MANREET KAUL

HISTORY IN THE DIGITAL AGE

VOL 1

History in the Digital Age: Vol 1

History in the Digital Age: Vol 1

Manreet Kaul

THE INFO LIBRARY

Published by The InfoLibrary,
4/21B, First Floor, E-Block,
Model Town-II,
New Delhi-110009, India

History in the Digital Age: Vol 1
Manreet Kaul
ISBN: 978-93-5590-089-0

Table of Contents

PART I
Introduction

Emergence of the New Digital History

Petri Paju, Mila Oiva and Mats Fridlund

Half a century ago, historian Emmanuel Le Roy Ladurie, when surveying the progress of quantitative history, prophesised that 'tomorrow's historian will have to be able to programme a computer in order to survive'.[1] Since then, computers and programming have indeed profoundly changed historians' practice through such digital tools as word processing, the internet, email, PowerPoint, Google, JSTOR, Facebook, Twitter and Zoom. They have made all of us historians into digital historians in one way or another. As these digital tools used by most historians illustrate, there are many ways that the digital has transformed the historian's craft beyond mere practical and administrative improvements. During the new millennium, the computer together with the internet have begun to change also the historian's research tools and methods in new and previously unforeseen ways into a novel kind of digital history. It is this new emerging digital history, together with some ever-significant approaches of the 'old' quantitative digital history, that is the subject of this book.

Digital history encompasses diverse historical practices, such as digitisation efforts at archives, libraries and museums, computer-assisted research, web-based teaching and professional and public dissemination of historical knowledge, as well as research on the history of 'the digital', computers and

digital technologies. One comprehensive definition capturing this diversity of practices was suggested more than a decade ago in a discussion between digital historians in the *Journal of American History*:

> Digital history is an approach to examining and representing the past that works with the new communication technologies of the computer, the Internet network, and software systems. On one level, digital history is an open arena of scholarly production and communication, encompassing the development of new course materials and scholarly data collections. On another, it is a methodological approach framed by the hypertextual power of these technologies to make, define, query, and annotate associations in the human record of the past. To do digital history, then, is to create a framework, an ontology, through the technology for people to experience, read, and follow an argument about a historical problem.[2]

While the digital embraces the whole spectrum of the historian's craft, this volume focuses on digital history as a form of scholarly research that uses digital sources and tools to produce new historical knowledge. This form of digital history research is part of the larger digital turn in academia, identified as digital humanities, culture analytics, computational social sciences and other concepts related to utilisation of computer-assisted methods for research.[3] By bringing together research contributions to the new digital history from historians, computer scientists, computational linguists and other scholars producing new empirical historical knowledge using digital methodologies, as well as conceptually focused perspectives on critical issues of the field's past, present and future development, this book provides digital histories that we hope will be read as laudable exemplars from within the emergent digital history research community. The digital histories collected here simultaneously represent various methodological applications of and themes within digital history research and thus an attempt to take stock of current research rather than providing a pedagogical textbook or programmatic manifesto. The new digital history has matured enough for us to instead be able to present historical work currently furthering historical research. Thus, the studies in this book take digital history beyond discussions of its future potential, proofs of concept and pedagogical examples to instead focus on digital history 'in action', to the making of new historical knowledge.

Through this focus on presenting results from digital history research projects, this book breaks new ground within the current wave of digital history. Other digital history books published so far have mainly been monographs focused on discussing how historians could use digital sources or methods to conduct and present research such as the pioneering *Digital history: a guide to gathering, preserving, and presenting the past on the web* (2006) by Daniel J. Cohen and Roy Rosenzweig, or anthologies such as *History in the digital*

age (2013) contributing discussions of the problems and possibilities of doing the new digital history, rather than research results of historical studies using new digital methodologies.[4] This is as expected, as it is only during the last couple of years that we have seen the first research publications using digital history research methodologies within mainstream academic historical publishing outlets. Matthew Jockers' *Macroanalysis* (2013) appears to be the first research monograph published by a university press, and Cameron Blevins' 'Space, nation, and the triumph of region' (2014) is the first peer-reviewed research article published in the *Journal of American History*.[5] In this way, this book aspires to pioneer and promote work within the new digital history by being a timely research anthology from the current third generation of digital historians that,[6] outside of digital spatial history,[7] focuses on contributing new historical knowledge from research using digital research methodologies.

Emergence of the New Digital History

The roots of exploiting modern data-processing equipment in humanities research date back at least to the 1950s, when Josephine Miles started using punch cards for literary analysis.[8] The development was continued in the 1950s with Father Roberto Busa utilising IBM mainframe computers and John W. Ellison using the UNIVAC I to produce lexical concordances.[9] Since then, computer-assisted history research has produced three 'generations', roughly following the advancement of computers and internet technologies. Simultaneously, there are continuities of parallel developments borrowed from, or developed together with, sister disciplines, such as text analysis in literary studies, statistical analyses in economic and social history, Geographic Information System (GIS) within geography, and digital image analysis in art history and visual studies. Allegedly, 'the first published work by an historian involving actual computerized research' came in 1963 with a 'scalogram analysis of voting patterns' in the British Parliament in the 1840s by William Aydelotte at the State University of Iowa.[10] A few years later, Viljo Rasila (Paju, this volume) did somewhat similar work in Finland. The first larger and more widespread application of computers was by the cliometricians of the 1960s, who were recognised as constituting the first generation of digital historians. They were followed by a second generation centred around the new 'personal' computers in the 1990s and were often seen as a part of the wider humanistic research field of 'humanities computing'.

The current third generation of digital history can be said to begin to emerge in the late 1990s and the early 2000s with the appearance of the first large digitised full text databases, such as Early English Books Online (EEBO) and Project Gutenberg,[11] and with the rebranding and expansion of humanities computing to digital humanities in the mid-2000s. Since the early 2000s, contemporary historians' toolkit has been expanded by an increasing volume of digitised sources and the swift development of computational analysis methods. This

was taking place at the same time as geographical history was going through a development from historical GIS (HGIS) to spatial history.

The snowballing growth in the amount of digital sources and the development of new research approaches and concepts has gradually increased the number of humanists using computational methods. One of the most frequently used concepts is *distant reading*, a perspective pioneered by the literary historian Franco Moretti. Distant reading can be understood as a counterpart to *close reading* that has been used extensively in humanities for distilling meanings from texts from the 1970s onwards. Distant reading has been used to extract meaningful patterns from textual sources, particularly when the number of texts are so numerous that it is impossible for a human to read them in a consistent manner.[12] The examples in this volume show that distant reading can also be a useful approach for exploring smaller amounts of text, as it provides another kind of approach to the texts in focus. Such machine or algorithmic reading provides 'another pair of scholarly glasses' and allows examining the sources from new perspectives. In the best case, close and distant readings complement each other.

Characteristic for this potentially paradigmatic digital history (Fridlund, this volume) is not just the introduction of new conceptualisations, such as 'distant reading', 'macroanalysis' or 'algorithmic reading', or the application of methodological tools such as topic modelling, but also the utilisation of novel practices for historians, new digitally augmented ways of working. Digital research brings along the collaborations of larger multidisciplinary group projects, the use of centralised technical infrastructures and machines. The changes that are taking place in history today are in several ways reminiscent of the changes that natural science disciplines such as physics and biology went through earlier with changeover from individuals' 'small science' tabletop experiments to interdisciplinary large team 'big science' collaborations.

The origin of this volume lays in an initiative to strengthen digital history research proposed by a collective of historians in Finland in 2015. That ambition was generously funded by the Kone Foundation through two interconnected projects 2016–2018, which brought together the majority of the authors in this volume. The first project, Towards a Roadmap for Digital History in Finland, aimed at identifying practical, professional and institutional obstacles and possibilities for developing digital history research. The second project, From Roadmap to Roadshow, built on the first one by bringing together digital historians to shape the best practices for disseminating knowledge about digital methods to historians so that in the end these would facilitate new digital history research. This was accomplished through a road tour to six major Finnish research universities, where the project organised presentations and workshops on emerging research and methodological developments within digital history.

Originally, the aim of the project was to end after the roadshow and to conclude with the subsequent publication of articles by the three main project researchers. However, the enthusiasm among the participants at the various

universities promised new digital research results in a not-so-distant future. Thus, towards the end of their shared work, the team decided to extend the project towards its logical conclusion by organising a workshop during the spring of 2018 that invited historians and specialists in digital research methods to come together to work collaboratively on formulating and answering a number of specific and concrete historical research questions. The historians who responded to the invitation brought their source materials and historical research questions, while the digital specialists contributed with their methodological expertise to jointly find answers to the research questions. At the workshop, the research teams analysed the sources to come up with preliminary solutions and answers and afterwards the teams were encouraged to keep working on their projects, and in this book, several of those projects are now brought to completion in the form of peer-reviewed research articles. They are complemented by articles from other digital historians, presenting results from a selection of the other recent research projects.

The majority of the research presented here is by digital historians active at Finnish universities. The rationale behind this is, in addition to the books' specific historical origins as explained above, that the emerging Finnish digital history community is both a representative and in many ways exemplary part of the larger international development of digital history. It is representative in that the used methodological research approaches correspond to the predominant directions of current digital history and thus the diversity and breadth of the studies presented in this volume, representing digital history research in a wide range of topics, from diverse disciplines from political, economic, cultural, intellectual and feminist history to history of science and technology and periods going from the Early Modern to the recent past. Taken together they provide a representative overview of the state-of-the-art of not just Finnish digital history research, but also of emerging digital history overall. Like most other research communities, the digital history landscape in Finland is diverse and dispersed, including bigger research groups, individual researchers and interdisciplinary and collaborative projects with national and foreign colleagues in Finland and abroad. This volume is exemplary in that digital history in Finland as a community and practice can be said to be more developed and institutionalised than in many other countries. In addition to several digital historians working at all levels of academic seniority, there are designated doctoral positions and professorships, textbooks, a regular digital history conference series and seminars and a digital history section within the national historical society. Compared to most other countries, the stage of digitisation of newspapers and archival documents is very advanced, which encourages digital history research. The common understanding of digital historians in Finland is that the focus of digital history research should be in finding answers to the research questions rather than utilisation of digital research tools just for their own sake. The contributions to this book, we feel, exemplify that critical evaluation of digital sources, metadata and research methods, and the results

they provide, are the basic components of good digital history research. Thus, the Finnish digital history research, together with the other contributions presented in this volume, should be a good representation of some of the most widely shared research practices emerging within the new digital history.

The New Digital and Distant Histories

This book contributes to advancing the field of history in primarily two ways, through new conceptual explorations of the past, present and future of digital history research and with new empirical historical knowledge coming out of research using digital methods. Through this, we aim to illuminate the new digital history's potential and pitfalls. We have divided the book into four parts. Part I 'The Beginning' consists of this introduction. Part II 'Making Sense of Digital History' starts with discussing the historical and methodological roots of digital history and contributes conceptual and contextual explorations of the current state of digital histories. Part III 'Distant Reading, Public Discussions and Movements in the Past' presents empirical case studies from various time periods that through the application of digital tools, primarily various forms of distant reading methodologies, demonstrate the further potential for expanding historical knowledge. The final Part IV 'Conclusions' draws the volume to an end by an exposition of the landscape of digital history and its future potential.

In the foreword, the late computer scientist and pioneering digital humanist Timo Honkela, draws on his wide experience of multidisciplinary cooperation using computational tools, to offer his thoughts on the digital future of history. In Chapter 2, providing a longer historical context for the new digital history, Petri Paju examines the history of computer-assisted history research from the 1960s until the 2010s. By focusing on one particular national development, that of historians' use of computers in Finland, he recognises how, although a particular national story, it was part of a larger, international and transnational pattern of development within digital history research.

After the overview of the roots of digital history, the subsequent chapters in Part II shed light on the fundamental components of digital history research: data, metadata and the mundane, often manual, work enabling the operation of our digital tools and resources. In Chapter 3, Jari Eloranta, Pasi Nevalainen and Jari Ojala exemplify how economic and business historians in many ways have been forerunners of digital history with computerised analysis of numerical and event code databases. They also share their experiences of the challenges to historical research of digitisation and uses of databases. Chapter 4 by Mats Fridlund attempts to conceptualise emerging historical practice by exploring the present state of digital history research according to two ideal types of digital history. Following Thomas Kuhn's theory of scientific revolutions, he describes them as 'normal' and 'paradigmatic' digital history. Further,

as a middle way between the two, he proposes one that is beyond the normal but still a less revolutionary form of semi-automatic digital history, described as 'digital history 1.5'.

This is followed by Chapter 5, which concerns research infrastructures, where Jessica Parland-von Essen calls for better data management and increasing the openness of data. She presents the FAIR (Findable, Accessible, Interoperable and Re-usable) approach to data, which would not only improve the efficiency, but also increase the trustworthiness and quality of historical research. The critical theme of the role of metadata in digital history research is taken up in Chapter 6 by Kimmo Elo, who points out that when focusing on data, we often neglect metadata, although it is a crucial part of the whole. In his chapter, he explores ways of improving the quality of the historian's metadata. Following this is a valuable reminder offered by Johan Jarlbrink in Chapter 7 on the importance of manual work to digital machine processing. In his chapter, he shows how digital research is far from automated, and that it actually requires countless hours of manual work which most of the time stays invisible and thus its problems and possibilities are often unnoticed and neglected.

The subsequent chapters offer a wide array of empirical case studies using a selection of digital research methods that exemplify how they can help us to reach for new understandings of the past. Beginning this series, in Chapter 8, Mirkka Danielsbacka, Lauri Aho, Robert Lynch, Jenni Pettay, Virpi Lummaa and John Loehr use statistical quantitative analysis to explore migration of Finnish individuals in the 20th century. Using a database that they have digitised and complemented with other historical data, they explore socio-demographic and environmental factors that can be combined with the domestic relocation and settlement of migrants. In Chapter 9, Heidi Kurvinen, in the vein of feminist history methodology, uses her personal experience of getting acquainted with historical text mining to explore traditional and not so traditional historians' experiences in encountering the new digital history methods. She notes that entering the field of digital history 'requires cultural and technological capital which marginalises researchers who do not have the skills to conduct digital analyses by themselves or do not have access to the organisational support'. Among the factors influencing the ability of researchers to participate, she identifies their gender. The next case study by Maiju Kannisto and Pekka Kauppinen in Chapter 10 illustrates the use of Named Entity Recognition (NER) to explore Finnish audio-visual history as it is presented in the public radio and television online archives. Their metadata analysis reveals interesting peculiarities in what kind of audio-visual imaginary of the past is provided by the dataset, and which elements of the national history it hides. In Chapter 11, Matti La Mela gives an excellent example of the opportunities of text analysis by tracing the history of the concept of *allemansrätten* (freedom to roam) in the Finnish parliamentary debates and argues counterintuitively to common knowledge that the present understanding of the concept has a surprisingly short history. His article also takes extra care in making the

methodological steps transparent to readers. In Chapter 12, Pasi Ihalainen, with the assistance of Aleksi Sahala, uses collocation analysis to study changes of the concept of 'internationalism' in 20th-century British parliamentary debates. By reconstructing the meanings attached to foreign political issues in the British Parliament from the early 19th century, they show that the 'international' has been associated in different ways during the various deliberations on the United Kingdom's membership in international organisations.

In Chapter 13, Melanie Conroy and Kimmo Elo, with the help of network analysis of the metadata of a large picture archive, explore the structure and temporal dynamics of the geospatial social networks of the East German opposition movement. They show how the network method can be used for exploring and visualising, as well as analysing, quantitative historical data. Reetta Sippola's contribution in Chapter 14 uses topic modelling to explore the evolution of the scientific discourse in the pioneering British scientific journal *Philosophical Transactions* in the mid-18th century. In her study, the method of arranging the data makes topic modelling reveal previously neglected themes and unnoticed temporal changes in the discourse. Heidi Hakkarainen and Zuhair Iftikhar also use this methodology in Chapter 15, in the expanded form of dynamic topic modelling, to focus on the formation of the concept of 'humanism' in the early 19th-century German-language press. They show how reaching reliable analysis results demands a deep understanding of the context, skills and time, but how the method has the potential to challenge established patterns of thought and underlying presumptions by providing a novel perspective on the sources. In Chapter 16, Reima Välimäki, Aleksi Vesanto, Anni Hella, Adam Poznański and Filip Ginter study author attribution and apply methods based on neural networks to explore their medieval cases of authorship recognition. Their intriguing results show how the uses of 'black-boxed' computational methods can potentially help us to solve centuries-long debates on the attribution of authorship. In the final case study in Chapter 17, Risto Turunen uses advanced collocation analysis to study Finnish labour newspapers during the late 19th and the early 20th centuries. With that material, he takes a macroscopic approach to study expressed temporality of the papers and especially the 'sun of socialism', which differed from the biblical sun shining on all and in this 'highlighted earthly problems'. Towards the end of his chapter, Turunen turns his discussion to the present situation and to future aims of digital history.

Jo Guldi concludes the volume in Chapter 18 by drawing a wide picture of the potential game-changing nature of digital history. She stresses the universal character and widely applicable nature of digital research methods: researchers of Chinese industrialisation can find a method used by a medievalist also useful to their research and vice versa. She also predicts that with the increasing number of digitised sources and utilisation of digital methods, we may see a rise of *longue durée* in history, which as she puts it could provide new findings that 'border on the breathtaking'.

New Historical Challenges and Criticisms

Digitisation and computer-assisted research tools open new possibilities, but also bring novel challenges and criticisms to the history discipline. There is a need for a wider methodological discussion on how digital research methods could and should be used in history research. To be able to take part in interdisciplinary collaboration, it is important for historians to have a discussion on what digital methods mean, and where they can lead us. The ambition is that the studies in this book will contribute to foster this discussion. Among the critical components of digital history research that are addressed in the following chapters are digitisation of sources, creating metadata for digital source materials, human–computer interaction and digital research methods. These are only a few of the critical issues troubling current digital history.

One of the most pressing questions in digital history research is access to and problems of *digitised sources*. Although also important to scholars in other disciplines, they are fundamental to historians. The availability of consistent digitised collections with long time series is one of the critical prerequisites of digital research. Simultaneously, the existing digitised sources invoke discussions of their availability and usability, and what overall should be digitised. Furthermore, digitisation also changes the object of research, as a digitised newspaper is not the same as the physical object of a newspaper. When digitising sources, we, as Mikko Tolonen and Leo Lahti have pointed out, also lose important elements of the physical objects.[13] The consensus of the scholars contributing to this book is that the readily available digitised sources should be used with the same or even higher level of source criticism than before. While the existence of easily accessible digitised sources is a crucial requirement of digital history research, non-problematised use of data—a kind of 'source myopia'—has the potential to skew the historiography towards the most readily available databases and source material, rather than the most important or representative, and thus possibly motivate researchers to study them instead of the sources that, digitised or not, would provide the best answers to the research question (Chapter 3, this volume). For example, the very popular usage of newspapers as sources, especially for historical studies before the 20th century, is not necessarily because they are the most relevant historical sources, but is rather due to the simple fact that newspaper collections have in many countries been extensively digitised.

In digitising historical sources, the digitiser faces several practical choices that have extensive effects on historical research. The first major question is what to digitise. In making such basic selections, there is a threat of repeating and amplifying the biases of the past knowledge constructions, leaving less prominent and marginalised topics aside. The sources chosen to be digitised, the ways in which they are digitised and shared, have far-reaching consequences.[14] Memory organisations, such as archives and libraries, often begin their digitising efforts from sources that are most often used by the general public and

researchers, and are thus considered to be more important. This common practice creates a threat to further marginalise less prominent topics and to exclude less studied materials. Therefore, alongside the use of digitised sources in readily available collections, the ability for historians to digitise their own sources is becoming an increasingly important skill. Learning how to digitise, and setting the best practices for digitising, data life cycles, and sharing digitised sources among historians are emerging important additions to the historian's toolbox. To increase the variety of the available digitised sources, it is valuable that historians learn to digitise sources on their own, and whenever possible, share their new data. The authors of this volume use both readily available databases and sources that they themselves have digitised. Digitisation is time consuming, and therefore the sharing of data is an important means of widening the base of digitised sources.

In addition to digitised sources, a key issue for digital history research is *metadata*, the data that describes and gives information about the digital data (sources), and especially concerning its varying quality. As Kimmo Elo points out in Chapter 6, 'more attention should be paid and more resources should be invested in metadata creation'. From this perspective, the real problem is not the structure of a data system itself (its 'ontology'), but rather the process of creating source material's metadata. The principles of adding metadata to the documents are often rather unsystematic and not transparent, and only too often the usefulness of (meta)data depends on the person creating and inserting the metadata. For example, at the workshop described above, one research team planned to work on metadata of images from a public source database (www.finna.fi). After some trials with that material, they ended up terminating their project because of the overly scattered and random character of the metadata collection. The large amount of processing necessary to enable digital methods to be applied would not have made it possible to finalise their project within a reasonable timeframe. However, this attempted project was not in vain, as it partly inspired one of its participants (Elo) to write a chapter on metadata and digital history for this volume.

The new kind of source material for historians in the form of digital data and metadata makes it important for digital historians to develop a new *digital source criticism*. Compared to the pre-digital era with large amounts of data in non-digital forms, the contributions in this volume demonstrate how digitisation instead of selective sampling allows historians to use all the available data in their analysis, and thus more systematic analysis. Interestingly, distant reading of large datasets often exposes the used databases' borders and restrictions better than traditional sampling for close reading. For example, the analysis of Kannisto and Kauppinen revealed the biases and partiality of the studied dataset. Using digitised sources demands deep understanding of what the data consists of because, as Eloranta, Nevalainen and Ojala point out in Chapter 3, straightforward and non-problematising data usage may lead to missing the key issues of the data and misleading interpretations of the

historical processes. In big data lie opportunities also for significant misinterpretations and falsifications.

As the contributions in this volume demonstrate, undertaking digital history research is often more time consuming and demands perhaps more conscious methodological choices than the traditional history approach. When one is undertaking digital history research, it becomes evident that alongside the algorithms used, the selection, creation, cleaning and filtering of the data heavily influence the results of the computer-assisted analysis. As Johan Jarlbrink shows, the digital research process at many stages demands manual work to be done, such as data cleaning. He demonstrates that this work is not only a necessary precondition for the analysis, but is actually in itself an important part of the analysis, as the researcher gets to work on and read through the material several times, and in this way learns to know the data in depth. While the quantitative digital analysis makes the conclusions more convincing, the in-depth knowledge of the data provides crucial qualitative understandings that guide the interpretations of the quantitative analysis.

Connected to this new source criticism, there is also a need to develop what has been described as a *digital resource criticism* (Chapter 4, this volume). This refers to the need, in order not to draw false conclusions, to be better aware of the internal technological logics of the digital resources used by historians, such as that of a database or a search engine. Similar questions of an awareness of the opportunities and limitations of the available resources and methods have lately been raised in reference to representation and visualisation of historical data.[15] One example of this is how Maiju Kannisto and Pekka Kauppinen in their study (see Chapter 10, this volume) found out that the frequencies of the search terms in the metadata did not reflect the actual frequencies of the audio-visual material to which the metadata referred, but that they were more an artifact of the processes of how the metadata had been produced. Both Elo (Chapter 6) and Kannisto and Kauppinen (Chapter 10) suggest in this book that archivists and historians should collaborate more and in this openly discuss the principles and practices of metadata formation, and how they could best serve all the parties.

Furthermore, the chapters of this book point out the methodological zigzag between distant and close reading of data, the repetitious adjustment of the algorithmic parameters, the evaluation of the means of the data formation, its broader context and preceding research, all involved in an overall research process of trial and error. Sippola, Kurvinen, and Hakkarainen and Iftikhar all show how the choices of the researcher influence the outcomes of the research. For example, when using topic modelling, the testing of the results with varying numbers of topics is a very important step in the process of analysis. Simultaneously, the scholar's understanding of the context is essential in identifying the meaningful results, and to be able to differentiate them from the potential nonsense produced by the computer, to discern the historical signal from the data noise. Usage of digital research methods amplifies the research findings, but

they also amplify the potential of false results. Computers and algorithms are important helpers, but they cannot operate on their own: they always require human guidance.

Despite all these challenges, the contributions to this volume demonstrate how computational analysis can disclose new and previously unnoticed patterns in history. For example, in Chapter 12, Ihalainen summarises the benefits of computer-assisted analysis for his study on conceptual history by stating that it revealed associations between the studied concepts, which made it possible to estimate trends in political attitudes and revealed particular and peculiar political issues that would have been very difficult to find with traditional methods. Along the same lines, Kurvinen states, in Chapter 9, that combining digital analysis and close reading allowed her to identify topics that might have remained unnoticed otherwise and exposed new ways of perceiving the material, ways that could prompt novel and previously unresearched questions.

The new digital history might also foster a wider rethinking of the parameters of historians' professional practice. Digital research methods create new and at times more stringent demands on accuracy, methodological thinking, self-organisation and collaboration than traditional historical research. As Kurvinen points out, digital environments could encourage historians to conduct their research in 'a more self-aware manner when every step of the process needs more thought than a traditional day with paper archives'. Similarly, Eloranta, Nevalainen and Ojala point out in Chapter 3 that collaborative research on digital data can lead to more efficient and accurate research, but it requires the development of a different professionalism from researchers. Jessica Parland-von Essen shows in Chapter 5 the importance of historians starting to manage their data in a more qualified manner to themselves so they become more like data curators and archivists, and including thinking about the preservation and reusability of research data from a longer-term perspective. To support the development of such new practices in historical disciplines, there is a need for historians to participate in developing new joint practices that support FAIR data and thus better research. This calls for collaboration among historians and memory organisation specialists, and for historians to reach outside of history to seek out ideas from other disciplines facing similar challenges.

Most of the chapters in this volume were written collaboratively. Along the process of our project, it was confirmed that digital history research demands interdisciplinary collaboration, since it is rare that a historian manages to combine in him- or herself both the skills of the historian and of the programmer. That said, it is not necessary for the historian to become a programmer. What is needed is the ability to collaborate and work together in an interdisciplinary manner with collaborators who bring expertise from the domains of computer and information science.[16] The above-mentioned workshop proved that fruitful collaboration with IT professionals is not only needed, but also feasible and beneficial. And this book proves that it can bring new knowledge, as well as conceptual developments, to the field of history.

One basic challenge, nevertheless, is that although multidisciplinarity is much-needed in the realm of digital humanities research, it is well known that not all computer-related questions or tasks carried out in digital history research are challenging enough to peak the interests of computer scientists. For example, the application of a ready code to a dataset is for a traditionally trained historian often too challenging a task, but rather trivial for a computer scientist. Thus, there is an increasing importance for universities and research institutions to be able to provide more mundane and routine technical support to historical researchers through their libraries, IT support facilities or other means, much like before the widespread availability of easily accessible online databases and online sources such institutional structures were central in assisting historians in finding research literature and source materials.

Conclusion

It seems evident that history research has been and will continue to be increasingly influenced by society's overall digitalisation. Still, the historians in general would benefit from being more aware than before of the interaction between historical research and the digitising world around them in order to stay both critical and constructive towards the changes and continuities of today. This includes taking advantage of the latest tools, as well as exploring their limitations to be able to keep our methods up to date and to gain a better understanding of the possibilities and pitfalls of historical research in the digital era.

As always, the future holds both promises and threats for historians, digital and otherwise. Although it is essentially an older condition, the skills and resources needed for digital history research could broaden the gap between history departments that are better positioned and those that are not, and consequently create more divisions among historians. One key issue for the digital historians is how to succeed to excel in using and developing new methods, while simultaneously avoiding overlooking the values of more traditional research. Doing and succeeding with the new explorations, while also respecting the older known and tried ways, has often shown to be the best working path towards the future.

In a similar vein as the encouragement by Jo Guldi and others in this book, one lesson from sociologists and historians of technology has been that users matter, that they, rather than being passive adopters of new technology delivered in black boxes, can have their say in influencing the direction of technological change, and at times even open up and reconstruct their tools so they better fit their particular needs and desires.[17] Historians as a group can and should be active in making choices and guiding their discipline towards an ever-more digital world of tomorrow, a tomorrow that soon will be a past and needs its born-digital history researched.

After almost 50 years, perhaps we have finally arrived at Emmanuel Le Roy Ladurie's 'tomorrow'. Or maybe we are already far beyond that—not least as most of the authors in this collection would not identify themselves as doing the quantitative kind of history Le Roy Ladurie expected future historians to be doing. As historians, we can recognise how difficult it is for history's actors to foresee future developments, and that while Le Roy Ladurie correctly predicted that historians needed to learn to harness computer technology for their work, neither he nor his colleagues could hardly have imagined the possibilities of the information technology at historians' disposal in the early 2020s. However, in the sense that historians should learn how to make the most of the 'computer', we feel that the historians in this book with their new digital and distant histories have tried to live up to his hopes by going towards and away from his tomorrow to reach our today and its past, present and future digital histories.

Notes

[1] Le Roy Ladurie 1979: 6. Rabb wrote: 'In I967, the basic posture of quantitative historians was a mixture of brashness and defensiveness. Le Roy Ladurie was sufficiently impressed by the discussions at Ann Arbor to predict that "the historian will be a programmer or he will be nothing"' (Rabb 1983: 591).

[2] William G. Thomas III quoted in Cohen et al. 2008: 454.

[3] See Jones 2014.

[4] For some of the major books published within the new digital history, see: Staley 2002; Cohen & Rosenzweig 2006; Galgano et al. 2008; Schmale 2010; Gantert 2011; Genet & Zorzi 2011; Haber 2011; Rosenzweig 2011; Clavert & Noiret 2013; Dougherty & Nawrotzki 2013; Jockers 2013; Weller 2013; Graham, Milligan & Weingart 2015; Bozic et al. 2016; Koller 2016; Brügger 2018.

[5] Jockers 2013; Blevins 2014. See also Guldi & Armitage 2014.

[6] As we well know, historical 'firsts' are often contested and contextual.

[7] The field of spatial history evolved from within Historical Geographic Information Systems research starting in the 2000s. See Gregory & Geddes 2014: x, xii, xiv–xv.

[8] Sagner Buurma & Heffernan 2018.

[9] Jockers 2013: 3; Vanhoutte 2013: 127–128.

[10] Swierenga 1970: 5.

[11] Although these collections also have much longer histories. See Lebert 2008.

[12] Moretti 2000, 2005, 2013. See also Underwood 2017.

[13] Tolonen & Lahti 2018.

[14] See, for instance, Jarlbrink & Snickars 2017.

[15] Foka, Westin & Chapman 2018.

[16] See also Fickers & van der Heijden 2020.

[17] See Oudshoorn & Pinch 2003.

References

Blevins, C. (2014). Space, nation, and the triumph of region: a view of the world from Houston. *Journal of American History, 101*(6), 122–147.

Bozic, B., Mendel-Gleason, G., Debruyne, C., & **O'Sullivan, D.** (2016, 25 May). *Computational history and data-driven humanities: second IFIP WG 12.7 international workshop, CHDDH 2016, Dublin, Ireland, revised selected papers*. Berlin and Heidelberg: Springer.

Brügger, N. (2018). *The archived web: doing history in the digital age*. Cambridge, MA: MIT Press.

Clavert, F., & **Noiret, S.** (Eds.). (2013). *L'histoire contemporaine à l'ère numérique—Contemporary History in the Digital Age*. Brussels: Peter Lang.

Cohen, D. J., & **Rosenzweig, R.** (2006). *Digital history: a guide to gathering, preserving, and presenting the past on the web*. Philadelphia, PA: University of Pennsylvania Press.

Cohen, D. J., Frisch, M., Gallagher, P., Mintz, S., Sword, K., Taylor, A. M., Thomas, III, W. G., & **Turkel. W. J.** (2008). Interchange: the promise of digital history. *Journal of American History, 95*(2), 452–491.

Dougherty, J., & **Nawrotzki, K.** (Eds.). (2013). *Writing history in the digital age*. Ann Arbor, MI: University of Michigan Press.

Fickers, A., & **van der Heijden, T.** (2020). Inside the trading zone: thinkering in a digital history lab. *Digital Humanities Quarterly, 14*(3). In M. Oiva & U. Pawlicka-Deger (Eds.), *Lab and slack: situated research practices in digital humanities*, special issue.

Foka, A., Westin, J., & **Chapman, A.** (Eds.). (2018). Technology in the study of the past. *Digital Humanities Quarterly, 12*(3), special issue. Retrieved from http://www.digitalhumanities.org/dhq/vol/12/3/index.html

Galgano, M. J., Arndt, C., & **Hyser, R. M.** (2008). *Doing history: research and writing in the digital age*. Boston, MA: Thomson Wadsworth.

