

Introduction: History and Data Science: Data-driven Inquiries into the Past

Gabor Mihaly Toth <gabor_dot_toth_at_maximilianeum_dot_de>, Center for Contemporary and Digital History, University of Luxembourg  <https://orcid.org/0000-0002-4301-1581>

DOI: <https://doi.org/10.63744/jgy3ruqzndgr>

Abstract

The integration of data science methods into historical research is an emerging joint focus of digital humanities and history. This new research area applies data science methods to study large historical datasets, such as correspondences, chronicles, oral history interviews, photos, and maps. As part of the overarching goal to further our knowledge of the past, the application of data science methods in historical research involves an array of epistemological questions as well. For instance, what can count as evidence and knowledge in the context of data driven inquiries into the past?

Introduction

After the Second World War, a young Jesuit embarked on an uncommon career path. The Italian professor of philosophy joined a team of American computer scientists working on the Pentagon's Machine Translation system. It was the beginning of the Cold War, and the Jesuit offered his knowledge of Russian. Later, in 1949, he became acquainted with Thomas J. Watson, the founder and the CEO of IBM; he convinced the entrepreneur to transform the oeuvres of Thomas d'Aquino into a gigantic database. His collaboration with IBM took about 30 years. By the end of the 1970s, it had completed the Index Thomisticus, the earliest computerised resource in the Humanities. ^[1] Today, the young Jesuit, Father Roberto Busa (1913 - 2011), is known as the founder of digital humanities and a pioneer of data-driven inquiries into the past. This special issue offers a glimpse into what data-driven study in history means eighty years after Busa set off for his unexpected journey into technology and computer science.

Since the pioneering work of Roberto Busa, there have been innumerable technological developments in computer science, digital humanities and beyond. However, two specific developments inspired this special issue. Today, we live in the so-called age of datafication, a development that profoundly impacts the way historians dig into the past and discover our history. With libraries, archives, and other cultural institutions digitising their collections, more and more machine-readable and ready-to-use historical data sets are becoming available. The abundance of historical data sets is a development begging for data-driven inquiries into the past.

Busa's decades-long collaboration with IBM highlights the other recent development that inspired this special issue. Busa and IBM needed many years to complete the Index Thomicus not only because the works of the medieval philosopher were not available in digital format. Their unusual collaboration lasted for decades because the technology to study vast amounts of textual data was not ready. It was not as easily accessible as it is today. By the mid-2010s, thanks to the emergence of humanities-oriented platforms, analysis tools, languages, and data representation systems, computational technology to study data has become widely available for anyone willing to use it. Today, studying a corpus of a few million words is possible with open-source tools on any personal computer. In addition to the technology needed for analyzing historical data sets, the knowledge to operate the technology has also become widely available through online tutorials and open courses. Finally, Large Language Models also contribute to the "democratization" of knowledge and technology needed to analyze cultural data. These recent developments ensure a promising future for

1

2

3

data-driven inquiries into the past. Our special issue aims to contribute to this future.

Even though the five papers published in the special issue do not form a complete survey of data-driven historical study, they still give a broad-brush overview of what this new line of study means. The essay by Julie Gravier *et al.* investigates the nineteenth-century development of Paris. To discover how urbanization shaped the history of the French capital, the authors use city directories, and the technology of geocoding along with spatial analysis. Stefanie Schneider offers a deep-learning-supported art-historical study of human body posture on thousands of digitised paintings. Ilona Pikkanen *et al.* harness network analysis and visualization to discover women's letter writing in nineteenth-century Finland; their datasets consist of catalogue records of correspondences from different cultural heritage institutions. The piece by William Turkel and Tim Hitchcock studies the history of crime by analysing the concept of manslaughter in the records of the Old Bailey using language technology. Finally, Ruilin Wang *et al.* write about book history, specifically about headpiece ornaments in tens of thousands of books published in the handpress era of book production (c. 1450–1800). The papers in the special issue come from a variety of fields of historical research and use very different datasets to study the past; they show how data-driven inquiries are reshaping the historical profession.

