version could highlight the letters exchanged without inferring the sequence.

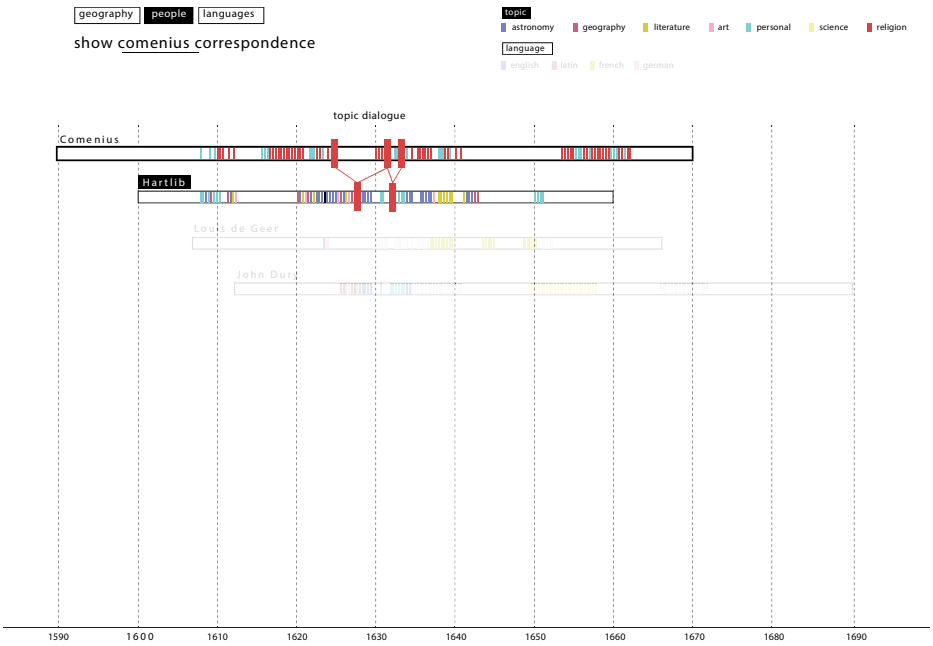


Figure 3: Indicating exchange of letters within a horizontal bar chart

Conclusion. Tools of this kind, for browsing correspondence metadata, are obviously quite different from tools for presenting the fruits of analysis in published visualizations. Like several of the proposals outlined in this volume, this section therefore represents a small illustration of the potential of data interaction design to enhance scholarly processes in the first instance and to enhance scholarly end-products only as a result.

1.3 'Maps with Memory': Capturing Chronology in Cartographic Form

Alex Butterworth, Howard Hotson, Christoph Kudella, Michele Mauri, and Serena Del Nero

One-dimensional timelines reading from left to right are a natural starting point for data interaction design in this field, but not an inevitable end point. The use of colour opens up the possibility of indicating temporal sequences on two-dimensional visualizations such as maps. This possibility may hold a key to overcoming the limitations of more traditional static and dynamic maps.

Static maps normally display all the correspondence of a lifetime as if it had taken place simultaneously. Although invaluable in providing an introductory overview of a correspondence, this mode of visualization fallaciously implies that the chief correspondent maintained contact with all of his contacts and their places simultaneously, and overlooks the possibility that a correspondence may have changed shape dramatically over the course of a lifetime.

Dynamic visualizations, on the other hand, cycling through a lifetime one year at a time, have the opposite disadvantage: they visualize a sequence of momentary states, typically retaining no impression of past transactions and no anticipation of future ones, in a manner that users cannot easily synthesise into a comprehensible image or narration.

Hence the need to experiment with means of creating 'visualizations with memory', that is, maps or other visuals that retain some fading indication of previous movements or transactions while distinguishing recent ones and highlighting current ones. More abstractly stated, the question is: how can visual conventions be devised to represent two or more network states, simultaneously and comparatively – that is, what is happening 'now' and what happened previously – within the same map?

Experimenting with answers to this question was the task set for the third group within the second data-design sprint convened by the Action in Como in the Spring of 2017. The group addressed this challenge by means of a case study of the ego-network of Desiderius Erasmus (1466?–1536) of Rotterdam, thanks to the meticulously curated catalogue of his over 3,000 letters contributed by Christoph Kudella.[7] Erasmus is traditionally regarded as the defining figure of the 'Erasmian humanism' of the immediately pre-Reformation era. His was perhaps the most genuinely pan-European correspondence in the whole history of the republic of letters, penetrating deep into both the Iberian peninsula and east-central Europe as well as England, the Low Countries, France, the Germanic lands, and Italy. His career peaked at exactly the moment when Luther's Ninety-Five Theses began a

[7] This data was created as part of his doctoral dissertation: Christoph Kudella, 'The Correspondence Network of Erasmus of Rotterdam: A Data-driven Exploration', Unpublished PhD Thesis, University College Cork, 2017.

process that led to the polarisation of Europe into antagonistic confessional blocks. As the focus for the case study, the eleven-year period between 1508 and 1518 was chosen, which saw his mobility peak alongside his prospects, influence, and the range of his correspondence, just before the Reformation controversy tore his humanist circles apart.

As Kudella's research had already shown, visualizing such a correspondence in the traditional manner is unsatisfactory for several reasons. One is that Erasmus's network changed shape dramatically in the course of his lifetime. Another is that Erasmus's constant movement makes any traditional map of his correspondence difficult to interpret. Of several possible responses to this problem, the group chose to explore a strategy that allowed some of the chronological development of a network to be captured in a cartographical representation, and the same set of data to be visualized in a variety of different ways while providing ready access to the underlying metadata and, potentially, even the letter texts.

Traditional cartographic representations aggregate multiple letters into a single symbol or icon, normally a circle, which expands or contracts in size depending on the number of letters it represents. The alternative mode of visualization explored here disaggregates these large circles in order to convey more information about the individual letters that they represent.

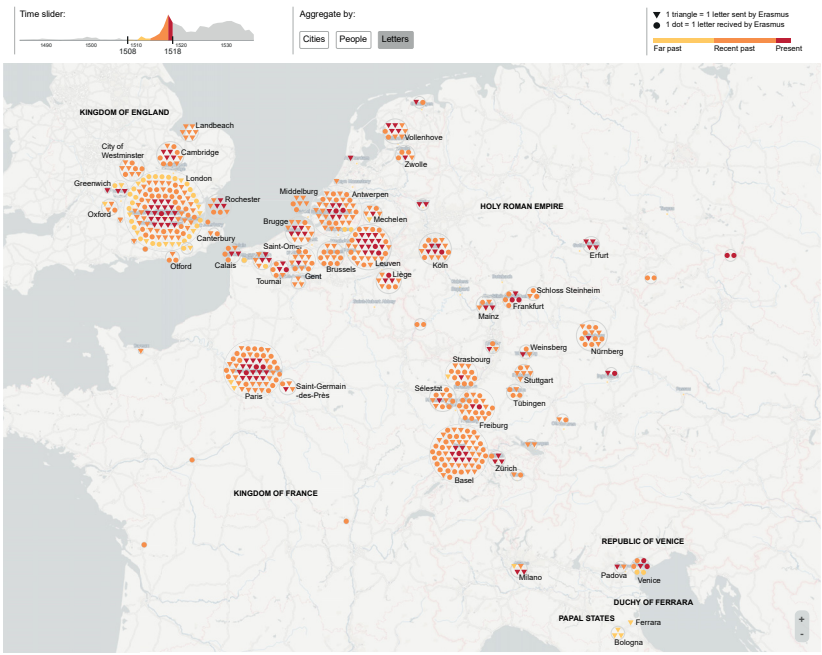


Figure 4: Mockup visualization of Erasmus's correspondence, 1508–18

In the mode of visualization used in figure 4, each individual letter is represented by a single *icon*, which is distinguished by its shape, order, and colour. The *shape* of the icons distinguishes letters sent from letters received by the principal correspondent: a circle represents one letter sent to Erasmus, and a triangle (the regular polygon most different from a circle) represents one letter sent by Erasmus. The *order* of icons is chronological: new letter-icons appear at the centre of these roughly circular clusters, displacing older ones, which spiral outward in anti-clockwise fashion. The *colour* of a letter-icon indicates how recently it was sent. Discrete shades are used instead of continuous colour gradients, because they allow comparison between places. In this case, the most vivid colour (red) indicates the most recent interval (1518); orange represents a longer, intermediate middle interval (1514–17), and the least vivid colour (yellow) represents the first half of the full interval (1508–13).[8] The size and number of these intervals, indicated by the legend that runs along the top of figure 4, could be changed at will by the user in a number of ways.[9]

Such an arrangement reveals considerably more about this decade of Erasmus's correspondence than the more traditional view. For instance, although Erasmus exchanged a similar number of extant letters during this decade with Basle (57) and Paris (55), and with Antwerp (38) and Leuven (37), the red icons indicate that his liveliest contact had recently shifted from Basle to Paris and from Antwerp to Leuven. Likewise, although Erasmus exchanged a similar number of known letters with Brussels (12) and with Bruges (14), the correspondence with Bruges appears to be active and ongoing in 1518, while that with Brussels is mostly passive and less recent, insofar as we can judge from extant correspondence. London is the only city with which Erasmus was demonstrably in frequent epistolary contact throughout this decade, but closer inspection reveals that the nature of this contact has changed: most of the letters in the earliest period (in yellow) were sent *to* Erasmus; all but one of the letters from 1518 (in red) were sent *by* Erasmus; and in the middle period (in orange) incoming and outcoming correspondence was equally balanced. Selecting a location might also produce a histogram showing all letters exchanged with that place, as well as those sent or received by Erasmus while he was resident there. To place the time-span selected in the context of the entire life, users could select a visual mode in which hair-thin circles appear around each place showing the size of the cluster of *all* the letters to and from there ex-

[8] Distinguishing letters individually in this way could also facilitate new functionality. For instance, mousing over a letter-icon might highlight the precise date in the timeline, the location of Erasmus at the time of sending, and all the other letters sent by or to that correspondent. Clicking on a specific letter-icon might reveal its metadata in a pop-up box, which could click through in turn to the letter text, translation, or image.

[9] The user would ideally be able to restructure the timeline for further study in three different ways: (1) by *moving pointers* along the timeline, to define a rough period; (2) by *manually inputting* precise start and end dates into boxes provided; or (3) by *selecting major events* from the life of the main correspondent from a list derived from the prosopographical metadata. These events might include major changes of place (e.g. Erasmus moves to Basle), career-changing events (e.g. Erasmus publishes the *Adagia*), or events in the wider world (e.g. the publication of Luther's Ninety-Five Theses).

changed with Erasmus over his lifetime. Rendering the data more transparent in this manner is, of course, a means of raising fresh questions as well as answering them. For instance, does this demonstrable shift represent a change in the pattern of Erasmus's correspondence with London, or in the pattern of the survival of his letters as he grows in fame during this crucial decade? The answer might be revealed by considering whether the yellow letters were sent to the relatively obscure Erasmus by more prominent figures and therefore deemed worthy of preservation.

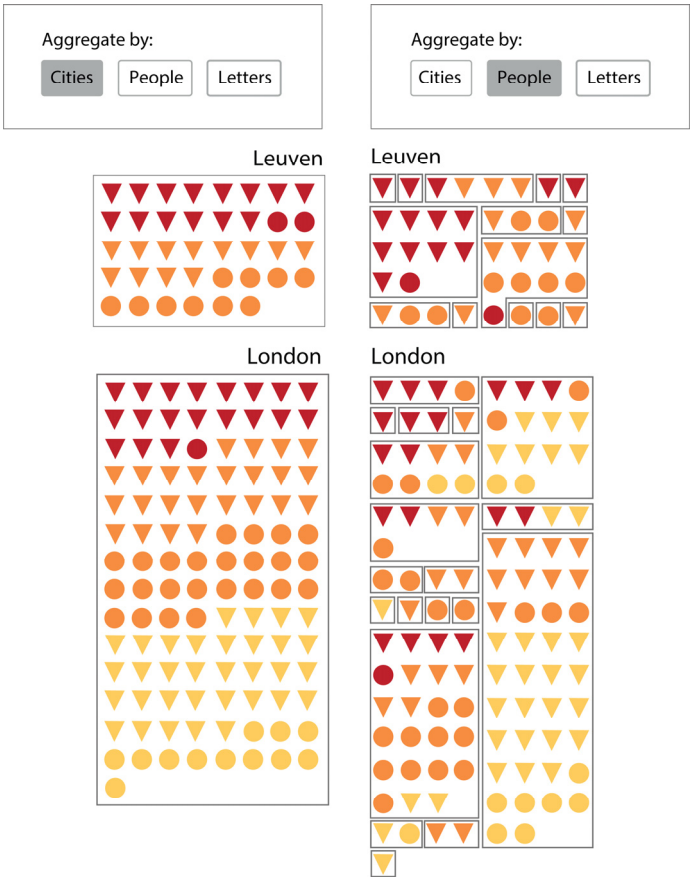


Figure 5: Erasmus's correspondence with London and Leuven, 1508–18, visualized in two different modes

In figure 4, the letter icons are arranged in a spiral in order to approximate the geometry of more conventional maps of correspondence. Figure 5 explores two other arrangements in which the letter icons are arranged in a rectangular cluster. In the left-hand pair, the letters to and from a given city are grouped together in a single large box, with the most recent in the top left and the oldest letter within the

time frame in the bottom right. This results in coloured bands giving a sense of the history of Erasmus's surviving correspondence with this place during the time-span selected. Again, the results are instructive. During the first half of the decade, Erasmus corresponded with London but not with Leuven; but in the second half of the decade, he sent not only forty-eight letters to London but also thirty-eight to Leuven. In 1518, sixteen letters were exchanged each with Leuven and London. At this point users could opt to use a larger number of colours to display smaller intervals of time, to see whether more differences of this kind were revealed.

Another option would be to change the manner in which the letters to or from a given city were clustered. In the right-hand pair in figure 5, the same icons have been regrouped into smaller boxes in the manner of a tree diagram, with each box containing icons for the letters exchanged between Erasmus and a single individual, again arranged chronologically, with the most recent letters in the top left and the oldest in bottom right.[10] This differentiates the two cases even more clearly. About 65 per cent of the London letters (75/113) were exchanged with just three people (Thomas More, John Colet, Andrea Ammonio), and most of this correspondence occurred early in the decade. In the second half of the decade, Erasmus's correspondence has become atomised: about 40 per cent of the letters from 1514-17 (20/48) are from exchanges consisting of only one or two letters during the entire decade. The Leuven correspondence is similarly atomised: the thirty-eight letters are exchanged with fourteen different people in Leuven, averaging only 2.7 letters per person in this entire decade.[11] Again, various interpretative possibilities emerge: is this a shift in the actual pattern of Erasmus's correspondence, or in its pattern of survival, as Erasmus's growing fame leads more recipients to preserve his letters?

Another advantage of this system of representation is that it might be instructively animated. Better still, arrows in the header strip might allow users to move the entire timeline forwards or backwards one year at a time with a single click. As the system advanced year by year through the lifetime of the correspondent – from 1508–18 to 1509–19 and so on – individual letter icons would appear at the centre of the spirals, while others disappear from its periphery. A further symbol icon could track the movement of Erasmus himself, with the same colour scheme used to indicate his pathway through previous locations within the selected time frame. Expanding the interval captured in the timeline and increasing the number of col-

[10] Once again, these boxes would provide a ready means of drilling down into the underlying data. Mousing over a person-box could reveal a histogram showing the chronology of their lifetime correspondence with Erasmus, with the selected period coloured in the same way as the timeline. Clicking on a person-box could also highlight other places from which that individual exchanged letters with Erasmus during the period being studied and the locations of Erasmus when the individual letters were sent.

[11] Clustering letter-icons in this way might also assist users in drilling down into the underlying data. For instance, selecting an individual person-box might reveal other locations of the individual during the time-span selected; a bar graph, displaying all letters sent to/received by this individual with Erasmus throughout the selected timespan (as highlighted within the entire correspondence); and the location(s) of Erasmus recorded in the metadata of these letters.

oured divisions within it would smooth the passage from one year to another and facilitate the perception of longer-term changes in the structure of the correspondence, which could then be studied in more detail with this and other means. The result would be a mode of visualization essentially midway between static visualizations (which do not represent change) and animated visualizations (which do not represent stasis).

This method of representing places as clusters of individual letters also provides a possible solution for another well-known problem. In a more traditional map of correspondence, when the density of data in a particular area is high, the circles representing all the letters to and from individual places overlap and occlude one another. In this alternative visualization mode, when two clusters of letter icons overlap, the clusters might merge to form a single, larger cluster. These merged clusters could be distinguished visually from single place icons, for instance, by a darker external border, or by being divided internally in the manner of a tree diagram. When users zoom in, these merged clusters could resolve into separate clusters for each place.[12]

This mode of visualization will nevertheless probably work best for relatively modest number of letters – hundreds rather than thousands. For that reason, the default cartographic option is likely to remain the kind of conventions made familiar by *Palladio*, in which the number of letters is indicated by size of circles, overlapping where necessary, with directionality indicated by curved edges. In the case of Erasmus, this initial mode would reveal crucial features of the correspondence, such as its remarkably pan-European scope and its extreme complexity especially in the Rhineland corridor. Users would then need a number of other modes of visualization in order to explore various features of this complex data set in more detail.

2 Correspondence Metadata: Collection-level Records

Riccardo Bellingacci, Carlo De Gaetano, Dirk van Miert, and Glauco Mantegari

Another dimension of epistolary chronology relates, not to individual letters or the networks formed by them, but to the collections in which they were preserved. As noted in section 3 of chapter III.1, the history of collections of manuscript letters needs to be better understood in order for their contents to be interpreted correctly; and the chronology of collecting could often provide the framework for such histories. Such collection-level chronologies would track when particular collections were assembled, or how long it took, focusing on the activities of figures like Zacharias Conrad von Uffenbach or Count Otto Thott.

[12] The viability of this solution can be studied here: http://glammap.net/glamdev/maps/1, accessed 20/03/2019.

Far easier to assemble is evidence of the chronology of collections of letters preserved in print. As discussed in section 2 of chapter III.1, printed letter collections are clearly situated in time by year of publication. This data can be mapped and graphed to reveal where and when printing epistolographies was popular. But printed collections typically also have one principal author, whose dates of birth and death are usually known. Relating the dates of death to publication dates will indicate whether a collection of letters was published during an author's lifetime or posthumously. In the latter case, it is easy to calculate how many years elapsed before an individual's correspondence was published. This provided the impulse for a small-scale research project within the COST Action. The basic data for the project was provided by the 1,874 titles of books holding printed letters assembled (as of June 2016) in *Epistolaries of the Republic of Letters* (EROL). The first data-design sprint in Como experimented with two types of visualization to enhance EROL.

The first experiment used *Palladio* to visualize the geography of printed letter collections, which could be dynamically mapped across time. The timeline showed that the majority of these books were published between the mid-seventeenth and mid-eighteenth centuries. This is unsurprising, since the bibliographies on which EROL was based at that time were focused mostly on that time period. But the graph also yielded more interesting granular information: within the time frame 1600–1750, which is evenly represented in the data set, the heyday of letter printing was in the first quarter of the eighteenth century. To make the resulting visualization even more interesting, the bar for each year was divided into segments, with each segment indicating a particular city. Hovering the mouse pointer over a segment lit up that city in the bars of the other years. This particular graph showed that Paris was a major centre for publishing letter collections throughout the early modern period, but that it ceased to be so for the second half of the eighteenth century. Alternatively, it would have been easy to add a facility for clicking on a location and yielding a timeline for the printing of epistolographies in that city, revealing when the printing of letters was popular in that particular place. Needless to say, such functionality could also be repurposed for other categories of print publication.

The second and more complex question regarded changes in sociocultural practices within the republic of letters. During the Renaissance, letters were regarded as a literary genre. Printed collections of letters had the status of major literary works and were often meticulously edited for publication by the authors themselves. By the eighteenth century, this literary status had been replaced by the conception of correspondences as the preserved discussions of leading men of science and letters. Such collections could often be prepared posthumously by sons or followers of the luminary in question. An obvious research question was therefore when the transition between these two different conceptions took place. One way to help pinpoint this was to see when posthumous editions began to outnumber those completed during the author's lifetime.

A quick manual count of EROL in its early stages gave rise to the hypothesis that posthumous publication increased from the start of the seventeenth century onwards and became a major genre of publication in the course of that century. The task was to find a type of visualization that would reveal the answer to this question in a concise and clear way.

We explored this type of chronology by using the so-called RAWGraph's parallel coordinates. This visualization disproved our intuition that, from the start of the seventeenth century, the humanist practice of publishing one's own letters (as initiated by Petrarch) gave way to a practice in which publication of correspondence was left posthumously to sons and students.

Figure 6 suggests that letter collections tended to be published shortly prior to the death of the author. This probably represents the intentions of the authors, conscious of their approaching end of life, to publish their letters and memoirs. It may also signal a change in the strategies of authors. Erasmus had famously 'constructed his charisma in print' when he was in the midst of his career in an attempt to control his image in the minds of contemporaries.[13] People like Justus Lipsius followed the same strategy. This strategy was gradually replaced, however, by the desire to control the image of posterity: Of course letter-writers since Petrarch had cared for their posthumous reputations, but big-data analysis makes it possible to transcend the individual examples and generalise.

A second observation can also be derived from this visualization: the passage of time sees an increase in the publication of books of authors who had died at least a century before the publication of their correspondences. This trend starts from the nineteenth century onwards, but more data will be needed to establish whether this trend can be confirmed. If so, it suggests that publication practices were becoming more historicist and less a strategy of active image-control.

[13] Lisa Jardine, *Erasmus, Man of Letters. The Construction of Charisma in Print* (Princeton: Princeton University Press, 2015).

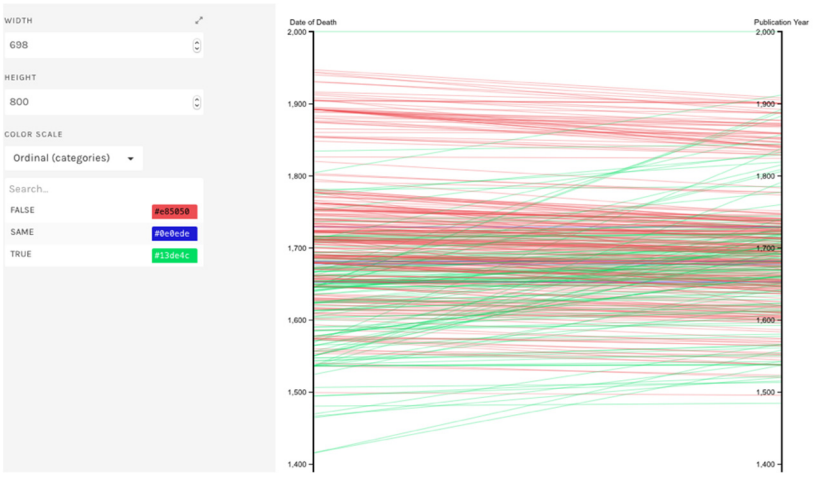


Figure 6: Use of 'parallel coordinates'[14] to show relation between date of death of an author and the date of his the publication of their correspondence (highest segment, from 2000 to 2000, is a dummy variable). Red segments are epistolaries published within the lifetime of the principal correspondent; green segments are published posthumously.

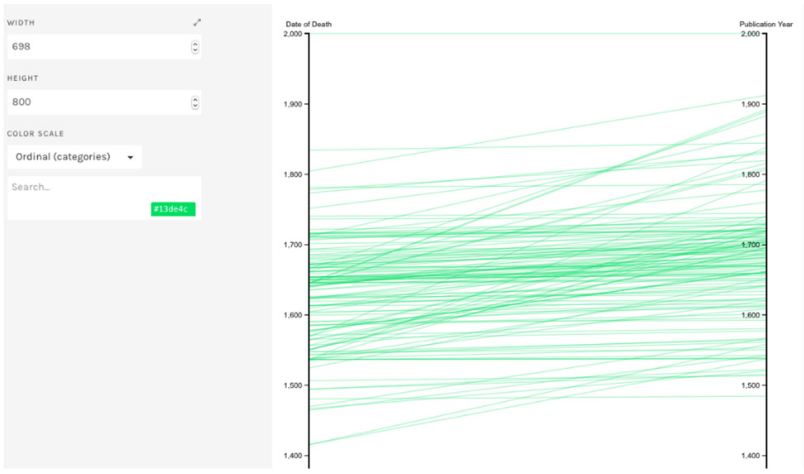


Figure 7: Only posthumous publications

[14] RAWGraphs: http://rawgraphs.io/about/, accessed 20/03/2019.

3 Prosopographical Data

Howard Hotson

Entire systems of intellectual exchange also move at several different chronological tempos. Postal systems have already been discussed by Alexandre Tessier in chapter IV.2. An excellent example of long-term evolution is provided by the animated map of the United States postal service in Tessier's figure 14. Two short-term postal chronologies track the time taken for a letter to travel from one place to another: the prescriptive chronology, based on regulations governing mature postal systems, indicates the time that *should* be required in principle, while a descriptive chronology, based on recorded dates of sending and receipt, indicates how long it actually takes in practice. A rather different cartographic representation of chronology is provided by what Tessier calls 'Minard diagrams' (such as that reproduced in his figure 12), which are based on the time taken to reach a whole network of places from a given point.

Mapping systems of learned exchange on the basis of correspondence metadata, however, is a perilous as well as laborious enterprise, because postal communication is so informal, capricious, episodic, and vulnerable to vicissitudes. In any case, the republic of letters was created by far more than just letters and needs to be charted with reference to the voluminous documentation of other forms of exchange as well. As discussed further in chapter IV.4 below, university records provide a rich and stable basis for mapping the evolution of long-term and large-scale patterns of intellectual exchange. An obvious point of departure is provided by the matriculation records kept meticulously in many universities in this period, which provide a mass of reasonably homogeneous and representative data running along a chronological axis.[15]

An obvious a case study for exploring the utility of matriculation data for tracking change over time is provided by the most tumultuous moment in the history of the densest concentration of universities in early modern Europe: namely, the universities of the Holy Roman Empire in the midst of the Thirty Years' War (1618−48).[16] Within the Empire, each prince and imperial free city wanted their own institutions of higher education, adapted to serving their political, economic, and religious needs without exporting students to enrich their neighbors by studying elsewhere. The result was a proliferation of university and sub-university institutions without parallel in Europe (see ch. IV.4, figures 1–2), which was mir-

[15] For a useful guide to this genre of sources, see Matthias Asche and Susanne Häcker, 'Matrikeln', in Ulrich Rasche, ed., *Quellen zur frühneuzeitlichen Universitätsgeschichte. Typen—Bestände—Forschungsperspektiven* (Wiesbaden: Harrassowitz, 2011), 243–67.

[16] For full documentation of the following discussion, see Howard Hotson, 'Catchment Areas and Killing Fields: Towards an Intellectual Geography of the Thirty Years' War', in Peter Meusburger, Michael Heffernan, and Laura Suarsana, eds., *Geographies of the University* (Dordrecht: Springer, 2018), 135–92, see https://doi.org/10.1007/978-3-319-75593-9_4.

rored by the rapid and sustained growth in student numbers. Between 1540 and 1620, student numbers grew eightfold in eight decades, placing the universities of the Empire at the centre of a gigantic catchment area which extended from Scotland via Scandinavia to the Baltic, south through Poland-Lithuania to Hungary-Transylvania while also attracting students from the Swiss Confederation to the south.

The Thirty Years' War, struck central Europe at the very apex of this boom, raising a long list of questions which have never been properly answered at either the domestic or the international level. How did the chronology of academic destruction and recovery unfold over the course of three decades? Can its effects on individual institutions be grouped together to understand its broader impact on whole regions and confessions? How did the disruption at the center of this huge catchment area affect the long-term development of neighboring university systems, the shifting patterns of international academic migration between them, and the intellectual influences communicated by them?

These questions can also be used to test a methodological thesis: namely, that matriculation data allow the impact of the war to be quantified in meaningful fashion, dated with some chronological precision, and therefore analysed comparatively between institutions in a manner which can produce sound generalisations regarding the differentiated impact of the war on whole confessions and regions. Investigating this thesis also helps illustrate the limitations of standard means of visualizing chronology when dealing with such complex data and the challenge of devising more suitable forms of data interaction design.

For illustrating the impact of the war on matriculations at a single university, the familiar histogram proves its value (figure 8). An obvious illustration is provided by Heidelberg, the first German university to be affected by the war. In the opening decades of the seventeenth century, Heidelberg was the most international university in the Reformed world. Yet, by accepting the crown of St Wenceslas from the rebellious Bohemian estates in 1619, the Elector Palatine Friedrich V provoked a massive Catholic reprisal which ruined Heidelberg for a generation. As early as 1620, Heidelberg's rate of matriculation was cut in half, as Spanish forces occupied the far bank of the Rhine; it halved again in 1621, after Friedrich's forces were defeated at the Battle of the White Mountain, and again in 1622, by September of which year the city, the university, and the fabled electoral library were in Catholic hands. After Maximilian I of Bavaria was granted hereditary status as the new Elector Palatine, Heidelberg reopened briefly as a Catholic institution, until Heidelberg fell to the Swedish army under Gustavus Adolphus in 1632. Attempts to reopen the Protestant university were in turn aborted by the Swedish defeat at Nördlingen in 1634. In short, it was not until the Peace of Westphalia restored a fragment of Friedrich V's lands to his son that the new elector, Karl Ludwig, could refound the once great university in 1652 virtually *de novo*, but with only a fraction of the magnetism it had enjoyed before the war.

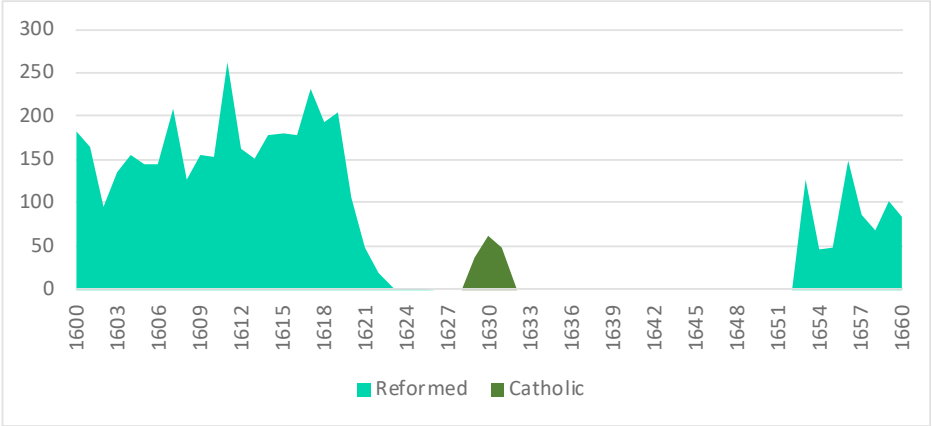


Figure 8: Matriculations in Heidelberg, 1600–1660

The general shape of this complicated narrative and most of the episodes within it are eloquently represented by a simple histogram of annual matriculations (figure 8). The problem is that there were over 30 other universities in the Holy Roman Empire in this period, for most of which the matriculation registers survive. Charting the fortunes of each of them in analogous fashion confirms the sensitivity of matriculation data as a barometer of military pressure. The difficulty is finding ways of aggregating the data from individual records in a manner which builds up similarly revealing impressions of the fortunes of clusters of universities.

One solution is to stack multiple histograms in the manner illustrated in Figure 9. Here Heidelberg (again in dark green), is embedded along with the other main Reformed institutions of higher education in the region. Marburg (represented by light green), having been Calvinised as recently as 1605, was regained by the Lutherans in the immediate aftermath of the sack of Heidelberg. The Reformed Landgraf of Hesse-Kassel established a new university to replace it in his *Residenzstadt* (in yellow), which remained very small until Marburg was restored to Hesse-Kassel in 1652, the same year in which Heidelberg reopened. The academy in Herborn (in blue), a university in all but name, was undermined by military occupation, plague, the departure of its international students, and restitution of the former monastic lands which provided finance. Just outside the boundaries if the Empire, Basel (in dark blue) was also affected by the disruption, not only of the immediate vicinity in 1634, but also of the academic trade route which had supplied it with international students. The mode of visualization employed in Figure 9 sacrifices the clarity with which these individual fortunes are depicted in order to tell the aggregate story in a clear and compelling way.

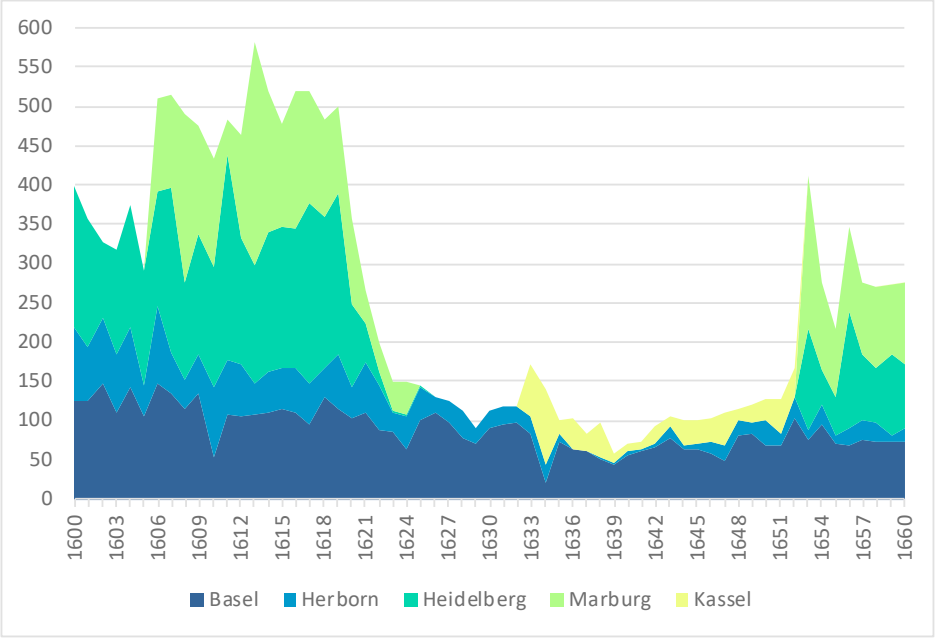


Figure 9: Matriculations in Reformed universities, 1600–1660

For relatively coherent data sets such as these, a further simplification is also revealing. In Figure 10, these same four German Reformed universities are plotted as a single light green line, which retains the same shape (aside from the absence of Kassel). The darker green line aggregates in similar fashion the combined annual enrollments of the three oldest Dutch universities: Leiden, Franeker, and Groningen. The two lines roughly mirror one another, suggesting that the stories of these seven institutions form a single, larger Reformed university system. The Dutch figures leap by over three quarters during the first five years of the war while the German Reformed numbers are cut by over two thirds. As the German Reformed line then gently settles to its low point during the second half of the war, the Dutch line also levels out, gradually peaking during its final decade. Enrollments in Leiden, Franeker, and Groningen fall sharply at the conclusion of the conflict, without losing much of their strength in the longer term, while the German ones bounce back with the re-establishment of Heidelberg and Marburg in 1652, without regaining their pre-war level. Meanwhile, between 1621 and 1652 the total number of annual matriculations in this system as a whole (indicated by the thin blue line) fluctuates within a very narrow band.[17]

[17] The main disruption to this system comes rather from Utrecht, which was raised to university status in 1636, and grew strongly after 1643, adding well over one hundred matriculations per year by the end of the war. But the new university's matriculation register is very unreliable for this period, averaging only five enrolments per year in some quinquennia and nearly 200 in others. Its arrival also

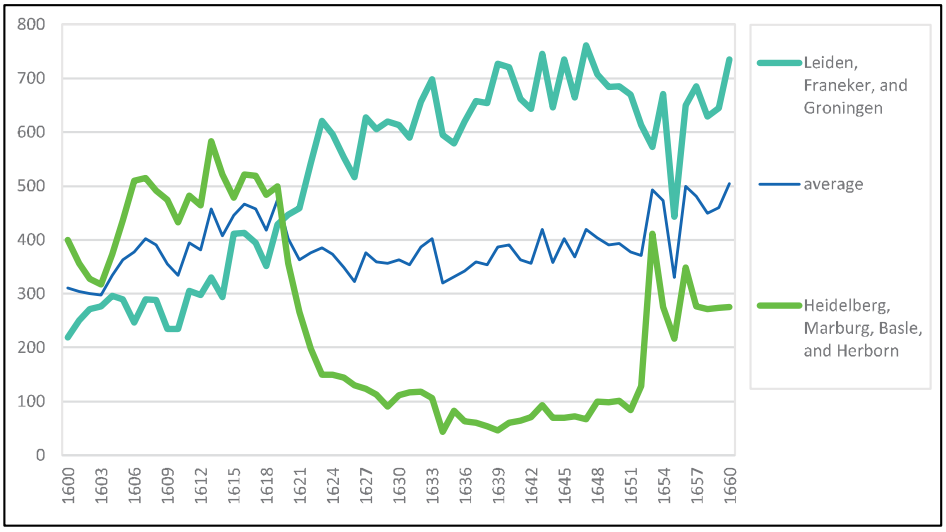


Figure 10: Matriculations in German Reformed and Dutch universities, 1600–1660

In pursuit of still more general impressions, further simplification is required. One option is illustrated by figure 11. In order to examine the relative growth or decline of different clusters of universities, this graph uses the pre-war size of each university as a baseline. More specifically, this graph express the aggregate matriculations of each cluster of universities, not in absolute terms, but as a percentage of their level during the five-year period 1616−1620. Matriculation levels below the 100 percent line are smaller than they were in 1616−1620, while those above the line are larger.