Gantert, K. (2011). *Elektronische Informationsressourcen für Historiker*. Berlin: de Gruyter.

Genet, J.-P., & **Zorzi, A.** (Eds.). (2011). *Les historiens et l'informatique: un métier à réinventer*. Rome: École française de Rome.

Graham, S., Milligan, I., & **Weingart, S.** (2015). *Exploring big historical data: the historian's macroscope*. London: Imperial College Press.

Gregory, I. N., & **Geddes, A.** (2014). Introduction: from historical GIS to spatial humanities: deepening scholarship and broadening technology. In I. N. Gregory & A. Geddes (Eds.), *Toward spatial humanities: historical GIS & spatial history*. Bloomington, IN: Indiana University Press.

Guldi, J., & **Armitage, D.** (2014). *The history manifesto*. Cambridge: Cambridge University Press.

Haber, P. (2011). *Digital Past: Geschichtswissenschaft im digitalen Zeitalter*. Munich: Oldenbourg.

Jarlbrink, J., & **Snickars, P.** (2017). Cultural heritage as digital noise: nineteenth century newspapers in the digital archive. *Journal of Documentation, 77*(6), 1228–1243.

Jockers, M. L. (2013). *Macroanalysis: digital methods and literary history*. Urbana, IL: University of Illinois Press.

Jones, S. E. (2014). *The emergence of the digital humanities*. London: Routledge.

Koller, G. (2016). *Geschichte digital: historische Welten neu vermessen*. Stuttgart: Kohlhammer Verlag.

Lebert, M. (2008). *Project Gutenberg (1971–2008)*. University of Toronto and Project Gutenberg. Retrieved from http://www.gutenberg.org/ebooks/27045

Le Roy Ladurie, E. (1979). *The territory of the historian*. Translated from the French original (in 1973) by B. Reynolds and S. Reynolds. Brighton: The Harvester Press.

Moretti, F. (2000). Conjectures on world literature. *New Left Review, 1*(1), 54–68.

Moretti, F. (2005). *Graphs, maps, trees: abstract models for literary history*. London and New York, NY: Verso Books.

Moretti, F. (2013). *Distant reading*. London and New York, NY: Verso Books.

Oudshoorn, N., & **Pinch, T.** (Eds.). (2003). *How users matter: the co-construction of users and technologies*. Cambridge, MA: MIT Press.

Rabb, T. (1983). The development of quantification in historical research. *Journal of Interdisciplinary History, 13*(4), 591–601.

Rosenzweig, R. (2011). *Clio wired: the future of the past in the digital age*. New York, NY: Columbia University Press.

Sagner Buurma, R., & **Heffernan, L.** (2018, 11 April). Search and replace: Josephine Miles and the origins of distant reading. *Modernism/Modernity Print Plus*. Retrieved from https://modernismmodernity.org/forums/posts/search-and-replace

Schmale, W. (2010). *Digitale Geschichtswissenschaft*. Vienna: Böhlau Verlag.

Staley, D. J. (2002). *Computers, visualization and history: how new technology will transform our understanding of the past*. London and New York, NY: Routledge.

Swierenga, R. P. (1970). Clio and computers: a survey of computerized research in history. *Computers and the Humanities, 5*(1), 1–21.

Tolonen, M., & **Lahti, L.** (2018). Digitaaliset ihmistieteet ja historiantutkimus. In M. O. Hannikainen, M. Danielsbacka, & T. Tepora (Eds.), *Menneisyyden rakentajat: teoriat historiantutkimuksessa*. Helsinki: Gaudeamus.

Underwood, T. (2017). A genealogy of distant reading. *Digital Humanities Quarterly, 11(2)*.

Vanhoutte, E. (2013). The gates of hell: history and digital | humanities | computing. In M. Terras, J. Nyhan & E. Vanhoutte (Eds.), *Defining digital humanities: a reader*. Farnham and Burlington, VT: Ashgate.

Weller, T. (Ed.). (2013). *History in the digital age*. London and New York, NY: Routledge.

Part 2
Meaning Making in the Digital Age

Uses of Computer Technology in Historical Research

Petri Paju

> Kranzberg's First Law reads as follows: Technology is neither good nor bad; nor is it neutral.[1]

Historians have rarely been associated with the latest IT, or the other way around. In broad terms, the same applies to all IT, both old and new, and history research; they seem a world apart, unless one counts things such as pens and books. In their publications, most historians make it look like their use of information technologies is unbiased and unproblematic. However, Melvin Kranzberg, who was a veteran historian of technology, reminded us that technologies always come with consequences. With digital history, and the growing use of computational methods in historical research, this practice and performance of neutrality vis-à-vis technological tools, as well as the old stereotype, could be changing.

In reality, IT such as computers has been utilised in history research since the 1960s, as in most other walks of life. At that time, a few historians in the United States (and elsewhere) started to explore the usability of mainframe computers for their work.[2] In over 50 years, computer-assisted history research has evolved, or graduated, from the tests of a very few scholars into an emerging

field of computational history, also called more broadly digital history research. Of course, one should inquire if those are phases and part of the same continuum or rather separate developments with no tangible influence from the former to the latter. In any case, this development seems to be something else than a straightforward progression.

This chapter focuses on the history of computer use by historians, drawing its evidence mostly from Finland, but with an emphasis on the researchers' transnational influences. To explore this evolution, this chapter asks: What have historians been doing professionally with computer technology, and when did that begin in Finland? What were their international influences in developing the use of computers in history research?

Here, 'computer technology' refers to the technological developments connected to computers and IT during the research period: in this case, its evolution from the relatively large mainframe computers to microcomputers, to internet and beyond. The focus in this chapter is on historical research, thus mostly excluding teaching history with the help of or via IT, as well as technological changes related to publishing.

Interviews and memoirs, various written documents, especially digitised history journals, and observations (since the late 1990s) are used in answering these questions.[3] With these materials, the chapter aims to examine this development from several different levels and viewpoints. These range from the individual scholar(s) to their collaboration and extend into libraries and archives, and institutional use and support of digital means to advance research in the field of history.

One important motivation behind these questions is to distance the researcher and readers from the present terminology concerning digital humanities and digital or computational history, which often seem to make studying their own development very confusing. Without these concepts, I hypothesise, we can better approach and understand historical events and trends on their own terms.

While research in historiography had tended to value and focus on the theoretical aspects of historical thinking and research, this chapter highlights the more practical side of carrying out historical research and thus contributes to a more balanced idea of how historians conduct their work. A better, increased understanding of the now mundane technologies and practices of historians is especially appropriate now that the discipline is facing yet another change towards an increasingly electronic and more digitised research process, with new and more powerful computational tools, which present challenges to historians themselves, but also to teaching and outreach to the public.[4]

Further, for the international discussion, this chapter serves as a reminder of and correction to the US-centric or Anglo-American view of history of computing-assisted history. This too was an international and transnational development.[5] In international comparison, the number of Finnish historians was fairly limited. After rapid growth in the 1960s, there were, in 1970, historical research units in six Finnish universities, employing a total of 32 professors.

Since then, the community grown to the extent that, in 2015, there were 56 history professorships in eight research units, but the profession has expanded greatly, especially when one counts all historians with doctoral education.[6] Nevertheless, from early on, this community of historians in Finland took part in most if not all transnational trends and developments in their field and adopted major new technologies used by historians in industrialised countries. In general, then, Finnish historians' experience of using computers can be thought of as rather representative of other Western countries. The few untypical aspects will be highlighted.

Computer Usage Starts in the Late 1960s

According to the digitised version of the *Historiallinen Aikakauskirja* (*Historical Journal*) in Finland, the word 'computer' (*tietokone*) was first mentioned on its pages in a book review in 1964.[7] One early Finnish historian to make use of computers, Pertti Huttunen, later wrote that he became interested in using computers during that same year, in 1964, while extending his studies and planning his doctoral dissertation in Rome, Italy. There, he first talked about such an option with a Finnish physicist and also visited a local computing centre.[8]

Following examples abroad, a small number of historians had started to familiarise themselves with mainframe computers in the mid-1960s. The first public discussion about computers by historians in Finland took place in the spring of 1967. At that time, the Historiallinen Yhdistys ry. (Historical Association), or younger generation of historians, had invited historians Kaarlo Wirilander and Pertti Huttunen, a well-known senior researcher and a doctoral candidate respectively, to talk about 'The historian and the computer'. At the meeting, an IT specialist from the Helsinki University's computing centre, Jorma Torppa, offered technical expertise.[9]

Before this seminar in Helsinki, historian Viljo Rasila had joined the first short, introductory course given by the new computing centre at Tampere University. The centre had installed its first computer in 1966. The following year, Rasila became the first historian in Finland to publish an article about using computers in the national *Historiallinen Aikakauskirja*. In it, he mentioned the work of Wirilander, Huttunen, the 'brick group' studying Roman brick stamps and his own as examples of history research involving computers in Finland. According to Rasila, this computer use by historians was just beginning.[10]

This use so far included collecting and inserting data into (punched) cards, which were meant for building databases (to create tables and to compile statistics) and performing calculations. Rasila himself was applying multivariable analysis, and specifically factor analysis, to weigh up the various reasons for the civil war in Finland. That same year (1967), Pertti Huttunen published an article outlining his ideas about how to use computers to study Roman social

history. His article was published as the first volume in the series *Studia Historica* from the young University of Oulu (founded in 1958) in northern Finland.[11]

The following year, Viljo Rasila was the first to publish a history book, a monograph where he applied computer-aided statistical methods to explore key themes in recent Finnish social history during the 1918 war. His main computational method, factor analysis, had been developed in the field of psychology. The book, *Kansalaissodan sosiaalinen tausta* (*Social background of the civil war*), appeared in 1968.

Heikki Waris, a professor and social historian at the University of Helsinki, reviewed the study for the *Historiallinen Aikakauskirja* and thanked Rasila especially for introducing new methods for historians to use.[12] In the same issue of the journal, however, Pertti Järvinen from the computing centre at Tampere University discussed Rasila's book and heavily criticised his choice of a statistical method. In his book's preface, Rasila acknowledged the computing centre and its 'mathematicians' who had helped him, but, importantly, Pertti Järvinen had not been involved in Rasila's project. Instead, Järvinen had taken an independent interest in this innovative approach to history and likely became the first computing professional to share his ideas in this journal.[13] All in all, Rasila's study accompanied many firsts simultaneously.

Issues of multidisciplinary soon impacted Pertti Huttunen. Based on an analysis by a colleague, it seems Huttunen's dissertation manuscript on Roman social history faced harsh criticism from a classical philologist in Helsinki, which led Huttunen to move his dissertation project to the University of Oulu.[14] For sure, such difficulties and change did not support finishing the study, but, importantly, they were not directly associated with the new, computerised method applied by Huttunen. He never returned to work in Helsinki, but forged a career in researching and lecturing (for instance, about the history of technology) in Oulu and in other universities.

Pertti Huttunen defended his doctoral dissertation and book *The social strata in the imperial city of Rome* in 1974. Arguably, Huttunen wrote the first Finnish doctoral dissertation in history to use computerised methods, although that same year (1974), Reino Kero also defended his doctoral dissertation of general history at the University of Turku, and he too had used a computerised method in his study on migration.[15]

Regarding the feedback surrounding his 1968 book, Viljo Rasila recalled in my interview with him that the method was widely noticed, but at that point in time it raised mostly confusion:

> The reception of the mathematical analysis was controversial. Researchers of economic and social history, Eino Jutikkala among them, welcomed it as opening new opportunities, but the school of historians following [Professor Pentti] Renvall and doing textual analysis ('*renvallilainen tekstianalyysiin nojaava koulukunta*') shunned it and doubted its usefulness.[16]

This ambiguity is relatively easy to understand when one considers the technological and data-processing options available at the time. Starting from mainframe computers and the programs available on them, computer technology for a long time worked mainly for quantitative research and did not really fit qualitative research designs. First and foremost, there was virtually no data to be processed in digital text formats. At this time, computers and the promise they represented undoubtedly encouraged historians (as well as social scientists before them) to carry out quantitative research, which grew more popular in universities during the 1970s. In certain history departments, this period left a relatively strong tradition of quantitative history research that has been more or less carried on ever since.

Nevertheless, it is important to note that historians had applied quantitative and computational methods in their research even before computers were available. In Finland, the breakthrough of these approaches occurred in the early 1960s, if not somewhat earlier.[17] In an interesting simultaneity to historians first learning about the use of computers, the first independent department of economic history in Finland, at the University of Helsinki, was established in 1966. Unlike the '"old" Finnish economic history' which was later seen as rather descriptive, the new economic history became characterised by 'systematic application of quantitative methods'.[18] From this perspective, embracing computers was not a beginning nor a revolution, but part of an evolutionary development in the scholarship of history. It was a step further, which later perhaps seems to us a bigger change than it actually was. However, this longer intellectual background of quantitative history, going back at least until the last decades of the 19th century, has been studied elsewhere.[19]

How did historians compare with social scientists in computer use? For instance, Kullervo Rainio, later Professor of Social Psychology at the University of Helsinki, visited Finland's first operational computer, an IBM 650, at a state-owned bank soon after the machine's inauguration in 1958. At that time, he took part in a visit arranged for the Suomen Psykologinen Seura (Psychological Association of Finland), and in 1960 he could learn using another computer in Helsinki with his complex mathematical calculations needed for simulating group behaviour in a computer program.[20]

In general, we can safely say that social science researchers started using computers well before historians.[21] In Tampere University, which until 1966 carried the name Yhteiskunnallinen korkeakoulu (College of Social Sciences), Viljo Rasila had for years been in the company of mostly social scientists and had become familiar with their statistical methods. This environment partly explains his early interest in and initiative to test and use a computer for scholarly work in history.

One could also surmise that Rasila was in a position to fully cooperate with social scientists at Tampere University, but that was not the case. When I interviewed him, he told me that there was a major political difference between himself (he was more conservative) and his colleagues who, for instance, in the

department of sociology, were politically quite left-wing. Despite the shared interest in using computers, this political dissimilarity caused them to maintain a working distance from each other.[22]

In this respect, Rasila was rather typical. For a long while in the 1970s too, I suspect, this was a more general pattern: when compared with social science departments, history departments were much more conservative, including politically. This points out, intriguingly, that many contextual, historical factors could have an effect on and limit the circulation and exchange of scientific and scholarly tools such as the use of computer programs.

Tellingly of this technological milieu and the options available, it was predominantly a few researchers in social and general history who first started making use of computers. In the 1960s and the 1970s, the group of active history researchers totalled a few hundred, so they all knew each other and knew what others were doing,[23] even if those using computers remained a tiny minority. Further, Viljo Rasila penned a textbook entitled *Tilastolliset menetelmät historiantutkimuksessa* (*Statistical methods in history research*, 1973, 2nd edition 1977), including examples of computer-assisted operations, and that book became widely known among the profession, and especially among history students.

In summation, during roughly the first decade of computer use by historians, they used IBM and other mainframe computers for statistics, saving collected data, evidence, storing and processing it, forming tables, and then carried out various kinds of calculations and statistical analysis.

Research Projects: The 1970s

The early 1970s saw a new phase in historians' use of computers when the technology was incorporated into research projects. Such projects were considered fashionable, and the reorganised Academy of Finland granted funds for up-to-date research projects in the field of history too. In 1971, for instance, Vilho Niitemaa, Professor of General History at the University of Turku, presented a newly funded project focusing on people who have emigrated from Finland to distant countries (known as *kaukosiirtolaiset* in Finnish). The project included what Niitemaa labelled the 'ADP department', or individuals working on data collecting and compiling statistics with automatic data-processing tools. To store data, they used punched cards. The first doctoral dissertation to emerge from this project was written by Reino Kero, who, as mentioned above, defended his thesis in 1974.[24]

Conducting research in organised projects had become more common in the sciences in postwar decades. In the leading history journal, *Historiallinen Aikakauskirja*, several Finnish researchers wrote about current historical research projects in Sweden from the late 1960s onwards, and these reports included a

few mentions of ADP systems which were either being tested or were already in use to store and handle information.[25]

Thus, historians continued to use computers for organising data and for statistical purposes in the 1970s, but, for them, making use of the 'computer' (as technology) had also become a tool for winning research funding. Using computers signalled taking part in advancing research with the latest ideas and technology, and being at the forefront of development.

Viljo Rasila's expertise in computers played a major role in encouraging a collaborative research project called *Muuttoliikeprojekti* (Migration Project), which focused on migration within Finland between 1850 and 1910, with a particular focus on industrialisation. That project was led by Professor of Finnish History, Pentti Virrankoski, from 1977. Virrankoski also directed one sub-project at the University of Turku while Rasila, now an appointed professor, led another research team at Tampere University, and Yrjö Kaukiainen a third team at the University of Helsinki. In this project, the workload for collecting data manually grew much larger than was anticipated. Still, the difficulties with the ADP programs and processing the data proved to be even more significant. Because of these surprises, the larger project ran out of funding in the early 1980s. Most of the human-collected and manually input data was never computerised.[26]

However, the sub-project team at Tampere constructed their database differently from that of the Turku team, and consequently the Tampere team and Rasila himself were able to use and process their materials with a computer, and publish research results. Importantly, the larger project had formed ties with the Swedish project already building a demographic database in the late 1970s, and they exchanged experiences in international seminars.[27] Surprisingly, there are hopes that this Tampere database could be used anew in the early 2020s, once again inspired by the Swedish example.[28]

In principle, such databases can have a very long lifespan. Nevertheless, the opposite seems to have been the rule, so that many Finnish projects collecting and processing data in history research have produced a very ephemeral legacy. Their datasets were left in archives with data formats that basically died out within a rather short period of time.

The international discussion concerning historians' use of computers was increasing from the late 1960s onwards. In that exchange, Finnish historians rarely contributed publications, although Viljo Rasila, at least, published two articles in international journals such as in the 1970 volume of *Economy and History*. Importantly, however, during the 1970s and continuing well into the 1980s, Finnish scholars had relatively dynamic transnational communications, especially with their Estonian colleagues from the Soviet Union who had pioneered using computers in history research. Juhan Kahk was one of several such Estonian colleagues who published studies (using both the Finnish and the English languages) also in Finnish history series.[29]

Microcomputers for Text Processing: The New Typewriter (Plus)

The impact of the computer on historians' practice was not only as a calculator, but even more so as a word processor. Typewriters were already being advertised for historians in *Historiallinen Aikakauskirja* in 1916. It took time, however, before they began to be widely used by historians. And relatively soon afterwards, the latest products of the IT industry emerged: smaller computers that could be used as an advanced typewriter. The spread of personal computers (PCs) or microcomputers opened up new possibilities for historians in the early 1980s.[30]

In Finland, Jussi T. Lappalainen was the first person to write to historians about the possibility of using a computer to write texts. He had heard of such a novelty from his son Vesa, who studied mathematics. Lappalainen explained that he first thought of writing archival notes on a computer in place of using the long-used edge-notched cards (or edge-punched cards, *neulakortti*). Father and son then co-wrote a short article entitled 'Historical research without papers', which was published in *Historiallinen Aikakauskirja* in 1983. At the time, Jussi Lappalainen, who had previously worked at the University of Jyväskylä, was as Associate Professor of Finnish History at the University of Turku.[31] When the first, still quite expensive, microcomputer landed in the history department's office in Turku, his colleagues were afraid of using it. Lappalainen, however, was convinced about the device's potential and wrote another article entitled 'Making text on the screen', after which his colleagues began to telephone him to glean some clarification. As a former publishing editor, Lappalainen also persuaded the popular Finnish novelist Kalle Päätalo to migrate to using a computer for his work. The learning phase involved some text vanishing from the computer's memory (or from the writing software) and this made the angry author revert to the typewriter for a while.[32] Despite the new technology, then, the (anticipated) main use of these new machines was familiar; it was typing. Computers replaced typewriters, and most of the historians started using computers as not-yet-so-advanced typewriters. Yet, social science historians soon discovered ways in which the PC could do more.

In 1985, a new historical research project at the University of Helsinki started using a microcomputer to save and study materials. Project members examined the Finnish famine of the late 1860s (*1860-luvun suuret nälkävuodet*) based on the latest developments in social science history. In that project, they utilised either quantitative or qualitative methods (or both) on a variety of materials. For both types of method, they developed new best practices using software for building databases and for word processing, including one project-member, Kari Pitkänen, writing a concise guide book for fellow historians entitled *Historiantutkija ja mikrotietokone* (*The historian and the microcomputer*, 1987).[33]

Many preferred to wait and see, however. Several historians have confessed that they themselves hesitated and postponed adopting the novel PCs in the

mid-1980s, but by the beginning of the 1990s, nearly all had started to at least write with microcomputers.[34] A significant factor in this transition was the increased user-friendliness of PCs in the form of graphical user interfaces (in place of the command line interface). At the same time, PCs became cheaper and consequently more common. Soon after, the media started to excite people about a new information network: the internet. Considering the changes recently introduced by microcomputers, it is unsurprising that for many (older) historians the new online world of information networks remained for most of the 1990s quite distant.

Compared to older mainframe technology, microcomputers opened up a whole new spectrum of uses for historians to choose. Typing or text processing was by far the most widely adopted of these new uses and thus in many ways the most important one. But, in addition, on a PC one could also keep records and notes, and later draw maps and graphs, and take time to learn other new uses. Again, much of the development was gradual.[35]

Meanwhile, many other people were using microcomputers too. These included genealogists, who launched their own journal *Sukutietotekniikka* (*Computer technology for family research*) in 1984, and who worked together to insert data in digital formats, and later digitised parish registers and made them available online (HisKi). In some universities, linguists developed corpus linguistics and even historical linguistics. In the early 1980s, the Helsinki Corpus of English Texts was initiated. This ground-breaking digital text collection was completed and publicly distributed in 1991.[36] Quite a few historians became aware of these endeavours, but they remained distant to historical research.

Overseas, groups of historians established for themselves organisations such as the Association for History and Computing (AHC), which was proposed at a conference at the University of London in 1986. The AHC was dedicated to the use of computers in historical research and in 'promoting the use of computers in all types of historical study, both for teaching and research'.[37] Unlike their colleagues in many other countries, Finnish historians did not form a national association for history and computing, and to the best of our knowledge, they consequently took part to a very limited extent in this international discussion.

With every major change, quite a few historians at first postponed adopting the new technology. Who were these non-users of the (new) technology? Until well into the 1980s, they were those historians who were relying on textual analysis—basically, the majority of people in most history departments. They could use card files to make archival notes and to store their data, and other such manual or mechanical tools, and they used typewriters or perhaps had the department's typist transcribe their writings.

Gradually, for instance, cultural historians also switched their typewriters to PCs. Perhaps it took them a few more years, but it did happen, and soon, in the 1990s, it was only the most senior historians who did not change to writing on a computer, but hung on to the typewriter.

At the same time, Finnish researchers committed to the new cultural history avoided numbers and statistics, and in general quantitative methods. For instance, their colleagues in Italy and Germany more often used numbers and calculations to study microhistory. This avoidance can be regarded as a counter-reaction towards the general emphasis on quantitative methods such as statistical approaches in the 1970s. Instead, cultural historians studied textual evidence in the light of the then recent linguistic turn. Their emphasis was on using qualitative methods, especially 'close reading' of texts, as well as discussing and exploring narratives. Over time in the late 20th century, close reading became a leading (often the main or even only) method for legions of historians and other people studying texts, so much so that the literary historian and Professor Franco Moretti termed his new and different computer-assisted method 'distant reading'. Inspired by the Annales School of historians, he coined the term in 2000.[38] It has subsequently gained popularity as a response and complement to the dominance of close reading.

Enter the Internet: Anticipating a Digital Revolution?

In the early 1990s, the younger generation of historians discovered the internet, or networks of computers, that had been first built in the United States in the 1960s for military purposes and only came into wider, academic use by scientists during the 1980s. Furthermore, some historians soon took part in creating a new, virtual dimension to the world. In Finland, they first tested Gopher-based internet pages (before the html language) which were in use by 1994. At that time, the World Wide Web, or the Web, after being created at CERN, had begun its successful expansion as the information medium over the internet.

One of the early Finnish projects was the Electronic Centre for History Research in Finland. It first opened in late 1995.[39] The following year, it joined forces with other related projects, and these were transformed into a new national cooperation. Named as the Agricola network, this was a joint effort among historians in the universities, libraries and archives, and it was officially launched in 1996.[40]

The new Agricola site brought together people working with or interested in history, created new avenues of communication and enabled them to discuss their relevant issues in a very popular email list, H-verkko, nationally. They aimed to inform others and share news, as well as publish online. Importantly, one key component for the network builders consisted of educating historians and keeping them abreast of the internet's latest relevant developments. This included thinking ahead and writing about the possible futures of history research in the digital era: an anticipated digital revolution and what that might entail.[41] Further in connection to the Agricola network, a group of historians started to study IT history, especially in Finland, thus improving the shared understanding of living in a society in which computer technology was

gradually applied everywhere.[42] Out of the Agricola network's publishing activities grew *Ennen ja nyt* (*Then & Now*), in existence since May 2001, which was the first national, refereed online journal in history.[43]

To summarise, historians were now using computers and their networks for searching and gathering information, including data about archives, and they sometimes even accessed the actual sources that someone had downloaded to their pages. This could easily be achieved transnationally, and for quite some time it seemed national borders were becoming less and less important. The burgeoning virtual world and its sites first complemented and then slowly began to replace former foundations of historians' work such as library indexes, travel to archives and archive guides, followed by books, phone books, etc. In scholarly communications, electronic mail or email correspondence instead of postal letters proved triumphant in the 'internet age'.[44]

For the first time, historians were also becoming familiar with sources that were 'born digital', such as email letters and digital art, and discussed the future of electronic sources. Two extreme questions surrounded whether everything would be saved electronically (a burden for historians) or whether the new electronic sources (such as early www-pages) would be deleted or otherwise lost within a relatively short time, leaving future historians without important materials from the 1990s.[45] Thinking about it now, the latter seems closer to what has actually happened. Furthermore, the digital revolution that took place proved to be slower than expected and transformed into a digital evolution that eventually invaded every aspect of life during the 2000s and onwards.

In the 50-year period examined here, the contextual changes for historians have been significant, ranging from the expanding universities to the evolution of the Finnish society at large. The historical profession in Finland in the early 1960s consisted of perhaps fewer than 100 people active in conducting research. The number of history professors in Finland was 17 in 1960, and it grew to 32 in 1970 to approximately 46 in 2000 and to 10 more in 2015, while the number of research units (larger university departments) rose from five to eight in the same time period. However, the number of university-educated history researchers (PhD) and lower-level positions grew much more extensively, particularly from the late 1990s onwards. In addition to universities, there were historians carrying out research elsewhere, especially in a few major institutions such as archives and the National Library.[46]

Starting in the 1990s, the Finland-based multinational corporation Nokia, selling new mobile phones, led the country's high-tech investments and image, and Finland became a leader in many IT developments. This probably encouraged also technologically open-minded historians to explore the new possibilities that the novelties might offer. Meanwhile, especially since 2000, the profession has both specialised further and internationalised heavily, and historians have in general perhaps become less and less knowledgeable of their domestic colleagues compared with experts abroad. Historians in the universities have also confronted an ever heavier competition for (external) research

funding, which has contributed to their willingness to adopt new methods and ideas.

Digitising Sources and Offering Them Online

In many ways, digitisation of historical sources had its roots in microfilming similar materials. The state (national) archive in Finland started a project to microfilm documents in the late 1940s. It was the new general manager of the archive, Yrjö Nurmio, who led ground-breaking efforts to film important sources abroad, first in Sweden and West Germany, and thus made these archival collections that were considered relevant for Finnish historians easily available to researchers in Finland, on microfilm readers. Later in the 1950s and 1960s, Finns could also microfilm Soviet materials.[47]

During a longer period of time, a large collection of historical newspapers was microfilmed in Finland. Foreign newspaper collections could be purchased for use in Finnish libraries and universities. Microfilming and their use had then continued for about three decades when automatic data processing (ADP) started to become another option to store and access primary sources. While the history of microfilming might sound ancient and wholly irrelevant for historical researchers in the 2020s, this legacy is in fact a pertinent background to the digital newspaper collection.

The National Library at Helsinki had already established the Centre for Microfilming and Conservation in 1990, located in the small town of Mikkeli in Eastern Finland. They aimed to create a comprehensive microfilm collection of Finnish newspapers and journals. Meanwhile, the internet made its first breakthrough as a new and exciting channel to distribute information in digital formats in the early and mid-1990s.

Digitisation of cultural heritage began in Finland after the mid-1990s, with the Mikkeli centre playing a central role. From the perspective of newspaper collections, an essential turning point was the launching of the Nordic project Tiden in 1998. In the Finnish case, the digital collection of newspapers is for the most part based on microfilms, which means that both the quality of the microfilm and the quality of the original newspaper have an important impact on the accuracy of optical character recognition (OCR), which varies from decade to decade. After a busy few years, the National library was able to open the Historical Finnish newspaper archive online in 2001.[48]

The first collection of digitised newspapers already covered several decades of the 19th-century press. Historians could now carry out some of their historical research using digitised original materials, over the internet, via their own computers in their own offices.

Since its inauguration in 2001, this major online press archive has been constantly expanded and its user interface, such as search options, improved. These significant investments have made the National Library's DIGI Collection of

newspapers and periodicals published in Finland arguably the most used historical digital source material in 2018.[49] In fact this collection is so complete especially regarding the 19th-century newspapers that in many cases they are enough for answering the researcher's question/s. This has made some researchers critical and asking if not the research questions where chosen so that one is able to limit his/her study into consulting only the digital materials, relying on keyword searches, and applying the rather conventional qualitative methods.

Evolving Digital Humanities and Emerging 'Digital History'

Gradually, in the 2000s and the early 2010s, an increasing number of historians became aware of and familiar with the massive amount of digital texts from primary sources that were processed by memory institutions such as libraries and archives around the world into digital formats and made available online. In retrospect, suddenly, there was an abundance of material suitable for qualitative and quantitative analysis online. Anyone could perform simple yet comprehensive keyword searches in these vast collections. It was (and is) easy to forget that such searches might be anything but perfect (due to the low quality of OCR results) because the accuracy of the search process was very difficult to assess.

Most researchers rapidly realised that one could only perform 'close reading' on a tiny fraction of those online sources because even just skimming them all went beyond anyone's capabilities time-wise. This gradually led progressive historians to think about obtaining and/or creating more adequate, computer-assisted methods and the means to get the most out of this wealth of digital sources. Among these, one can count the above-mentioned literary historian Franco Moretti.

Meanwhile, computerised methods and software with a longer development history such as GIS came to be used by a few historians in Finland in the 2000s. They used GIS to place and study historical information on maps of various kinds. Compared to GIS, textual analysis with computational tools and the newly emerging 'big data' was still very much being invented and developed during the early 2000s. Nevertheless, researchers of AI had made important progress in cooperation with linguistics since the 1980s, and a research field called natural language processing (NLP) was advancing. Based on complex statistical mathematics and algorithms, this work promised new tools for analysing texts too. The first peer-reviewed journal article where the rather recent method of 'topic modelling' was applied for historical materials was published in 2006.[50]

In Finland, too, the early 2000s witnessed inventions in software turned into new digital tools that historians could use. For instance, in the late 1990s, a group of medievalists and the National Archives had built an electronic version of Finland's medieval sources (*medeltidsurkunder*), producing an online database called Diplomatarium Fennicum.[51] In the mid-2000s, Tuomas Heikkilä

joined forces with some IT specialists and together they started developing computational methods to group medieval texts. Their aim was to create a family tree, a stemma, based on the dis/similarity of those early scripts, in order to better study their origins as well as influences on each other.[52] Over the years, this new interdisciplinary cooperation has led to several international scholarly meetings called Studia Stemmatologica, as well as publications developing further stemmatological analyses.[53]

The availability of these digital materials combined with the introduction of new tools sparked many developments during the 2010s that are changing and will renew history research. Starting towards the middle of the decade, several national conferences and seminars have been organised to discuss such new research. The first two textbooks concerning historical research and digital methods were published in Swedish and Finnish, in 2014 and 2016, respectively.[54] In 2015, the major research funder for historians, the Academy of Finland, opened a call for projects to The Digital Humanities Academy Programme (2016–2019), which encouraged many to pay more attention to developments going on in this new research area. Some, but not quite all, of the outcomes of this wave of new research are presented in this book.

