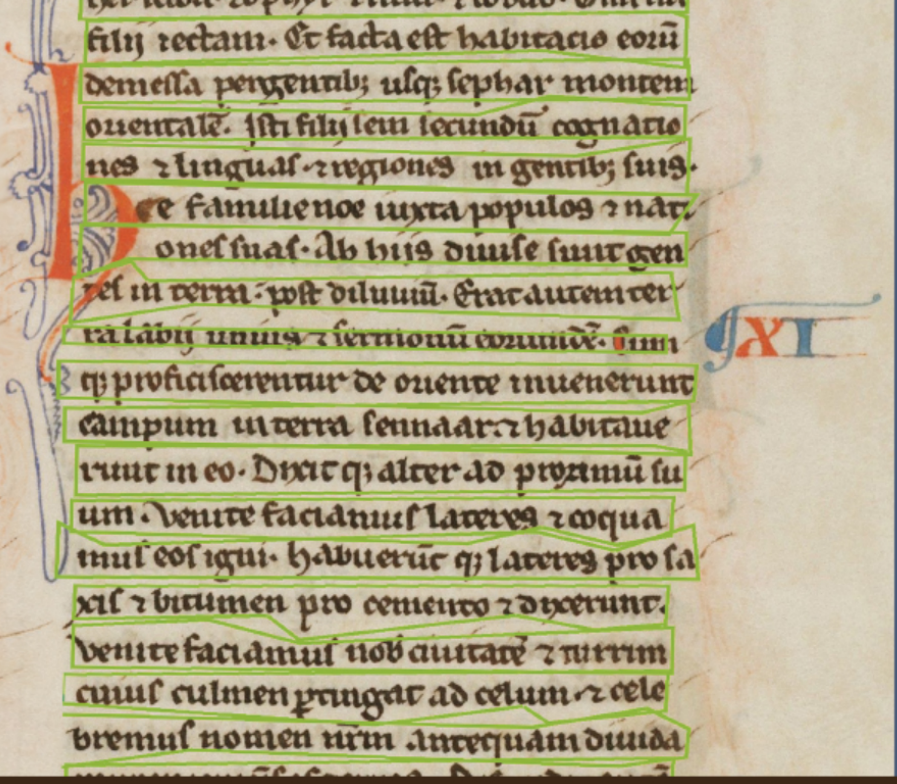


Book Cultures, Medieval to Modern

MEDIEVAL MANUSCRIPTS AND THE COMPUTATIONAL HUMANITIES

BIG DATA, SCRIBES, AND THE "PARIS BIBLE"

by DAVID JOSEPH WRISLEY
and ESTELLE GUÉVILLE

ARC HUMANITIES PRESS

BOOK CULTURES, MEDIEVAL TO MODERN

Further Information and Publications
www.arc-humanities.org/series/book-series/

MEDIEVAL MANUSCRIPTS AND THE COMPUTATIONAL HUMANITIES

BIG DATA, SCRIBES, AND THE "PARIS BIBLE"

by

DAVID JOSEPH WRISLEY
and **ESTELLE GUÉVILLE**

ARC HUMANITIES PRESS

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library.

ISBN (Hardback): 9781802702439
ISBN (Paperback): 9781802704501
e-ISBN (PDF): 9781802704488
e-ISBN (ePUB): 9781802704495

DOI: 10.17302/mfsk1586

www.arc-humanities.org

Printed and bound in the UK (by CPIGroup [UK] Ltd), USA (by Bookmasters), and elsewhere using print-on-demand technology.

Publisher (manufacturer) details: Arc Humanities Press, 14 Clifton Moor Business Village, James Nicolson Link, York YO30 4XG, United Kingdom.

EU Authorized Representative details (for GPSR purposes): Amsterdam University Press, Nieuwe Prinsengracht 89, 1018 VR Amsterdam, The Netherlands. www.aup.nl

ARC HUMANITIES PRESS

CONTENTS

DOI: 10.17302/mfsk15861

LIST OF ILLUSTRATIONS

Figures

DOI: 10.17302/mfsk15862

Tables

ACKNOWLEDGEMENTS

THIS BOOK EMERGED from a collaborative effort made possible by the intellectual generosity and shared commitment of many individuals and institutions. First, since this co-authored book is the fruit of many years of sustained collaboration, we would like to thank each other for the engagement and enthusiasm with which we each approached this project.

We would also like to acknowledge the support of our current and past respective employers, Louvre Abu Dhabi, New York University Abu Dhabi, and Yale University, whose infrastructure and community provided resources and space for the reflection that gave rise to our work. This project was generously supported by research funds at New York University Abu Dhabi for infrastructure, manuscript digitization, a sabbatical that provided the time, travel, and data work required for our research, and funded research assistant positions. We are grateful to Niccolò Acram Cappelletto, Alice Fournier, Amanda Robin Hemmons, and Joshua Isaac, who served as our research assistants and for their spirit of dialogue that shaped every stage of this work. Open access publication of this book was also supported by institutional funding from New York University Abu Dhabi.

To the teams behind the handwritten text recognition software we used, Transkribus, we are grateful for the open, iterative ethos of digital scholarship that made this work possible. We relied on the sustained discussions with members of the Transkribus community: Nora Barakat, Tobias Hodel, Suphan Kirmizialtin, Annemieke Romein, and Melissa Terras. We are especially indebted to the librarians and curators at New York University Abu Dhabi and Yale University, as well as to Matthew Heintzelman at the Hill Museum and Manuscript Library and Charlotte Denoël at the Bibliothèque nationale de France, whose expertise in manuscript collections, digital collections, and preservation enriched both our analysis and our methods.

We would like to extend a special thanks to Frédéric Spagnoli of the Université Marie-et-Louis Pasteur in Besançon, France, and all the students from the Master's in "Rare Books and Digital Humanities" and "Editions Numériques et Patrimoine de l'Antiquité" who participated in the Paris Bible Correct-a-Thon we organized in January 2023: Robert Lloyd, Gauri Bhagwat, Alice Fournier, Amanda Robin Hemmons, Alexandre Keyes, Lucia Sol

Bezzecchi Petroff, Marie Noirot, Anna Chemisova, Benedicta Arthur, Sharon Hassive Guerra Álvarez, Nina Jacobson, Sumeyye Topkara, Sonaj Kailas, Kateri Soulard, Jesus David Macchi Franco, Elia Coulot, Serhat Acar, Úna Faller, and Diego Rodriguez. We would also like to thank Pierre-Emmanuel Guilleray of the Bibliothèque d'étude et de conservation at the Bibliothèque municipale de Besançon, France, for opening the library's collections to our project and hosting our participants.

We would also like to acknowledge Laura Morreale for spearheading transcription challenges in the medievalist community. Our experience participating in them helped shape our own challenge and supported our thinking which resulted in the fourth chapter of this book.

We would like to thank our editors for their patience and skill in ushering the manuscript through to completion as well as the reviewers of the manuscript for their generous comments.

David would like to thank Jeff Richards and Ray Siemens, for their constant encouragement and intellectual exchange. Estelle would like to thank Jessica Brantley, Alex Gil, and Agnieszka Rec for their generous guidance, patience, and support, and for the freedom to pursue intellectual paths extending well beyond her dissertation.

Finally, we thank our partners and communities for their patience and encouragement during the long hours of writing, coding, and revising this manuscript.

ABBREVIATIONS

Aarau KB	Aarau, Aargau Kantonsbibliothek
BAV	Città del Vaticano, Biblioteca Apostolica Vaticana
Beinecke	New Haven, Yale University, Beinecke Library
BL	London, British Library
BnF	Paris, Bibliothèque nationale de France
BnP	Lisboa, Biblioteca nacional de Portugal
CCC	Cambridge University, Corpus Christi College
Cologny	Cologny, Fondation Martin Bodmer
DigiVatLib	Digitized Collections of the Biblioteca Apostolica Vaticana
Free Library	Philadelphia, Free Library of Philadelphia
HMML	The Hill Museum and Manuscript Library
HTR	handwritten text recognition
Lambeth	London, Lambeth Palace
LAD	Abu Dhabi, Louvre Abu Dhabi
Mazarine	Paris, Institut de France, Bibliothèque Mazarine
NT	New Testament
OT	Old Testament
Sarnen KB	Sarnen, Kollegiumsbibliothek, Stiftsarchiv Muri-Gries
Schaffhausen MB	Schaffhausen, Ministerialbibliothek
St. Gallen KB	St. Gallen, Kantonsbibliothek
Stegmüller	Stegmüller, *Repertorium biblicum medii aevi*, Vol. 1
UPenn	Philadelphia, University of Pennsylvania Libraries

DOI: 10.17302/mfsk15864
ARC HUMANITIES PRESS

INTRODUCTION

THIS BOOK IS about the so-called Paris bible—both a material and textual form of the Vulgate Bible—which flourished in Western Europe during the thirteenth and fourteenth centuries. Researchers have identified almost two thousand examples in global libraries and museums, although the full extent of the Paris bible corpus has not been documented.[1] Paris bibles have been studied by codicologists, art historians, and historians, but most attention has been paid to their decoration and illumination, as one way of localizing or dating them, as well as their annotations for the study of readership, provenance, production, correction, and use.[2] These codices have also been studied from the perspective of theories of the miniaturization of biblical manuscripts for use in specific social and religious contexts.[3] The evidence used in quantitative codicological studies has largely employed metadata from archival catalogues and on site assessment.

This book is not the result of a premeditated plan, however, to retell the history of the Paris bible, but emerges from a combination of circumstances. Had we not both been living in the United Arab Emirates, working on the same island for two different institutions a few kilometres apart—New York University Abu Dhabi and the Louvre Abu Dhabi—we may have never met. Saadiyat Island hosts a number of cultural and educational institutions, including a full degree-granting campus of New York University where one of this book's authors teaches digital humanities. The island also hosts the first universal museum in the Arab world where the other author previously worked as an assistant curator and researcher. Among the many items collected by the Louvre Abu Dhabi, a Paris bible in two volumes, Abu Dhabi, Louvre Abu Dhabi, 2013.051 (hereafter LAD, MS 2013.051), was acquired in 2013. In 2020, the planning of an exhibition about sacred books in the Abrahamic traditions brought this manuscript to our attention. Our decision to launch a digital project exploring Paris bibles, using the Louvre Abu

1 Ruzzier, *Entre université et ordres mendiants.*

2 Linde, *How to Correct the Sacra Scriptura?*; Magrini, "Production and Use of Latin Bible Manuscripts."

3 Light, "What Was a Bible For?"; Ruzzier, "The Miniaturisation of Bible Manuscripts."

DOI: 10.17302/mfsk15865
ARC HUMANITIES PRESS

Dhabi codices and other manuscripts spread throughout global collections, was shaped as much by our intellectual interests as medievalists, as by a desire to model a form of digital scholarship between the university and the museum in a place where such collaboration was not common.[4] Seen from this perspective, our book also offers reflections on the kinds of work that can be done as medievalists using primarily manuscript surrogates, working both outside of large European capitals and between national paradigms for digital research. In the end, our interest in scaling transcription and using computational textual analysis emerged not from a prior plan, but rather from the convergence of content, context, and methodological curiosity.

Paris bibles have long held a privileged position in manuscript collections across time and space, collected by historical entities as varied as the Ottoman Sultan Abdulhamid II, the American philanthropist and lawyer John Frederick Lewis, and the Louvre Abu Dhabi.[5] Important museums, libraries, and private collections often have at least one copy of a Paris bible and sometimes several of different origin: English, French, and Italian. Paris bibles are attractive for collections today, since they are numerous, illuminated, and an example of a core book in the history of civilization. Moreover, the rare book market has a limited inventory of pre-fifteenth century illuminated manuscripts. Paris bibles, mainly dating from the thirteenth century are some of the few earlier illuminated manuscripts that appear in dealers' catalogues, making them a must-buy for collecting institutions.

Despite this rich history of collection and circulation, the text itself of the Paris bibles rarely sparks scholarly interest. Although there is much to be said about scriptural variability across codices in the period we are studying, again, our work does not address the question. Instead, our starting point that predated the Covid-19 pandemic by some months was an experiment in training a computer to transcribe French and Latin scripts from digitized manuscripts. LAD, MS 2013.051 had been fully digitized and was available to the in-house researcher so, when the pandemic set in, we had an unusual opportunity for deep, exploratory work, as well as the time to devote to the initial investment required to begin training handwritten text recognition (HTR) models.

---

4 Guéville and Wrisley, *Paris Bible Project*; Wrisley, "Enacting Open Scholarship"; Wrisley, Guéville, and Cappelletto, "Creating New Audiences."

5 Respectively, Budapest, Eötvös Loránd Tudományegyetem Könyvtára, MS Cod. lat. 18; Philadelphia, Free Library of Philadelphia (hereafter Free Library), MS Lewis E242; and LAD, MS 2013.051.

This book is also a meditation on the current state and future of the computational humanities, its promises and its challenges, from the specific vantage point of medieval manuscripts and practices of automated transcription. It does propose some analysis (notably in Chapter 3) that will be useful in shedding light on unresolved questions about some Paris bibles, and perhaps will model how analysis of other digitized manuscripts from other periods or in other languages might be done, but it is not first and foremost a technical discussion of those methods. In the initial analysis and training of an HTR model based on the single biblical manuscript, we were able to devise a customized transcription scheme that we describe in Chapter 2. In turn, the project evolved, as research often does, following the digitized record. We balanced our research questions with what we were able to find and use remotely. In this way, our work sits at the intersection of what we wanted to ask and what we were able to access.

The Digital and the Computational

Just as this book is not a new history of the Paris bible, it is also important to say that it is not a general overview of computational approaches in the field of medieval manuscripts. So many exciting pathways have emerged in recent years for working with manuscripts and new technologies, that it would be impossible to cover them with critical depth in the space that a short book can afford. Some of these methods include 3D and multispectral imaging to study inks or palimpsests, bio-codicology to detect the kinds of animal skin used to put together a codex, computational approaches to modelling collation, unseen species models for assessing manuscript production, linked open data to connect collection and provenance metadata, computer vision to tackle problems such as keyword spotting in large digitized corpora, image embeddings to study patterns in illumination corpora, and HTR to automate the transcription of documents copied by scribes. This book largely focuses on the latter, an approach that allows—with significant training and human input—for transcription to be automated from digitized copies of manuscripts. Whereas a full discussion of medieval manuscripts and the computational humanities would be necessarily more inclusive of a variety of emergent methods, we hope that this book can nonetheless open up fruitful avenues for discussion about how the stakes of computational methods and, in particular, machine learning (ML) and artificial intelligence (AI), are changing—and challenging—the field of medieval studies.

The field of medieval studies has undergone notable transformation in recent decades, with a growing number of scholars working at the intersection of the humanities and digital or computational methods. To begin, it is important to clarify what we mean by "the digital" and "the computational," and how these terms relate to one another. Although often used interchangeably, they are, strictly speaking, not synonymous. Broadly, "the digital" encompasses scholarly activities that involve digital technologies or digitized formats. This includes digitized content such as scanned or photographed manuscripts, digital images of illuminations, and a range of file formats—from portable network graphics (PNG) and the tagged image file format (TIFF) to JavaScript Object Notation (JSON) and the portable document format (PDF)—as well as the tools and platforms that facilitate their use. Examples include websites, databases, viewers for the International Image Interoperability Framework (IIIF), texts encoded according to the guidelines of the Text Encoding Initiative (TEI), and editions, as well as what may be described as digital workflows, such as cataloguing, metadata creation, and the curation of digital exhibitions. The digital, then, primarily concerns questions of modelling, representation, and interfaces—that is, how we remediate, encode, and navigate (often born analogue, but not always) cultural materials in virtual formats.

By contrast, we use the term "computational" to refer to the use of algorithms, programming, and quantitative methods such as statistical analysis to explore and interpret data. This term encompasses a wide array of techniques and approaches, including natural language processing (NLP), ML and AI, text mining, network analysis, and various forms of data visualization. Unlike the digital, which emphasizes access and representation, the computational approach involves active data processing or transformation, often—but not necessarily—at scale. Computational methods frequently require customized software solutions or the use of scripting languages and algorithmic models tailored to specific research questions. In many cases, the computational can be considered a subset of the digital, since computational analysis depends on digital data—whether textual or visual, structured or unstructured—and takes place within digital environments. However, it is important to note the distinction in emphasis: while the digital is often concerned with preservation, format, and interface (e.g., digitizing a manuscript and making it accessible online), the computational focus is on interpretation and transformation, applying code or mathematical models to generate new insights and meanings.