Replotted in this way, the collapse of German Reformed matriculations (not counting the bi-confessional Frankfurt an der Oder but including Basel) follows a trajectory reminiscent of the previous two figures. The main difference is that the use of five-year intervals softens both the precipitous decline in the 1620s and the meteoric if partial recovery after 1652.

heightens the need to distinguish matriculation rates from student numbers, since individual students increasingly matriculated in more than one Dutch university (as they had done previously in the German Reformed university system in its heyday).

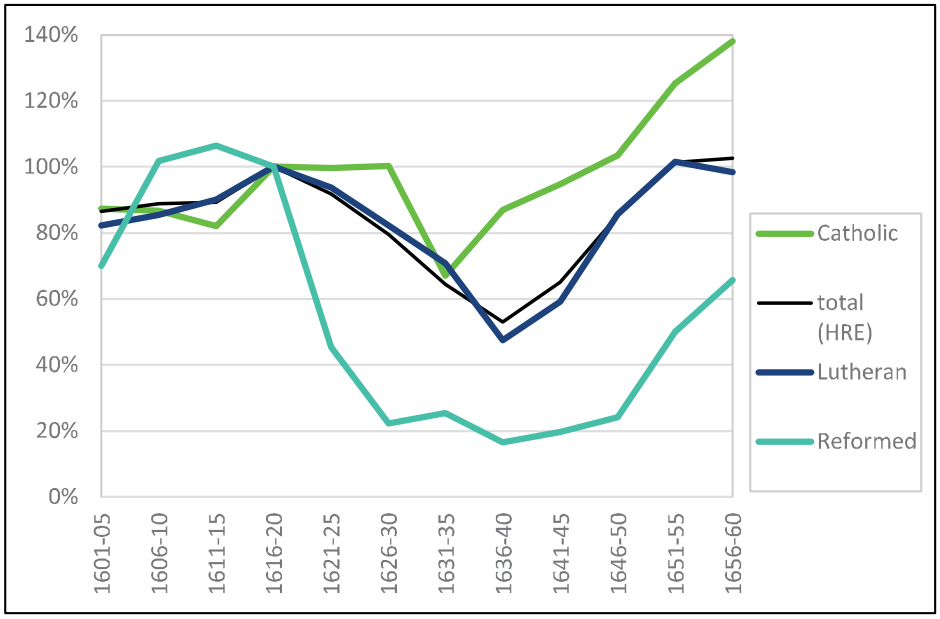


Figure 11: Three main confessions within the Holy Roman Empire compared. Quinquennial percentages of matriculations in 1616–1620

Catholic enrolments within the Empire provide a stark contrast. They were essentially unaffected by the war until 1630, when the invasion of the Swedish armies under Gustavus Adolphus differentiated their fortunes markedly. Five Catholic universities in the path of this onslaught – Würzburg, Mainz, Ingolstadt, Dillingen, and Freiburg im Breisgau – saw their numbers plummet in 1630–31, after which they recovered slowly. Meanwhile, the gentle growth of Vienna, Graz, and Salzburg accelerated, as students fled the warzone to areas free of maurading armies. The relatively sheltered northwestern corner of the Empire fell between these two extremes: Cologne's gentle growth before and after the war was interrupted only by a period of stasis during it, and Leuven was almost unaffected. All told, aside from the sudden shock of the early 1630s, Catholic universities in the relative safe havens to the northwest and southeast managed to absorb the students displaced from the southwest, leaving overall numbers surprisingly unaffected by the war. Finishing the war marginally stronger than they began it, the Catholic figures for the Empire as whole then rose sharply during in the first post-war decade to a level 38 percent higher than their status ante bellum.[18]

[18] Due to its large and relatively stable size, including Leuven flattens the curve (after 1616–20) without transforming its overall shape. The missing data for Douai would disrupt this curve, since the university was radically affected by a French invasion from the south.

In the northeastern quarter of the Empire where Lutheran universities predominated, the impact of war was felt in an even more differentiated fashion, as the main theatre shifted from place to place. Helmstedt suffered worst during the Lower Saxon phase of the war; Rostock and Greifswald during the Swedish landing; the giant Saxon universities after the failure of the Peace of Prague; and Frankfurt an der Oder was hammered twice: once in the latter 1620s and again in the latter 1630s. Due to the huge size of these northeastern universities, their aggregated matriculations closely track the trajectory of the Empire as a whole. This means that the decline of this region was far more gradual than that of the southeast and its recovery in the final years of the war was more rapid but also only partial.

Two significant methodological results can be derived from this brief survey. On the positive side, matriculation registers within the Holy Roman Empire appear to provide a sensitive barometer of the pressure of military events on local universities and on the confessions organized around them. A relatively small and easily assembled data set – containing only 2160 data points over a sixty-year period – provides a surprisingly revealing impression of the impact of three decades of conflict across most of the Empire. On the negative side, standard modes of visualizing time are not really suited to tracking the changes implicit even within this relatively small data set. Simple spreadsheets and the graphs generated by them reveal the shape of simple data series with ease; but when dozens of different data series are involved, the researcher needs to be able to toggle easily between many different views in order to form an accurate overall impression. The primary difficulty in this case is not that this humanistic data is highly nuanced, uncertain, or incomplete: merely that there is too much of it to be readily captured by off-the-shelf visualization tools.

Such tools will be even less well adapted to exploiting the full potential of hundreds of thousands of individual matriculation records. Even the most basic of these records typically include the matriculant's place of origin as well as name and date. In some cases, this is supplemented by other information, including age, social status, and subject of study. The indexes of many of the German registers translate both surnames and place names from Latin to German, thereby providing the basis for comprehensive, bi-lingual authority files for university-educated people and the places from which they came. Simple algorithms can provisionally link records of the same student matriculating in multiple universities during the course of a *peregrinatio academica*, with the process of inference and degree of certainty recorded on the system. For the late medieval period, a major collaborative project has been supplementing such basic data with further archival records to create a *Repertorium Academicum Germanicum*, a detailed prosopography of all the graduated scholars of the Holy Roman Empire between 1250 and 1550. The result, published as an online database and atlas, will provide comprehensive 'who's who' of late

medieval scholars in the region.[19] To date, nothing similar exists for the post-Reformation period.

Once coherent bodies of data have been assembled, they will need to be analysed in many different ways. Analyses of the origins of students at individual universities merely requires the digitization of a single register. Analyzing the records of competing clusters of universities will show how catchment areas wax and wane in response to political, confessional, and military events, as well as the foundation of competing institutions in the region. Alternatively, the destinations of students from an individual city, territory, or region could be displayed, in order to understand how these shifted over time. A third data view could reconstruct the routes followed by students who visited more than one university in the course of their *peregrinatio academica*, revealing how academic trade routes shifted, in some cases dramatically, as military conflict moved from one theatre to another. Those matriculation registers which systematically record social status, age, or subject of study will allow even more complicated, multi-dimensional analysis. Such a data set would provide the starting point for multiple, comparative studies of academic mobility, such as that undertaken by Mikkel Munthe Jensen for Scandinavian professors in the eighteenth century (ch. IV.4 below).

Understanding the movements of hundreds of thousands of students between thousands of places of origin and dozens of different institutions against the background of complicated physical, political, and confessional geography and constantly changing military events will require a much richer variety of interactive, dynamic, animated, multi-dimensional, full color visualizations designed to allow both expert and non-expert users to explore all the dimensions of the data at a variety of different tempos and scales. This represents an ideal arena for future collaboration between historians, data analyses, and experts in data interaction design.

[19] Rainer C. Schwinges, 'Das Repertorium Academicum Germanicum (RAG). Ein digitales Forschungsvorhaben zur Geschichte der Gelehrten des alten Reiches (1250-1550)', in Rüdiger vom Bruch, Martin Kintzinger, Oliver Auge, and Swantje Piotrowski, eds., *Professorenkataloge 2.0. Ansätze und Perspektiven webbasierter Forschung in der gegenwärtigen Universitäts- und Wissensgeschichte* (Stuttgart: Steiner, 2015), 215–32; and most recently: Kaspar Gubler and Rainer C. Schwinges, eds., *Gelehrte Lebenswelten im 15. und 16. Jahrhundert* (Zürich: vdf Hochschulverlag AG an der ETH Zürich, 2018): http://www.rag-online.org/, accessed 20/03/2019.

4 Textual Data

Dirk van Miert

The chronologies discussed above relate mostly to metadata: to letter records, to bibliographical records, to prosopographical data, and to data on the origin and development of postal systems themselves. Yet letters are designed to convey messages primarily in textual form; and this opens up a vast new realm for investigation; or rather it positions the traditional preoccupation of scholarship with the texts of letters in a new framework, and potentially supplies it with new tools and approaches. Chronologies of the topics contained in texts are a traditional focus of conceptual history. What large digital corpora and text-mining techniques offer is automated means of determining the changing frequency with which certain terms, topics, or names were mentioned in the available letters (see also ch. IV.6).

The potential of this approach has recently been illustrated by a comparison of the use of certain terms and their changing semantic fields in the works of Kant, Fichte, and Schelling. For instance, in Kant's works the word 'experience' (*Erfahrung*) is associated with 'matter', 'space', and 'object', whereas Fichte used the term more frequently in relation to 'nature' and 'essence'. Schelling is the first to use the word 'experience' in relation to 'empiricism' (*Empirismus*), a term that does not occur in the semantic fields of 'experience' in the texts of Kant and Fichte. The overall suggestion is that the notion of 'empiricism' gains ground only in Schelling's philosophy.[20] There are several caveats to such conclusions: these results are based on the use of the available texts, which include other texts as well as correspondence. Moreover, these texts are not tied to dates, so the chronological dimension of this comparison is a crude one, based on the simple fact that Kant was the oldest of the three philosophers and Schelling the youngest.

Such trajectories can be visualized in several conventional ways, such as histograms of occurrences per year; but the DensityDesign Research Lab in Milan has developed several more appealing visualizations. One of these is the 'streamgraph', which traces the chronology of the frequency of key concepts in Kant's work. Figure 12 gives a fragment of this visualization, which is developed through Minerva, 'a web tool for supporting philosophical historiography research'.[21]

[20] Tom Giesbers, Timmy de Goeij, Daniel Meijer, Dirk van Miert, Peter Sperber, and Paul Ziche, 'Mining for Associated Words in Philosophical Texts', *Schelling-Studien. Internationale Zeitschrift zur klassischen deutschen Philosophie* 2 (2014): 215–31, at 221.

[21] Paolo Ciuccarelli and Valerio Pellegrini, 'Minerva – Data visualization to support the interpretation of Kant's work', posted on 6 August 2013 at https://densitydesign.org/2013/08/minerva-data-visualization-to-support-the-interpretation-of-kants-work/, accessed 20/03/2019.

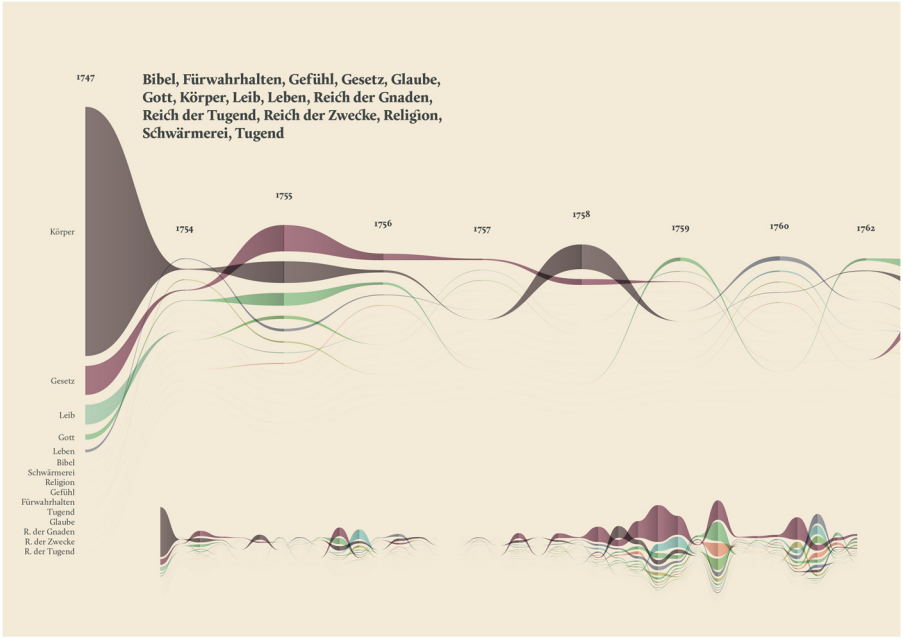


Figure 12: Streamgraph of several key concepts in the oeuvre of Immanuel Kant

As Paolo Ciuccarelli and Vallerio Pellegrini point out, 'The streamgraph has been figured out as the most effective visual model, since its ability to show the evolution of lemmas (in quantitative terms) across the works (and the time) and, at the same time, to compare them work by work'.

Whereas such visualizations work well for individual concepts, the chronology of semantic fields would require a more complex graph, perhaps with changing colours indicating the entry and exit of certain concepts in the set of words associated with a key term. It is imaginable, for example, that the word *merita* (services) is a loyal satellite to the republic of letters, but the history of the word 'God' in the semantic field of the 'republic of letters' shows fluctuations across time.

With precedents such as these in mind, the ERC *Consolidator* project SKILLNET has undertaken to trace the chronology of the concept 'republic of letters' itself. When, where, and by whom was this phrase used most frequently? Applying conceptual history to the term 'republic of letters' promises to reveal when the idea of a virtual transnational scholarly community appealed to individual scholars and perhaps the trajectory of its rise and fall within the broader intellectual community as a whole. The prospects for answering that question with reference to correspondence are particularly appealing: if the individual texts constituting the corpus to be mined have been assigned metadata including years (or even specific dates, in the case of letters), places, and authors, the overall geographical trajectory of the 'republic of letters' can be traced throughout time, and the chronology of its use

within specific time frames can be mapped, for example by considering one author's use of the term throughout his career. Moreover, such questions can be answered in absolute terms and also in relative ones, compensating for the varying numbers of letters available. Different vernacular translations of the term and the history of their use can be compared, and the interrelation of terms can be measured.

Like most fresh scholarly undertakings, however, such an inquiry is more difficult that it initially seems. Two challenges currently beset such a project. The first is to create a clean corpus of texts provided with rich metadata. Even in the case of one such corpus, the *ePistolarium*, it proved impossible to chart the career of the 'republic of letters', since the *ePistolarium* was not built with an eye to conducting conceptual historical research. The first step is therefore to repurpose the corpus of the *ePistolarium* to do just that, and then to enrich it with other readily available digitized texts of letters, or even of printed treatises and scholarly journals.

A second challenge is that a concept that consists, not of one word, but of at least two and possibly three, is less easy to locate than one would anticipate. Taking into account spelling variants, word order, and different endings, the Latin phrase *respublica litteraria* theoretically could appear in at least 192 different variants. The best-known form of the expression is no doubt *respublica litteraria* (learned republic). Sometimes the variant *respublica litterarum* (republic of letters) is mentioned in secondary literature: this form is common in English, Dutch, Spanish, and Italian, but far less so in Latin. However, letter-writers occasionally speak of a *respublica litteratorum* (republic of the learned) or even a *respublica litterata* (literate republic or republic versed in letters): this variant is suggested by the variant term, *orbis literatus* (the literate world), which also frequently occurs in the letters. Moverover, the word *respublica* can be split into two separate words, *res publica*, or abbreviated as *resp.*; in each variant, the radix *lit(t)era-* can occur with single or double 't' and the word order can be reversed. Fortunately, the theoretically possible formats *res literaria publica* and *publica res litteraria* do not appear in the *ePistolarium*.

This proliferation of variants renders the search for material for a conceptual history of this term difficult. Ideally, a machine should be able to deal with a complex combination of wildcards and double quotes, such as: "re*publica* lit*erar*" OR "lit*erar* re*publica*". Moreover, *respublica litteraria* is by no means the only Latin term used to capture this basic concept: within the *ePistolarium*, the phrase 'learned world' (*orbis litterarius, -tus, -torum*) occur eleven times, the collocation 'all the learned men' (*omnes litteratores*) another ten time; and both of these are susceptible of generating multiple variants in turn.

To compound difficulties further, the corpus of correspondence in the *ePistolarium* is multilingual; and this forces the researcher to translate the word into several vernaculars and negotiate the spelling variants which proliferate in the early modern period, such as Republic(k) with 'c' or 'k', or differences such as *Gelehrtenrepublik* and *gelehrte Republik*. In fact, the *ePistolarium* contains one instance of 'Republick of Letters' and one of 'Commonwealth of Learning' (despite the fact that

there are very few English letters in the corpus), seven usages of the phrase *république des lettres* , and one of the Italian *repubblica letteraria*. The most surprising result is the small number of French hits: 26.5 per cent of the corpus is in French (32.8 per cent in Latin). Although 37.1 per cent of the corpus is in Dutch, the term *republiek der letteren* does not feature in any of these Dutch letters, but the words *geleerde wereld*, learned world, do occur.

Fortunately, for the purposes of big data analysis, not every occurrence of a term must be detected in order to produce significant results. In the 20,020 letters in the *ePistolarium*, the majority take the form *respublica lit(t)eraria*; the format *respublica literatorum* is found only once, and the word *respublica* is never separated, at least not in combinations denoting the republic of letters. Surprisingly, the wellknown variant *respublica literarum* is used only once. The form *republica litteraria* occurs in two further letters, but those are in Italian, and constitute a spelling variant of *repubblica letteraria* (which occurs once, in another Italian letter by the same author). Note that some variants are neutralised by editorial decisions to expand the abbreviation *resp.* or to standardize the use either of single or of double 't'.

If all of these variations are included, only forty-four uses of the phrase were detected in the 20,020 letters in the *ePistolarium*. This number is unexpectedly small. It suggests that the idea of a republic of letters was not a vital concept for some of the people we like to regard as self-aware citizens of this virtual community. As in the case of published letter collections, one would want to organize the occurrences chronologically and by author, although with so few hits, an automated visualization is likely to overshoot the target. Hugo Grotius is a case in point. A universal scholar who corresponded in three languages, a widely travelled diplomat, an irenicist as well as a polemicist, and often compared with Erasmus, Grotius is typically regarded as an exemplary citizen of the republic of letters. Yet he used the concept only once, in an official letter to the States General in which he underscored his own services to the commonwealth of learning. In the correspondence of Joseph Scaliger, however (which is not included in the *ePistolarium*), the term 'republic of letters' occurs fifty-eight times between 1576 and 1609, evenly distributed between twenty-seven letters to Scaliger and thirty-one letters from Scaliger. Already, these two contrasting examples raise important questions. Was the concept more popular, for instance, around 1600 than around 1650? These few results suggest how much fundamental work remains to be done before we can understand the chronology of the republic of letters as an idea as well as a reality. In order to map the trajectories of key concepts in the history of learning through an automated chronological visualization of any kind, larger bodies of clean textual data and reliable metadata will be required.

IV.4 Prosopographies of the Republic of Letters

Howard Hotson, Thomas Wallnig, Mikkel Munthe Jensen, Gabriela Martínez, and Dagmar Mrozik

1 Introduction: Where to Begin?

Howard Hotson and Thomas Wallnig

A data-driven analysis of the republic of letters requires robust data models for the many different kinds of intellectual exchange which bound the commonwealth of learning together. It also requires tried and tested prosopographical models for describing the kinds of careers that sustained intellectual activity throughout this period. But developing models of sufficient detail and reliability is a difficult challenge which still awaits a satisfactory solution.

Part of the reason is that the structures which shaped learned careers, and the terminology in which those structures were described, varied greatly with time and space. Political offices provide an extreme case of this variability: the manner in which such offices are labelled and described, distinguished from one another, and arranged hierarchically varies enormously from one part of Europe to another, and within most European polities from the beginning of the early modern period in the later fifteenth century to its end in the later eighteenth. As a consequence, the diplomats and other political office-holders who played important roles in the republic of letters throughout this period do not provide an attractive point of departure for developing a durable model for capturing the career structure of a significant portion of early modern intellectuals. The same goes for 'gentlemen

virtuosi', that is, the independently wealthy and often noble amateurs so active in the intellectual networks of the later seventeenth century: the very lack of professions which provided the liberty for them to experiment with new forms of intellectual activity often lends their careers an amorphous quality difficult to capture as structured data.[1] The limiting case is provided by learned women, who often lacked access to formal educational institutions as well as to professional careers.

More appropriate points of departure could be provided by those careers that conform to four key characteristics: namely, careers which (1) are relatively stable in structure throughout the early modern period (c. 1450–1800); (2) are relatively uniform from one end of Europe to the other (north, south, east, and west); (3) are played out within institutions that document their activities meticulously; and (4) unfold especially within institutions that have preserved this documentation in large quantities and published some of that documentation in easily accessible printed form.

To meet all four of these criteria, such careers must be structured by institutions of a sort founded within the Middle Ages, which spread across Europe in relatively standard forms, and endured with relatively little change throughout the early modern period. Such institutions are not, almost by definition, the most forward-looking and innovative of the early modern period and are not necessarily central to the way in which the republic of letters functioned, conceived of itself, or has been described in the literature. But they did help to provide a stable foundation for much of the intellectual work of the period, and provide the best conditions for developing data models which will not need to be developed incrementally through trial and error and repeatedly overhauled before they are fit for purpose. Elements of those models, the experience gained in creating them, and the culture established when multiple projects are using them, may then help show the way forward for developing models for careers which lack some of these four key features.

This chapter concentrates primarily on two institutions of this kind: universities and religious orders, broadly understood. After establishing their suitability as points of departure and showing how they might provide new bases for data-driven scholarship on the republic of letters, the chapter concludes by considering how prosopographical data models might be extended to other communities less firmly anchored by institutions of this kind.

[1] Stephen Shapin, 'A Scholar and a Gentleman: The Problematic Identity of the Scientific Practitioner in Early Modern England', *History of Science* 29:3 (1991): 279–327, see https://doi.org/10.1177/007327539102900303; id., 'The Man of Science', in Katharine Park and Lorraine Daston, eds., *Early Modern Science* (Cambridge: Cambridge University Press, 2006), 179–91; Michael Hunter, 'The Scientific Community', *Science and Society in Restoration England* (Cambridge: Cambridge University Press, 1981), ch. 3; Craig Ashley Hanson, *The English Virtuoso: Art, Medicine, and Antiquarianism in the Age of Empiricism* (Chicago: University of Chicago Press, 2009); Mar Rey-Bueno and Miguel López-Pérez, eds., *The Gentleman, the Virtuoso, the Inquirer: Vincencio Juan de Lastanosa and the Art of Collecting in Early Modern Spain* (Cambridge: Scholars Publishing, 2008).

2 Modelling Academics

2.1 Universities as Points of Departure

Howard Hotson and Mikkel Munthe Jensen

Of all the learned institutions of early modern Europe, universities most readily fit the four criteria outlined above. The university is one of the most successful institutional innovations of the European Middle Ages. Already by 1500, nearly seventy universities were active simultaneously, the oldest of which was already 400 years old. Geographically, they were scattered across the European landscape relatively uniformly from the middle of the Iberian and Italian peninsulas northward to southern Scotland and Scandinavia (fig. 1).

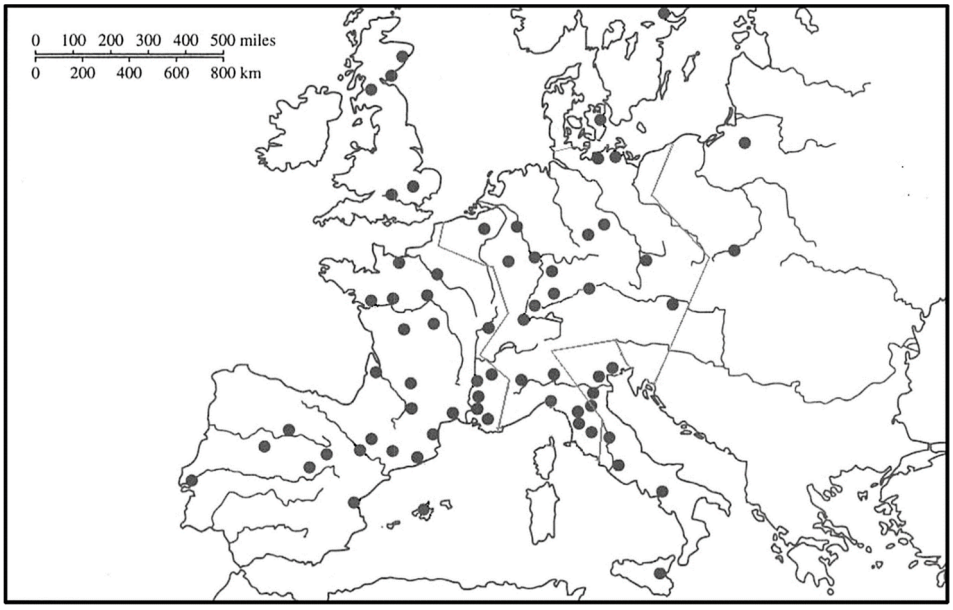


Figure 1: Universities active in 1500

The Reformation accelerated the rate of university foundations still further: by 1650 the number of universities in Europe had trebled and been joined, especially in the politically fragmented and confessionally contested heartland of central Europe, by a large number of immediately sub-university institutions as well (fig. 2). For nearly a century and a half, new foundations continued but at a markedly slower rate: between 1651 and 1800, forty-three universities and sixty-two sub-university institutions appeared. Finally, in the decades either side of 1800, the

number of universities in Europe actually fell: between 1793 and 1811, over fifty universities were suppressed, first in France and then in a rationalization of educational provision imposed across Napoleonic Europe.[2] The new institutions arising from the subsequent wave of university foundations in the nineteenth century increasingly deviated from the pattern set by their medieval and early modern forebears. So the demise of the *république des lettres* in the latter part of the eighteenth century coincides closely with a major watershed in European university history.

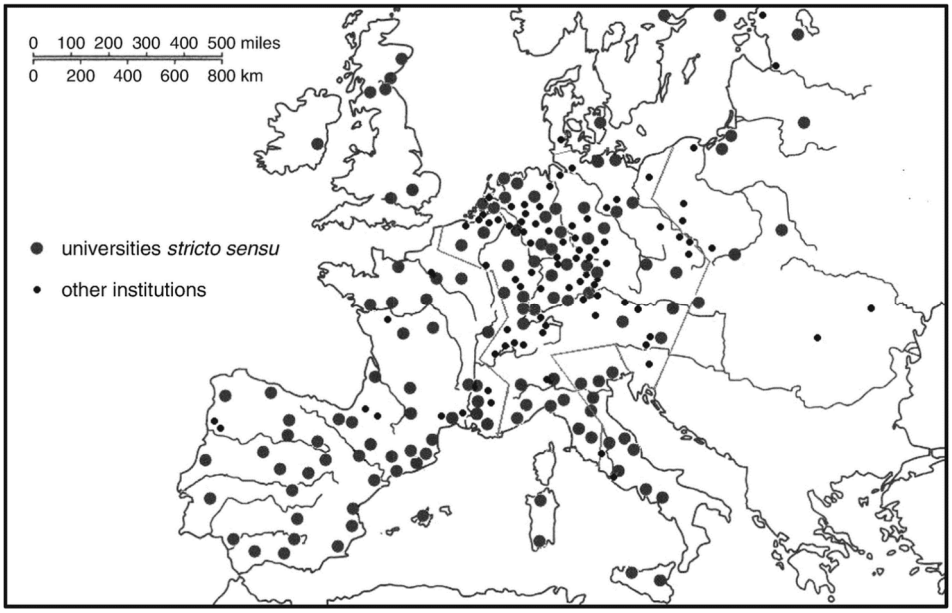


Figure 2: Universities and sub-university institutions active in 1650

As they spread and proliferated, these institutions remained surprisingly uniform in structure. The basis of this structure was the division of the university into four faculties: the three higher faculties of theology, law, and medicine, and the lower faculty of philosophy or arts.[3] Disciplinary subdivisions within these faculties also remained remarkably stable: even within the rather miscellaneous lower faculty,

[2] Map derived and data taken from Willem Frijhoff, 'Patterns', in Hilde de Ridder-Symoens, ed., *Universities in Early Modern Europe (1500−1800)*, vol. 2: *A History of the University in Europe* (Cambridge: Cambridge University Press, 1996), 43−110, at 80−9. Revised map published with permission of Willem Frijhoff and the Cambridge University Press.

[3] For university and faculty structures see Aleksander Gieysztor 'Management and Resources', in Hilde de Ridder-Symoens, ed., *Universities in the Middle Ages*, vol I: *A History of the University in Europe* (Cambridge: Cambridge University Press, 1992), 108–43; and Hilde de Ridder-Symoens, 'Management and Resources', in de Ridder-Symoens, ed., *Universities in Early Modern Europe*, 155–8. Not all universities possessed all four faculties.

which taught all the subjects propaedeutical to higher study, a basic Aristotelian division was maintained between theoretical philosophy (or *scientia*, including metaphysics and natural philosophy), practical philosophy (or *prudentia*, concentrating on ethics and politics), and productive philosophy (*ars*, including the seven liberal arts of grammar, rhetoric, logic, arithmetic, geometry, music, and astronomy), to which were added history and the sacred languages.[4] It was not until the institutionalization of new disciplines and more specialized subdisciplines in the later eighteenth century that the traditional disciplinary structure of the university began to change radically in a process that substantially intensified with the Humboltian reforms and the new German research universities in the early nineteenth century.[5]

This common disciplinary structure helped lend stability to the hierarchy of academic degrees as well. Although the requirements for academic degrees and ranks varied during the period and across the Continent, enough uniformity was maintained to create stable categories for a prosopography of academics. The three-level degree system – consisting of baccalaureate, licentiate, and doctorate – was recognized internationally (especially at the upper end), although the intersection of the hierarchies of degrees with that of faculties caused complications resolved differently in different institutions. Likewise, the main ranks of academic teachers retained a close family resemblance, rising from *docentus* and *adjunctus* to extraordinary and ordinary professors. The roles of university officers – including consistorials, faculty deans, rectors, and chancellors – were also widely replicated and mutually understood with few substantial variations. In short, the institutional matrix remained remarkably stable throughout the entire period in which the republic of letters flourished. This stability was one of the preconditions for the relatively 'frictionless' academic commerce of this period, and it is also the prerequisite for capturing this important province of the commonwealth of learning within a uniform data model.

Documentation with which to populate such a model is also abundant. Universities recorded many of their activities meticulously and have preserved this documentation in large quantities to the present day. Indeed, since the seventeenth century, these institutions have devoted considerable resources to publishing material both from and on their students and teachers, founding their own printing houses partly for this purpose. The most important of these sources for prosopographical purposes can be divided into institutional sources, scholarly productions, and ego-documents. Among institutional sources, matriculation registers and lecture catalogues are of particular value: the former provide structured data on the

[4] Olaf Pedersen, 'Tradition and Innovation', Wilhelm Schmidt-Biggemann, 'New Structures of Knowledge', and Laurence Brockliss 'Curricula', all in de Ridder-Symoens, ed., *Universities in Early Modern Europe*, 451–88, 489–530, 565–620.

[5] Walter Rüegg, 'Themes', in Walter Rüegg, ed., *Universities in the Nineteenth and Twentieth Centuries (1800–1945)*, vol. 3: *A History of the University in Europe* (Cambridge: Cambridge University Press, 2011), 3–20. See also William Clark, *Academic Charisma and the Origins of the Research University* (Chicago: University of Chicago Press, 2006).

broad base of the university hierarchy composed of students (typically including place of origin as well as date of matriculation, and sometimes also including indication of age, discipline, and social status), while the latter document a key activity at the peak of the pyramid enjoyed by professors. In addition, disputations and dissertations, mandatorily published in many institutions, offer concrete biographical information alongside abstract theses: their title pages are festooned with the authors' academic and political degrees, titles, and positions; their dedications often reveal the patrons who funded the students; and celebratory verses are sometimes included by student friends. Since cultural representation and institutional memory were also central to academic culture, large quantities of ego-documents survive as well: biographies and autobiographies in script and print, travel diaries and *alba amicorum*, eulogies, panegyrics, and funeral sermons in great number. Moreover, from the early seventeenth century, institutions began to compile biographical dictionaries of their professors; and with the advent of the *historia litteraria* towards the end of that century more comprehensive lexica of scholars were compiled which represent the major predecessors of modern national biographical dictionaries.[6]

The medieval origins, transcontinental spread, stable structure, and rich archival documentation of Europe's universities therefore provide an excellent basis for developing a robust and homogeneous framework for organizing detailed data on students and academics across the Continent. Fortunately, the work of creating a formal prosopographical data model need not begin from scratch. Many individual institutions are already constructing databases of their own former students, graduates, and professors. A major research project, the *Repertorium Academicum Germanicum*, is assembling prosopographical data on an estimated 60,000 masters or licentiates of arts and graduates of the three higher faculties within the Holy Roman Empire in the late medieval period (1250–1550).[7] At the uppermost level of generality, the *European Network on Digital Academic History*, also known as *Heloise*, is negotiating common data standards and structures within a community of projects collaboratively building open-access databases of students, graduates, and academics.[8]

Building on these efforts, a data model of academics could form a crucial part of a more general prosopographical model for the republic of letters because universities enjoyed a special relationship to the commonwealth of learning at the level both of the individual and of the collective. On the individual level, although

[6] For a good overview of the sources for university history, see Ulrich Rasche, ed., *Quellen zur frühneuzeitlichen Universitätsgeschichte: Typen, Bestände, Forschungsperspektiven* (Wiesbaden: Harrassowitz, 2011).

[7] For a recent overview of this project (www.rag-online.org), see Rainer Christoph Schwinges, 'The Repertorium Academicum Germanicum (RAG) and the Geography of German Universities and Academics (1350–1550)', in Peter Meusburger, Michael Heffernan, and Laura Suarsana, eds., *Geographies of the Universities* (Cham: Springer, 2018), 23–42, see https://doi.org/10.1007/978-3-319-75593-9_2.