All this technological development and expectations for ever faster and wider analysis of the historians' 'big data' has also re-emphasised 'old' problems (stemming from the 1990s), such as the poor quality of OCR-processed digital texts. How can we overcome this obstacle to the use of these latest computational research methods? Challenges like this partly motivated some historians to plan the project Computational History and the Transformation of Public Discourse in Finland, 1640–1910, funded during 2016 to 2019, in which the low OCR accuracy in the digitised newspapers and periodicals was circumvented by basically using a method originally designed for bioinformatics—in this case, modified to recognise the reoccurrences of similar text passages systematically in several millions of pages of primary sources.[55]

These challenges are highlighting our need for developing novel ways of digital source criticism, but also for taking new, fresh perspectives on the digital evolution that surrounds us. An eye-opening example is offered by Johan Jarlbrink and Pelle Snickars, who studied the specific ways in which newspapers are transformed in the digitisation process, and concluded that in fact the massive digitisation has created large amounts of digital noise: 'that is millions of misinterpreted words generated by OCR, and millions of texts re-edited by the auto-segmentation tool', resulting in a new—and, moreover, unevenly distributed—layer being added to the shared cultural heritage.[56] This reinterpretation suggests and confirms, first, that we need to learn to live and come to terms with that digital noise and, second, that a totally new and so to speak born-digital (that is, generated by computer technology) demand for historians' tools in computer technology will be to reduce that digital noise.

Meanwhile, this emerging 'digital history' research has also been explored. In one inquiry, Finnish historians raised doubts about this new concept and/or identifying themselves with it. In other words, many responders expressed

uncertainty about whether or not they were digital historians and/or digital enough, meaning that, as of 2016, few historians saw themselves as digital historians.[57] Among the critical issues that were identified through the inquiry were the importance of creating better, up-to-date information channels of digital history resources and events, providing relevant education, skills and teaching by historians, and the need to help historians and IT specialists to meet and collaborate better and more systematically than before.

One can hypothesise that two camps of historians were formed in the late 20th century, distinguished by their use of computer technology. On the one hand, everybody was more or less taking advantage of text processing (working with text files and mainly writing), PCs in general and the internet, in various ways. On the other hand, there were those sub-fields that (had) also continued with quantitative methods, such as statistics, for a long time. But many historians concentrated mainly on text processing. It is important to note that the new methods of digital humanities, based among other things on developing NLP (technology), were more eagerly adopted, and embraced even, by those researchers who focused on processing texts. To be more precise, it was a fraction of those historians who embraced the latest methods and also appropriated the term 'digital history', while the social and economic historians adhered for a longer period to their seasoned ways in quantitative methods.

Further, these new ideas and the digital humanities scholarship have in Finland, as elsewhere, been brought together in new laboratories for humanistic research. By far the largest effort nationally in this field, the Helsinki Centre for Digital Humanities, or HELDIG, was established at the University of Helsinki in 2016. By 2020, HELDIG has evolved into a vibrant centre of teaching and research in digital humanities, including digital history. The centre's multidisciplinary research groups, led by Eero Hyvönen and Mikko Tolonen among others, have concentrated on semantic web and building linked open data portals, such as the Sampo series, intended also as historians' research tools, and on using large but overlooked collections of library metadata to quantitatively examine the evolution of book publishing and the press over hundreds of years, respectively. In addition, a group of Finnish historians has been actively involved in the association Digital Humanities in the Nordic Countries and its DHN conference series held annually since 2016. In 2018, HELDIG was one of the key organisers of the third DHN conference, this time arranged in Helsinki. The overarching theme of the conference was Open Science, which challenges current and future historians in yet other ways. Historians and other scholars involved in the field of digital humanities may expect all of this to further advance their digital research capabilities in the future.[58]

Conclusion

To better understand where the present digital and computational history has come from and its place in the historical discipline, this chapter has studied

the historians' use of computer technology, together with some associated technological influence in history research in the case of Finland. It is argued here that such an open and broad approach to these phenomena serves best to expose the complex and already quite extensive roots of the present-day digital history approaches.

Certainly, historical research has many layers of history with the digital, and this relationship continues to be formed in the mutual shaping of the research field, including its people and ways of doing things, technology and the society at large. Perhaps we can even say that the field of digital history today has not one but many histories, and its history remains open to a variety of interpretations.

On the one hand, it is difficult to exaggerate the changes that computer technology has brought to the work of historians (too) during the recent decades. Combined with other changes, the technological advances have positively enabled and enlarged historians' study options in unforeseen ways and scale, while they have also guided and reformed the research designs (see Table 2.1). On the other hand, it has been a long and circuitous route from computers being used for processing statistical data in the late 1960s (Viljo Rasila) and thereafter being used mostly by historians undertaking quantitative research, up until several technological advances and also disruptions (microcomputers, the internet and the World Wide Web, and related software), to the present day, where historians are able to perform their whole research process digitally, from planning to gathering materials, carrying out the analysis, including statistics (if any), writing their interpretation and then publishing the results online.

Nevertheless, it is evident that the use of IT was heavier in some sub-fields than in others, for many reasons. Those reasons range from theoretical underpinnings to copyright law, which has slowed both digitising and distributing certain primary sources from the 20th century.

From early on, divisions were created by different approaches to understanding history and consequently how the research was done. For a long time, starting from mainframe computers and the programs available on these, computer technology worked better for quantitative than qualitative research. That, in turn, might be one reason why the new 'digital history' was, albeit decades later, more eagerly welcomed by (some of the) historians analysing texts. This type of source had been the focus of their qualitative work for decades, and by the 2010s they needed new tools to handle the massive amounts of textual sources that organisations such as major libraries around the world had digitised and made available online during the last 15 to 20 years.

What remained the same during the 50 years in between was that the interpretations were made by the human mind of the historian. Unless perhaps those interpretations also changed while the technological environment and tools for making them were transformed? This is quite conceivable, which reminds us that we still know very little about the impact that computerisation has had on history as a field of study and its products from historical narratives to its theories of change and continuity. It is also time for the students of historiography and even philosophers of history to take a serious, deep look into the

Table 2.1: Milestones of computer use by Finnish historians.

1967	Two articles on using computers for scholarship: Viljo Rasila in *Historiallinen Aikakauskirja*; Pertti Huttunen on Roman social history.
1968	First monograph to use computer-aided statistical methods: Viljo Rasila, *Kansalaissodan sosiaalinen tausta* (*Social background of the civil war*).
1974	First two history PhDs using computerised methods: Pertti Huttunen and Reino Kero.
1970s	Computers in research projects: focused particularly on migration and mobility.
1983	First article (Lappalainen and Lappalainen) about PCs for historians' use.
1990	Centre for Microfilming and Conservation established in Mikkeli.
1996	The Electronic Centre for History Research in Finland (SHEK) for internet use and digitising sources begins (in the Mikkeli centre and elsewhere).
2001	Historical newspapers opened for research online and *Ennen ja Nyt* journal established online.
2014–2016	First two textbooks about digital history published in Finland.

Source: Author.

practical aspects of 'doing history',[59] where computer technology has become so central.

Whether embracing the new tools or shunning them, we should, however, remember what Melvin Kranzberg (a leading historian of technology) famously formulated as his first law. In our case, Kranzberg's rule, quoted as the epigraph to this chapter, means that we should take historians' thoughts and feelings about technology seriously. At times, they probably saw the computer technology as good, bad or both. More importantly, it reminds us that the computer has never been 'just a tool', and this is why we should collectively think more about using these changing products of IT developers and their bearing on our work.

Notes

[1] Kranzberg 1986: 545.
[2] Thomas 2004; see also Kahk 1984.
[3] Specifically, I have studied and observed the field of digital history from 2015 onwards in two research projects funded by the Kone Foundation.
[4] See also Kaiserfeld 1998; Jarlbrink 2015; Haapala, Jalava & Larsson 2017.
[5] See also Paju 2019. For the Anglo-American milestones, see Thomas 2004.
[6] Karonen 2019: esp. 19.
[7] Tirranen 1964: 225–234.

[8] Huttunen 1992: 21, 28. This book by Huttunen includes republished articles and the ones relevant here were originally written in the late 1960s.
[9] Historiallinen Yhdistys ry. 1966–1967. *Historiallinen Aikakauskirja* 1/1968, 89–90; Åberg 2010: passim.
[10] Rasila 1967: 145; Viljo Rasila, interview on 17 May 2016. The 'brick group' (*tiiliryhmä*) was a coordinated research effort focused on studying Roman brick stamps and led by Jaakko Suolahti, Professor of General History at the University of Helsinki. See Bruun 1992: 133–134.
[11] Huttunen 1967; Rasila 1967: 145. See also Rasila 1970.
[12] Waris 1969: 73–74.
[13] Järvinen 1969: 57–59; Rasila 1969a: 60–61; Pertti Järvinen, email letters, 26 October 2018.
[14] Bruun 1992: esp. 135–136.
[15] Huttunen 1974; Kero 1974.
[16] Viljo Rasila, email letter 21 March 2016.
[17] Tommila 1998: passim.
[18] Mauranen 1988.
[19] See Iggers 2012: 43–45 and passim; Hudson & Ishizu 2017: ch. 2.
[20] Paju 2008; Rainio 2013.
[21] See Heyck 2015.
[22] Viljo Rasila, interview on 17 May 2016.
[23] See, for instance, Strömberg 1998.
[24] Niitemaa 1971; Reino Kero, email letter 6 June 2016. ADP stood for automatic data processing.
[25] Lindberg & Sovio 1969: 134–142.
[26] Virrankoski 1982: 23–28, passim. On manual work behind the digital, see Jarlbrink, Chapter 7, this volume.
[27] Rasila 1982; Virrankoski 1982; Haapala 1986. See also Nygren, Foka & Buckland 2014.
[28] Tampere Research Group for History of population, environments and social structures.
[29] Rasila 1969b, 1970; Kahk 1973; Virrankoski 2013: passim. See also Kahk 1984; Paju 2019.
[30] See, e.g., *Historiallinen Aikakauskirja* 5/1916, 73; Kirschenbaum 2016.
[31] Neulakortit—jokamiehen reikäkorttijärjestelmä. (Kirjoittanut K.) *Tekniikan Maailma* 1/1955, 30; Lappalainen & Lappalainen 1983; Lappalainen, email letter 26 February 2016.
[32] Lappalainen 1985; Lappalainen, email letter 26 February 2016.
[33] Häkkinen et al. 1989.
[34] Virrankoski 2013: esp. 314. See also Paju 2016.
[35] Lappalainen, email letter 26 February 2016.
[36] Rissanen & Tyrkkö 2013.
[37] Denley & Hopkin 1987.
[38] See Hackler & Kirsten 2016: 6. See also Kiiskinen 2010; Salmi 2011.
[39] Onnela 1995.

[40] Kallio, Kari: Agricolasta Suomen historiaverkko Internetiin. *Digitoday*. Julkaistu 12.8.1996 15:41; Tapio Onnela, oral information 10 October 2018.
[41] See esp. Onnela 1998.
[42] See, for instance, Suominen 2000; Paju 2008.
[43] See Ennen ja Nyt 2001.
[44] See Paju 2016.
[45] Suominen & Sivula 2016: passim.
[46] Karonen 2019: esp. 19 and passim.
[47] Nurmio 1952; Nuorteva & Happonen 2016: passim. See also Jarlbrink 2015.
[48] Bremer-Laamanen 2006.
[49] See Kettunen, Pääkkonen & Koistinen 2016.
[50] Brauer & Fridlund 2013.
[51] See Diplomatarium Fennicum's history.
[52] Tuomas Heikkilä, interview on 15 August 2016. See Roos, Heikkilä & Myllymäki 2006.
[53] See Heikkilä & Roos 2016.
[54] Parland-von Essen & Nybergh 2014; Elo 2016. See also Guldi & Armitage 2015.
[55] See Vesanto et al. 2017.
[56] Jarlbrink & Snickars 2017.
[57] See Paju 2016.
[58] Hyvönen 2018; Matres, Oiva & Tolonen 2018; Tolonen et al. 2019. See also Mäkelä & Tolonen 2018.
[59] See also Paul 2011, who suggests the study of historians' 'doings'.

References

Interviews and correspondence (all by the author)

Pertti Järvinen, email letters, 26 October 2018.
Tuomas Heikkilä, interview, 15 August 2016.
Reino Kero, email letter, 6 June 2016.
Jussi T. Lappalainen, email letter, 26 February 2016.
Tapio Onnela, oral information, 10 October 2018.
Viljo Rasila, email letter, 21 March 2016; interview, 17 May 2016.

Journals

(Mostly from the National Library of Finland: https://digi.kansalliskirjasto.fi/aikakausi/)
Digitoday 1996
Historiallinen Aikakauskirja 1916, 1968
Tekniikan maailma 1955

Literature

Brauer, R., & **Fridlund, M.** (2013). Historicizing topic models: a distant reading of topic modeling texts within historical studies. In L. V. Nikiforova & N. V. Nikiforova (Eds.), *Cultural research in the context of digital humanities: proceedings of international conference 3–5 October 2013* (pp. 152–163). St. Petersburg.

Bremer-Laamanen, M. (2006). Connecting to the past—newspaper digitisation in the Nordic countries. *Journal of Digital Asset Management, 2*(3–4), 168–171.

Bruun, C. (1992). Nyt Rooman historia! *Historiallinen Aikakauskirja 90*(2), 129–143.

Denley, P., & **Hopkin, D.** (Eds.) (1987). *History and computing*. Manchester: Manchester University Press.

Diplomatarium Fennicum. *Tietoa hankkeesta*. Retrieved 3 October, 2018 from http://df.narc.fi/info/project

Elo, K. (Ed.) (2016). *Digitaalinen humanismi ja historiatieteet*. Turku: Turun historiallinen yhdistys.

Ennen ja Nyt. (2001). *Arkistot kuukauden mukaan: toukokuu 2001*. Retrieved 22 February, 2019 from http://www.ennenjanyt.net/2001/05/.

Guldi, J., & **Armitage, D.** (2015). *The history manifesto*. Cambridge: Cambridge University Press. 1st edn. 2014. Retrieved from http://historymanifesto.cambridge.org/

Haapala, P. (1986). *Tehtaan valossa: teollistuminen ja työväestön muodostuminen Tampereella 1820–1920*. Tampere and Helsinki: Osuuskunta Vastapaino and Suomen historiallinen seura.

Haapala, P., Jalava, M., & **Larsson, S.** (Eds.) (2017). *Making Nordic historiography: connections, tensions and methodology, 1850–1970*. New York, NY: Berghahn Books.

Hackler, R., & **Kirsten, G.** (2016). Distant reading, computational criticism, and social critique: an interview with Franco Moretti. *Le foucaldien, 2*(1), 7. DOI: http://doi.org/10.16995/lefou.22

Heikkilä, T., & **Roos, T.** (2016). Thematic section on Studia Stemmatologica. *Digital Scholarship in the Humanities, 31*(3), 520–522. DOI: https://doi.org/10.1093/llc/fqw038

Heyck, H. (2015). *Age of system: understanding the development of modern social science*. Baltimore, MD: Johns Hopkins University Press.

Hudson, P., & **Ishizu, M.** (2017). *History by numbers: an introduction to quantitative approaches*. 2nd edn. London: Bloomsbury.

Huttunen, P. (1967). Tietokoneet Rooman sosiaalihistorian tutkimuksessa (The computer in the study of Roman social history). *Studia Historica: Acta Societatis Historicae Ouluensis* (pp. 29–64). Tomus I. Oulu: Oulun historiaseura.

Huttunen, P. (1974). *The social strata in the imperial city of Rome: A quantitative study of the social representation in the epitaphs published in the Corpus inscriptionum Latinarum, volumen VI*. Oulu: University of Oulu.

Huttunen, P. (1992). *Työ—tekniikka—historian muutos: kirjoituksia työn ja tekniikan historiasta*. Oulu: Oulun historiaseura.

Hyvönen, E. (2018). *Semanttinen web: linkitetyn avoimen datan käsikirja* (Semantic web: handbook of linked open data). Helsinki: Gaudeamus.

Häkkinen, A., Ikonen, V., Pitkänen, K., & **Soikkanen, H.** (1989). *1860-luvun suuret nälkävuodet: tutkimus eri väestöryhmien mielialoista ja toimintamalleista. Loppuraportti*. Helsinki: Helsingin yliopisto.

Iggers, G. G. (2012). *Historiography in the twentieth century: from scientific objectivity to the postmodern challenge*. 2nd edn. Middletown, CT: Wesleyan University Press.

Jarlbrink, J., & **Snickars, P.** (2017). Cultural heritage as digital noise: nineteenth century newspapers in the digital archive. *Journal of Documentation, 73*(6), 122–1243.

Jarlbrink, J. (2015). Historievetenskapens mediehantering. In M. Hyvönen, P. Snickars & P. Vesterlund (Eds.). *Massmedieproblem: mediestudiets formering* (pp. 225–247). Lund: Mediehistoriskt arkiv 30, Lunds universitet.

Järvinen, P. (1969). Voidaanko historiaa tutkia tietokoneella? *Historiallinen Aikakauskirja 67*(1), 57–59.

Kahk, J. (1973). New possibilities of using computerized historical analysis in the study of peasant households. In *Turun Historiallinen Arkisto 28* (pp. 375–389). Turku: Turun Historiallinen Yhdistys.

Kahk, J. (1984). Quantitative historical research in Estonia: a case study in Soviet historiography. *Social Science History, 8*(2), 193–200.

Kaiserfeld, T. (1998) Historikerna och tekniken: om betydelse av tekniska hjälpmedel för historieforskningen. In M. Hedin & U. Larsson (Eds.). *Teknikens landskap: en teknikhistorisk antologi tillägnad Svante Lindqvist* (pp. 365–377). Stockholm: Atlantis.

Karonen, P. (2019) Historiantutkimuksen ja yhteiskunnan yli puolitoistavuosisatainen vuoropuhelu: resurssit, rakenteet ja tulokset. In P. Karonen (Ed.), *Tiede ja yhteiskunta: Suomen Historiallinen Seura ja historiantutkimus 1800-luvulta 2010-luvulle* (pp. 13–44). Helsinki: Suomalaisen Kirjallisuuden Seura.

Kero, R. (1974). *Migration from Finland to North America in the years between the United States Civil War and the First World War*. Turku: Institute for Migration.

Kettunen, K., Pääkkonen, T., & **Koistinen, M.** (2016). Kansalliskirjaston digitoitu historiallinen lehtiaineisto 1771–1910: sanatason laatu, kokoelmien käyttö ja laadun parantaminen. *Informaatiotutkimus 35*(3), 3–14.

Kiiskinen, H. (2010). Talous, käytänne ja kvantitatiivinen analyysi kulttuurihistoriallisesti suuntautuneessa tutkimuksessa. In H. Rantala &

S. Ollitervo (Eds.), *Kulttuurihistoriallinen katse* (pp. 80–97). Turku: k&h, Turun yliopisto.

Kirschenbaum, M. G. (2016). *Track changes: a literary history of word processing*. Cambridge, MA: Harvard University Press.

Kranzberg, M. (1986). Technology and history: 'Kranzberg's Laws'. *Technology and Culture, 27*(3), 544–560. DOI: https://doi.org/10.2307/3105385

Lappalainen, J. T. (1985). Tekstintekoa näytöllä. *Historiallinen Aikakauskirja*, *83*(4), 283–286.

Lappalainen, J. T., & **Lappalainen, V.** (1983). Historiantutkimusta ilman papereita. *Historiallinen Aikakauskirja, 81*(1), 75–78.

Lindberg. D., & **Sovio, P.** (1969). Katsaus Ruotsin tutkimusprojekteihin. *Historiallinen Aikakauskirja 67*(2), 134–142.

Matres, I., Oiva, M., & **Tolonen, M.** (2018). In between research cultures: the state of digital humanities in Finland. *Informaatiotutkimus, 37*(2), 37–61. DOI: https://doi.org/10.23978/inf.71160

Mauranen, T. (1988). Review of research in economic and social history in Finland in the 1970s and 1980s. *Scandinavian Economic History Review, 36*(3), 23–41. DOI: https://doi.org/10.1080/03585522.1988.10408125

Mäkelä, E., & **Tolonen, M.** (2018, 7–9 March) *DHN2018—an analysis of a digital humanities conference: Digital Humanities in the Nordic Countries 3rd Conference*. Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (pp. 1–9). Helsinki.

Niitemaa, V. (1971). Kaukosiirtolaishistorian tutkimusprojekti. *Historiallinen Aikakauskirja, 69*(2), 146–150.

Nuorteva, J., & **Happonen, P.** (2016). *Suomen Arkistolaitos 200 vuotta: Arkivverket i Finland 200 år*. Helsinki: Kansallisarkisto.

Nurmio, Y. (1952). Valtionarkiston toimesta ulkomailla suoritettavat mikrofilmaustyöt. *Historiallinen Aikakauskirja, 50*(4), 268–274.

Nygren, T., Foka, A., & **Buckland, P. I.** (2014). The status quo of digital humanities in Sweden: past, present and future of digital history. *H-Soz-Kult*. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-88730

Onnela, T. (1995). Email to H-verkko email list: 'Suomalaisen historiantutkimuksen elektroninen keskus' 14 November 1995. Retrieved 12 November, 2018 from http://historia.utu.fi/h-verkko/arkisto/0137.html

Onnela, T. (1998). Historiantutkimus internetin ja digitaalisen kumouksen aikakaudella. *Agricolan Tietosanomat* 5/1998. Retrieved 1 September, 2016 from http://agricola.utu.fi/julkaisut/tietosanomat-1998-2005/numero5-98/historiantutkimus.html

Paju, P. (2008). National projects and international users: Finland and early European computerization. *IEEE Annals of the History of Computing, 30*(4), 77–91.

Paju, P. (2016, 1 September). *Digitaalinen historiantutkimus: kyselytuloksia*. Report from the project Towards a Roadmap for Digital History in

Finland. Retrieved from https://digihistfinlandroadmapblog.wordpress.com/2016/09/01/raportti-kyselyvastauksista/

Paju, P. (2019, 6–8 March). International collaboration and Finland in the early years of computer-assisted history research: combining influences from Nordic and Soviet Baltic historians. In *Proceedings of the 4th Conference of the Association Digital Humanities in the Nordic Countries* (pp. 349–357). Copenhagen. CEUR Workshop Proceedings 2364.

Parland-von Essen, J., & **Nyberg, K.** (2014). *Historia i en digital värld*. Online book. Retrieved 10 September, 2018 from http://digihist.se/

Paul, H. (2011). Performing history: how historical scholarship is shaped by epistemic virtues. *History and Theory, 50*(1), 1=19. DOI: https://doi.org/10.1111/j.1468-2303.2011.00565.x

Rainio, K. (2013). Kuusitoista kilotavua—humanisti koodinmuuntimien ajassa. In *HY: Tietotekniikkapalvelut: 1960-luku*. Retrieved 28 June 2018 from http://www.helsinki.fi/atk/50v/1960-luku.html.

Rasila, V. (1967). Tietokone historiantutkimuksessa. *Historiallinen Aikakauskirja, 65*(2), 140–146.

Rasila, V. (1968). *Kansalaissodan sosiaalinen tausta*. Helsinki: Tammi.

Rasila, V. (1969a). Edellisen johdosta. *Historiallinen Aikakauskirja, 67*(1), 60–61.

Rasila, V. (1969b). The Finnish civil war and land lease problems. *Scandinavian Economic History Review, 17*(1), 115–135. DOI: https://doi.org/10.1080/03585522.1969.10407659

Rasila, V. (1970). The use of multivariable analysis in historical studies. *Economy and History, 13*(1), 24–53. DOI: https://doi.org/10.1080/00708852.1970.10418879

Rasila, V. (1982). Projektin hankkiman aineiston käyttömahdollisuudet. In *Muuttoliikkeiden ja sosiaalisen kehityksen väliset yhteydet Suomen teollistumisen alusta maan itsenäistymiseen* (pp. 42–60). Turku: Turun yliopiston historian laitos.

Rissanen, M., & **Tyrkkö, J.** (2013). The Helsinki corpus of English texts (HC). In A. Meurman-Solin & J. Tyrkkö (Eds.), *Principles and practices for the digital editing and annotation of diachronic data*. Studies in Variation, Contacts and Change in English 14. Helsinki: Varieng.

Roos, T., Heikkilä, T., & **Myllymäki, P.** (2006). A compression-based method for stemmatic analysis. In ECAI 2006: Proceedings of the 17th European Conference on Artificial Intelligence.

Salmi, H. (2011). Traditions of cultural history in Finland, 1900–2000. In J. Rogge (Ed.), *Cultural history in Europe: institutions—themes—perspectives* (pp. 45–62). Bielefelt: transcript Verlag.

Strömberg, J. (1998). Avhandlingarna. In P. Tommila (Ed.), *Historiantutkijan muotokuva* (pp. 55–80). Helsinki: Suomen historiallinen seura.

Suominen, J. (2000). *Sähköaivo sinuiksi, tietokone tutuksi: tietotekniikan kulttuurihistoriaa*. Jyväskylä: Jyväskylän yliopisto.

Suominen, J., & **Sivula, A.** (2016). Digisyntyisten ilmiöiden historiantutkimus. In K. Elo (Ed.), *Digitaalinen humanismi ja historiatieteet* (pp. 96–130). Turku: Turun Historiallinen Yhdistys.

Tampere Research Group for History of population, environments and social structures. Retrieved 19 December, 2018 from https://research.uta.fi/hopes-en/database/

Thomas, W. G. III (2004). Computing and the historical imagination. In S. Schreibman, R. Siemens & J. Unsworth (Eds.), *A companion to digital humanities* (pp. 56–68). Malden, MA: Blackwell.

Tirranen, H. (1964). Katsauksia: biologisia historianselityksiä. *Historiallinen Aikakauskirja 62*(3), 225–234.

Tolonen, M., Lahti, L., Roivainen, H., & **Marjanen, J.** (2019). A quantitative approach to book-printing in Sweden and Finland, 1640–1828. *Historical methods: a journal of quantitative and interdisciplinary history, 52*(1), 57–78. DOI: https://doi.org/10.1080/01615440.2018.1526657

Tommila, P. (Ed.) (1998). *Miten meistä tuli historian tohtoreita*. Helsinki: Suomen historiallinen seura.

Waris, H. (1969). Kansalaissodan taustatekijät. *Historiallinen Aikakauskirja, 67*(1), 73–74.

Vesanto, A., Nivala, A., Rantala, H., Salakoski, T., Salmi, H., & **Ginter, F.** (2017, 23–24 May). Applying BLAST to text reuse detection in Finnish newspapers and journals, 1771–1910. In Proceedings of the 21st Nordic Conference of Computational Linguistics. Gothenburg, Sweden (pp. 54–58). Linköping Electronic Conference Proceedings. Retrieved from http://www.ep.liu.se/ecp/133/010/ecp17133010.pdf

Virrankoski, P. (1982). Projektin historia. In *Muuttoliikkeiden ja sosiaalisen kehityksen väliset yhteydet suomen teollistumisen alusta maan itsenäistymiseen* (pp. 7–29). Turku: Turun yliopiston historian laitos.

Virrankoski, P. (2013). *Historian professori ja laulajapoika: kulttuurin kuvia suuren muutoksen ajalta*. Turku: Memnon-kirjat.

Åberg, V. (2010). *Lisää muistia! 50 vuotta tietotekniikkaa Helsingin yliopistossa*. Helsinki: Helsingin yliopiston tietotekniikkakeskus.

3

Big Data in Economic and Business History: Quantitative and Qualitative Methods

Jari Eloranta, Pasi Nevalainen and Jari Ojala

Prologue

An ambitious project was initiated in 2002 and concluded by 2007 by Finnish economic and business historians to analyse digitised news agency data in order to create a model to predict the behaviour of business enterprises. This project, entitled MetaSignal (later MetaAlert), was a joint venture between historians, journalism researchers, engineering scholars and economists working at the University of Jyväskylä and the Tampere University of Technology. The aim was nothing less ambitious than to create an artificial intelligence (AI) that could learn from the past to predict the future.

The AI was intended to compile automatically, categorise and analyse available online information to find so-called weak signals from a massive flow of information. To 'teach' the AI, the project used a massive news agency database, including roughly 20 million business newsfeeds from the early 1970s to the early 2000s. For the first time in Finnish historical research, the project also used digitised full-text *New York Times* newspaper data from the 1850s onwards, together with databases containing information about listed companies and stock market prices over an extended period of time.

Needless to say, this bold initiative failed as the project did not have sufficient human resources or computational power circa 15 years ago to reach its goals. Nevertheless, as for the outcomes, the project did identify publications and networks that were valuable at the time and at least interesting even from today's perspective. One must bear in mind that the internet was still a newcomer at the turn of the millennium; thus, there were still many uncertainties as to which direction it would develop and which would be the most usable tools to find information on various topics. Moreover, the databases on emerging markets of the internet were also at a developing stage, and so was the price of information: the price of annual use of the databases used in the project roughly doubled every year.

By analysing the data available from open sources at the time and comparing it to the data purchased from the databases in the market, the project found, for example, that the very origins of the contemporary newsfeeds could be traced to few, well-established and old news agency firms or media companies.[1] It was not until the emergence of the digital camera, smartphones and social media when the supremacy of these companies began to collapse, at least to a certain extent.

The project members did not necessarily even notice at the time how fast the environment around them was changing. The project participants travelled to Stanford to learn about the latest trends in Silicon Valley and report their findings to the steering group. Therefore, it was the historians and the other humanists who were the first to inform the others in the meetings with the funding agency Tekes (Finnish Institute of Technology and Innovation, nowadays known as Business Finland) about interesting emerging companies in the United States, like Facebook.[2]

. . .

The MetaSignal project was just an outcome of a long tradition of compiling and using massive databases, distant reading methods and, most importantly, sophisticated methods among economic and business historians to analyse numerical and textual data. The use of a massive database to predict future trends in the MetaSignal project was not, obviously, a ground-breaking idea. On the contrary, computerised methods have been used in social sciences in this respect at least from the early 1960s, when the first attempts were made at the RAND laboratories.[3]

Economic and business historians have been the forerunners in the digital history data gathering and analysis for decades. This chapter attempts to discuss the major developments internationally and, in some specific cases, in Finland in the fields of digital economic and business history, concentrating on some of our own projects, as well as research outcomes by economic and business historians at the University of Jyväskylä and within our networks. We are not claiming that our projects are unique or ahead of their time in the field of economic and business history—on the contrary. However, we feel that these

projects are indeed illustrative cases (such as the aforementioned MetaSignal) about the possibilities and challenges facing historians in the digital era.

After a section introducing the use of digitised data in economic and business history, we will briefly discuss the methodological challenges in the use of these methods, followed by sections concentrating on event data analysis and challenges involved in using various databases (with some examples). Thereafter, we focus our narrative on the use of digital sources and methods in business history. In the concluding discussion, we will address the challenges and opportunities offered by digitised sources, followed by some exposition of the remaining challenges.

Big Data in Economic and Business History

Big data is at the heart of economic history research, and has already been so for decades.[4] *Big structures, large processes, huge comparisons*, by Charles Tilly, a famous historical sociologist, was a book published in the mid-1980s that highlighted some of the early efforts in such scholarship. Tilly's classic studies urged researchers to study the macro-level societal structures systematically, to better understand large processes of change.[5] Tilly was also one of the forerunners of 'social science history', pushing sociological understanding to advance historical research. Economic historians were also part of this process and, to a certain extent, the first ones to explore and exploit the possibilities of social scientific methods and data in historical research.

Since the time of publication of Tilly's book, the datasets compiled and used by economic historians have become larger and more varied: numeric data is nowadays more often 'born digital'; and besides numbers, even economic historians are today more often using high-resolution digital images and digitised texts. The quantity of available data has increased dramatically, whereas the costs of storage have decreased—even though there is now a new challenge for academia arising from the costs of the best datasets and digitised library collections.[6] As Guttman and colleagues (2018: 269) note: 'A key characteristic of modern "big data" is that the volume of stored data exceeds human analytic capacity and pushes against the boundaries of currently-available computing power. For that reason, the magnitude of "big" is continually growing.'