This distinction has been further theorized by Gil, who differentiates between "Architectures of Knowledge" and "Algorithmic Approaches to the

Humanities."[6] The former refers to the construction of a digital scholarly infrastructure—such as digital libraries, archives, editions, and databases—that reshapes the cultural human record into a networked ecosystem of sources. If one thinks about digital humanities in phases, the architectures of knowledge represent a first approach, focusing on the digitization of collections and the creation of environments for humanistic inquiry. The development of algorithmic approaches comes in a second phase involving the application of computational methods to analyze, interpret, or classify digitized cultural objects in the search of patterns. The latter might be seen through what Berry has called a "computational turn," in which epistemological shifts arise from digital mediation—transforming not only our objects of study, but also the very conditions of scholarly reasoning. This turn transforms the nature of humanities research, lending it what Berry calls "computationality," a reconfiguration of humanistic knowledge, increasingly structured and understood through computational systems, privileging modalities such as the discrete, the encoded, and the algorithmic.[7]

The digital has always harboured the potential for computational transformation, a potential that is becoming increasingly realized as infrastructures, tools, and methodologies evolve. While the digital has long been associated with modes of representation, preservation, and access—such as TEI XML for richly marked-up texts or IIIF for standardized image delivery—these same frameworks also provide the structured data and interfaces necessary for computational modelling and processing. TEI, for example, not only enables the creation of digital editions but also supplies machine-readable corpora for natural language processing (NLP). Platforms like FromThePage or Zooniverse, initially developed to support public transcription and annotation, now also serve as sites from which human inputs are leveraged to train and refine machine learning models. Geographic Information Systems (GIS), once primarily used for analyzing and visualizing spatial data, are used for increasingly specific geospatial processing tasks. Similarly, while IIIF is often considered as a mere delivery mechanism for images, it is also becoming a conduit for computational enhancement—supporting tasks such as image segmentation, feature recognition, and annotation pipelines. In short, the boundary between the digital and the computational is not only porous, but increasingly busy: digital infrastructures are becoming computational platforms, and their potential for analysis and interpretation is expanding accordingly.

6 XP Method, "Architectures of Knowledge."

7 Berry, "The 'Computational Turn.'"

Handwritten Text Recognition (HTR) and Medieval Studies

The digitization of medieval bibles has unfolded alongside evolving scholarly and archival cultures of digitization since the late twentieth century. In Chapter 2, we illustrate how working with digitized manuscripts—or, more accurately, transcriptions derived from them—can generate new layers of data that allow us to examine the role of scribes in the production of Paris bibles. While many of the distant reading and computational approaches in literary studies have relied on precompiled corpora of edited sources or digitized versions of printed materials, we shift the focus to transcribing directly from digitized medieval manuscripts.[8] Our approach is informed by emerging computational work that uses HTR-generated transcriptions, both within and beyond the field of medieval studies.[9] We explore how corpus-building through automated transcription remains an imperfect process, shaped by the evolving standards and technologies of digitization.[10] Yet despite the challenges posed by the materiality and variability of manuscripts, we argue that HTR ushers in exciting possibilities for engaging with this broader archive—offering evidence that enriches our ability to describe and interpret complex features of medieval scribal culture.

Developments in HTR mark a paradigm shift, both for us as scholars and for medieval studies in general, as it is fundamentally transforming both the scale and nature of the questions we can ask. No longer limited to large institutions, HTR has become increasingly accessible—empowering small research teams and individual scholars to train, refine, and adapt bespoke models (although we will discuss limitations of access in Chapters 2 and 4). What was once prohibitively labour-intensive and inconsistent—the process of transcribing texts from manuscript—can now be achieved within a research group with concerted effort in a matter of months, enabling new kinds of inquiry into texts and at new scales. HTR can also support the generation of multiple layers of text representation—diplomatic, semi-diplomatic, and normalized—offering a more textured basis for philological analysis. These models can be tested, iteratively improved, and within the span

8 Bode, *A World of Fiction*; Cronk and Roe, *Voltaire's Correspondence*; Piper, *Enumerations*; Underwood, *Distant Horizons*; van Dalen-Oskam, *The Riddle of Literary Quality*.

9 Massot et al., "Transcribing Foucault's Handwriting"; Kirmizialtin and Wrisley, "Exploring Gulf Manumission Documents"; Franzini et al., "Attributing Authorship"; Vandyck and Kestemont, "Abbreviation Application."

10 Salmi, *What is Digital History?*

of our project, we have observed significant advances in both quality and adaptability.

HTR represented a leap in method, allowing us to bridge traditions of close reading and philology with the scaled ambitions of computational research, opening up possibilities that were once beyond reach. The Paris bible corpus poses a particular editorial challenge discussed in Chapter 2, on account of the high degree of orthographic and scribal variance between copies, reflecting complex historical processes of co-creation and geographical diffusion. HTR has enabled us to engage with that complexity directly, rather than flattening it, allowing us to observe textual phenomena without sacrificing details previously almost impossible to study before. The question is no longer who is better at transcription—the machine or the human—but what can be achieved more rapidly with a similar accuracy and how research time is best spent. Human labour does not disappear from this scenario, rather it is channelled toward different, specialized tasks: research design, ground-truth creation, quality assessment, and computational analysis. It must be said that HTR has enabled the productive convergence between medieval studies and computational humanities, and that it marks a decisive shift, freeing digital scholarship from its dependence on printed critical editions and enabling direct analysis of manuscript evidence.

Epistemic and Social Implications of Computational Methods in Scholarship

Building on the idea of an increasing entanglement between the digital and the computational, it is now possible to speak of the creation of text corpora not simply as a digital editorial task carried out by humans, but as an increasingly computationally rich process that, depending on the importance of absolute accuracy, renders traditional editing of historical documents somewhat obsolete.[11] Workflows that use AI are not only leveraged to carry out editorial tasks such as creating transcriptions, but can also perform tasks from NLP, such as named entity recognition (NER). By provocatively asking if the "end of the edition" is upon us, Hodel is, of course, not declaring the death of the editor. Instead, he is arguing for a transformation of contemporary historical praxis by embracing its "computationality," to use Berry's term. This transformation does not eliminate the role of the editor, but repositions it within a broader set of automated and semi-automated

11 Hodel, "Das Ende der Edition?"

processes. The historical edition, like the diplomatic transcription of a medieval manuscript, is no longer conceived as a static, final product, but as a step of data preparation within a longer path of computer-assisted research. For the mass of early modern source materials (the period Hodel is writing about) editorial labour is one research task among many interconnected ones: structuring, linking, and preparing historical data for analysis and interpretation. We add our voices to this broader shift described by Hodel, advocating for a reconceptualization of the editorial work of transcription as a data-rich intervention enabling new forms of critical inquiry.

Important studies integrating digital and computational approaches are underway for many sources bases from the medieval period—manuscripts, charters, coins, seals, maps. The ability to process manuscripts in an automated fashion is already a mature technology with a diversity of approaches and considerable theoretical debate. No one, to our knowledge, argues that technologies of HTR displace human labour altogether in the study of book cultures. As we will discuss in the chapters of this book, HTR extends human labour—automating certain tasks through innovative workflows and equipping scholars with new lenses through which to reconsider both familiar and previously unexamined aspects of our source material.

Yet in medieval studies—as in other fields—our ability to pursue such algorithmic approaches remains constrained by the unevenness of both the digitized record and research infrastructure. For questions such as the development and dissemination of Paris bibles, we lack a comprehensive corpus that could support large-scale comparison, modelling, or distant reading. A significant portion of relevant archival material remains undigitized or inconsistently described, limiting the potential for computational methods to contribute to interpretation. As we demonstrate in Chapter 3, computational approaches can serve as powerful heuristics, but they should never be mistaken for neutral findings about manuscript realities. That is to say, the results of such approaches need to be considered for what they might help us to understand, without exaggerating their epistemic claims.

Another important dimension of the computational within the digital humanities consists in the socio-technical and networked dimensions of such scholarship. This notion includes infrastructures of collaboration—crowdsourcing platforms, open peer review platforms, version-controlled repositories, digital research infrastructures, and commons for data sharing—that reconfigure how scholars and publics engage with knowledge. Berry suggests that this movement toward collaborative, code-mediated research reflects a deeper transformation: a shift from seeing computation only as a tool, to understanding it as the medium that shapes the production

and transmission of knowledge.[12] Tilton, Mimno, and Johnson add an important ethical dimension to this turn in the humanities by emphasizing that computational humanities are not merely about applying quantitative methods to humanistic data and turning to collective platforms, but about reimagining what counts as data, whose labour supports its production, what kinds of infrastructures make such work possible, and for whom.[13] They reject hard boundaries between close and distant reading, qualitative and quantitative methods, and academic versus community labour.

We discuss these issues at more length in Chapter 4, emphasizing that in medieval studies, as in many other fields, the infrastructures that shape digital scholarship are far from neutral. What gets digitized, who gains access, and which projects receive institutional backing are shaped by both power and disciplinary tradition. We have been fortunate that a handful of manuscript-holding institutions—such as the Bibliothèque nationale de France, the University of Pennsylvania Libraries, the Biblioteca Apostolica Vaticana and the Free Library of Philadelphia (hereafter BnF, UPenn, BAV and Free Library)—have digitized many of their manuscripts and made them openly accessible. Their efforts have enabled new forms of scholarly inquiry such as our analysis in Chapters 2 and 3 that would otherwise be impossible. Yet so many other collections remain closed, semi-closed, or altogether inaccessible. Questions of what we call "collections bias" loom over the materials we are able to locate and reuse, as not every cultural institution has equally constituted and disseminated its digital collections. This asymmetry in access mirrors broader structural inequalities in the field where technical capacity, funding, and digital infrastructure are unevenly distributed amongst institutions.

There have been important advocates of open scholarship, fighting for transparency, reproducibility, and equitable participation within digital humanities, but these perspectives are not without critique. Openness is not a neutral value, as it can shift burdens onto precarious scholars, replicate colonial dynamics by placing cultural heritage into open infrastructures without engaging the scholars who created it or the communities to whom it belongs.[14] For medievalists, the nature of the collections and the datasets derived from them poses particular challenges. Since data can be provisional or incomplete, and since it is always shaped by scholarly judgement,

12 Berry, "The 'Computational Turn.'"

13 Tilton et al., "What Gets Counted."

14 Risam, *New Digital Worlds*.

treating such data as open and reusable without sufficient contextualization, there is always the risk of it being considered authoritative by others unfamiliar with the scholarly limitations, or by non-human agents that cannot understand context.[15] We will discuss openness and its limitations more extensively in Chapter 4.

In conclusion, the four chapters that follow explore the evolving relationship between medieval studies and computational approaches, especially as shaped by ML and AI against a backdrop of increased collaboration between medievalists. We begin by revisiting longstanding editorial and interpretive challenges faced by medievalists when dealing with texts in manuscript, and we ask how new tools might recalibrate—rather than replace—traditional modes of scholarly attention. Far from treating distance and scale as neutral benefits of computation, we describe how when working with a corpus, one quickly encounters material constraints. In this book, scale is not first and foremost a matter of the number of manuscripts consulted and processed, but instead of the large number of micro-features—abbreviations, letterforms, brevigraphs, punctuation habits—that scribes leave in the fabric of texts, and which computational analytical techniques are now being designed to capture and to analyze. As such, ML and AI do not sever us from the materiality of manuscripts; instead, they offer new ways to engage with their "handmadeness," especially at the level of scribal orthography. In Chapter 3, we present four case studies that combine HTR-based transcription with stylometric analysis across a dozen manuscripts, assessing questions of authorship, scribal attribution, and collective (medieval) labour. But as D'Ignazio and Klein remind us, data cannot be disentangled from the social and political contexts of its creation.[16] The data we generate from medieval manuscripts, and the platforms we use to create and analyze them, must be situated within a critical appraisal of the wider transformations brought on by AI across society and academe—including shifts in labour and authority. We conclude by returning to the question of how interdisciplinary collaboration and digital infrastructures might shape the future of medieval studies—not simply by expanding what individual scholars can do, but by rethinking what kinds of work, and whose work, count within a rapidly changing research landscape.

15 Flanders and Jannidis, *The Shape of Data*.

16 D'Ignazio and Klein, *Data Feminism*.

Chapter 1

DO THE HUMANITIES DREAM OF PATTERNS?

IN THE THIRTEENTH and fourteenth centuries, a version of a Latin bible emerged, commonly known as a "Paris bible." Whereas the scholarly literature on Paris bibles often underscores their similarity, perhaps even their uniformity, the text found in them was hardly a stable copy of the Vulgate. In this first chapter we discuss the creation of thousands of these book-objects and different kinds of variance—textual and material—found among the corpus. Indeed, Paris bibles varied in many ways: the size of the codices, the size of the textblock, the order of the books, the placement and choice of prologues, the number of hands present in the codex, the presence or absence of catchwords, or the number of lines per column. Such variance is to be expected in a culture of handmade books.

The overwhelming degree of variance in Paris bibles involves the scribal contribution to the manuscript. When you examine the same section of two Paris bibles side by side, to the trained eye the differences in medieval Latin are distinctive. When the conventions of scribes—a wide variety of abbreviations, brevigraphs, and pre-modern letterforms—are eliminated, that is, expanded into the unabbreviated and normalized Latin, the text is much the same from codex to codex, with some notable exceptions. Paris bibles also exhibit variance in word order and content: eye-skip, slight differences when copying from memory and even echoes of the pre-Vulgate text (*Vetus Latina*) found in some manuscripts. One of the major questions we ask in this book is how a machine—accustomed to matching strings and counting words—could process and compare non-normalized transcriptions made from these book objects. Indeed, the complex entanglement of textual and material difference lends the corpus a particular allure for the contemporary fields of material philology, quantitative codicology, distant reading, and digital humanities.

We argue that the study of medieval production of bible manuscripts needs to address these complex forms of variance, with both pragmatism and rigour. Codicologists might study collation or bindings, while palaeographers would study variation in pen strokes and the shape of letters, two highly specialized methods of understanding the complex context of a codex's fabrication. Both approaches could be, and have been, modelled

DOI: 10.17302/mfsk15866

computationally. Adopting a form of "software intensive humanities,"[1] we build a corpus of transcriptions from Paris bibles using HTR, and we consider our method a *computationally enhanced material philology*.[2]

Our method is capable of capturing certain, but not all, features of scribal behaviour. With the resulting transcriptions we use computational methods to see if we are able to discover patterns in the transcribed text. With contemporary HTR, the well-trained algorithms are certainly able to outperform human transcription in terms of speed, but they perform with differing degrees of accuracy, especially if we move across different types of sources. While machine transcribed text is not perfect, it is worth noting that human-created transcription (especially diplomatic transcription) is not flawless either, and so the practice of using technology to do the work of transcription with far greater speed and with comparable accuracy to human work is not only our choice, but it is gradually becoming a practical method adopted by many scholars.

Digitization of manuscripts is a necessary starting point for this research and the kinds of questions that one is able to ask of a digitized collection are necessarily different from those one would ask from in-person consultation of a physical collection. For this reason, in this chapter and subsequent ones, we reflect on the time we are able to dedicate to digitized manuscripts and the quantity of data we can create from them outside the reading room. While scale is an important quality of computational work, we have attempted to leverage it here without illusions of grandeur. Scaled approaches to manuscripts have their own limitations, not only due to what algorithmic analysis can reveal, but also to what manuscripts are available for our study. Luckily, of the thousands of extant Paris bibles, many hundreds are digitized, and with some effort we are able to gain access to them.

Manuscripts are not, however, always digitized or available for remote scholarly access. Even though a handful of national or university libraries have made great efforts to digitize their collections, and to make them open for consultation and reuse, it does not mean that Paris bible manuscripts everywhere in the world are open and findable. For this reason, the rate determining step in studying transcriptions made from medieval

1 Smithies, "Software Intensive Humanities."

2 "Material philology takes as its point of departure the premise that one should study or theorize medieval literature by reinserting it directly into the *vif* of its historical context by privileging the material artefact(s) that convey this literature to us: the manuscript"; Nichols, "Why Material Philology?" 10–11 (italics in the original text).

manuscripts often is not linked to knowledge of advanced methods of automated transcription, but rather to our access to quality digitized copies from archives. This situation has required us to temper our desire to comprehend the tradition to its fullest extent. Digital inaccessibility of much of our potential research corpus may change in coming decades, but at the time we are writing, computational accessibility to collections of medieval manuscripts is advancing only slowly.