[8] See http://heloisenetwork.eu, accessed 20/03/2019.

academics were not regarded as citizens of the republic of letters *ex officio*, most such citizens were university-educated alumni, that is, former members of the university community. The reason is obvious. Letters presuppose literacy. Learned letters normally presuppose formal education. Early modern Latin letters typically presuppose the kind of learning cultivated at university. Most full participants in the *respublica litteraria* had in fact attended university, and the first documentary trace to be located of the more obscure of such participants is often in a university matriculation register. Modelling academic interactions will therefore help to document the formative stages in the careers of most early modern men of learning, as well as the more extended careers pursued within the academy itself.

No less importantly, the European university system laid one of the foundations for the collective life of the republic of letters as well. The transnational republic of letters was an imagined community and citizenship within it was little more than a learned conceit. The transnational academic community, by contrast, was a legally recognized institution. From the twelfth century onwards, matriculation bestowed a specific form of citizenship which released the student from ordinary civil jurisdiction and subjected them instead to the jurisdiction of the local university. Moreover, this legal status was portable: it guaranteed the university student the freedom to travel from one institution to another in pursuit of learning, and guaranteed that academic degrees obtained at one institution were recognized in all the others.[9]

The fact that many of the individuals who personified the republic of letters between Erasmus and Voltaire defined themselves over against the universities should not disguise the fundamental nature of this debt. Long before the fiction of a republic of letters was first conceived, many generations of European men of learning had experienced the reality of life in the academic republic. The universities not only replenished the ranks of European learning, generation after generation: they also represented the lived experience of a legally recognized commonwealth of learning of which the *respublica litteraria* was an imaginary projection. Yet the concrete character of the academic republic – its instantiation in legal institutions located in specific places and requiring continuous injection of funding – also made it more vulnerable to direct, political manipulation than the more ethereal republic of letters. This, at least, is the impression created by the following case study of the Nordic universities in the eighteenth century, which shows how the prosopographical approach advocated here can be put into practice.

[9] Generally on matriculation in Europe see Rainer Christoph Schwinges 'Admission', in de Ridder-Symoens, ed., *Universities in the Middle Ages*, 171–94; and Maria Rosa de Simone 'Admission', in de Ridder-Symoens, ed., *Universities in Early Modern Europe*, 285–325. Specifically on academic citizenship, jurisdiction, rights, and privileges see also Friedrich Stein, *Die akademische Gerichtsbarkeit in Deutschland* (Leipzig: C. L. Hirschfeld, 1891), and the more recent work on the topic by Lukas Ruprecht Herbert, *Die akademische Gerichtsbarkeit der Universität Heidelberg – Rechtsprechung, Statuten und Gerichtsorganisation von der Gründung der Universität 1386 bis zum Ende der eigenständigen Gerichtsbarkeit 1867* (Heidelberg: heiBOOKS, 2018), see https://doi.org/10.11588/heibooks.348.481.

2.2 Charting the Travels of Nordic Academics via VIA (*Virtual Itineraries of Academics*)

Mikkel Munthe Jensen

Viewed over *la longue durée*, the story of the European university is in large part the history of a transition from a relatively international and European system to a set of independent national university systems. At the outset, when only a few universities existed, international travel was inevitable for many. In some disciplines, internationalism persisted into the early modern period: in 1500 one still travelled to Paris to cap off a theological education, to Bologna to study civil law, and to Padua (into the seventeenth century) to access the best of medical education. A huge arc of educationally peripheral countries with few large universities – ranging from Ireland and Scotland via Scandinavia to east-central Europe – continued to send their sons abroad for study well into the eighteenth century. These large-scale and long-term patterns provided the deep tectonic plates on which more superficial intellectual exchange often rested; but they will only be fully appreciated when large pools of homogeneous structured data can be readily processed with interactive digital tools.

A case in point is provided by the Scandinavian universities during the long eighteenth century.[10] Among the Nordic universities, there existed a widespread understanding and acceptance that their peripheral region relied on foreign academic experience and expertise.[11] A cornerstone in the Nordic universities' recruitment practices was therefore the emphasis on foreign academic experience, and in order to obtain it many, generous travel scholarships were granted to talented students.[12] But which continental universities were preferred? What were the reasons for those preferences? And to what extent did they shift over time?

In order to answer these basic questions, a substantial amount of structured data is required. For this purpose, data was collected on the 592 university professors who occupied an ordinary or extraordinary chair at one of the six Nordic universities (Copenhagen, Uppsala, Lund, Åbo, Greifswald, and Kiel) between 1700 and 1799. Of these, over half (290) had studied abroad, some more than once; and their total of 332 foreign voyages involved over 1,200 visits to learned

[10] This case study is drawn primarily from Mikkel Munthe Jensen, 'From Learned Cosmopolitanism to Scientific Inter-Nationalism – The Patriotic Transformation of Nordic Academia and Academic Culture during the Long Eighteenth Century', 2 vols., Doctoral Dissertation, European University Institute, Fiesole, 2018.

[11] Besides Jensen, 'From Learned Cosmopolitanism to Scientific Inter-Nationalism', vol. 1, 226ff, 249–50, see also Vello Helk, *Dansk-Norske studierejser*, vols. 1–2 (1536–1813), (Odense: Odense Universitetsforlag, 1987, 1991). On academic travel in general see also Justin Stagl, *A History of Curiosity: The Theory of Travel, 1550–1800* (Australia: Harwood Academic Publishers, 1995).

[12] On travel expenditures in Europe see Helk, *Dansk-Norske studierejser*, 61, 65–6. On the socio-economic background of Nordic academic travellers, see Jensen, 'From Learned Cosmopolitanism to Scientific Inter-Nationalism', vol. 1, 72–4.

cities. Data was assembled on each of these travellers under six main headings: 'Personal Data', 'Education', 'Travel and Academic Career', 'Information on the Father', and 'Other Professions, Titles and Memberships'. Each of these categories was further subdivided: under the section 'Education', for instance, the subcategory 'doctoral degrees' contained information on 'degree name', 'year', and 'institution'.

The most obvious way of analysing this data is by means of static maps and other visualizations. This method is adequate to reveal basic features of the intellectual geography of the Scandinavian academics (fig. 3), meaning, in this case, those who held chairs in Copenhagen, Uppsala, Lund, and Åbo.[13] The most obvious finding is the convergence of most travel on the Protestant regions of northern-central and western Europe, particularly in the newly reformed Saxon and Prussian universities and the fashionable Dutch ones, as well as the European scientific and cultural hot spots of Paris and London. This finding confirms the impression, gained by previous studies of individual academics, of the places in which most Scandinavians studied in this period; but only a comprehensive data set of this kind also allows us also to see the places in which Scandinavians did *not* study; and this finding already yields an important qualification to traditional notions of the *respublica litteraria*. While early eighteenth-century Scandinavians still imagined an unbounded republic of letters embracing the whole of Europe, in practice they restricted their academic travels primarily to one portion of it, avoiding the Iberian peninsula completely and neglecting France outside Paris, Italy south of Rome, and east-central Europe aside from Vienna, including the Protestant university in Königsberg, as well as the thriving Scottish universities.

This first finding naturally generates further questions. Most obviously, why was Scandinavian academic travel circumscribed in this way? Answering this question requires far more complicated analysis, which relates the temporal and spatial data on university travel to other, economic, confessional, and historical data. The socio-economic background of the academics is one crucial factor: judging from their fathers' professions, the majority originated from the middle and lower-middle classes, and finance clearly imposed one constraint on the length and distance of their educational travels. Confessional considerations were also important: many of them held scholarships that required them to study and be examined in specific and exclusively Protestant universities in Germany. Long-established tradition played a role as well: the safest and most dependable course was to follow in the footsteps of previous travellers, including the itineraries recorded in published guidebooks. In short, the intellectual space within which Scandinavian academics primarily travelled was limited for reasons of economy, confession, and tradition.[14]

[13] Map reproduced from Jensen, 'From Learned Cosmopolitanism to Scientific Inter-Nationalism', vol. 2, 112.

[14] Jensen, 'From Learned Cosmopolitanism to Scientific Inter-Nationalism', vol. 1, 235–41.

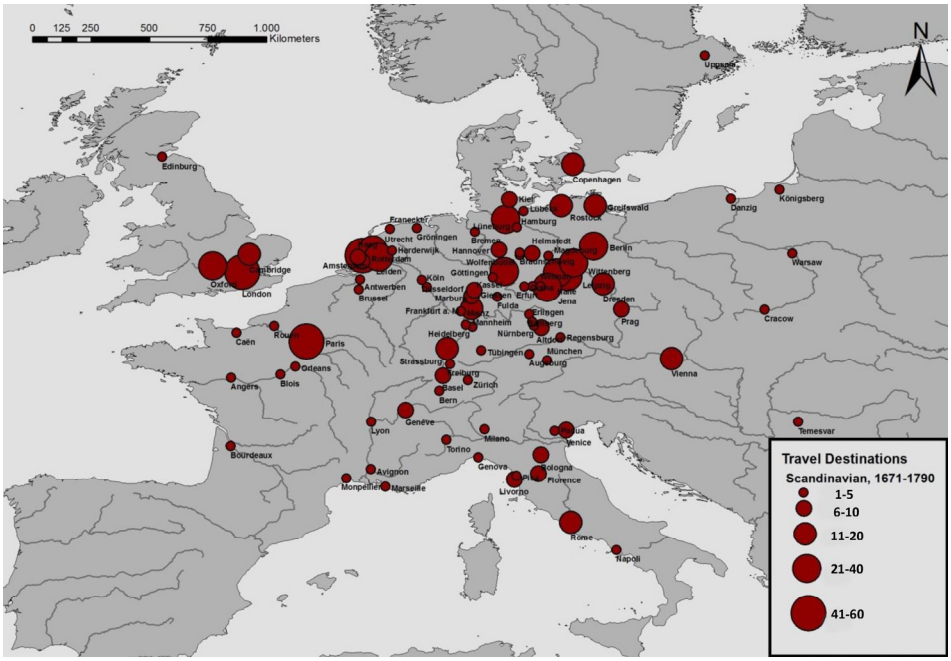


Figure 3: Academic travels by Scandinavian university professors, 1671–1790

Such multivariant analysis is extremely cumbersome to conduct using static visualizations, and this difficulty is increased still further when a temporal dimension is introduced, to determine whether these factors vary in importance over time. Moreover, laboriously produced visualizations are also limited by the researcher's prior assumptions and predefined questions. In order to explore the many dimensions of this data efficiently, new, interactive digital tools are required capable of performing Explorative Data Analysis (EDA).[15]

With this objective in mind, a group was formed in the first COST-funded design sprint in Como in the spring of 2016 in which the digital designer Marco Quaggiotto (Politecnico di Milano) and the historian Joëlle Weis (Université du Luxembourg) collaborated in using data on Scandinavian academics to develop the digital visualization and exploration tool known as VIA or *Virtual Itineraries of Academics*.[16] Unlike the static visualizations produced by more conventional means,

[15] For introductory articles on EDA, see David R. Brillinger, 'Data Analysis, Exploratory' in Bertrand Badie, Dirk Berg-Schlosser, and Leonardo Morlino, eds., *International Encyclopedia of Political Science* (Los Angeles: SAGE, 2011) and John T. Behrens, 'Principles and Procedures of Exploratory Data Analysis', *Psychological Methods* 2:2 (1997): 131–60, see https://doi.org/10.1037/1082-989X.2.2.131.

[16] For a more detailed elaboration on VIA, see Mikkel Munthe Jensen, Marco Quaggiotto, and Joëlle Weis, 'VIA – Virtual Itineraries of Academics. A Digital Exploration Tool for Early Modern Academic Travels', in *Jahrbuch der Österreichischen Gesellschaft zur Erforschung des 18. Jahrhunderts* (Vienna 2019). VIA is available at: http://knowledgecartography.org/via2/#travels, accessed 20/03/2019.

VIA allows the temporal and spatial dimensions of the academics' travels to be analysed visually along with 5,000 other data points regarding their ages, confessions, funding, university affiliations, faculties, degrees, and eventual chairs. As shown in figure 4, each of these three basic parameters (time, space, and person) is displayed in a separate panel of VIA's user interface. The geographical dimension is depicted on a map on the left of the screen, which displays the data either proportionally (in the manner of fig. 3) or as a heat map (as in fig. 4). Time is indicated in the bar chart below the map: the time-scale both indicates the number of documented students travelling in any given year and allows the user to expand or contract the period under analysis at will. The complex personal data is then displayed in the series of smaller statistical visualizations on the right half of the screen: histograms for continuous numerical data, stacked bar graphs for categorical properties with few values (where the interest is mostly on the relative values), and bar charts including search functionality for other categorical properties. The key virtue of this arrangement is that each of these three panels adjusts instantly to any change of parameters in any of the others, allowing the user to explore the interrelationships of various aspects of the data with a facility impossible using more conventional tools.

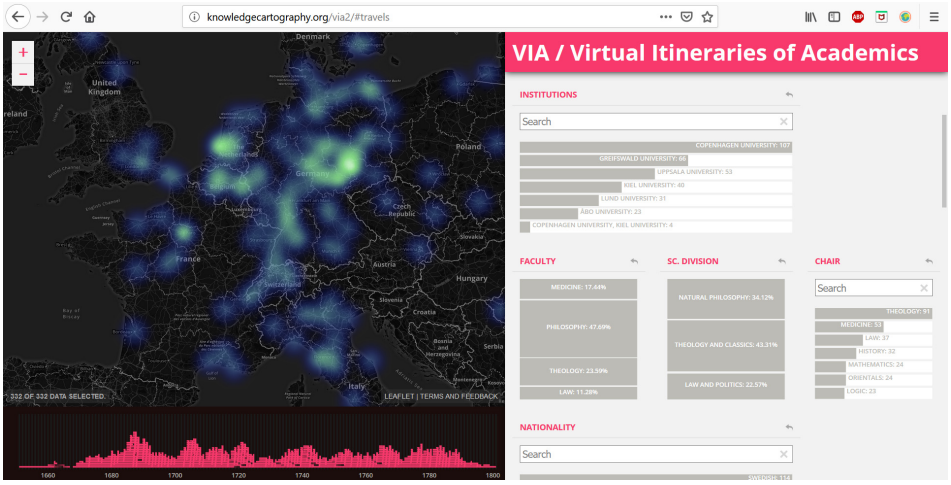


Figure 4: VIA's interface with its geographical, chronological, and prosopographical frames

Exploring the data in this manner readily reveals further patterns within the data set. One example is the differences between the professors in the properly Scandinavian universities (Copenhagen, Uppsala, Lund, and Åbo) and the two German universities (Greifswald and Kiel) subject to the Swedish and partly the Danish Crown in this period. Whereas the Scandinavian academics travelled for a long time (on average four years) and ranged more widely across the broad north-

western European space, the professors at Kiel and Greifswald rarely studied outside the German academic space to which they were more strongly and exclusively tied culturally, confessionally, and intellectually.

The Explorative Data Analysis facilitated by VIA also helps both to reveal long-term changes in these patterns over time and, still more importantly, to explain them. The basic trend is clear: toward the end of the eighteenth century, the frequency and duration of Scandinavian academic voyages abroad markedly declined. The causes at first sight appear innocent: as the Scandinavian universities expanded and the quality of instruction available within them improved, there was less need to travel abroad in search of state-of-the-art education. Yet this improvement was funded by state investment, undertaken to meet a steadily increasing perceived need, not merely for educated office-holders, but for 'home-grown' civil servants, fashioned to serve the fatherland. And these changes were culturally consolidated in turn by a kind of academic mercantilism, that is, by the new ideal of belonging exclusively to a single political community coupled with new notions of academic self-sufficiency, and buttressed by a growing patriotic critique of foreign academic travel. In these and other ways, the organic growth of Scandinavian universities to maturity was compounded by a much larger change: the patriotic transformation of academia from learned cosmopolitanism to scientific internationalism.

In this important respect, the limited case study of Scandinavian academic travel during the long eighteenth century has much to contribute to the broader question of the flourishing and demise of the republic of letters as a whole. For centuries, European academia had constituted a learned network of academic institutions, grounded in the shared idea of belonging to one learned world, upheld by similar legal and institutional structures, and enforced through common rituals, symbolism, and academic representation. During the course of the eighteenth century, however, the older and more cosmopolitan idea that the arts and sciences served to enhance human wellbeing *generally* was displaced by a new societal ideal of patriotic citizens, united in one 'corpus fatherland', offering educational and scientific 'utility for the fatherland' rather than the learned world as a whole. Henceforward, each university, each academia, was transformed into an academia for King and Country. The rise of the patriotic state thus not only disrupted the autonomy of the learned estate and invalidated the meaning of academic citizenship in favour of a patriotic affiliation to one political community; it also nationalized the universities themselves, as precedence was given to naturalized fellow citizens over even the most distinguished foreigners. International collaboration and exchange continued, but the cosmopolitan academia of the earlier period disintegrated into a multiplicity of international (in its literally meaning), competing, politicized *academiae* (in the plural), in which each academic and each institution served and followed its own king and fatherland.[17]

[17] Jensen, 'From Learned Cosmopolitanism to Scientific Inter-Nationalism', vol. 1.

Although inevitably limited in scope, this case study is intended to suggest something of the potential of data-driven analysis of this kind for providing a vivid and precise impression of crucial features of early modern international intellectual exchange with important implications for our understanding of the republic of letters. A homogeneous set of prosopographical data documenting the other peripheral regions of Europe together with the main centres of Europe's university landscape would be a desirable scholarly goal. As well as revealing one of the deepest and most significant underlying systems of transnational intellectual exchange, the capacity to analyse such data through even more sophisticated interfaces will reveal the substructure within which the more celebrated individuals, discoveries, and controversies played themselves out. Such would provide the precondition of a new historiography which is both more democratic, in revealing the experience and influence of more ordinary members of the educated elite, and more transnational, in making visible the massive intellectual transfers which have shaped the intellectual history of every European country.

3 Modelling Religious Orders

3.1 Religious Orders as Points of Departure

Thomas Wallnig

Like universities, religious communities were among the central protagonists of learning in pre-modern Europe. They were also among the first institutions to practise systematic collection of knowledge, specimens, and artefacts in the context of a globalizing world. Among the huge number and great variety of early modern religious orders, Benedictines, Franciscans, Dominicans, and Jesuits particularly demand attention, but it should be remembered that they were part of a larger framework of regular clergy which can be best described by the term 'religious communities'. Not all such communities were 'orders' in the canon law sense of the word: the Benedictines, for example, would not obtain that status before the 1890s; nor were all of them were monastic: the Jesuits were not, because they did not profess retreat from the world.

Almost all (male) religious communities engaged with learning, but only a few of them had intersection points with the republic of letters. This means that only a few of their members would shape their biographies in both religious and humanist terms and fashion themselves along the tropes of *scientia* and *virtus*.[18]

While the first religious communities date back to late Antiquity, a first effort at enforcing transregional structures took place around 800, and was related to the organization of power within the Carolingian Empire. The following centuries saw order filiations and secessions in the context of ongoing Church reforms. The mendicant orders (Franciscans and Dominicans), founded in the thirteenth century, first targeted urban audiences, and came to embrace academic learning. Another landmark was the foundation of the Society of Jesus (1534), which after the Council of Trent (1545–63) became the spearhead of the Counter-Reformation and Catholic reform. Through the Jesuits' quasi-monopoly on learning, their synthesis of humanist scholarship and Catholic mission became a model that most other communities sought to emulate to some degree. Actually, the intention of learned clergymen, and in particular monks, to participate in the republic of letters was determined by their interpretative claim on medieval history.

Religious orders or communities and the 'secular clergy' (that is, the episcopal hierarchy) share the practice of documenting the lives of their members by standardized writing practices. They also share this feature with the early modern universities, and their histories, to a certain extent, run parallel. These writing practices consist in applying the same standard form of documentation, and the same stable

[18] Thomas Wallnig, *Critical Monks. The German Benedictines, 1680–1740* (Leiden and Boston: Brill, 2019), 32–6, 76–86.

vocabulary, to each individual: profession books record date and place of birth and profession of the respective vows, as well as offices held during a lifetime (with a stable nomenclature). This standardized biography helps eliminate individual traits and creates the possibility for timeless models. This is as much a part of the communities' spiritual agenda as of their practices of documentation.

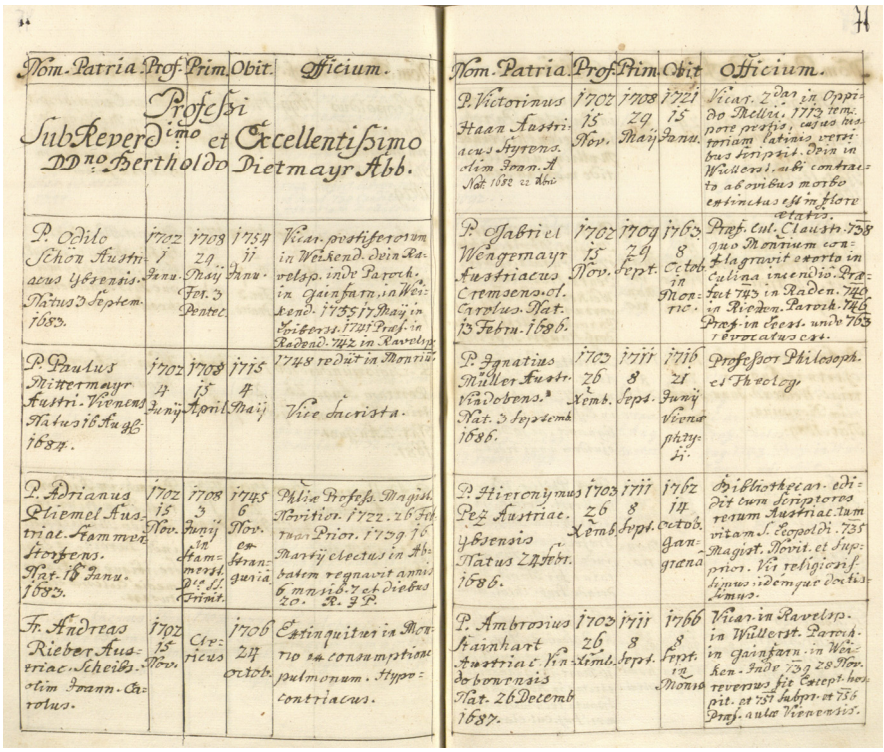


Figure 5: Page from the profession book of Melk Abbey, including the note on Hieronymus Pez: Stiftsbibliothek Melk, Cod. 493, 75v–76r

Orders (or communities) date back to different points in time, and they consequently follow different rhythms in their development. This also relates to the degree to which a community is organized on a transregional structure: Orthodox monasteries were highly autonomous, as were Benedictine communities. A Jesuit college, or a Franciscan house, instead, depended on their provincial and order superiors, the latter based in Rome. Thus each order had its own specific geography and chronology, while at the same time collecting standardized prosopographical data of its members: imagining the standardized individual as not depending on the cultural context of the 'world' was another intrinsic intention of monastic communities. A Jesuit should be recognizable as a Jesuit, irrespective of whether he is active in Mexico or Trnava.

In this sense, the life of a monk was structured according to an existing ontology, with some limited space for 'free text'. It remained variable how membership in the republic of letters could be formulated and integrated into this account: a more contemplative order was likely to downplay erudition, while those with scholarly and scientific claims tried to frame intellectual activity as a service to God. For outside observers, learned members of religious orders were thus often reinterpreted as exceptions to the rule of otherwise silly, lazy, and frivolous communities.

Some prominent examples document this potential conflict of two models of self: the polymath and Jesuit answer to the Scientific Revolution, Athanasius Kircher; the Minim natural philosopher, Marin Mersenne; or the Benedictine inventor of diplomatics, Jean Mabillon. In each of these cases, modelling of biographical data not only determines the shape of their interaction with the republic of letters, but also that with the other members of their order.

The relation between learning and religious life is thus in the data itself. In databases focusing on erudition, monks cannot always be recognized as such; and in databases on religious populations they cannot always be recognized as members of the republic of letters.[19]

However, the problem starts with the missing interoperability of those databases of religious personnel that are already in place. Even within the Catholic sphere, they do not yet make use of the possibilities offered by the transnational and chronologically stable nomenclature, because they come from different institutional and national backgrounds. Each of them represents a specific viewpoint: the *Bio-bibliographic Database of Monks in the Czech Lands in the Early Modern Age* is related to cataloguing efforts of the Czech National Library;[20] the *Germania Sacra* database builds on the respective printed book series, with its wide scope and consequent conceptual limitations;[21] and *Who Were the Nuns?* represents a clearly question-driven data set.[22]

First attempts have been made at creating an API to make these and other databases interoperable;[23] there are also ontologies that could be used for these purposes.[24] Results can be expected in due course.

One of the challenges in this process will be to find a vocabulary that maintains the multifaceted nomenclature of individual settings (e.g. the difference be-

[19] Thomas Wallnig, 'Cistercienser in der Res Publica Literaria. Überlegungen zu Datenmodellierung', in *Analecta Cisterciensia* 68 (2018): 299–312.

[20] See https://www.lib.cas.cz/dh/en/db/bio-bibliographic-database-of-monks-in-the-czech-lands-in-the-early-modern-age/, accessed 20/03/2019.

[21] See http://personendatenbank.germania-sacra.de/, accessed 20/03/2019.

[22] See https://wwtn.history.qmul.ac.uk/search/howto.html, accessed 20/03/2019.

[23] See https://github.com/GVogeler/prosopogrAPhI, and also https://pezworkshopdotorg.files.wordpress.com/2017/01/2017-02–21_chc.pdf, accessed 20/03/2019. A project (*Nuns and Monks. Prosopographical Interfaces*) starting in 2019 at the University of Vienna will continue working in that direction.

[24] See http://symogih.org/, accessed 20/03/2019.

tween the use of the word 'prior' in the Carthusian and Benedictine contexts) while summing them up in generic, yet historically grounded meta-categories (like 'superior', 'parish priest', or, indeed, 'erudite').

Once such interoperability is achieved, it will become possible to assess the interaction between religion and learning across the alleged boundary of the year 1500, and across all religions and denominations. Part of this will be a comparison of the learned portfolios of different Catholic religious communities. An example of how to approach such a study is given in the following section, in which 'mobile' Jesuit scientists are compared to their 'stable' Benedictine counterparts in the context of the 'Jesuit Science Network'.[25]

3.2 The Jesuit Science Network[26]

Dagmar Mrozik

The underlying research interest of the Jesuit Science Network (JSN) concerns the *who*, *what*, *where*, and *when* of Jesuit science, i.e. the biographical information of those members of the Society of Jesus who taught, experimented, or wrote about what we now call 'early modern sciences'. With existing research by Steven Harris from the late 1980s suggesting that the group of interest counts about 1,600 scholars,[27] a computer-based prosopographical approach was necessary in order to deal with the expected amount of data. The *Person Data Repository*, a now unfortunately defunct research project at the Berlin-Brandenburg Academy of Sciences and Humanities, provided the desired digital infrastructure and the corresponding tools.[28] As for the data itself, Carlos Sommervogel's nine-volume *Bibliothèque de la Compagnie de Jésus* (1890–1932),[29] the standard reference for Jesuit research, also quite naturally constitutes the central source of the JSN, and is expanded by Charles O'Neill's four-volume *Diccionario histórico de la Compañía de Jesús* (2001)[30] as well as several smaller, regionally focused sources.[31]

[25] Dagmar Mrozik, 'The Jesuit Science Network. A Digital Prosopography on Jesuit Scholars in the Early Modern Sciences', Doctoral dissertation, Bergische Universität Wuppertal, 2018. See http://elpub.bib.uni-wuppertal.de/servlets/DocumentServlet?id=8981, accessed 20/03/2019.

[26] Work for this section was supported by a COST STSM for Dagmar Mrozik entitled 'Digital Prosopographies of Religious Orders in Early Modern Europe', see http://www.republicofletters.net/wp-content/uploads/2017/03/STSM_MrozikReport.pdf, accessed 20/03/2019.

[27] The JSN, which is partially oriented by Harris's work but uses a different methodology and pursues a different goal, covers about 1,000 scholars. See ibid., 82 and 113, also Steven Harris, 'Jesuit Ideology & Jesuit Science: Scientific Activity in the Society of Jesus, 1540–1773', PhD Dissertation, University of Wisconsin-Madison, 1988, 139.

[28] See http://pdr.bbaw.de/english, accessed 20/03/2019.

[29] Carlos Sommervogel et al., eds., *Bibliothèque de la Compagnie de Jésus*, vols. I–XII (1890–1932; Louvain: Éditions de la Bibliothèque S.J.: Collège Philosophique et Théologique, repr. 1960).

[30] Charles O'Neill et al., eds., *Diccionario histórico de la Compañía de Jesús*, vols. I–IV (Madrid: Universidad Pontifica Comillas, 2001).

[31] I use *source* in terms of *data source*, not *historical source*.

Taking into account all the requirements, possibilities, and limitations of the research interest, the infrastructure, and the sources, the biographical information to be collected was structured as follows:

Name	First name, last name, VIAF ID
Biograpical data	Date and place of birth, date and place of death, entry in the order, resignation, expulsion
Education	Date, place, subject
Career	Date, place, subject, occupation
Miscellaneous	Relations, true miscellaneous

Figure 6: Information to be collected

The resulting tabular CVs therefore focus on the scholars' recorded activities and do not consider their publications. Using the example of the Jesuit scholar Stanislav Vydra (1741–1804), figure 7 shows how such a CV is implemented on the project website jesuitscience.net.[32] In addition to the biographical data displayed on the left, Vydra's data sheet[33] also features a short summary of his subjects and relations on the right, a map view underneath, and a bibliography of the corresponding sources at the very bottom.

[32] See http://jesuitscience.net, accessed 20/03/2019.

[33] The data sheet contains all the information collected about a person and can be accessed via a permanent URL, here http://jesuitscience.net/p/100, accessed 20/03/2019. It is also possible to download the data as XML or JSON.

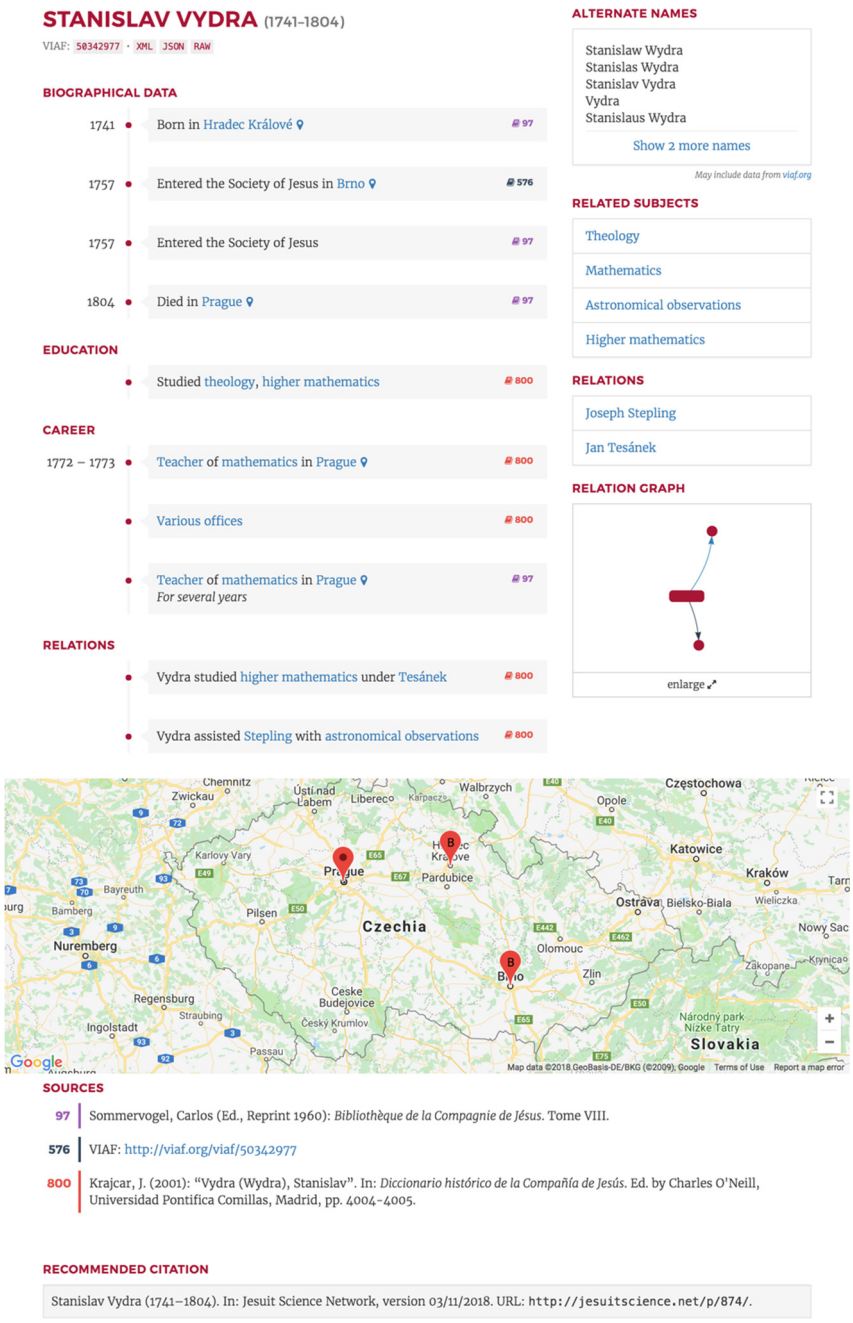


Figure 7: JSN data sheet for Stanislav Vydra (1741–1804)

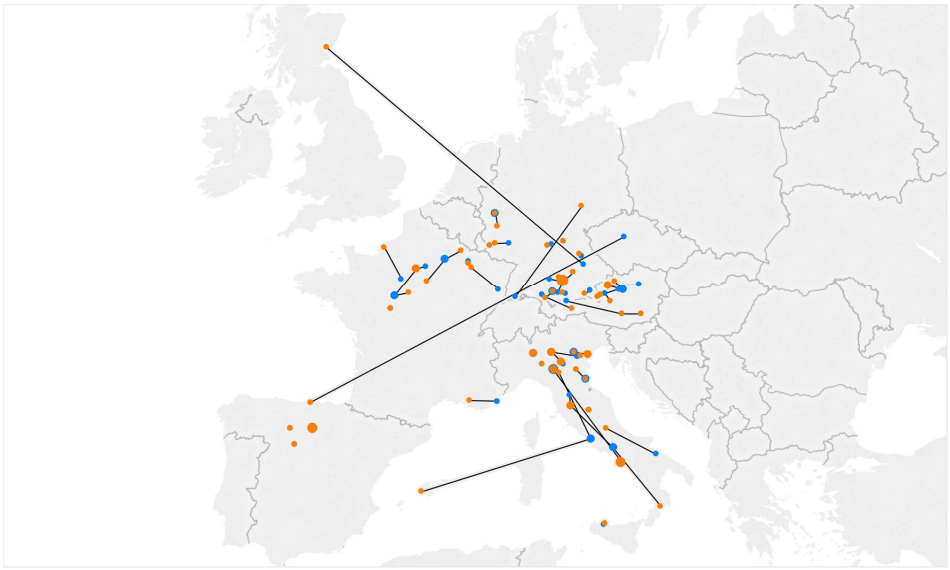
But the data from the underlying JSN database that feeds the project website can also be used in other ways. One example, prepared in collaboration with Thomas Wallnig, is the first pilot of a comparative study which visualized the places of birth and places of entry for a sample of Benedictines and Jesuits (fig. 8).[34]

Taken just by themselves, these images do not reveal any new insights, but they do illustrate and confirm impressions from the literature in a more directly accessible visual manner. Figure 8(b) demonstrates that the Jesuits came from all over Europe, had numerous centres scattered over a large area, more disparate points of entry into their order, and demonstrated some of the mobility that their additional vow would require even before entering the order. Figure 8(a) gives a contrasting graphical representation of the Benedictines' relative confinement to their geographical localities of origin and subsequent immobility. A prospective member born in France was supposed to enter a monastery in France, which resulted in the national congregations keeping mostly to themselves. This of course had consequences for the nationalizing tendencies within the Catholic Church, above all Gallicanism, opposing Jesuit 'universalism'.