By its principles, economic history research does not differ substantively from other types of historical research: economic historians compile data from original (archival) sources to provide answers to questions posed by scholars. What differs, though, is that the questions asked are often based on testable theoretical frameworks originating from social sciences and usually require a massive amount of data that, in turn, cannot be analysed without sorting the data into a database format, as well as by using some sort of quantitative methods. However, economic historians were forced to compile these types of datasets themselves for decades, whereas today there is a large amount of readymade data available,

starting with various text corpuses (for example, digitised newspapers), statistical data provided by different national and international authorities (such as census records) and databases compiled by researchers, authorities and private enthusiasts in different fields, including genealogical associations. The latter type of 'citizen participation' or 'citizen science' to compile data will most likely increase in the future, as well as different kinds of official, linked register data. Nevertheless, even today, researchers studying especially the ancient and early modern eras are forced to mainly compile the datasets by themselves, whereas those concentrating on the more contemporary periods and topics have to face the challenges associated with the already existing datasets.

Using digitised sources is at the very core of international economic history. Computerised methods were embedded into the economic history research during the 'Cliometric Revolution' in the 1960s and 1970s, when the so-called 'historical economics' tradition emerged first in the United States, then also later in Europe. The first researchers in this tradition were mostly trained as economists—such as Alfred Conrad and John Meyer, and then Robert Fogel and Douglass C. North—using their theories, models and econometric methods to study and understand controversial topics in history, like the productivity and profitability of slavery. Obviously, mainstream historians were not totally convinced about their studies and methods, especially as some of the advocates of the 'new economic history' took historians head on vis-à-vis many big topics.[7] By the turn of the millennium, this battle had settled down, as more historians have adopted cliometric methods to be a part of their toolkit and as 'social science history' has become more common. Simultaneously, economists are taking history research more seriously. Nevertheless, the major journals in economic history today are more oriented towards economics than they were back in the 1950s.[8]

The most obvious outcomes of the 'new economic history' have been the historical growth studies in different countries, compiled together in the Maddison Project database maintained at the Groningen University.[9] Historical national account series and other long-run societal and economic time series form a basis for all comparative macroeconomic studies of history. These include data on population, prices, wages, structure of the economy (size of agriculture, industry and services), foreign and domestic trade, urbanisation, central (and local) government expenditures and, finally, GDP (per capita) that is based on all the other data series listed above. Historical national accounts have made comprehensive comparisons over long periods of time more credible between a growing number of countries. These datasets have been game changers in the field and have occupied a substantial role in the debates over long-run economic growth. Angus Maddison (2001) published his initial global growth figures spanning 2,000 years at the turn of the millennium, but he had already started putting these numbers out in various publications from the 1980s onwards. Obviously, his early figures were rather tentative, and the GDP per capita estimates in general for many developing states were too low. Recent

efforts, for example, by Stephen Broadberry[10] and others, have exposed some of the flaws in these figures and extended our knowledge of not just European and Western development patterns, but also economic performance in Asia and Africa. These figures are now changing the debate over global trade and the so-called Great Divergence; that is, when and how China fell behind the West in the last 500 years.[11] In recent studies, the focus has shifted to account for new areas of interest, such as well-being and inequality.[12] Consequently, the existing Finnish historical national accounts from 1860 onwards were compiled by Riitta Hjerppe and the growth studies research group in the 1970s and 1980s, comprising 13 volumes in total, and they are still the benchmark in the study of Finnish economic history.[13]

Business historians, in turn, have been more focused on actors and related activities in the economy, whether by private persons, entrepreneurs, business enterprises or other groups. These actors represent the 'visible hand' of the aggregate economic system. Research on these actors, in turn, helps us to understand the evolution of economic structures. By looking at the American 19th-century railway companies, Alfred Chandler Jr. (1977) created the basic framework for the business strategy research. The methods used by modern business historians are more often qualitative, and the quantitative methods used are typically more descriptive than statistical ones.[14] Nevertheless, big data and methods used to analyse digitised databases have become more important also for business historians. This is simply due to the fact that either the data produced by entrepreneurs and enterprises over time are in most cases in numerical form and/or the volume of data is massive.[15] Even the early modern businessmen such as 13th-century *Commenda* traders in Genoa or late-18th-century Finnish businessmen produced a massive amount of letters and ledgers; some of those have lately been converted into a digitised format. The recent historical business data is already of digital origin. The shift to increasingly digitised material has enabled researchers to utilise larger quantities of material in qualitative research in future studies, including new ways to collect and analyse the material, including the use of AI in data mining and analysis.

Use of Quantitative and Qualitative Methods to Tackle Digital Sources

The use and analysis of quantitative data has been a hallmark of economic history research, especially since the turn towards more quantitative economic history, as we have already discussed. The aim of this more economics-influenced research has often been to attempt to find causal relationships between different phenomena; namely, to measure what were the factors explaining changes in phenomena proxied by various time series, cross-sectional or panel data. For example, during the past decades, there have been many attempts to compile data on, better measure and understand the dynamics of pre-industrial

economies; for instance, to clarify the role of women, children and families in the pre- and early-industrialising societies.[16] Alongside the time series (or panels) of economic development, much attention has been placed on the study of equal or unequal distribution arising from this development.[17]

From the 1950s onwards, econometric tools such as regression analysis have emerged as a typical way of estimating the relationships between economic variables. Regression analysis is today a common tool both in economics and social sciences, and also in economic history. Thus, in order to understand what has been written in the field during the past decades, one has to be familiar with at least the basics of this method; or, rather, the set of regression and other econometric techniques for modelling and analysing several variables. More commonly, regression analysis estimates the conditional expectation of the dependent variable vis-à-vis a set of independent variables; for example, what was the importance of education, investments or policy indicators for the economic growth or, as we have done, the effect of new technologies for wages of different skill levels of employees.[18]

Certain aspects of regression analysis have also been criticised, such as the over-reliance on measures of statistical significance.[19] Historians are particularly worried how such methods are suited to the analysis of time series as the observable and unobservable factors might change over time, and also the sources of data are similarly subject to change. Some of the research has become perhaps even overly technical by nature, thus losing its relevance for broader historical narratives.[20] Finally, causal relationships are hard to pinpoint, especially from more qualitative data,[21] and in econometrics the very idea that causality could be ascertained from regression analysis has become quite contested.

Another way to analyse causal relationships is by using counterfactual modelling: namely, to analyse a scenario of 'what if' the phenomena had not have occurred or a different historical trajectory had taken place. Economic history also has a long tradition of counterfactual analysis, starting from the early writings of Nobel-prize-winning economic historian Robert Fogel. Those models have, however, been criticised time and again by historians.[22]

Event Data Analysis

Although the methods used by economic historians could and should be criticised for certain shortcomings, they are nevertheless something that other historians might wish to emulate when using digitised, 'big data' sources. These methods can also be used when analysing qualitative, textual datasets, by introducing 'binary thinking' to the analysis; that is, coding the textual data to enable quantitative analysis. We have used, for example, 'event data analysis' to code actions and activities found in historical data, like the 'strategic actions' of companies. The basis for event data analysis can be found in historical events

that are arranged according their sequences. The coding of events (for example, strategic actions) enables comparing different actors, such as companies or business groups.[23]

While reducing texts to ones and zeroes might lead to over-simplifications, the use of more open methods, such as fuzzy set Qualitative Comparative Analysis (fs/QCA), has proved to be suitable for historical inquiries, as the set-theoretic relations frequently reveal more plausible causal relations than simple correlations.[24] Moreover, these types of methods can also be used to extrapolate larger datasets from smaller samples, in which typically statistical analysis has been near impossible. Often, the dichotomy between small-N qualitative case studies and large-N statistical studies has been overstated.[25] Essentially, they follow the same underlying logic of research. The best way to avoid the pitfalls of each is to engage in both or combine the strengths of each approach. These types of methods have been further developed by some Finnish business history and management scholars in particular.[26]

In international comparisons, comparable data, contexts and how the data helps make broader points about processes all play a role. For Finnish historians, though, even the question of the relevance of comparisons might sometimes alter the way in which we think about the sources and data. One of our own examples is from some years ago when we were using a large-N database which comprised information on Finnish and Swedish sailors. Thus, an obvious perspective for us was to compare these two countries in our analyses. For readers outside Scandinavia, however, this did not make much sense: the reviewers and editors of journals saw Sweden and Finland rather as complementary than interesting comparative cases in terms of our research question, and the paper was rejected time and again, before we fully realised this challenge and changed the paper accordingly.[27]

This type of categorisation is something we have tried to develop further also in our bibliometric work focusing on analysing trends in business history scholarships. As categories of the contents of journal articles in the ready-made databases (such as WoS or Scopus) are always subjective, we introduced certain measures to make such categorisations more objective in our study. Obviously, these are again methods used previously in other fields, but ones that can also be adapted to the study of economic and business history debates. For example, we engaged several researchers to do categorisations of previously published business history articles simultaneously, and then either used 'consensus' or average categorisations, or results of 'voting'. In the latter case, the 'votes' (zeros or ones by each individual doing the categorisation) for each category were summed up, and thereafter these sums of votes were calculated as a percentage of the maximum possible number of votes. These percentages were then taken to be the share of each category and as basis for further statistical analysis, namely to study why certain business history articles received the most citations.[28] The next obvious step is to introduce these bibliometric techniques to book-format publications, which would help us gauge the trends

in a publication format that historians prefer, again broadening the analysis of interdisciplinary transference.

Making Big Data Work: Databases and Their Challenges

As we have shown here, economic and business historians have been engaged in creating their own databases for a long time by using a variety of primary sources.[29] The data collected from the original primary source material has typically been stored as digital images, Word and Excel files on the researchers' own computers, and perhaps distributed via email or cloud services, when sharing was needed, for example, to make a common writing project easier. That is the case even today in many instances. Regardless, currently there is a growing number of ready-made databases that have to a certain extent eased the work of economic and business historians, yet at the same time they have provided new types of challenges. First of all, the availability of these databases has motivated researchers to study topics for which the data is (easily) available, and to find connections between those variables for which we have information. To study Finnish economic and business history, it might be challenging to use some of the international datasets, as information on Finland might be lacking, or is otherwise irrelevant or even incorrect. Some of the most important international databases, however, do have some data for Finland as well, like the Maddison Project database described above; Clio-Infra (http://www.clio-infra.eu/), EH-net databases (http://eh.net/databases/), Global Price and Income History (http://gpih.ucdavis.edu/) and Swedish historical monetary statistics 1668–2008 (http://www.riksbank.se/research/historicalstatistics).

The challenge is, however, that in many of these datasets the data on Finland is to a certain degree confusing and even misleading. This, in turn, relates to the fact that the data has been compiled from national statistical sources or from previous research. In the Finnish case, we simply still lack some of the basic research; thus, the datasets are using the existing figures for Finland. The Maddison database, for example, uses the growth figures for Finnish GDP (per capita), for certain benchmark years, for the last 2,000 years by using inter- and extrapolation methods. Nonetheless, Finnish growth studies have produced more exact figures so far only from the 1860s onwards. Currently, though, there is project at the University of Jyväskylä to fill the gap from the 1500s to mid-1850s in order to have more reliable, internationally comparable time series for Finland as well.[30] This will, hopefully, make Finland more appealing as a unit to be used in international comparisons: currently, Finland is lacking from a number of international studies simply due to the fact that comparable data does not exist yet.

Some of the international databases have been especially valuable also for Finnish economic and business historians. Beside those noted above, two specific datasets recently used by Finnish scholars are worth noting: the Soundtoll Registers Online (STRO) compilation (soundtoll.nl) and the Swedish Seamen's House enrolment database.

The STRO compilation is a good example of how digitised, large databases can be constructed with reasonable costs and in a limited amount of time.[31] The STRO database is based on the archival data created in the Danish Elsinore in the Sound Toll that was established in the late 15th century and lasted until 1857. The STRO database includes roughly all the ships and their cargoes that passed the Danish Sound from 1634 to 1856, comprising 1.4 million ships. Of these ships, roughly 2.4%, that is 35,000 ships, came from or headed towards Finland. In order to understand Finnish international trade and shipping, the STRO is especially important as the Danish Sound was the only route for Finnish export and import trade to markets beyond the Baltic for centuries. The Baltic trade as a whole, in turn, was of utmost importance in understanding the early modern and modern growth of Europe, as this trade was, as Milja van Tielhof puts it, 'the mother of all trades'.[32] The Danish Sound data used in previous research[33] was mainly based on the Sound Toll Tables compilation by Nina Ellinger Bang and Knud Korst in the 1920s and 1930s.[34] Their data, though, covered only the period up until the early 1780s, and later Hans Christian Johansen extended the period up until the mid-1790s.[35] Thus, from the Finnish perspective, the STRO is fascinating as it covers the era from the late 18th century until the mid-19th century, which was in many respects an emerging era for Finnish export trade and shipping.

Nevertheless, although the STRO data is highly valuable for research in general and for Finnish history research in particular, it also entails many challenges that can at the moment only be partly solved in the online dataset. The names of places and commodities are currently being made uniform, as well as the different units used (weights, sizes, etc.), and, moreover, there are a number of mistakes in the dataset that might have been present already when the entries of the original customs data were made or later during the data-entry process of the database. At the moment, there is an extensive project in Leipzig being overseen by Dr. Werner Scheltjens to modify the data further; this version, STRO 2.0, will be launched in the coming years. Finnish economic historians are also collaborating closely with this work in order to have even better data to use to study Finnish long-term trade patterns.[36]

Another important database used by Finnish economic historians is the Swedish Seamen's House enrolment dataset. This database was compiled at the turn of the millennium by the Swedish National Archives in collaboration with the Swedish Genealogical Association. The database includes roughly 650,000 enrolment cases and 26 million data points from nine Swedish coastal towns and one Finnish town (Kokkola). Researchers at the University of Jyväskylä gained full access to the database more than 10 years ago, only to find out that there were many challenges with the data. Indeed, the database is a good, or bad, example of the challenges inherent in these types of databases.

First, the researchers did not have full access to the data in the beginning, which made quantitative analysis impossible. Many similar genealogical databases have been designed to help users find detailed information on, for example, their ancestors—not to perform statistical analyses. Second, the data did not need to be exact in terms of values and figures to serve genealogical inquiries

and, therefore, in the datasheets sometimes numeric and textual data became mixed. This all meant that it took almost a decade for the researchers first to clean up the data, enrich it with additional information, and then standardise the monetary and other units (especially tonnage measures of ships) before it could really be used. This led to the third challenge that, again, is unfortunately rather common in many research projects using ready-made digitised databases. Namely, the database used in the research is to some degree different from the one that is available online at the Swedish National Archives website, and the researchers cannot, in accordance with the signed contract, publish the data they are using. Thus, hopefully in the future, the Swedish National Archives will publish the modified dataset separately on their website; this would be helpful for the research community at large, as this database is certainly highly valuable and the results have already been published in some of the most notable publishing forums.[37] There is already an initial agreement between the project researchers and the archive to publish the data in one form or another.

Digitising Business History

The magnitude of 'big' is also continually growing in the field of business history.[38] In practice, qualitative researchers can utilise much larger volumes and types of data than before and, on the other hand, different tools of analysis. The major development trend of recent decades is the diversification of the research field. Although the mainstream debate is still focused on businesses, entrepreneurs and entrepreneurship, the perspectives of research have widened over the last decades to cover a broad range of business-related themes. For example, the importance of interest groups, entrepreneurship of women and minorities, developing economies and environmental issues as part of business practices have emerged as major topics of discussion.[39] Even though most of the research is still being carried out in corporate archives, relying largely on textual material such as minutes and memos, it is because of the broadening of the scope of inquiry that the source material is quite sparse.

Finland's strength in business history research has traditionally been a comprehensive and open public archival service, which has guaranteed access to first-class material. One of the most important of these institutions is the National Archives (*Kansallisarkisto*), which has provided access to, among other things, abundant government documents, but also many archive collections of individuals and some private organisations. In Finland, the state has a strong position in society, and state documents, for example, contain not only information on legislation and administration, but also a huge variety of useful reports produced by various government organisations. The availability of sources has been supported by legislation under which a public authority document is in principle public.[40] Moreover, this also covers state-owned

enterprises, depending on their legal form of action. Such archives also cover very interesting research sites that are difficult to access in many other countries. The archives of the state-owned telecom company (PTL Tele/Sonera) are available to researchers until the year 1994, when the company changed its legal form from a public authority into a limited liability company. An even more important archive for Finnish business historical research is the Central Archives for Finnish Business Records (*Elinkeinoelämän Keskusarkisto*), where the archives of many Finnish companies are currently located and easily accessible to scholars. Often, such archives require a licence to use, which typically does not form an obstacle to academic research. For example, a large number of private telecoms documents are available up until the 2010s.

Despite the quality of the archive service, access to archival material and its quality are still key issues. For a private company, handing over the archive to the archival establishment is voluntary. The quality and usability vary on a case-by-case basis. At worst, even the material of important companies has been virtually lost. For example, when a large company, whose older archive sources are conveniently located in the National Archives, was asked about its late-1990s archives, it became clear that the company had outsourced the management of these archives to a private archive management company, which in turn had transferred the material to its own repositories. Worst of all, there was not even a list available for that material. On the other hand, the private archive management company does not provide any 'extra services' without an extra charge. Hence, to even find out whether the archives are relevant for scientific use would require a laborious and costly preliminary inquiry. On the other hand, some companies have already digitised their archives. However, even if the material is in digital format, there is no guarantee that it will be accompanied by a proper search engine and metadata, or that the archive would be properly organised and/or that the researcher would have full access to the database.

Business history has a long tradition of using digital images and optical character recognition (OCR) techniques, similar to economic history. In this way, scholars themselves have digitised a considerable amount of material. These have already greatly accelerated the utilisation of broader amounts of information. These are mostly private, rather limited databases. When talking about the possibilities of these images and personal collections, it should be borne in mind that these are not usually complete sets. A business history scholar, rarely paid for their efforts in this regard, usually has to photograph only the 'necessary' documents. For this reason, these private collections usually serve specific research questions. It is clear that large-scale digitisation of the material should be done by archives or large, well-funded projects in a professional and systematic way, leading to a publication of the data in a commonly used form. Unfortunately, the digitisation projects of aforementioned key Finnish institutions are still only in their infancy. Digitisation has, first and foremost, captured the oldest material. On the other hand, new machine reading technologies are

promising and will surely improve the usability of data in the future. Up until this point, very positive developments have taken place vis-à-vis search engines, making it easier for the researcher to find material from traditional archives.

Discussion of the business history method has touched upon the usability of the history research method in social sciences (such as organisational research), and how business historians can contribute to these discussions. Qualitative history research that takes into account temporal processes, contexts and coincidences has also been seen to be instrumental in building and modifying theoretical understanding.[41]

However, defining the method of historical research has become a problem: instead of a clearly defined method and source series, qualitative history research often takes advantage of different perspectives and sets of sources that may change as the research process evolves. The problem arises because historians are not accustomed to describing these research processes with the precision that is customary in social sciences, which in turn has begun to take replicability seriously. This debate has highlighted the need for business historians to pay more attention to describing their methods.[42] This requirement can also be viewed against the development of digital analytical methods. Since the idea of such methods is to automate the work, this requires event data coding in different ways, which in turn requires precision as well as continuous justification of choices. In this way, methodological precision and connection to theoretical models will be a more central part of the historian's daily routine.

Digitisation of business history sources and methods allows not only the use of qualitative data in larger quantities, but also the more intensive research collaboration. A particularly interesting example of using digital methods in business history research pertains to the 'Digital History of Telco and Exchanges in Finland and Sweden' consortium.[43] The project includes researchers, social scientists and historians from Aalto University and several Swedish universities, including the Stockholm School of Economics. Moreover, one of the authors of this chapter has participated in this collaboration. At the heart of this 'DigHist' project is a database, which includes the digital business archives of four business enterprises. Two telecommunications companies and two exchanges from Sweden and Finland have been selected for the project. These archives have been digitised for their most relevant parts. The coded digitised material is shared between the members of the consortium. For example, the database contains key sections of the Finnish state-owned telecom company's (PTL Tele/Sonera) archives (95 digitised archive boxes). Some of these have been digitised from the collections of the Finnish National Archives, but the others have been digitised from the material held by the current company Telia. Consequently, in one project, we were able to perform searches on all of the 764 Executive Team meetings (including attachments) that took place between 1981 and 1998. The software used also allows for the indexing of material and linking different documents to each other. Materials related to an interesting event can be assembled into a set of materials that make it easy to view relevant documents

together. The sheer amount of data and the search functions make it possible to efficiently compile information on the desired topics.

At best, this type of working method enables quantitative exploitation of qualitative material and analysis. In addition, by working closely together, the project scholars have been able to develop a unique research design centred around collaboration across institutions and disciplines. Data availability, a common desktop and teamwork enable a highly effective and accurate research process combining different areas of expertise. Practical experience has shown that such a method also poses challenges. Finding information about a huge amount of data requires good knowledge of the case and the materials. To know what kind of potentially interesting things have happened in the company being researched (namely, the terms used in the company at different times), it is important that someone in the research group is knowledgeable about the subject and sources of the case. Again, easy and partially mechanical availability of the material may blind the scholar. Too narrow a focus on certain source series and 'relevant' documents may obscure the importance of the historical process and context, leaving the strengths of historical research untapped. In any case, such a way of working has proved to be a promising way of combining digital tools and theoretical knowledge with methods of historical research.[44]

Discussion and Further Challenges

Digitisation is part of the development of technology and society, and hence something that naturally enhances economic and business history research. Its direct impact is related to the available material, the amount and usability of which are greatly improved as digitisation proceeds further. In many cases, such as in business archives, digitisation could potentially proceed much faster, but such efforts have been hampered by the lack of funding and expertise. Although digitising research materials and methods does not bring anything other than more efficient tools for managing the research process, at its best it can also be used as a tool for speeding up and facilitating the development of methodology and science, as well as international collaborations.

For some, digitisation itself changes human and social sciences. At the heart of such 'Google of archives' thinking are massive increases in the amount of data and improvements in search functions. According to Berry and Fagerjord (2017), up until now, digitisation in human sciences can almost completely be understood as a mechanism for sorting availability and dissemination of material in large quantities. The discussion has dealt with technical issues that are considered to be part of the archive's or library's work. In fact, even the tools are not new in principle. As we have discussed here, databases and quantitative methods have been used for a long time, and even before PCs became available. Economic history has been a forerunner in this and the lessons learned by economic and business historians—both successes and failures—could and

should, we argue, be used also more broadly among other historians. As Berry and Fagerjord conclude, the actual contribution of digitisation has to 'move beyond the purely instrumental and mechanical automation of processing of humanities materials'.[45]

Excessive and straightforward trust in digitisation is methodologically problematic. Using, for example, keywords, the desired documents can be found quickly from a large cache of data, yet a poor choice of keywords can lead the scholar to miss key contributions. Moreover, context and other areas are easily missed. The same applies to quantitative research, in which the researcher still needs to understand what dimensions and weights meant at different times and contexts. In that case, the researcher may unknowingly twist the history of an event in a way to reinforce his or her own hypotheses.[46] In reality, the need for someone to know the empirical material thoroughly does not disappear with the new digital collections and large-N methods. This is also an important starting point in all historical studies using, for example, regressions analysis: you need to know the units you are analysing before you analyse them, and what information you might still be missing from your analysis.

Furthermore, the sheer amount of information is a methodological problem, because the researcher needs to separate the necessary pieces from a massive amount of data. The committee which explored options for developing Finnish state-owned businesses published a 154-page report in 1985. However, the same committee delivered into the National Archives material that takes up about two shelf metres. Most of these consisted of unorganised documents, which contained numerous versions of the same memoirs, meeting invitations and drafts.[47] If this material were to be digitised and searched for, in practice, the same document would appear among the results tens of times, but only a few documents would increase our knowledge of the subject itself. We had a similar challenge with the MetaSignal project: using the large database containing news agency newsfeeds actually delivered the same information in the worst cases dozens of times. On the one hand, we could use this information to show the 'hype' around various topics, but on the other, it hindered the possibilities of performing proper quantitative analysis.

The creation and use of large digital collections require collaboration between several state and private actors. In the Finnish case, a specific role is played by the Finnish National Archives, which is responsible for the official documents created by the different state authorities. The Central Archives for Finnish Business Records (*Elka*), in turn, is the most important institution vis-à-vis private business archives and collections. The official collections (and to a certain extent also private archives) can be divided into roughly three groups, each entailing some specific challenges in today's digital world.

The first group consist of the 'old' paper archives, the total volume of which today is roughly 220 shelf-kilometres at the Finnish National Archives. Only a small fraction of this 'old' archival data has been or will ever be digitised; today, there are already 85 million digitised pictures available at the National Archives.

The goal is to digitise 20 per cent from this old material; mainly archives from the 1920s onwards. Nevertheless, the bulk of this material will also remain in paper format in the future.

Second, a large amount of paper format official documents resides with different state authorities that have been created since the 1970s, during the era of bureaucratisation, which are to be moved to National Archives in the coming years. The volume of these documents is around 135 shelf-kilometres. As it would not be sensible to build new warehouses to house such archives, the material will presumably be digitised *en masse* and the originals will be destroyed.[i] On the whole, this will mean that future historians who are looking for official documentation from 1970 onwards have to contend with *only* digitised archives. To a certain degree, the same is occurring in the private sector as well.

The third challenge relates to the so-called born-digital documents. The new service acquire for born-digital material is to be launched in year 2021; moreover a pilot project for private archives was under construction in 2020. Whatever the archival solution will be, both regarding public and private documents, the format will be digital. Thus, future historians will definitely need versatile skills to be able to use the digitised data and archives effectively.

In general, historical sciences have been the pioneers in the utilisation of information technology since the 1970s. Since then, digitisation has inevitably progressed, but as we have noticed, often slowly and sporadically. However, the advancement of digitisation has embodied undeniable advantages. Economic history at large has been forging ahead of other fields of history in using big data, digitised sources and quantitative methods. By using the rhetoric embedded in the theoretical debates of the discipline itself, economic history might eventually lose its comparative advantage as other, larger fields of history are catching up in using these data and methods, which in itself would be a good turn for debates about big issues in history, such as trade, slavery, development, environment, conflicts and so on.

There is a plethora of other challenges ahead for the field of history too, as well as economic and business history in particular. First, what materials should be digitised and when? This reflects the priorities among the scholars and the institutions that produce and maintain such records. Often, those priorities are not the same, which can create friction among the stakeholders. It also concerns resources and technologies available to facilitate such processes. Second, who has access and to what? While many archival collections are open access, some are not. And most published articles and books are not open access, which limits their use among scholars who are not institutionally linked and especially those who are located outside Western academic institutions. The same, of course, applies to the first concern about what materials are digitised; namely, for example, are business and economic records from the developing world less accessible those from the West? Third, new methods are emerging to analyse both the data itself as well as research trends, including bibliometric and AI

methods. These methods can offer great insights, but they can also be used to direct funds towards the most 'popular' types of research at the expense of, at least in terms of perception, more marginal topics. This can foster groupthink and could be detrimental to smaller, interdisciplinary fields like economic and business history. Finally, there are great challenges among the various fields of history to remaster quantitative techniques to be able to make use of the new 'big data', given that the so-called cultural turn from the 1980s until the early 2000s had no real interest in quantitative analysis and that the 'Cliometric Revolution' often took economic historians to departments of economics. Now, there is a greater demand to bring back quantitative historians, who have the requisite skill set to work with these types of data and methods. However, to achieve that, humanities will have to compete with other fields, with higher wages and better resources, so this process will likely take some time.

. . .

As stated at the outset of this chapter, our MetaSignal project failed 15 years ago. Would it be possible to create such an AI with historical data to predict future today? A number of similar software solutions have already been created, using various kinds of data sources. However, with similar sources and algorithms that we were using, it is highly unlikely that the project would succeed even with today's computational power. Moreover, although digital methods in humanities and social sciences have developed significantly over the past 15 years, the use of these methods is still lacking behind the digitisation of sources. Nevertheless, it would certainly be beneficial to have historians on board to develop similar kinds of projects also in the future. AI methods are certainly already in use to deal with large datasets and analytical projects, and eventually they will become the cornerstones of historical analysis more broadly, although historians will have to exercise careful control over these efforts and remember the points of caution we have reflected on in this chapter.

Notes

[1] Ojala 2005: 19.
[2] The early project outcomes are summarised in Ojala & Uskali 2005.
[3] See esp. Andersson 2012: 1411–1430.
[4] See, e.g., Calafat & Monnet 2017; Ojala 2017: 446–456.
[5] Tilly 1984. Tilly's research focused heavily on finding structural patterns in history in the long term, as evidenced in his classic study of European urbanisation and warfare (Tilly 1990).
[6] Gutmann et al. 2018: 270, 280.
[7] McCloskey 1978. See also Conrad & Meyer (1958). The debate about slavery came to a head over the book by Fogel & Engerman 1974, which was criticised by many, including Gutman 2003 and Sutch 1975. See

also Kolchin 1992 on critique of the follow-up book. For a review of the Cliometric Revolution and its achievements, see, e.g., Goldin 1995; Greif 1997; and Carlos 2010.

[8] Whaples 1991; Eloranta, Ojala & Valtonen 2010.

[9] See the Maddison Project database, version 2018, https://www.rug.nl/ggdc/historicaldevelopment/maddison/. On this database and its use, see Bolt & van Zanden 2014; Bolt et al. 2018.

[10] Broadberry & Gupta 2006; Broadberry, Custodis & Gupta 2014; Broadberry et al. 2015.

[11] See, e.g., Pomeranz 2009; de Vries 2010.

[12] See esp. van Zanden, et al. 2014.

[13] Hjerppe 1989.

[14] Eloranta, Ojala & Valtonen 2010; Ojala et al. 2017.

[15] For a broader discussion of data and methods in business history, see, e.g., Decker et al. 2015.

[16] de Vries 2008; Humphries & Weisdorf 2015.

[17] See, e.g., Hoffman et al. 2002; Milanovic, Lindert & Williamson 2010; Piketty 2015.

[18] See esp. Allen 2001; van Zanden 2009; Ojala, Pehkonen & Eloranta 2016.

[19] Ziliak & McCloskey 2016.

[20] See esp. Sala-i-Martin 1997; Reckendrees 2017: 3.

[21] Mahoney 2000; Ketokivi & Mantere 2010.

[22] See, e.g., Atack 2018.

[23] See esp. Lamberg & Ojala 2006: 22–25; Lamberg, Laurila & Nokelainen 2006: 307–312.

[24] For an introduction to this method, see Fiss 2011: 393–420.

[25] See, e.g., Mahoney & Goertz 2006; Jordan et al. 2011.

[26] See esp. Pajunen 2008: 652–669; Järvinen et al. 2009: 545–574.

[27] The final published article is Ojala, Pehkonen & Eloranta 2016.

[28] On the models, see, e.g., Ojala & Tenold 2013: 17–35; Ojala et al. 2017: 305–333.

[29] In the English case, we can go as far back as the 17th century; see Broadberry et al. 2013 for further discussion.

[30] See jyu.fi/growth.

[31] About the project, see: Gøbel 2010: 305–324; Veluwenkamp & Scheltjens 2018.

[32] van Tielhof 2002.

[33] See, e.g., Åström 1962; Åström 1963; Åström 1988.

[34] Bang & Korst 1930.

[35] Johansen 1983.

[36] See, e.g., Eloranta, Moreira & Karvonen 2015; Moreira et al. 2015; Ojala & Räihä 2017; Ojala et al. 2018.

[37] See esp. Ojala, Pehkonen & Eloranta 2016.

[38] Cf. Gutmann et al. 2018.

[39] See, e.g., Amatori & Jones 2003; Scranton & Fridenson 2013.
[40] The Freedom of Information Act (621/1999).
[41] Kipping & Üsdiken 2014; Üsdiken & Kipping 2014; Wadhwani & Bucheli 2014; Decker 2017.
[42] See, e.g., Yates 2014; Decker, Kipping & Wadhwani 2015; de Jong, Higgins & van Driel 2015; Stutz 2019.
[43] See https://blogs.aalto.fi/digihist/.
[44] One of the papers resulting from this project (Cheung, Aalto & Nevalainen 2019) was selected as one of the best research method papers at the Academy of International Business conference in 2019.
[45] Berry & Fagerjord 2017: 14.
[46] See, e.g., the discussion of the problems involved with a study conducted by Timothy Leunig and Hans-Joachim Voth on smallpox deaths, concerning both the sources they used and the methods involved. See Vervaeke & Devos 2018.
[47] See the Committee Report 1985:2 (*Valtion liikelaitoskomitean mietintö*); committee archives in the Finnish National Archives.

References

Allen, R. C. (2001). The great divergence in European wages and prices from the Middle Ages to the First World War. *Explorations in Economic History, 38*(4), 411–447.

Amatori, F., & **Jones, G.** (2003). Introduction. In F. Amatori & G. Jones (Eds.), *Business history around the world*. Cambridge: Cambridge University Press.