In this chapter, we reflect on the challenges of building a research corpus of digital medieval manuscripts generally speaking, and of the Paris bible in particular, beginning with the availability of digitized collections necessary to carry out our analysis. Some computational textual studies have a tendency of choosing textual scenarios that work well with established methods, or with research questions arising from method-first inquiry. Rather than limiting ourselves to method-driven analysis, our computational engagement with Paris bibles is driven by both focused hypotheses and more exploratory questions—grounded in genre, scribal culture, and the histories of manuscript creation.

On Big Data from the Middle Ages

Medieval Manuscripts as a Corpus

Big historical data, it has been argued, provides a unique opportunity to understand large-scale patterns using materials from both the distant and not-so-distant past.[3] Working with large amounts of research data necessitates a fluency with a quickly evolving set of analytical methods appropriate for such scale, but also, importantly, a double understanding: both of how such data came into being and of the algorithms by which we have found the data.[4] Web-based records or email archives of organizations are said to be *born-digital*, meaning that they never existed in an analogue format, nor did they have to be retro-digitized. By contrast, materials that were printed, written on paper, or copied on animal skin were *born-analogue*. Some of these archives have had digital copies made of them, with both institutional strategies of preservation and access in mind, and these copies are a boon to researchers in as much as they allow for expanded access to otherwise fragile and inaccessible documents. Their digitization does not automatically enhance access to archives and provide an unmediated relationship to

3 Graham et al., *Exploring Big Historical Data*; Cohen and Rosenzweig, *Digital History*.

4 Salmi, *What is Digital History?*; Fridlund, Oiva, and Paju, *Digital History*.

the originals; instead, digital infrastructure changes the nature of books and our relationship with them.[5]

For some periods of history, the notion of abundance has been used to describe the amount of data available to the researcher,[6] whereas in other archives and other parts of the world, digital abundance is not the norm, or it takes on a different form than in large libraries of the West. In some ways, since this book addresses a genre, the Paris bible, that is found in different sorts of global collections, we simultaneously confront questions of abundance and scarcity. Some of the world's largest libraries have dozens of Paris bibles digitized and available for study. In other libraries, one existing copy made of a bible manuscript may have been made in the 1960s on microfilm and it is only available in digitized form of the microfilm consultable in a reading room. As we discuss in Chapter 3, some key examples of Paris bibles would be interesting to study, but not all have been digitized, even though they are housed in the same large libraries that champion digitization. Truly, digitization is not only about preservation and enabling new research possibilities such as remote reading or computational analysis, it is also a deeply situated process that requires us to understand how archives came to be in the first place, how they have been digitized, and how they have been made accessible on the web.[7]

Let's imagine—for the sake of argument—that all of the surviving collections of medieval manuscripts had already been digitized and were accessible in the form of a digital universal library, using the most convenient contemporary protocols for image transfer on the web, the International Image Interoperability Framework (IIIF). Of course, it is not the case of medieval manuscripts, but let us say that they have been. In such a case, users would require facilities and infrastructure in order to be able to "check out" parts of the digitized collection for study. Scale, in such a case, could easily become as much of a liability as it is an advantage; it is easy to imagine how individual researchers could not have enough hours in the day to read the thousands of folios of digitized manuscripts, let alone find where they want to begin reading. Similar to the physical counterpart of the "total library" imagined by Laßwitz or Borges, the complete copy of surviving digitized medieval manuscripts might be as useless as it would be complete.[8] Scholars

5 Warren, *Holy Digital Grail*.

6 Rosenzweig, "Scarcity or Abundance?"; Milligan, *History in the Age of Abundance?*

7 Zaagsma, "Digital History and the Politics of Digitization."

8 Laßwitz, *Die Universalbibliothek*; Borges, *La biblioteca de Babel*.

would need finding aids and infrastructure like keyword spotting, standardized metadata, or search functionality in manuscripts to help them navigate their research materials in such a large collection.

Still in the realm of the imagination, we could limit ourselves to a smaller corpus: large collections of religious or philosophical texts to be studied looking for differences in form or content by region, monastic order, or time period. Alternatively, scholars could attempt to track iconographic or compositional differences across illuminated manuscript production. All of these projects speak to the possibility of large-scale analysis revealing insights about a given genre, place, or time, insights that would not be possible without digitization. Such grand studies are also highly unlikely, given the way that institutional collections have been, and continue to be, digitized.[9] For a computational medievalist, research necessarily follows the digitized record. Indeed, there are many computational research projects using manuscript collections that one might want to carry out, but unfortunately their incomplete or inaccessible digitization makes it impractical to do so. Design of a scholarly corpus for computational analysis thus requires a significant amount of reflection about the relationship between what you can reliably find and what you ideally are looking for to address your research questions.

Contemporary historians of migration might have a large amount of census data, on par with the size and complexity of the field data of environmental historians or archaeologists. Other disciplines in the humanities and social sciences have large datasets that stem from, and shape, future understanding of research in their field. But what would it mean to speak about big data from the Middle Ages? In 2013 Holsinger published seven short posts to his blog (six of which were guest authored) in which the question of "big data" in the context of medieval studies was already being explored in a North American context. Opinions varied on what the impact(s) would be, but the scholarly exchange is memorable and, fortunately, it has been saved by the Internet Archive. Whereas one of the authors simply rejected the idea of big data altogether in favour of bespoke smaller data,[10] the other contributors explored the question of scale, albeit cautiously, for what it implies from the perspective of our need to organize large amounts of information,[11]

9 Vandendorpe, *Du papyrus à l'hypertexte*.

10 Gillespie, "In Praise of Small Data."

11 Foys, "How I Learned to Stop Worrying."

to gain access to digitized corpora,[12] to carry out automated transcription and palaeography,[13] to study textual *mouvance*,[14] to explore possibilities for new forms of pedagogy,[15] and, last but certainly not least, to train new generations of medievalists about the aggregated resources of our field while exploring new forms of social impact.[16] What seemed like speculative reflections on scale and access in 2013 are today pressing concerns as the expanding reach of artificial intelligence (AI) and machine learning (ML) in cultural collections raises the stakes of scholarly inquiry and reopens old questions with new force.

Institutional Collections and Scholarly Corpora

The concept of big data in medieval studies reveals distinct tensions between institutional priorities and the evolving practices of digital scholarship. Institutional collections are curated with historical, cultural, and preservation imperatives in mind—often shaped by centuries of acquisition, cataloguing, and decades of digitization practices. Cataloguing within these repositories typically prioritizes provenance, conservation status, and accessibility for general users or specialized, case-based research. In contrast, scholarly corpora are constructed to serve specific research questions. They aggregate and structure data in formats that enable large-scale analysis and comparative approaches. As a result, their metadata emphasize standardization across items and compatibility with computational tools—features that institutional catalogues often lack. Indeed, although this situation is slowly evolving, many institutional systems do not provide the level of standardized, detailed metadata required for cross-collection comparison. In a word, their collections are not ready as data for scholarly research.[17] As we elaborate below, obstacles to computational analysis remain substantial, particularly because each institution tends to maintain its own cataloguing system, creating challenges for interoperability. Projects such as Europeana, Biblissima, and a recent iteration of Digital Scriptorium have made important advances by aggregating metadata across repositories and emphasizing common data fields, but this process is far from complete.

---

12 Stinton, "An Unrevolutionary Revolution."

13 Holsinger, "The Googlization of … Paleography???"

14 Nichols, "It's the Manuscripts, Stupid!"

15 McGrady, "Change in the Age of Big Data."

16 Treharne, "'I Tag Bad.'"

17 Padilla, "On a Collections as Data Imperative."

On the other hand, while corpora curated by scholars may have the volume of data for cross-collection analysis, they lack the established infrastructure, funding, and platforms that institutional collections possess. In that case, project teams may need to create their own infrastructure. One such example of a scholarly corpus is the *Horae* (Hours) project. Developed at the Institut de recherche et d'histoire des textes (IRHT) with the support of a number of institutional partners, this research project assembled a collection of digitized, dispersed manuscripts and made it both accessible and usable for large-scale research.[18] With 1158 books of hours currently online (adding up to some three hundred thousand pages), their research corpus assembled about a third of the five thousand or so known witnesses of the genre in the world, providing a standardized framework for analysis. The IIIF proved essential for integrating resources from various institutions, allowing for more uniform metadata and imaging standards, despite the scattered physical locations of the manuscripts. For example, the research team ran an analysis of the structure of the page layouts (miniatures, initials, text lines, headings, and decoration), and produced rough automatic transcriptions of the handwritten texts contained therein. Using these transcriptions, they were able to proceed with automatic identification of texts into categories.

Turning to our own corpus of Paris bibles, we currently lack the type of infrastructure developed for the funded *Horae* project. Nonetheless, we have so far collected metadata from a few hundred digitized manuscripts. What we have as a corpus is large sample of the genre, but clearly nowhere near what was produced in the Middle Ages.[19] The main difference between our project and the *Horae* project lies in the textual variety and focus: while the *Horae* project centred on the same structural typology of a book of hours, which included a variety of different texts brought together, when we study Paris bibles, it is always the text of the Vulgate with minor textual deviations. A Paris bible includes elements such as prologues and textual variants, but as we have already mentioned, for us the main type of variance is graphetic. Congruent with our genre of study, in this book we examine in Paris bibles what we call *variance in uniformity*, focusing on the "fingerprint" of the scribes of the tradition, while also paying careful attention to material details across this corpus.

18 See Boillet et al., "HORAE"; Teklia, *The Horae Project*.

19 Buringh, *Medieval Manuscript Production*.

Do the Humanities Dream of Patterns?

One of the refrains of digital scholarship is that by remediating source materials into digitized form, and by using algorithmic analysis to examine carefully defined elements of them, we may find patterns in our corpus that have not previously been identified by scholars. Modelling humanities data, by which we mean the formalization of sources in a research process that allows specific aspects of them to be analyzed, classified, and explored, is a rich and well discussed topic in digital scholarship.[20] We will discuss modelling in detail in Chapter 2, as it is, in our opinion, the main way that context specific knowledge can be embedded into computational work in the humanities.

As we move from the features that are gleaned during physical access to a codex to others which require more attention, skill, and time to record, the appeal of automation becomes stronger. Some research has looked beyond material evidence of the codex to graphetic features in an attempt to establish the order of scribal contribution or to assess portions of text available to scribes.[21] In orthographically unstable vernacular traditions, such variance might not only include special letterforms, brevigraphs, or abbreviations but also a highly unstable spelling due to evolving language and regional dialectal forms.[22] In our experience with Latin bible manuscripts, however, the former are predominant. Computational forms of analysis of such graphetic patterns, we argue, are an appropriate match for automated approaches such as HTR.

Wanting to discover patterns or underlying principles is not exclusive to digital or computational research, but is found in many knowledge-making disciplines, especially in the humanities.[23] Lemercier and Zalc have argued, for instance, that any source can be quantified if researchers carefully select relevant features and apply systematic methods.[24] They advocate for the use of quantitative analysis in historical research to identify patterns and differentiate the common from the exceptional, arguing that quantitative methods discourage reliance on anecdotal evidence or vague generalization. While this approach reflects the perspective of the quantitative historian, other humanists have raised concerns about what is lost when the interpretive

20 McCarty, *Humanities Computing*, 151.

21 Stubbs, "A Study of the Codicology of Four Early Manuscripts," 23.

22 Bourgain, "Sur l'édition des textes," 15.

23 Bod, "Modelling in the Humanities," 79.

24 Lemercier and Zalc, *Quantitative Methods*.

process is framed primarily around pattern discovery. Herrnstein Smith has critiqued the use of metaphors found in distant reading of uncovering or revealing patterns in texts as an alienating form of extractivism.[25] She insists that such metaphors misrepresent traditions of close reading as a mechanical act, setting up a false contrast with distant reading as a more neutral process. She argues that close and distant reading both attend to micro-features, in ways that always involve subjective human judgement.

We tend to agree on this last point. For us, micro-features are not arbitrary, and they could be gleaned from a close (albeit very slow) examination of a codex. In the case of manuscripts, attention to graphetic micro-features may not illuminate scribal intent, in the way that choice of words in poetry or digressions inserted into texts do for an author, but they do help reconstruct the embodied, collaborative nature of textual transmission. Rather than placing textual analysis and quantitative history in opposition with each other, the contrast we have drawn between Lemercier and Zalc and Herrnstein Smith illustrates how different critical voices within the human and social sciences conceptualize notions of evidence, intentionality, and pattern in distinct ways. We expect that these different notions exist elsewhere, even within the community of medieval studies.

While it is not uncommon to view archival documents such as medieval manuscripts as unique—as a product of a specific time and place and the fruits of human labour—it is also not surprising when art historians classify a painting as typical of a specific atelier, from a specific city, or created by the students of a specific painter, on account of common features they possess. In a somewhat uniform genre such as the Paris bible, in which a rough production template is followed, there are most definitely parameters of that template that can be systematically recorded to study variance. It is the hope of this book that, in defining micro-features of interest and tracing them through many thousands of words and folios, not only can patterns help us to offer novel interpretations of scribal contribution to the corpus, but also that the exercise can open up discussion into the critical reflexivity required when working with computational methods that engage with differing concepts of the pattern.

The exploration of the pattern may be "the strongest point of intersection between the computational structures of text analysis and the open-ended interpretative landscape of literary studies."[26] Ramsay has argued how the

25 Herrnstein Smith, "What was Close Reading?," 70–71.

26 Ramsay, "In Praise of Pattern," 177.

humanities and the sciences leverage forms of evidence in different ways to produce scholarly judgements, with the rhetorical and exegetical practices in the humanities being distinct from factual and denotative approaches in the sciences. Both fields value the discovery of patterns in their materials, and he hopes that future humanities research methodologies can help to take some of the "serendipity out of the process" of discovering unnoticed patterns.[27]

While patterns can yield valuable insights, an important question arises: to what extent do we value the identification of patterns as either a distinctly human or non-human activity? The interpretive weight assigned to patterns depends not only on their recognition, but also on a deep contextual understanding of the material under analysis—an understanding often cultivated through active participation in the creation and curation of the datasets themselves. Data analysis in the humanities involves navigating "the spurious, the contingent, the inexact, the imperfect, and the accidental in a state of almost guaranteed incompleteness"—a description Ramsay argues aptly characterizes the central task of criticism.[28] Yet in computational humanities projects involving cultural datasets, the boundaries between human and machine recognition of patterns can blur.[29] Computational models are trained by example—using data created mostly by humans—yet a machine possessing a dataset of a finite number of known facts cannot be said to possess comprehensive understanding. Instead of a stark dichotomy between human judgement and machine detection, we believe it is more productive to understand this relationship as a synergistic continuum—and one that continues to evolve with the increasing integration of computational tools into humanities research. In the end, the value of patterns in the humanities remains closely linked to our human capacity to define and to discern significance, to separate signal from noise. That is, computational results must be carefully assessed and translated into new critical questions, arguments, and interpretive frameworks.

The computational turn in the humanities, on which scaled pattern searching depends, has been accompanied by significant political and institutional critique. It has been written, for example, that

> [c]omputational techniques are not merely an instrument wielded by traditional methods; rather they have profound effects on all aspects of the

27 Ramsay, "In Praise of Pattern," 181.

28 Ramsay, "In Praise of Pattern," 186.

29 Kaplan, "A Map for Big Data Research," 5.

> disciplines. Not only do they introduce new methods, which tend to focus on the identification of novel patterns in the data as against the principle of narrative and understanding, they also allow the modularisation and recombination of disciplines within the university itself.[30]

We also agree with this claim, and in Chapter 3 we will consider examples of pattern searching in transcriptions of medieval manuscripts, arguing that we do not see such methods as oppositional to scholarly narrative practices, but as complementary to, even constitutive of, new forms of scholarly judgement. In Chapter 4, we extend the discussion of disciplinary recombination, arguing that while computational methods realign scholarly practices in ways reminiscent of the sciences, they also open new pathways for innovation—provided that such recombination is pursued with disciplinary self-awareness and active critical engagement. These shifts present an opportunity for the medievalist community to reflect on how the profession might adapt—not by resisting change, but by shaping it in ways that sustain intellectual rigour and methodological plurality. In our opinion, if the medievalist community, and in particular its scholarly societies, does not take an active role in shaping how new computational and interdisciplinary approaches are incorporated into the future of the field, they do so at their own peril.