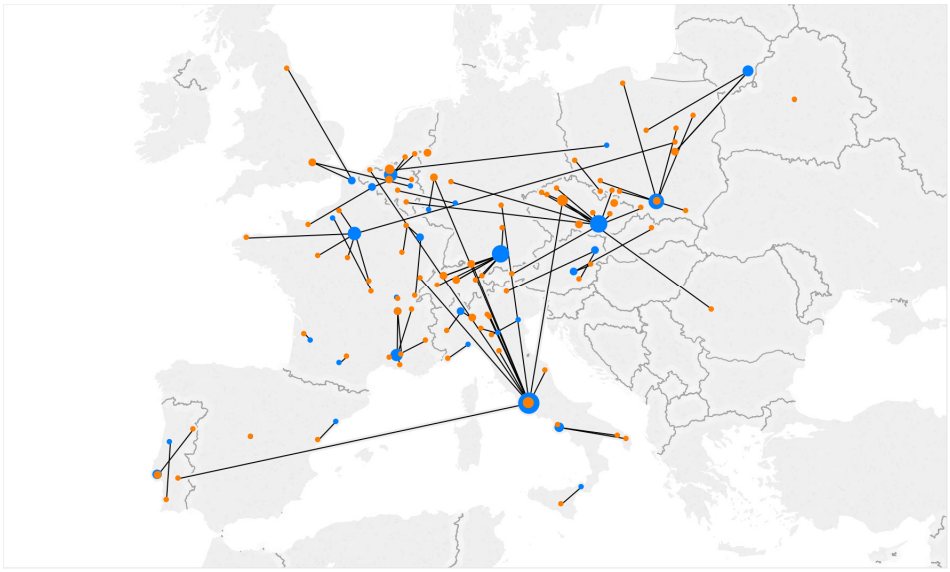
Individually, these visualizations facilitate the understanding of fundamental properties of each order, but their greatest utility lies in the comparison: the particular case shown here demonstrates the supremacy of the Jesuits over the Benedictines concerning the number of members active in the early modern sciences, and it also hints at distinct Jesuit centres of these activities. This could be further fine-tuned to focus on particular periods or subjects, expanded to include scholars from other religious orders, or altered to embrace other areas of knowledge.

A comparison of this kind is a meaningful exercise in the first place because both Benedictines and Jesuits are early modern religious orders. Despite all their differences, they can be documented within similar prosopographical structures. At the same time, they represent strikingly divergent approaches to scholarship as well as to religious life: the Benedictines, rooted in early medieval traditions and decentralized in structure, maintained a local and contemplative form of knowledge preservation, whereas the Jesuits, a relatively recent foundation highly hierarchical in its overall structure, were the primary missionary order of the Counter-Reformation: they acted globally and also conceived of knowledge in global terms. During the seventeenth and eighteenth centuries, these different intellectual cultures at times also brought the two communities into conflict, most famously over Jansenism and the theology of grace, but also over matters of historical research.

[34] The Benedictine sample, provided by Thomas Wallnig, was put together from a 1754 scholarly history of the order (Magnoald Ziegelbauer and Oliver Legipont, *Historia rei literaria ordinis sancti Benedicti*, 4 vols., Augsburg, 1754); the Jesuit sample was taken from the JSN. For more details, see Mrozik, 'The Jesuit Science Network', ch. 5.5.



(a): Benedictines (data courtesy of Thomas Wallnig)



(b): Jesuits

Figure 8: Places of birth (light blue dots) versus places of entry (dark blue dots) for select members of the Benedictine and Jesuit orders. Dot size scaled to the amount of people who were born and entered at this place, respectively. Visualization generated with *Tableau.* Map copyright: OpenStreetMap contributors

The brief experiment in data-driven comparison carried out above inevitably only hints at aspects of this broader picture, but it serves to indicate the way in which further research into these questions could be framed. Populating a shared model with the relevant data would allow comparison of the mobility of Jesuits and Benedictines throughout their lives; the two orders' global dimensions; the position and possible overlapping of their religious centres; and the relationships of these centres to important places of learning, printing, and political power. This would also assist a closer investigation of the situations in which the two communities came into conflict with each other. Adding data on other orders would further enrich the picture and multiply comparative perspectives. Eventually, the picture will become even richer if we include other orders – and if we then relate the results back to the broader geographies and chronologies of the republic of letters.

4 Modelling Other Learned Communities

4.1 Modelling the Learned Professions

Howard Hotson and Thomas Wallnig

One of the principal difficulties arising from the attempt to create a generic prosopographical data model relates to the variation of terms and categories with time and space. Partly because they emerge out of the three higher faculties of the medieval university, the learned professions of medicine, law, and theology tend to retain a quite similar structure across Europe during the early modern period; and this common structure may provide an opportunity for extending the academic data model to other domains relevant to the republic of letters as a whole.

The clearest case is probably theology. On the one hand, the situation is complicated by the manner in which the Reformation fractured the ecclesiastical structures of the medieval Church. On the other, new churches were erected in Protestant states and territories with their own internal structures, which remained fairly stable over time and often shared characteristics across international borders within the same religious confession. In the case of the Lutheran and Reformed clergy, for instance, endless proliferation of variety was constrained by the forces of confessionalization. As far as the English Church is concerned, data is already available in the *Clergy of the Church of England Database*.[35]

More caution is in order regarding law. While the degrees were stable, professions varied. There were different traditions across Europe of interpreting Roman law, customary law, and canon law in relation to each other, and these different interpretations were central in the creation of national states. This means that dif-

[35] See http://theclergydatabase.org.uk/, accessed 20/03/2019.

ferent kinds of legal counselling could acquire different shapes and meanings, for example, in post-civil-war France in the late sixteenth century, and in the Holy Roman Empire after the Peace of Westphalia. Moreover, in some parts of Europe (like Italy) the ancient institute of the public notary survived up to the end of the Ancien Régime, and the *parlement* of Paris was an institution of jurisdiction while at the same time displaying traits of a body of representation.

More coherence may again be achieved in the field of medicine: the basic profession of 'physician' and the terminologies used to describe it differ less radically with time and place, and the prospect of creating a reasonably neat and simple typology which captures the different types of medical practitioner reasonably accurately across Europe throughout the early modern period seems more probable than doing the same for political or administrative office-holders. In this domain, leadership has been taken by the database of 'Frühneuzeitliche Ärztebriefe', which uses the generic term *Arzt* as a label for profession, while also differentiating several different kinds of physician, including *Leibarzt* and *Chirurg*.[36]

In each case, collaboration will be needed with specialist communities devising precise means of modelling these groups of people whose biographical accounts 'intersect' with the world of learning and the fashioning of the learned 'persona'. It will be necessary then to move out from there to less well-documented, stable or uniform subsets of the learned community as a whole. These must include practitioners of more artisanal forms of learning – including printers, cartographers, and instrument makers. The example of Baruch de Spinoza serves to establish this point: a philosopher of central importance to modern thought, he earned his living as a lens grinder.

In all of this work, the important distinction between profession and sphere of activity must always be borne in mind. Profession can be indicated by the holding of professional qualifications, membership of professional organizations, or practice of various arts and sciences as the basis of livelihood. Spheres of activity can range far outside professions, above all for gentlemen virtuosi who were proudly amateurs rather than professionals of any kind. A given individual might be a professional in some of these senses and not others, might have professional status in more than one discipline, but might engage with the republic of letters in spheres quite different from his official professional domain. Leibniz, for instance, was qualified as a lawyer and employed as a court councillor, librarian, and historiographer, worked less formally in political and ecclesiastical diplomacy, but was known within the republic of letters primarily as a philosopher and mathematician. The extraordinary breadth and diversity of his intellectual engagements makes him a valuable test case for measuring the viability of any future prosopographical model for this domain.

[36] See http://www.medizingeschichte.uni-wuerzburg.de/akademie/index.html, accessed 20/03/2019.

4.2 Modelling Learned Women[37]

Gabriela Martínez

If academics provide a good starting point for generating prosopographical models for the republic of letters because their lives were 'institutionalized' within structures that were relatively stable across early modern Europe, the people least well integrated into learned institutions of this kind (at least outside the monastic orders) were women. In order to assess the difficulty of the prosopographical enterprise, some attention therefore needs to be directed at the challenges of modelling early modern learned women.

At the outset it must be stressed that, in proposing to develop prosopographical categories capable of capturing data on women who participated in the republic of letters, the intention is not to perpetuate the 'othering' of female learning. On the contrary, the aim is to ensure that the gender biases so pronounced in the early modern period are not thoughtlessly replicated in modern data models, the data assembled within those models, and the scholarship based on that data. Early modern European societies, for instance, typically considered it unnecessary and even undesirable to educate individuals of both genders equally and still less desirable to provide them with the same access to learned institutions;[38] so a prosopographical model designed to assemble only data on formal education and public institutional activities would guarantee the erasure of women from future scholarship based on it.

This situation poses several fundamental difficulties which need to be confronted at the outset. First, a data model capable of capturing women's activities cannot be populated primarily from stable, institutional sources, such as matriculation registers or membership lists of learned societies. As a consequence, second, the resulting data cannot be compiled in an equally systematic fashion from a finite list of basic sources. This means that, third, a wider range of alternative sources will have to be consulted to piece together, for instance, evidence of women's educational experience. And fourth, this implies that the 'male-centric' model will need to be supplemented with additional categories devised to capture evidence of women's informal education and intellectual activity, which frequently developed outside what might be called 'official' circles or formal institutions.

Among these sources of evidence, none is more important than letters themselves. The informal, personal, and semi-private nature of correspondence was well

[37] Work for this section was supported by a COST STSM for Gabriela Martínez Pérez entitled 'Adapting Cultures of Knowledge's Prosopographical Data Model for Women' [this text is not yet on the COST website].

[38] For a general vision on early modern women education, see Barbara J.Whitehead, ed., *Women's Education in Early Modern Europe: A History, 1500 to 1800* (New York and London: Garland, 1999). For the Spanish case, not included in Whitehead's volume, see Elisabeth T. Howe, *Education and Women in the Early Modern Hispanic World* (Aldershot: Ashgate, 2008).

adapted to expressing and preserving insights into the lives and interests of women. A variety of archives throughout Europe – including public repositories, convents, and private collections – preserve huge quantities of missives written by and to women during the early modern era. Together, these letters document the considerable female contribution to epistolary cultures in general, and to elite and learned correspondence more particularly. Pioneering initiatives have begun to assemble catalogues of this material in readily accessible and navigable form, notably in WEMLO (*Women's Early Modern Letters Online*),[39] a community cataloguing project, directed by James Daybell and Kim McLean-Fiander and supported by *Cultures of Knowledge*, which offers an open access catalogue of nearly 15,000 letters to, from, or mentioning women, including catalogues of the extant correspondences of some two dozen notable women. This project has both benefited from and contributed to the production of a new wave of literature in recent decades in which female participation in the republic of letters is increasingly well documented and carefully examined.[40]

WEMLO represents an obvious point of departure for collaboratively populated infrastructure in this field because there is no need for a separate *epistolary* data model for women's letters, which have the same core attributes as men's. A far larger problem is how to devise a *prosopographical* data model capable of accommodating the more fragmentary and informal data on women's education and participation in intellectual affairs.

The starting point for addressing this problem must be to recognize that this task need not begin *de novo*. Several projects have already begun collecting structured data on early modern literate and learned women. These include *Bibliografía de Escritoras Españolas* (BIESES);[41] *Epistolae: Medieval Women's Letters*;[42] *Escritoras: Women Writers in Portuguese before 1900*;[43] *Monastic Matrix: A Scholarly Resource for the Study of Women's Religious Communities*;[44] *NEWW Women Writers*;[45] *The Reception and Circulation of Early Modern Women's Writing, 1550–1700* (RECIRC);[46] *Who Were the Nuns? A*

[39] The EMLO project page at http://emlo-portal.bodleian.ox.ac.uk/collections/?page_id=2595 is supported by a Facebook page at https://www.facebook.com/WEMLO-Womens-Early-Modern-Letters-Online-325357000924586/, both accessed 20/03/2019.

[40] For an overview on this issue from an international perspective, see James Daybell and Andrew Gordon, eds., *Women and Epistolary Agency in Early Modern Culture, 1450–1690* (London and New York: Routledge, 2016); Carol Pal, *Republic of Women. Rethinking the Republic of Letters in the Seventeenth Century* (Cambridge: Cambridge University Press, 2012); Julie D. Campbell and Anne R. Larsen, *Early Modern Women and Transnational Communities of Letters* (London and New York: Routledge, 2006); Jane Couchman and Ann Crabb, *Women's Letters across Europe, 1400–1700. Form and Persuasion* (London and New York: Routledge, 2005). For specific national cases, see James Daybell, *Women Letter-Writers in Tudor England* (Oxford: Oxford University Press, 2006) and Gabriella Zarri, ed., *Per lettera: la scrittura epistolare femminile tra archivio e tipografia. Secoli XV–XVII* (Rome: Viella, 1999).

[41] See https://www.bieses.net, accessed 20/03/2019.

[42] See https://epistolae.ctl.columbia.edu, accessed 20/03/2019.

[43] See http://www.escritoras-em-portugues.eu, accessed 20/03/2019.

[44] See https://monasticmatrix.osu.edu, accessed 20/03/2019.

[45] See http://resources.huygens.knaw.nl/womenwriters, accessed 20/03/2019.

[46] See http://recirc.nuigalway.ie, accessed 20/03/2019.

Prosopographical Study of the English Convents in Exile (1600–1800);[47] and *Middlebrow Enlightenment: Disseminating Ideas, Authors and Texts in Europe (1665–1830)* (MEDIATE).[48] A serious assault on this problem must begin by pooling the resources and experience of established projects such as these.

As a preliminary scoping exercise, a COST short-term scientific mission was organized to bring a few of these existing resources together for study. The objective was to assemble a collection of diverse women correspondents in order to determine which aspects of their lives could be readily accommodated within the prosopographical data model being developed by the *Cultures of Knowledge* project, which aspects of the model required modification for dealing with women, and what missing features could profitably be added to it.

To provide an initial test case, nine women were selected, all of them either contained in the WEMLO repository or being studied in BIESES: Bess of Hardwick (c. 1521–1608, amongst the wealthiest and more redoubtable Elizabethan noblewomen); Anne Bacon (1528–1610, scholar, translator, and mother of Francis); Margaret Clifford (1540–1596, an aristocratic literary patron interested in alchemical medicine); Ana de san Bartolomé (1549–1626, a Spanish Carmelite prioress in France and Flanders); Penelope Rich (1563–1607, accomplished courtier and muse of Sir Philip Sidney); Luisa de Carvajal (1566?–1614, a Spanish mystic poet and Catholic missionary in England); Anna Maria van Schurman (1607–1678, the most prodigiously learned Dutch woman of her age); Madame Françoise de Graffigny (1695–1758, the celebrated novelist, playwright, and Parisian salon hostess); and Lucía Carrillo de Albornoz (1735–1805, key member of a relevant aristocratic family in viceregal Lima). This selection not only covered a fairly large area chronologically and geographically, it also brought together a quite diverse set of female correspondents: a large amount of data is available for some of these women (such as van Schurman and de Graffigny), but for others (Bess of Hardwick, Lucía Carrillo), very little (a distinction relevant to the discussion of core versus supplementary data in ch. II.4, sect. 3.6). 'Trial fitting' the biographical data available on these women was undertaken as a preliminary means of exploring the requisites of an enhanced prosopographical data model.

Some of the results of this exercise were predictable. The 'core' prosopographical data – including place and date of birth and death, names and status of parents, etc. (ch. II.4, sect. 3.6) – naturally applied to all female protagonists, as did generic categories relating to confession. Equally predictable was the fact that many other categories central to the definition of male citizens of the republic of letters had little bearing on their female counterparts: these included the standard stages of a formal, extra-mural education, membership of learned societies, political activity, and most categories of professional activity.

[47] See https://wwtn.history.qmul.ac.uk, accessed 20/03/2019.

[48] See http://mediate18.nl/, accessed 20/03/2019.

More difficult was the question of what alternative categories might be devised to capture the female counterparts to some of these experiences and activities. For instance, where formal educational qualifications and political roles are lacking, what precise categories could be devised to indicate a woman's degree of literacy or empowerment? Since the domestic environment is of crucial importance to the female subject, what categories might be developed for differentiating between different kinds of family environment?

Despite its brief duration and preliminary nature, this exercise provided plenty of food for thought. Amongst the major categories of activity requiring refinement in order to accommodate data on women are the following:

- *Maritial, familial, and domestic events* require further elaboration: these include not only formal changes of status (such as marriage, widowhood, religious profession, and the birth and death of children), but also informal developments (such as pregnancy, miscarriage, extramarital relationships, marital separation, and celibacy). Likewise, dealings relating to the management of households and estates (including formal litigation) loom large, particularly in the biographies of heiresses and propertied widows.
- *Religious life* is a sphere for which abundant formal institutional records exist: categories needed include religious confession, religious order, convent, formal roles within order and convent, change(s) of status and profession, foundation of convent, and potentially sanctification/beatification.
- *Education* is a domain in which more differentiation is needed between the various alternatives to formal education, including private schooling (whether under a relative or private tutor, whether alone or with others, whether with relatives or others, whether inside or outside the parental home), formal extramural education (e.g. in a convent), and the status of autodidact.
- *Literacy* is a category for which existing distinctions need to be implemented (including between the ability to read/to write one's name/to write more generally) and applied to all other languages read, spoken, or written.
- *Literary activity* can be documented with reference to books owned, works quoted, extant marginalia, literary and non-literary writings, diffusion strategies (such as the use of anonymity, pseudonyms, and restricted circulation in manuscript), evidence of their diffusion and reception (in manuscript and print, including quotation in other texts), as well as receipt of dedications and literary productions by others and other evidence of patronage activity.
- *Social life and networks* are related categories requiring careful modelling, including roles such as mentor, gatekeeper to cultural circles, intermediary, and hostess.

In essence, capturing many of these forms of literary and intellectual activity simply requires a more precise and nuanced accounting of activities that women shared with men. In other words, the development of means of capturing female activities

in and around the republic of letters will ultimately benefit the detailed study of their male counterparts as well. This being the case, the development of a prosopographical model capable of documenting female literary and intellectual activities will not distort that needed for their male contemporaries: it will enhance, enrich, and deepen it.

IV.5 Networking the Republic of Letters

Ruth Ahnert and Sebastian E. Ahnert

With contributions from Per Pippin Aspaas, Howard Hotson, Christoph Kudella, Ikaros Mantouvalos, Alexandra Sfoini, and Anna Skolimowska

In recent years it has become common to speak about the republic of letters as a network. But this was not always the case. Rather, it is the product of a specific set of conditions: the confluence of readily available digitized documents, computational power to analyse that data, and a ready acceptance of the 'network perspective' in the popular consciousness. In our increasingly interconnected world we encounter networks at every turn. The Internet, public transport networks, and power grids make our everyday lives possible; our careers are dependent on networking; and social networking sites provide an online account of our professional and personal capital. Networks have become a metaphor for connectedness, but also a concrete framework for visualizing and measuring complex systems of knowledge in the era of big data.

Although scholars working in the humanities might not realize it, the network turn is due to the emergence of 'network science' as a field of interdisciplinary study. In a series of key publications in the late 1990s and early 2000s, scholars such as Albert-László Barabási, Reka Albert, Duncan J. Watts, and Steven Strogatz showed that a huge variety of real-world networks – such as, for example, neural networks, transport networks, biological regulatory networks, and social networks – share an underlying order, follow simple laws, and therefore can be analysed

using the same mathematical tools and models.[1] These publications build on work from various different disciplines, such as sociology, mathematics, and physics, which stretches back some decades; but the emergence of network science as a field in its own right was the product of certain conditions that did not exist before. Barabási and Albert explicitly cite the computerization of data acquisition as essential to their research. In other words, what they needed was numerous examples of big network data, which they could compare, and the computational power to analyse that data. In this field, thousands of publications every year describe the development of new quantitative network analysis methods, and the analysis of new types of network data.

The advent of large-scale digitization efforts in the humanities has given scholars unprecedented access to their research materials. Perhaps more importantly, however, it has also put quantitative analysis methods within the reach of this community. This is particularly true of large collections of metadata, as these represent structured information that is easier to abstract and quantify. Correspondence metadata, such as the data collected by the constituent members of the COST Action *Reassembling the Republic of Letters*, lends itself particularly well to quantitative analysis, as it is exactly the kind of data that network analysis was designed to study – a set of well-defined relationships, namely letters sent and received, between well-defined entities, namely individuals. As discussed in chapter II.4, some work may be necessary to establish the identities of the individuals, but correspondence is a social relationship that is particularly clearly defined, due to its physical manifestation in the form of the manuscript letter.

The value of the COST Action *Reassembling the Republic of Letters* additionally relies on a 'network effect' – a term employed in the context of modern technology companies, which means that the value of a software product rises with the number of people using it, as such products typically facilitate interactions between users in some way. By combining the metadata of a wide range of historical correspondence projects, and by making them compatible with each other, their combined value to the scholarly community is greatly increased. Consistent metadata allows for much more wide-ranging searches across correspondence collections, and the power of quantitative network analysis grows rapidly with the size and scope of the network under study.

[1] See Duncan Watts and Steven Strogatz, 'Collective Dynamics of "Small-world" Networks', *Nature* 393 (1998): 440–2, see https://doi.org/10.1038/30918; Albert-László Barabási and Reka Albert, 'Emergence of Scaling in Random Networks', *Science* 286 (1999): 509–12, see https://doi.org/10.1126/science.286.5439.509; and Reka Albert and Albert-László Barabási, 'Statistical Mechanics of Complex Networks', *Reviews of Modern Physics* 74 (2002): 47–97, see https://doi.org/10.1103/RevModPhys.74.47.

1 Letters as Data

While correspondence is an ideal form of data to analyse using network analysis, there are a number of obstacles that we face when applying these methods to the republic of letters. The biggest of these is what we might call 'data silos'. For hundreds of years vast resources have been invested in collecting, cataloguing, editing, annotating, and translating the letters exchanged between leading political and intellectual figures scattered across and beyond early modern Europe. These collections might be divided into two separate types: the physical archive and the virtual archive. The former, the physical archive, is determined by the actual location of the document, in a particular institutional or national repository. In the case of letters that were actually sent (as opposed to drafts or copies), their final resting place usually correlates with the location to which a missive was sent. Some of these locations would have been institutional, but most would have ended up in the personal records of their recipients, many of which later found their way into local or national libraries. The concept of the personal archive is often the basis too for the 'virtual archive'. We use this term here to think about the mission behind edited collections of correspondence: these were traditionally published between boards and brought together the unified personal archives of a named individual's received letters with their sent letters, which were, necessarily, scattered in perhaps as many locations as the number of people to whom the original author wrote. The task of reuniting these scattered letters often became the life's work of a given scholar, or, in some cases, whole communities of scholars.

While digitization efforts create great promise for the use of computational methods, like network analysis, the digitization of historical documents has for the most part only reinscribed these silos. While large sums of money have been invested to make letters available online, these tend to be available either through online archives that are accessed through institutional websites or virtual online archives focused around a particular identity (such as the *Hartlib Papers*, or *Bess of Hardwick's Letters*).[2] Such repositories have transformed the way research is done, and have been used both by traditional scholars and, more recently, by digital humanists. For both groups, however, the reliance on these data silos as sources means that the way we ask research questions is often circumscribed by the contours of those archives. More importantly for this chapter, these silos act as a barrier to network analysis.

Historians and literary scholars easily see the problem when the barrier is introduced by the contours of a physical archive; they already understand that individuals with letters contained therein may have other letters held in numerous other archives. However, this is in fact less of a problem than it initially seems: below we examine a network analysis of the letters held in the Tudor State Papers, and how such work can tell us some powerful things about an archive's making. By

[2] See https://hridigital.shef.ac.uk/hartlib/, and https://www.bessofhardwick.org/, both accessed 20/03/2019.

contrast, scholars feel confident that the collected correspondence of a named individual will be able to yield important insights about his/her network. In fact, this silo is much more difficult to analyse because it constitutes what we might call an ego-network. The standard definition of an ego-network is one that consists of a focal node ('ego') and the nodes to whom the ego is directly connected to (these are called 'alters'), plus the ties or edges among the alters.[3] Of course, the networks we have in edited collections of correspondence actually contain even less data than this because we lack those connections or edges between the alters. We can of course visualize that network; but without those edges between alters there are very few quantitative measures that can be derived. All we can count are: the degree of the ego (i.e. how many unique correspondents s/he has), the ego's in- and out-degree (the total number of people s/he writes to, or receives letters from), and the strength or weight of the edges the ego shares with their alters (i.e. how many letters passed along those edges in each direction). To derive these statistics, however, you do not really need network analysis.

The COST Action, however, presents an opportunity both to overcome these silos of knowledge, and to undertake more interesting network analysis of the republic of letters. The solution is the meta-archive (in this instance, hosted by *Early Modern Letters Online*, or EMLO). The concept of the meta-archive is an online resource that collects together metadata[4] from many different sources, both by creating metadata files for early modern letters that currently only exist in material forms, and by integrating metadata from numerous other digital projects, to create a powerful research hub for early modern researchers. While many of the correspondences that members of the Action are working on might be described as ego-networks, by bringing them together we create overlapping archives that provide those cross-links between the alters within the constituent ego-networks. The main challenge in establishing this meta-archive, as outlined in the foregoing chapters, is reconciling the metadata fields, and in particular of person identities, across different correspondence projects. This is why a substantial proportion of the overall time and energy of this COST Action has been spent on the development of technical resources for metadata disambiguation, de-duplication, and reconciliation (see chapter III.2). However, once this is achieved the composite archive presents exciting opportunities for analysis.

[3] Stanley Wasserman and Katherine Faust, *Social Networks Analysis: Methods and Applications* (Cambridge: Cambridge University Press, 1994), 41–3.

[4] Metadata is a set of data that describes and gives information about other data. For a letter this would be the name of sender, name of recipient, date and place from which it was sent, description of contents, and reference information such as shelf-mark.

2 What Can Networks Offer?

In abstract terms, and in its simplest form, a network is simply information about the presence or absence of connections (often termed 'edges' or 'links') between entities (often termed 'nodes' or 'vertices'). Wherever we encounter a definable set of entities – such as people, objects, institutions, or devices – and definable relationships – such as letters, phone calls, face-to-face interactions, or affiliations – we can cast a set of relationships in the language of network analysis. A network need not be binary. We can move beyond the presence or absence of connections and include information about the number of interactions, the frequency, the exact timings, or the length of each communication. The higher the resolution of the data in this regard, the more complex and therefore restricted the scope of quantitative analysis becomes. This trade-off between analytical power and resolution is the inevitable consequence of any process of abstraction.[5]

The number of properties that can be measured is vast and ever-expanding. Simple examples include the number of connections of a node: its 'degree' (already mentioned above). More complex, and often more interesting, examples include 'clustering coefficient', which measures the density of connections among the network neighbours of a node, and 'betweenness centrality', which measures the number of shortest paths through the network that pass through a given node. Newer analysis methods can take into account the temporal nature of a network, and can calculate the accessibility of information as a result of the time ordering of connections. Simply put, if B stops talking to C before A starts talking to B, information cannot pass from A to C. More basic analysis can be done using off-the-shelf software tools, such as *Gephi* and *Cytoscape*, whereas quantitative analysis that is tailored to a specific historical research question, or investigates more complex network measures, such as temporal ones, needs to be programmed in languages such as Python or R. More information on the former language can be found in the lesson written for *The Programming Historian* on 'Exploring and Analyzing Network Data with Python' written by John Ladd, Jessica Otis, Chris Warren, and Scott Weingart.[6]

There is a growing body of scholarship that demonstrates the power of such methods to uncover new findings in the humanities. The highly cited *Science* article 'A Network Framework of Cultural History' reconstructed aggregate intellectual

[5] There is not space here to outline the subtleties of this huge interdisciplinary field. For a more thorough introduction, there are several options. For an overview designed for a popular readership, see Albert-László Barabási, *Linked: The New Science of Networks* (Cambridge, MA: Perseus, 2002); for the mathematically literate there is Mark E. J. Newman, *Networks: An Introduction* (Oxford: Oxford University Press, 2010); and for humanists, Scott Weingart has produced a number of resources, including a series of blogposts called 'De-mystifying Networks', available at http://www.scottbot.net/, accessed 20/03/2019, and chapters in the book by Shawn Graham, Ian Milligan, and Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscope* (London: Imperial University Press, 2015).

[6] See https://programminghistorian.org/lessons/exploring-and-analyzing-network-data-with-python, accessed 20/03/2019.

mobility over two millennia through the birth and death locations of more than 150,000 notable individuals.[7] The resulting network of locations provides a macroscopic perspective of cultural history, which helped to retrace cultural narratives of Europe and North America using large-scale visualization and quantitative dynamical tools and to derive historical trends beyond the scope of specific events or narrow time intervals. In particular, they used this data to show the changing locations of cultural centres over time. There are several other communities of scholars that are making incremental developments, including: *The Connected Past*, a community that has held regular conferences and published outcomes in articles and books;[8] the Arts Humanities and Complex Networks Symposia, which have led to a large number of contributions in *Leonardo Journal*;[9] and the contributors behind the newly launched *Journal of Historical Network Research*,[10] among others. In the latter we see how those working on the republic of letters are already making key contributions: Ingeborg van Vogt's article 'Using Multilayered Networks to Disclose Books in the Republic of Letters', appeared in the inaugural issue.[11]

A common misconception surrounding the application of network analysis – and more generally, of quantitative methods – to the humanities is the idea that quantitative methods by themselves offer wholly new outcomes and insights. What these new approaches do best however is to *facilitate* new outcomes and insights in the context of traditional scholarship. Much like aerial photography enables archaeologists to gain an unprecedented large-scale overview of structures concealed underground, quantitative approaches such as network analysis can place an individual, group, or institution of historical interest into a much larger context in which their role can be examined from an entirely new perspective. Aerial photography also offers the opportunity to discover entirely unknown structures in overlooked areas of the landscape, just as quantitative analysis can use a variety of measurements to highlight the infrastructural roles of understudied individuals in a network. In both scenarios the quantitative analysis outcomes do not represent an endpoint, not least because the data they rely upon is inevitably an incomplete and biased representation of the social network at the time. Rather, these outcomes should be understood as stepping stones in an iterative process between large-scale analysis and detail-focused scholarship in the traditional vein. Just as the archaeologists must eventually return to the ground to actually dig up the structures they

[7] Maximilian Schich et al., 'A Network Framework of Cultural History', *Science* 345:6196 (2014): 558–62, see https://doi.org/10.1126/science.1240064.

[8] Anna Collar, Fiona Coward, Tom Brughmans, and Barbara J. Mills, eds., *The Connected Past: Critical and Innovative Approaches to Networks in Archaeology,* a special issue of *Journal of Archaeological Method and Theory* 22:1 (2015)*;* and Tom Brughmans, Anna Collar, and Fiona Coward, eds., *The Connected Past: Challenges to Network Studies in Archaeology and* History (Oxford: Oxford University Press, 2016).

[9] See, for example, Special Sections in *Leonardo Journal* issues 43:3 (2010), 44:3 (2011), 45:1 (2012), 45:3 (2012), 46:3 (2013), 47:3 (2014).

[10] *Journal of Historical Network Research*, https://jhnr.uni.lu/index.php/jhnr, accessed 20/03/2019.

[11] Ingeborg van Vugt, 'Using Multilayered Networks to Disclose Books in the Republic of Letters', *Journal of Historical Network Research* 1:1 (2017): 25–51, see https://doi.org/10.25517/jhnr.v1i1.7.

have mapped or discovered from above, the humanities scholar has to dig down into the outcomes of the quantitative analysis.

An example of this iterative process can be found in the work of Ruth Ahnert and Sebastian E. Ahnert (the lead authors of this chapter), who studied the underground network of a Protestant community during the reign of Queen Mary I of England.[12] From the metadata and content of almost 300 letters the authors extracted a network of correspondence relationships and other social interactions. The leaders of this community were the well-studied Protestant martyrs documented in contemporary writings such as *Foxe's Book of Martyrs* (1563, and later editions), and they unsurprisingly represent the nodes with the most connections in the network. By using more sophisticated network measurements, however, such as the aforementioned betweenness centrality, other figures came to the fore. These included women who provided important infrastructural support to the network in the form of money and shelter, as well as the letter couriers who formed the postal infrastructure. Both have largely been written out of the histories of this time, often already in the versions of the letters printed by Foxe, where references to women were disguised by reducing their names to initials or, in some cases, even changing their gender. However, these same figures rise again to the surface when their importance is measured using a network approach. Importantly the output of the quantitative analysis here is tied back to the underlying history – the numbers in themselves are not a final outcome.

The term 'network analysis' is often understood as 'network visualization'.[13] The field of quantitative network analysis as described above, however, does not necessarily overlap with visualization. This is because visualization offers a complementary approach, with its own opportunities and challenges (as described in ch. IV.1). A visual representation of a network can provide an intuitive overview of a network data set. The dominant hubs of the network are likely to stand out immediately, as are largely disconnected sub-communities, and parts of the network with a particularly high density of connections. Moreover, visualization can provide guidance in understanding the structure of sets of data too heterogeneous for formal network analysis, particularly when network visualizations are combined with cartographical and other perspectives on the data. Visualizations can therefore offer a powerful way to gain first intuitive insights into a network data set. It also offers a powerful rhetoric of its own for supporting scholarly arguments with concision and clarity – sometimes a picture really is worth a thousand words. The downside of visualizations is that their legibility for the purposes of interrogation

[12] Ruth Ahnert and Sebastian E. Ahnert, 'Protestant Letter Networks in the Reign of Mary I: A Quantitative Approach', *English Literary History* 82:1 (2015): 1–33, see https://doi.org/10.1353/elh.2015.0000.

[13] On the distinctions between network visualization and quantitative network analysis, see Ruth Ahnert, 'Maps Versus Networks', in Noah Moxham and Joad Raymond, eds., *News Networks in Early Modern Europe* (Leiden: Brill, 2016), 130–57; Shawn Graham, Ian Milligan, and Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscope* (London: Imperial College Press, 2015), 250.

decreases as the data set grows; whilst very large data sets may produce very beautiful images, it is often difficult for an untrained eye to intuit much more than the sprawling complexity of that data. In addition, many network visualizations are highly arbitrary, as there are multiple ways in which the same network can be displayed. Even the same network layout algorithm may generate very different visualizations from two identical or near-identical networks. In summary, network visualization offers an intuitive means of exploring small and medium-sized data sets, with the drawback of arbitrariness and therefore limited reproducibility. Quantitative network analysis by contrast produces no visual outputs, and few directly interpretable intuitive insights. It does however offer a plethora of well-defined, reproducible network measurements in order to guide scholarly enquiry in new directions.

In the following we will illustrate how a number of research projects in the COST Action can both contribute to and benefit from the large-scale collection and quantitative analysis of historical correspondence metadata from the republic of letters.