More precisely, the papers published here demonstrate the new ways historians interact with, analyze, and explore historical sources as large data sets. Using the papers published here as inspiration, in this short introduction I will account for some of these new ways and point out how data-driven inquiries are reshaping the profession of historians. Even though 80 years have elapsed since the beginning of Busa's trail-blazing work, the data-driven study of the past still remains somewhat elusive. For instance, we still lack comprehensive handbooks that define data-driven historical study. It is still unclear what know-how historians willing to pursue data-driven inquiries need. The new body of knowledge that traditional historians expect to see is also unclear. By presenting some of the research practices and analytical concepts underlying this new line (or perhaps emerging sub-field) of historical research, I hope to contribute to its consolidation in the near future.

Data-driven interactions with historical sources

The authors in the special issue work with a great variety of historical sources: images, book ornaments, city directories, proceedings of a criminal court, and letters. Even though these historical sources had all been digitised before the authors started to work with them, the resulting historical data sets still needed to undergo complex transformations so that they could be analyzed in data-driven inquiries. As some of the papers published here demonstrate, the pre-processing steps preceding data analysis are not merely technologised interactions with sources.

Pre-processing a historical data set is also an interpretative practice bridging the gap between epistemological regimes in the past and the present. To give a name to this practice, Gravier *et al.* harness the concept of re-appropriation. The city directories they work with in their paper were originally meant to be used as finding aids and facilitators of business relations in a rich urban environment; the main purpose of the Parisian city directories was to help Parisians find businesses in their city. By contrast, Gravier *et al.* use the Parisian city directories to study the urban development of the French capital. Therefore, the data needed to be re-appropriated and the authors had to bridge a gap between two epistemological regimes. The first one is the conceptual framework under which the original dataset was collected and put together in the nineteenth century. The second one is the modern analytical and conceptual framework that the authors use to study city directories as historical data sets. Re-appropriating city directories as data sets for the study of urban development is challenging because the editors of the Parisian directories had their own nineteenth-century spatial conception of Paris, which they never explained. Pikkanen *et al.* use the practice of repurposing to describe what Gravier *et al.* call reappropriation. They offer further details about the challenges that arise from the epistemological and cultural gaps between the past and present. To investigate female epistolary in nineteenth-century Finland, they use digitized catalogue records from libraries and archives. These catalogue records were created by successive generations of professional and semi-professional archivists, which leads to a key problem. There is "no way of assessing their (the archivists' from the past) inherent biases or distortions, or the representativeness of the collections in relation to nineteenth-century epistolary cultures." The paper by Pikkanen *et al.* thoroughly describes how historians need to take the biases inherent in a historical data set into account when re-appropriating it for contemporary research

purposes. As the authors in the special issue remind us, when confronted with the gap between epistemological regimes in the past and the present, the historian needs to rely on the interpretative practices provided by the field of data criticism and critical archival studies.

The authors also reflect on how to engage critically with historical data. Gravier *et al.* describe the practice of data criticism as an iterative process that is "a constant back-and-forth process between the pipeline chain, data analysis and the original source." This constant "back-and-forth" helps untangle the complex relations between the data and the original historical source it represents. Other authors in the special issue describe critical engagement as a set of practices aimed at assessing the quality of data, including its inconsistencies, ambiguities, and relationships with the original analogue source materials it represents. Stephanie Schneider, whose paper deals with canonical representations of the human body in paintings, argues for the importance of comparing data with the existing canon; in her view, by studying a historical data set along with the canon it belongs to, historians can avoid "reinforcing that canon in a data-driven study." This is also an effective way to address those "inherent biases and distortions" that Pikkanen *et al.* discuss. In short, the papers in this special issue emphasise the importance of contextual knowledge in data-driven historical research. As one of the anonymous reviewers noted, the integration of contextual knowledge into the practice of data criticism can "enrich the understanding of the humanistic changes revealed by the data."