Why Paris Bibles?

With these introductory considerations in mind, the Paris bible corpus provides us with a unique opportunity to explore scribal practices through a computation lens, but this exploration needs to be accompanied by careful attention to material detail.[31] Since each Paris bible is a handcrafted object—often copied by multiple scribes—its material execution inevitably departs from idealized notions of uniformity in ways that are both visible and innumerable, often exceeding the limits of what can be consistently observed or recorded through human inspection alone. When approached computationally and at scale, these micro-variations offer a compelling framework for the quantitative analysis of scribal practices, allowing us to detect commonalities and divergences between scribal behaviour that would otherwise remain undocumented, or only superficially studied.

Such patterns of variation across single manuscripts and the larger corpus prompt us to adapt frameworks initiated from quantitative (or statistical) codicology, a field that considers the codex not simply as a container for

30 Berry, "The 'Computational Turn.'"

31 Deploige and De Gussem, "Medieval Authorship and Canonicity."

text, but as a complex cultural and material apparatus. A convenient definition from quantitative codicology is that a codex is a

> device that functions in a complex way, [*machine au fonctionnement complexe*] which is to say it is a handcrafted object designed to transmit a text in a most enduring and legible way, whilst at the same time serving as an expression of a specific "cultural and social fabric."[32]

Quantitative codicological research endeavours to collect material evidence from handcrafted book-objects (ideally, first-hand in the archive) in view of gaining a deeper idea of how they were fabricated, modified, expanded, manipulated, or damaged over time by the various actors in the object's history.[33] We say ideally, since some details do require close physical access to the book-object, yet the study of codices, in-person and in-depth, is not realistically a scalable endeavour. The engagement of researchers in archival collections is limited by time and resources, and especially for globally scattered corpora such as that of Paris bibles. As such, it is unreasonable for a team of researchers, let alone two researchers, to engage first-hand with the totality of the corpus.

Practitioners in quantitative codicology have understood the value of digitized collections, including both digital reproductions of manuscripts and the metadata that captures some of the evidence they would like to glean from such objects. They have played, as Maniaci observes, an important role in the creation and dissemination of data about manuscripts for the community of scholars. Yet, in our opinion, that role has not been without significant divergence from contemporary data standards. Indeed, publishing data about manuscripts in an open and machine readable way, as OPenn or Digital Scriptorium 2.0 have, is an exception in the field.[34] Quantitative codicological research drawing on large genres must rely on metadata or descriptions taken from a variety of sources, print or digital catalogues, rendering the process of organizing data about corpus another important rate determining step.

Whereas it is impossible to claim comprehensive access to the many, many thousands of extant medieval manuscripts found around the world, quantitative codicology need not be restricted to physical examination. New

---

32 Maniaci, *Trends in Statistical Codicology*, 1, partially citing Bozzolo et al. "Une machine au fonctionnement complexe," 69–78.

33 Maniaci, "Introduction: Statistical Codicology," 2.

34 University of Pennsylvania, OPenn; Digital Scriptorium, Digital Scriptorium Search Portal.

practices in digital codicology are emerging, but like our general claims above about medieval manuscripts and computational humanities, they depend on a rich, ethical reconsideration of scholarly labour.[35] This perspective necessitates a commitment to the traceability of data, which includes giving credit to cataloguers and their work in constructing and transmitting that metadata. It also requires a commitment to infrastructures of fair collaboration, a point on which we will elaborate in Chapter 4.

Instead of assuring a continuity with knowledge of the past, the turn to the digital has erased a part of institutional and cataloguing histories. When catalogues were printed, they were (and continue to be) treated as publications in their own right, that scholars use and cite accordingly. But the transformation of these printed catalogues into large online databases and catalogues has changed the way we conceive of the metadata about objects, manuscripts, and books. Despite presenting some of the same information, although not always with the same degree of detail, the printed catalogue and the digital one are not the same work: one is a scholarly endeavour undertaken by researchers whom we credit, the other is merely a gathering of information produced by an institution. Identities are lost and with them, authority and prestige can be as well.

However, as exemplified by the BnF's digital catalogue "Archives et Manuscrits" (Archives and Manuscripts), cataloguing practices can evolve to acknowledge the research labour of creating and updating catalogue descriptions explicitly, embedding an accountability that encourages scholarship to retain its connection to the individuals who produced them. By citing the authors of catalogue descriptions and recognizing the metadata they document as both a tool and a text of its own, we can strengthen the bond between physical manuscripts and their digital surrogates, all the while enriching the scholarly dialogue that enables computational approaches to large manuscript corpora.

In this book we reflect on some of the concerns of quantitative codicological approaches mentioned above, while deepening its relationship with advances made in the computational methods. The research questions we would ask of a physical collection are obviously not the same as those that can be asked of a digital collection. Specifically, scribal practices cannot be studied at scale using manual quantitative codicological and palaeographical methods. When one is primarily working with digitized collections, as we have for this study of Paris bibles, one of the most compelling advances in

35 Whearty, *Digital Codicology*.

technology in the 2020s is the availability of automated HTR. In Chapter 2, we discuss how text can be automatically transcribed from digitized medieval manuscripts, using features defined by scholarly judgement, and how this method helps us reflect on the "handmadeness" of the medieval codex.

Variance in Uniformity

What is a Paris Bible? A Corpus in Search of a Definition

The twelfth-century Renaissance witnessed a rise in both the number and the quality of manuscripts. These objects were created at an unprecedented scale, with the production of manuscripts being greatly influenced by the creation of universities.[36] Medieval scholars needed large quantities of affordable and standardized texts, a demand which led to a thriving commercial book trade as well as several innovations by manuscript makers. The *pecia* system,[37] whereby university booksellers loaned out individual sections of a manuscript to students or scribes for copying, was first developed in Bologna before spreading to Paris, and it changed the way manuscripts were produced. It brought about lasting changes to the format, layout, and organization of manuscripts, and bibles in particular, resulting in compact single-volume manuscripts with a more or less standardized layout, iconographic programme, chapter numbers and running titles. Evidence such as illumination patterns and faint marginal notes about payment have led manuscript historians to suggest that such bibles were produced for a commercialized trade, in what can be described as one of the first mass-produced objects.[38]

The production of these Bibles has been linked to the rise of literacy. Their existence and spread reflect a shift in how readers—clerical or lay, scholarly or pious—engaged with the written word.[39] Similarly, for the first time in the Middle Ages, a bible could become a personal possession used by a range of individuals: from students and professors at emerging universities or bishops and priests focused on expanding pastoral care, to itinerant preachers.[40] During this period, books in general were increasingly designed for personal rather than institutional use, and the rise in book production

36 Clanchy, *From Memory to Written Record*.

37 Destrez, *La pecia*.

38 Rouse and Rouse, "Book Production in Paris."

39 Magrini, "Production and Use of Latin Bible Manuscripts," 222.

40 Light, "The Thirteenth Century and the Paris Bible."

Table 1. A list of the sixty-four common prologues, with the English names of the biblical books and the corresponding Stegmüller identifier, based on London, Lambeth Palace (hereafter Lambeth), MS 1364. Adapted from Ker, *Medieval Manuscripts in British Libraries*, 1:96–98.

No.	Book of the Paris Bible	Stegmüller ID	No.	Book of the Paris Bible	Stegmüller ID
1	Epistle of Saint Jerome to Paul	S.284	35	Habakkuk	S.531
2	Pentateuch	S.285	36	Zephaniah	S.534
3	Joshua	S.311	37	Haggai	S.538
4	1 Kings (a.k.a. I Samuel)	S.323	38	Zechariah	S.539
5	1 Paralipomenon (a.k.a. I Chronicles)	S.328	39	Malachi	S.543
6	2 Paralipomenon (a.k.a. II Chronicles)	S.327	40, 41, 42	1 Maccabees	S.547, S.553, S.551
7	1 Ezra	S.330	43, 44	Matthew	S.590, S.589
8	Tobit	S.332	45	Mark	S. 607
9	Judith	S.335	46	Luke	S.620
10, 11	Esther	S.341, S.343	47	John	S.624
12, 13	Job	S.344, S.357	48	Romans	S.677
14	Proverbs	S.457	49	1 Corinthians	S.685
15	Ecclesiastes	S.462	50	2 Corinthians	S.699
16	Wisdom	S.468	51	Galatians	S.707
17	Isaiah	S.482	52	Ephesians	S.715
18	Jeremiah	S.487	53	Philippians	S.728
19	Baruch	S.491	54	Colossians	S.736
20	Ezekiel	S.492	55	1 Thessalonians	S.747
21	Daniel	S.494	56	2 Thessalonians	S.752
22	Minor prophets	S.500	57	1 Timothy	S.765
23	Hosea	S.507	58	2 Timothy	S.772
24, 25	Joel	S.511, S.510	59	Titus	S.780
26, 27, 28	Amos	S.515, S.512, S.513	60	Philemon	S.783
29, 30	Obadiah	S.519, S.517	61	Hebrews	S.793
31, 32	Jonah	S.524, S.521	62	Acts	S.640
33	Micah	S.526	63	Catholic Epistles	S.809
34	Nahum	S.528	64	Apocalypse	S.839

from the thirteenth century onward can be seen as a response to a sustained demand from a growing readership.[41]

Indeed, in the mid-thirteenth century, newly formed friar orders, such as the Dominicans and Franciscans, adopted the Paris bible model, leading to further innovations in the bible codex, producing even smaller formats. Around 1230, both orders established schools in Paris, with the Dominicans attending the theology faculty at the University of Paris, where they encountered these new bibles. Friars favoured the Paris bible for a few reasons: it was portable, definitive, searchable, and commercially available. Unlike monks, friars did not live in seclusion but travelled and preached widely, thus spreading these biblical manuscripts across Europe. We find evidence of French bibles appearing in Italian convent libraries by the first half of the thirteenth century.[42] Since the friars renounced personal possessions and moved frequently, they needed small, portable bibles (known as saddle or pocket bibles) and many such codices were produced to meet their needs, resulting in their relative abundance today: around 1800 are still extant in public collections around the world, albeit often undigitized.[43]

The development of the Paris bibles is linked to another important phenomenon, that of the revision of the Christian scriptures around 1220 by the University of Paris. The result was a corrected, organized text and is often evoked as a defining feature of Paris bibles. In terms of biblical transmission, this revision is second in significance only to St. Jerome's establishment of the biblical canon in the fifth century, and it is the template upon which the structure of the modern Bible is based.[44] It includes a new book ordering system independent of the liturgical year and a common set of sixty-four prologues introducing individual (or sets of) biblical books.[45] The books were divided into new standardized chapters, and these bibles often end with the text of the Interpretations of Hebrew Names.

When we use the term "Paris bible," we do so in a deliberately broad and inclusive sense: the term does not refer strictly to bibles produced in Paris, but to a codex written in a particular style and containing a specific textual configuration that emerged from a revision of the Scriptures. While Paris was undeniably one of the intellectual centres of this development,

41 Ornato, "Les conditions de production," 59.

42 Ruzzier, "Qui lisait les bibles portatives fabriquées au XIII^e siècle ?"

43 Ruzzier, *Entre université et ordres mendiants*, 2.

44 Morard, "À la recherche de la 'Lettre commune.'"

45 Ruzzier, *Entre université et ordres mendiants*, 68.

especially in the thirteenth century when many of these manuscripts were copied there, visually similar bibles were also produced in other major university towns such as Oxford and Bologna, as well as in regions including Catalonia, southern Germany, and other Italian cities. In this book, and in our digital research project launched in 2020, we use the term "Paris bible" as a flexible category—a loosely defined container, a genre, if one can call it that—encompassing manuscripts exhibiting varying degrees of formal, material, and textual coherence. Our aim is not to construct a typology of late medieval Latin bibles, as projects like Gloss-e have attempted to trace the evolution toward the Clementine Vulgate.[46] Instead, the breadth of the term "Paris bible" affords us methodological freedom to include a diversity of Latin biblical manuscripts in our corpus, acknowledging that this inclusivity may stretch conventional definitions. This openness is central to our commitment to computational analysis, where capturing variation and uncertainty is as important as identifying regularity.

Indeed, many assumptions about the uniformity of Paris bibles have been made in the scholarly record over time. Given the visual quality of these manuscripts, there are some features that are more easily tracked and quantified than others. Paris bibles are pandects, or, in other words, they are bound in one volume—or two in rare cases—unlike bibles from previous periods that were often bound in several volumes, usually three, four, or as many as fourteen.[47] In fact, only about 3 per cent of extant pocket bibles are currently bound in two or three volumes.[48] However, for most of these, the division seems to have appeared long after the creation of the manuscript, on the occasion of a rebinding for example.

Paris bibles also share a unique layout with two columns, a gothic *textualis* script,[49] the presence of alternating red and blue coloured running titles in the upper margin, chapter numbers in alternating red and blue coloured Roman numerals, and the general absence of in-line glosses or commentaries. They are also significantly smaller compared to the previous periods of bibles. In this way, the corpus of Paris bibles provides not only a textual but also a visual standard, codifying a recognizable aesthetic that would shape biblical production well beyond the fourteenth century in Germany and

46 Morard, "À la recherche de la 'Lettre commune.'"

47 On the library catalogue of the monastery of Saint Riquier (831), see Ganz, "Carolingian Bibles," 326.

48 Ruzzier, "Continuité et rupture," 165.

49 Derolez, *The Palaeography of Gothic Manuscript Books*.

Eastern Europe, as well as in the form of the Gutenberg Bible. This consistency in visual features raises questions about the assumption of uniformity across the corpus—an assumption that our project seeks to test with analytical precision.

Quantitative Codicology and the Corpus

Beyond the visual coherence that characterizes the genre, a wide range of variance can be observed across the corpus. Some of the most apparent differences among the bibles reflect the time and place of their production, yet such information is often only approximate, since these manuscripts are rarely dated or localized precisely in a colophon. The field of quantitative codicology has developed methods for identifying and measuring physical features of codices to formulate informed arguments about their fabrication.[50] In some cases, this variation is well documented through observable features that occasionally appear in printed catalogue descriptions or digital records. Such features can include the overall dimensions of the codex, the size of the text block, the number of lines per column, the number of quires, or their collation structure.[51]

From a technical point of view, the creation of Paris bibles was facilitated by the manufacture of large quantities of high quality, fine parchment, the development of a condensed script (described in detail by Derolez[52]), and the extensive use of abbreviations. Yet, the technique that allowed the creation of such fine parchment is still largely unknown. Some scholars have suggested the use of the skins of various, small animals, such as rats, squirrels, rabbits, or aborted calves.[53] To solve this mystery, Fiddyment and colleagues studied seventy-one pocket bibles, but were not able to identify such rare skins; rather, only skins from calves, sheep and goats were found.[54] Other scholars have suggested that the parchment was made by splitting

50 Bozzolo and Ornato, *Pour une histoire du livre manuscrit* ; Julien, "Construction et composition"; Maniaci, *Trends in Statistical Codicology*; Ornato, "La codicologie quantitative"; Reynhout, "Codicologie quantitative et paradigmes scientifiques."

51 Quantitative codicology is particularly interested in the sum of the height and width of a codex. See Bozzolo and Ornato, *Pour une histoire du livre manuscrit*, 217; and Muzerelle, "Pour revenir sur et à la 'taille' des manuscrits."

52 Derolez, *The Palaeography of Gothic Manuscript Books*.

53 Thompson, "Technology of Production," 75–84.

54 Fiddyment et al., "Animal Origin of 13th-Century Uterine Vellum," 15,066.

skins to create two leaves of parchment from one individual skin.[55] Whatever the technique used to create the fine parchment of Paris bibles may have been, the data point to patterns in the use of certain animals. In their study, Fiddyment and colleagues noticed that Italian bibles were exclusively made from goats whereas English bibles were made from three different species: calves, sheep and goats.[56] Interestingly, bibles made in Paris were made mainly from calf skins, with only 15 per cent of the total being made from goat skins. No sheep skin was identified.