Andersson, J. (2012). The great future debate and the struggle for the world. *The American Historical Review, 117*(5), 1411–1430.

Atack, J. (2018). Railroads. *Handbook of Cliometrics*, 1–29.

Bang, N. E., & **Korst, K. E. P.** (1930). *Tabeller over skibsfart og varetransport gennem Øresund 1661–1783, og gennem Storebælt 1701–1748: udarbejdede efter de bevarede regnskaber over Øresundstolden og Bælttolden; udgivne paa bekostning af en international indsamling*. Copenhagen: Gyldendalske Boghandel.

Berry, D. M., & **Fagerjord, A.** (2017). *Digital humanities*. Cambridge: Polity Press.

Bolt, J., & **van Zanden, J. L.** (2014). The Maddison Project: collaborative research on historical national accounts. *Economic History Review, 67*(3), 627–651.

Bolt, J., Inklaar, R., de Jong, H., & **van Zanden, J. L.** (2018). Rebasing 'Maddison': new income comparisons and the shape of long-run economic development. Maddison Project Working Paper, No. 10. Retrieved from http://www.ggdc.net/maddison

Broadberry, S., & **Gupta, B.** (2006). The early modern great divergence: wages, prices and economic development in Europe and Asia, 1500–1800. *Economic History Review, 59*(1), 2–31.

Broadberry, S., Campbell, B. M., & **van Leeuwen, B.** (2013). When did Britain industrialise? The sectoral distribution of the labour force and labour productivity in Britain, 1381–1851. *Explorations in Economic History, 50*(1), 16–27.

Broadberry, S., Campbell, B. M., Klein, A., Overton, M., & **van Leeuwen, B.** (2015). *British economic growth 1270–1870*. Cambridge: Cambridge University Press.

Broadberry, S., Custodis, J., & **Gupta, B.** (2015). India and the great divergence: an Anglo-Indian comparison of GDP per capita, 1600–1871. *Explorations in Economic History, 56*, 58–75.

Calafat, G., & **Monnet, É.** (2017). The return of economic history? Books & Ideas.net. Retrieved from http://www.booksandideas.net/IMG/pdf/2017-30-01-economic-history.pdf

Carlos, A. M. (2010). Reflection on reflections: review essay on reflections on the cliometric revolution: conversations with economic historians. *Cliometrica, 4*(1), 97–111.

Chandler Jr., A. D. (1977). *The visible hand: the managerial revolution in American business*. Cambridge, MA: Harvard University Press.

Cheung, Z., Aalto, E., & **Nevalainen, P.** (2019). *Changing criteria for internal legitimacy and the internationalization process of a state-owned enterprise*. Paper presented at AIB conference, Copenhagen.

Conrad, A. H., & **Meyer, J. R.** (1958). The economics of slavery in the ante bellum South. *Journal of Political Economy, 66*(2), 95–130.

de Jong, A., Higgins, D. M., & **van Driel, H.** (2015). Towards a new business history? *Business History, 57*(1), 5–29.

De Vries, J. (2008). *The industrious revolution: consumer behavior and the household economy, 1650 to the present*. Cambridge: Cambridge University Press.

De Vries, P. (2010). The California School and beyond: how to study the Great Divergence? *History Compass, 8*(7), 730–751.

Decker, S. (2017). Paradigms lost: history and organization studies. *Management and Organizational History, 11*(4), 364–379.

Decker, S., Kipping, M., & **Wadhwani, R. D.** (2015). New business histories! Plurality in business history research methods. *Business History, 57*(1), 30–40.

Eloranta, J., Moreira, M. C., & **Karvonen, L.** (2015). Between conflicts and commerce: the impact of institutions and wars of Swedish–Portuguese trade, 1686–1815. *Journal of European Economic History, 44*, 9–50.

Eloranta, J., Ojala, J., & **Valtonen, H.** (2010). Quantitative methods in business history: an impossible equation? *Management & Organizational History, 5*(1), 79–107.

Fiss, P. C. (2011). Building better causal theories: a fuzzy set approach to typologies in organization research. *Academy of Management Journal, 54*(2), 393–420.

Fogel, R., & **Engerman, S.** (1974). *Time on the cross*. Boston, MA: Little, Brown & Co.

Gøbel, E. (2010). The sound toll registers online project, 1497–1857. *International Journal of Maritime History, 22*(2), 305–324.

Goldin, C. (1995). Cliometrics and the Nobel. *Journal of Economic Perspectives, 9*(2), 191–208.

Greif, A. (1997). Cliometrics after 40 years. *American Economic Review, 87*(2), 400–403.

Gutman, H. G. (2003). *Slavery and the numbers game: a critique of time on the cross*, Vol. 82. Champaign, IL: University of Illinois Press.

Gutmann, M. P., Klancher Merchant, E., & **Roberts, E.** (2018). 'Big data' in economic history. *Journal of Economic History, 78*(1), 268–299.

Hjerppe, R. (1989). *The Finnish economy 1860–1985: growth and structural change*. Helsinki: Bank of Finland.

Hoffman, P. T., Jacks, D. S., Levin, P. A., & **Lindert, P. H.** (2002). Real inequality in Europe since 1500. *Journal of Economic History, 62*(2), 322–355.

Humphries, J., & **Weisdorf, J.** (2015). The wages of women in England, 1260–1850. *Journal of Economic History, 75*(2), 405–447.

Järvinen, J., Lamberg, J. A., Murmann, J. P., & **Ojala, J.** (2009). Alternative paths to competitive advantage: a fuzzy-set analysis of the origins of large firms. *Industry and innovation*, *16*(6), 545–574.

Johansen, H. C. (1983). *Shipping and trade between the Baltic area and Western Europe 1784–1795*. Odense: Odense University Press.

Jordan, E., Gross, M. E., Javernick-Will, A. N., & **Garvin, M. J.** (2011). Use and misuse of qualitative comparative analysis. *Construction Management and Economics, 29*, 1159–1173.

Ketokivi, M., & **Mantere, S.** (2010). Two strategies for inductive reasoning in organizational research. *Academy of Management Review, 35*(2), 315–333.

Kipping, M., & **Üsdiken, B.** (2014). History in organization and management theory: more than meets the eye. *Academy of Management Annals, 8*(1), 535–588.

Kolchin, P. (1992). More time on the cross? An evaluation of Robert William Fogel's without consent or contract. *Journal of Southern History, 58*(3), 491–502.

Lamberg, J.-A., & **Ojala, J.** (2006). Evolution of competitive strategies in global forestry industries: introduction. In J.-A. Lamberg, J. Näsi, J. Ojala & P. Sajasalo (Eds.), *The evolution of competitive strategies in global forestry industries: comparative perspectives*. World forests Vol. IV (pp. 22–25). Dordrecht: Springer.

Lamberg, J.-A., Laurila, J., & **Nokelainen, T.** (2006). Competitive activities of forestry industry firms: a coding manual for event history analysis. In J.-A. Lamberg, J. Näsi, J. Ojala & P. Sajasalo (Eds.), *The evolution of competitive strategies in global forestry industries: comparative perspectives*. World forests Vol. IV (pp. 307–312). Dordrecht: Springer.

Mahoney, J., & **Goertz, G.** (2006). A tale of two cultures: contrasting quantitative and qualitative research. *Political analysis, 14*(3), 227–249.

Mahoney, J. (2000). Strategies of causal inference in small-N analysis. *Sociological methods & research, 28*(4), 387–424.

McCloskey, D. N. (1978). The achievements of the cliometric school. *Journal of Economic History, 38*(1), 13–28.

Milanovic, B., Lindert, P. H., & **Williamson, J. G.** (2010). Pre-industrial inequality. *Economic Journal, 121*(551), 255–272.

Moreira, C., Eloranta, J., Ojala, J., & **Karvonen, L.** (2015). Early modern trade flows between smaller states: the Portuguese–Swedish trade in the eighteenth century as an example. *Revue de l'OFCE, 140*, 87–109.

Ojala, J., Karvonen, L., Moreira, M. C., & **Eloranta, J.** (2018). Trade between Sweden and Portugal in the eighteenth century: assessing the reliability of STRO compared to Swedish and Portuguese sources. In J. W. Veluwenkamp & W. Scheltjens (Eds.), *Early modern shipping and trade: novel approaches using sound toll registers online* (pp. 151–173). Leiden and Boston, MA: Brill.

Ojala, J., & **Räihä, A.** (2017). Navigation acts and the integration of North Baltic shipping in the early nineteenth century. International Journal of Maritime History, 29**(**1**),** 26–43.

Ojala, J., & **Tenold, S.** (2013). What is maritime history? A content and contributor analysis of the *International Journal of Maritime History*, 1989–2012. *International Journal of Maritime History, 25*(2), 17–35.

Ojala, J., & **Uskali, T.** (2005). *Mediajättien aika: uusia heikkoja signaaleja etsimässä*. Helsinki: Inforviestintä.

Ojala, J. (2005), Mediaimperiumien heikot signaalit. In J. Ojala & T. Uskali (Eds.), *Mediajättien aika: uusia heikkoja signaaleja etsimässä*. Helsinki: Inforviestintä.

Ojala, J. (2017). Taloushistorian paluu ja liiketoimintahistorian nousu. *Historiallinen Aikakauskirja, 115*(4), 446–456.

Ojala, J., Eloranta, J., Ojala, A., & **Valtonen, H.** (2017). Let the best story win—evaluation of the most cited business history articles. *Management & Organizational History, 12*(4), 305–333.

Ojala, J., Pehkonen, J., & **Eloranta, J.** (2016). Deskilling and decline in skill premiums during the age of sail: Swedish and Finnish seamen, 1751–1913. *Explorations in Economic History, 61*(July), 85–94.

Pajunen, K. (2008). Institutions and inflows of foreign direct investment: a fuzzy-set analysis. *Journal of International Business Studies, 39*(4), 652–669.

Piketty, T. (2015). About capital in the twenty-first century. *American Economic Review, 105*(5), 48–53.

Pomeranz, K. (2009). *The great divergence: China, Europe, and the making of the modern world economy*. Princeton, NJ: Princeton University Press.

Reckendrees, A. (2017). Economic history in times of transition. *Scandinavian Economic History Review, 65*(1), 1–5.

Sala-i-Martin, X. X. (1997). *I just ran four million regressions* (No. w6252). Cambridge, MA: National Bureau of Economic Research.

Scranton, P., & **Fridenson, P.** (2013). *Reimaging business history*. Baltimore, MD: Johns Hopkins University Press.

Sutch, R. (1975). The treatment received by American slaves: a critical review of the evidence presented in Time on the Cross. *Explorations in Economic History, 12*(4), 335–438.

Stutz, C. (2019). History and organizational theorizing blended: Insights from exploring the corporate social responsibility field (Doctoral dissertation). Jyväskylä: Department of history and ethnology, University of Jyväskylä. Retrieved from http://urn.fi/URN:ISBN:978-951-39-7981-2

Tilly, C. (1984). *Big structures, large processes, huge comparisons*. New York, NY: Russell Sage Foundation.

Tilly, C. (1990). *Coercion, capital, and European states, AD 990–1990*. Cambridge, MA: Basil Blackwell.

Üsdiken, B., & **Kipping, M.** (2014). History and organization studies: a long-term view. In M. Bucheli & R. D. Wadhwani (Eds.), *Organizations in time* (pp. 33–55). Oxford: Oxford University Press.

Van Tielhof, M. (2002). *The 'mother of all trades': the Baltic grain trade in Amsterdam from the late sixteenth to the early nineteenth century*. Leiden: Brill.

Van Zanden, J. L. (2009). The skill premium and the 'Great Divergence'. *European Review of Economic History, 13*(1), 121–153.

Van Zanden, J. L., Baten, J., Mira d'Ercole, M., Rijpma, A., Smith, C., & **Timmer, M.** (Eds.) (2014). *How was life? Global well-being since 1820*. Paris: OECD Publishing.

Veluwenkamp, J. W., & **Scheltjens, W.** (Eds.) (2018). *Early modern shipping and trade: novel approach using sound toll registers online*. Leiden: Brill.

Vervaeke, A., & **Devos, I.** (2018). Much ado about nothing? Reconsidering the smallpox effect: height in the nineteenth-century town of Thielt, Belgium. *Tijdschrift voor Sociale en Economische Geschiedenis, 14*(4), 56–83.

Wadhwani, R. D., & **Bucheli, M.** (2014). The future of the past in management and organization studies. In M. Bucheli & R. D. Wadhwani (Eds.), *Organizations in time* (pp. 3–30). Oxford: Oxford University Press.

Whaples, R. (1991). A quantitative history of the *Journal of Economic History* and the cliometric revolution. *Journal of Economic History, 51*(2), 289–301.

Yates, J. (2014). Understanding historical methods in organization studies. In M. Bucheli & R. D. Wadhwani (Eds.), *Organizations in time* (pp. 265–283). Oxford: Oxford University Press.

Ziliak, S. T., & **McCloskey, D. N.** (2016). Lady Justice versus cult of statistical significance. In G. E. DeMartino & D. N. McCloskey (Eds.), *The Oxford Handbook of Professional Economic Ethics* (pp. 352–365). New York, NY: Oxford University Press.

Åström, S. E. (1962). *From Stockholm to St. Petersburg*. Helsinki: Suomen Historiallinen Seura.

Åström, S. E. (1963). *From cloth to iron: the Anglo–Baltic trade in the late seventeenth century*. Helsinki: Societas Scientiarum Fennica.

Åström, S. E. (1988). *From tar to timber: studies in Northeast European forest exploitation and foreign trade, 1660–1860*. Helsinki: Societas Scientiarum Fennica.

The Modern Paradigms of Explanation: A Comprehensive Study of Digital History 1.5

Mats Fridlund

History is one of the oldest and most conservative humanist disciplines, which begs the question how it could react to the current third 'generation' or 'wave' of digital history and its new potential to transform the practice of historians' research. History as a discipline is according to some digital historians at a crossroads, 'in a transitory moment'[1] and 'standing on the edge of a conceptual precipice'. The 'understanding and practice of traditional history' has been said to be 'facing a fundamental "paradigm shift"' and 'straddling a line between revolution and continuity' and that the resolution of 'this tension is going to be a central part of historians' tasks over the coming years'.[2] Some historians claim that 'digital history has become the buzz-word for avant-garde historical scholarship in the digital age',[3] while others worry about external interests and pressures from funders, governments and industrial stakeholders and the possibilities of reallocation of resources and 'fear for the hermeneutic character of the humanities, and a reduction of humanities research to data crunching or to a view that proclaims the search for underlying patterns and structures in human history and culture to be its essence'.[4] The overall concern is that history will be transformed into a new primarily quantitatively focused discipline

where traditional 'analogue history' focused on narrative and close and deep reading of primary sources will be marginalised.

This chapter want to take these hopes and fears of a paradigm shift in history seriously and I will use my training as a historian and theorist of modern science and technology to analyse and conceptualise what such a paradigmatic change of historical science might mean. To do this, I will discuss what I have elsewhere identified to be the main methodological strands of computational digital history and in this use research from history and philosophy of science on revolutionary and paradigmatic change within science, and especially Thomas Kuhn's historical and philosophical research on scientific revolutions.[5] In doing this, I have made the choice to, rather than provide an empirical case study of the practices of current digital historians, combine a description of some of the current practices within historical research with a larger conceptualisation of what I and other digital historians have identified as some of the central methodological elements of the new digital history.

The reason for this is that I consider it to be crucial for current and future digital historians to analyse and think reflectively about their new emergent historical practices. We need empirical descriptions of current historical practice, but we need critical reflections and conceptualisations even more. As a conceptually minded historian, it is crucial for me to have conceptual tools that helps us better see and better understand. In this, I am inspired by Joseph Schumpeter's statement on the foundation of historical analysis:

> Analytic effort starts when we have conceived our vision of the set of phenomena that caught our interest, no matter whether this set lies in virgin soil or in land that had been cultivated before. The first task is to verbalize the vision or to conceptualize it in such a way that its elements take their places, with names attached to them that facilitate recognition and manipulation, in a more or less orderly schema or picture.[6]

Thus, the central task of this chapter is to attempt to conceptualise and attach names to some of the central elements of the new emerging digital history practices so that we can start our analytic efforts to better understand the new emerging digital history.

Paradigmatic Change in Sciences, History of Science and Historical Sciences

There are especially two main areas of Thomas Kuhn's research on scientific revolutions that are of relevance to understanding the current changes within digital history. The first is Kuhn's research on what he described as 'the second Scientific Revolution' of the 19th century and on the historical impact of quantification of earlier qualitative research fields. Quantification, Kuhn argued, was central for understanding the historical development of scientific research and,

in 1961, in an article published just before *Structure of scientific revolutions* and at the same time as the historical sciences were entering their first quantitative 'Cliometric Revolution', Kuhn investigated 'the effects of introducing quantitative methods into sciences that had previously proceeded without major assistance from them'.[7] Kuhn starts his article describing how the Social Science Research Building at the University of Chicago on its facade

> bears Lord Kelvin's famous dictum: 'If you cannot measure, your knowledge is meager and unsatisfactory.' Would that statement be there if it had been written, not by a physicist, but by a sociologist, political scientist, or economist? Or again, would terms like 'meter reading' and 'yardstick' recur so frequently in contemporary discussions of epistemology and scientific method were it not for the prestige of modern physical science and the fact that measurement so obviously bulks large in its research?[8]

In his article Kuhn studies how the physical sciences achieved this exemplary and aspirational character for other sciences to follow, something which still is very much with us in the current debate on digital humanities and digital history. The reason for physics' status as the contemporary model science, Kuhn posited, could be understood as coming from that

> physicists, as a group, have displayed since about 1840 a greater ability to concentrate their attention on a few key areas of research than have their colleagues in less completely quantified fields. In the same period, if I am right, physicists would prove to have been more successful than most other scientists in decreasing the length of controversies about scientific theories and in increasing the strength of the consensus that emerged from such controversies. In short, I believe that the nineteenth-century mathematization of physical science produced vastly refined professional criteria for problem selection and that it simultaneously very much increased the effectiveness of professional verification procedures.[9]

And the reason for this in its turn came from how the physical sciences "came to make use of quantitative techniques at all'.[10] Perhaps surprisingly to some, then and now, the physical sciences had not always been based on measurements and mathematics. Some parts of physics, what Kuhn described as the 'traditional sciences' in the form of astronomy, optics and mechanics, had developed considerably quantitatively before the first scientific revolution, while the relatively new 'Baconian sciences', 'the study of heat, of electricity, of magnetism, and of chemistry', had not been a systematic field of inquiry previously, but 'owed their status *as sciences* to the seventeenth century's characteristic insistence upon experimentation and upon the compilation of natural histories, including histories of the crafts'.[11] Their quantification and a wider

and more thorough mathematisation of physics overall took place during the first half of the 19th century and was accompanied by a number of new instruments, conceptualisations, theories and institutionalisations, which was part of what Kuhn described as a second scientific revolution of the sciences. The larger question in focus of this chapter is whether the historical sciences is currently in such a Kuhnian moment.

The second relevant area of Kuhn's research is his more widely known general theory of scientific change that was first presented in *Structure of scientific revolutions* (1962) and that he continued to revise and refine for the remainder of his career. Kuhn's theory uses the history of scientific development especially during the first scientific revolution from the 15th to the 17th centuries to design a theory that outlines how a traditional or 'normal science' through a scientific revolution transforms into a new science, a radically different paradigm of knowledge practice. In this perspective, the response of a scientific community to 'crisis' in the form of a major epistemological disruption usually follows either of two main paths, what can be described as the reintegration and domestication of the new disruption as part of the existing framework of traditional 'normal' science, or the revolutionary transformation of the traditional science into a new science.

Kuhn's theory of scientific revolutions has been important in not just helping historians of science conceptualise changes within the natural sciences, but also in helping historians in general to better understand change within their respective domains. It is difficult to exactly translate Kuhn's terminology to other areas and as I. Bernard Cohen points out, there are many problems with using Kuhn, such as that 'historians and philosophers of science do not agree on what constitutes or defines a revolution in science; they do not have an objective test for the occurrence of such a revolution', and that 'there are certain kinds of revolutions in science that do not exactly fit Kuhn's schema'.[12] Nevertheless, despite these obstacles, several historians have used Kuhn's conceptualisations to understand change also within historical disciplines. As David Hollinger has pointed out, 'Kuhn's terms have been employed explicitly by historians of art, religion, political organization, social thought, and American foreign policy'.[13] Those historians also include Thomas Kuhn himself, as is clear from his remark on an upcoming academic discussion of Martin Bernal's 'Black Athena' theory of ancient history, when he stated that it 'was being held far too soon and that disciplines did not usually respond so quickly to fundamental challenges'.[14]

Aware of these problems, I use Kuhnian terminologies as ideal types (in a Weberian sense) to help me conceptualise the recent past, present and future developments within digital historical practice and to outline two major responses to the challenges of the new computational digital history, as well as sketch a possible methodological middle way navigating between the two. This is an extension of previous research of mine where I, as a part of an empirical digital history study, identified and outlined what I saw as the

major methodological strands within current digital history research. Following Kuhn, I have described the two main ideal type responses towards the new disruptive digital methodologies as them either being domesticated and naturalised as part of traditional history, what Kuhn would describe as 'normal science', which I have termed *digital history 1.0*, or taking the second more revolutionary route in the form of a paradigmatic *digital history 2.0*, radically transforming and disciplining the practice of historical research.[15] However, as an alternative to these two main routes of conservation or revolution, I also outline a potential third 'middle way' between the 'normal' practice of historical science and a potentially 'paradigmatic' digital history. The overarching question is whether the new digital historians will want to transform, and succeed in transforming, the historical discipline overall, to break off and form a new historical discipline, or whether they prefer to remain part of history's 'disciplinary mosaic'.[16]

Our Invisible Digital History

The digital has already changed historians' practice so that today 'all historians are already digital' whether or not they 'self-identify as digital historians',[17] although perhaps in ways invisible to or at least not reflected upon by most historians. History is already changed through historians' everyday use of digital tools and materials, something which can roughly be divided into the production, communication, presentation and administration of historical research.[18] The following description might to some appear trivial, banal or mundane, but that should not diminish its importance; on the contrary, this ordinariness makes it even more important for understanding the wider impact of the digital on the historians' craft.

The first and most important influence of digitisation is on historians' production. Like other office workers, the overwhelming majority of historians have since the 1980s been relying on digital computers as their foremost research tool. Most importantly, computers are used for writing and note-taking and since at least the 1990s also for organising and storing primary and secondary digital source materials, often in such portable digital document formats as photographed, scanned or born-digital images of archival documents, texts, photographs, artifacts, journal articles and books. The existence on most historians' computers of hundreds or thousands of files with names ending in suffixes such as .doc, .pdf, .xls and .jpg provides ample material evidence of the impact on historians' practice from reading, watching, manipulating and writing of digital materials.

Digitisation's second major impact is on how we historians communicate with institutions and individuals that provide access to source materials for our research, such as archives and libraries, as well as with other historians and non-historical researchers within our research fields. Since the 1990s, emails,

mobile and smart phones, text messaging and social media has afforded historians ever faster and wider communication possibilities. Third, the digital has impacted the historians' practice through making possible their research results to be communicated through new digital forms of representations. This is through presentations at academic conferences, seminars and talks, primarily through much easier and efficient use of digital images, figures and graphs, as well as the increased use of digital presentation software programs such as PowerPoint, Keynote and Prezi as well as online presentations and meeting using digital applications such as Skype and Zoom. In addition, preliminary and finished research is routinely presented in the form of digital documents to colleagues, conferences and publishers, as conference and seminar papers, manuscripts, preprints and offprints of articles, chapters and books. The final way in which historians' research practice has been impacted is with regard to its practical organisation and administration, through the various ways in which the digital tools and formats described above, together with the internet, have changed the possibilities for conducting research more effectively and (mostly if not always) with less costs in time and money. This includes all the ways in which we use the internet and especially search engines, such as Google, Bing and Baidu, to gather practical information about locations, access and opening hours for archives, libraries and museums, as well as conducting practical matters such as booking travel and buying books, source materials and artifacts through services such as Amazon, eBay and Alibaba, and registering and paying online for conferences or memberships in professional organisations.

This normal everyday digital impact on the historian's craft is most often invisible. The hidden digital tools and computational algorithms built into these various applications enabling our research are probably not much reflected upon by most historians, but these concealed tools have enhanced traditional history by making it faster, easier and cheaper in money as well as in time and energy. However, there are also other domesticated forms of digital methods and tools that in more conscious, reflective and visible ways have influenced historians' practice, something which I describe as digital history 1.0.

Domesticated Normal Science: Digital History 1.0

By conceptualising various aspects of historians' practice as 'digital history 1.0', I mean to accentuate that already today many historians, in addition to the invisible application of digital tools discussed above, have intentionally although often without much apparent thought appropriated digital methodologies as a part of their standard historical research practice. Digital history 1.0 includes how historians have integrated the use of digitally enhanced tools and materials as a part of their normal research practice, such as digital databases and resources such as Google, Wikipedia and JSTOR for digitally augmenting their historical research.[19] Such historians might, however, not see themselves as doing 'digital' history, but just 'history', as these digital applications have often

been domesticated and seamlessly incorporated into 'normal history'. This digital ignorance or blindness is a returning complaint of digital historians, with statements such that 'the average historian is at most a passive user of digitised sources in which he/she mostly sees a substitute for the material original' and 'carrying out fairly traditional research as if the [digital] resource was not there (but hopefully citing it nevertheless)'.[20]

In the vocabulary of the historian and philosopher of science Thomas Kuhn these historians have augmented their 'normal science'—'history'—of historical research with the use of various forms of digital sources, tools and methods. By normal science, Kuhn means the established and dominant scientific tradition of conducting research existing within a scientific discipline which 'often suppresses fundamental novelties because they are necessarily subversive of its basic commitments'.[21] The ignorant attitude among normal historians referred to above can be seen as exemplifying this. Another example is when one digital historian complains about traditional historians' blindness to how the digital has changed the historians' practice, how most historians today 'combine traditional/analogue and new/digital practices, at least in the information gathering stage of their research'. However, 'reflection is often missing. On more than one occasion I have heard historians proclaim to be non-digital, as if this were something of which to be proud, while evidently making use of digital resources in their research.'[22] Yet another digital historian describes 'a degree of condescension and suspicion towards digital resources from many mainstream historians'.[23] These examples could easily be multiplied.

And still, digital history 1.0 has already visibly changed historians' practice: first, by increasing the number of citations and the diversity of primary sources used, as well as a disproportionate use of citations to online sources.[24] One example is from Canada, the first country to have two of its major newspapers the *Toronto Star* and the *Globe and Mail* digitised in 2002. Research on history doctoral dissertations uploaded to the ProQuest database between 1997 and 2010 showed a 991% increase in citations to the *Toronto Star* after it had been digitised, 'as opposed to minor increases and even decreases for other newspapers'.[25] Connected to this, digitisation has also changed how historians think of their *archives*. Traditionally, for most historians, an emblem of becoming a 'real' historian and marking something of a rite of passage is to carry out research in a physical archive located in a particular (often remote) place where you sit and go through dusty and perhaps previously unread pages of primary sources in the form of paper documents such as letters, minutes, reports, etc. In the digital age, these traditional archives are often supplemented or surpassed by online document archives that you can access from your office chair at your home institution. But even when the historians do visit physical archives, their practice has been changed by the digital in that 'analytical work is displaced from the archives'. This is also due to new digital tools, as the

> use of digitized finding aids, digitized collections, and digital cameras [that] have altered the way that historians interact with primary sources.

> While the centrality of archives to the research process remains, the nature of interactions with archival materials has changed dramatically over time; for many researchers, activities in the archives have become more photographic and less analytical.[26]

By changing the possibilities of access to distanced primary materials, the new digital resources have transformed history.

One striking example of how the digital history practices can be transformative while almost methodologically invisible comes from the research by historians Sönke Neitzel and Harald Welzer on the politics and world view of German Second World War soldiers that was based on a previously unused source material in British and American archives in the form of several hundred thousand pages of transcripts of interrogations with German POWs. This groundbreaking in-depth research on this 'mind-boggling amount of material' was only made possible through the use of digital methodologies and was described in the following way in their monograph *Soldaten*: 'We were able to digitize all of the British documents and most of the American material and sort through it with the help of content-recognition software.'[27] This is all that is said. No further words on their digital research methodology such as what software, search methods or keywords that were used. The choices made and opportunities created by the digital tools have been made almost totally invisible.

It appears that Toni Weller is correct in stating that 'for most historians, the challenges of the digital age are not ones that are seen to directly concern their research' and that the suggestion by an author commenting on the tenure, promotion and review process 'that "learning to use a database, scan materials, and query that database all consume time that could be used to write" is probably a reasonably accurate reflection of the way the majority of historians perceive digital scholarship'.[28] However, there are those historians where the digital is a primary methodological focus in their research practice and who are practising a more radical form of 'digital history 2.0'.

Revolutionary Paradigmatic Science: Digital History 2.0

Some digital historians appear to see digitisation's 'profound transformation' of history as inevitable, in that they state that as 'datasets expand into the realm of the big, computational analysis ceases to be "nice to have" and becomes a simple requirement'.[29] This new paradigmatic digital history practice 'offers a stark contrast to what has become standard historical practice'.[30] The current revolutionary enthusiasm is in some ways reminiscent of digital history's first wave in the 1970s when 'it looked like history might move wholesale into quantitative histories, with the widespread application of math and statistics to the understanding of the past' and resonate with the past 'hyperbole that saw computational history as making more substantial "truth" claims, or the invocation

of a "scientific method" of history'.[31] The question is whether also the current putative computational revolution will live up to the high hopes and hypes or if it also will wane to become just another small specialised sub-discipline of the historical discipline or that it perhaps will abandon history and emigrate, like many of the first generation of digital historians who left the humanities for the social sciences and its new, more quantitatively inclined sub-disciplines, such as social and economic history.

The question is whether this new potentially revolutionary historical paradigm can be described, in Thomas Kuhn's words, as the outcome of a scientific revolution 'from which a new tradition of normal science can emerge'.[32] Kuhn described 'what all scientific revolutions are about' in that they

> produced a consequent shift in the problems available for scientific scrutiny and in the standards by which the profession determined what should count as an admissible problem or as a legitimate problem-solution. And each transformed the scientific imagination in ways that we shall ultimately need to describe as a transformation of the world within which scientific work was done.[33]

After a paradigm shift, it is not just what is valued as good research that has shifted, but the discipline's core elements are transformed and the field is reconstructed 'from new fundamentals, a reconstruction that changes some of the field's most elementary theoretical generalizations as well as many of its paradigm methods and applications'.[34] What is accomplished in this is the transformation of the 'disciplinary matrix'—what is considered as the relevant and central methods, significant data, instruments, theory, methods, concepts and working practices. Below, some of the major elements of the possible disciplinary matrix of digital history 2.0 will be outlined.

Digital history 2.0 is taken to represent research practices with a potential to form a new digital historical paradigm primarily focused on new quantitative and computational methods to undertake text analysis and manipulations and visualisations of historical data. Its research systematically use various digital applications and quantitative methodologies for big-data text and data mining, calculations and visualisations, such as topic modelling, network analysis and text and data scraping. Most of these methods necessitate investments in acquiring expertise in or collaborators skilled in coding and database methodologies.