8

The word "enrich" brings us to an example of technologized interactions with historical sources, namely to the practice of data enrichment. In simple terms, through data enrichment, a practice often mentioned in the papers published here, historians connect, enhance and improve their data with external resources and collections. As a result, isolated collections can enter into dialogue with each other and foster opportunities for cross-analysis of fragmented and dispersed pieces of information. For instance, Pikkanen *et al.* - computationally - connect their nineteenth-century correspondence data with biographical information about the authors of the correspondence in external databases. This allows them to map the social background of each person appearing in the correspondence data. Consequently, they can "enrich the data with categories not explicitly present in the archival metadata, like the gender of letter writers and recipients, and observe relations that are simply not detectable in the traditional archive of printed finding aids or organisational archival databases." As part of the enrichment process, Gravier *et al.* add geocoding data to the Parisian city directories they work with. Wang *et al.* use LLMs to enrich their data set and create short descriptions of headpiece ornaments in historical books. The fact that different papers studying different types of historical sources widely rely on data enrichment signals the overall importance of this practice in data-driven historical research.

9

Generally speaking, by now data enrichment has become a computational practice widely accessible for historians and digital humanists. To expand data sets with new pieces of information, historians have a great variety of computational resources at their disposal. The most obvious sources are Wikipedia and other open encyclopaedias. In addition, historians and humanists can work with bibliographic, biographical, and spatial databases to obtain additional information about their subjects. As Wang *et al.* demonstrate, LLMs are also outstanding resources for data enrichment. At the same time, LLMs and other external resources are not free of risk; they can for instance reinforce existing biases in historical datasets. Our authors therefore emphasize the role of human judgement in the assessment of the outcome of computational data enrichment.

10

Altogether, data processing, data enrichment, and critical data study illustrate novel avenues for interacting with historical sources. Nonetheless, as the papers in the special issue highlight, these technologized interactions still require non-technological practices and expertise from the historian embarking on a data-driven study of the past.

11

Large-scale analysis of historical data sets

When Roberto Busa gave an interview in 2001, he underlined how he worked with millions of words in different languages and alphabets. ^[2] One of the motivations for Busa to work with millions of words must have been the same as that for the authors in the special issue. This is the large-scale analysis of historical sources: investigating tens of thousands of documents, images, and words as part of a single large ecosystem.

12

The essays in the special issue extensively discuss large-scale analyses of historical data sets and how this novel

13

analytical approach informs their work. In recent years, some high-level concepts were suggested to describe the large-scale exploration of humanistic datasets. They include influential concepts such as "distant reading" and "distant viewing." Even though some of our authors mention these concepts, for the large-scale analysis of their historical sources they rely on another unexpected concept.

The unexpected concept often empowering data-driven inquiries is space. The authors speak about a great variety of spaces: embedding space, document space, perceptual space, and space of possibilities. How is the concept of space related to data-driven inquiries? How does it support the large-scale analysis of historical data sets?

14

In data-driven research, the concept of space is often unrelated to the conventional three-dimensional geographical space in which we live. Instead, it refers to the concept of feature space used in data and computer science. In non-technical terms, a feature space is populated by "objects" with a predefined set of features. The "objects" populating a feature space can, in fact, be a great variety of entities: artifacts, texts, or even humans. In the context of a feature space, the word "object" is not used literally (to indicate this usage, I enclose the word in apostrophes). The core idea underlying the concept of feature space is that "objects" have quantifiable properties, which can be viewed as dimensions. As a result, the properties that an "object" has determine its place in a feature space. In practice, feature spaces are mathematical constructions; they are high-dimensional vector spaces that are populated by "objects" represented as vectors. For instance, if the "objects" are texts, the features, giving rise to document vectors, are those keywords that they contain. In short, a given feature space populated by a certain set of "objects" is a representation of a complex and large ecosystem to be explored and analyzed computationally. Schneider gives a to-the-point description: the feature-space-based approach enables "researchers to quickly navigate and reduce large data sets to tractable sizes, making possible studies that would otherwise be logistically unfeasible due to the sheer volume of the data involved."