The dimensions, relatively easy to study, are also important. Ruzzier has argued that the size of these books is closely tied to their intended use and audience.[57] She distinguishes three primary categories: small portable bibles, larger ceremonial ones, and medium-sized volumes. Portable bibles were designed for personal use and, at times, for travel, making them practical for individuals on the move. In contrast, large-format bibles retained their traditional role as ceremonial objects, usually housed in religious institutions where they were as symbols of status or served for liturgical purposes. The third group, medium-sized bibles, fulfilled a variety of functions. Many of these medium-sized codices contain traces of exegetical commentary, a feature facilitated by their manageable size, which provided enough space for marginal notes without being cumbersome. Ruzzier's analysis highlights how material aspects of manuscripts, such as size, were not arbitrary, but instead reflected the practical, institutional, and intellectual needs of their users.

She also highlighted a series of codicological and textual indicators that are considered reliable to locate these manuscripts and help create typologies to measure the diffusion of the Parisian model over Europe.[58] For example, she demonstrated that, in bibles produced in France, chapter numbers occur more frequently within the text, while in the English and Italian manuscripts they are more frequently placed within the margins. She also noticed that gatherings of twenty-four folios are characteristic of French pocket bibles. Additional textual indicators include the order of the books and their variability compared to the "Parisian text" mentioned above, presence/absence and localization of psalms, number and choice of prologues, capitulation type, presence or absence of the interpretation of the Hebrew names, use and

55 Clarkson, "Rediscovering Parchment," 5–26.

56 Fiddyment et al., "Animal Origin of 13th-Century Uterine Vellum," 15,070.

57 Ruzzier, "Continuité et rupture," 158.

58 Ruzzier, "Continuité et rupture," 156.

frequency of abbreviations, or the place and colours of peritextual elements such as running titles or chapter numbers. From a material and visual perspective, number of folios, dimensions, thickness of the parchment, ruling pattern, structure of the quires, type of signatures or catchwords, or the extent to which space on the page is used are important indicators as well. Looking at the order of the books, variance is widespread; many exceptions and older orders to the Parisian text persisted. Ruzzier organized them into three main categories, perfect Parisian order, Parisian order with minor gaps or variations, and other orders, looking for patterns in their use.[59]

Ruzzier also looked at the variation in script and graphical features across 122 manuscripts.[60] However, her analysis raises methodological concerns. One issue of concern to us is a lack of data about the specific manuscripts included in the sample, and their breakdown in production by regions or countries. The limited number of filters used in the study can restrict the scope of the findings, reducing our ability to draw conclusions about regional or temporal variations. Moreover, Ruzzier identifies only one instance of abbreviation, suggesting that the treatment of abbreviations is not a major issue of concern. While her study addresses significant differences in script styles and graphical elements, its data selection and classification make it difficult to draw more general conclusions. We aim to fill this gap, not for the entire tradition, but for a sample of manuscripts, by engaging in some specific and well described computational experiments concerning scribal practices in Chapter 3. The next section will look, however, at one aspect of uniformity in Paris bibles to introduce the reader to some forms of statistical and computational analysis.

Prologues in a Corpus-Based Analysis of Paris Bibles

To explore variance within a seemingly uniform corpus, we focus on the sixty-four prologues that Ruzzier identifies as commonly belonging in Paris bibles. A cataloguing system for biblical prologues, the *Repertorium biblicum medii aevi*, was developed by Stegmüller.[61] In it, he assigned unique reference numbers to incipits and explicits of various prologues found in medieval Latin Bibles, particularly those associated with the Vulgate tradition. His references have become standard identifiers, often cited in manuscript catalogue descriptions today as "Stegmüller," "Stegm.," "RB" (for *Repertorium*

59 Ruzzier, "Continuité et rupture," 159.

60 Ruzzier, *Entre université et ordres mendiants*, 161.

61 Stegmüller, *Repertorium biblicum medii aevi*, 1:253–310.

biblicum) or more simply "S." followed by the catalogue number (e.g., S.468). Examples of these are listed in the rightmost column of Table 1.

This data about the composition of prologues has been used convincingly by Ruzzier to carry out analysis of patterns in Paris bibles. She argued that this corpus lacks systematic analysis, highlighting two important observations: first, that the Psalms had not been preceded by a prologue in what she calls typical Paris bibles, and second, that three prologues systematically appeared in these biblical codices: S.468 before the book of Wisdom, S.327 before II Chronicles and S.551 before Maccabees. More importantly, she discovered that a particular pattern of six prologues that is typical of Parisian Paris bibles: S.462 before Ecclesiastes, S.513 before Amos, S.547 and S.553 before Maccabees, S.589 before the Gospels, and S.839 before the Apocalypse. Moreover, she documented that only a third of the manuscripts she studied exhibited the ideal sequence, leading her to create a typology: "non Parisian," for a manuscript which does not contain any of the six Parisian prologues, "mixed" for manuscripts which contain only a portion of the six prologues combined with additional ones, "incomplete Parisian" for manuscripts which contain only the Parisian sequence but is incomplete, and "sixty-four + others" for manuscripts which contain both the full set of the Parisian prologues and the full set of "traditional" sixty-four prologues.

Pursuing this structural analysis, Ruzzier observed that many Bibles contain only a few Parisian prologues alongside numerous non-Parisian ones, indicating that the Parisian sequence did not achieve widespread success, especially outside of France. She mentions that only 6.8 per cent of English bibles and 2.7 per cent of Italian bibles include the full Parisian sequence, which is highly significant since English bibles otherwise resemble French manuscripts closely. England has the highest percentage (56.8 per cent) of bibles with none of the new Parisian prologues. France seems to be the only country where the Parisian sequence saw real success, yet there remains a clear difference between manuscripts known to have been copied in Paris (80.6 per cent) and those copied elsewhere or of uncertain origin (41.4 per cent). These differences are noteworthy since they are much more important than the differences observed in other textual and codicological elements, such as the order of the books or the chapters divisions. She also noted a correlation between the presence of Parisian prologues and the size of the manuscripts: the smaller the manuscript is, the higher chance it seems to have to contain this specific set of prologues, highlighting a link between the prologue structure and the type of audience. She explains these differences by noting that the success of the Parisian chapter divisions and, to a lesser extent, the order of books, can be attributed to their practical

functionality, whereas the adoption of the Parisian prologues was driven by exegetical choices. The latter were more open to debate and could clash with existing local traditions.

Although Ruzzier highlights the presence or absence of the Parisian prologues and divergences in bibles originating outside of Paris, she does not, however, offer a detailed analysis of these prologues, how they correspond to each other, nor if patterns can be observed beyond the Parisian sequence. For example, do other manuscripts produced, say, in Germany or England, share a number of common prologues not present in the traditional sixty-four French set? To offer preliminary answers, and to demonstrate the value of a data-centred approach, we gathered the Stegmüller references for the prologues contained in eighteen manuscripts of the general Paris bible tradition, drawing primarily on data from manuscript catalogues. A critical comparison with Ruzzier is, however, not as straightforward as one might hope, since Ruzzier's dataset is not openly available.

Identifying features such as the incipits and explicits of prologues and their corresponding Stegmüller reference requires a significant amount of effort. Looking at prologues across a sample of the corpus led to our preliminary analysis of patterns in the distribution, frequency, and variance of prologues. The manuscripts chosen span the thirteenth and fourteenth centuries and originate from France, England, Spain, and southern Germany. To identify patterns in prologue usage, we performed some statistics and network analysis linking Stegmüller numbers to biblical books and other metadata across manuscripts. Some of the key findings include results about the variability or stability of the prologues, some book-specific associations as well as some geographical and temporal clustering.

Perhaps not surprisingly, the manuscripts, including almost all of the sixty-four prologues from the Parisian sequence, have been located in Paris (Table 2). The farther we move from Paris as a centre of production, the lower the percentage of prologues from this set becomes. In Germany, manuscripts include only 57 and 65 per cent of the sixty-four Parisian prologues while English manuscripts possess less than 50 per cent of the prologues. However, other patterns can be found in a network analysis (Figure 1). While most books exhibit a high degree of consistency, always or nearly always associated with the same prologue, others exhibit considerable variation. For example, Wisdom consistently appears with S.468 and Philippians with S.728. In contrast, Romans appear with seven different prologues (S.669, 679, 678, 670, 674, 654, and 677); Psalms with ten (S.430, 414, 418, 394, 398, 450, 443, 382, 445, and 456) and Matthew with seven (S.601, 595, 589, 590, 596, 591, and 581).

Table 2. A selection of eighteen Paris bibles, their origins and dates, and a calculation of the percentage of prologues they contain of the most common sixty-four prologues. Data by authors.

Shelfmark	Origin of Manuscript	Dating	% of 64
New Haven, Yale University, Beinecke Library (hereafter Beinecke), MS 387	England	second half, thirteenth	37.5
Paris, Les Enluminures, MS TM 1226	England	second quarter, thirteenth	48.5
Aarau, Aargau Kantonsbibliothek (hereafter Aarau KB), MS MsWettF 11	southwest Germany	third quarter, thirteenth	57.8
Schaffhausen, Ministerialbibliothek (hereafter Schaffhausen MB), MS Min. 6	southwest Germany	first quarter, fourteenth	65.6
St. Gallen, Kantonsbibliothek (hereafter St. Gallen KB) MS VadSlg 332	northern France	third quarter, thirteenth	70.3
Paris, Les Enluminures, MS TM 844	Spain	mid-thirteenth	78.1
LAD, MS 2013.051	northern France	third quarter, thirteenth	82.8
Göttweig, Benediktinerstift (hereafter Göttweig BS), MS Cod. 116	northern France	mid-thirteenth	87.5
Paris, Sotheby's, June, 27 2024, "Livres et manuscrits," Lot 1	Paris	second quarter, thirteenth	89.1
Université du Québec à Montréal, MS without shelfmark	southern France	mid-thirteenth	89.1
Paris, Bibilothèque nationale de France (hereafter BnF), MS latin 10421	Paris	thirteenth	90.6
Cologny, Fondation Martin Bodmer (hereafter Cologny), MS Cod. 28	northern France	fourth quarter, thirteenth	92.2
BnF, MS latin 10426	Paris	third quarter, thirteenth	93.8
BnF, MS latin 11935	Paris	second quarter, fourteenth	93.8
Lambeth, MS 1362	Paris	second quarter, thirteenth	96.9
Paris, Les Enluminures, MS TM 1327	northern France	second quarter, thirteenth	96.9
Beinecke, MS 433	Paris	mid-thirteenth	98.4
Lambeth, MS 1364	Paris	second quarter, thirteenth	100

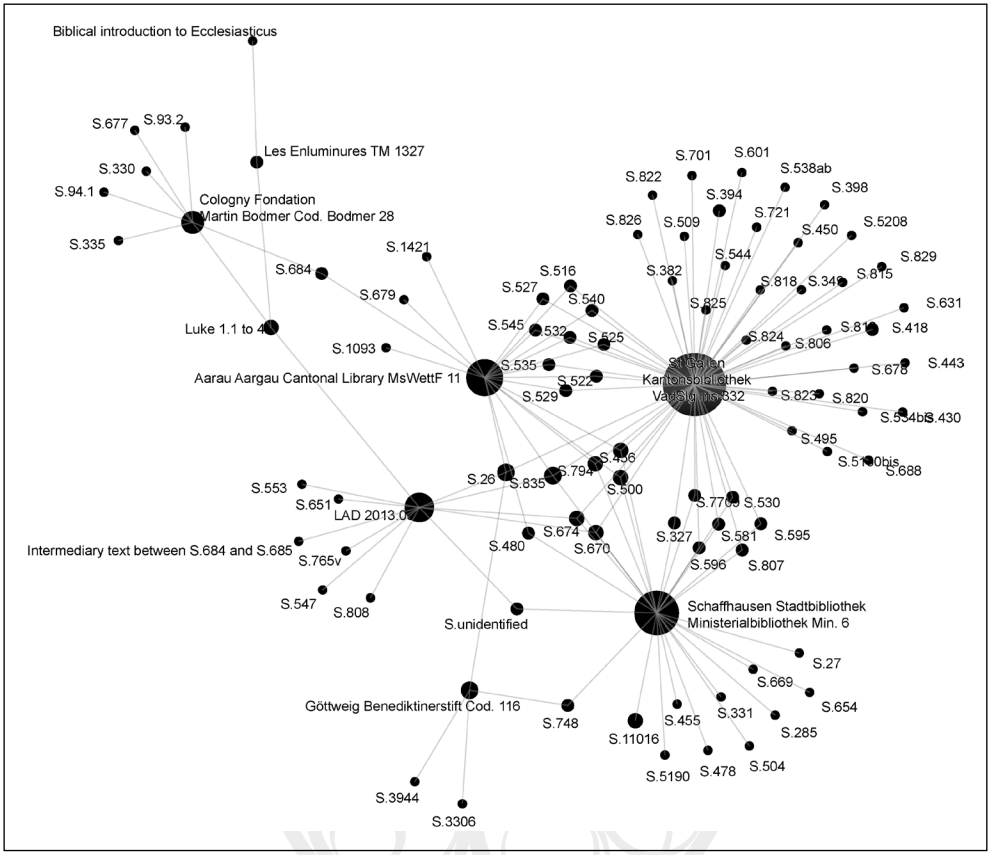


Figure 1. A network visualization illustrating prologues shared (beyond the sixty-four common prologues) by seven manuscripts: five localized to northern France (Cologny, MS Cod. 28, Göttweig BS, MS Cod. 116, St. Gallen KB, MS VadSlg 332, LAD, MS 2013.051, and a manuscript sold by Paris, Les Enluminures, MS TM 1327) and two to southwest Germany (Aarau KB, MS MSWettF 11 and Schaffhausen MB, MS Min. 6). Data by authors. Visualized in Palladio by authors and reproduced with permission.

In fact, most prologues are closely associated with a single biblical book. In some instances, however, a prologue is shared between two or three books, typically from the same genre or biblical grouping. Examples include S.11016, shared by Mark, John, and Luke (three of the four Evangelists), and S.327, shared by I Paralipomenon and II Paralipomenon, while S.328 is present exclusively with I Paralipomenon and S.546 and S.455 are both linked to Proverbs and Psalms. Similarly, although a large number of prologues are shared by all or almost all the manuscripts—mainly the set of the sixty-four Parisian prologues—initial analysis shows that others correlate with particular locations or time periods. For instance, S.794 appears in only three

manuscripts, two of them having been produced in southwest Germany and the other one, identified as originating in northern France. The manuscripts St. Gallen, Kantonsbibliothek (hereafter St. Gallen KB), VadSlg 332 of northern French origin and Schaffhausen, Ministerialbibliothek (hereafter Schaffhausen MB, Min. 6) of southwestern German origin, share S.807, 581, and 530—three prologues that are present in no other manuscript of the studied corpus. Furthermore, these two manuscripts also share a number of other prologues that do not belong to the traditional sixty-four: S.670 and S.674 which both have three occurrences, the third in LAD, MS 2013.051 (northern French origin); S.595 and S. 596 which both have three occurrences, the third in Beinecke, MS 387 (English origin); S.456, shared by five manuscripts, including the two German ones; and S.835 also shared by five manuscripts including the two German manuscripts and one English manuscript. Interestingly, S.327, which is traditionally part of the sixty-four Parisian prologues placed before II Paralipomenon, is located before I Paralipomenon in three instances: St. Gallen KB, VadSlg 332, Schaffhausen MB, Min. 6 and LAD, MS 2013.051.

These observations, based on a small subset of only eighteen manuscripts, suggest strong relationships between the German manuscripts and two additional manuscripts produced in northern France: St. Gallen KB, VadSlg 332 and LAD, MS 2013.051. However, these two manuscripts reveal few similarities with other manuscripts located in northern France (Cologny, MS Cod. 28, Göttweig BS, Cod. 116 and a manuscript sold by Paris, Les Enluminures, MS TM 1327), all of which display a stronger resemblance to Parisian manuscripts. Based on the use of prologue incipits alone, an argument could thus be made to suggest a change in the place of production of St. Gallen KB, VadSlg 332 and LAD, MS 2013.051 to southwest Germany. While preliminary, these patterns suggest localized textual traditions or scriptorium-specific practices in the use of specific prologue texts. However, given the relatively small sample size per region and for the fourteenth century, caution is warranted: such patterns will be nuanced as the corpus grows. A larger sample scaling up the corpus would be necessary to refine the analysis.