3 From Ego-network to Network

An example of an ego-network is that of the Korais's Correspondence Project (1777–1833), overseen by Ikaros Mantouvalos and Alexandra Sfoini.[14] Adamantios Korais (1748–1833), the most prominent scholar of the Modern Greek Enlightenment, was born in Smyrna into a prosperous merchant family of Chiot origin. He worked unsuccessfully as a merchant in Amsterdam (1771–6) and subsequently studied at the Medical School of the University of Montpellier. From 1788 until his death in 1833, Korais lived in Paris, where he was a member of the Société des Observateurs de l'Homme, and where he produced *inter alia* many critical editions of Ancient Greek authors. Korais may be considered to belong to the European republic of letters, with whose members he had developed relations and corresponded on issues of Greek interest. His six-volume correspondence (1777–1833) contains 1,511 letters, 1,286 of which were authored by Korais. These were sent to a total of 149 persons: 100 of them were Greek scholars, merchants, politicians, and military officers; and the other forty-nine were non-Greeks, mainly Hellenists – scholars and editors – such as Chardon de la Rochette, J.-F. Thurot, d'Anse de Villoison, A.-M. Bandini, J.-F. Boissonade, and Fr.-A. Wolf, but also philosophers and politicians such as Jeremy Bentham, Thomas Paine, and Thomas Jefferson. As is shown by a letter of his to Chardon de la Rochette (27 July 1793), he considers

[14] Mantouvalos and Sfoini have contributed the following two paragraphs to this chapter. For some of their research on Korais, see Ikaros Mantouvalos, "'The Great Korais died on April 6": An Unpublished Letter from Philip Fournarakis to Thomas Spaniolakis (1833)', *Eranistis* 27 (2009): 149–63 (in Greek); Alexandra Sfoini, 'Korais and Michaelis: The Democracy of the Language', *Eranistis* 29 (2016): 229–55 (in Greek).

no part of Europe as his homeland, but rather feels like a 'citizen of the world', his fellow citizens being a very small number of scholars who recognize the role of Ancient Greek texts in disseminating the Lights in Europe and commiserate with the enslavement of the Greeks.

Korais's communication with classical scholars and Philhellenes shows the long-distance intellectual community of the age of Greek Enlightenment, a world of literary figures that stretched across geographical and social boundaries. If we examine the location of his correspondents on the maps designed by Eleni Gadolou we can see sent letters to sixty-six cities and towns (forty-one in Europe, three in America and twenty-two in Greece – see fig. 1), and he received letters from eighty-nine letter-writers, fifty of whom were Greek and thirty-nine non-Greek scattered across various cities (fig. 2).

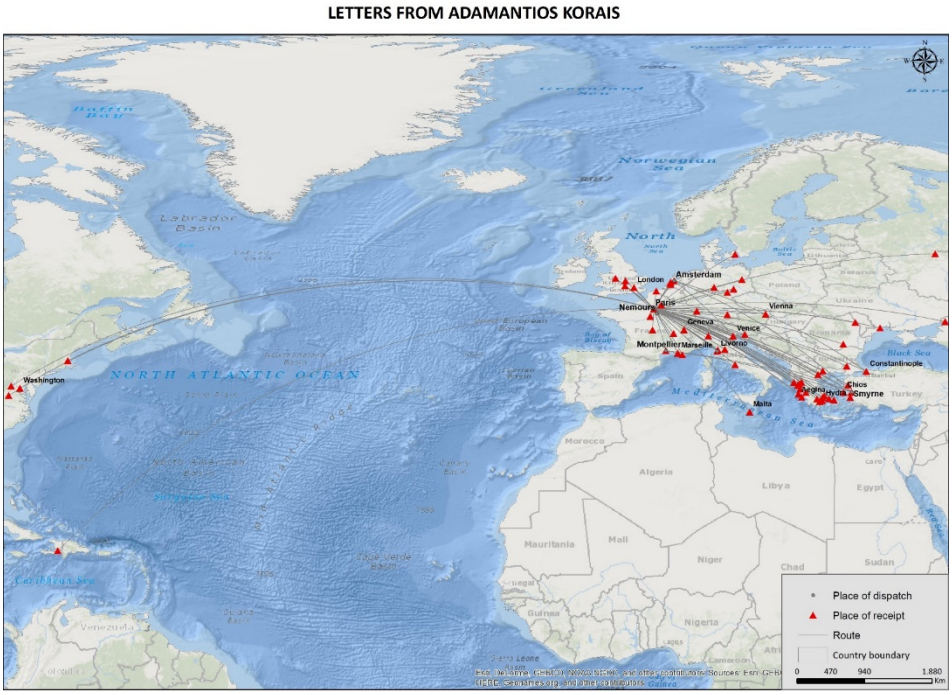


Figure 1

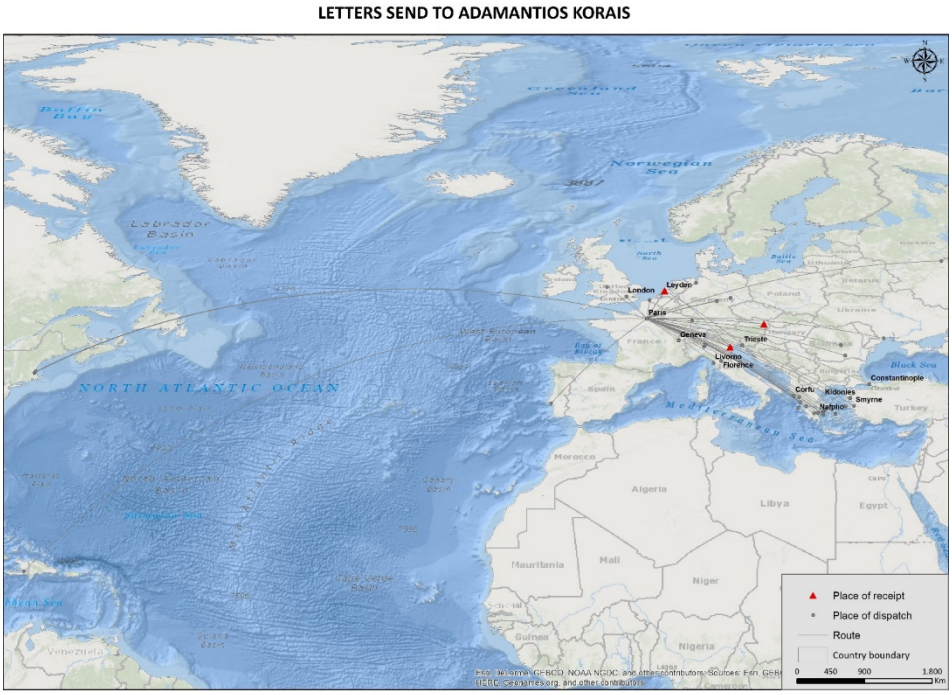


Figure 2

These maps quickly show us the the wide geographic dispersal of Korais's epistolary network. The varying intensity of these exchanges helped create centres of intellectual life, mostly in Europe, defining which regions were more involved in cultural exchange and intellectual debate.

Through this example we can begin to see how ego-networks like this form the essential building blocks of the collective effort of the COST Action. At least some of the 149 people Korais sends letters to, and the eighty-nine he receives correspondence from, are likely to appear in other correspondence projects of the period. In the framework of the COST Action, as a result of the reconciliation of person identities, these networks become connected to each other. This enables both a re-examination of individual ego-networks in the light of their correspondents' own correspondences, as well as a larger-scale analysis of their infrastructural roles in a much larger network. The potential of these overlaps to generate important discoveries has been a major point of exploration in various working groups within the COST Action, including the visualization working group who have explored avenues for visualizing large correspondence corpora in their relations to one another, using the metadata of four or five correspondences contained in EMLO, and to do this in a readily comprehensible way. Their aim was to help users navigate intersections of such corpora, especially when seeking to explore new aspects

of the data, such as the role played by sub-networks or gain new insights into knowledge exchanged between third parties (see ch. IV.1).

The power of examining these intersections between ego-networks is demonstrated by the ongoing work of Christoph Kudella and Anna Skolimowska on Erasmus of Rotterdam (1466?–1536).[15] Erasmus's oeuvre is deservedly famous for its size, and its geographical, and social, scope. Of Erasmus's epistolary exchange, 3,098 letters, written exclusively in Latin, are preserved between 1484 and 1536 (of which 37 per cent are letters to him). Erasmus corresponded with almost all the eminent figures of his time, whose respective corpora of correspondence provide researchers with insights into the interlinked nature of the early modern republic of letters. The intersections between individual networks of correspondence in the Erasmian republic of letters can be exemplified by the case of Ioannes Dantiscus (1485–1548), a diplomat in the service of the king and queen of Poland, bishop in Kulm and Ermland, and a patron of scholars and artists. Of his correspondence, 6,120 letters from the years 1500–48 are preserved (of which 72 per cent are letters to him), written predominantly in Latin and German. While a direct epistolary contact between Erasmus and Dantiscus is evidenced solely by a single letter, they had twenty-four correspondents in common. These two dozen individuals constitute only 3–4 per cent of the total correspondents of Erasmus and Dantiscus, but they serve to illustrate how even two geographically disparate correspondents with minimal direct contact can be linked by multiple third parties ('alters') who were in contact with both. Analysing the 'alters' connecting multiple correspondents is one of the obvious opportunities arising from the collection of multiple correspondences in EMLO.

The analysis of multiple intersecting correspondence is also a means of testing hypotheses difficult to assess through traditional means. A promising example is Per Pippin Aspaas's exploration of the correspondence of several eighteenth-century astronomers, funded by an STSM within the COST Action.[16] It is now well established that, by the middle of the eighteenth century, the theories of Kepler and Newton had gained de facto acceptance in all quarters. As a result, an increasing number of observatories popped up across Europe and beyond. With these institutions there followed a degree of professionalization that has led to eighteenth-century astronomy being described as a scientific discipline *avant la lettre*.[17] This discipline was driven forward by collaboration: in order to test new

[15] This paragraph has been contributed by Kudella and Skolimowska. Kudella's contribution draws on his unpublished PhD Thesis, 'The Correspondence Network of Erasmus of Rotterdam: A Data-Driven Exploration', University College Cork, 2017. Skolimowska's contribution draws on her work directing Internet publication of the 'Corpus of Ioannes Dantiscus: Texts and Correspondence' at the University of Warsaw, see http://dantiscus.al.uw.edu.pl/, accessed 20/03/2019.

[16] Aspaas, 'Astronomia disciplina maxime oecumenica?', an STSM hosted by Fritz Nagel at the Bernoulli-Euler Zentrum, Basle, from 19 February to 2 March 2017. The following two paragraphs derive from this work.

[17] Irène Passeron, René Sigrist, and Siegfried Bodenmann, 'La république des sciences: Réseaux des correspondances, des académies et des livres scientifiques', *Dix-huitième siècle* 40:1 (2008): 5–27 (esp. 20), see https://doi.org/10.3917/dhs.040.0005.

instruments and observational procedures, fix the longitude, calculate trajectories of planets and comets, etc., widespread exchange of 'corresponding observations' became necessary. A question that has rarely been raised is to what extent individual astronomers crossed linguistic, political, and – above all – denominational borders in their pursuit of corresponding observations. It is widely attested that they did so, but the exact extent and duration of such trans-denominational collaboration, and how it may have fluctuated over time, has not yet been the object of scrutiny. Within the framework of the COST Action, five astronomers from the latter half of the eighteenth century have been singled out for analysis. The primary correspondence collections (where the incoming and/or outgoing correspondence remains largely intact) are those of Placidus Fixlmillner OSB, head of the observatory of the Kremsmünster Monastery from 1762 to 1791; the Protestant Pehr Wilhelm Wargentin, secretary of the Royal Swedish Academy of Sciences and head of Stockholm Observatory from 1753 to 1783; and Johann III Bernoulli of Huguenot stock, head of the observatory of the Berlin Academy of Sciences from 1767 to 1787. They are enriched by two correspondences that are more fragmentarily preserved: those of Franciscus Weiss SJ, head of the University Observatories in Tyrnavia (Trnava) and Budapest from 1755 to 1785, and Maximilianus Hell SJ, imperial and royal astronomer of Vienna and head of the University Observatory from 1755 to 1792.

The resulting collection comprises several thousand letters, exchanged between astronomers from all over Europe. Basic metadata on all these letters (date of composition, names, geo-coordinates, and denomination of both sender and recipient) will, in due course, be entered into the EMLO database. By extracting the metadata and using both visual and quantitative network analysis one can study how the various correspondents' networks developed over time, with the hope that two corollaries may be achieved. First and foremost, one may expect that the analysis will illustrate the implications of pivotal developments, such as the abolition of Jesuits from Portugal, Spain, and France beginning in the late 1750s and culminating with the universal suppression of the Society of Jesus by the pope in 1773. A likely assumption is that (ex-)Jesuit astronomers will either be less visible, or disappear altogether from the map, as these developments unfold. Secondly, more basic questions of historical methodology may be tested, including: to what extent can visual and quantitative network analysis help pinpoint trends and ruptures that cannot be observed through more traditional methods of hermeneutics?[18]

Bringing together even larger collections of overlapping correspondences potentially opens up the possibility of understanding the structure and formation of increasingly large portions of the republic of letters more generally. A good example is ongoing work designed to understand the network of the Anglo-German

[18] A preliminary result, based on the study of a subset of this data is currently in press, see Per Pippin Aspaas and Katalin Pataki, 'Did astronomy constitute a denominationally neutral space within the Republic of Letters? An outline for the use of visualization tools in the study of astronomical correspondence', *Jahrbuch der Österreichischen Gesellschaft zur Erforschung des 18. Jahrhunderts* 34 (2019).

intelligencer of mid-seventeenth-century London, Samuel Hartlib (c. 1600–1662).[19] For almost half a century, historians have frequently labelled this network as 'the Hartlib circle', a designation which seems to imply that Hartlib himself is both the centre of that circle and the agency which brought it into being. The difficulty is that these assumptions are dangerously tautological. Hartlib is naturally the central figure in his own ego-network and the archive of it which he collected, and that archive is the key source of documentation of 'his circle'. But did that 'circle' have a robust reality outside his archive? Was he as central to the intellectual activity of the 1630s, 1640s, and 1650s as naturally appears when we view that period through the lens of his archive?

In order to answer this question, it is necessary to step outside the archive and immerse it within a representative cross-section of data documenting the intellectual commerce of England and neighbouring regions during Hartlib's active period. With that prospect in mind, EMLO has gradually assembled inventories of the letters' numerous contemporary intellectuals who corresponded with Hartlib: these currently include Johann Valentin Andreae, Elias Ashmole, John Aubrey, the Dutch Church at Austin Friars, Johann Heinrich Bisterfeld, Antoinette Bourignon, Robert Boyle, Johannes Coccejus, Jan Amos Comenius, Elisabeth Stuart, René Descartes, Abraham von Frankenberg, Hugo Grotius, Athanasius Kircher, Marin Mersenne, Henry Oldenburg, Nicolas-Claude Fabri de Peiresc, Johann Permeier, Henricus Reneri, and John Wallis. With this newly amassed data, we will be able to test that contention for the first time. Network analysis is ideally suited to quantifying the centrality of Hartlib's correspondence within this much larger body of data, and to determine the degree to which members of 'his circle' were independently connected with one another. Moreover, a chronologically organized series of studies may also help to reveal the process in which Hartlib's network was formed, and his own centrality – or otherwise – to that process.

Broader insights may then be gleaned by immersing this entire composite data set within a still larger catalogue. In existing historical literature, Hartlib is typically listed alongside Mersenne, Peiresc, Kircher, and Oldenburg as one of the key intellectual networkers or 'intelligencers' of the seventeenth-century republic of letters. Yet the parliamentary pension he was granted between 1645 and 1660 for his intelligencing activity was not for services to the republic of letters: it was 'in regard of the intelligence and correspondence maintained by him abroad' on behalf of the

[19] The following three paragraphs have been contributed by Howard Hotson. Although the idea that Hartlib was an important linking figure, central to several important groups, is much older, the formulation 'the Hartlib circle' seems to have been first used by Charles Webster's pioneering collection of source material, *Samuel Hartlib and the Advancement of Learning* (Cambridge: Cambridge University Press, 1970), vii and *passim*. It was further developed in what remains the central study of the topic: Webster's *The Great Instauration: Science, Medicine and Reform, 1626–1660* (London: Duckworth, 1975). See also Mark Greengrass, Michael Leslie, and Timothy Raylor, eds., *Samuel Hartlib and Universal Reformation: Studies in Intellectual Communication* (Cambridge: Cambridge University Press, 1994).

Commonwealth and Protectorate.[20] More specifically, between 1654 and 1661 Hartlib conducted a news agency, collecting excerpts from letters from the Continent, very often of a military, political, or diplomatic character, for delivery to Cromwell's secretary of state, John Thurloe, many of them from the newsbooks of the day, including *The Moderate Intelligencer* and *The Public Intelligencer*. For that reason Hartlib presents a fascinating site of experiment: he straddles the international, intellectual 'intelligencing' characteristic of the republic of letters, and the more pragmatic intelligence gathering central to the formation of the English state. A large body of correspondence representative of this kind of political intelligencing within the Commonwealth and Protectorate are readily available within the English seventeenth-century State Papers. Analysing the manner in which Hartlib's intelligencing activities cut across these two intersecting data sets might open up fresh perspectives on the manner in which the intellectual intelligencing within the republic of letters both contributed to and was superimposed on the information-gathering of the early modern state.[21]

The above cases outline briefly the potential of bringing together multiple ego-networks with others kinds of archives. However, such a narrative falls into the common pattern of digital humanities scholarship of speaking in the future tense: of what could, or should, or will be possible; of outlining work in progress, or methodologies developed that will be able to solve problems. As Franco Moretti has observed: 'Somehow digital humanities has managed to secure for itself this endless infancy, in which, it is always a future promise'.[22] Moretti, with others, has complained of the relative lack of completed research that has demonstrated unequivocally the value of digital methods to uncover new findings or to establish grand theories. There is often a good reason for this: the scale of ambition in projects like the COST Action means that a lot of preparatory work is required. We can either get quick and dirty results, or take the time to clean and prepare data meticulously so that we can have faith in our findings. As the previous chapters have thoroughly documented, the particular problems of historical humanities data clearly shows why there have been few interventions demonstrating the application of quantitative network analysis to early modern letters. Nevertheless, despite the considerable groundwork required, long-standing projects on large-scale early

[20] George Henry Turnbull, *Samuel Hartlib: A Sketch of His Life and His Relations to J. A. Comenius* (Oxford: Oxford University Press, 1920), 49.

[21] In order to pursue this possibility, Hotson and the *Cultures of Knowledge* project have joined forces with the lead authors of this chapter in the pursuit of the funding necessary to amass this body of data on EMLO and subject it to network analysis.

[22] Melissa Dinsman, 'The Digital in the Humanities: An Interview with Franco Moretti', *LA Review of Books*, https://lareviewofbooks.org/article/the-digital-in-the-humanities-an-interview-with-franco-moretti/, accessed 20/03/2019.

modern letter networks are beginning to yield results. The following is a preview of forthcoming work by the lead authors of this chapter.[23]

4 A Test Case: *Tudor Networks of Power*

The benefits offered by the large-scale collection and analysis of historical correspondence data are demonstrated by the AHRC-funded *Tudor Networks of Power* project, which examines the correspondence network formed by 132,747 letters in the Tudor State Papers from the period 1509–1603. The archive comprises the accumulated papers of the secretaries of state relating to home affairs, the papers produced or received by the secretaries as a result of their conduct of British diplomacy abroad, as well as petitions written to the government by ordinary people like farmers and widows, and bodies of letters seized or intercepted for the benefit of government intelligence. The epistolary archive implicates 20,663 unique people, either as senders or recipients. The project underwent an extensive disambiguation and de-duplication effort to map variant spellings, changing titles, name changes, and aliases to the correct individuals, and a similar process to clean the fields of place names and map them to geo-coordinates. It is now employing a range of network analysis measures as well as textual and geographical analysis to study a wide variety of historical research questions, such as: What is the changing role of the early modern 'intelligencer' during the Tudor period? What infrastructural roles did women occupy in the Tudor networks of power? Who were the individuals bridging disparate political communities? Can we use networks to make new predictions about the true identities of aliases? Which individuals weathered the mid-sixteenth-century political and religious changes better than others, and why? Which individuals were talked about by others, and how do the networks of those who were talked about relate to the networks of those talking about them?

So what can network measures reveal about this archive? Starting with the most basic observations, the ranking of nodes by their degree (the number of unique people with whom a given node shares edges) is able to show the prominence of certain hubs. Unsurprisingly, the nodes with the very highest degree are the Tudor monarchs, secretaries of state, foreign leaders, and key statesmen. The measure of betweenness centrality (which measures the number of a times a shortest path travels through any given node) is a valuable measure for highlighting figures who act as bridges, crossing 'structural holes' in a network and are therefore good at highlighting the Tudor diplomatic corps:[24] resident ambassadors, special ambassadors and commissioners, and intelligencers (often soldiers, or merchants,

[23] The monograph *Tudor Networks of Power* is a work in progress; the majority of the findings below draw on material reported in Ruth Ahnert and Sebastian E. Ahnert, 'Metadata, Surveillance, and the Tudor State', *History Workshop Journal*, dby033, https://doi.org/10.1093/hwj/dby033.

[24] On structural holes, see Ronald S. Burt, *Structural Holes: The Social Structure of Competition* (Cambridge, MA: Harvard University Press, 1992).

but sometimes travelling academics) sending weekly news bulletins to the secretaries of state). More interesting, however, are those nodes with the statistical combination of high betweenness centrality and relatively low degree, i.e. those who only have a few connections within the epistolary network, but nevertheless still have a high bridging function. If we look at the 1570s–1590s, a large number of the people who fulfil this condition are recognizable as spies, double agents, and conspirators. The clustering of similar figures is intriguing and implies that there may be a specific network profile for those trading in secrets; it seems unlikely that such a striking trend can be attributed merely to chance.

If there is a network profile for spies and conspirators, then a predictive model can also be developed. The discovery that such figures have this specific combination of statistical features led to an exploration of whether that information could be used to predict other likely spies and conspirators. Such methods could tell us which of the 20,656 people in the archive were most likely to have been involved in, or the focus of, Tudor surveillance, and therefore which of the 132,747 letters were worth reading in closer detail. By bringing in six further measures in addition to degree and betweenness centrality (in-degree, out-degree, strength, in-strength, out-strength, and eigenvector centrality) it is possible to assign each node a network 'profile' based on their individual scores and ranking for each of these eight measures: a kind of signature. It is then possible to measure the distance between these signatures (using Euclidean distance on the logarithms of the ranks), and thereby construct a measurement of network similarity between individuals. The result is a ranked list of people most similar to a given individual in terms of their network profile.

The results are striking. If we begin with Cardinal William Allen, who was leader of the English Catholic exiles and implicated in various conspiracies to dethrone Elizabeth I and replace her with a Catholic monarch, the fifteen most 'similar' people writing in Elizabeth I's reign include seven Catholic conspirators from the British Isles, and five continental Catholics, four of which are Spanish men in positions of diplomatic and military leadership.[25] What unites them is that all of these were perceived to present foreign threats to England's security, and the majority of their correspondence entered the archive through interception. These were people who were being carefully watched by the Tudor government, and this kind of surveillance leaves behind a particular kind of network profile in the archive.

This distance measurement not only finds patterns of conspiracy and interception, however. Its use is more general, helping us to understand the commonalities in network properties within and between particular groups of people. In this way we can, for example, find clusters of diplomats sharing network attributes. For example, if we look at Tommaso Spinelli – one of England's earliest resident am-

[25] These are William Douglas, earl of Angus, Robert Persons, Francis Dacre, Anthony Babington, Hugh Owen, Thomas Paget, Gilbert Curll; and Antonio de Guaras, Don Juan d'Idiaquez, Pedro de Zubiaur, and Charles of Lorraine, duke of Mayenne.

bassadors, serving at the court of Margaret of Austria – we find that nineteen of the twenty most similar individuals in the reign of Henry VIII all served on diplomatic missions during this reign.[26] Similarly, we can use the method to highlight a category of extra-diplomatic 'intelligencers' working in the Elizabethan period. Here our starting point is one Pietro Bizzarri, who offered himself to William Cecil, Lord Burghley (the principal secretary to Elizabeth I) as an intelligence-gatherer in Venice, in return for permission to travel.[27] It was an offer Burghley readily accepted, having no diplomatic presence in Venice at that time; and so began Bizzarri's lifelong career as an intelligencer, passing political and diplomatic information to the Tudor government. We find that the fifteen most similar people to Bizzarri in the Elizabethan period include fourteen who also provided the government with intelligence.

What is perhaps notable about this list of fourteen intelligencers is that only five of them have any kind of biography, either in the *Oxford Dictionary of National Biography*, *Wikipedia*, or *The History of Parliament*. Rather, the majority of these men are the kind of figures who only get a single sentence in reference books, normally saying something along the lines of 'X sent a letter to Walsingham/Burghley/Cecil with the information that …'. The focus is on the events reported on by these men, rather than on the men themselves and their intelligence roles. The men individually may not have been deemed worthy of their own histories (although the potted histories above suggest that some are), but one might contend that, considered as a group, they are. By using the similarity score we are encouraged to understand the commonalities between those men, and the way that the government employed them to supplement the information gathered through formal diplomatic arrangements. As a group they greatly influenced the foreign policy of the Elizabethan government, as is evident in their substantial contributions to its collected archives. This predictive approach, then, has the additional benefit of suggesting to us not only individual men and women whose letters may merit closer attention, but also of proposing to us new categories of writers whose significance perhaps only emerges when understood as a group.

5 Conclusion

This brief outline of the application of quantitative network analysis to the republic of letters is a narrative of trade-offs and pay-offs. In the application of quantitative network analysis, one such trade-off is between data complexity and computational power. The kinds of analysis undertaken on the *Tudor Networks of Power* project

[26] On Spinelli, see Betty Behrens, 'The Office of the English Resident Ambassador: Its Evolution as Illustrated by the Career of Sir Thomas Spinelly, 1509–22', *Transactions of the Royal Historical Society* 16 (1933): 161–95 (esp. 162), see https://doi.org/10.2307/3678668.

[27] The only book-length study on the intelligencer and historian is Massimo Firpo, *Pietro Bizzarri: esule italiano del Cinquecento* (Torino: Giappichelli, 1971).

takes place often at the most abstract level: namely, when network data is abstracted as a system of nodes and directed edges. The majority of the algorithms used do not take account of the weight of the edges (i.e. number of letters that passed), or any incidental information which enriches our understanding of those nodes or edges (such as roles held by node, or additional information about relationships between nodes, such as kinship). By ignoring that additional information in the first stage of analysis, the project has been able to find overarching patterns and trends, to identify anomalies that require closer analysis and discover people who might have been overlooked, and to develop predictive models and an understanding of commonalities between nodes. But in the humanistic context, network analysis is not necessarily undertaken as an end in itself. Rather, it can serve to open up revealing new perspectives on historical data in all its richness. The abstract, quantitative findings act as prompts to return to the concrete peculiarities of the individual letter, where close reading is needed to explain and illuminate these quantitative results, which in turn can help to form new large-scale questions that can be asked and answered with network analysis.

Further trade-offs will be needed to apply similar methods to analysing data pertaining to the republic of letters. The basic precondition for moving beyond ego-centred archives and the analyses based on them is to create data sets where we can add those all-important edges between alters. Before we can undertake meaningful computational analysis, in other words, a great deal of foundational work is required, of the kind outlined above and in previous chapters. This will require trade-offs in the scholarly environment more broadly, in order to commit to sharing data, collaborating, and undertaking the unglamorous curatorial work of reconciling name and place data across these archive silos. But the pay-offs for such a cultural shift are potentially transformative: if the work is undertaken properly, we will be able to navigate between multiple archives, executing computational measures that leverage all this data to give us an overview of the early modern social, political, and intellectual networks that is greater than the sum of its parts.

IV.6 Text-mining the Republic of Letters

Charles van den Heuvel, Jan Bloemendal, Robin Buning, Mihai Dascalu, Simon Hengchen, Barbara McGillivray, Sinai Rusinek, Lucie Storchová, Stefan Trausan-Matu, and Vladimír Urbánek

With contributions from Tommaso Elli, Giovanni Moretti, and Ludovica Marinucci

1 Introduction

Charles van den Heuvel

Navigating the ocean of textual data potentially assembled to document the epistolary exchanges central to the republic of letters will require the adaptation and application of multiple text-mining techniques. Proper explanation of the methodologies underlying these techniques and thorough exploration of their application to large quantities of early modern learned correspondence would vastly exceed the time and resources made available by the COST Action and the space available in this volume. Instead, this chapter will attempt merely to illustrate the potential applicability of several of these techniques to the kind of problems central to the study of the republic of letters.

In the second section, following this introduction, one of the founding fathers of the *respublica litteraria*, Desiderius Erasmus, is used as a basis for discussing how stylometrics can be used to ascertain the authorship of anonymously or pseudonymously published texts and, by extension, of letters whose authorship is uncertain. The third section explores the way in which text-reuse algorithms can be used

to detect the reuse of rhetorical features and other textual passages by the same author as well as the hypertextual reproduction of classical and contemporary authors within the Latin discourse of the republic of letters. The fourth section recounts a basic experiment conducted to determine whether topic modelling software can detect a shift in the topics of discussion within the correspondence of Samuel Hartlib in the years before and after the cessation of a series of interconnected military conflicts in the years 1648–9. The final case, in the fifth section, investigates whether a variety of Natural Language Processing (NLP) techniques can be deployed to determine whether Jan Amos Comenius used different rhetorical styles to address three different circles of correspondents in his rapidly expanding network during the 1630s.

2 Stylometrics: Verification of Authorship in the Case of Erasmus

Jan Bloemendal

Stylometry is the study of linguistic style, often used to attribute certain anonymous or disputed texts to certain authors. One of the earliest examples of stylometry is the study of the *Donatio Constantini* by the Italian humanist Lorenzo Valla (1407–1457), proving it to be a forgery on historical and linguistic grounds, but without using quantitative methods (*De falso credita et ementita Constantini donatione declamatio*, 1440). In modern times this kind of investigation is done with the help of computers for their capacity to analyse large quantities of data. This requires an adaptation of the definition of style that combines traditional and computational research, which has recently been developed by Hermann, Van Dalen-Oskam, and Schöch: 'Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively'.[1]

The case of Valla itself already indicates the importance of style for humanists. One of the main features of early modern humanism was its preoccupation with language, originating in the idea that a proper use of language is closely connected to right thinking. In an elegant style, its representatives communicated with each other by means of letters. They did so in their universal language, Latin, and thus formed a supranational republic of letters. They shaped a more or less shared Latin style based on the ancient use of the language, and each humanist also had his own style, partially depending on the model used. With regard to prose, a 'battle' was fought between proponents of purely Ciceronian Latin ('Ciceronianism') and those who practised a more eclectic way of dealing with the Latin language and its style.

[1] J. Berenike Herrmann, Karina van Dalen-Oskam, and Christof Schöch, 'Revisiting Style, a Key Concept in Literary Studies', *Journal of Literary Theory* 9:1 (2015): 25–52, at 44. See https://doi.org/10.1515/jlt-2015-0003.

Such quarrels mattered, the more since for humanists Latin was not their mother tongue. Yet, the members of the republic of letters developed their own style. This tension between the common language and style and each humanist's individual style, makes this *respublica litteraria* an interesting case for stylometric analysis.

One of the main representatives of this humanist movement was Desiderius Erasmus (1466?–1536). Stylistically, he was an eclectic, who shaped his own Latin style on the basis of the styles of many ancient authors. Moreover, he wrote about stylistics and the art of letter writing in, among other texts, *Antibarbari* (written 1495, printed 1520), *Lingua* (1525) and *Ciceronianus* (1528). He also wrote and received many letters and wrote about the art of letter writing in *De conscribendis epistolis* (1522). Here, plans for such stylometric research into his style will be presented in pursuit of additional evidence on the authorship of certain texts which have been insecurely attributed to him.

Erasmus wrote over 100 works on the fields of education, ethics, theology, and polemics; also, a vast correspondence of over 3,000 letters survives. All of these texts are written in Latin and the humanist himself undisputedly wrote most of them. However, the authorship of three texts by Erasmus is controversial: *Julius exclusus e caelis*, *De duplici martyrio*, and *Dialogus bilinguium ac trilinguium*.

The first of these, the *Julius exclusus e caelis*, was written between 1513 and 1514, and published in 1517. Its subject is Pope Julius II (d. 1513), who is denied access to heaven when he arrives, drunk, at St Peter's Gate, because he has waged too many wars. Moreover, the keys to heaven that a pope as a successor to Peter has in his possession do not fit. The satire was published anonymously, but rumour spread quickly that Erasmus was the author, which he himself always denied, at least in public discussion. A likely explanation is that he wished to avoid upsetting Julius's successor and Erasmus's patron, Pope Leo X (1475–1521). Historical research has suggested two other candidates for this satirical essay, the German humanist Ulrich von Hutten (1488–1523) and Publio Fausto Andrelini (c. 1562–1518), an Italian humanist and friend of Erasmus. Peter Fabisch in particular advocated the authorship of Fausto Andrelini. The editor of volume I.8 of the *Erasmi opera omnia* (*ASD*), Silvana Seidel Menchi, however, has excellently proven on linguistic and historical grounds that the text was written by Erasmus. An exploration of the text applying authorship attribution software might nevertheless give some extra clues for the otherwise quite certain authorship of Erasmus.

The second text of which Erasmus's authorship is uncertain is the satirical *Dialogus bilinguium ac trilinguium* (1519). Its title page mentions the name of Conrad Nesen (1495–1560) as the author, but Erasmus himself might well be the author of this dialogue in which the *magistri* in Louvain are ridiculed. Here our investigation may corroborate the uncertain attribution to Erasmus.

The third text is the Pseudo-Cyprian *De duplici martyrio*, which only appears in Erasmus's fourth edition of the works of Cyprian of 1530. It is believed that Erasmus himself was the author; but alternatively one of the correctors of the Froben Press, many of whom were scholars themselves, might have been the au-

thor instead. Of these correctors, Sigismundus Gelenius could be the most likely candidate. In this case digital exploration may provide a higher level of certainty for the authorship of Erasmus, Gelenius, or Cyprian himself. In this case, the authorship of Erasmus is not unlikely. Neil Adkin tried to attribute this text by comparing it to the *Paraphrases*.

Scholars agree that Erasmus wrote Latin in his own style. It is characterized by a choice for a more ancient (or 'classical') use of Latin in comparison to medieval Latin. In the debate about the best Latin prose he is an eclectic, and no Ciceronian. Moreover, he has a predilection for diminutives. Tunberg[2] lists some features of Erasmus in grammar, expressions, and vocabulary, from which it becomes clear that in spite of his return to classical Latin, he also uses medieval words. Stylometric analysis may reveal some features that in modern research have previously remained unnoticed.

The following texts will be taken into account: for Erasmus's *Julius exclusus*: *Julius exclusus*, *Utopia*, *Colloquia*, texts by Hutten and by Fausto Andrelini, and Thomas More's *Utopia*; for *De duplici martyrio*: *De duplici martyrio*, *De immensa Dei misericordia*, *Virginis et martyris comparatio*, *Paraphrases*, texts by Cyprian; for *Dialogus bilinguium ac trilinguium* the same texts as for the *Julius exclusus*. For each of the three texts the corpus will be expanded with some more texts. The texts will be analysed using the Stylo Package for R for authorship verification (Eder, Rybicki, and Kestemont),[3] in combination with *AntConc*, developed by Laurence Anthony.

For analysing Erasmus's style, a substantial corpus of texts will be selected and analysed with the use of these programmes. For our analysis we shall look, for instance, at specific vocabulary used by Erasmus and others, as well as the frequency of conjunctions such as *quod* and *quia* (that), *quod*, *quia*, and *cum* (because), *et*, *atque*, *ac* (and), *at*, *autem*, *sed* (but, however), and *quamvis* and *quamquam* (although). For example, one of Erasmus's idiosyncrasies seems to be a preference for *ac* above *et*. This stylistic research may also discover additional particularities of Erasmus's style.[4]

These methods may also be useful in helping to identify the writers of anonymous letters, provided that the body of letters in question is substantial enough for these methods to be reliably applied and that an equally large body of securely attributed letters is also available. As in the case of entire works, it will also be necessary that computational evidence is substantiated with reference to more traditional historical scholarship and vice versa. These techniques can be applied for any language; and it may also sometimes be possible to deduce the nationality of a

[2] Terence Tunberg, 'The Latinity of Erasmus and Medieval Latin: Continuities and Discontinuities', *The Journal of Medieval Latin* 14 (2014): 147–70, at 162–6. See https://doi.org/10.1484/J.JML.2.304219.