15

What makes the large-scale analysis of "objects" in a feature space possible is that similar "objects" will be in close proximity in the feature space. In other words, spatial proximity in the feature space, measured with tools of vector algebra, translates into structural similarity. Turkel and Hitchcock, who created a feature space from legal records, pinpoint how the spatial approach makes the observation of similarity possible: the "spatial metaphors make it easier to visualize and reason about document similarity and difference. Items that are like one another are positioned nearer in space; those that are less similar are more distant." The methods used to discover groups of similar "objects" in a feature space belong to a broader class of data-driven research practices called clustering. Without going into technical details, this allows the mapping of connections between thousands of different objects. Schneider for instance analyses how depiction of human posture changed over time with cluster analysis. To accomplish a large-scale study of posture, she transforms human figures in thousands of paintings into objects in a feature space. She then discovers clusters of paintings that are similar in terms of the postures of the figures depicted. This leads her to the exploration of the "developmental trajectories of different postures," their variability over the centuries; it also helps her identify non-canonical human body depictions. Wang *et al.* also rely on cluster analysis. They use this analytical tool to identify groups of similar headpiece ornaments in books. As a result, they can identify "more comprehensive sets of related images than by a manual approach alone." As a whole, the spatial framework supports large-scale data analysis by uncovering clusters of similar "objects", as well as the relationships between them. But how does it promote novelty and innovation for data-driven exploration of historical sources?

16

Novelty of data-driven explorations

Papers in the special issue do reflect on these questions; the authors highlight three novel forms of historical explorations supported by data-driven inquiries.

17

First, some of them suggest that data-driven approaches allow historians to pursue a bottom-up exploration of the past. The study by Wang *et al.* explores this avenue. In their study on book history, they show how the data-driven approach enabled them to "adopt an agnostic and large-scale perspective on headpieces, organising them based on the comprehensive data available in ECCO, a source of approximately 200,000 digitised books." By focusing on the data itself, they can effectively challenge previous scholarship. They argue that book ornaments did not necessarily belong to

18

specific printers; instead, they were often exchanged between printers. The data-driven approach thus helps us focus on the sources themselves and set aside the preconceptions and biases inherited from previous scholarship.

Second, some of the authors discuss a dual perspective facilitated by a data-driven approach. As they argue, by focusing on the data, they can investigate both micro and macro phenomena inherent in a data set. For instance, Gravier *et al.* show how the data-driven study of Paris city directories led them to explore long-term and large-scale urbanization processes, as well as micro-level trends in different neighbourhoods. They could thus identify "new places of investigations, from a social and spatial micro-history perspective." Wang *et al.* also highlight the dual perspective defined by the macro and the micro. In their view, data-driven exploration is "a starting point for further investigations, investigations which themselves need a new set of practices and approaches, involving both close and distant 'reading' (...)." The result of the dual approach is a novel dialogue between individual case studies and a larger system that these case studies represent.

19

A third novelty of data-driven explorations is also mentioned in the papers of the special issue. The authors point out that the data itself is a source of theoretical innovation and intellectual stimulation. It forces historians to rethink their concepts and approaches to sources. Pikkanen *et al.* give a thorough description of this novelty: "But perhaps even more important is how the modelling process forces us to think about the 'things' being modelled: what are the catalogued collections of letters, and how can we conceptualise the phenomena, such as epistolary culture, that we are studying with the letter metadata."

20

Finally, we need to mention a fourth novelty provided by data-driven inquiries. When historians adapt this research strategy, they inevitably need to collaborate with computer scientists, statisticians, linguists, and representatives of other fields working with data. Sometimes, as Gravier *et al.* discuss, the collaboration takes place in the form of borrowing methods from other fields. In this sense, data-driven inquiries enhance interdisciplinarity in history and beyond; with one exception, the papers published here are results of collaborative efforts.

21

What next?