Importantly, our initial results challenge the uniformity thesis of the Paris bible tradition as well as the claim of sixty-four traditional prologues associated with it. In this sample of eighteen manuscripts, more than 150 different prologues appear, including a few that have not yet been identified. While it appears that the Parisian sixty-four constitute a core set of them, considerably more research needs to be done to gain a more global view as to their diversity. A corpus-based approach allows us to confirm some aspects of standardization, while simultaneously revealing local, temporal,

or codicological exceptions. This examination of prologues in a small sample of Paris bibles demonstrates the importance of structured, shareable data in manuscript studies. Looking forward, the use of IIIF-based tools combined with web annotation platforms like hypothes.is could allow scholars to collaboratively tag and trace Stegmüller prologues across digital facsimiles. We could imagine tapping into HTR-created transcriptions and machine-assisted approaches to automate some of the process of data creation. Being explicit about whatever workflow one might choose to solve this problem in the future would not only support transparency and reproducibility in the research, but would also enable scaling the analysis to a much larger sample. We will discuss the possibilities and limitations of collaboration for such a research project later in Chapter 4.

Obstacles to Computational Analysis of Paris Bibles: Discoverability and Use

What Has Been Produced? What Has Been Preserved?

One of the main questions we are faced with when doing computational analysis of premodern collections is: what portion of manuscripts created in the medieval period is extant? Scholars have grappled with this question for some time, yet it is quite difficult to determine precisely the number of remaining manuscripts in the world today, and it is more challenging to estimate the total number of manuscripts produced in the Middle Ages. Depending on a multitude of parameters, the percentage of lost manuscripts can be difficult to evaluate. To estimate the currently surviving number of manuscripts in the world, researchers have used several methods, and, as we could expect, conclusions varied immensely. Neddermeyer evoked one out of fifteen manuscripts produced that survived for the fifteenth century while Cisne estimated the survival rate of ninth-century manuscripts to be one out of seven.[62] On the other hand, Bozzolo and colleagues were not able to provide a number, or even an estimate, of the survival rates of medieval manuscripts.[63]

Kestemont and colleagues used models from biodiversity to estimate the loss of Middle Dutch chivalric epics.[64] In a similar vein using statistical

62 Neddermeyer, *Von der Handschrift zum gedruckten Buch*, 81; Cisne, "How Science Survived."

63 See Bozzolo and Ornato, *Pour une histoire du livre manuscrit*, 83.

64 Kestemont et al., "Forgotten Books."

methods based on numbers of extant manuscripts, Buringh argued that more than eleven million books were produced in the Latin West during the Middle Ages between the sixth and fifteenth centuries.[65] Their chronological distribution follows an exponential curve, with a slight increase in production around the ninth century and significant growth from the twelfth century, peaking between the thirteenth and the fifteenth centuries. More than four million manuscripts were thought to have been produced during the fifteenth century alone, while the production of the thirteenth to fifteenth centuries make up almost 90 per cent of the total. What would these predictive models mean for the production of Paris bibles? We know that over 1800 Paris bibles are documented as extant, and we can draw on Buringh's methodology to estimate the total number originally produced.[66] He calculated a "reciprocal survival factor" (RSurvi), which quantifies the factor by which surviving manuscripts must be multiplied to approximate the total number originally produced in a given century. For the thirteenth century, this factor is 13.4.[67] Using Buringh's calculation and Ruzzier's original figure of 1800 extant manuscripts, we can estimate that over 24,000 Paris bibles were likely to have been produced in the thirteenth century alone. Faced with the potential of such a large number of Paris bibles, our network approach above, while provocative, must be considered a preliminary finding.

What is Discoverable? Politics of Mass Cataloguing and Digitization

These numbers are most likely underestimated. What if we try to determine the number of extant manuscripts, or, in our case, to gather data about all the surviving copies of Paris bibles? One of the major obstacles to computational analysis of Paris bibles resides in their discoverability and use. Mass cataloguing and the digitization of sources have significantly altered historical research practices. While online catalogues and digital surrogates have increased the availability of some data, they do not inherently ensure the accessibility of all data. The discoverability of dispersed manuscripts remains a question as many collections continue to be partially or entirely unknown, inaccessible due to both conscious and unconscious decisions made during cataloguing and digitization. What portion is extant, but is only accessed in exceptional circumstances (sometimes called "dark archives")? What portion is digitized and available to select individuals in archives,

65 Buringh, *Medieval Manuscript Production*.

66 Ruzzier, *Entre université et ordres mendiants*.

67 Buringh, *Medieval Manuscript Production*, 261.

or for which a digital copy is available only on request ("dim archives")? And finally, what portion of them exists in some digitized format, easily findable online and openly available for computational use ("light archives")? After all, digitization projects and cataloguing practices sometimes reflect institutional priorities, which, in turn, reflect what is interesting to the public and decision-makers. Other times what institutions choose to digitize may only be what is straightforward to do. Priorities are frequently aligned with national European and North American institutions whose collecting practices were established in the nineteenth century, and support traditional research fields such as literary studies and questions of authorship or provenance. Mass digitization is a more political act that reinforces the existing gendered and racialized structures embedded within both cultural institutions and the tech industry.[68] As such, digitization projects often overlook underrepresented or understudied collections and other methodological approaches.

Fragments and non-European collections are usually more difficult to discover due to varying cataloguing standards and geographical disparities: European and North American institutions tend to be more advanced in cataloguing their collections, thus increasing existing inequalities in access rather than diminishing them. The variegated landscape of digitization shapes the direction of research by providing some materials that are, comparatively, easier to access, while vast quantities of historical sources remain undiscoverable and inaccessible. And yet D'Ignazio and Klein have argued that simply accumulating more data is not always a better practice, particularly when the quality or certainty of the data is questionable.[69] Seen from this perspective, the rush to digitize or expand holdings of collections such as medieval manuscripts can have the inadvertent and undesirable effect of amplifying biases in archival practices. We use a term to describe this aspect of the uneven landscape of digitized medieval manuscripts: collections bias.

Collections bias can easily have a geographic dimension. In the French context for example, from which we drew a number of our working manuscripts, public collections are protected by inalienability laws, preventing their sale or deaccession. Given the multiplicity of public collections, institutions, museums, libraries, and other archive centres holding manuscripts, gathering all the catalogues containing medieval manuscripts is a near

68 Thylstrup, *The Politics of Mass Digitization*.

69 D'Ignazio and Klein, *Data Feminism*.

impossible task. Not all manuscripts have been catalogued and digitization of said catalogues and their availability to researchers are not a given: "paywalls, huge file sizes, and impenetrable catalogues kept all but the most assiduous expert researchers out."[70] To add to the problem, say one could succeed in the former endeavour, there is still a largely unknown part in private collections and smaller or poorer institutions without the means to have a curator, a librarian, or a cataloguer who could make their collection easily available to the public. If we rely on a heterogeneous set of institutions, necessarily the metadata about them will be of variable granularity and accuracy.

What is Usable?

The usability of manuscript collections today involves navigating various tensions inherent within the corpus itself. These tensions are not limited to the manuscripts that have been preserved, but extend to the biases that exist in how collections are curated, catalogued, and made discoverable. In particular, the thoroughness and methodology of their descriptions reflect subjective decisions influenced by institutional, national, and historical priorities, or what might be called methodological nationalism or institutionalism. Methodological nationalism suggests the challenges faced when looking for medieval—or any other—objects within digital collections that have been shaped by distinct national traditions. Catalogue entries for the same type of manuscript can vary depending on the cataloguing standards and historical practices of each country, affecting both the consistency and usability of these records for international research. Thus, the national traditions embedded in cataloguing practices create barriers to cross-cultural research and of objects held in many national collections.

Additionally, within a seemingly uniform national space, such as a very centralized country like France, methodological institutionalism is an obstacle: the institutional context—a museum, library, or university—shapes the study and interpretation of the object. For example, museums may catalogue an object as an art piece, focusing on its materiality and aesthetic details, libraries as a text, and universities for academic use. By focusing on only parts of an object and using different methodologies, vocabularies and ontologies, institutions tend to introduce biases in each case, limiting the discoverability and accessibility of these objects. The institutional setting determines not only what information is prioritized, but also frames

70 Treharne et al., eds., *Medieval Manuscripts in the Digital Age*, 1.

potential objects of analysis within that context. For instance, typical catalogue entries for thirteenth- and fourteenth-century bibles often press the medieval objects to fit modern digital catalogue standards. Whereas libraries attribute biblical manuscripts to the author "Saint Jerome," "Bede the Venerable," or to translators, most museum entries lack detailed material descriptions, such as collation, that are crucial for manuscript studies. Such details are often relegated to specialized listings rather than appearing in more accessible institutional catalogues. In a project such as ours, there is a need to work around methodological nationalism and institutionalism by re-cataloguing and standardizing the metadata obtained as these limitations can be a common hindrance to projects in the computational humanities.

Despite these obstacles, there is a growing need among scholars for flexible and creative ways to create clear and coherent online manuscript descriptions.[71] Although they focus on manuscript provenance research, Burrows and colleagues emphasized the importance of discoverability and reuse, highlighting the need for Linked Open Data (LOD) identifiers to enable effective data integration across collections.[72] Relying on traditional identifiers—such as owner, collection, and shelfmark—is challenging on account of inconsistencies in formatting, sometimes within the same institution. The authors point to the International Standard Manuscript Identifier (ISMI) initiative as a promising step toward creating a universal manuscript identifier, though its progress has been limited so far.

When it comes to images of these manuscripts, many collections remain accessible only on microfilm, especially at long-standing research centres. In the day, microfilming projects were often constrained by the physical limitations of manuscripts, which sometimes affected the quality of copies. Access to manuscripts has evolved from microfilms to digitized copies: some libraries chose to digitize the microfilms, while others chose to digitize the manuscripts again, in high resolution. Some, such as the BnF, did both. However, libraries prioritized digitization based on their resources or the importance of specific works, such as bible codices with unique insignia or historical provenance marks.

71 Andrist et al., "New Digital Strategies"; Koho et al. "Harmonizing and Publishing"; Coladangelo et al., "Leveraging the Power of Crowdsourcing."

72 Burrows et al., "A New Model."

Conclusion

In the mid-2020s, we find ourselves at a pivotal moment that the Paris bible corpus exemplifies particularly well: we cannot access or process every medieval manuscript of interest to us, and yet with the computational resources currently available, we are able to scale transcription efforts in ways that were unimaginable a decade ago. This book offers a reflection on how computational tools and collaborative practices are transforming the study of medieval manuscripts and what still lies beyond our grasp. Standards and methods play an important role in navigating the immense, chaotic world of digitized materials and in analyzing the data they contain. They allow us to impose a measure of order on a fundamentally heterogeneous corpus. Yet given the complexity of the task, especially with a corpus as diverse as the Paris bibles, it cannot be fully addressed by any one scholar or even a small team. It must be a collective endeavour.

Medieval bibles, and Paris bibles among them, are far from uniform, as we discussed in this chapter. The number of volumes, the style of decoration, the systems of lining and ruling, the order of the books, and the choice of prologues can all vary significantly. Descriptions of medieval bibles also typically omit important scribal information: either they do not address the number or identity of the hands involved with a precise description, or they assume multiple scribes without necessarily specifying how many and where they contributed to the codex. When colophons are present, they sometimes preserve the names of individual scribes, but the existing lists of known scribes remain incomplete.[73] For example, Paris, Institut de France, Bibliothèque Mazarine (hereafter Mazarine), MS 6, which we analyze in Chapter 3, preserves a colophon at the end of the Old Testament (hereafter OT): "Explicit Vetus Testamentum per Johannem de Cristemanneford scriptum, cui Deus reddat premium" ("Here ends the Old Testament, copied by John of Cristemanneford, whom God will reward.") Such traces are rare and incomplete, reminders of how much has been lost and how much work remains to be done.

Four major ways that digital technologies are reshaping the humanities will become apparent in the chapters that follow: (1) they transform how scholars interact with primary materials; (2) they introduce new technology-driven methodologies; (3) they alter the relationships between scholars, libraries, and publishers; and (4) they foster collaborative, interdisciplinary

73 Ruzzier, *Entre université et ordres mendiants*.

work across fields and institutions.[74] Our research addresses each of these four shifts, emphasizing that no single actor, whether human or machine, can process the medieval record alone. While we assess new ways to transcribe medieval manuscripts using artificial intelligence in Chapter 2, Chapter 3 implements other computational methods to analyze manuscripts in four case studies. Finally, Chapter 4 looks both toward the past and the future, asking how trends in collaboration might impact the future of medieval studies. Could groups of medievalists collaborate to map, transcribe, and analyze medieval corpora? If so, how would they agree on how to do so and how to share their data? What could non-specialists contribute to the next generation of medieval manuscript studies? Will we be able to craft sustainable, human-centred tools that tap into the power of artificial intelligence for the purpose of studying medieval manuscripts more holistically? These questions remain open, and we hope that this book will begin to stimulate some discussion of them.

74 Brett, "Why the Digital Humanities?"

Chapter 2

TRANSCRIPTION, SCRIBAL MODELLING, AND ARTIFICIAL INTELLIGENCE IN MEDIEVAL STUDIES

WHILE DIGITAL LIBRARIES do not rival fields like healthcare or space science in terms of the sheer volume of data produced and managed, digitized cultural collections have been described as a form of "big data" in their own right.[1] Although there is a natural upper bound to the total corpus of digitizable, pre-modern materials, the landscape is far from static. Libraries and archives around the world continue to release digitized content at varying paces, and there are ongoing efforts to surface materials from dark archives, meaning that new medieval manuscripts are regularly being made accessible. As a result, the field of medieval studies is increasingly confronted with both the challenges and the opportunities of scale, particularly concerning how this vast body of data will be explored, analyzed, and interpreted.[2]

Medieval studies have entered a dynamic new phase in which digitized manuscripts have become sites of both increased experimentation and innovation.[3] The growing accessibility of these materials has opened new avenues for research, including approaches to textual analysis that would have been impractical only a few decades ago. Scholars now routinely combine close and distant reading in creative ways, blending traditional humanistic inquiry and computational analysis. We argue that working with digitized manuscripts needs to be situated within the broader domain of the computational humanities that we have discussed in the previous chapter.

A transformative infrastructural development in the premodern humanities is the rapidly evolving field of handwritten text recognition (HTR), a subset of the broader field of automatic text recognition (ATR).[4] While ATR

1 Marciano, "Afterword," 205–6.

2 This chapter is a reworked and expanded version of a previous publication: Guéville and Wrisley, "Transcribing Medieval Manuscripts."

3 Gilsdorf and Morreale, eds., *Digital Medieval Studies*.

4 Although there are other more specialized approaches, large user groups in the field of ATR have emerged in the last decade around two general ecosystems:

DOI: 10.17302/mfsk15867

refers more generally to the automated processing of printed and digital texts in various formats and layouts, HTR specifically addresses the complexities of interpreting historical handwritten materials, which in the case of medieval manuscripts is notably complex and multi-faceted. The general field of ATR is not only about the process of transcription, however. It combines at least two non-trivial steps of automation: on the one hand, segmentation and layout analysis, by which images of digitized manuscripts are algorithmically divided up into semantically significant subparts, and on the other the identification, decoding and transcription of zones of written text. One action within ATR ecosystems crucial for its use for specialized philological analysis is the ability to train AI models to decode historical scripts of interest to the researcher. For the study of Paris bibles, we are fortunate that the "cognitive layout of manuscript folio layouts" is straightforward.[5] The regularization of the form and layout of the bibles we are studying in two uniform columns—notwithstanding the variance we introduced in Chapter 1—makes segmentation and layout analysis a relatively straightforward task. Were we to move to another related genre, say, that of the glossed Bible, the process of the layout analysis would be significantly more complex.

Although researchers working on forms of automated transcription have not adopted the same approaches and platforms—and these are rapidly evolving—the wider availability of community-based software has made it feasible for inclusion in a broad range of research projects and pipelines. It is increasingly scalable, offering new possibilities for the searchability and analysis of premodern manuscript and book cultures.[6] Indeed, there are many reasons that researchers might want to automate the transcription of the manuscripts. For institutional or multi-institutional research projects, as in the case of a library or archival collections, transcriptions are desirable for reasons of keyword searchability. Finding a one-size-fits-all transcription workflow for processing many different kinds of documents, with different

eScriptorium and Transkribus. See Kiessling et al., "eScriptorium," 19–24; and Muehlberger et al., "Transforming Scholarship in the Archives," 954–76.