[3] Maciej Eder, Jan Rybicki, and Mike Kestemont, 'Stylometry with R: A Package for Computational Text Analysis', *R Journal* 8 (2016): 107–21, https://journal.r-project.org/archive/2016/RJ-2016-007/index.html, accessed 20/03/2019.

[4] With thanks to Karina van Dalen-Oskam and Charles van den Heuvel.

writer from his Latin style, since Latin usage sometimes reveals the syntactical and grammatical habits imprinted by a mother tongue. As the available corpus of digital texts expands, stylometrics and author recognition will have much to offer.

3 Text Reuse in the Republic of Letters

Sinai Rusinek

With contributions from Tommaso Elli, Giovanni Moretti, and Ludovica Marinucci

'Text reuse' denotes a broad range of phenomena, from citations and direct quotations through paraphrases and hidden references to dissemination of information, whether verbatim or not. Its detection, therefore, also presents a bigger computational challenge than mere plagiarism, which normally refers to the use of more or less exact phrases or extended textual passages. In recent years, text reuse algorithms and tools have been applied to various textual genres – from ancient poetry and literature[5] to the *Encyclopédie*,[6] to modern fiction,[7] legal documents,[8] and newspaper collections,[9] but not, to our knowledge, to correspondence collections.

[5] E.g. Bridget Almas and Monica Berti, 'Perseids Collaborative Platform for Annotating Text Re-uses of Fragmentary Authors', in *DH-CASE 13: Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies and Techniques in the Digital Humanities, 10 September 2013, Florence, Italy* (Florence: ACM Press, 10 September 2013), article no. 7, see https://doi.org/10.1145/2517978.2517986; Neil Coffee, Jean-Pierre Koenig, Shakthi Poornima, Christopher W. Forstall, Roelant Ossewaarde, and Sarah L. Jacobson, 'The *Tesserae* Project: Intertextual Analysis of Latin Poetry', *Literary and Linguistic Computing* 28:2 (1 June 2013): 221–8, see https://doi.org/10.1093/llc/fqs033; Patrick J. Burns, 'Measuring and Mapping Intergeneric Allusion in Latin Poetry Using *Tesserae*', *Journal of Data Mining and Digital Humanities* (2016), see https://hal.archives-ouvertes.fr/hal-01282568/document, accessed 20/03/2019.

[6] Mark Olsen, Russell Horton, and Glenn Roe, 'Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections', *Digital Studies/Le Champ Numérique* 2:1 (17 May 2011), see http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/190, accessed 20/03/2019.

[7] Grace Muzny, Mark Algee-Hewitt, and Dan Jurafsky, 'Dialogism in the Novel: A Computational Model of the Dialogic Nature of Narration and Quotations', *Digital Scholarship in the Humanities* 32 Supplement 2:1 (2017): ii32–ii52, see https://doi.org/10.1093/llc/fqx031; Douglas Ernest Duhaime, 'Textual Reuse in the Eighteenth Century: Mining Eliza Haywood's Quotations', *Digital Humanities Quarterly*, 10:1 (2016), see http://www.digitalhumanities.org/dhq/vol/10/1/000229/000229.html, accessed 20/03/2019.

[8] Lincoln Mullen, 'Detecting Text Reuse in Nineteenth-century Legal Documents', blog, 11 March 2015, see http://lincolnmullen.com/blog/detecting-text-reuse-in-legal-documents/, accessed 20/03/2019.

[9] Ryan Cordell and David Smith, *Viral Texts: Mapping Networks of Reprinting in 19th-century Newspapers and Magazines* (2017), see http://viraltexts.org, accessed 20/03/2019; Giovanni Colavizza, Mario Infelise, and Frédéric Kaplan, 'Mapping the Early Modern News Flow: An Enquiry by Robust Text Reuse Detection', in Luca Maria Aiello and Daniel McFarland, eds., *Social Informatics. SocInfo 2014* (Cham: Springer, 2015), 244–53, see https://doi.org/10.1007/978-3-319-15168-7_31.

The motivations for studying text reuse are as many, as varied, and as central to the humanities as they have been for over two millennia of textual scholarship: the social character of language and culture are embedded in the art of text reuse, where creativity is in dialogue with tradition. This is all the more true when text reuse is studied in early modern humanistic culture, where cross references, shared and exchanged textual traditions constitute the warp and woof which wove the republic of letters together. The broad and elusive nature of text reuse then creates a further challenge for visualization, exploration, and analysis with computational means and methods.

The case study described here commenced in a COST-funded short-term scientific mission (STSM) in Göttingen and Leipzig in early 2016. The STSM was dedicated to experimenting with the use of the *Tracer* tool,[10] a text reuse detection package developed by Marco Büchler, on the epistolary corpus created in the *ePistolarium* by the project *Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic*.[11]

Text reuse detection is typically motivated by the search for commonplaces, quotations, and citations, or evidences for various practices of recycling larger units of text.[12] Examples for these types of text reuse were found also in the *Tracer* results of our corpus: for example, the words 'Amare liceat, si potiri non licet' were found in two of Hugo Grotius's letters – one addressed to his brother Willem and the other to Ludwig Camerarius, the German statesman who at the time was representing Sweden in The Hague. This is a quotation of the opening and closing line of the poem *Anechomenos*, by Apuleius. Longer sections of what appeared to be 'reuse' were also found: one case revealed two copies of the same letter by Hugo Grotius to Paul du May from 9 April: one dated 1635 and the other 1637 (one of which is obviously erroneous). The current edition preserved the duplication that had previously appeared in Grotius's own edition of his correspondence.[13] Another example is a reattachment, for reference, of the same postscript to two separate letters.[14]

The overwhelming majority of the detection results, however, yielded a rather different type of text reuse which is abundant in letter writing and remains, to our knowledge, unexplored by digital means: namely, formulaic salutations. The large

[10] Marco Büchler, Philipp R. Burns, Martin Müller, Emily Franzini, and Greta Franzini, 'Towards a Historical Text Re-use Detection', in Chris Biemann and Alexander Mehler, eds., *Text Mining, Theory and Applications of Natural Language Processing* (Switzerland: Springer International Publishing, 2014), 221–38, see https://doi.org/10.1007/978-3-319-12655-5_11.

[11] Huygens ING-CKCC, *Project Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic. A Web-based Humanities Collaboratory on Correspondences* (Geleerdenbrieven, 2011), archived version d.d. 2013-07-23. DANS. See https://doi.org/10.17026/dans-xfd-n8y5.

[12] Respective examples for these motivations are the projects https://commonplacecultures.org, http://americaspublicbible.org/, and http://viraltexts.org, all accessed 20/03/2019.

[13] Philip C. Molhuysen, Bernardus L. Meulenbroek, Paula P. Witkam, Henk J. M. Nellen, and Cornelia M. Ridderikhoff, letters 3018 and 2053. I thank Henk Nellen and Charles van den Heuvel for clarification of the matter. See http://grotius.huygens.knaw.nl, accessed 20/03/2019.

[14] Ibid., letters 6350 from 7 August 1643 and 6696 from 6 February 1644.

number of these recurring formulae called for distant reading approaches, combining quantitative analysis with the power of visualization.

In the digital humanities, network analysis has been used, first and foremost, to study historical relations. Edges in these networks often represent social relations, or more abstract relations such as citation or co-citation in a document. In most cases the nodes are human actors. More recently, network analysis software has been found helpful also in modelling texts and relations between them, as in the Stylo package for computational stylistics,[15] which enables exporting results as edges for a network analytic study. In this case too, network visualization enables not only the evaluation of the results of detection but also the analysis and visualization of the flow of ideas, tropes, motifs, quotes, and commonplaces in the corpus.

In a network representation of text reuse, phrases, rather than correspondents, can serve as nodes, and their similarity is expressed by edges. In the following two figures, *Cytoscape* software[16] was used to visualize results of text reuse detection in the correspondence of Hugo Grotius. Two parameters for visualization help in revealing patterns in the result. The first one is the colour: in figure 1, a cluster of salutations of Grotius by the queen Christina of Sweden is coloured red. The cluster in dark and light blue and in light and dark grey represents salutations within Grotius's circle of statesmen and diplomats, including the privy councillor Johann Oxenstierna and Schering Rosenhane. Correspondence with his father and brother create patterns such as the lower cluster in dark and olive green. The cluster of nodes in light green are not a proper text reuse, but repetition of the five-word title pattern 'De Iure belli et/ac pacis', from references Hugo Grotius makes to his book.

To readers of early modern correspondences, the repetitiveness of these textual nodes is no surprise. It represents the practice of using regular templates for salutations (e.g. by the queen's secretary), or the formalities expected in addressing specific correspondents (even father or brother); but a graph visualization offers an encompassing view of these textual-social relations and enables a study of the social fabric of the correspondents more refined than a mere network of correspondents.

---

15 Eder, Rybicki, and Kestemont, 'Stylometry with R'.

16 Paul Shannon et al., 'Cytoscape: A SoftwareEnvironment for Integrated Models of Biomolecular Interaction Networks' *Genome Research* 13:11 (November 2003): 2498–504, see https://doi.org/10.1101/gr.1239303, analyse modularity and visualize the results. Deeper exegesis will reveal the connections between the cluster parameters and the characteristics of the textual phenomena they represent.

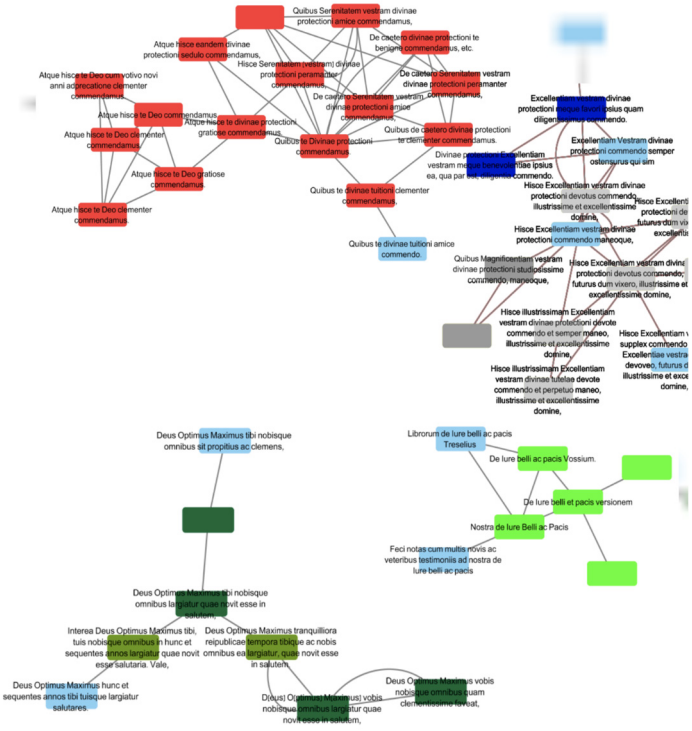


Figure 1: Detail of *Cytoscape* graph visualization: clusters coloured by author

In April 2016, a team composed of a designer (Tommaso Elli), a developer (Giovanni Moretti), and two humanities scholars (Ludovica Marinucci and Sinai Rusinek) met in Como for a design sprint.[17] We conceived of a pipeline of interactive visualization tools that would enable a more interactive and flexible interface for the study of text reuse in a correspondence corpus. In addition to the network view in the first phase of visualization, the pipeline included a second phase in which each cluster of text repetition detected in the network can open to a variation graph, which visualizes similarities and differences between the text units. The graph, on the left side of figure 2, is using an implementation of Stefan Jaenicke's *TRAViz* tool.[18] The interrelated graph to the right shows that the formulae visual-

[17] COST Action IS1310 *Reassembling the Republic of Letters*, Visualization Meeting, COMO, IT (April 2016) Group 1: Seeing Echoes: Visualizing Text Reuse in Correspondence. For the event description and group report, see http://www.republicofletters.net/wp-content/uploads/2017/02/Como-Notes-COST-Action-IS1310-Reassembling-the-Republic-of-Letters.pdf, accessed 20/03/2019.

[18] Stefan Jänicke, Annette Geßner, Marco Büchler, and Gerik Scheuermann, 'Visualizations for Text Re-use', in Luca Maria Aiello and Daniel McFarland, eds., *Social Informatics: SocInfo 2014 International Workshops, Barcelona, Spain, November 11, 2014, Revised Selected Papers* (Cham: Springer, 2015), 59–70, see https://doi.org/10.5220/0004692500590070.

ized is repeated in a correspondence of the queen of Sweden with her brother and with Hugo Grotius. Clicking a node in the graph to the right would colour the respective variant in the graph to the left, and vice versa.

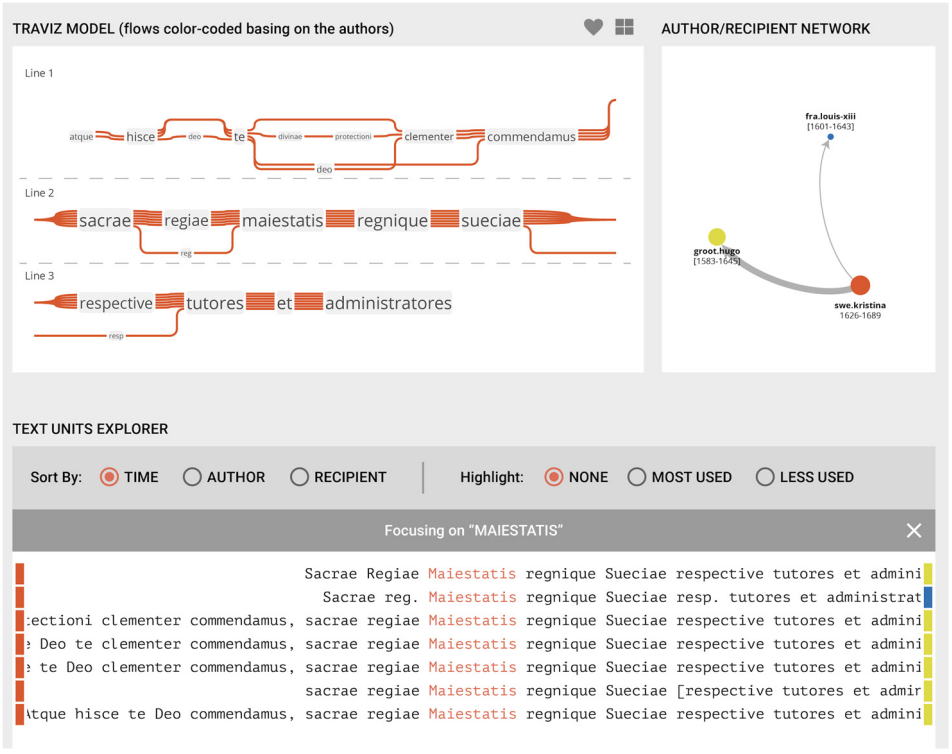


Figure 2: Integrated view of variation visualization, graph, and concordance for explorations of text reuse. Designed by Tommaso Elli, Giovanni Moretti, assisted by *TRAViz* developer, Stefan Jänicke

In the lower section of this screen, the reuse cluster can be viewed in KWIC (*Key-Word in Context*) display, with various sorting and highlighting options available and side panels coded to indicate the author's and recipient's identities.

This experiment was carried on a specific ego-network correspondence yielding over 60,000 text reuse candidate results. Larger corpora of letter collections will provide an ample field for the study of text reuse phenomena, but will also require both macroscopic and microscopic tools for exploration. In using this set of three interrelated tools, a future platform for studying large corpora of letters could provide a smooth combination of and passage between distant and close reading of text reuse detection results.

4 Topic Modelling: Hartlib's Correspondence before and after 1650

Barbara McGillivray, Robin Buning, and Simon Hengchen

The Hartlib Papers is the collection of correspondence and other papers of the 'intelligencer' Samuel Hartlib (c. 1600–1662). The epistolary corpus consists of 4,833 letters written by some 325 correspondents, spread mostly over western and central Europe, and consisting of people with diverse socio-economic backgrounds, such as refugees, merchants, churchmen, academics, inventors, diplomats, and artisans. Not only are the letters informative of Hartlib's network, they also are a rich source for the study of intellectual life in seventeenth-century Europe, covering a wide spectrum of topics and spanning the relatively lengthy period of forty-two years, from 1620 to 1662. The geographical spread of the network and the large number of non-academics among the correspondents also accounts for the multilinguality of the corpus, consisting of 3,015 English letters, 527 German letters, 859 Latin letters, eighty-four French letters, three Dutch letters, and 345 letters mixing multiple languages.

The size of the collection, its multilinguality, and the great variety of subjects discussed in the Hartlib Papers, however, make the collection difficult to access for the individual researcher. Exploring this large quantity of epistolary data by using methods from NLP and Computational Linguistics to analyse the letters' topic distribution and evolution will help researchers to mine the content of the Hartlib Papers and to understand better the interests shared by Hartlib's correspondents. More broadly, the development of such methods, while being tested in the challenging context of the Hartlib Papers, offers the opportunity to develop a general methodology that can be applied to other epistolary corpora and languages.

Following automatic preprocessing of the letters in English (cleaning the annotation markup, lemmatization, language detection, and part-of-speech tagging), work initially focused on the 3,070 files in (early modern) English.[19] The lemmatization was achieved through *Morphadorner*,[20] although tests were also carried out with historical spelling normalization tool *VARD2*,[21] which yielded very similar results. In the context of a short-term scientific mission of our COST network, topics were extracted from the collection of all the verbs, adverbs, nouns, and adjectives present in the original files using Latent Dirichlet Allocation (LDA) in its JAVA implementation with a graphical user interface.[22] After excluding a stopword list (consisting of *MALLET*'s standard list plus *man*, *thing*, *mr*, and *sir*), we ran LDA

[19] The code created in the framework of this project is available at https://github.com/alan-turing-institute/hartlib (https://doi.org/10.5281/zenodo.1040682), accessed 20/03/2019.

[20] For details, see http://morphadorner.northwestern.edu/morphadorner/ accessed 20/03/2019.

[21] Available on http://ucrel.lancs.ac.uk/vard/about/, accessed 20/03/2019.

[22] See https://doi.org/10.5281/zenodo.30704.

on the data with different configurations for the number of topics: ten, twenty, thirty, and forty. After an analysis by our team's domain expert, we decided to proceed with the forty-topic configuration.

In order to detect changes in the topics that were discussed in the letters, the corpus was divided into two time periods based on historical grounds: the first period spans the years between 1620 and 1650, and the second one the years 1651 to 1662. The year 1650 was chosen as the dividing line because around the middle of the seventeenth century there were major changes in the political situation of Europe, which had some direct consequences for Hartlib and the people in his network, most importantly the end of the Thirty Years' War and the Dutch Revolt in 1648, and the civil wars in Scotland, Ireland, and England in 1651.

This analysis confirmed some expected results, such as a decrease in the extent to which war was discussed. Others cannot be immediately explained and require further research, such as an increase in discussion of letter writing. Generally, a shift can be seen from preoccupation with war and religion towards discussion of physics, chemistry, medicine, agriculture, and horticulture. For example, a significant decrease is seen with regard to the topic that we labelled *army/military engineering/troop movement* with associated words *army town place horse ship soldier war engine bring good*. An obvious explanation for this decrease is the end of the wars in Protestant Europe and the British Isles. A significant increase is seen with regard to the topic *medicine: (iatro) chemical experiments* with the associated words *id make clodius great cure experiment excellent lb kind secret*. The top ten documents associated with this topic actually are not letters, but sections of Hartlib's *Ephemerides*, the diary of the information he received, which, on the other hand, largely consists of excerpts from his correspondence, making the *Ephemerides* a relevant source. In six of the top ten sections that our analysis yielded, Hartlib's son-in-law Friedrich Clodius, who as of 1653 carried out chemical experiments in Hartlib's house, is prominently present. In the *Ephemerides* Hartlib indeed frequently refers to chemical experiments carried out by Clodius or mentions him as a source of news on experiments carried out by others.

Although preliminary in their nature, these results – obtained with very modest means over a short period of time – suggest that far more might be accomplished by more concerted efforts sustained over a longer period of time. This research has also shed light on several technical aspects that should be considered in the framework of a large-scale digital humanities project on correspondence. Specifically, the format and availability of metadata, as well as limited consistency of the annotation and metadata linking, presented challenges for the cleaning and preprocessing of the texts. For example, consistent annotation would have made automatic identification of editors' notes from the original letter texts quicker, while metadata linking would have allowed us to connect the files that stored the letters' texts with information on the language represented in them. This experience is particularly valuable for the planning of future digital edition projects, where such consistency and linking would be recommended.

Future research could also expand in several directions. One would be to extend the methodology to the letters written in other languages, starting with Latin. This is made possible by the automatic nature of our research process and will allow us to compare the distribution of topics across languages, shedding new light on the association between language of letters and topic coverage. Another opportunity is to study semantic change in the letters by analysing the words associated with each topic over time, thus exploring linguistic choices, conceptual change, and language change in the letters, similarly to what is proposed in the following study of Hartlib's close associate, Jan Amos Comenius.

5 Natural Language Processing: Shifting Rhetorical Strategies in Comenius's Correspondence with Three Separate Communities

Mihai Dascalu, Lucie Storchová, Stefan Trausan-Matu, and Vladimír Urbánek

The extant correspondence of Jan Amos Comenius (1592–1670) includes 569 letters of which 450 are sent and 119 received. With the exception of two early dedicatory epistles, these letters all fall within the period between 1628 and 1670. As the focus of a brief, collaborative project, a selection of the complete correspondence was chosen: namely, those letters sent between 1630 and 1642.[23] This was a crucial period in Comenius's career: especially after publication of his enormously popular *Janua linguarum reserata* (1631), Comenius's correspondence contacts grew rapidly and he became more integrated into the European republic of letters.

The expansion and diversification of his correspondence posed a rhetorical challenge. As Comenius developed very different relationships with several distinct scholarly communities, he needed to diversify his epistolary styles and to develop rhetorical strategies appropriate to each. This posed the key research questions for this brief project. Did Comenius employ different rhetorical and stylistic strategies and discuss different major topics with each of these communities?[24] Can these differences be related to the formal characteristics of these three groups of letters? More specifically, can the qualitative conclusions of traditional historical modes of interpretation be combined with and enriched by the application of a variety of quantitative, NLP techniques to this diverse epistolary corpus?

[23] Most of the letters from this period have been published in the *J. A. Comenii Opera Omnia*, vol. 26/I (Prague: Academia, 2018). We used the texts prepared for this critical edition by Martin Steiner, Marcela Slavíková, Kateřina Šolcová, and Markéta Klosová.

[24] Vladimír Urbánek, 'J. A. Comenius and the Practice of Correspondence Networking: Between the Office of Address and the Collegium Lucis', in Wouter Goris, Meinert A. Meyer, and Vladimír Urbánek, eds., *Gewalt sei ferne den Dingen! Contemporary Perspectives on the Works of John Amos Comenius* (Wiesbaden: Springer VS, 2016), 291–308, see https://doi.org/10.1007/978-3-658-08261-1_19.

To sharpen up the profile of the study, a total corpus was chosen consisting of thirty-three letters addressed to three distinct communities.

One community consisted of German educational reformers (hereafter GER), notably Sigismund Evenius (1587–1639), Johannes Docemius (d. 1638), Martin Moser (d. 1636), and Georg Winkler (b. 1566). Letters to this group focused on three major topics: a new method of teaching languages, re-publications of *Janua linguarum*, and work on *Didactica magna*. Comenius developed various strategies for presenting the idea of an educational reform as a collective project, using a number of rhetorical figures to express his enthusiastic attitude, the necessity of cooperation, and mutual agreement among scholars. He stressed civility, temperance, and the other virtues of the republic of letters and presented himself as a modest person, who owes his success not only to himself but also to his learned friends and God. At the same time, however, he fashioned himself in a self-confident way and never hesitated to criticize other scholars' methods of teaching and learning Latin.

A second group chosen was the Danzig circle (hereafter DC) surrounding Johann Mochinger (1603–1652), Wojciech Niclassius (1592–1651), and Martin Opitz (1597–1639). Comenius used a brief, strict, and self-confident style in his letters to the DC, which typically conveyed instructions and recommendations regarding the translation and publication of his didactic works. For example, he praised Mochinger's German translation of the *Janua linguarum* but also sent him detailed instructions on how to improve it. Other letters contain expressions of disappointment with the delay in publication, insistence on exclusive collaboration with a specific printer, and instructions for checking the quality of the printer's work. The only person in DC with whom Comenius communicated in a more eloquent humanist style was Martin Opitz, the most renowned German poet of this time.

A third crucial group comprised the circle of Samuel Hartlib (hereafter HC) and his close associates, including Joachim Hübner (1611–1666), Godefroid Hotton (1596–1656), and Johann Moriaen (c. 1591–1668). Unlike the letters to GER and DC, those to HC covered a broader agenda, including Comenius's early pansophic plans. Two typical features of his letters to Hartlib are the eloquent style and a shared emotional code, which differ considerably from his other correspondence of this period. Stylistically, the structure of sentences becomes more complicated. Rhetorically, Comenius addressed Hartlib as his 'most beloved brother' or 'honey-sweet friend' and describes his personal attitude to him as 'open hearted'. This emotional code refers, for instance, to kisses using phrases like 'you are my soul' or 'I love you so much'. Letters become highly desired objects because they substitute for direct contact with the most beloved person.

After undertaking this traditional, historical, and rhetorical analysis, these same texts were analysed using the NLP framework *ReaderBench*.[25] The preliminary step

[25] *ReaderBench* (http://readerbench.com, accessed 20/03/2019) is an open-source advanced, multilingual, NLP framework centred on comprehension prediction that integrates a wide range of textual complexity indices. *ReaderBench* introduces Cohesion Network Analysis (CNA) that considers cohesive links quantified using different semantic models between different text segments and operation-

was to adapt *ReaderBench* for use on Latin texts. Three dimensions of analysis were considered. First, a comprehensive list of textual complexity indices was selected in order to reflect specific features of writing style. Second, specific metrics for text cohesion and discourse connectivity were introduced and tailored for the current Latin language processing. Third, new semantic spaces denoting concepts and semantic distances among them were trained using state-of-the-art methods on a collection of approximately 7,000 letters written in Latin (approximately 2.7 million words) containing a part of Comenius's extant letters as well as Latin letters from prominent Dutch scholars of the same period (taken from the *ePistolarium*).

Afterwards, statistical analyses focused on lexical and semantic features were performed to investigate the differences in the writing style of letters addressed to the three communities. After some preliminary checks, a multivariate analysis of variance was conducted to examine whether the lexical and semantic features differed between the three groups of letters.[26] These analytical methods revealed significant differences between the three groups, and in six out of the seven measures (illustrated in fig. 3) Comenius's letters to HC were clearly differentiated from those to the other two groups. In the first place, they were considerably longer, whether measured simply by counting either the words (3.a) or the sentences (3.b). Second, they had a more varied vocabulary containing more word types (3.c): known as 'higher entropy', this variety is measured by the logarithm of the probability distributions of different word occurrences. Third, the letters to HC are more cohesive overall (3.d), that is, they display a higher cohesion in terms of semantic relatedness between paragraphs. Fourth, they contain more elaborate sentences: this was measured using Cohesion Network Analysis (3.e), which is based on the number of connectives and conjunctions (3.f). In all these respects, the letters addressed to members of HC stood apart from the rest. The only exception to this pattern is the number of reason and purpose connectors per paragraph (3.g), in which the highest values were exhibited for the GER community. This finding is in line with the scholarly interpretations of the three communities, since the letters to GER were devoted to specific projects (such as the publication of the various versions of *Janua linguarum*), together with their corresponding justifications. The lowest index values were observed for the DC, which is in keeping with Comenius's strict and self-confident style in his letters to this group.

alizing Trausan-Matu's polyphonic model. For a general introduction, see Mihai Dascalu, Philippe Dessus, Maryse Bianco, Stefan Trausan-Matu, and Aurélie Nardy, 'Mining Texts, Learner Productions and Strategies with *ReaderBench*', in Alejandro Peña-Ayala, ed., *Educational Data Mining. Applications and Trends* (Cham: Springer, 2014), 345–77, see https://doi.org/10.1007/978-3-319-02738-8_13. On the polyphonic model, see also Mihai Dascalu, Stefan Trausan-Matu, Danielle S. McNamara, and Philippe Dessus, '*ReaderBench:* Automated Evaluation of Collaboration Based on Cohesion and Dialogism', *International Journal of Computer-Supported Collaborative Learning* 10:4 (December 2015): 395–423, see https://doi.org/10.1007/s11412-015-9226-y.

[26] MANOVA was employed for this purpose: a statistical method to identify differences on one continuous dependent variable by an independent grouping variable.

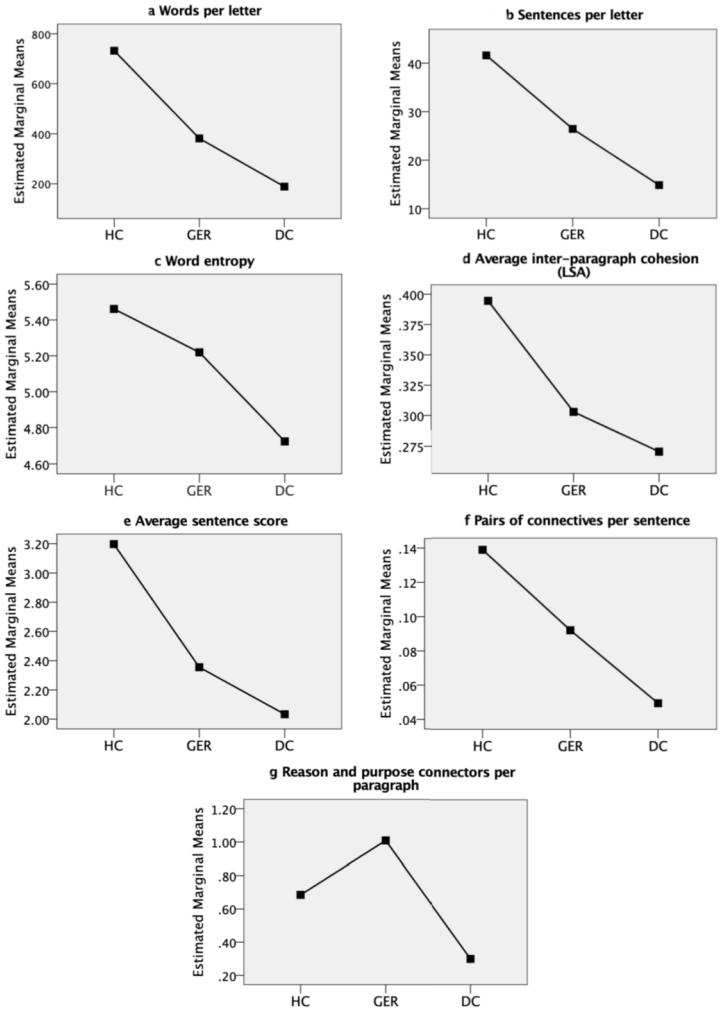


Figure 3.a–g: Comparative views of writing styles between the three communities reflected in textual complexity indices

A further phase of analysis provided an even more synthetic comparison. In this phase, a stepwise Discriminant Function Analysis (DFA) was performed in order to determine the community of a given letter based on the underlying writing style features. A DFA is in general used to classify individuals into predefined groups (in our case the three communities) and can be considered a multivariate analogue to variance analysis. The DFA retained two canonical discriminant functions (i.e. latent variable created as a linear combination of independent variables – textual complexity indices) based on three indices that were identified as significant

predictors, namely: (1) word entropy; (2) standard deviation of sentence scores; (3) average number of reason and purpose connectors per paragraph. The resulting two canonical discriminant functions lead a 66.7 per cent accuracy in the separation of writing styles between communities, depicted in figure 4. In laymen's terms, this means that there is very little overlap in figure 4 between the three clusters of coloured dots representing Comenius's letters to each of these three communities.

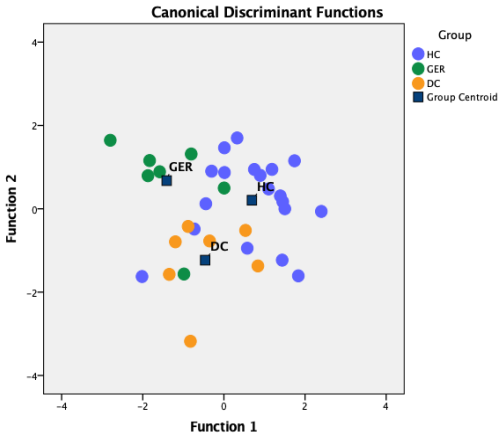


Figure 4: Separation of communities using the two canonical discriminant functions obtained from the DFA

This project posed two basic questions, one substantive, the other methodological. First, did Comenius employ different rhetorical and stylistic strategies in addressing different groups? And second, can NLP techniques provide a quantitative foundation for a qualitative scholarly answer to the first question? The affirmative answers to both of these questions open up a broad field for further experimentation. One option would involve testing the same approach on differently configured groups of correspondents, defined, for example, by family relations, confessions, disciplines, or common sources of patronage. Another possibility would be to look for long-term changes in lexical choices, rhetoric techniques, and other formal characteristics of language over the course of decades-long epistolary exchanges, such as those between Comenius and Hartlib. At the most general level, combination of these two analytical approaches may advance the study of broader intellectual constellations and networks within the republic of letters as a whole by detecting differences and similarities in the linguistic means used by different groups of scholars in different times and places to represent themselves, their allies, and their enemies.

IV.7 Virtual Research Environments for the Digital Republic of Letters

Meliha Handžić and Charles van den Heuvel

1 Introduction: Digital Letter Editions and Research

Digital technology is transforming scholarship in the humanities. Increased engagement with technology is giving scholars unprecedented opportunities for significant intensification and diversification of their research activities. At the same time, translating traditional humanistic objectives, materials, methods, and activities into the digital domain poses fresh challenges for humanistic study and practice. Seizing these opportunities and overcoming these challenges fully will require the transformation of the environment within which much humanistic research and study is undertaken, especially in order to facilitate engagement with unprecedented quantities of complex data and metadata. More particularly still, a Virtual Research Environment (VRE) is needed to support that subset of the humanities community engaged in innovative and collaborative study of the republic of letters.

The main aim of the COST Action *Reassembling the Republic of Letters* (RRL) has been to devise means for bringing together as many early modern learned letters and related documentation as possible in a digital format as the precondition for analyzing wider networks of intellectual exchange, eventually expanding to embrace the republic of letters as a whole. New tools are needed for every stage of this process, from assembling, exchanging, and reconciling catalogue metadata to the transcription and annotation of letters to create new, 'born digital' collections (see chs. III.1–5).

At this point, however, we confront a paradox, which necessitates further reflection regarding the nature and function of a VRE. The outcome of these plans will be to inundate the field with digital data; in order to handle this deluge of data, methods and tools such as text mining and topic modelling will be needed; and yet (as established in ch. III.4) these tools and methods can only yield meaningful results if even larger quantities of data are available for mining, modelling, and analysis. When confronted by data sets large enough for reliable text mining and topic modelling, scholars will urgently need new strategies and environments in order to explore and analyse these unprecedented bodies of data and metadata effectively. It is therefore necessary to consider at the outset the manner in which big data on the republic of letters must be organized and accessed if future scholars are going to be able to process this information individually and collectively, cognitively and tacitly, to make it explicit by documenting, publishing, and sharing it.