Like with paradigm change within the sciences, the new digital history practice transforms the existing practice by introducing new focus and altering what is valued, making some of the existing ideals and standards less relevant or obsolete in favour of new values and concepts salient to the particular characteristics of the new history. One such new key aspect of the digital history 2.0 can be described as *compression*, which characterises methods that allow the historian 'to begin with the complex and winnow it down until a narrative emerges from the cacophony of evidence'.[35] This is in contrast to 'normal history'

where historians, 'like good detectives, test their merit through *expansion*: the ability to extract complex knowledge from the smallest crumbs of evidence, that history has left behind. By tracing the trail of these breadcrumbs, a historian might weave together a narrative of the past.'[36] Some historians even question whether the digital turn will so much change history's foundational concepts to 'render the word "narrative" too confining for describing what historians produce' and to make *historiographies* into a 'more encompassing term'.[37]

Normal historians prefer to describe the empirical foundations of their conclusions in terms of documents, sources and at times even 'facts', while the new digital historians often prefer to talk about 'data'. Jim Mussell describes perhaps *the* core aspect of the new digital history just in that it 'requires a change in focus from document to data'.[38] Data as information, in forms that are able to be processed by computers, is central to the new digital history, qualitatively as well as quantitatively. Its qualitative effect is the view favouring 'data' to signify what counts as the preferred and proper basis for constructing a historical argument. The quantitative impact lies in that the new digital texts provide copious and often very easily accessible source materials for historians. In 2008, a senior digital historian stated with special reference to the recently started digitisation efforts by Google Books, online digital image collections and the creation of digital newspaper archives that it was 'now quite clear that historians will have to grapple with abundance, not scarcity' and that 'nearly every day we are confronted with a new digital historical resource of almost unimaginable size'.[39] In that sense, history could be seen as having entered the era of *big data* or perhaps better 'biggish data'.[40] How much data it takes to makes it 'big' has been described as 'in the eye of the beholder', in that if 'there are more data than you could conceivably read yourself in a reasonable amount of time, or that require computational intervention to make sense of them, it's big enough!'[41] One example of such big data for historians are the online Old Bailey records (*www.oldbaileyonline.org*), which consist of almost 200,000 criminal trials between 1674 and 1913 and 127 million words.[42]

The rise of online archival research and the loss of the manual physical handling of original primary sources is one example of how the material practice of the historian is changing in the digital era. Another example of a radically new social dimension consists of *multidisciplinary teamwork*. This might be one of the most challenging aspects of the new history to many traditional historians. Although many examples exist of co-authored works in history, it is still far from the norm, and when it does occur it is rarely with collaborators from outside historical disciplines. Another changing practice is a shift to totally new activities in that 'less than 5% of the time spent on a project will be time spent analyzing and visualizing data', with the majority 'spent on collecting, cleaning, and interpreting'.[43] Another aspect of the changing historical practice is new digital forms of *publications* as the traditional paper forms of historical publications are not seen as 'suited to the fast-changing discourses of the digital age—demonstrated by the fact that most pure digital history texts tend to be in the form of websites, blogs and online articles and journals rather than the

traditional historical outlet of the monograph'.[44] Such new digital forms of publications also make possible new dynamic and interactive forms of presentations with inclusion of digital sound and video files, as well as scalable images, maps and network graphs.

To conclude this discussion of the changing practice of the new digital history practice, I will quote the two computer scientists behind the Culturomics project who also helped to develop the Google Ngram Viewer and when criticised for not having included any historians in their project explained it thus:

> Even when we found historians who shared our enthusiasm, there were still great barriers to working together. For instance, [a meeting was convened with] about a dozen interested history students and faculty. The historians who came to the meeting were intelligent, kind, and encouraging. But they didn't seem to have a good sense of how to wield quantitative data to answer questions, didn't have relevant computational skills, and didn't seem to have the time to dedicate to a big multi-author collaboration. It's not their fault: these things don't appear to be taught or encouraged in history departments right now.[45]

In short, history had failed in being willing to work like computer science.

Semi-Automatic History: Digital History 1.5

Some digital historians propose a less radical transformation than that promised by digital history 2.0, where 'historians do not need to learn new technologies or computer codes; they do not need to become computer scientists'. They disagree with those advocating a revolutionary transformation of the historical practice and argue that a part of 'the problem thus far has been too much emphasis on historians becoming something they are not; to the detriment of the fundamental skills and expertise that is the craft of the historian'.[46] The real challenge lies, such historians argue, 'in persuading the vast majority of historians of the benefit of even relatively simple information technology, not in developing specialist historical tools and methods that would only ever be of relevance to a minority of historians'.[47] Some like Gerben Zaagsma want to go somewhat further and consider that the 'real challenge is to be consciously hybrid and to integrate "traditional" and "digital" approaches in a new practice of doing history'.[48] 'Digital history 1.5' aligns itself with such views and can be described as an acknowledged and reflective digital history 'without the programming' that consist of the use of semi-automatic historical methodologies in between normalised 'digital history 1.0' and paradigmatic 'digital history 2.0' research methods.[49]

Digital history 1.5 is a hybrid or mixed methodology in that it is a combination of quantitative and qualitative historical research methodologies, and semi-automatic as it combines a large amount of manual evaluation with the

systematic use of automatic analysis vested in pre-programmed offline and online calculation and visualisation applications and tools using digital text and databases, such as Google Books, Early English Books Online (EEBO) and digitised historical newspaper archives. That this digital history is without programming is of course not absolutely true in that it does use digital applications based on a lot of computer code and many mathematical algorithms, but this coding and programming is invisible as it is pre-packaged in the various applications and tools: it is 'black-boxed' to the historian user.[50]

What differentiates digital history 1.5 from digital history 1.0 is that it consists of a systematic use of digital tools and sources where the digital methodology is the central method enabling the investigation. Furthermore, it incorporates a conscious reflexivity about the digital sources, resources and methods used in the investigation and is being reflective about its respective strengths and weaknesses. At the same time, it is not 'digital history 2.0' in that in its investigation it is using pre-programmed applications and resources without any additional coding of software, advanced programming of applications or tuning of digital techniques and methodologies. Some specific digital history 1.5 methodologies are semi-automatic text extraction and presentation, which combine quantitative computer-enabled 'distant reading' of big data digital text corpora and qualitative 'close reading' of extracted individual texts.[51] This takes the use of semi-automatically extracted and processed databases where the individual texts can be newspaper and journal articles that could be collected using various online search interfaces such as those that exist at various online newspaper and journal archives.

To conclude this treatment of the hybrid practice of digital history 1.5, two of its central methodological elements will be conceptualised. This is inspired by Ted Underwood's article 'Theorizing research practices we forgot to theorize twenty years ago', which argues the need for digital humanists to 'think more rigorously and deliberately about existing practices'.[52] The first central element is its key technology, as well as a central engine of the potential digital history revolution, in the form of the *search engine*. One problem with talking about 'search' for digital historians is that it is, as Underwood states, 'a deceptively modest name for a complex technology that has come to play an evidentiary role in scholarship'.[53] By 'search' is meant the algorithmic mining of large electronic databases that since the 1990s has been used by humanists. Furthermore, the term 'search' only points to its use as a finding tool and leaves out its wider methodological implications and—echoing digital historians' criticism of traditional historians' negligence of their digital tools as discussed above—that the 'scholarly consequences of search practices are difficult to assess, since scholars tend to suppress description of their own discovery process in published work'.[54] Therefore, as a way of contributing to digital history's conceptual development and to make the existing digital history methodologies more explicit and reflective, I have elsewhere described and named an already existing quali-quantitative digital history methodology. I thus proposed the term *readsearch*

for the methodology of using online keyword searches as being 'a new hybrid concept denoting a quali-quantitative methodology combining targeted close manual and machine distant reading through the use of search engines on large digital text corpora'.[55]

Furthermore, I have attempted to further explicate the various forms of readsearch methodologies and problematise the use of search for research. Taking inspiration from Underwood, who explains that 'a full-text search is often a Boolean fishing expedition for a set of documents that may or may not exist',[56] and in line with this I differentiate between different readsearch methodologies by categorising them into three main forms: *spearfishing*, *angle* and *trawl readsearch*. 'Spearfishing readsearch' designates a form of search consisting of browsing through a large text corpora close to what can be described as 'online microfilm browsing', in that the search interface is using various keywords or dates to focus the search, but at the same time allow the reader to immerse him- or herself in the text until he or she comes across any relevant findings. When using 'angle readsearch', the researcher searches for texts referring to one specific unique event, person or place and thus like an angler adapts the angles (the search terms) to tailor them for best catching a particular fish (an event or entity). Finally, in the use of 'trawl readsearch', the search is used to find many hits of a general term, word or phenomena and this is the form of readsearch where the distant machine reading plays the largest part. Like when fishing using a trawl, this is a combination of machine and manual reading. After a large fishing trawler makes a catch in its trawl, it hoists it up and empties the catch onto the vessel and then manually goes through the catch to sort out and 'throw back' the unwanted catch: fish of the wrong species or too small to matter, as well as garbage caught up in the trawl. Similarly, the texts found through a search's machine reading is in a trawl readsearch examined manually to sort out the valuable and searched-for texts. This is a methodology especially used when tracing the change of a concept or a term over time. Some readers might find these methodological neologisms as too idiosyncratic to be meaningful and whether digital historians in the future will follow in adapting the specific readsearch terminologies is of less importance. What is crucial for them to follow, however, is in reflecting on their digital epistemology, what their use of digital methods does to the historical knowledge being produced and to explicitly conceptualise and theorise their practice as historians using digital tools and resources.

The second main element of digital history 1.5 connects to historian Andreas Fickers' claims that as a response to the salience of the new digital sources, the discipline of history needs 'a new digital historicism'. This historicism should be 'characterized by collaboration between archivists, computer scientists, historians and the public, with the aim of developing tools for a new digital source criticism'.[57] Along with many digital historians, I would add to this the need for a *digital resource criticism* that extends historians' critical faculties to the digital resources they use, such as the search engines, algorithms, programs

and applications. Overall, a digital historian 'requires a more advanced understanding of the affordances of the digital in order to perform more advanced research'.[58] Historians, like most users of digital technologies, use technology 'without reflection, without understanding how it *actually* works' and thus need to develop a new digital reflexivity. Like historians are trained to consider and look for the contextual and authorial biases of our historical sources we need to think about 'the worldviews built into our tools',[59] as too often we tend to forget 'that our digital helpers are full of "theory" and "judgement" already. As with any methodology, they rely on sets of assumptions, models, and strategies. Theory is already at work on the most basic level when it comes to defining units of analysis, algorithms, and visualization procedures.'[60] In doing this, the traditional skills of historians

> are still necessary, but the focus on practice—on doing things with data—extends their application, forcing a recognition of the constructed nature of evidence and its relation to the absent past. Necessarily speculative, the historian must bring his or her expertise to bear on these digital environments and evaluate the plausibility of what they both embody and imply.[61]

When we historians start 'to think digitally', we can gain a better understanding of the underlying mechanisms, algorithms, programmed omissions and choices of our digital tools and allow the historian 'to be a better critic, a better consumer of digital data, a better user',[62] and thus a better historian.

Conclusions: Business as Usual or Going Fully Digital?

This chapter has in many ways gone against historians' normal practice. Instead of trying to see the patterns and causes of past events after the dust has settled it has tried to discern the contours of emerging phenomena and to conjecture about possible future outcomes. This it has done to try to better understand which way or ways history will take in our ever increasing digital age. Will it be the old-trodden one or a new and radically different path? This has been a necessarily speculative exposition of three routes for digital historians that could be summarised as unreflective normalisation, paradigmatic transformation and reflective appropriation. In this, it has tried to point to the third middle way as a wider route for historians who are neither satisfied with just continuing with their historical 'business as usual' by staying agnostic about its already existing digital methodological dimensions nor prepared to join the specialised minority of historians who will go 'fully digital' by learning to code or enter into collaborations with computer and information scientists. In this, I align myself with previous digital historians, such as Toni Weller, who have argued that 'part of the "them and us" problem thus far has been too much emphasis

on historians becoming something they are not; to the detriment of the fundamental skills and expertise that is the craft of the historian'.[63]

To conclude, let us return to Thomas Kuhn and take some solace from his statements 'that there can be small revolutions as well as large ones, that some revolutions affect only the members of a professional subspecialty'[64] and on rare occasions 'two paradigms can coexist peacefully'.[65] Furthermore, history teaches us that revolutions, scientific as well as political, always come at a cost and bring losses as well as benefits, such as in

> the transition from an earlier to a later theory, there is very often a loss as well as a gain of explanatory power. Newton's theory of planetary and projectile motion was fought vehemently for more than a generation because, unlike its main competitors, it demanded the [conceptual] introduction of an inexplicable force that acted directly upon bodies at a distance. Cartesian theory, for example, had attempted to explain gravity [mechanically] in terms of the direct collisions between elementary particles. To accept Newton meant to abandon the possibility of any such explanation.[66]

However, although the new ways of understanding the world were triumphant, 'the price of victory was the abandonment of an old and partly achieved goal. For eighteenth-century Newtonians it gradually became "unscientific" to ask for the cause of gravity'.[67] The task ahead for us historians is to make sure that, whoever will succeed in shaping the apparently inevitable further digitisation of the historical discipline, into a domesticated or revolutionary historical practice or something in between, that history's rewards outweigh its losses.

Notes

1 Graham et al. 2015: 35.
2 Weller 2013b: 1; Graham et al. 2015: 35. William Cronon in 2012 as President of the American Historical Association said that he 'increasingly believe[s] that the digital revolution is yielding transformations so profound that their nearest parallel is to Gutenberg's invention of moveable type' (see Cronon 2012).
3 Weller 2013a: 195.
4 Zaagsma 2013: 24; Weller 2013a: 195.
5 Fridlund 2017; Fridlund & La Mela 2019.
6 Schumpeter 1954: 42.
7 Kuhn 1961: 162.
8 Ibid.: 161.
9 Ibid.: 190.
10 Ibid.: 185.

[11] Ibid.: 186, emphasis in the original.
[12] Cohen 1987: 24, 31.
[13] Hollinger 1989: 108.
[14] Bernal 1991: xix.
[15] My distinction between digital history 1.0 and 2.0 is similar to but more general than that of Jim Mussell, who primarily discusses changing digital history practice in relation to the digitisation of source materials. See Mussell 2013: 80–91.
[16] Graham et al. 2015: 4.
[17] Ibid.: xvii.
[18] This description focuses on the historian as a researcher and does not include changes to the historian's practice as a teacher, administrator or public historian.
[19] Besides using 'invisible' domesticated digital tools such as word processing, email, search engines and electronic articles, pictures and documents in their normal professional research practice.
[20] Zaagsma 2013: 18; Mussell 2013: 90.
[21] Kuhn 1970: 5.
[22] Zaagsma 2013: 17.
[23] Weller 2013b: 4.
[24] Bilansky 2017: 517.
[25] Graham et al. 2015: 48.
[26] Rutner & Schonfeld 2014: 8.
[27] Neitzel & Welzer 2012: ix–x.
[28] Weller 2013b: 3.
[29] Graham et al. 2015: 4.
[30] Ibid.: 1.
[31] Ibid.: 23.
[32] Kuhn 1970: 84.
[33] Ibid.: 6–7.
[34] Ibid.: 85.
[35] Graham et al. 2015: 2.
[36] Ibid.: 1, emphasis added.
[37] Ibid.: 32.
[38] Mussell 2013: 81.
[39] Daniel J. Cohen in Cohen et al. 2008: 455. Cohen was echoing and answering the question posed in 2003 by his digital history predecessor Roy Rosenzweig in an article entitled 'Scarcity or abundance?'.
[40] Graham et al. 2015: 264.
[41] Ibid.: 3.
[42] Hitchcock et al. 2012.
[43] Graham et al. 2015: 235.
[44] Weller 2013b: 4.
[45] Aiden & Michel 2011.

[46] Weller 2013b: 1.
[47] Anderson 2008.
[48] Zaagsma 2013: 17.
[49] My designation of digital history 1.5 and 2.0 is close to what Zaagsma describes as 'plain IT' and 'enhanced IT' respectively (see Zaagsma 2013: 12).
[50] Fridlund 2017; Fridlund & La Mela 2019: 12.
[51] Moretti 2000; Moretti 2005; Moretti 2013.
[52] Underwood 2014: 64.
[53] Ibid.
[54] Ibid.: 65.
[55] Fridlund & La Mela 2019: 13. This is similar to 'critical search' as described by Jo Guldi (see Guldi 2018).
[56] Underwood 2014: 64.
[57] Fickers 2012: 26.
[58] Mussell 2013: 91.
[59] Graham et al. 2015: 54.
[60] Rieder & Röhle 2012: 70.
[61] Mussell 2013: 91.
[62] Graham et al. 2015: 267.
[63] Weller 2013a: 195.
[64] Kuhn 1970: 49.
[65] Ibid.: xi.
[66] Kuhn 1961: 184.
[67] Ibid.

References

Aiden, E. L., & **Michel, J.-B.** (2011). Thoughts/clarifications on Grafton's 'Loneliness and Freedom'. *Culturomics*. Retrieved from http://www.culturomics.org/Resources/faq/thoughts-clarifications-on-grafton-s-loneliness-and-freedom

Anderson, I. (2008). History and computing. Blog *Making history: the changing face of the profession in Britain*. Retrieved from http://www.history.ac.uk/makinghistory/resources/articles/ history_and_computing.html

Bernal, M. (1991). *Black Athena: Afroasiatic roots of classical civilization*, Vol. II: *The archaeological and documentary evidence*. New Brunswick, NJ: Rutgers University Press.

Bilansky, A. (2017). Search, reading, and the rise of database. *Digital Scholarship in the Humanities, 32*, 511–527.

Cohen, I. B. (1987). Scientific revolutions, revolutions in science, and a probabilistic revolution 1800–1930. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution*, Vol. 1: *Ideas in History*. Cambridge, MA: MIT Press.

Cohen, D. J., Frisch, M., Gallagher, P., Mintz, S., Sword, K., Murrell Taylor, A. M., Thomas, III, W. G., & **Turkel. W. J.** (2008). Interchange: the promise of digital history. *Journal of American History, 95*, 452–491.

Cronon, W. (2012, 1 January). The public practice of history in and for a digital age. *Perspectives in History Online*. Retrieved from http://www.historians.org/publications-and-directories/perspectives-on-history/january-2012/the-public-practice-of-history-in-and-for-a-digital-age

Fickers, A. (2012). Towards a new digital historicism? Doing history in the age of abundance. *VIEW: Journal of European Television History and Culture, 1*, 19–26.

Fridlund, M. (2017, 18 May). *Digital History 1.5: Historical research between domesticated and paradigmatic digital methods*. Video fil]. Umeå: Hum Lab, Umeå University. Retrieved from https://web.archive.org/web/20200510185709/http://stream.humlab.umu.se/?streamName=digital_history_1_5

Fridlund, M., & **La Mela, M.** (2019). Between technological nostalgia and engineering imperialism: digital history readings of China in the Finnish technoindustrial public sphere 1880–1912. *Tekniikan Waiheita, 35*(1), 7–40.

Graham, S., Milligan, I., & **Weingart, S.** (2015). *Exploring big historical data: the historian's macroscope*. London: Imperial College Press.

Guldi, J. (2018). Critical search: a procedure for guided reading in large-scale textual corpora. *Journal of Cultural Analytics*. DOI: https://doi.org/10.22148/16.030

Hitchcock, T., Shoemaker R., Emsley, C., Howard, S., McLaughlin J., et al. (2012, 24 March). *The Old Bailey proceedings online, 1674–1913*. Version 7.0. Retrieved from http://www.oldbaileyonline.org

Hollinger, D. A. (1989). *In the American province: studies in the history and historiography of ideas*. Baltimore, MD: Johns Hopkins University Press.

Kuhn, T. S. (1961). The function of measurement in modern physical science. *Isis, 52*, 161–193.

Kuhn, T. S. (1970). *The structure of scientific revolutions*. 2nd edn. Chicago, IL and London: University of Chicago Press.

Moretti, F. (2000). Conjectures on world literature. *New Left Review* n.s. *1*(1), 54–68.

Moretti, F. (2005). *Graphs, maps, trees: abstract models for a literary history*. London: Verso.

Moretti, F. (2013). *Distant reading*. London and New York, NY: Verso Books.

Mussell, J. (2013). Doing and making: history as digital practice. In T. Weller (Ed.), *History in the digital age* (pp. 79–94). London and New York, NY: Routledge.

Neitzel, S., & **Welzer, H.** (2012). *Soldaten: on fighting, killing, and dying: the secret World War II tapes of German POWs*. New York, NY: Simon & Schuster.

Rieder, B., & **Röhle, T.** (2012). Digital methods: five challenges. In D. M. Berry (Ed.), *Understanding digital humanities* (pp. 67–85). Houndmills: Palgrave Macmillan.

Rutner, J., & **Schonfeld, R. C.** (2014). *Supporting the changing research practices of historians*. New York, NY: Ithaka S+R.

Schumpeter, J. A. (1954). *History of economic analysis*. Oxford: Oxford University Press.

Underwood, T. (2014). Theorizing research practices we forgot to theorize twenty years ago. *Representations, 127*, 64–72.

Weller, T. (2013a). Conclusion: a changing field. In T. Weller (Ed.), *History in the digital age* (pp. 195–205). London and New York, NY: Routledge.

Weller, T. (2013b). Introduction: history in the digital age. In T. Weller (Ed.), *History in the digital age* (pp. 1–19). London and New York, NY: Routledge.

Zaagsma, G. (2013). On digital history. *BMGN—Low Countries Historical Review, 128*(4), 3–29.

Historical Literacy, Knowledge and Byte: Conceptual Approach to FAIR Data

Infrastructures and Data Management in Modern Scholarship

Jessica Parland-von Essen

Introduction

Historians are very good at source criticism, but in the digital era this requires good provenance data. Historians should also step up to the demand for transparency and open scholarship that comes with digital humanities. Research and knowledge has to be well documented and reliable. This means we need good data management, but also better and more integrated services and infrastructures.

Despite often exceptionally rich descriptive metadata in the cultural heritage sector, research life cycle data management is not easy and finding sources might be difficult due to questions of metadata formats or granularity of publication. The humanists' workflow and practices regarding use of sources is often hybrid and only partly digital.[1] In this chapter, I will analyse different digital data types and infrastructures from the point of view of a historian and discuss the needs of historical research and knowledge creation. Questions about data

management and information structures are important to solve, so that it is possible to formulate service needs and user stories for historical research data services. I will propose a model for planning research data management and data publication for historians. The chapter focuses on the Finnish research sector, but includes relevant international infrastructures and initiatives.

The Concept of FAIR Data

FAIR data was minted as a concept in an expert meeting among science data experts, and resulted in a seminal article on research data management published in 2016.[2] The concept, which was a more than needed completion to the Open Science, Open Access and Open Data rhetoric, won immediate approbation within the European Union and other data-aware stakeholders. It was obvious that open data or access was not by far enough to solve the issues with science reproducibility, let alone the efficiency goals of the Digital Single Market. Data cannot always be open and there were other, more technical hurdles, too. Data needed to better managed, and the money invested in research should not be wasted by sloppy planning. To make the most of our data, it has to be organised and taken well care of. Only then can we combine datasets and build digital knowledge by linking publications and data in sustainable and trustworthy ways.

In short, the FAIR principles state that data should be Findable, Accessible, Interoperable and Re-usable. It turns out that these fine words in practice result in very technical definitions. When going into details, we soon exceed the level most scholars in the humanities should have to be bothered with. We should simply have workflows and services that support these principles, but for that to happen, all stakeholders have to raise their awareness and understand what is necessary to accomplish regarding services and infrastructures.

Let's take a short look at the principles and how they could be translated into a relevant form for our purposes. F stands for *Findable*. What this actually means is machine-readable. The amount of data today is so immense that it is important that computers cannot only sort out data, but also act upon it and find what is really relevant. This means, for instance, not only that digitising text so that it is only in image form is not sufficient, but also that the content of text needs to be organised in more specific, semantic ways: it requires structured metadata and keywords, as well as common and persistent identifiers for concepts like persons or place names. Furthermore, the metadata has to be available for and utilised by different kinds of indexing and search tools.

A means *Accessible*. This, in practice, today means data that can be downloaded over the web, or at least the internet. Both machines and humans should be able to understand the information the data represents or contains, and it should not be transferred or changed in non-transparent or undocumented ways. I, as in *Interoperable*, is a tough one. It means you should be able to combine datasets and copy metadata smoothly, without losing any information.

This means you should comply with existing standards and formats. As research data management in many ways is in its infancy and the information systems are still largely insufficient or impractical, this is difficult. It is necessary to balance the needs of the research and serve the actual research use, which must be prioritised. Unfortunately, many researchers are inclined to think that their data is far more different and unique than it actually is or needs to be. Usually, it is possible to find *some* aspect of the data that somehow relates to something else, be it source, structure or some semantics of the content. As people tend to understand how much effort they have put into their own work and development, it is too easy to underestimate the value of other people's work. The *not invented here* syndrome[3] can easily trump real creative openings and slow down research. Particularly in the life sciences, there have been many important insights and tools developed (bioinformatics might be the oldest domain-specific field within research data management). We should copy as much and as fast as we can from other, successful domains.

R, which is *Re-usable* data, means that it has a functioning licence or rights statement, but also that it has been thoroughly documented so that another researcher, or the composer of the dataset in 10 years for that matter, can take a dataset and use it again. Often, researchers spend up to 80% of their time creating or cleaning their data.[4] Therefore, careless documentation can be considered an inexcusable waste of resources and time.

The utmost goal, besides efficiency, is of course trustworthy, high quality research. The digital environment has the unfortunate quality of being simultaneously dynamic and unreliable. Links, even in scientific publications, tend to break.[5] This phenomenon is called link rot. Similarly, the content behind the link might change in a devious, unnoticeable way, which is called content drift. To address this problem, one of the main building blocks of FAIR data are *persistent identifiers*. Above, I mentioned identifiers for different kinds of concepts, which makes it easy to trace and link information. Researchers might have their own identifiers in the form of an ORCID, which is personal, unique and resists changes in name form or affiliation, and makes it possible to differentiate people with the same or similar names. Correspondingly, the datasets and articles should have their own identifiers, a URN or a DOI, which makes citing clear and unambiguous. The point is then the persistence; namely, the sustainability of this identifier. This means that we need platforms and services that provide and manage them on a long-term basis. This has a direct connection to the importance of infrastructure, which I will address later in this text.

To a historian, it is obvious that one has to address problems of sustainability in the long-term perspective, as well as that the sources need to be well documented. Are there other means for evaluating the trustworthiness or suitability of the data for our needs? Or to ensure that the data are authentic and have maintained their integrity? We need to know who said what, where and when. *Simultaneously, we also need to accept that our own research outputs should meet these requirements*.

The example of citations, the ultimate goals and tests for the data, demonstrates well the problem of sustainability. We should ask ourselves how can I cite (link to) my (digital) source in a persistent and unambiguous way and how can someone else cite the data I have created? There are recommendations for this, but they are not obviously sufficient or easy to implement. The national Finnish guideline for citing research data offers principles for citing a dataset, but how to cite more dynamic resources and what to do[6] when the resource does not provide identifiers or possibilities to download or save (partial) snapshots? Or even if the researchers manage to download the needed data, where do they archive it conveniently? The questions of data management during research are inescapable for all these practical, technical reasons. However, data management is even more complex for historians, because of questions about personal data regulation, ethical issues and copyright.

The Historian's Data Life Cycle

In Finland, the government and major research funders have promptly adopted the Open Science ideology, and research data was included in the policies at an early state.[7] There has been quite extensive work done on a national level regarding services, formats and recommendations. In parallel, there has been an effort for interoperability and digital preservation within the cultural heritage sector. This has produced services like the search portal Finna.fi and the national preservation services.[8] These and their future development are of course both important from a historian's point of view. Still, the situation for research data is quite different, since research data does not come with a clear legislation, accountability and centuries-old tradition of long-term or even short-term management. Responsibilities are often unclear when it comes to both rights and costs. In the humanities, researchers are used to expecting free or subsidised services when it comes to sources and information management. On the other hand, the research outputs are clearly considered to be the property of the researcher, at least concerning copyright. The work within humanities is considered creative and personal and thus often falls under intellectual property rights legislation.

The problem is, of course, that ownership is not a simple concept when it comes to digital resources. There are many kinds of rights and responsibilities entailed in 'owning': who has the right to access, copy, use, give access, agree on use, alter or destroy a dataset? Who has the responsibility to keep the platforms running, create metadata, plan for migrations, manage access for the next decades and curate the metadata or data if errors are found? It sometimes seems that some believe that the researcher herself should have all the rights with no responsibilities, even after the research has ended. This obviously does not work. There has to be an agreement and a balance in responsibilities and rights management. The researcher might have to give up some of the control

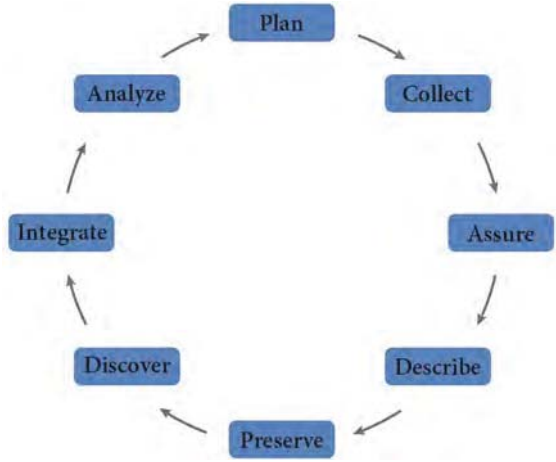

Figure 5.1: The DataONE data life cycle. Source: Author.

of her data in return for someone taking care of it. This calls for trust from both parties and concordance on common interest and explicit agreements. This is usually not a problem, but problems tend to arise from insufficient research data management planning. The agreeing would best be done in advance, preferably not by dictates from one party or the other, but by joint interests which should be easy to identify. Since the historian rarely makes up the data, but refines existing digital or non-digital data, there are usually concerns that need to be taken into account already when the data is created. Therefore, the data management life cycle always starts with planning.

There are different interpretations of the research data life cycle, but generally they tend to be variations of models that reflect the traditional way of understanding how the research process works in theory (see Figure 5.1). The idea is that there is always a project and one or several funders. Although often presented as a circular, never-ending process, one premise seems to be linear progression of the research process, as well as of science and knowledge building. This is, as any historian or other researcher knows, obviously a construct that nicely resonates with the way in which scientific publication traditionally works, with outputs that are corresponding, constructed narrations about the research process. The reality is much more chaotic and unorganised, which any data librarian will also willingly admit. The traditional publishing comprises snapshots, reports frozen in time, documenting what has been done, for dissemination and future reference. Still, these knowledge bytes are cumbersome, ambiguous and digitally discrete from the sources.

Thus, the single 'byte' of new knowledge has actually been quite open for future interpretation, often difficult to spot and point to. Even though the novelty might be a new interpretation or insight, there might also be included other new information or 'factoids', all of which become buried within an

extensive narration impenetrable for computers. Much of the information now being digital, there might be an opportunity to critically assess how we communicate our knowledge and are open to transformation within scientific publishing. Often, digitisation has meant diversification as well as convergence.[9] When we now bring data into the world of publication, there are immense possibilities for opening the whole process, enhancing documentation and sharing knowledge in new ways.

The historian has to decide upon how many of the sources can or should be linked to, in other words how many should be digitised, if the sources aren't digital and how digitisation should be achieved. Or perhaps the links are all external, linking to existing trustworthy digital sources? Data collection and creation is more complex in digital humanities than in traditional humanities research, since questions of documenting provenance and deciding on data and metadata formats will affect the research in profound ways. There are some cases where established standards exist, like TEI (xml format by the Text Encoding Initiative) for encoding text. But TEI in itself will not solve problems of interoperability on a deeper semantic level. It would, for instance, always be advisable to use good external references as identifiers for all concepts whenever possible. Also, the plan for publication might set limits to what the researchers should do, since the platform they choose might have some bearing on the formats, metadata and granularity of the publication. If the researchers use other people's digital resources (OPEDAS or Other People's Existing Data and Services, as named by a leading FAIR data expert Barend Mons[10]), they obviously need to find out extensive information about them, not only the technical and historical provenance, but also about how the data is structured and coded. Often, a historian uses OPEDAS created not by researchers, but by heritage institutions. As the use context changes, the data provider institutions generally do not have readymade generic solutions for managing and publishing research data, especially when it is produced by outsiders.

One of the unfortunate traits of the traditional data life cycle model is that publication turns up as a distinct step in a specific and late stage of the process. This hides the fact that the most efficient and impactful way of doing research might be doing it transparently *all the way*. Since this both forces the researcher to implement some type of data management and opens up for collaboration and spotting quality issues at early stages, this can accelerate the work and enhance the quality of the research. After publishing raw versions of data, unforeseen help can turn up, when colleagues become aware of what the researcher is doing. Close collaborations have not always been an option in historical research, which carries the heavy burden of romantic lonely genius syndrome, but luckily times are changing. Stealing other people's ideas and data is not the first thing most researchers think of. Rather, by publishing raw data, the researchers can get their work registered at an early stage, instead of waiting for the final peer review. Better collaborating and coordinating than working in silence.