In his 2001 interview, Roberto Busa claimed that computational and statistical approaches enable the researcher to go deeper into large textual collections such as the works of Saint Thomas. His claim echoes some of the papers published here that also discuss a deeper meaning to be discovered with a data-driven approach. Nevertheless, the readers will see that the deeper meaning remains an unfulfilled promise. As we have seen, the large-scale study of big data in history innovates and reshapes the craft of the historian; it also offers enhanced access to large collections of documents and artefacts. But all this can hardly be viewed as an expression of deeper meaning, which explains why the data-driven study of the past is still elusive.

22

As the authors themselves recognize, to solidify the data-driven study of the past, much more research is needed. In my own opinion, to find the deeper meaning in data, the recent technological innovations need to be complemented with epistemological advancements. Specifically, I believe that epistemological advancements are needed in two areas.

23

The first one is the domain of concepts. Even though new concepts arise with the data-driven approach, these concepts need to be made compatible with the agenda of historical, and more broadly speaking humanistic research. Take for instance the concept of feature space explained above. What does the discovery of thousands of documents represented as "objects" in a feature space say about human experience and culture?

24

The second domain where advancement is needed is related to the question of evidence. The overall goal of historical scholarship is to build a solid body of knowledge about the past, the cornerstone of which is the process of gathering evidence based on the sources we have. How do data-driven inquiries enable the development of historical evidence? What types of historical evidence are we hoping for when studying large collections with the help of data science? To answer these questions, we need not only new methods and concepts but also an epistemological foundation of data-driven historical research.

25

Acknowledgements

Notes

[1] A modern edition of the Index Thomisticus is available on <https://www.corpusthomicum.org/it/> (last accessed, 12 January, 2026).

[2] "I did not work only on the words of Saint Thomas but also on millions of other words in 22 other languages and different alphabets: the Qumran Caves Scrolls, which are in the Hebrew alphabet (also in three languages: Hebrew, Aramaic, Nabataean), the entire Quran in Arabic, Finnish, Cyrillic, Bohemian, and obviously all the European languages and Greek." Translated by Gabor Mihaly Toth "Io non ho lavorato solo le parole di san Tommaso, ma altrettanti milioni di parole in altre 22 lingue e alfabeti diversi: i rotoli di Qumrân, che sono in alfabeto ebraico ma in tre lingue (ebraico, aramaico e nabateo), tutto il Corano in arabo, il finnico, il cirillico, il boemo e ovviamente tutte le lingue europee e il greco" <https://www.ilsussidiario.net/news/emmeciquadro/emmeciquadro-n-13/2001/12/20/scienza-in-atto-il-computer-usato-bene-intervista-a-roberto-busa-s-j/397260/> (last accessed, 14 January 2026).

Works Cited

Recommendations

DHQ is testing out three new article recommendation methods! Please explore the links below to find articles that are related in different ways to the one you just read. We are interested in how these methods work for readers—if you would like to share feedback with us, please complete our short evaluation survey. You can also visit our documentation for these recommendation methods to learn more.

SPECTER Recommendations

Below are article recommendations generated by the SPECTER model:

1. Unknowable Facts and Digital Databases: Reflections on the Women Film Pioneers Project and Women in Film History, 2020, Sarah-Mai Dang, Institute of Media Studies, Philipps-Universität Marburg
2. A Macroscopic for Global History: Seshat Global History Databank, a methodological overview, 2016, Pieter François, University of Hertfordshire, University of Oxford; J.G. Manning, Yale University; Harvey Whitehouse, University of Oxford; Rob Brennan, Trinity College Dublin; Thomas Currie, University of Exeter, Penryn Campus; Kevin Feeney, Trinity College Dublin; Peter Turchin, University of Connecticut
3. A Critical Collection History of Nineteenth-century Women's Letters: Overcoming the Occluded Archive with Data-Driven Methods, 2026, Ilona Pikkanen, The Finnish Literature Society; Matti La Mela, Uppsala University; Hanna-Leena Paloposki, Independent Scholar; Jouni Tuominen, University of Helsinki and Aalto University
4. The Age Old Question: A Review of What is Digital History? by Hannu Salmi, 2021, Tracy L. Barnett, University of Georgia
5. Building the Women in Book History Bibliography, or Digital Enumerative Bibliography as Preservation of Feminist Labor., 2019, Cait Coker, University of Illinois at Urbana-Champaign; Kate Ozment, California State Polytechnic University, Pomona