In one sense, this problem is not new. As Ann Blair has made clear in her seminal book, *Too Much to Know*, scholars since Antiquity have developed strategies for producing and managing information in a collaborative way. Similar strategies are under development today, in the early digital age. For the *production* of knowledge, for instance, crowdsourcing is becoming a common method: to advance our knowledge of the republic of letters, whole communities of scholars are cataloguing, transcribing, disclosing, annotating, and sometimes even coding collections of letters. For the *analysis* of the resulting data, similar collaborative methods will be needed. To this end, this chapter conceptualizes a future VRE for analysing digitally reassembled data on the republic of letters. The general objective is to sketch the outline of an environment in which scholars can navigate, explore, search, analyse, visualize, interpret, and validate the unstructured and structured big data of the digital republic of letters. Although many of the current methods and tools might be obsolete by the time a comprehensive data set has been assembled, the fundamental principles of knowledge organization and management will still be needed to handle large quantities of letters and contextual texts for scholarly purposes.

2 Delimiting the Knowledge Space

A natural starting point for this discussion is provided by the most common way of dealing with the overload of information, namely, abstraction. Collections are abstracted by making indexes composed of keywords considered to be representative of the contents of larger texts. This creates metadata, that is, information about information, or data about data. Metadata-based indexing of this kind can then be used for navigating, searching, and analyzing the essential elements of huge quantities of information.

Many of the basic resources fundamental to large-scale research on the republic of letters consist primarily of catalogues of metadata of this kind. Examples

include the *Catalogus Epistularum Neerlandicarum*, *Early Modern Letters Online*, and *Kalliope* (see ch. III.1). Other resources which also include images and texts of letters – including *ArchiLet*, the *Electronic Enlightenment*, *e-manuscripta*, the *ePistolarium* – use metadata of this kind as the primary means of searching, sorting, and navigation.

The Linked Open Data paradigm is a logical step in a further integration of structured historical data of the republic of letters. Yet this promising approach is nevertheless subject to two important limitations.

On the one hand, for the time being at least, Linked Open Data has notoriously bad user interfaces, especially for humanities scholars not familiar with the complex SPARQL languages to query these data. Promising experiments are currently being undertaken with 'data lenses' or 'data scopes' that allow the interactive exploration from various perspectives of multiple representations of the same object within a large cloud of linked data on the republic of letters.[1] However, to get beyond the stage of metaphor in organizing and exploring the big data of the republic of letters, a more holistic approach is needed.

More important, however, is a second limitation: the access to the textual resources offered by systems based on structured metadata is intrinsically limited by the fact that such metadata is an abstraction of the textual contents of the future reassembled republic of letters itself. In other words, the greater part of the contents of the republic of letters cannot be disclosed by structured metadata. Once again, this insight is not new. Already over half a century ago, key figures in the historiography of library and information sciences, such as Vannavar Bush in 'As We May Think' (published in *Atlantic Monthly* in July 1945), declared that indexing would be far too limited for disclosing the deluge of information in the digital era and that search by association rather than keyword queries better reflects the cognitive processes necessary for accessing the desired knowledge.

It is for this reason that we need both methods for the disclosure of large quantities of unstructured textual data via the development of intuitive user interfaces.[2] Combinations with the *ePistolarium* and *ReaderBench* based on text analyses and Natural Language Processing techniques are necessary to get access to the much larger contents of the correspondences and their contextual information in unstructured data.

In order to implement a VRE around the digitally assembled republic of letters we are first of all in need of a knowledge space in which structured and unstructured data can be organized. For this purpose, further reflection on the 'knowledge space' within which research in this field must be situated is helpful. To get a grip

[1] Eetu Mäkelä, Eero Hyvönen, and Tuuka Ruotsalo, 'How to Deal with Massively Heterogeneous Cultural Heritage Data-lessons Learned in CultureSampo', *Semantic Web* 3:1 (2012): 85–109, see https://doi.org/10.3233/SW-2012-0049.

[2] Charles van den Heuvel et al. 'Deep Networks as Associative Interfaces to Historical Research', in Florian Kerschbaumer, Linda von Keyserlingk, Martin Stark, and Marten Düring, eds., *Power of Networks. Prospects of Historical Networks Research* (Oxford: Routledge, 2019, in press).

on big data and to organize knowledge in a meaningful way, library and information scientists have made use of all sorts of metaphors drawn from a long tradition. As already noted in chapter II.5, since Antiquity, knowledge has tended to be organized and classified in spatial terms. More recently, since probably the first modern historiographical overview of the library sciences by Henry Bliss in 1929, this ancient tradition has been further developed using the 'universe of knowledge' metaphor,[3] which has been further elaborated as a 'universe of concepts'[4] and the 'multiverse of knowledge'.[5] Still more useful for present purposes is the proposal to extend the metaphor of knowledge spaces to attempt to formulate the analogue to the laws of physics operating within those spaces. For instance, the 'gravitational forces' in these knowledge universes are used metaphorically to explain important concepts in the theory of classification such as 'likeness' and 'likeliness',[6] which are explained at further length in section 4 of this chapter. These metaphors might seem at first sight far-fetched, but in computer and information virtual spaces are quite common. For instance, vector space models are used in topic modelling of text to organize and measure the distance between large quantities of words that more likely and less likely belong to each other to explain specific concepts and topics. The creation of 'virtual space models' is essential to bridge the gap between tacit knowledge (in the minds of researchers) and explicit knowledge, between close and distant reading of unstructured and structured data of the republic of letters in digital format.

In short, we are in need of conceptualizations of how to organize the big data of the humanities, and more particularly, of the republic of letters, to make them

[3] Henry E. Bliss, *The Organization of Knowledge and the System of the Sciences* (New York: H. Holt and Co., 1929); and Charles van den Heuvel, 'Multidimensional Classifications: Past and Future Conceptualizations and Visualizations', *Knowledge Organization* 39:6 (2012): 446–60, see https://doi.org/10.7152/nasko.v3i1.12795.

[4] Shiyali R. Ranganathan, *Prologomena to Library Classification*. 2nd edn. (London: The Library Association, 1957); Francis L. Miksa, 'The Concept of the Universe of Knowledge and the Purpose of LIS Classification', in Nancy J. Williamson and Michele Hudon, eds., *Classification Research for Knowledge Representation and Organization: Proceedings of the 5th International Study Conference on Classification Research* (Toronto, Amsterdam, and Würzburg: Ergon Verlag, 1992), 161–78; Clare Beghtol, 'From the Universe of Knowledge to the Universe of Concepts: The Structural Revolution in Classification for Information Retrieval', *Axiomathes* 18:2 (2008): 131–44, see https://doi.org/10.1007/s10516-007-9021-0.

[5] Charles van den Heuvel and Richard P. Smiraglia, 'Concepts as Particles: Metaphors for the Universe of Knowledge', in Claudio Gnoli and Fulvio Mazzocchi, eds., *Paradigms and Conceptual Systems in Knowledge Organization: Proceedings of the Eleventh International ISKO Conference, 23–26 February 2010, Rome, Italy* (Würzburg: Ergon Verlag, 2010), 50–6; Richard P. Smiraglia, Charles van den Heuvel, and Thomas M. Dousa, 'Interactions between Elementary Structures in Universes of Knowledge', in Aida Slavic and Edgardo Civallero, eds., *Classification & Ontology : Formal Approaches and Access to Knowledge. Proceedings of the International UDC Seminar, 19 - 20 September 2011, The Hague, The Netherlands* (Würzburg: Ergon-Verlag, 2011), 25–40.

[6] Charles van den Heuvel and Richard Smiraglia, 'Likeness and Likeliness: Exploring Multidimensional Classification for the Multiverse of Information', *Advances in Classification Research Online* 23:1 (2013): 35–7. https://doi.org/10.7152/acro.v23i1.14235.

available for research. To make these conceptualizations of organization meaningful, we need to examine the scholarly practices of researchers carefully.

3 Conceptualizing a VRE for the Digital Republic of Letters

The main objective of the design for this inclusive knowledge space proposed below is to integrate all relevant digital assets, services, and tools that support the user experience. It is expected to serve as a central hub for humanities scholars in the digital production and usage of relevant knowledge of the republic of letters. We therefore turn to the activities of the researchers that need to shape the design of knowledge spaces.

So far, there have been several attempts to model research practices of individual digital scholars. Some of these models focus on research processes in the analysis[7] and visualization of data.[8] Other models relate humanities data with computing tools.[9] A more comprehensive knowledge management (KM) approach[10] that unifies knowledge stock, process, and enabling technology aspects is used here as a theoretical basis for proposing a VRE for the digital republic of letters.

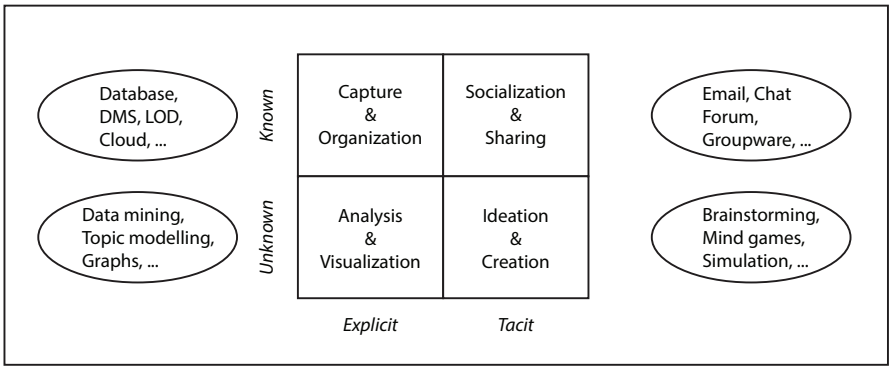


Figure 1: KM model of virtual knowledge space (adapted from Handzic, Knowledge Management)

[7] John Unsworth, 'Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?', *Symposium on Humanities Computing: Formal Methods, Experimental Practice*, London, 13 May 2000; Smiljana Antonijevic, *Amongst Digital Humanists* (New York: Palgrave MacMillan, 2015).

[8] Ben Fry, *Visualizing Data: Exploring and Explaining Data with the Processing Environment* (Sebastopol, CA: O'Reilly Media, Inc., 2007).

[9] Martyn Jessop, 'Computing or Humanities?', *Ubiquity* 5:41 (23–31 December 2004), see https://doi.org/10.1145/1040560.1040561.

[10] Meliha Handzic, *Knowledge Management: Through the Technology Glass* (Singapore: World Scientific Publishing, 2004).

According to a generic KM model of virtual knowledge space presented in figure 1, a holistic VRE design should provide technological support for knowledge exploitation (of what is known) and exploration (of unknown domains). Technology should also support knowledge codification (focused on explicit knowledge contained in digital objects), and personalization (orientated towards people and tacit knowledge held in their heads).

Following the generic KM principles, the proposed VRE design for the republic of letters incorporates the following four basic components: (3.1) Capture and Organization: the ability to capture, organize, and access content in structured and unstructured knowledge repositories (e.g. digital documents, images, metadata); (3.2) Analysis and Visualization: tools for knowledge discovery and presentation from data (e.g. data mining, topic modelling, visualization); (3.3) Socialization and Sharing: mechanisms for communication and knowledge sharing with other researchers (e.g. email, wikis, virtual meeting rooms); and (3.4) Ideation and Creation: support for creativity and new ideas generation (e.g. simulation games, mind mapping, brainstorming). In order to understand and appreciate the proposed VRE model for the republic of letters better, the following discussion has been arranged around the above four themes.

3.1 Capture and Organization: Repositories of Structured and Unstructured Digital Data

An enormous amount of humanities data has been produced in various places over a long period of time. These data need to be represented in digital form before computing techniques can be applied to them. According to Jessop (see note 9), typical digital forms for representing humanities data include the following: digitally born or digitized texts; numerical data, such as those resulting from the textual analyses; digital images representing various objects and materials, such as early manuscripts; digital moving images in films and videos, often used for teaching purposes; spatial data in the sense of geography or sound recordings, of particular interest in language study, etc. In addition to these unstructured digital documents and objects, it is also necessary to assemble more structured data representations in the form of metadata. The primary role of metadata is to identify and classify digital objects, and to make their content visible to computing techniques, thus facilitating their retrieval and analysis.

Within the sphere of the republic of letters, an impressive collection of digital repositories already exists. These are undergoing further development which might jointly be regarded as the 'visible universe' of the current humanistic knowledge in captured and organized form. Among the most notable of these repositories are *ArchiLet*, the *Catalogus Epistularum Neerlandicarum*, the *Corpus Epistolicum Recentioris Aevi*, *Early Modern Letters Online*, *Electronic Enlightenment*, *e-manuscripta*, the *ePistolarium*, and *Kalliope*. In addition, numerous projects initiated by individual scholars and institutions provide the wealth of untapped knowledge on topics as diverse as ge-

ographies, chronologies, prosopographies, networks, and topics in the republic of letters.

These resources provide the evidential foundation upon which much of the future large-scale work in this field will rest. A holistic understanding of this field will require the progressive integration of these resources, and the eventual easing of usage restrictions imposed for a variety of reasons. Some of these limitations can be overcome by the application of the principles of the Linked Open Data;[11] but creating a large, homogeneous, and interoperable pool of open data will not remove all the difficulties confronting this field. On the contrary, it will draw attention to the additional challenges that remain in our understanding of user interaction, application architectures, data fusion, link maintenance, licensing, trust, quality and relevance, and privacy (some of which are discussed in ch. III.5).

3.2 Analysis and Visualization: Tools for Knowledge Discovery and Presentation

Another important piece of VRE architecture is tools for knowledge discovery and presentation. By definition, knowledge discovery involves the non-trivial process of identifying valid, novel, useful, and understandable patterns in data.[12] The uncovered patterns – in the form of clusters, categories, associations, or trends – are often described and presented in a visual mode understandable by humans. For discovering and presenting such patterns, a plethora of data mining and visualization tools are available.

Alan Liu, in his online resource *DH Toychest: Digital Humanities Tools*[13] maintains a comprehensive library of tools that are currently prevalent, canonical, or hot in the digital humanities community, and other tools with high power or general application. Among the most popular ones mentioned are: *Gephi* for network analysis; *Python* and *R* as programming languages that facilitate data analysis; *AntConc*, *TaPOR*, *TXM*, and *Voyant* as text analysis tools; Overview for clustering by topics; and *MALLET* for topic modelling. With respect to knowledge presentation, *Gephi* and *R* are mentioned again as the most popular general multi-purpose visualization tools.[14]

Other tools which have proved useful within the RRL Action also indicate the kind of toolkit needed for the VRE. For example, *NodeGoat* facilitates the visualiza-

[11] Christian Bizer, Tom Heath, and Tim Berners-Lee, 'Linked Data – The Story So Far', *International Journal on Semantic Web and Information Systems (IJSWIS)* 5:3 (2009): 1–22, see https://doi.org/10.4018/jswis.2009081901.

[12] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, 'Knowledge Discovery and Data Mining: Towards a Unifying Framework', in Evangelos Simoudis, Jiawei Han, and Usama Fayyad, eds., *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 2–4 August 1996, Portland, Oregon, USA* (Menlo Park, CA: AAAI Press, 1996) 82–8.

[13] http://dhresourcesforprojectbuilding.pbworks.com/, accessed 20/03/ 2019.

[14] More information on these resources can be found at http://dhresourcesforprojectbuilding.pbworks.com/w/page/69244243/FrontPage.

tion of multilayered networks and will be augmented with analytical algorithms similar to *Gephi* to describe and measure various topological features of networks (such as centrality and betweenness). *Palladio* has been extensively tested and found to be beneficial in temporal, spatial, and relational data analysis and presentation. *ReaderBench* is another promising tool explored in the context of topic modelling based on the polyphonic model.[15] Place should also be reserved for novel visualization tools specifically developed for the republic of letters.

3.3 Socialization and Sharing: Mechanisms for Communication and Knowledge Sharing

In *Virtual Research Environments: From Portals to Science Gateways* (Oxford: Chandos Publishing, 2009), Robert Allan describes VREs as web-based portals to various services designed for use by scientific researchers. According to Allan, besides access to digital repositories and computation services, such portals should include tools for communication and collaboration tools, designed for use by researchers.

The RRL Action connected large number of scholars, libraries, engineers, and designers from various institutions and countries who now act as a community of practice (CoP). For the lifetime of the Action, this community was organized in topical Working Groups and supported by the COST networking tools including meetings, workshops, conferences, training schools, short-term scientific meetings, conference grants, and a website. Most of these involve face-to-face communication, supplemented by email, Skype, and experimentation with *ActiveCollab*.

To stay connected and continue networking and collaborating, this community needs to transform itself into a virtual CoP supported by the members' preferred electronic communication channels and groupware applications. These may be in the form of email lists, wikis, bulletin boards, chatrooms, whiteboards, audio and video conferencing, and more. While these technologies lack the emotional richness and depth of direct, personal interaction, they are often more practicable and considered no less effective in many situations. Experiences and recommendations from research networks such as *ResearchGate* and *Academia.edu* may be useful for selecting the most suitable means for maintaining active knowledge sharing among CoP members in VRE.

3.4 Ideation and Creation: Tools for Stimulating Creativity and Generating Ideas

Following these general principles, the RRL Action organized two successful design sprints that brought scholars and designers with different research needs and

[15] Stefan Trausan-Matu, 'A Polyphonic Model, Analysis Method and Computer Support Tools for the Analysis of Socially-Built Discourse', *Romanian Journal of Information Science and Technology* 16:2–3 (2013): 144–54.

different skill sets to a beautiful historic setting in Como, where they applied the five-part method developed by the Politechnico di Milano (understand, sketch, hypothesize, prototype, present) to the goal of producing novel means of visually analyzing and presenting data on the republic of letters (see ch. IV.1).

The VRE should provide similar support digitally. The VRE developed at the Delft University of Technology[16] shows that technology can support researchers during all stages of the research cycle, from idea creation, through grant pursuit and experimentation to dissemination of research outputs. In this case, idea creation services assist in finding important prior research and people with relevant expertise, while funding services help in searching for and managing research grants; experimentation services support both the virtual research environment and research data management; and publishing services help increase the visibility and impact of a researcher.

Another group of technologies worth considering for inclusion in a VRE includes mind games that foster creativity and innovative problem solving based on the principles of associations, memory retrieval, and the use of analogy and metaphor.

4 VRE Usage Example

A large-scale empirical study of users' behaviour failed to develop a portrait of the virtual researcher in the current VRE.[17] This is not surprising given that different individuals exhibit different styles of enquiry and may not follow the linear workflow models suggested by the humanities literature.[18] On the contrary, humanistic research can be grouped into a number of different styles. According to Handzic and Lin,[19] Type 1 enquirers rely heavily on the study of documents; Type 2 like to share their observations and create consensus; Type 3 seek knowledge by scanning and combining ideas from a wide variety of resources and unusual associations; Type 4 tend to construct and debate different viewpoints and generate new solutions; while Type 5 'are most flexible and comfortable with all systems of enquiry'. Therefore, it is argued here that a VRE must be flexible and interactive enough to accommodate different research needs and the styles of individual researchers.

A fresh approach to this challenge was pursued in 2017 within a COST-funded design sprint in Como in 2017 which brought together experts from the domains of history, knowledge management, computer science, and data interaction design

[16] See http://researchsupport.tudelft.nl/no_cache/, accessed 20/03/2019.

[17] Lynn S. Connaway and Timothy J. Dickey. *Towards a Profile of the Researcher of Today: What Can We Learn from JISC Projects?* (Higher Education Funding Council for England, UK, 2009).

[18] Unsworth, 'Scholarly Primitives'; Antonijevic, *Amongst Digital Humanists*.

[19] Meliha Handzic and Joanne C. Y. Lin, 'K-space and Learning', *Proceedings of the Australasian Conference on Information Systems ACIS*, Perth, 28–29 November 2003.

from the DensityDesign Research Lab at the Politechnico di Milano.[20] Instead of seeking to design visual interfaces for a small set of metadata, this group chose to visualize an interactive interface to a VRE for this field designed to facilitate the exploration of primary texts, images, visualizations, and the connections between them. The thought experiment they conducted covered two of the four quadrants of the model of a virtual knowledge space for the humanities proposed here: namely, 3.1 Capture and Organization and 3.4 Ideation and Creation.[21]

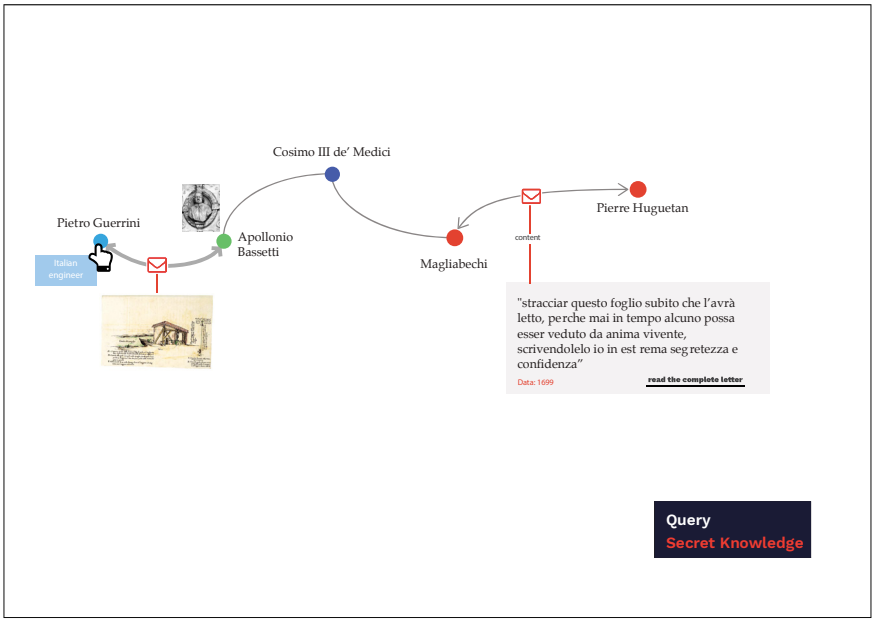
As their point of departure, this group considered the metaphor of the universe of knowledge discussed above. For this purpose, however, this metaphor needed further development, since it implied that the republic of letters was a single, continuous information space in which the parameters (in our case reciprocity in correspondence) with which to explain its specific features would be based on one common law. Instead, in this thought experiment, a multiplicity of knowledge spaces (multiverse of knowledge) was postulated, each with laws of their own. Inspired by Vannavar Bush's search by association, the group postulated the 'gravitational forces' in these knowledge spaces as words that would be perceived as belonging to each other with greater or lesser likelihood in order to explain specific concepts and topics. To illustrate this multiverse of the republic of letters and its physical laws, the potential gravitational pull between two cases of secrecy around technological inventions in correspondences and administrative documents was simulated.

Figure 2 presents a mock-up logbook created on the basis of a study of secrecy, confidentiality, and espionage in letter correspondence[22]. In this case, an Italian engineer, Pietro Guerrini, who was sent by Cosimo III de' Medici to Germany, the Netherlands, and England, figures differently in the context of espionage and in a religious context. Images of Guerrini's inventions stand close to other secret inventions such as Cornelis Drebbel's torpedo-submarine tested on the River Thames in London. In this case, there is also a difference between the intellectual and technical knowledge networks. These networks do not overlap fully, and the non-overlapping parts have a different nature of confidentiality. Typically, the letter exchanges among aristocrats are reciprocal, but not between aristocrats and artisans.

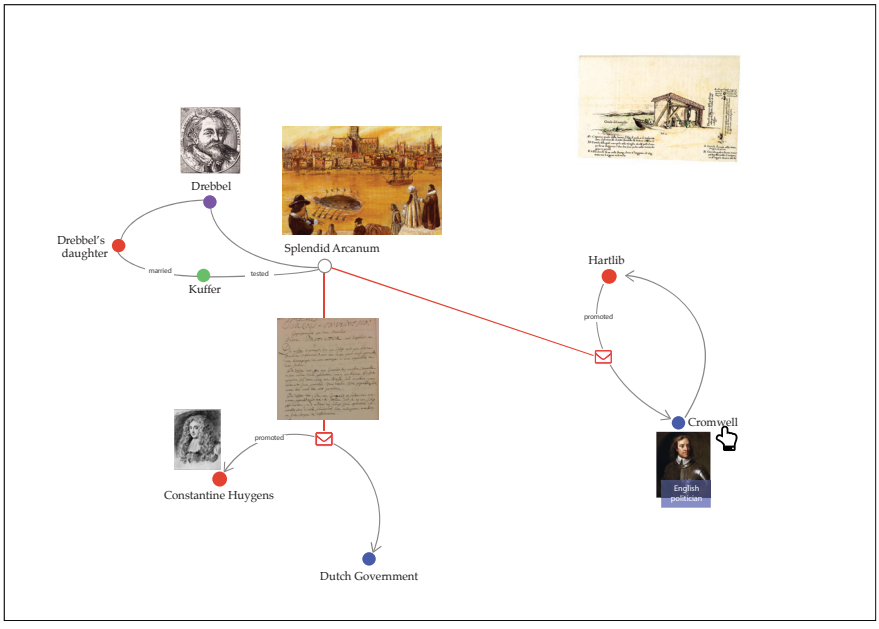
[20] This group consisted of Angeles Briones, Celine Fohn, Meliha Handzic, Charles van den Heuvel, Michele Invernizzi, and Stefan Trausan-Matu.

[21] For a more elaborate discussion of the model in relation to the case, see Meliha Handzic and Charles van den Heuvel, '"Humanists"' Virtual Knowledge Space: Model and Usage', in Meliha Handzic and Daniela Carlucci, eds., *Knowledge Management Arts and Humanities* (Berlin: Springer Verlag, 2019, in press).

[22] Charles van den Heuvel et al., 'Circles of Confidence in Correspondence. Modeling Confidentiality and Secrecy in Knowledge Exchange Networks of Letters and Drawings in the Early Modern Europe', *Nuncius* 31 (2016) 78–106, see https://doi.org/10.1163/18253911-03101002 and van den Heuvel et al, 'Deep Networks as Associative Interfaces to Historical Research'.



(a) Guerrini's inventions



(b) Drebbel's inventions

Figure 2: Mock-up logbook of a researcher's journey through VRE

Rather than focusing on information retrieval in one integrated knowledge universe, this thought experiment explored other ways of interacting with the multiple levels of abstraction of concepts of confidentiality and secrecy by a journey through not one but multiple knowledge spaces. This journey involved three steps: interaction with the search interface, retrieval of data (structured, unstructured, metadata), and application of the associated analysis. The imagined VRE was supplied with a telescope, a spaceship, and a ship's log, allowing it to move back and forth in all the dimensions of the knowledge space while recording its movements. At any point of the journey, the researcher could move back and forth and keep a log of previously visited elements and searches. From both cases it became clear that one person can be part of several networks at the same time. Drebbel was a common citizen from Alkmaar, but had stimulated the interest of James I in London and Rudolph II in Prague, who according to Huygens watched experiments with the *perpetuum mobile* which were also discussed by scholars in the republic of letters. Guerrini tried to filch secret information, but at the same time conversed with scholars and merchants. In short, Drebbel and Guerrini both existed simultaneously in multiple 'knowledge universes', each of which had different 'gravitational forces' in the sense that the exchanges between the people about their knowledge production were different (for instance, more or less confidential or secret) in different contexts. Once the user after this associative visual exploration arrives at a certain point of interest in the knowledge space of the republic of letters, the retrieved information can be further analysed and visualized in the VRE.

5 Discussion and Conclusions

This paper contributes a novel conceptual design of a VRE for the republic of letters. The proposed model pushes forward the current state-of-the art in digital humanities by integrating relevant digital assets, services, and tools that support the research process. It is envisaged as an integrated portal for humanities scholars in the digital production and usage of relevant humanistic knowledge.

Following KM principles, the proposed VRE design incorporates the ability to build and access repositories of structured and unstructured knowledge, tools for knowledge discovery and presentation, support for knowledge sharing and collaboration with other researchers, as well as for generation of ideas and the creation of new knowledge.

The current conceptual work suggests that KM technology has the potential to change the way humanities scholars interact with their data and share their insights. In particular, the paper suggests that the proposed conceptual VRE model may serve as reference for implementing various digital humanities projects relevant to a wide range of humanistic disciplines. However, these implications need to be interpreted with caution due to the current lack of empirical research on adoption and usage of VREs.

With respect to future work, there are a number of major challenges for the designers and adopters of VREs. They include different needs across disciplines (e.g. text-orientated features for literary disciplines versus image-orientated tools for the visual arts), difficulty in systems use (e.g. the need for specialized computer skills), privacy concerns (e.g. how trust can be established among the users, applications, and devices), and the need for advocacy (i.e. explaining the value of VRE and prompting its use among scholars). These findings can serve as an incentive for improving VREs as well as for future research into their adoption and use.

Overall, from the experience gained so far, it can be concluded that the right VRE needs to be dynamic (enabling the addition of new data on the republic of letters), trustworthy (providing some indication of data quality), interactive (via an exploratory interface), flexible (supporting different research needs, practices, and styles), and easy to use (including by scholars who are not technologically proficient).

V Epilogue

Synopsis and Prospects

Howard Hotson

The twenty-two chapters of this book are intended to sketch the outline of 'a digital framework for multilateral collaboration on Europe's intellectual history'. Given the scale of this sketch, a concluding synopsis is needed to provide an introduction to the project as a whole and orientation for closer study. This synopsis is divided into three parts, which describe use-cases operating at different levels of generality. Together, they seek to explain how the distributed infrastructure proposed in this volume can facilitate (1–2) the reassembling of scattered correspondence in particular; (3) the study of the early modern republic of letters more generally; and ultimately (4) collaborative work on cultural and intellectual exchange in all periods and regions. The concluding section (5) considers how this scholarly infrastructure could affect developments outside the domain of scholarship itself, including the vexed question of European identity.

1 Networked Infrastructure for Correspondence Networks

This volume proposes a distributed structure of nodes and hubs as the basis for the next generation of digital infrastructure for this field. Ultimately, this infrastructure must be distributed because that is the way in which intellectual and cultural exchange is structured. In theory, letters could be sent from any point in Europe to any other point; but in practice, the points of origin and destination of early modern learned correspondence were clustered predominantly in a smaller number of major intellectual nodes, and letters travelling any distance normally passed

through major hubs en route to their destinations. The current geographical distribution of the early modern manuscript letters is broadly similar in structure: although some letters have been archived in virtually the same locations in which the correspondents assembled them, most surviving letters have been clustered together in a number of archives and libraries far smaller than the number of places from which they were sent and received.

These scattered archives and a wide range of related digital resources are typically described as 'silos' rather than 'nodes' because there is no efficient communications network connecting them with one another. Such communication is needed for many reasons. The most basic of these is that every single letter pertains, not to one, but to at least two different correspondences: that of the sender and that of the recipient. Since one correspondent may have exchanged letters with hundreds of others, this means that letters to and from any one person may have been scattered through hundreds of different places and may pertain to hundreds of different correspondences. A radically improved network of communication is needed for these scattered materials to be efficiently reunited digitally.

The most obvious solution to this problem, however, is not the best. The collection of data from all these scattered silos into a single, centralized, monolithic data store is politically as well as practically unworkable. Politically, many heritage institutions are reluctant to relinquish control over the catalogues of the cultural patrimony entrusted to them. Practically, no single institution has the expertise, the incentives, or the funding necessary to reconcile the records of all the learned letters scattered across and beyond Europe.

What is needed instead is a system which mirrors the structure of the republic of letters itself: that is, an efficient system of communication linking the large number of archival and other nodes in which large quantities of data reside with a small number of data-collecting and -curating hubs. More specifically, this communications system must involve two basic components: a standard means of describing letters, on the one hand, and a standard means of exchanging these standardized descriptions, on the other. At the moment, we have neither,[1] and the core task of this volume is to begin to describe both.

The second section of this book, 'Standards', outlines the various dimensions of a standard description of the letter. It begins by providing a scholarly definition of what a letter is, and identifying the different textual characteristics, physical states, and literary genres into which letters can be divided (ch. II.1). It then carefully considers how best to model the places to and from which letters are sent (ch. II.2), the dates at which they are written and received (ch. II.3), the people who

[1] The most encouraging experiment of this kind currently is the Correspondence Metadata Interchange Format in the TEI-community, fundamental to correspSearch (https://correspsearch.net/), which demonstrates both the utility of standards of this kind and the willingness of an important subsection of the relevant scholarly community to adopt them. To carry this experiment further, this volume proposes to transition the underlying technology to the Semantic Web, which allows far richer event-based modelling of people, places, and activities.

write and receive letters (ch. II.4), and the topics discussed within them (ch. II.5). The development of a generic model for events (ch. II.6) is the final task before proposing a process-based conception of epistolary communication, which integrates much of the work of section II into a basic letter model (ch. II.7).

The third section of the book, entitled 'Systems', describes key features of the tools and infrastructure required for exchanging, reconciling, and ultimately analysing such letter records. The various sources of epistolary metadata are discussed in chapter III.1, and semi-automated means of reconciling and enhancing them are outlined in chapter III.2. In analogous fashion, chapter III.3 describes several innovative tools for transcribing and editing the texts of letters, while chapter III.4 discusses digital means for modelling texts and topics. Once again, the concluding chapter (III.5) integrates much of the foregoing discussion into a synthetic account of the distributed infrastructure of nodes and hubs capable of transforming the existing silos into an instantly accessible network of increasingly rich and homogeneous data.

Finally, the fourth section, 'Scholarship', explores some of the ways in which the efficient exchange and curation of these standardized records could transform the ways in which scholarly results are generated and presented. Here again, the basic dimensions of the data model provide a structure for considering the geographies (ch. IV.2), chronologies (ch. IV.3), prosopographies (ch. IV.4), networking (ch. IV.5), and text-mining (ch. IV.6) of the republic of letters, while the concluding chapter (IV.7) discusses how tools for handling all these data dimensions might be integrated into a digital research environment.

2 Benefits to Contributing Institutions and Users

The key objective of this whole exercise is to devise a system which best meets the needs of both a wide range of potential data providers and a new generation of data users. To begin with the benefits to data providers, such a network would enable cultural heritage institutions to share core epistolary metadata without relinquishing full control of it; it would allow commercial publishers to increase the discoverability of their publications while controlling access to copyrighted texts; it would permit crowdsourcing and other cataloguing projects to release the fruits of their work immediately for reuse throughout the entire network; and, if enhanced with individual user accounts of the kind already piloted in EMLO, it would also allow contributing projects to curate their material in the context of all the data available in the network while withholding their research results from publication until the time is right and appropriate licences are in place.

For scholarly users, the benefits of such a system would be even more palpable. The principal benefit would be the ability to access through the interface of a single hub all the relevant data distributed throughout the entire network of participating nodes. With the participation of the full range of data providers discussed

in chapter III.1, such a system could provide access to well over a million letter records from the early modern period in relatively short order. Needless to say, with such huge quantities of data flooding in from so many different sources, the task of standardizing, reconciling, and disambiguating it would be immense. Automatically distinguishing the level of metadata standardization and curation of each incoming data set would help prioritize data cleansing and equip users for deciding which collections could be reliably used for which purpose. Thereafter, successive generations of tools would be developed – deriving from those already described in chapter III.2 – to accelerate the processes of data cleaning, normalization, and disambiguation, reconciliation with existing letter records, and enhancement with reference to authority files both old and new.

At an even higher level of granularity, the data contained on places, dates, and people in each letter record would be standardized and reconciled with reference to a new generation of data management tools. One set (called *EM Places* in chapter II.2) would allow the scholarly community to collaborate in assembling the historical gazetteer of places and hierarchies needed to analyse historical data effectively. Another toolset (called *EM Dates* in chapter II.3) would assist in assigning calendars to individual letters and synchronizing the several different calendars used simultaneously within the early modern period. Most useful of all would be a prosopographical tool (discussed in chapter II.4) designed to assemble and coordinate all the biographical details recorded on the system into prosopographical event streams, which can serve both as the basis for disambiguation and as a research resource in its own right. Data curated with the aid of these tools will be progressively enhanced through a series of quality stages far beyond its raw, incoming form, with all automated enhancements clearly identified and all scholarly emendations properly attributed.