5 On the question of cognitive layouts and the challenges that other traditions might face, see Somfai, "Medieval Manuscript Layouts," 1–35; Somfai, "Visual Thinking," 19–27.

6 A wide range of viewpoints on the state of the art for the transcription of historical documents can be found in a 2024 special issue of the *Journal of Data Mining and Digital Humanities*, prefaced by Pinche and Stokes, "Historical Documents and Automatic Text Recognition." See also Nockels et al., "Understanding the Application of Handwritten Text Recognition"; Nockels et al., "The Implications of Handwritten Text Recognition."

layouts and in different states of language, would be challenging, however. By contrast, the individual scholar interested in an unedited text for a book project might choose to transcribe it, instead of waiting for a critical edition to be published someday. A text editor might want to make transcriptions in order to establish a stemma or to create a synoptic edition. More generally, medievalists might turn to automatic transcription to create a first draft of a transcription so that their time is better spent on correcting or interpreting more interesting details of a text. Scholars in certain fields, especially historical research, might choose to work with transcriptions of digitized manuscripts as their main source material as part of a computational workflow, eschewing editions altogether.[7]

Moreover, medievalists with an eye for how algorithms are becoming increasingly embedded in our study of the past, and how twentieth-first century archival digital infrastructures are evolving rapidly, might take an interest in advances in AI for the simple reason that it is fundamentally altering how we access textual culture. Although HTR and other AI-based modes for creating data from manuscripts are not explicitly mentioned by Warren in *Holy Digital Grail*, her critical framework of "tech medievalism" is most definitely relevant in our case. Tech medievalism encompasses a variety of ways that digital infrastructures encode assumptions about history, language, and authority. Warren's analysis of how economic and colonial practices persist in digitization reminds us that the technologies mediating access to the medieval are not neutral. Tech medievalism not only concerns what becomes visible online, but interrogates how the design of digital access systems reinscribes longstanding power hierarchies.[8]

The Paris Bible Project is a case study for the application of handwritten text recognition (HTR) and machine learning analytics in museum and archival contexts, but it also examines how the evolving landscape of automated transcription is reshaping the infrastructures of access and authority that determine what counts as data in the humanities. Crucially, the success of such projects depends not simply on the successful application of AI, but on the informed contribution of critical medievalists—scholars whose deep expertise in the structure, content, and context of the manuscripts is indispensable for training models to decode handwriting, and whose attention to shifting media environments keeps us attuned to evolving nature of tech medievalism. As scholars have noted, a vast amount of valuable knowledge

7 Hodel, "Das Ende der Edition?"

8 Warren, *Holy Digital Grail*.

is embedded in and around cultural archives—knowledge from which AI systems also stand to benefit.[9] While much of the current discourse around AI centres on the capabilities of large-scale, general-purpose models with commercial appeal, these approaches can be ill-suited to the specific needs and interpretive aims of humanities research. In this book, we argue instead for the value of smaller, domain-specific datasets, curated and shaped by human expertise. These models are not only more practical for specific projects, but are also more ethically and epistemologically grounded, reflecting the particularities of the cultural artifacts that we design them to work with, and reminding us that interpretation, not automation, remains at the heart of humanistic knowledge-making.

Archives are not merely inert locations of cultural data, but are becoming active sites of computational innovation. They provide richly contextualized, expert-curated material that can inform specialized and transparent practices in ML and AI practices.[10] Moreover, scholars who work with cross-archive problems such as the study of a genre distributed across global collections, like our own investigation into Latin bibles, are acutely aware of the challenges of assembling data from different sources in meaningful, reliable and sustainable ways. Our research engages squarely with debates within the computational humanities, which advocate for approaches that are critically engaged, power-aware, and interdisciplinary, attuned to ethical use of computation in cultural analysis.[11] Rather than relying solely on large, generic corpora or corpora that have already been created for us—most often detached from their context and relevant interpretive frameworks—researchers are calling for bespoke, curated and adaptable datasets that arise from a rich dialogue with domain expertise.[12] We believe that applications of HTR exemplify the possibilities of working with smaller, meaningful datasets with critical reflexivity. Although some might consider HTR systems to be "stochastic parrots"[13] that generate output based on probabilistic models without grounded reasoning or genuine understanding, we do not

9 Thiel and Bernhardt, *AI in Museums*, 24.

10 Jaillant, "Introduction."

11 Tilton et al., "What Gets Counted," vii–xviii.

12 There is a paradox in the desire to create bespoke datasets that we will discuss in Chapter 4, namely the citation advantage, by which is meant that scholarship that centres certain kinds of practices (open access, publication in pre-print in an institutional repository or a high-impact journal, in English or with a large group of international/collaborative co-authors) ends up with more citations over time.

13 Bender et al., "On the Dangers of Stochastic Parrots."

choose to see them (or use them) this way. HTR platforms have ushered in opportunities for methodological experimentation, allowing medievalists to exert control and precision in both the model and data creation process using medieval manuscript collections. Deployed thoughtfully, these tools do not replace humanistic interpretation but extend its reach, opening up new possibilities for the analysis and circulation of historical texts.

As medievalists, we engage with automated transcription technologies for a clear and compelling reason: the variations from hand to hand and manuscript to manuscript, when counted across thousands of folios and codices, offer an immense and textured body of textual data which enables new forms of inquiry into the modes of production and transmission of handwritten objects. Rather than succumbing to what has been called "buttonology"—the superficial use and teaching of tools without critical engagement—medievalists must develop a nuanced understanding of how HTR works, what it can and cannot do, how to shape it to serve specific scholarly goals and how to teach it to new generations of researchers.[14] The AI behind models we train to recognize medieval Latin in different hands is undoubtedly sophisticated, but remains, at its core, profoundly artificial. Engaging with it, we confront what Andler has called the "double enigma": we are compelled not only to understand the nature and limits of AI, but also to grapple with the complexities of human (and scholarly) intelligence.[15] Human intelligence for Andler is not merely about problem-solving with data, but involves consciousness, contextual understanding, and affective dimensions, the importance of which all medievalists who work with manuscripts can easily appreciate.

Working with HTR systems to create handwriting models requires a significant reconfiguration of expertise amongst medievalists. Machines are able to surface patterns in archival documentation, but they lack the nuance of understanding to interpret them with rigour. One set of algorithms can be trained to pay attention to graphetic differences, but they cannot assert anything about the material intricacies of the material object (say, flesh, hair or ink) from a digitized copy.[16] Any given technology-centred approach must be trained to detect the specific features of value to a medieval research project, and it may be that for some details the logic of machine learning may not be practical at all. In our case, it is an appropriate match for studying varieties

14 Russell and Hensley, "Beyond Buttonology."

15 Andler, *Intelligence artificielle, intelligence humaine*.

16 Treharne, "Fleshing Out the Text."

of graphetic difference, but it must be underscored that training a machine that has never had experience carrying out these tasks relies intimately on informed human judgement. Without such expert input, general-purpose AI tools will offer limited benefit to the field.

Were we to adapt one of the recent critical provocations from the humanities for generative AI advanced by Klein and colleagues but are also ("Models make words, but people make meaning") to the context of HTR technologies, it would need to be split into two parts. First, "HTR systems transcribe words from historical documents, but people make the HTR data on which those models are trained" and second, "HTR can transcribe a lot of words, but scholars make meaning from those transcribed words."[17] As such, we believe the best posture to adopt as a community of medievalists with respect to forms of applied AI such as HTR is as a complementary tool that enhances research without supplanting human expertise. When these models are deployed with critical reflexivity, they open up many new research pathways. The scalability and repeatability of automated transcription with HTR, we believe, has potential to change manuscript studies deeply. It enables the creation of a large volume of relatively consistent transcriptions at a pace far greater than what is possible through manual labour alone.

Our Kind of Big Data: Transcription at Scale

Having many pages of transcriptions from different medieval bibles might seem like a counterintuitive goal, since the text of the various bibles would be for the most part the same. When we factor in all of the different orthography and abbreviations in the manuscript—the kinds of detail that material philology encourages us to study—nothing could be farther from the truth. In the Paris Bible Project, we use HTR to automate the process of transcription and then methods from classical machine learning to carry out our analysis. ML and AI for medieval texts has particular promise, we believe, for the comparative analysis of scribal hands, quires and other material and textual features of manuscripts, but it does rely on a generation of medievalists both aware of trends in data science and able to co-design computational experiments. Far from displacing expertise, as we have argued in the previous section, ML and AI can be leveraged to work on larger, complex problems in manuscript studies. It becomes possible not only to describe manuscripts by creating more data about them, but also to attempt a new

17 Klein et al., "Provocations from the Humanities."

understanding of them, perhaps to categorize them according to common features they possess. Computational methods are not foolproof, but they are promising for how they help us to confirm, nuance or overturn human-created hypotheses about how manuscripts were put together.

Features such as letterforms, abbreviations or other punctuation marks vary from codex to codex and from scribe to scribe. They are measurable and countable features, which comprise a unique imprint of the hand-copied documents.[18] While documenting these features across a large number of codices is complex and labour intensive, doing so opens the door to the possibility of a medieval big data that can reveal how scribes, and groups of scribes, copied their documents. As we argued in Chapter 1, at the time we are writing, it is impractical to assemble a digitized corpus of the thousands of manuscripts of the Paris bible tradition for computational analysis. The sheer scale of thousands of manuscripts is not what makes it impossible to build a research corpus; it is the varied archival infrastructure of all the institutions holding them. They have not all been digitized, and when they have been, the state of digitization or the access to the copies sometimes impede us from using them fully.

To offer our readers a notion of scale, however, from the full automated transcription of only one manuscript, we are able to transcribe many words and even more characters. Let us take the example of Cambridge, Corpus Christi College (hereafter CCC), MS 49, a manuscript that we describe in detail in Chapter 3.[19] Remembering that a Paris bible is typically presented in a two-column format of about fifty lines of text each, an automated transcription of a full manuscript generates on average about one hundred thousand lines of text. Including hyphenated words at the end of the line, that same text contains about 665 thousand words, of which about 102 thousand are unique. In total, a transcription of a similar manuscript would be made up of about four and a half million characters (including spaces). By comparison, the medieval French edition of Christine de Pizan's *Le livre de la cité des dames* contains about 120 thousand words, and the Latin edition of Augustine's *De civitate Dei* comes in at about 275 thousand words.

Given how variable the abbreviated Latin text of a Paris bible is, with each line in the manuscript having a half dozen or so variants, many millions

18 The idea of measurable and countable features makes up what some in computational textual studies have deemed the way that style can be redefined. See Herrmann et al., "Revisiting Style," 25–52.

19 Guéville and Wrisley, "From Localization to Chronological and Geographical Prediction."

of words and characters can easily be created. Only a few decades ago, brought on by a "face-to-face confrontation with the medieval manuscript," it became clear that new philology was opening the door to a range of new interpretative problems in medieval studies.[20] Whereas in the 1990s creating big data from manuscripts was probably not on the mind of most, we now face a research environment where automated transcription is a mature process. In the age of HTR, AI and ML, transcribing a few dozen, or even a few hundred manuscripts, is within the realm of possibility in the timespan of a PhD dissertation. We believe that these developments should provoke a sense of excitement, but also of urgency.

There is much to be gained in accumulating more research data, but scalability is not without its discontents, as it requires careful attention to method. We concur with Hodel and colleagues on this important point, and we contend that using a computer both to transcribe and to analyze documents requires medievalists to pay significant attention to material detail about manuscripts.[21] Put another way, when studying manuscripts computationally, we must pay careful attention to the *manu-scriptedness* of our book-objects. We must decide which of those details specific to handwriting are most important, so that we can embed them in our transcription norms. It goes without saying that these features will vary according to language and text tradition. From the plethora of details found in the book-object, we must decide which ones we will capture and which ones we will not.

Incorporating ML and AI into medieval studies necessitates important changes in the ways that we interpret our objects of study. Automation can capture forms of evidence such as the presence of a given set of abbreviations or letterforms in manuscripts, but we must underscore that automated transcription technologies are not able to capture an infinite number of unique features at once. If we choose to use such tools—and not all medievalists will choose to do so—we must set up our research projects to take advantage of methods we have at our disposal, while also being accountable for their shortcomings. To borrow an expression from business computing, HTR is most definitely not a "turn-key solution," that is, it is not a tool that we can seamlessly integrate into current research practices. While it does indeed automate the process of manual transcription, adopting it also disrupts the ways in which we might typically work. HTR changes the way that we think about, and work with, archival documents as research objects.

20 Nichols, "Why Material Philology?," 11.

21 Hodel et al., "General Models."

It is not enough to focus on mastering new techniques from data science simply to introduce new scales at which we can research.[22] As researchers in medieval studies, we must reflect critically on the creation of our research corpus between and beyond individual digitized collections, assessing how representative they are to our subjects of study. HTR models embody the kinds and forms of knowledge we hope to reproduce in our automated transcription, striking a compromise between all the information we would ideally capture and the most important information for our research questions. HTR systems are capable of creating an abstraction of the handmade object,[23] and they are increasingly being used in digital research pipelines. HTR is best grounded, however, in the specificities of historical production. This need becomes especially urgent in our study of the collaborative scribal work we began to detail in Chapter 1. By collaborative, we do not mean in the codicological sense: creating the parchment, lining the folios, illuminating the folios, binding the codex, etc. Instead, we are interested to what extent we can understand the *collaborative effort of copying manuscripts*. In an age of computational criticism, our transcription systems must be reimagined to foreground the codex not merely as a textual container, but as a layered artifact accessible through the combined lenses of material and digital philology.[24]

Diplomatic Transcription: Scientific Method or Movable Feast?

Transcribing Without Normalization

In this section, we focus on transcription as a scholarly encounter with written, rather than oral, sources. All actors who participate in the transmission of texts, medieval scribes as much as modern editors, engage in some form of transcription, leaving a detectable trace.[25] As we suggested in the previous section, transcriptions are made for a variety of reasons, and they allow for the dissemination of language in media beyond their original form, and ultimately include some form of normalization. Transcribers follow practices which embody assumptions about textuality and unacknowledged norms, catering to literacies of their assumed audiences. Sometimes, although not always, transcriptions made from historical sources become part of editions. Editors, publishers and professional societies will often set forth such

22 Jaillant, "Introduction," 8–9.

23 Widner, "Toward Text-Mining the Middle Ages."

24 Guéville and Wrisley, "Transcribing Medieval Manuscripts."

25 Guéville and Wrisley, "Everyone Leaves a Trace."

norms to do so, again pragmatically corresponding to issues and challenges found both in sources and in the communities to which they belong.

The ability to define new transcription practices, and to adapt one's traditional ways of working with agility in order to adopt new norms for automated transcription, not only has great potential in a future of computational medieval studies, but, we believe, is a necessary component of its success. In this section, we turn to a method known as diplomatic transcription, a term commonly used by many who work with historical documents. Diplomatic transcription ostensibly takes its name from a field known as diplomatics within the archival sciences, born from pre-modern schools of critical documentation that focused on the understanding of the complex modes of historical document creation. Diplomatics has been described as the analysis of the "genesis, inner constitution and transmission of documents, and of their relationship with the facts represented in them and with their creators."[26]

The basic idea of a diplomatic transcription could be said to be an attempt at capturing forms of graphic representation found in archival documents that vary from the typographic norms of technologies in which a given researcher is working; the perspective of the diplomatic transcriber, it has been claimed, is that of the historian.[27] Although the purpose of such detail in the case of diplomatics has traditionally focused on the authentication of the large number of short legal and official documents that circulated in the medieval world, the clear division between the focus on detail in either manuscripts or documents seems to have faded in the age of digital editing.[28] We suspect that this distinction will become increasingly blurred in the age of automated transcription.