DHQ Keyword Recommendations

Below are article recommendations generated by DHQ Keywords:

1. Lost Spaces, Lost Technologies, and Lost People: Online History Projects Seek to Recover LGBTQ+ Spatial Histories, 2020, Alex D. Ketchum, McGill University
2. Networked Cross-Dressing: A Digital Refashioning of Shakespearean Gender Subversion, 2022, Erik Simpson, Grinnell College; Hannah L.P. Brown, Department of Psychology, University of Florida; Lana Sabb, Grinnell College; Olly Shortell, Prescott College; James Lee, University of Cincinnati
3. Studying Large-Scale Behavioral Differences in Auschwitz-Birkenau with Simulation of Gendered Narratives, 2022, Gábor Mihály Tóth, USC Shoah Foundation / USC Viterbi School of Engineering; Tim Hempel, Freie Universität Berlin, Department of Mathematics and Computer Science / Department of Physics; Krishna Somandepalli, Google Research; Shri Narayanan, USC Viterbi School of Engineering
4. Ticha: Collaboration with Indigenous communities to build digital resources on Zapotec language and history, 2020, George Aaron Broadwell, University of Florida; Moisés García Guzmán, pueblo of San Jerónimo Tlacoahuaya; Brook Danielle Lillehaugen, Haverford College; Felipe H. Lopez, Haverford College Libraries; May Helena Plumb, University of Texas at Austin; Mike Zarafonitis, Haverford College Libraries
5. Agent-Based Modeling and Historical Simulation, 2014, Michael Gavin, University of South Carolina

TF-IDF Recommendations

Below are article recommendations generated by the TF-IDF Model:

1. Evaluating and Understanding the Geocoding of City Directories of Paris (1787-1914): Data-Driven Geography of Urban Sprawl and Densification, 2026, Julie Gravier, Laboratoire ThéMA UMR 6049, CNRS, Université Marie et Louis Pasteur; Stéphane Baciocchi, Centre de Recherches Historiques, EHESS-CNRS UMR 8558; Pascal Cristofoli, Centre de Recherches Historiques, EHESS-CNRS UMR 8558; Bertrand Duménieu, Centre de Recherches Historiques, EHESS-CNRS UMR 8558; Edwin Carlinet, Laboratoire de Recherche de l'EPITA; Joseph Chazalon, Laboratoire de Recherche de l'EPITA; Nathalie Abadie, Université Gustave Eiffel, ENSG, IGN, LASTIG; Solenn Tual, Université Gustave Eiffel, ENSG, IGN, LASTIG; Julien Perret, Université Gustave Eiffel, ENSG, IGN, LASTIG
2. Image Reuse in Eighteenth-Century Book History: Large-Scale Data-Driven Study of Headpiece Ornament Variants, 2026, Ruilin Wang, University of Helsinki; Lidia Pivovatova, University of Helsinki; Yann Ryan, Digital Humanities and postdoctoral researcher, Leiden University; Enes Yilandiloğlu, University of Helsinki; Mikko Tolonen, University of Helsinki
3. A Critical Collection History of Nineteenth-century Women's Letters: Overcoming the Occluded Archive with Data-Driven Methods, 2026, Ilona Pikkanen, The Finnish Literature Society; Matti La Mela, Uppsala University; Hanna-Leena Paloposki, Independent Scholar; Jouni Tuominen, University of Helsinki and Aalto University
4. Facets of Friction: Investigating epistemological friction between computing and the humanities to support Digital Humanities computing education, 2026, Anna Sollazzo, University of Glasgow
5. Categorising Legal Records – Deductive, Pragmatic, and Computational Strategies, 2023, Marlene Ernst, University of Passau; Sebastian Gassner, University of Passau; Markus Gerstmeier, University of Passau; Malte Rehbein, University of Passau



This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.