Version control is the next crucial aspect of the data cycle. If you ask an archivist, they would probably want to save every version of everything. Even worse, this might mean not just saving the information you need to recreate the needed version of a dataset, but saving complete copies of each version, independent of all redundancy that would create. Version control is generally not that well developed in traditional archives. However, every version that is published needs metadata and, preferably, a persistent identifier. But this does not mean that the researchers have to save everything, every single byte. The researchers simply have to be sure that the dataset can be presented in an exactly identical form when asked for at a later point in time. In case somebody made a citation or important conclusion based on it, it should be possible to reconstruct what has happened. It is very important to be clear about it, if this is something the researchers do *not* commit to, when they publish data.

Managing research data is not the same thing as archiving it, and handling digital data requires a somewhat different approach. Here, storage and data management are relevant components building trustworthiness of the documentation. Citation is one of the main functions of persistent identifiers in research. The researchers should be mindful creating them though, since every persistent identifier is a commitment to manage the dataset or at least its metadata forever. It will cost somebody a substantial amount of effort and work. And even if the dataset is deleted, a tombstone page should be maintained. Here, the well-managed research infrastructures and data services come into the picture as essential supporters of research.

Generally, one could consider there to be three different types of datasets that are relevant for historians (see Figure 5.2). First, there is the master data produced and often published by government institutions, like the cultural heritage data. Unfortunately, it is not always well versioned or documented (red in Figure 5.2). It could be data of any kind for any use, but it might be relevant for a historical research question due to a long time series or for some other reason. Second, there are generic research datasets, which are produced by researchers for scientific use (green). Here, you find datasets like corpuses or some of the surveys published by the Finnish data archive. Much data of this kind can also be found, for instance, with the National Institute of Health or other domain-specific research institutes or government bodies. These datasets are validated and often cumulative. The third type of research data is a research output, created to underpin a specific study or article (blue). These data need to be saved, albeit the interest for reuse might be minute, for the simple reasons of reproducibility of the research and merit for the creator.

The historian often finds her digital sources within the first or second category of data. But as she proceeds with her work, the question of publishing second- or third-type data becomes increasingly pressing. Now, there is no single clear path to publishing this kind of data, which is often a derivate of cultural heritage data. Additionally, researchers within the humanities many times deal with sensitive data or data under copyright, which makes storing

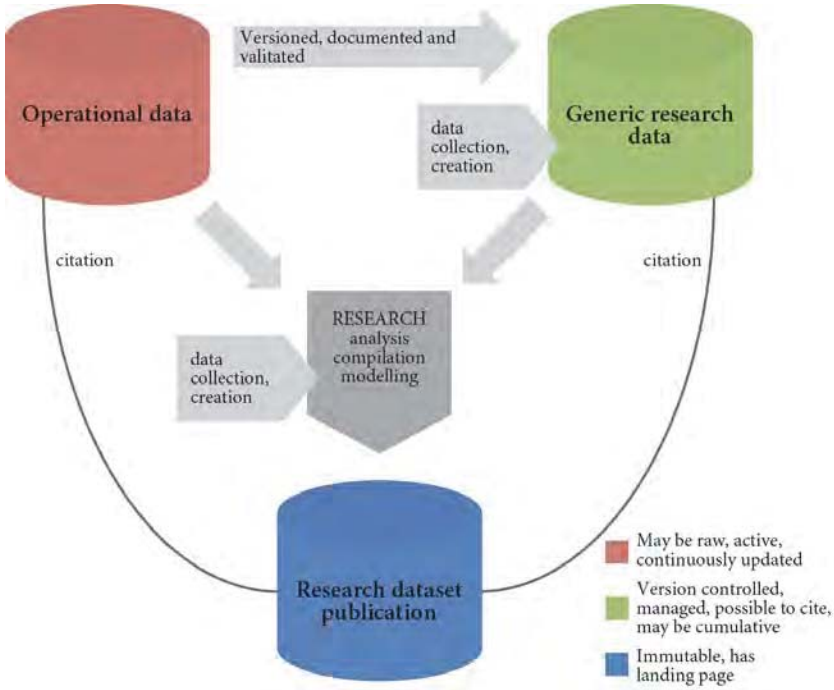

Figure 5.2: Main types of data used by historians and how those are interrelated. Source: Author.

and publishing even more difficult. I will discuss the options later in this chapter when discussing research infrastructures.

There is often more to a digital humanities research outcome than just a dataset and a result explained in a narrative form. It needs to be pointed out that the historian often handles a double narration: one of the research process and then another, which is the actual new knowledge. This is the normal situation when carrying out qualitative research or being unable to present or refer to the actual tools and methods used. However, when using computers and computational methods, the process and outcomes like dynamic databases or visualisations could and should also be included in the outputs, in addition to information about the sources or actual data. For this, the existing solutions are few and the methodology is very thin. Preservation of databases has developed somewhat, but documentation and preservation of dynamic user interfaces and other kinds of complex code is still in its cradle. It is well known that they need an extensive amount of curation to be kept usable for more than some years. This means that they are both risky and costly to preserve. Still, some effort to save these is better than just abandoning digital projects at the end of project funding. The problem is usually to find the party willing to take the responsibility. Therefore, this is also one thing that best would be solved at the point of planning the research.

Source Criticism and Research Assessment

Assessing digital sources requires a substantial amount of metadata. I need to discuss this theme more closely to explain why and how data management planning and infrastructures are relevant not only for creating FAIR data, but also for carrying out high quality research in history in a digital environment.

A digital document does not have an 'original' copy. Instead, it is recreated every time the source is rendered from a digital file consisting of zeroes and ones. Everything is just copy, while the analogue versions, which are the ones we can perceive with our senses on the screen or in our ears, are generated by software and hardware that have a decisive effect on what they actually apprehend. The calibration of the screen or the sampling frequency of an audio file might affect how one interprets what is represented or real. In cases where a physical original exists, one can always check it, but if the source is born-digital, this becomes impossible. Therefore, there is a need for technical metadata.

The best way to evaluate the trustworthiness of a digital source, as is commonplace for the historian, is to check its provenance. In practice, the researchers need to assess the organisation or person who has delivered the source. Can they show documentation about the technical and administrative life cycle of the source? Do they comply with the Open Archival Information System (OAIS) standard or do they have other certificates for digital preservation?[11] Do they use and manage persistent identifiers that are globally unique and persistent? Can they present extensive metadata, including checksums? The checksums are important digital seals for calculating the integrity of files, but they do not work across file formats, which is why the researchers need to have a good trail of documentation and management of persistent identifiers. The formats might have changed during the life cycle of the data. What else has happened in terms of migrations, curation, cleaning and enhancing the data? Is everything convincingly documented?

There are several kinds of metadata. To be able to represent a digital source in a similar or corresponding way we need technical and structural metadata that helps one choose the right tools and understand possible offset. We also need administrative metadata that informs about the rights and responsibilities attached to the data. Furthermore, we need descriptive metadata, which helps with finding and organising the data, as well as with the usual historical source criticism around what, who, when, why and other contextual information. This is the part of information that is most threatened in research data, due to reasons of personal data. Data archives often prefer anonymised data, which means crucial historical information is permanently lost from the historian's point of view. This is also the reason why the current research data archives do not provide sufficient services for many historians. The organisations that do this best are institutions like the Swedish and Finnish literary societies, which have a profound understanding of the importance of the personal and unique as part of the greater whole and of the research processes within cultural studies and history.

It is important to understand the ephemeral nature of digital information, not only when it comes to the historian's own sources, but also concerning working with data. If research is to be possible to repeat, the digital operations undertaken should be well documented. Code should be documented and saved and versions of the dataset have to be managed. Not everything has to be saved, but one should consider versioning and documentation when significant changes are made.[12] Conversions, cleaning and mapping need to be accounted for, since they may affect the outcome of the research. And as technologies become obsolete over time, all types of metadata are necessary. Otherwise preservation will not be possible.[13] This part of the data management should be planned together with data librarians and professional data stewards.

Infrastructure and Services

Reliable and good quality research craves good citations and linking. The historian's digital sources can be found in cultural heritage institutions or in many other places that sometimes, but not always, offer possibilities to create FAIR data by pointing to the sources in exact and sustainable ways. Often, the researcher needs to clean and organise the data, which in turn creates a new dataset.

According to the European Commission, research infrastructures (RIs) are facilities, resources and services used by the science community to conduct research and foster innovation. The Finnish Academy is lengthier in its definition:[14]

> Research infrastructures refer to a reserve of instruments, equipment, information networks, databases, materials and services enabling research at various stages. Research infrastructures may be based at a single location (single-sited), scattered across several sites (distributed), or provided via a virtual platform (virtual). They can also form mutually complementary wholes and networks. Europe hosts several large-scale research infrastructures that are open to collaborative use across national boundaries.

The Open Science and Research Initiative report addressed RIs.[15] This report distinguished between services, data and equipment. This classification has also been implemented in the national Research Infrastructure catalogue, which provides persistent identifiers for these (https://research.fi/).[16] Many infrastructures provide two or three of these types of resources. The national strategy for RIs[17] demonstrates that we have advanced infrastructures for linguistics, register research and social sciences. The national consortium for supplying digital publications for the research libraries within all domains is also, for some reason, considered a humanities and social sciences infrastructure.

The cultural heritage sector is left out, except for the shared search portal Finna, which serves the public as well as the research community at large when it comes to traditional research publications (namely, articles and monographs). This means the search portal aggregates some relevant research data for historians, like individual photographs or archival collections, research dataset metadata from the Data archive (a patchwork with very few internal links and of highly varied granularity) and research literature from all fields. The European cultural heritage portal Europeana does the same, except leaving out the literature and focusing on the traditional but digitised sources.

The main problem is, besides missing sufficient persistent identifier management, the lacking information structures. The digital objects vary in size from single photographs to archival collections and corpuses with almost nonexisting descriptive metadata from a historian's point of view. Saying this, I do not want to belittle the enormous and important work that has been done to bring all this metadata together. It has been an extremely valuable effort, with thorough implications for the cultural heritage sector in Finland, which has taken huge steps towards openness and digitisation. However, for research we still require better representations of the sources and their internal relations. Important digital sources are omitted, including databases provided by the institutions themselves, not to mention historical research databases elsewhere, whose producers often face great difficulties getting hold of sustainable funding or sufficient data management for their digital research outputs. The cataloguing of these resources, documentation and linking datasets derived from cultural heritage data in general is today left to the researcher, who generally has few possibilities to maintain these after the funding ends. Today, the historian most often has to be content with publishing discrete research datasets as simple files, which have weak and only human-readable links to other digital resources. Also, the reuse value is less than it probably would have to be, due to this approach and meager machine-readable semantics.

Both the Language Bank of Finland and the data archive have juridical mandates to store this kind of data, but the researcher has an extensive responsibility too. The slightest flaw in consents or rights questions easily becomes an insurmountable hurdle for archiving or sharing the data. There are also reasons to question whether this kind of publishing is the one and only, or whether there could be more suitable platforms or structures than the currently available solutions.

Digital media are not only unstable and diverse, but they are also often more disposed for interactivity and a dynamic communication that happens in dialogue, even co-creation with the readers/users.[18] In fact, it might be a mistake not to consider this kind of publishing and knowledge creation in a research domain that is so relevant and open to popularisation and popular culture. Different kinds of map and wiki applications can be used for sharing historical knowledge. Wikis are especially suitable due to their very transparent and clear version management. They also enable very good structuring and linking of

data.[19] In fact, the wiki technology combined with careful data management would offer an almost out-of-the-box solution for FAIR data.

The historian needs to carefully plan her data management. Questions of personal data, consent and copyright need to be addressed at an early stage before even starting the research. This does not mean that one has to decide on every detail or stick to the plan whatever happens. In fact, the opposite is often true: the plans have to be modified or redone, when new issues arise. The research process in digital humanities is often iterative, oscillating between qualitative and quantitative methods, and research questions sometimes have to be adjusted or revised.

From the very beginning, it is important to plan for managing data files, backups and versions. Also consider the types of data that will be included and analyse the need for documentation needed for citations and reproducibility. It is not necessarily a good idea to get a resolvable persistent identifier for every single data object. Instead, one should be pragmatic and consider the dataset as a part of the surrounding information universe and try to create meaningful, machine-readable and sustainable relations to that universe. Do not produce new data objects where you can reliably point to external ones. Also, one should be mindful about the granularity: Which are meaningful entities to make findable and for which to create metadata?

When it comes to infrastructures, we have to operate with what we have got, but historians could also give valuable input in creating a meaningful larger network of digital historical knowledge by engaging even more in questions of common or interoperable infrastructures. There are large infrastructure initiatives like DARIAH-EU, CLARIN-ERIC, Europeana and the European Open Science Cloud (EOSC), but there is still not a suitable solution that would serve historians well in publishing and linking their research outputs. It is essential that historians discuss these questions with other stakeholders, the cultural heritage institutions, the scientific libraries and their own research institutions and funders to find sustainable solutions and drive infrastructure development in directions that serve knowledge creation, not only as separate projects, but as a linked network of information.

Notes

[1] Antonijevic & Stern Cahoy 2018.
[2] FORCE11; Wilkinson et al. 2016.
[3] Not invented here 2018.
[4] Data science report 2016.
[5] Klein et al. 2014; Jones et al. 2016.
[6] Finnish Committee for Research Data 2018; Research Data Alliance 2015.
[7] Parland-von Essen 2017; see also openscience.fi.
[8] See Finna.fi, kdk.fi and digitalpreservation.fi.

[9] Anderson 2007; Manovich 2013.
[10] See Mons 2018.
[11] See, e.g., the DCP online guide on OAIS: Lavoie 2014 and the standard **ISO 16363:2012**.
[12] Language Bank of Finland.
[13] PREMIS preservation metadata.
[14] Academy of Finland 2018b.
[15] Avoimuuden politiikat tutkimusinfrastruktuureissa: Selvitys 2015.
[16] RIs, https://research.fi/.
[17] Academy of Finland 2018a.
[18] Salgado 2009; Nygren 2013; Marttila 2018; Viinikkala et al. 2016.
[19] See, e.g., Wikisources, Wikimedia, Wikidata and Tieteen termipankki. See also Wikidocumentaries and Wikimaps.

References

Academy of Finland (2018a). *Finland's strategy and roadmap for research infrastructures 2014–2020*. Interim report. Retrieved from http://www.aka.fi/globalassets/tiedostot/aka_infra_tiekartta_raportti_en_030518.pdf

Academy of Finland (2018b). *Research infrastructures*. Retrieved from http://www.aka.fi/en/research-and-science-policy/research-infrastructures/

Anderson, C. (2007). *The long tail*. Random House, London.

Antonijevic, S., & **Stern Cahoy, E.** (2018). Researcher as bricoleur: contextualizing humanists' digital workflows. *Digital Humanities Quarterly, 12*(3). Retrieved from http://www.digitalhumanities.org/dhq/vol/12/3/000399/000399.html

Avoimuuden politiikat tutkimusinfrastruktuureissa: Selvitys. (2015). *Avoin tiede ja tutkimus -hanke, Avoimuuden politiikat -työryhmä*. Retrieved from http://urn.fi/URN:NBN:fi-fe2016122731714

Data science report. (2016). *Crowdflower*. Retrieved from http://visit.crowdflower.com/r/416-ZBE-142/images/

European Commission. *About research infrastructures. What are research infrastructures?* Retrieved from https://ec.europa.eu/research/infrastructures/index.cfm?pg=about

Finnish Committee for Research Data. (2018). *Tracing data: data citation roadmap for Finland*. Retrieved from http://urn.fi/URN:NBN:fi-fe201804106446

FORCE11. *The FAIR data principles*. Retrieved 29 September 2018 from http://www.force11.org/group/fairgroup/fairprinciples

ISO. 16363:2012: *space data and information transfer systems. Audit and certification of trustworthy digital repositories*. Retrieved from http://www.iso.org/standard/56510.html

Jones, S., Van de Sompel, H., Shankar, H., Klein, M., Tobin, R., & **Grover, C.** (2016). Scholarly context adrift: three out of four URI references lead to

changed content. *PLoS One, 12*(1), e0171057. DOI: https://doi.org/10.1371/journal.pone.0167475

Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., & **Tobin, R.** (2014). Scholarly context not found: one in five articles suffers from reference rot. *PloS One, 9*(12), e115253. DOI: https://doi.org/10.1371/journal.pone.0115253

Language Bank of Finland. *Life cycle and metadata model of language resources*. Retrieved from https://www.kielipankki.fi/support/life-cycle-and-meta-data-model-of-language-resources/

Lavoie, B. (2014). *The Open Archival Information System (OAIS) reference model: introductory guide*, 2nd edn. DPC Technology Watch Report 14-02. DOI: https://doi.org/doi.org/10.7207/twr14-02

Manovich, L. (2013). *Software takes command*. New York, NY: Bloomsbury Academic.

Marttila, S. (2018): *Infrastructuring for cultural commons*. Espoo: Aalto University Series, doctoral dissertations.

Mons, B. (2018). *Data stewardship for open science: implementing FAIR principles*. New York, NY: Chapman and Hall/CRC.

Nygren, T. (2013). Digitala material och verktyg: möjligheter och problem utifrån exemplet spatial history. *Historisk Tidskrift, 133*(3), 474–482.

Parland-von Essen, J. (2017). Från open access till open science. Framväxten av öppen forskning och vetenskap. *NORDICOM-INFORMATION, 39*(1), 97–103. Retrieved from http://nordicom.gu.se/sites/default/files/kapitel-pdf/von_essen_97-103.pd

PREMIS preservation metadata. Retrieved from http://www.loc.gov/standards/premis/

Research Data Alliance. (2015). *Data citation of evolving data*. Recommendations of the Working Group on Data Citation. Retrieved from https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf

Salgado. M. (2009). *Designing for an open museum: an exploration of content creation and sharing through interactive pieces*. Taideteollisen korkeakoulun julkaisusarja A 98.

Viinikkala, L., Yli-Seppälä, L., Heimo O. I., Helle, S., Härkänen, L., Jokela, S., Järvenpää, L., Korkalainen, T., Latvala, J., Pääkylä, J., Seppälä, K., Mäkilä, T., & **Lehtonen, T.** (2016). *Reformation representation. Mixed reality narratives in communicating tangible and intangible heritage*. DIHA & NODEM Special Session at 22nd International Conference on Virtual Systems and Multimedia VSMM, Kuala Lumpur.

Wikipedia. (2018). *Not invented here*. Retrieved from https://en.wikipedia.org/wiki/Not_invented_here

Wilkinson, M., Dumontier, M., Aalsbersberg, J. J., Hoekstra, H. E., & **Boyer, D. M.** (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, 160018. DOI: https://doi.org/10.1038/sdata.2016.1

History and Digital Age: Annotated with Metadata

Kimmo Elo

Introduction

During the past decade, digital humanities has emerged as a new paradigm seeking to gather scholars interested in applying computational methods on their research materials. This development has been supported by the almost exponential growth of either born-digital or digitised materials currently available for researchers. Further, the availability of computational research tools is much better today than, say, five or 10 years ago.

New terminology like big data, data mining and text mining well illustrate the massive growth of digital data available for research purposes. At the same time, the digital research agenda is filled with huge expectations regarding exploratory research, the growth of scientific and societal knowledge or new forms of data analysis. Some scholars have rather strong expectations about how digital humanities should change our whole understanding of knowledge and how knowledge is presented.[1]

This chapter supports the general understanding of digital humanities as an important, computational field of research for the Humanities and social

sciences in general, and for historical research in particular. The chapter stems from the deep conviction of a scholar rooted in the intersection of computational, historical and social scientific research that exploring digitised historical sources could help us to gain new insights and improve our understanding of the past.

At the same time, however, this chapter is motivated by my worry that, as regards historical research, thus far much attention has been paid to the creation of digital research material, but too little has been to paid to the creation of research data. To clarify my point, with *research material*, I refer to original, primary sources like documents, letters, photographs, etc. With *research data*, I refer to corpora consisting of both the original material and additional, descriptive information derived from the original material. To put it bluntly, we are almost over-flooded by the former, but there still is no shared or common strategy about how to cope with the latter. The importance of the latter is, however, reflected by the fact that many universities are developing research data management practices.[2]

The main thesis of this chapter is that more attention should be paid and more resources should be invested in metadata creation. The next section introduces the very concept of metadata and tackles the question of why metadata matters. The second section presents arguments about why metadata should be considered as an important part of digitising projects. The chapter is rounded up with concluding remarks related to the future work in digital history.

What Is Metadata and Why Do We Need It?

Due to the limited space available for this chapter, I refrain from a literature review and just point out some of the most important aspects related to metadata and discussed (mostly) by librarians or archivists. Metadata is widely understood and defined as 'data about data' and, thus, is expected to provide information about the content of the material it is linked with. In other words, metadata should summarise the most important content. According to *The metadata handbook*, metadata should be constructed in a way which 'fully supports findability and discovery'.[3]

According to Allen Benson, metadata is a descriptive model, a summary report to present the main content according to a formalised structure consisting of information-bearing entities.[4] Richard Pearce-Moses defines metadata creation as the 'process of creating a finding aid or other access tools that allow individuals to browse a surrogate of the collection to facilitate access and that improve security by creating a record of the collection and by minimizing the amount of handling of the original materials'.[5] Hence, metadata is an ontological model providing a structure for information arrangement. At the same time, metadata creation is a descriptive process aiming at filling in the ontological model with material-related, descriptive information.

I am quite convinced that the ontological side is not the core problem. Several well-developed models exists as to how metadata should be structured or what descriptive elements are available in order to guarantee a standardised, formalised metadata.[6] Further, as regards born-digital materials specialists have been discussing from the late 1990s onwards how this development affects ontological requirements for the metadata.[7]

Hence, the real problem is the metadata creation process, especially when this process must be started from scratch and/or with only limited previous knowledge about the full content of the material to be modelled and summarised into metadata. Although the metadata should fulfil a relatively straightforward task (namely, support findability and discovery), at least three main pitfalls should be taken seriously.

First, what or who determines the elements included in the metadata structure? The answer to this question widely determines the content described and formalised in the metadata. At the same time, however, it has a strong impact on both findability and discovery, since metadata queries are limited to the fields used in the model. A more complicated issue relates to hierarchies or sub-categories typical for historical sources (for example, 'building'–'house' or 'building'–'church'). Two examples should clarify the point. Let us first consider a novel. A standard metadata includes the author(s), the title, the publisher, the year of publication, the genre and a few keywords used to summarise the main content. In most cases, these elements suit well the needs of a reader looking for certain novels. But how about a researcher looking, for example, for novels with a certain type of protagonist or a certain person/figure? Or, second, a photograph collection. Once again, many elements to be included in the metadata are quite straightforward and obvious (timestamp, photographer, title), but how about persons, places or abstract elements like gestures, memes or visual effects? The answers depend on the supposed group of end-users and, thus, make the material unusable or unfindable for certain groups.

Second, what or who determines the terminology (for example, keywords, descriptions) used to describe content? Once we have determined what content should be summarised in the metadata, we need to determine how different content-related aspects are described. Once again, standardised dictionaries, keyword indices, etc. exist, so there is rarely a need to reinvent the wheel by creating a new vocabulary. The challenge is to maintain coherence; that is, to ensure that the same (or similar) content is described in the same terms. To use a simple example, if there are bunches of photographs all having different kinds of buildings in them, all of these photographs should be found if one searches for 'building'. But should the end-user be able to find buildings of the type 'church' as well? Once again, findability should guide the process of metadata creation.

And, third, who creates and maintains the metadata? Prior to the digital era, collection management and metadata creation have been almost solely in the hands of librarians and archivists, especially when it came to the creation

and maintenance of large document collections.[8] Today, many collections are created, maintained and made available by private organisations, institutions and companies. This is partly due to the limited resources of public institutions like state archives or libraries, but also thanks to the reduced costs of digitisation, the increase of easy-to-use solutions for data management and hosting, and to the growth of data-sharing platforms like cloud-based services. The other side of the coin is that a majority of these platforms is rather weak and underdeveloped in metadata creation and maintenance, especially as regards the content description. One solution enjoying growing popularity is 'crowdsourcing', a process where 'ordinary people' help the maintainers to create descriptive metadata. There are many examples ranging from 'tagging' over 'person identification' to 'linked data creation', all of them producing interesting and promising results, but also highlighting many problems mostly related to the heterogeneous quality of the resulting metadata and difficulties in ensuring the correctness of input.[9]

Why Digital History Should Take Metadata Seriously

A quick survey in recent literature around digital history reveals that questions related to metadata creation have rarely been debated among digital historians. Instead, historians seem to be educated to use metadata when searching for sources, not to question the metadata itself. In other words, we are used to relying on metadata created by archivists or librarians.[10] This was a good practice in the times when collections were mainly and dominantly housed by libraries and archives.

The digital era has already changed this division of labour, and there is no evidence whatsoever that this would change in the future. Quite the contrary, billions of gigabytes of born-digital textual and visual materials are produced and made available without any, or with only weak and incomplete, metadata. However, without a proper metadata, materials 'are simply a meaningless collection of files, values and characters'.[11] And as Edelstein and colleagues point out: 'Historians increasingly find themselves utilizing digital databases as the idea of the searchable document and the virtual archive reorganize how libraries, research institutes, teams of scholars, and even individual researchers present and share interesting sources.'[12]

Quite much effort, money and time have been invested in the digitising of historical textual materials like manuscripts, documents, letters, etc. As a result, historians have access to a vast number of digitised text and can view and query digitised indexed document collections and editions online. One of the most prominent examples is the 'Republic of Letters' project, focusing on historical networks of correspondence between scholars from all around the world.[13] Another similar project is the 'Letters of 1916 Digital Edition' project, one of the first crowdsourced humanities projects, as well as histoGraph, which also uses crowdsourcing for metadata creation.[14]

In their evaluation of the 'Letters of 1916' project, the authors note that '[t]he meaning of the term "metadata" was unclear for most participants'.[15] This seems to be linked to a wider aspect, namely that '[m]uch attention in the past fifteen years has been directed toward text digitization',[16] forcing 'scholars to access historical sources in a new way: through specific words'.[17] As a result, most digitised collections available online are 'focused on searching, not browsing'.[18] Hence, findability might be good (thanks to the power of full text search in digitised text documents), whereas discovery might be poor.

Modern text mining methods can be of help when historians are dealing with digitised textual corpora. Further, computational methods like (semi-) automated document classification or indexing can make the metadata creation process easier and more effective. However, the current tendency to make old documents available as PDF collections worsens the situation. The positive thing in using the so-called layered PDF format is that end-users can see the original document, but also use search and copying functionalities through the text layer. The negative side is that in most cases the text layer is an exact, character-based reconstruction of the page (mostly based on the corrected results from the optical character recognition (OCR) process), not a raw text laid out and paginated according to the original design. As a result, hyphened words, to give an example, on two lines are not understood as one, but as two separate words (of which the first ends with a hyphen!). My reader can imagine what kinds of limitations result from this kind of practice for document discovery, even if the research interface offers expanded search capabilities like regular expressions. This is because most search engines are based on pattern matching, whereas, for example, irregularly split words do not have a distinct pattern.

Another growing challenge is that sources relevant for historians and social scientists include not only textual collections, but also visual or audio materials like photographs, music, films and so on. Although the question of metadata creation is relevant for all digitised collections, the real challenge relates to non-textual materials. Since the share of information delivered in non-textual, mostly visual forms is steadily growing, the problem of findability and discovery of such materials is of increasing relevance also for historians. There exists already vast collections of such materials, but at the same time our tools to directly query visual or audio materials are very limited, yet slowly improving.[19] For example, many digitised historical photographs include non-recognised persons or places, but the problem is also relevant for today. According to de Figueirêdo and Feitosa '[a]pproximately 350 million photos are added to Facebook each day[, but most of them] are not annotated'.[20] The problem here is not just about forgetting, but also about findability and discovery. Non-annotated photographs cannot be queried, and they do not appear in search results, even if their content was relevant for the query. How are we expected to find, for example, photographs with 'Konrad Adeanuer' on them if we lack both techniques to identify (that is, to name) persons behind recognised faces and metadata containing information about persons shown on the photographs?

Many recent articles point out that digitised collections and online resources affect the way in which scholars discover and access historical sources. Instead of selecting research material from the sources by close reading, research material is increasingly selected by using search engines or by applying methods of computer-aided, distant reading. Two biasing consequences seem worth being noted. First, the use of search engines and other online resources might influence and steer scholars to favour materials available online and, consciously or unconsciously, to change their research questions to suit digitally available materials. Second, scholars might not be aware of missing or incomplete metadata possibly affecting and limiting research results. This second aspect is especially relevant for non-textual material collection, but has at least some relevance also in regard to textual data offered as simple, non-indexed PDF document collections. Another problem is that many collections do not provide any information about the completeness (or better: incompleteness) of their data.

Discussion

This chapter has tackled the question of the relevance of metadata for historical research. Metadata is understood as 'data about data', an ontological model summarising the main content of the data. The very idea of metadata is to make the source material findable and discoverable. In the current digital era characterised by the exponential growth of digitised materials and the availability of vast online resources, both goal-settings gain in importance also for historical research.

Based on the arguments presented above, I conclude that metadata is extremely relevant also for historians. On the one hand, historians increasingly use and explore online resources like historical document collections or photograph corpora. Most of these online portals offer search engines or other possibilities to query the collections. Instead of selecting material by the process of reading the material document by document, material selection is increasingly based on search results. Since there is no reason to believe that this will change in the future, historians should be interested in ensuring that all relevant aspects are searchable, findable and discoverable.

On the other hand, the whole collection management is in flux, as digitised collections are made available by a wide variety of actors. If there exist no standards for quality management of data collection, how can findability and discovery be guaranteed? Once again, the ontological side is not the problem. The problem is the process of creating annotations and metadata.

A third aspect should be added to the two points above. Historical digitisation projects often deal with materials of which only trained historians possess knowledge. With all respect to librarians and archivists, we cannot expect them to have an in-depth knowledge of historical persons, events or eras. Despite this, these two groups are still in charge when national, governmental and official collections are digitised and annotated with metadata.

Although there is no easy patent solution regarding how to ensure metadata quality for historical collections, historians should be encouraged to engage in digitisation projects in their own fields of expertise. As Reilly point out, libraries, but also archives, 'must ensure that they maximize the visibility of their collections—not just to the general public but to those in the education system'.[21] In this respect, historians should engage as mediators between the research community and libraries and archives.

Historians value original documents and are trained to source criticism and to work in archives. At the same time, they are quite reliable on what is involved in the quality of collection management and hosting in archives, and many archivists and librarians enjoy a high respect for their expertise. A good archivist can fill the gaps in a researcher's inquiry and, thus, find relevant and reliable sources.

The shift from this human-to-human interface towards a human-to-computer interface replaces the 'silent knowledge' of an archivist with algorithms run by the computer. The search process itself might be more effective and quicker, but the other side of the coin is that the user has only limited possibilities to explain her intentions. As pointed out above, a scholar is forced to figure out correct terms and words for his query, but still he cannot be sure whether he receives all (or even the most) relevant materials.

To round up my argument: it is by far not sufficient to digitise original sources if we cannot ensure findability and discovery. Digitised original sources must be processed into research data consisting of the original content plus descriptive metadata summarising the essential content of the material. Metadata creation should not be disparaged, nor should it be seen as a quick, dirty task to be completed as soon as and as inexpensively as possible. Research data is the most valuable content of a vast material collection, since it enables both findability and discovery. If scholars cannot rely on getting reliable results when committing searches in online collections, the digital leap manifested by proponents of digital humanities might end with a belly flop.

Notes

[1] See, e.g., Burdick et al. 2012.
[2] See, e.g., https://www.helsinki.fi/en/research/research-environment/research-data-management.
[3] Register & McIlroy 2015.
[4] Benson 2009: 161–162.
[5] Pearce-Moses 2005: 112–113.
[6] Benson 2009; Gonzales 2014; Valentino 2017.
[7] Langdon 2016.
[8] Edelstein 2017: 401.
[9] See, e.g., Stvilia 2009; Reilly 2012; Stvilia 2012; Turin 2015; Valentino 2017; Wusteman 2017.

[10] Edelstein 2017: 401.
[11] See https://www.fsd.uta.fi/aineistonhallinta/en/data-description-and-meta-data.html.
[12] Edelstein 2017: 401.
[13] Stanford University 2013.
[14] Letters 1916–1923 Consortium 2016; University of Luxembourg 2018.
[15] Wusteman 2017: 133.
[16] Edelstein 2017: 417.
[17] Huistra 2016: 220.
[18] Ibid.: 222.
[19] See, e.g., Huang, Ma & Gong 2014; Ries & Lienhart 2014; Ko & Lee 2015; Vinyals et al. 2015; Li, Wang & Zhang 2016; Osadchy, Karen & Raviv 2016; Wang, Wang & Liu 2016; Zhong, Liu & Hua 2016; Li et al. 2017.
[20] de Figueirêdo & Feitosa 2015: 203.
[21] Reilly 2012: 39.