The benefits of this scholarly work would then flow back from the hubs to the nodes which contributed the data. Emendations to contributed records would be treated as annotations using the W3C Open Annotation model, which allows the attribution of scholarly work and the monitoring of workflows and quality control. The enhanced records would then be automatically returned to the originating nodes as overlay metadata. The originating data contributors could then opt to handle these emendations in one of three basic ways: they could disregard the enrichments entirely and display only the original records unchanged; they could disregard the original records and display only the enriched versions in their primary institutional catalogues; or they could display the original, unchanged records alongside those enriched by the scholarly community. The attractiveness of this third option is already apparent from the pilot undertaken on EMLO, where users can see digital facsimiles of the index cards from the original catalogue of literary correspondence in the Bodleian Library alongside far more informative, standardized records which are undergoing a continuous process of incremental improvement. Implemented on a larger scale, this networked approach to data curation would effectively put the combined expertise of the global scholarly community at

the service of an entire network of repositories wanting to increase the discoverability and enhance the quality of their internal catalogues.

The vast size of the data pool generated in this way will stimulate the creation of new tools for searching, analysing, and visualizing the data as well as cleansing and reconciling it. As well as traditional keyword and full-text searches and more recently developed faceted search methods, far more sophisticated inferential and semantic queries will be developed for interrogating this growing knowledge graph (ch. III.2, sect. 5). Much of the investment currently wasted in building and maintaining simple, stand-alone data silos can likewise be devoted to creating far more sophisticated tools for analysing and visualizing the growing pool of homogeneous metadata (ch. IV.1). At the micro level, tools can be developed to track the itineraries of individual scholars (ch. IV.2, sect. 3) and to capture more of the chronological dimension of individual correspondences as they wax, wane, and change shape over time (ch. IV.3, sect. 1.3). At the intermediate scale, the potential of this framework is even more transformational. The ongoing application of quantitative network analysis to correspondence networks (ch. IV.5) will be enhanced in many ways, to deal with the evolution of networks over time and the analysis of networks documented by rich prosopographical as well as purely epistolary data (ch. IV.3 sect. 1.1 and 1.3). The patient development of rigorous data models for all the main learned institutions which participated in the republic of letters (ch. IV.4) can provide a new framework in which correspondence networks are superimposed on top of more robust systems of academic exchange evolving over centuries and responding to major political and confessional crises (ch. IV.3, sect. 3). Entire postal systems will also be modelled, along with the shifting landscapes of postal connectivity created by them (ch. IV.2, sect. 4). On an even larger scale, the growing and increasingly representative data set will provide an entirely fresh basis for making data-driven assertions about the origins of the republic of letters, its geographical spread, its chronological development, and the different geographical, confessional, and disciplinary domains within it (ch. I.2).

The opportunities for easy exploration of epistolary metadata will be further enhanced when even larger quantities of machine-readable text are added to the framework. Vast numbers of images of printed and manuscript sources have already been published on open access in resources like *e-manuscripta* (for manuscript letters) and *archive.org* (for printed letter collections: see further ch. III.1, sect. 2). Experiments in crowdsourcing the extraction of metadata from printed letter collections are already underway (ch. III.1, sect. 2) and could eventually be extended to manuscript letters as well. Optical Character Recognition (OCR) already allows the automated transcription of printed letters, creating vast quantities of machine-readable text, much of which has already been enhanced by manual correction (for instance in the case of CERA). Meanwhile, the enhancement of OCR with machine learning is being pioneered by *Transkribus* and applied to manuscript as well as print (ch. III.4, sect. 3). Transcription tools already in widespread use can accelerate the genesis of high-quality, machine-readable texts as well, especially in the

crowdsourcing of transcriptions of vernacular documents (ch. III.3, sect. 2). Digital editing platforms then facilitate the creation of collaboratively produced and fully annotated editions (ch. III.4, sect. 4).

By a combination of these means, the creation of large quantities of high-quality, machine-readable text will prepare the groundwork for the further development of topic modelling of large, polyglot corpora of learned letters. Although much work will be required to perfect these techniques, promising results have already been documented in this volume, including those obtained with the *ePistolarium* (ch. III.4), the *Edinburgh Geoparser* (ch. IV.2, sect. 2), stylometrics (ch. IV.6, sect. 2), text reuse software (ch. IV.6, sect. 3), and other Natural Language Processing techniques (ch. IV.6, sect. 4) and pipelines such as *ReaderBench* (ch. IV.6, sect. 5). The organization and exploration of the resulting topical data can then be further enhanced with reference to multiple, parallel, evolving topical gazetteers derived from early modern sources themselves (ch. II.5). Ultimately, a rich suite of tools for the curation and analysis of both structured and unstructured data will need to be brought together in an integrated digital research environment (ch. IV.7). At the same time, the sustainability of the products of this work will also be guaranteed by the system-agnostic character of the underlying data.

3 From Correspondence Networks to the Republic of Letters

The utility of such a system, however, need not be confined to the relatively limited task of reassembling and analysing letter records and texts. As emphasized at the outset of the volume and at intervals throughout it, there is more to the republic of letters than letters. The sending and receipt of letters were just two of the many different kinds of learned exchange which knit the commonwealth of learning together. In order to move from a paper-thin approach to a more three-dimensional conception of the republic of letters, we need to treat these other activities in a manner analogous to the activities of sending or receiving a letter: that is, as carefully modelled events undertaken by specific individuals pinpointable in time and space. With such models in place, the distributed infrastructure created for reassembling correspondences will be equally effective in assembling documentation on all of these other forms of intellectual exchange as well.

Three broad categories of documentation require particular attention before a more rounded and data-driven conception of the republic of letters can be achieved. Correspondence is just one of the *media of communication* which translated ideas from one place to another, so the large quantities of documentation of other media of this kind need to be approached in analogous fashion. Like letters, printed books systematically record their own places and dates of publication and the names of the people – authors, printer, publishers, and patrons – associated with them, as well as the subjects treated within them. Data models for describing books are already highly advanced and widely adopted within the library science

community. Huge quantities of structured data conforming to those models already exist in digital form in library catalogues and national bibliographies. Further data of this kind can be extracted from printed and manuscript sources, including printers' lists, book fair catalogues, booksellers' stock-lists, probate inventories, and auction catalogues. Bibliographical data of this kind can be used to document events in the lives not only of authors, printers, and publishers of those books, but also of the people who discussed them in their letters and books, taught them in the lecture theatre, collected them for private and public libraries, and indeed documented their responses to them in marginalia and commonplace books. Unlike books, the learned journals which proliferated after 1665 provide a relatively untapped source of data on the shifting configuration of international communities of contributors in time and place. The reason is that the technology has hitherto been lacking for assembling the huge treasure trove of highly granular data which they contain. When combined with the archives assembled by the learned societies which edited some of these journals, data on the publications themselves can be further enhanced with documentation of the entire process of submission, review, and distribution.

Institutional records also document international exchange over long periods and wide areas. Academic institutions were created to facilitate intellectual exchange on the international as well as national and local levels. University matriculation registers record formative events in the lives of innumerable intellectuals. Aggregating these records can reveal shifts in the patterns of national and international student migration extending over centuries and whole continents (ch. IV.3, sect. 3). Studies of academic travel to and from specific cities and countries have already amassed large quantities of high-quality data ready for aggregation (ch. IV.4, sect. 2.2). Professorial biographies, available in large quantities, can be used to map international exchange in structured ways. Ecclesiastical institutions kept detailed records of often highly international clerical and monastic orders, synods, and councils (ch. IV.4, sect. 3). During the early modern period, the spread of Christian missions created the first learned networks in history that were truly global in scope. Other learned organizations with rich records can be treated in analogous fashion, including formal literary and artistic academies, scientific societies, and professional associations as well as more informal sodalities and salons. As in the case of media, the potential scope of this structured approach to analysing learned institutions is immense.

Less formal documentation of travel and exchange can be gleaned from many supplementary sources as well. One particularly rich category is 'ego documents', which include autobiographies, funeral sermons, travel diaries, and *alba amicorum* (the 'books of friends' or *Stammbücher* collected during travels in the early modern period, which still survive in large numbers today). Involuntary displacement, forced migration, and exile produced less happy but equally informative records on an episodic basis.

The feasibility of this approach depends on two key points which need to be made explicit. The generic point is that all of these activities – matriculating at a university, signing an *album amicorum,* purchasing a book, submitting a journal article, participating in a learned gathering, or writing a letter – are events undertaken by specific people at specific times and places. This means that the tools built for the purpose of disambiguating and reconciling the people, places, and dates involved in exchanging correspondence can also be repurposed for all these other networking activities as well. The more specific point is that the self-same people and places are often involved as well: the citizens of the republic of letters who exchanged letters also participated in these other forms of learned exchange and did so in the same places. This means that the authority files established and populated for people and places exchanging early modern learned letters will already include many of the people and places involved in other forms of learned exchange as well. In other words, once this set of tools has been developed to handle scholarly correspondence, it can be adapted to handle data on any other form of intellectual exchange by adding only one additional component: namely, a data model for each new form of intellectual activity, based on the generic event model proposed in chapter II.6, and analogous to the letter model proposed in chapter II.7. Moreover, data models for many of these other 'basic event types' will be less complicated than those for letters: an individual record in a university matriculation register, for instance, is far simpler in structure than the process-based letter model proposed in chapter II.7.

Once these resources and their accompanying reconciliation tools are available, huge quantities of historical data can be efficiently processed as events involving an expanding company of people, places, and dates. These events can then be interlinked into event streams, which in turn can form the basis of broader narratives. Such event streams can be usefully distinguished into several different kinds. *Single data series* – such as the place of origin of books in a given early modern library, or the origin and destination of an individual's correspondence – will be assembled, analysed, visualized, animated, and narrated with unprecedented facility. *Multiple homogeneous data series* will be combined to study similar patterns on a larger scale: for instance, visualizing multiple matriculation registers could show the expansion of a new university's catchment area at the expense of neighbouring institutions. *Multiple homologous data series* could also be combined or compared: for instance, databases on different religious orders or learned societies could be juxtaposed for comparison with one another and with secular clergy from different confessions or members from different learned disciplines. *Multiple heterogeneous data series* could be mined for data relating to a specific entity: a simple example would be the automatic collection of a prosopographical event stream consisting of all the events on the system relevant to a single person (for instance, all incoming and outgoing letters, all records in diaries of learned acquaintances, all records of participation in learned societies, and so on); a more complex example might assemble data on a particular place during a specified period (for instance, all students, professors and visitors in

Leiden in 1637). *Person and place authority files* could likewise be populated with data from multiple series (such as the Latinized names from matriculation registers, learned correspondence, and travel diaries).

Just as important as narrating developments involving many people will be the capacity of this system to assemble precise prosopographical data on individuals. A robust prosopographical data model will consist principally of a stream of events of this kind which are carefully modelled and precisely documented. A learned life is composed largely of a stream of such events; a learned network is composed largely of the intersection of many such lives; and the *respublica litteraria* is composed largely of the sum total of all these networks. To these private events need to be added events in the public life of the republic of letters.[2]

4 Beyond the *Respublica Litteraria* to Global Cultural Exchange

This volume has focused exclusively on the early modern period, primarily on correspondence, and secondarily on the other forms of intellectual exchange that constituted the republic of letters. The relatively small space of this domain is represented by the pale area in the top centre of the figure below. But the approach proposed here is potentially of much wider applicability: it can be expanded chronologically to earlier and later periods, thematically to other forms of exchange, and geographically to regions beyond Europe.

	ANCIENT	MEDIEVAL	EARLY MODERN	MODERN	CONTEMPORARY
LEARNING			correspondence		
			books and journals education and travel		
THE ARTS			literature, drama, music painting, sculpture, architecture technology, popular culture		
COMMUNICATION & COMMERCE			diplomacy news and intelligencing trade and commerce		

Figure 1: Chronological and thematic expansion of systems for capturing data on intellectual, cultural, diplomatic and commercial exchange

[2] Events and periods of consequence to the republic of letters might be natural (such as the appearance of the new star in Cassiopeiae in early November 1572, or the Lisbon earthquake at around 09:40 local time on 1 November 1755), military (such as the fall of Constantinople on 29 May 1453, or the Thirty Years' War in central Europe, from the defenestration of Prague on 23 May 1618 to the proclamation of the Peace of Westphalia on 24 October 1648), political (such as the accession and death of successive monarchs), religious (whether ephemeral, like the posting of Luther's Ninety-Five Theses on 31 October 1517; episodic, like the Council of Trent; successive, like the series of Roman pontiffs; or vague and indistinct, like the advent of Pietism); or related to stages in the history of learned institutions (like the founding of the Royal Society on 28 November 1660).

Chronological expansion. This volume has been limited to the early modern period of European intellectual history for good reason. Devising standards, tools, and infrastructure to facilitate collaboration requires intensive and sustained dialogue with a wide range of interested parties. Since sustaining such a dialogue requires a tight focus, COST Action IS 1310 concentrated chronologically and thematically on one of the most widespread and coherent systems of intellectual and cultural exchange in European history: the so-called 'republic of letters' between 1500 and 1800. But any chronological boundaries are, to some degree, arbitrary and unsustainable, most obviously because any given beginning or end date will fall mid-way through the working lives of a whole generation of correspondents. Fortunately, standards, tools, and systems designed to facilitate collaboration on early modern correspondence can readily be repurposed for work on earlier and later periods. In short, we need infrastructure for assembling and analysing correspondence metadata from the nineteenth and twentieth centuries and from the contemporary digital era, as well as from earlier periods insofar as such data is available. Likewise, the infrastructure created for dealing with early modern people and places can be adjusted for use with earlier and later periods as well.

Thematic expansion. Less obvious but equally transformative is the potential for thematic expansion. Any infrastructure developed to serve the study of intellectual exchange could be repurposed, with somewhat more modification, to serve other forms of exchange. Diplomacy was likewise conducted largely through travel and the scribal exchange of letters and reports which were precisely located in time and space. The trade in all manner of commodities was accompanied by commercial correspondence which likewise survives in enormous quantities. Cultural exchange was conducted largely through the traffic of artefacts and artificers. Not all of these forms of exchange were as systematically and voluminously documented as the best institutionalized activities central to the republic of letters: the fact that letters, books, and many other artefacts documenting intellectual exchange typically include precise information about the agents, dates, and places involved in their creation explains why they provide the best starting point for devising a system to model cultural exchange in detail. But the approach developed to model the one can be adapted for the purpose of modelling the others. In this way, the 'republic of letters' can be regarded as a suitably capacious testbed – three centuries in duration, pan-European in scope, incipiently global in reach, encyclopaedic in thematic breadth, and manageably large in terms of sheer volume of materials for treatment – for standards, tools, systems, policies, and collaborative cultures which could be applied to the study of diplomatic, cultural, and commercial exchange in earlier and later periods as well.

Geographical expansion. Equally obviously, as such resources move from the early modern to the modern period, their European focus must become more global. Even without abandoning a focus on intellectual exchange, this becomes increasingly necessary from the sixteenth century onwards, thanks to Portuguese, Spanish, Dutch, French, and English commerce and trade, Catholic and eventually also

Protestant missionary work, and eventually New World colonies with a unique and active role in key debates in the eighteenth century. During the nineteenth century, moreover, the rapid spread of empires and global trade and the rising importance of colonies and former colonies in intellectual life makes a narrowly European focus inappropriate even for resources concentrating on networks rooted in the West.

5 Social and Political Outcomes

The social and political outcomes of this transformation will be as significant as the scholarly and technical ones. This digital framework has been conceived from the outset as a vehicle for facilitating radically multilateral collaboration, and designed to have a palpable effect on the conduct of scholarship. One of its basic effects will be to break down disciplinary barriers between custodians of libraries and archives, the scholars who use them, and the systems developers and designers co-creating tools and methods necessary to process and analyse unprecedented quantities of data. No less important will be the breakdown of barriers between individual institutions and indeed individual countries: the application of digital technology to humanistic research has now reached a stage in which ambitious, cutting-edge projects can no longer be self-contained, or conceived and developed in isolation. High-level innovation is now too demanding for a single project, institution, or even country to undertake alone: instead, it will belong in the future to networks of projects and institutions pursuing the most ambitious objectives collectively. The resulting collaboratively designed, built, and populated infrastructure will also address the problem of sustainability, since shared systems can more economically be maintained and upgraded than a large number of wholly independent ones.

Sharing data via the Semantic Web and tools for processing it via distributed infrastructure will also serve to level the scholarly playing field at both the international and the more local levels. At the moment, the capacity of some European countries to invest far more than others in creating digital tools and data leaves less prosperous regions mere onlookers in the digital transformation of scholarship. Shared and distributed infrastructure will not only raise the standard of tools and processes in the most advantaged countries: it will simultaneously share those advanced with less favoured regions. At the more local level, this process will also democratize scholarship by providing state-of-the art research facilities in areas distant from leading heritage institutions. At both the national and the continental level, these processes will contribute to the rewriting of intellectual history in a manner which is not only transnational but also decentralized, because it has progressively escaped from the structures imposed by national scholarly infrastructure in the nation-building process of previous centuries.

As well as being distributed more equitably throughout the scholarly community, the fruits of this arrangement will overflow the academic domain and benefit

wider publics as well. When intellectual exchange is regarded as just one variety of cultural and commercial exchange, the potential impact of this approach on the consolidation of a European level of identity becomes most apparent. For millennia, Europe has been knit together by complex patterns of continental traffic and exchange. The ceaseless movement of people, goods, and ideas across borders has created, transplanted, and transformed much of what Europeans most love and cherish. Understanding this shared heritage is fundamental to the construction of broad transnational narratives, which can supplement existing regional, national, and local identities. Nation-states depend for their coherence on shared narratives about a common past. The future solidarity of Europe can likewise be strengthened by transnational stories about its shared history. In this respect as well, collaboration will be necessary to develop means of projecting that history beyond the realm of specialized scholarship into the public domains of universities, schools, the media, heritage institutions, and general culture in a manner capable of contributing to the formation of urgently needed transnational levels of identity.

The data needed to create such narratives is embedded in vast numbers of documents scattered throughout cultural heritage institutions across the Continent. The sheer quantity of these resources overwhelms human readers while most also remain inaccessible to computers. Europe now urgently needs new shared protocols and tools which can help to identify relevant data within these local and national resources, convert them to a machine-readable form, make them openly available, contextualize them with reference to one another, analyse the resulting patterns, knit them together into powerful new narratives, and project those stories into the public sphere. These needs can best be addressed by means of a distributed, open-source system for aggregating transnational cultural heritage data into narratives capable of transforming identities.

The key to achieving these objectives is the reconceptualization of metadata systems to make them compliant with FAIR principles (findable, accessible, interoperable, and reusable).[3] This requires analysing these stories into their elemental units. Compiled over centuries, the historical data required to construct these narratives takes many different forms; but much of it shares three basic features. A merchant's account books, a university's matriculation list, a library catalogue, an inventory of correspondence: these records describe very different kinds of activities and events, but all ascribe those activities to specific *individuals* in specified *times* and *places*. Narratives are composed, in part, of events and their representations. Events, in turn, involve an agent (typically a person or people) undertaking a specific activity at a specific place and time. The fundamental task is therefore rigorously to develop systems and authorities for handling historical people, places, and dates; and models for an expanding catalogue of basic event types. Rigorously modelling and populating an expanding knowledge-graph of interlinked basic events is therefore the precondition for generating new transnational narratives

[3] See https://www.force11.org/group/fairgroup/fairprinciples, accessed 20/03/2019.

from masses of granular data. Deriving rich historical narratives from these event streams will remain an artisanal process, involving collaboration between experts in computer science, data analysis, visual communication, and humanistic scholarship. But the capacity to generate streams of hundreds of thousands of events will supply abundant material on which fresh transnational narratives can be constructed, presented visually, disseminated inexpensively, and preserved as the basis for further research.

It is to be hoped that the infrastructure development – like its planning and use – can be undertaken on a broad and distributed basis. Since the infrastructure is not monolithic, it need not be developed by a single consortium financed with a single, massive, e-infrastructure development grant. Substantial investment will be necessary to create the core system for exchanging data within a distributed network of nodes and hubs; but meanwhile many details can be worked out and many tools developed independently by other project teams. In principle, collaborative systems design should be followed by collaborative systems development in order to produce the conditions for collaborative population of the system. In other words, the continuous engagement of a broad community is necessary at every stage of the process if that community is to embrace the resulting systems as its own.

As this volume goes to press, Europe confronts many daunting challenges. In order to surmount them, the Old World will need to seize the fresh opportunities for enhanced self-understanding offered by the newest of technologies.

Contributors

Dr Ruth Ahnert, Senior Lecturer in Renaissance Studies,
Queen Mary University of London.
r.r.ahnert@qmul.ac.uk

Dr Sebastian E. Ahnert, Gatsby Career Development Fellow,
University of Cambridge.
sea31@cam.ac.uk

Dr Rebekah Ahrendt, Associate Professor of Musicology,
Utrecht University.
r.s.ahrendt@uu.nl

Dr Nadine Akkerman, Reader in Early Modern English Literature,
Leiden University.
n.n.w.akkerman@hum.leidenuniv.nl

Dr Gábor Almási, Researcher, Ludwig Boltzmann-Institute for Neo-Latin Studies,
University of Innsbruck.
almasi.gabor@gmail.com

Dr Per Pippin Aspaas, Senior Academic Librarian,
University of Tromsø – The Arctic University of Norway.
per.pippin.aspaas@uit.no

PD Dr Jan Bloemendal, Senior Researcher,
Huygens Institute for the History of the Netherlands and Privatdozent
of Neo-Latin, Ruhr-University Bochum.
jan.bloemendal@huygens.knaw.nl

Dr Elizabethanne Boran, Librarian of the Edward Worth Library, Dublin.
eaboran@tcd.ie

Arno Bosse, Digital Project Manager, Cultures of Knowledge,
University of Oxford.
arno.bosse@history.ox.ac.uk

Dr Robin Buning, Postdoctoral Researcher,
Huygens Institute for the History of the Netherlands.
robin.buning@huygens.knaw.nl

Dr Alex Butterworth, Research Fellow,
Sussex Humanities Lab, University of Sussex.
alex@alexbutterworth.co.uk

Prof Clizia Carminati, Professor of Italian Literature,
University of Bergamo.
clizia.carminati@unibg.it

Prof Paolo Ciuccarelli, Associate Professor,
DensityDesign Research Lab, Dipartimento di Design, Politecnico di Milano.
paolo.ciuccarelli@polimi.it

Dr Roberta Colbertaldo, Postdoctoral Researcher,
Goethe-University of Frankfurt.
colbertaldo@em.uni-frankfurt.de

Dr Per Cullhed, Strategic Development Manager,
Uppsala University Library.
per.cullhed@ub.uu.se

Jana Dambrogio, Thomas F. Peterson (1957) Conservator,
Massachusetts Institute of Technology (MIT) Libraries.
jld@mit.edu

Prof Mihai Dascalu, Associate Professor of Computer Science,
University Politehnica of Bucharest.
mihai.dascalu@cs.pub.ro

Prof Antonio Dávila Pérez, Senior Lecturer in Latin Philology,
University of Cádiz.
antonio.davila@uca.es

Carlo De Gaetano, Designer and Researcher,
Amsterdam University of Applied Sciences.
c.a.m.de.gaetano@hva.nl

Serena Del Nero, M.Sc Candidate,
DensityDesign Research Lab, Dipartimento di Design, Politecnico di Milano.
srn.delnero@gmail.com

Tommaso Elli, PhD Student in Design,
DensityDesign Research Lab, Dipartimento di Design, Politecnico di Milano.
tommaso.elli@polimi.it

Dr Vittoria Feloa, Assistant Professor of Early Modern History,
University of Padua.
vittoria.feola@unipd.it

Gertjan Filarski, Director of Digital Infrastructure,
Humanities Cluster of the Royal Netherlands Academy of Arts & Sciences.
gertjan.filarski@di.huc.knaw.nl

Dr Andreas Fingernagel, Director of the Department of Manuscripts
and Rare Books at the Vienna National Library.
andreas.fingernagel@onb.ac.at

Prof Ian Gregory, Professor of Digital Humanities,
Lancaster University.
i.gregory@lancaster.ac.uk.

Dr Claire Grover, Senior Research Fellow,
University of Edinburgh, School of Informatics.
c.grover@ed.ac.uk

Prof Meliha Handzic, Professor of Information Systems,
International Burch University, Sarajevo.
meliha.handzic@ibu.edu.ba

Dr Simon Hengchen, Postdoctoral Researcher,
University of Helsinki.
simon.hengchen@helsinki.fi

Prof Howard Hotson, Professor of Early Modern Intellectual History
and Director, Cultures of Knowledge project,
Faculty of History, University of Oxford.
howard.hotson@history.ox.ac.uk

Prof Eero Hyvönen, Professor of Semantic Media Technology,
Aalto University;
Director of Helsinki Centre for Digital Humanities (HELDIG),
University of Helsinki.
eero.hyvonen@aalto.fi

Neil Jefferies, Head of Innovation, Bodleian Digital Libraries,
University of Oxford.
neil.jefferies@bodleian.ox.ac.uk

Dr Mikkel Munthe Jensen, Junior Fellow,
Max Weber Centre for Advanced Cultural and Social Studies /
Gotha Research Centre, University of Erfurt.
Mikkel.Jensen@uni-erfurt.de

Christoph Kudella, DARIAH-DE & Scientific Coordinator 'Digital Editions',
Göttingen State and University Library.
kudella@sub.uni-goettingen.de

Dr Ad Leerintveld, Keeper of Modern Manuscripts, responsible for
Catalogus Epistularum Neerlandicarum,
National Library of the Netherlands.
ad.leerintveld@kb.nl

Miranda Lewis, Editor, Early Modern Letters Online,
Cultures of Knowledge Project,
University of Oxford.
miranda.lewis@history.ox.ac.uk

Prof Eetu Mäkelä, Professor in Human Sciences–Computing Interaction,
University of Helsinki;
Docent (Adjunct Professor) in Computer Science,
Aalto University; Helsinki Centre for Digital Humanities (HELDIG).
eetu.makela@helsinki.fi

Dr Glauco Mantegari, Independent Researcher and Consultant
in Data Science and Visualization.
mantegla@gmail.com

Dr Ikaros Mantouvalos, Faculty Member,
Democritus University of Thrace, Department of Education Sciences
in Early Childhood.
imantouv@psed.duth.gr

Dr Ludovica Marinucci, Postdoctoral Researcher,
Semantic Technology Laboratory (STLab), Istituto di Scienze e Tecnologie della Cognizione - Consiglio Nazionale delle Ricerche (ISTC-CNR), Italy.
ludovica.marinucci@istc.cnr.it

Dr Marie Isabel Matthews-Schlinzig, Dunfermline.
whatisaletter@gmail.com

Michele Mauri, Research Fellow,
DensityDesign Research Lab, Dipartimento di Design, Politecnico di Milano.
michele.mauri@polimi.it

Dr Barbara McGillivray, Turing Research Fellow,
University of Cambridge and The Alan Turing Institute.
bmcgillivray@turing.ac.uk

Gabriela Martínez, Predoctoral Fellow,
Universidad Nacional de Educación a Distancia.
gabrielamartinez@flog.uned.es

Prof Bruno Martins, Assistant Professor, Data Management
and Information Retrieval,
Instituto Superior Técnico, University of Lisbon.
bruno.g.martins@ist.utl.pt

Giovanni Moretti, Senior Software Developer, Digital Humanities Group,
Fondazione Bruno Kessler / Trento.
moretti@fbk.eu

Dr Yves Moreau, associated researcher,
Laboratoire de Recherche Historique Rhône Alpes, University Lyon III.
yvesmoreau99@msn.com

Dr Dagmar Mrozik.
kontakt@dagmar-mrozik.de

Dr Günter Mühlberger, Director of the Research Center 'Digital Humanities',
University of Innsbruck; Co-ordinator, READ Project.
guenter.muehlberger@uibk.ac.at

Gerhard Müller, Director of Kalliope Verbund,
State Library Berlin.
gerhard.mueller@sbb.spk-berlin.de

Dr Patricia Murrieta-Flores, Lecturer in Digital Humanities
and Co-Director of the Digital Humanities Hub,
Lancaster University.
p.murrieta@lancaster.ac.uk

Dr Chiara Petrolini, Research Project Member 'The Oriental Outpost
of the Republic of Letters',
University of Vienna, Institute for Austrian Historical Research.
chiara.petrolini@univie.ac.at

Dr Azzurra Pini, Research Fellow,
DensityDesign Research Lab, Dipartimento di Design, Politecnico di Milano.
azzurra.pini@polimi.it

Dr Catherine Porter, Research Fellow,
School of Natural and Built Environment,
Queen's University Belfast.
c.porter@qub.ac.uk

Dr Montserrat Prats López, Assistant Professor of Information Systems,
Open Universiteit / Heerlen.
montserrat.pratslopez@ou.nl

Dr Alexa Renggli, Coordinator of e-manuscripta.ch,
Zentralbibliothek Zürich.
alexa.renggli@zb.uzh.ch

Dr Sinai Rusinek, Digital Humanities Program,
Haifa University; OMILab, The Open University / Israel.
sinai.rusinek@mail.huji.ac.il

Patryk Sapala, Senior Librarian,
National Library of Poland.
p.sapala@bn.org.pl

Dr Alexandra Sfoini, Senior Researcher of Modern Greek History,
Institute of Historical Research / National Hellenic Research Foundation.
alexsfin@eie.gr

Dr Anna Skolimowska, Head of the Laboratory for Source Editing
& Digital Humanities,
Faculty of 'Artes Liberales', University of Warsaw.
as@al.uw.edu.pl

Dr Daniel Smith, Lecturer in Early Modern English Literature,
King's College London.
daniel.s.smith@kcl.ac.uk

Dr Elena Spadini, Postdoctoral Researcher in Digital Philology,
University of Lausanne.
elena.spadini@unil.ch

Prof Thomas Stäcker, Director of the State and University Library Darmstadt.
thomas.staecker@ulb.tu-darmstadt.de

Dr Lucie Storchová, Research Fellow,
Institute of Philosophy, Czech Academy of Sciences, Prague.
storchova@flu.cas.cz

Dr Alex J. Tessier,
Université d'Évry-Val d'Essonne/ IDHES-Évry.
alexandre-tessier@laposte.net

Prof Stefan Trausan-Matu, Professor of Computational Linguistics,
Human-Computer Interaction, and Algorithms,
University Politehnica, Bucharest.
stefan.trausan@cs.pub.ro

Dr Jouni Tuominen, Postdoctoral Researcher,
University of Helsinki, and Aalto University.
jouni.tuominen@helsinki.fi

Dr Vladimír Urbánek, Senior Researcher, Head of the Department of Comenius Studies and Early Modern Intellectual History,
Institute of Philosophy, Czech Academy of Sciences, Prague.
urbanek@flu.cas.cz

Prof Charles van den Heuvel,
Huygens Institute for the History of the Netherlands, Head of Department History of Science and Scholarship;
University of Amsterdam, Chair of Digital Methods in Historical Disciplines.
charles.van.den.heuvel@huygens.knaw.nl

Dr David van der Linden, NWO VENI Post-Doctoral Research Fellow,
University of Groningen.
d.c.van.der.linden@rug.nl

Dr Dirk van Miert, Assistant Professor of Early Modern Cultural History,
Department of History and Art History, Utrecht University.
d.k.w.vanmiert@uu.nl

Dr Justine Walden, Postdoctoral Fellow,
Department of History, University of Toronto.
justine.walden@utoronto.ca

PD Dr Thomas Wallnig, PI and Privatdozent (Adjunct Professor) of Modern History,
University of Vienna.
thomas.wallnig@univie.ac.at

Prof Axel E. Walter, Professor,
Faculty of Communication and University Library, University of Vilnius.
axel.walter@mb.vu.lt

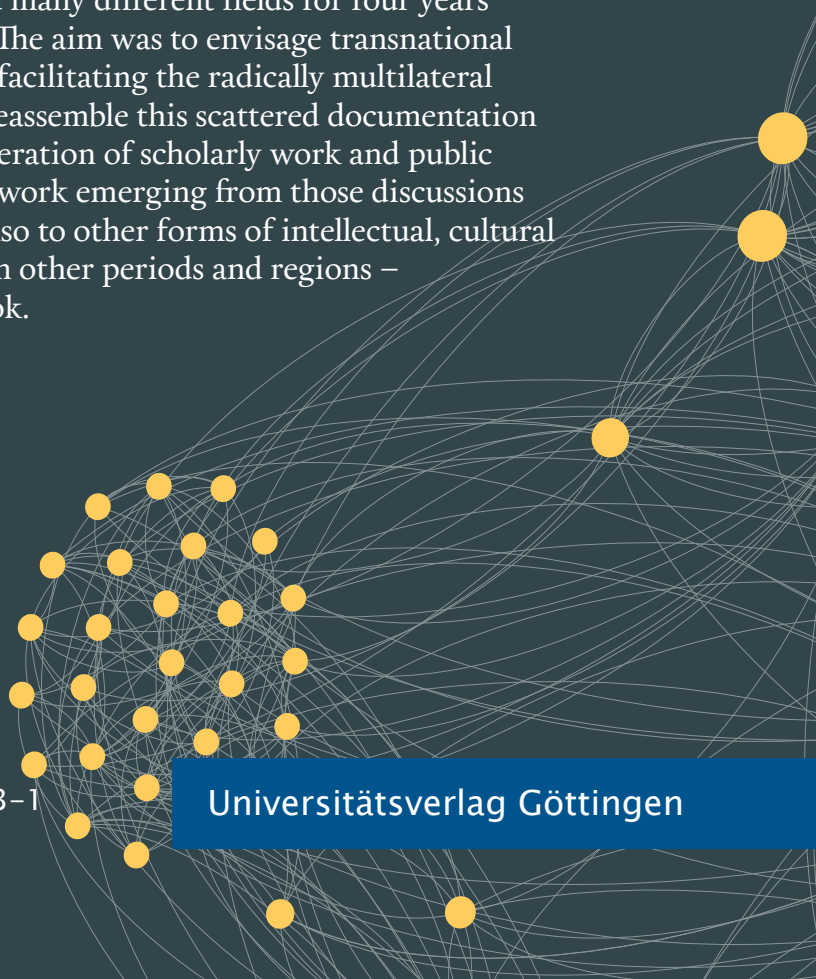
Prof Ruth Whelan, Professor of French,
Maynooth University.
ruth.whelan@mu.ie

Dr Elizabeth R. Williamson, Research Fellow in Digital Humanities,
University of Exeter.
e.r.williamson@exeter.ac.uk

Between 1500 and 1800, the rapid evolution of postal communication allowed ordinary men and women to scatter letters across Europe like never before. This exchange helped knit together what contemporaries called the 'respublica litteraria', a knowledge-based civil society, crucial to that era's intellectual breakthroughs, formative of many modern values and institutions, and a potential cornerstone of a transnational level of European identity.

Ironically, the exchange of letters which created this community also dispersed the documentation required to study it, posing enormous difficulties for historians of the subject ever since. To reassemble that scattered material and chart the history of that imagined community, we need a revolution in digital communications.

Between 2014 and 2018, an EU networking grant assembled an interdisciplinary community of over 200 experts from 33 different countries and many different fields for four years of structured discussion. The aim was to envisage transnational digital infrastructure for facilitating the radically multilateral collaboration needed to reassemble this scattered documentation and to support a new generation of scholarly work and public dissemination. The framework emerging from those discussions – potentially applicable also to other forms of intellectual, cultural and economic exchange in other periods and regions – is documented in this book.



BOOK COVER DESIGN
Tommaso Elli and Beatrice Gobbo
DensityDesign Research Lab, Milan

ISBN: 978-3-86395-403-1

Universitätsverlag Göttingen