References

Benson, A. C. (2009). The archival photograph and its meaning: formalisms for modelling images. *Journal of Archival Organization, 7*(4), 148–187.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & **Schnapp, J.** (2012). *Digital_Humanities*. Cambridge, MA: MIT Press.

de Figueirêdo, X., & **Feitosa, H.** (2015). Semi-automatic photograph tagging by combining context with content-based information. *Expert Systems with Applications, 42*(1), 203–211.

Edelstein, D. (2017). Historical research in a digital age: reflections from the mapping the republic of letters project. *American Historical Review, 122*(2), 400–424.

Gonzales, B. (2014). The conversion of MARC metadata for online visual resource collections: a case study of tactics, challenges and results. *Library Philosophy and Practice (e-journal)*, 1–64.

Huang, M., Ma, Y., & **Gong, Q.** (2014). Image recognition using modified zernike moments. *Sensors & Transducers, 166*(3), 219–223.

Huistra, H. (2016). Phrasing history: selecting sources in digital repositories. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 49*(4), 220–229.

Ko, C.-N., & **Lee, C.-M.** (2015). Image recognition using adaptive fuzzy neural network based on lifting scheme of wavelet. *Artificial Life and Robotics, 20*(4), 353–358.

Langdon, J. (2016). Describing the digital: the archival cataloguing of born-digital personal papers. *Archives and Records, 37*(1), 37–52.

Letters 1916–1923 Consortium. (2016). *Letters of 1916 digital edition*. Retrieved from http://letters1916.maynoothuniversity.ie/

Li, K., Wang, F., & **Zhang, L.** (2016). A new algorithm for image recognition and classification based on improved bag of features algorithm. *Optik—International Journal for Light and Electron Optics, 127*(11), 4736–4740.

Li, W., Chen, L., Xu, D., & **Gool, L.V.** (2017). Visual recognition in RGB images and videos by learning from rgb-d data. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PP*(99), 1–1.

Osadchy, M., Keren, D., & **Raviv, D.** (2016). Recognition using hybrid classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(4), 759–771.

Pearce-Moses, R. (2005). *A glossary of archival and records terminology*. Chicago, IL: Society of American Archivists.

Register, R., & **McIlroy, T**. (2015). *The metadata handbook*. Retrieved from http://themetadatahandbook.com/wp-content/uploads/2015/01/Metadata-Handbook-Preview-Revised.pdf

Reilly, S. K. (2012). Collaboration to build a meaningful connection between library content and the researcher. *New Review of Information Networking, 17*(1), 34–42.

Ries, C. X., & **Lienhart, R.** (2014). A survey on visual adult image recognition. *Multimedia Tools and Applications, 69*(3), 661–688. Copyright: Springer Science+Business Media, New York, 2014; last updated 30 August 2014.

Stanford University (2013). *The republic of letters*. Retrieved from http://republicofletters.stanford.edu/

Stvilia, B. (2009). User-generated collection-level metadata in an online photo-sharing system. *Library and Information Science Research, 31*(1), 54–65.

Stvilia, B. (2012). Establishing the value of socially-created metadata to image indexing. *Library and Information Science Research, 34*(2), 99–109.

Turin, M. (2015). Devil in the digital: ambivalent results in an object-based teaching course. *Museum Anthropology, 38*(2), 123–132.

University of Luxembourg (2018). *histoGraph*. Retrieved from http://histograph.eu/

Valentino, M. (2017). Linked data metadata for digital clothing collections. *Journal of Web Librarianship, 11*(3–4), 231–240.

Vinyals, O., Toshev, A., Bengio, S., & **Erhan, D.** (2015, June). Show and tell: a neural image caption generator. In Computer Vision and Pattern Recognition conference (pp. 3156–3164).

Wang, Y., Wang, X., & **Liu, W.** (2016). Unsupervised local deep feature for image recognition. *Information Sciences, 351*, 67–75.

Wusteman, J. (2017). Usability testing of the letters of 1916 digital edition. *Library Hi Tech, 35*(1), 120–143.

Zhong, S.-H., Liu, Y., & **Hua, K. A.** (2016). Field effect deep networks for image recognition with incomplete data. *ACM Trans. Multimedia Comput. Commun. Appl., 12*(4), 52:1–52:22.

Need of Manual Labor in Digital Age

Johan Jarlbrink

Automation is a temptation and a promise, and perhaps a threat. Old jobs disappear as robots and software do what human workers used to. Is this also the case with research within the humanities? Computers can process datasets of texts so large that it would take several lifetimes for scholars just to read it. Computers are excellent in finding patterns that are hard to recognise for human eyes and brains. What should researchers do when computers are much better in doing what scholars used to?

In this chapter, I will argue that digital research is far from automatised.[1] A human being is still needed to make sense of results, of course. I will focus on something else, not on the creative ways in which scholars interpret data outputs, but on the dull tasks that make data outputs possible. Most datasets need cleaning, editing and error checking. The outcome of automatic processes needs to be examined by someone who goes through the results; sometimes it needs to be corrected manually. Such procedures are often left out completely or only mentioned in brief when digital methods are discussed. Yet, they have a significant impact on results and need to be taken seriously.

I will mainly focus on various forms of text analysis, based on my own experiences and what colleagues have told me, as well as cases described in the

literature. The cases are meant to shed light on an often neglected part of digital methodologies, but the mundane aspects of data cleaning and curation are also significant beyond the field of digital humanities. Such procedures can be understood as 'a crucial part of the materiality of how scholarly and scientific work is done'.[2] The manual work needed to feed, improve and evaluate digital processing belongs to a long history of little tools and (supposedly) insignificant back-end operations that have made different kinds of research output possible. Digital scholarship, as traditional archival research and experimental work in laboratories, involves material and conceptual actors as well as human ones.[3]

In the first section, I will give a short background and explain why I think manual digital work matters. Three empirical sections will exemplify various kinds of manual operations. First, I describe human-assisted computational analysis in the humanities in the 1950s, 1960s and 1970s. Second, I present my own experiences from a project based on 19th-century newspapers. Third, I tell the story of how a colleague of mine used digital Named Entity Recognition (NER) in combination with pen and paper.

Invisible Work

Glimpses of the manual work that makes digitisation and computational analysis possible are sometimes given by accident. Google Books preserves a large part of our printed cultural heritage in a digital form, but also some of the hands that were needed to operate the scanners and handle the printed volumes. Just as the secretaries of the early 20th century, who left traces of themselves in the typewritten texts only as a result of errors, accidents make Google employees become visible in the digital database. Index fingers covered in condom-like pink gloves are included in many of the images now available online. They serve as a reminder of the people and work that feed the digital infrastructures.[4] Part of the workforce digitising printed materials is less visible. Much of the post-processing needed to produce useful digital surrogates is being outsourced to companies hiring low-wage workers in Cambodia and India.[5]

This kind of hidden work makes digitisation seem more straightforward and automatised than it is. The same goes for various forms of computer-assisted analysis. Tamraparni Dasu and Theodore Johnson has stated that:

> In our experience, the tasks of exploratory data mining and data cleaning constitute 80% of the effort that determines 80% of the value of the ultimate data mining results. Data mining books … provide a great amount of detail about the analytical process and advanced data mining techniques. However they assume that the data has already been gathered, cleaned, explored, and understood.[6]

Much of the cleaning can be done with software. Even an easy-to-use program such as Microsoft Excel allows you to search and replace, filter, merge, separate

and delete different kinds of data. More advanced or custom-made tools allow you to fine-tune the process. Still, such procedures need to be monitored in order to ensure the quality of the outcome. Sometimes software fail, and sometimes they need assistance from human pattern recognition. With a limited dataset, it can be more efficient to correct and edit by hand instead of spending time finding and running a software that will require additional and manual error checking anyway.

Algorithms solve problems according to specified rules. That is why they may be of limited use if a dataset is noisy and patterns are irregular. 'Signals are always surrounded by noise, even to the extent that we cannot always decipher which is which.'[7] Hadley Wickham explains (alluding to Leo Tolstoy) that 'tidy datasets are all alike but every messy dataset is messy in its own way'.[8] A dataset can be corrupt in numerous ways, but there is only one way in which it is flawless. The multiple possibilities of errors, and the irregularity of their occurrence, can make it difficult to specify the rules on how to solve problems algorithmically. In some cases, the fastest way may be to do some of the work manually.

As Dasu and Johnson point out, cleaning has a significant impact on results. Yet, detailed discussions on cleaning and error-checking processes are rare in introductions and chapters on methodology. Introductions usually describe digital tools, not manual or semi-manual tasks.[9] The role of digital tools and models is often discussed in terms of black boxes, with an input and an output and an obscure software in the middle. Such black boxes must be opened up in order to make research processes transparent.[10] Manual and semi-manual procedures can be said to represent another black box, however, perhaps even more opaque. They can be difficult to describe in a transparent way since they rely on human pattern recognition, a sensitivity to individual cases and the ability to make informed distinctions between information and noise.

A History of Manual Labour

As Markus Krajewski has pointed out in his media history of service, before digital servers there were human servants: human calculators, research assistants and secretaries.[11] The birth of automatised data processing did not do away with them. When Vannevar Bush speculated on the future research potentials of computers in 1945, he described a machine that 'will take instructions and data from a roomful of girls armed with simple keyboard punches, and will deliver sheets of computed results every few minutes'.[12] Father Robert Busa is often referred to as the first scholar to use the capabilities of computers within the humanities. However, his project also involved 'a roomful of girls'. His interest started in the 1940s when he studied the preposition 'in' in the works of Thomas Aquinas. This research would clearly benefit from the technologies developed to speed up data processing in business and administration. Busa partnered with IBM and during the following decades they constructed

an index of the full vocabulary in the works of Aquinas, published in 1974 as *Index Thomisticus*. In words that echo in recent publications on distant reading, Busa stated that: 'What had first appeared as merely intuition, can today be presented as an acquired fact: the punched card machines carry out all the material part of the work of (making a concordance).'[13]

The process was far from automatic though. The mainframe computers available at the time required 'a constant procession of human servants'.[14] In 1964, Busa had a team of 60 people assisting him with editing, programming and machine operations. Around 35 staff members were required for key-punching texts, verifying, listing, sight-checking and punch-card processing (the data was later transferred to magnetic tapes). In 1951, he estimated that it would take four years to complete the index. The reason why the project was not finished until the mid-1970s was mainly the laborious work of pre-editing and proofreading. 'Busa calculated that the thirty years of work he and others had spent on it amounted to roughly one million man hours.'[15] The foundational project of what would become digital humanities was truly a manifestation of the manual work needed to process data with computers.

The labour-intensive process did not discourage other scholars from using computers in their research (perhaps because those who introduced new methods seldom emphasised the importance of manual tasks). When the *Index Thomisticus* was completed in 1974, Busa was no longer alone. Linguists were among the early adopters, as well as some historians. Swedish historians were introduced to the idea that 'Clio faces automation' in an article by Carl Göran Andræ from 1966. Andræ explained that modern computers provided solutions to problems related to massive source materials. With data coded onto punch cards, or optical and machine-readable paper forms, it was possible to sort large amounts of data mechanically or electronically. In many cases, the systems were used as search engines, but they could also perform statistical calculations. The examples he gave included databases of coded newspaper articles, correlations between election results and census data, and the geographical distribution of unions and memberships in popular movements. Andræ concluded, as Busa before him, that: 'The mechanical work can now be left to computers.'[16]

Details on the actual research process are rare in publications by the first generation of computer-using Swedish historians.[17] Assistants, secretaries and machine operators may have been essential parts of the research process, but they were rarely acknowledged in the end results. Some clues can be found, however, and the impression they give is quite different from Andræ's optimistic view. The most laborious tasks concerned coding, in this case referring to the transfer of data from source documents to machine-readable formats (punch cards or optical markings on paper forms). A Swedish pioneer, the press historian Stig Hadenius, explained in 1967 that it took 'not more than 16 people' to extract the data needed for a pilot study on political news between 1896 and 1908.[18] A large project on Sweden during the Second World War had

a group of researchers investigating newspaper debates during the war. In order to render the newspaper material searchable, they coded 165,000 articles to create an index based on punch cards. The research team manually coded 138 variables for every article.[19] In the 1970s, a series of dissertations from Lund University used similar methods to process newspaper articles on various topics during the postwar era. Gunnel Rikardsson, who wrote about *The Middle East conflict in the Swedish press* (1978), did not elaborate on the manual tasks, but explained that six people had been involved in the process and that the 'coding work was experienced as exacting, mainly due to the high degree of concentration needed'. When the newspaper data was finally coded, however, the computer took over the workload: 'Manual processing had not been possible.'[20]

In his article from 1966, Andræ speculated on future research possibilities. Governmental agencies were already using computers to store and process data. Thus, for future historians who wanted to analyse the data, computer skills would be an absolute necessity. Most of the sources that historians worked with in the 1960s and 1970s were not 'born digital' though. The technologies (such as Optical Character Recognition, OCR) transferring analogue data to digital media showed promising results, but the majority of the research projects relied on manual labour. Millions of hours were spent on manual coding, punching and proofreading. The name of the research centre founded by Busa in the early 1960s was *Centro per L'Automazione dell'Analisi Letteraria*. Yet, and contrary to the automation emphasised in the name, photographs from the centre show what was often left unnoticed when research output was presented: rows of human operators, most of them young women.[21]

Struggling with Noisy Newspapers

The manual tasks needed today are of a different kind. The digitisation of sources is part of many research projects, but with scanners and software for OCR the digitisation of printed texts can be more or less automatised. Even handwritten texts can to some degree be digitised with the help of OCR technology. A significant difference, though, is that archives and libraries do much of this work for us. This is especially true for newspapers and books, parliamentary records and collections of audio-visual media, paintings and maps, and other museum artifacts. As long as the copyright allows for it, texts and images are made available online. In most cases, we do not need (and cannot afford) 35 assistants transferring data from one medium to another. Full-text search, topic modelling and tools for text analysis often make it unnecessary to code individual texts manually.

And yet, not all datasets are ready for processing out of the box; many of them can be very messy. As Carl Lagoze has pointed out, traditional archives and libraries used to guarantee the integrity of their records, at least in principle.

Control and curation were meant to facilitate the provenance and stability of data. The digitisation of collections and archival records has meant a fracturing of this control zone.[22] When millions and millions of pages are transferred (or translated) into digital formats, no one can guarantee the integrity of the data anymore. For large datasets of non-canonical texts in particular, libraries have spent less resources on curation, leaving researchers with much of the cleaning and preparation. Newspaper databases are notorious in this respect. Frequent OCR errors are well known, problems related to text segmentation less so, but both kinds of errors make it difficult to process the texts without manual interventions.

In one of my projects, I wanted to analyse discursive patterns in newspaper reports about the electrical telegraph in mid-19th-century Sweden.[23] From the National Library of Sweden, I was able to download a complete dataset covering one major newspaper from 1830 to 1862, about 10,000 pages. A systems developer helped me to penetrate the data (the first person who was asked refused to work with a dataset this noisy). Our first goal was to find every article containing the words 'electrical' and 'telegraph' ('elektrisk' and 'telegraf' in Swedish). Since we expected a high frequency of OCR errors, we used a Levenshtein distance to identify corrupted versions of our keywords, allowing three characters to be added, replaced or missing. In this way, we got 489 different hits for 'electrical' and 4,017 for telegraph. This was all done with a few simple commands, and the result came quickly.[24]

Not all of the hits had anything to do with the electrical telegraph though, and in order to filter out the false positives I had to go through the lists manually. That 'dialektisk' and 'apoplektisk' referred to something else was easy to figure out, but what about 'pelektriska' and 'elepris'? What about 'tograf', 'tfiesraf' and 'ttlefrnf'? Such combinations of characters can only be interpreted in the context of their appearances in the newspaper. In order to single out the proper keywords, I had to search the database and read the texts. It turned out that many of the incomprehensible words generated by the OCR actually referred to the electrical telegraph. My corpus would have been much smaller if I had not spent some time on this semi-manual step.

With an edited list of keywords, it was possible to locate every 'textblock' in the XML-files where 'electrical' and 'telegraph' co-occurred. A textblock is a unit of text identified as a coherent text by the text segmentation tool used in the digitisation process. However, nineteenth-century newspapers are difficult to process for the tool. The small print, the lack of headlines and the packed columns give few graphical clues on where one text finishes and another one starts. Human eyes can see it quite easily, while digital tools make several mistakes. Many libraries send the auto-segmented newspaper pages to private firms with outsourced divisions in Eastern Europe, Cambodia and India. The job of the staff is to correct the segmentation where it has failed.[25] The National Library of Sweden have skipped this crucial step in the process, however. I had to do the job myself.

We soon discovered that the textblocks generated by the tool had little to do with the texts as they were printed in the paper. Short news items from the same column were regularly merged into one single textblock, and longer texts chopped up into shorter pieces. The only way to single out the texts I wanted was to read through the whole corpus of identified textblocks and delete the unrelated parts. I also deleted text lines and combinations too difficult to decipher, such as 'lPlApfos2kOS2viKfSbmNAl' and 'rilet4R12bin1dPRRmo-8botoFrfutmfsOMMFgpgFvf'. I did not read the texts as carefully as I would have done if close reading was my main research method. But still, I had to read them.

With a somewhat clean dataset we could finally start to explore what the texts had to say about the electrical telegraph. We used a fairly simple and transparent method to identify semantic patterns. We looked at words co-occurring in a sliding window, and used the network analysis tool Gephi to find clusters of frequently co-occurring words. We still had some problems with noise though. Our method identified co-occurring words no matter the quality of the OCR, but for the final analysis we wanted to merge corrupted versions with the uncorrupted (for example, 'oeanen' and 'oceanen' (the ocean), 'Mo«se' and 'Morse'). Once again, we used a Levenshtein distance to pick out the most likely candidates to be merged, but I went through the lists to confirm the results manually.

In the end, we came up with some new and fascinating results. Many of the ideas we frequently associate with the electrical telegraph were more or less absent in the newspaper reports. Very few mentioned anything about the utopian potential of the new medium, it was not seen as an immaterial way of communicating and the idea that it 'freed communication from the constraints of geography' must be contextualised.[26] A bureaucratic discourse on regulation was much more prominent than a utopian on liberation, many of the articles described the material components of the new network instead of immaterial flows of electrical signals, and the geographical prerequisites (such as ocean floors and mountains) that determined where cables could be laid out were described in detail. I recognised much of this already when I read the texts in order to delete the noise, but I believe the quantitative analysis made the conclusions more convincing.

Scholars writing about computational text analysis usually emphasise the need to combine distant and close reading.[27] You need to switch between different perspectives to get an understanding of general patterns, as well as individual cases. In my own research, I already had to read the texts more or less closely in order to clean and prepare the corpus. When I reviewed the lists and graphs of frequently co-occurring words, I had an in-depth knowledge about the dataset on which they were based, making it easier to interpret the output. The time I spent reading and editing turned out to be well invested, but the process was very different from what I had imagined when I started the project.

Recognising Named Entities

What media technology we consider to be the first one ever invented depends on our definition of media. One common definition emphasises that a medium is a technology for the storage and/or transfer of information.[28] In that case, the tally stick might be the oldest media technology in human history. A tally stick keeps track of things you want to count (days, people, objects, etc.) and makes it possible to save the counts for later and to transport them from one place to another. The oldest one found, a bone from a baboon with carved markings, is at least 40,000 years old. 'Although our ancestors could not have known it, their invention of the notched stick has turned out to be amongst the most permanent of human discoveries.'[29] That my colleague Erik is using their invention to keep track of an imprecise digital tool in 2018 would definitely be beyond their imagination. Erik counts on paper though, not a bone from a baboon.

Tools for NER make it possible to identify and extract names of persons (even mythological creatures), organisations and places in digitised texts, as well as expressions of time (1857, 'next week'), monetary values and so on. The extracted data can be used for geographical visualisations, for network analysis, in timelines and as building blocks in other kinds of text analysis. HFST-SweNER, a language-processing technology developed to extract named entities from Swedish texts, is based on a dictionary as well as rules for identifying entities not in the dictionary, but likely candidates based on their contexts.[30] Tests have shown that it works fairly well for a curated corpus of texts from the 1990s, but will it work for 19th-century newspaper texts?

Erik Edoff is a media historian interested in geography. In one of his projects, he tries to figure out how new communication technologies in the 19th century reorganised the notion of space.[31] Was the world getting smaller when telegraphs, railroads, canals and steamships made it possible to communicate across space in a shorter time or in no time at all? Did far-away places come closer as a result of a time-space compression? One way to examine this (but certainly not the only one) is to identify and count place names in newspapers before and after the introduction of the new technologies (Erik selected papers from 1850 and 1890). Were names of distant locations printed more frequently when news travelled faster? The first results generated with NER indicated that places in the local region were in fact getting relatively more attention when new connections made communication faster, compared to places outside of the region. These were exciting results, since they seemed to show that the impact of the new technologies was different from what is usually believed. The question then became whether these numbers could be trusted. Did the tool find all the place names printed in the papers? If not, was it biased towards local Swedish place names?

In order to calculate the precision and recall, Erik chose a few newspaper issues for every title and year in the corpus. He read through the NER-tagged text files manually, and kept track of valid hits and false negatives in two

Figure 7.1: How many named entities (locations) did NER find, and how many more did Erik find? New locations not tagged by the tool were recorded on post-it notes. Source: Author.

columns on a couple of paper sheets. The method of counting was basically the same as the one used by our distant ancestors making notches on a bone: one mark for every word counted (see Figure 7.1). The brackets enclosing some of the counts separate place names mentioned in advertisements and lists, such as weather reports, stock market prices, etc. Those entities were more difficult to identify for the digital tool. There are other and perhaps more sophisticated ways to count occurrences of place names. But pen and paper are often efficient tools for minor tasks. No downloading or installing is required, and no special training. The interface makes the paper easy to use, and it is highly flexible.

The manual control revealed that the tool had left several place names untagged. For some reason, it did not recognise locations such as Paris, Kiel or Swinemünde, nor the Swedish towns Gävle (in the 19th century: Gefle), Växjö (Wexiö) and many minor towns and villages. One explanation might be the old spelling, but in some cases (when the spelling changed between 1850 and 1890), the tool recognised the old spelling, but missed the new. And the spelling does not explain the case of Paris. One geo-administrative category was left untagged almost completely: the parish. Today, it is hardly used outside of the Swedish church, but in the 19th century it was one of the most common ways in which Swedish locations were identified. Apart from these place names, Erik

found several locations untagged because of OCR errors. All of the entities identified manually were fed back into the system in order to make the final hit list more complete.

It turned out that the trend indicated by the first results was even more prominent once the false negatives were included. The relative frequency of places in far-away countries did not increase with the introduction of new communication technologies. Rather, locations close to the towns where the newspapers were published got more attention in 1890 compared to 1850. Erik's close reading of the sample issues provided him with some possible explanations. New places were put on the map thanks to new communication technologies: railway intersections, telegraph stations, bridges where steamships picked up passengers and goods, locks connecting canals and lakes. The places most frequently mentioned were those in the region, such as neighbouring towns and villages connected by railway, harbours close to home and regional centres nearby where telegrams were sent. New communications brought neighbours together. What was already close came even closer, while distant places were as far away as they were before. The repetitive task of recording place names on paper paid off in an interesting and convincing analysis. NER was a helpful tool, but it needed human assistance.

Troubleshooting Black Boxes

Digital models and tools will continue to improve. In the future there will, hopefully, be no need to carry out many of the manual tasks described in this text. OCR is getting more accurate every year; for some languages, NER seems to work fine already. On the other hand, as digital research practices are becoming more widespread, researchers will try to use the methods for new kinds of materials and in new areas—even areas where they will not run as smoothly. If we limited our research to clean datasets, very little would be accomplished. Many of the manual tasks carried out by research assistants and undergraduates in the 1960s are automatised today. New tools can achieve things unthinkable 50 years ago, but not always without human interventions. New problems seem to arise as old ones are taken care of.

The long history of information management can be seen as a series of new solutions generating old problems. In a fascinating article about the paper technologies used by Carl Linneus, in his big data-project on the natural system, Staffan Müller-Wille and Isabelle Charmantier note a 'curious dynamic' in the attempts to master information overload. 'The many technologies that were designed to contain information actually fuelled its further production, partly by providing platforms for more efficient data accumulation, partly by bringing to the fore new structural relations and patterns within the material collected.'[32] The result of technologies, developed to create order, overview and searchability, is often a new information overload. The digital media of today have other

capabilities than Linneus' paper slips and lists, but their operations are not as precise and clean as we might think. Rotten data, spam and noise thrives in a digital habitat (an interesting research topic in itself).[33] As shown by libraries' digitisation efforts, new technologies are far from perfect and human assistance is sometimes needed to keep them on track.

To edit, clean and validate large datasets manually or semi-manually may seem highly ineffective. In many cases, however, these procedures can be quite effective. Reading, counting, deleting and merging texts and other kinds of data in a manual or semi-manual fashion is a way to bridge distant and close reading. Insights from such encounters with data can be fruitful in the final analysis. It might also be a way to dig deeper into the inner workings of the digital tool on which the researcher is relying, to figure out how a specific dataset was processed and why the output turned out as it did. Troubleshooting is a good way to start if we want to examine what is inside the black boxes.

Notes

[1] The research presented here is part of the project 'Digital Models: Techno-historical collections, digital humanities & narratives of industrialisation', funded by the Royal Swedish Academy of Letters, History and Antiquities.
[2] Star 2002: 109.
[3] On the role of marginal (and yet central) figures, actions and technologies in the history of science, see Becker & Clark 2001 and Krajewski 2018.
[4] Thylstrup 2018: 42–43. See also Price & Thurschwell 2005.
[5] Fyfe 2016.
[6] Dasu & Johnson 2003: ix.
[7] Parikka 2012: 111.
[8] Wickham 2014: 2.
[9] See, e.g., Jockers 2013; Graham, Milligan & Weingart 2016; Rockwell & Sinclair 2016.
[10] Rieder & Röhle 2012.
[11] Krajewski 2018.
[12] Bush 1945: 104.
[13] Robert Busa quoted in Burton 1981: 1.
[14] Krajewski 2018: 308.
[15] Burton 1981: 3.
[16] Andræ 1966: 96.
[17] Jarlbrink 2015.
[18] Hadenius 1968: 68.
[19] The coding manual is now available online. See Åmark 2013.
[20] Rikardsson 1978: 59–60.
[21] Jones 2018.
[22] Lagoze 2014.

[23] Jarlbrink 2018.
[24] The newspaper noise is further explored in Jarlbrink & Snickars 2017.
[25] Fyfe 2016: 565.
[26] Carey 2008: 157.
[27] Jockers 2013: 26; Blevins 2014: 126; Hitchcock & Turkel 2016: 953.
[28] Mitchell 2017.
[29] Ifrah 2000: 64.
[30] Kokkinakis et al. 2014.
[31] Edoff, forthcoming.
[32] Müller-Wille & Charmantier 2012: 4.
[33] See Parikka & Sampson 2009; Eriksson 2016.

References

Andræ, C. G. (1966). Clio inför automationen. *Historisk Tidskrift, 86*(1), 47–79.

Becker, P., & **Clark, W.** (Eds.) (2001). *Little tools of knowledge: historical essays on academic and bureaucratic practices*. Ann Arbor, MI: University of Michigan Press.

Blevins, C. (2014). Space, nation, and the triumph of region: a view of the world from Houston. *Journal of American History, 101*(1), 122–147.

Burton, D. M. (1981). Automated concordances and word indexes: the fifties. *Computers and the Humanities, 15*(1), 1–14.

Bush, V. (1945). As we may think. *The Atlantic Monthly*, July, 101–108.

Carey, J. W. (2008). *Communication as culture: essays on media and society*. London and New York, NY: Routledge.

Dasu, T., & **Johnson, T.** (2003). *Exploratory data mining and data cleaning*. Hoboken: John Wiley.

Edoff, E. (forthcoming). Revolutions in communication? Digital methods and nineteenth century Swedish press.

Eriksson, M. (2016). Close reading big data: the echo nest and the production of (rotten) music metadata. *First Monday, 21*(7). DOI: https://doi.org/10.5210/fm.v21i7.6303

Fyfe, P. (2016). An archaeology of Victorian newspapers. *Victorian Periodicals Review, 49*(4), 546–577.

Graham, S., Milligan, I., & **Weingart, S.** (2016). *Exploring big historical data: the historian's macroscope*. London: Imperial College Press.

Hadenius, S. (1968). En kvantitativ innehållsanalys av dagspressen: teknik och användning i modern historisk forskning. In *Opinion och opinions bildning som historiska forskningsobjekt: Föredrag vid Nordiska fackkonferensen för historisk metodlära på Hässelby slott 4–6 maj 1967*. Oslo: Universitetsforlag.

Hitchcock, T., & **Turkel, W. J.** (2016). The *Old Bailey proceedings, 1674–1913*: text mining for evidence of court behavior. *Law and History Review, 34*(4), 929–955.

Ifrah, G. (2000). *The universal history of numbers: from prehistory to the invention of the computer*. New York, NY: John Wiley.

Jarlbrink, J. (2015). Historievetenskapens mediehantering. In M. Hyvönen, P. Snickars & P. Vesterlund (Eds.), *Massmedieproblem: mediestudiets formering*. Lund: Lunds universitet.

Jarlbrink, J. (2018). Telegrafen från distans: ett digitalt metodexperiment. *Scandia, 84*(1), 9–35.

Jarlbrink, J., & **Snickars, P.** (2017). Cultural heritage as digital noise: nineteenth century newspapers in the digital archive. *Journal of Documentation, 77*(6), 1228–1243.

Jockers, M. (2013). *Macroanalysis: digital methods & literary history*. Urbana, Chicago and Springfield, IL: University of Illinois Press.

Jones, S. (2018). Reverse engineering the first humanities computing center. *Digital Humanities Quarterly, 12*(2).

Kokkinakis, D., Niemi, J., Hardwick, S., Lindén, K., & **Borin, L.** (2014). HFST-SweNER: a new NER resource for Swedish. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. No. 391. Reykjavik, Iceland: European Language Resources Association.

Krajewski, M. (2018). *The server: a media history from the present to the Baroque*. New Haven, CT: Yale University Press.

Lagoze, C. (2014). Big data, data integrity, and the fracturing of the control zone. *Big Data & Society, 1*(2), 1–11. DOI: https://doi.org/10.1177/2053951714558281

Mitchell, W. J. T. (2017). Counting media: some rules of thumb. *Media Theory, 1*(1), 12–16.

Müller-Wille, S.. & **Charmantier, I.** (2012). Natural history and information overload: the case of Linnaeus. *Studies in History and Philosophy of Biological and Biomedical Sciences, 43*(1), 4–15. DOI: https://doi.org/10.1016/j.shpsc.2011.10.021

Parikka, J. (2012). *What is media archaeology?* Cambridge and Malden, MA: Polity Press.

Parikka, J., & **Sampson, T. D.** (Eds.) (2009). *The spam book: on viruses, porn, and other anomalies from the dark side of digital culture*. Cresskill, NJ: Hampton Press.

Price, L., & **Thurschwell, P.** (2005). Invisible hands. In L. Price & P. Thurschwell (Eds.), *Literary secretaries/secretarial culture*. Aldershot and Burlington, VT: Routledge.

Rieder, B., & **Röhle, T.** (2012). Digital methods: five challenges. In D. M. Berry (Ed.), *Understanding digital humanities*. Houndmills: Palgrave Macmillan.

Rikardsson, G. (1978). *The Middle East conflict in the Swedish press: a content analysis of editorials in three daily newspapers 1948–1973*. Stockholm: Esselte studium.

Rockwell, G., & **Sinclair, S.** (2016). *Hermeneutica: computer-assisted interpretation in the humanities*. London and Cambridge, MA: MIT Press.

Star, S. L. (2002). Infrastructure and ethnographic practice: working on the fringes. *Scandinavian Journal of Information Systems, 14*(2), 107–122.

Thylstrup, N. B. (2018). *The politics of mass digitization*. London and Cambridge, MA: MIT Press.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software, 59*(10), 1–23. DOI: https://doi.org/10.18637/jss.v059.i10

Åmark, K. (2013). *Sverige under andra världskriget: pressregister 1938–1945*. Stockholms universitet, Historiska institutionen: Svensk nationell data tjänst (SND). Retrieved from snd.gu.se/catalogue/file/3386