Another school of scholarly editing, genetic criticism, relies on diplomatic transcription to document and interpret the process of textual creation tracing the reworking of a writer's materials—drafts, notes, and revisions—into a final text, offering a meticulous presentation of these sources to illuminate their variations.[29] These critics focus on different features of documents for their documentation. Diplomatic editing sometimes respects the spatial layout on the page of all original graphic elements of documents, but these

26 Giorgio Cencetti, "La Preparazione dell'Archivista," 285; cited in translation by Duranti, *Diplomatics*, 7.

27 Bourgain, "Sur l'édition des textes," 5–49.

28 Gallo, *Diplomatics*.

29 Shillingsburg, *Scholarly Editing*, 174.

original elements do not all have to be placed in the space of reading—some can be relegated to an apparatus.[30] Line breaks are a key element of transcription, along with the documentation of errors and abbreviations.[31] For some, a diplomatic transcription is a record of textual genetics via a machine adaptation (typing or typesetting) of an original manuscript.[32] D'Iorio offers a synthesis of these perspectives, underscoring the importance of fidelity to page layout, and specifying particular ways that machine adaptation might achieve a mimetic effect: font size of characters, font colour for types of ink, orientation of writing, or the use of certain characters to imitate or stand in for other diacritics or writing conventions.[33] D'Iorio goes on to define a related process, the ultra-diplomatic transcription, that blurs the boundary between facsimile and transcription, with congruent typographical characters used in the place of all glyphs in historical writing. Diplomatic transcription is indeed a diverse editorial field that asserts the importance of the unique historical and documentary nature of the object, while rejecting fully contemporary spelling and layout in favour of the alterity of historical writing.

Diplomatic transcription styles are highly variant, and are rooted in the assumptions of projects, or the conventions of textual genres and languages. In the 1990s some medievalists were attempting to theorize the concept of "levels of transcription," to come to grips with the complexities of medieval textual scenarios.[34] Additional pressure had been placed by literary critical approaches such as New Philology to include "historical context by privileging the material artifact(s) that convey this literature to us: the manuscript."[35] Print formats imposed numerous limitations on how much historical detail editions could include, so print editors needed to choose what to represent. Pierazzo famously asserted that the rise of digital editing became so permissive, that "editors need new scholarly guidelines to establish 'where to stop.'"[36] In the 2020s with the situation of automated transcription using HTR technologies, as we mentioned above, many different graphic or

30 Grésillon, *Éléments de critique génétique*; Kline, *A Guide to Documentary Editing*.

31 Plachta, *Editionswissenschaft*.

32 Shillingsburg, *Scholarly Editing.*

33 D'Iorio, "Qu'est-ce qu'une édition génétique numérique ?," 49–53.

34 Robinson and Solopova, "Guidelines for Transcription," 19–52.

35 Nichols, "Why Material Philology?," 11; Rigg, "The Editing of Medieval Latin Texts."

36 Pierazzo, "A Rationale of Digital Documentary Editions," 463.

graphetic features can be captured, but the technologies themselves impose limits on how far we can go.

It is therefore useful to examine the various ways transcription features have been categorized, particularly in relation to documentary-style editions. While the concept of transcription levels is not fixed—and may never be—general patterns emerge. Most frameworks recognize three principal levels, ranging from the most diplomatic to the most normalized, with intermediate gradations. The diplomatic level closely reproduces all visible features of the archival page, while the most normalized level typically expands abbreviations, regularizes layout, and applies editorial conventions to produce a more standardized text. Normalization, however, is not a neutral or purely objective process. It entails a series of editorial decisions—both conscious and unconscious—that shape the representation of the text. When poorly executed, this process risks obscuring linguistically significant features, making it difficult for readers to distinguish what originated with the scribe from what was introduced by the editor.[37]

Taken together, these perspectives underscore a significant shift in the editorial priorities of some digital medievalists—from producing streamlined, readable texts to capturing the complex, layered materiality of manuscript sources in ways that respect their historical specificity. Transcription without normalization, or without full normalization to be more exact, as we are suggesting, resists the erasure of scribal practices. Rather than concealing variation for the sake of uniformity, it embraces irregularities, abbreviations, and idiosyncrasies as meaningful data, both for the historian and the philologist. Digital encoding standards such as Unicode and MUFI have made it increasingly feasible to transcribe texts at the character level without imposing modern orthographic norms, enabling representations that preserve abbreviation systems, unusual characters, and diacritical marks.[38] In this way, contemporary digital infrastructure revives and extends the possibilities once offered by facsimile-based reproduction technologies like lithography and hectography, but with vastly greater precision, consistency, and analytical potential.

Our approach aligns with a growing methodological interest within digital and computational philology to document instead of overwriting textual variation. Within this paradigm, transparency becomes central: diplomatic editions must clearly define which features of the manuscript

37 Cugliana and Barabucci, "Signs of the Times," 7–8.

38 MUFI: The Medieval Unicode Font Initiative.

are preserved—spacing, punctuation, erasures, marginalia—and which are standardized or omitted.[39] Such detailed encoding serves not only scholarly rigour but also renders texts amenable to computational analysis, fostering new forms of inquiry into textual transmission, scribal habits, and linguistic change.[40] Moreover, the convergence of these practices with advances in digital palaeography has opened the door to a more systematic study of historical scripts and scribal hands, enabling the creation of complementary datasets and tools that support the identification, classification, and documentation of scriptoria and individual scribes across corpora and collections. This point is especially important since scribes could be proficient in multiple scribal styles and could shift scripts depending on context, commission, or material constraints. Furthermore, a scribe's hand might evolve over time due to factors such as age, illness, or training, complicating efforts to link script to identity.

Implementing Levels of Transcription: Three Case Studies

It is beneficial to scrutinize approaches in editing to see how specific textual details can be translated into explicit levels of transcription.[41] The examples below demonstrate how the same types of information can be organized differently, depending on editorial choices and their underlying frameworks. While the genetic-critical approach offers important insights, its application in medieval studies presents unique challenges due to the inherent fluidity and variance of the documentary base and the variety of historical states of language.[42] Whereas normalized methods have been established for Middle

39 Li, "Critical Diplomatic Editing," 305.

40 Piotrowski, *Natural Language Processing*; Widner, "Toward Text-Mining the Middle Ages."

41 While the examples in this section are all digital examples, it is worth noting that there is a print tradition of representing medieval notional systems in typeset text in late nineteenth-century German-born medievalists that is worthy of scholarly attention. In an era before widespread reproduction of texts with microfilming or digitization, documentary editions were being made of texts in manuscripts, while being attributed the status of a "linguistic monument" (*Sprachdenkmäler*). They used typeset notation to approximate medieval *scripta* with a dual purpose in mind: to teach students to read in manuscript and to preserve the geographic and temporal specificities of the witness. Two works that theorize and demonstrate this editing style include Koschwitz, *Les plus anciens monuments*, and Koschwitz, *Commentar*.

42 Zumthor, "Intertextualité et mouvance."

High German by Karl Lachmann, similar norms are not only still missing for Early New High German, but also not sought after.[43]

Some projects create more granular distinctions. The Canterbury Tales project has been producing transcriptions of all known manuscripts of Chaucer's work since 1996 as part of a systematic genetic study of the text. In the "Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue" project editors defined not three, but four levels of transcription: graphic, in which "every mark in the manuscript, every space, is represented in the transcription, to the point of decomposition of letterforms into discrete marks"; graphetic, in which "every distinct letter-type is distinguished (as: r 'short' is transcribed apart from r 'round' and r 'long descender,' etc.)"; graphemic, in which "every manuscript spelling is preserved (as: 'she,' 'sche') without distinction of separate letterforms as in a graphetic transcription"; and regularized, in which "all manuscript spellings are regularized to a particular norm, perhaps the spelling of a manuscript considered authoritative." In 2021, Bitner and Kyle revisited such transcription guidelines and updated them, notably to include additional encoding characters and TEI XML elements.[44]

On the other hand, some have called for a reduction of transcription to two main levels: "graphique" (graphical) and "allographétique," (allographetic) emphasizing the distinction between form and meaning in palaeographic analysis.[45] Graphical transcription replicates visual features of manuscripts—combining metadata, photos, coordinates of letterforms—prioritizing form over meaning and enabling detailed palaeographic datasets useful for scribal hand recognition. However, Stutzmann notes that perfect graphical replication is impossible and requires advanced technological resources. Allographetic transcription, by contrast, classifies letter variants (allographs) using a controlled vocabulary, offering structured descriptions over transcriptions. Aligned with methodological claims of New Philology, it focuses on script diversity, but demands methodological rigour and standardization for target corpora, unlike simpler graphemic approaches.

The transcription of the Early New High German Marco Polo introduces three levels—diplomatic, semi-diplomatic, and interpretative—focused largely on punctuation.[46] At the diplomatic level, original punctuation and

43 Cugliana and Barabucci, "Signs of the Times," 6.

44 Bitner and Dase, "A Macron Signifying Nothing."

45 Stutzmann, "Paléographie statistique."

46 Cugliana and Barabucci, "Signs of the Times," 14.

allographic features are preserved and encoded with Unicode for precise analysis in stylometry and phylogenetics. The semi-diplomatic level simplified punctuation to a middle dot and virgula, expanded abbreviations, and corrected trivial errors while retaining key graphemic features. The interpretative level modernizes punctuation and graphemes entirely for readability. Unlike Stutzmann's form-based approach, this model balances readability with analytical depth by layering transcriptions for different uses. Its feasibility, however, depends on the manageable scale of the corpus.

As these examples suggest, little consensus exists on what constitutes a level of transcription, or what such levels should include. Instead of constituting a one-size-fits-all standard, transcription practices tend to align with specific research questions and the anticipated reuse of project data. They are shaped not only by available resources and scholarly expertise, but also by the coordination and interoperability of technical infrastructures. The rise of digital tools has significantly expanded the possibilities for transcription, with technologies such as HTR and guidelines such as TEI now supporting more nuanced diplomatic representations of both authorial and scribal contributions. As McGillivray foresaw in 2005, image-recognition-based transcription tools have helped to reduce the labour intensity of editing, shifting labour to new challenges and opening up new avenues for scholarship. She points to a (then) emerging field of "the study of scribal behaviour and analysis of the interrelationship between the writing system used by a scribe and the process of production of the manuscript."[47] Almost a decade before McGillivray, Robinson argued that future scholars will inevitably require greater granularity, creating transcriptions that preserve details we have traditionally overlooked. The long-term value of such transcriptions remains an open question, as the notes of a text editor in creating a critical edition may also be. He predicted that in 100 years (i.e., in the then-year 2097), legacy editions may hold insignificant value except as archival curiosities, surpassed by those capable of capturing the full material complexity of the text.[48]

47 McGillivray, "Statistical Analysis."

48 Robinson, "What Text Really Is Not," 50.

Medieval Ground Truth

Elsewhere we laid out some adaptable general guidelines for creating a transcription scheme for medieval manuscripts.[49] In them, we underscored that at present there are only so many distinctions that transcription systems are able to make beyond modern alphabets, and practitioners of HTR limit additional characters to about forty or so beyond those typically used in any given alphabet set. Our general approach is to encode allographetic variance (following Stutzmann) by using special characters and combining characters in Unicode to approximate the special letterforms, brevigraphs and abbreviations found in our manuscripts.

Currently AI-based HTR technologies do not automatically generate unnormalized diplomatic transcriptions capturing the kinds of variance visible on the document page; instead, they need to be trained to do so. The notion of *ground truth* refers to the data that scholars create from examples of digitized manuscripts that serves as the reference standard for the training process. In practice, ground truth is painstakingly prepared line by line and encodes not only the words themselves in a normalized orthography, but can also include abbreviation choices, forming the baseline that defines what correct recognition means. It is on the basis of this agreed-upon ground truth that we are able to provide a quantitative assessment of how well HTR-created transcriptions perform.[50] Some groups of scholars publish their ground truth for others to use for purposes of time-saving or reproducibility. Other scholars might also document the details of the extended character set they have chosen to replace allographetic features in manuscript, in the interest of transparency about decisions they have made about modelling characters in historical documents.[51]

---

49 Guéville and Wrisley, "Transcribing Medieval Manuscripts." Since such transcription guidelines inevitably evolve over time, we chose to provide a summary, rather than a definitive version. The guidelines can be found on the website of the Paris Bible Project: https://parisbible.github.io/guidelines/.

50 See Guéville and Wrisley, "Everyone Leaves a Trace."

51 For some time, users working with HTR have been sharing their models and data in a Zenodo community named OCR/HTR models: https://zenodo.org/communities/ocr_models. Two examples from that community illustrate the unevenness in description for the encoding decisions which are included in models and ground truth. On the one hand, in the data paper for v0.1.2 of the CREMMA Medii Aevi data, Clérice et al. included a detailed mapping out potential variants of medieval Latin brevigraphs and abbreviations using sample images to the Unicode characters used to encode them (Clérice et al., "CREMMA Medii Aevi"); on the other hand, Weil published a model for German, but simply lists the Unicode codepoints that were

Extensive numbers of pages of digitized archival documents paired with manually created transcriptions made by researchers are the basis of what are called "HTR models." When humans make such transcriptions for the HTR system to align with the images and to learn, of course, they do not always agree on the same reading in the text, or at a more basic level, they also can make mistakes, since specialists do not always see the same letters when faced with a variety of examples of premodern handwriting. In a project in which a bespoke model is being assembled to transcribe many pages, typically a few dozen pages of ground truth are manually created. Such ground truth and the transcription norms that are contained in it (for example, what pen stroke is equivalent to a macron, or what Unicode codepoint will represent the pen stroke replacing either the letters "er" in a word or the rotund r) usually emerge after researchers have studied their target documents in depth.

Applied computer vision technologies such as HTR work because they have "seen" many different examples of handwriting and have learned to predict and to transcribe what is written on other digitized texts that the system has not yet encountered. They do so with variable accuracy depending on how different the "unseen" data are. One of the persistent issues with HTR models is that they tend to overfit—that is, they perform well on the specific manuscripts or scribal hands they were trained on, but falter when confronted with new, unseen examples. An algorithm tends to recognize and reproduce patterns it has already encountered, including in places where those patterns do not exist. Even though transcription norms are designed to be as inclusive and adaptable as possible, they are still shaped by prior examples and the contours of existing data. Models inevitably carry the biases of their training sets, a limitation that becomes especially apparent in our case when a model trained on a handful of Parisian bibles from northern France is used to transcribe a manuscript from southern Italy, where scribal practices differed markedly. A major challenge, then, lies in how we adapt HTR-based transcription norms to accommodate the variation in scribal behaviour across manuscripts. This way of thinking about training HTR models can appear at odds with traditional humanities training that has long emphasized the value of attending to the exceptional, the unexpected, and the rare. Scholarly instincts notice what resists the pattern—not what conforms to it. Attending to this tension—between the drive for uniformity

used for training leaving a mapping to be done by users of the data (Weil, "HTR Model for German Manuscripts"). In HTR-United, a commons for publishing metadata for ground truth we will discuss in Chapter 4, there is a field for transcription guidelines that is also used by the various contributors with differing degrees of precision.

in our models and the pull of singularity in our sources—can sharpen our sense of what computational methods can reveal when placed in the service of the humanities. Moreover, it underscores one of the critical stakes of computational humanities: how we choose to define, encode, and ultimately value differences in the data we create and the analysis we do. At stake here, as in other domains of artificial intelligence, is the persistent question of bias: whose data gets represented, whose practices are normalized, and whose differences are elided.

Given the challenges of preparing and analyzing data directly from manuscript, our approach to creating ground truth and HTR models has followed what has been deemed the "single book paradigm," by which is meant that using digital surrogates of the manuscripts under study, we proceed one by one, using HTR models with ground truth sourced from those very same manuscripts.[52] Put more simply, we train the computer to recognize samples of the general kinds of handwriting found of the dozen or so manuscripts in question and then we ask the computer to generalize based on what it has seen thus far. This single, or in reality, small group, book paradigm provides acceptable results for the kinds of analysis we do. Creating domain-specific training data imitates in this respect the close analysis that one might do with careful examination of individual bibles. It is, in effect, an attempt to reconcile the machine's need for stable regularities with our humanistic commitment to retaining the unpredictable traces that large models, optimized for generalization, might erase—precisely the traces that make fine-grained philological analysis possible.

We do encounter variance in the allographetic behaviour found across our manuscripts. Some of these letterforms and abbreviations prove remarkably stable, while others fluctuate in their usage. In practice, we only needed to retrain our HTR model a handful of times before it produced transcriptions of sufficient quality for the analyses we describe in Chapter 3. The slippage of certain abbreviating practices is, in our view, an inevitable feature of premodern documents: manuscripts are handcrafted objects, and copyists trained in different places and on different exemplars inevitably carried forward their own scribal habits. To some extent, such variation can be accounted for as our research progresses, but a project encompassing a wider range of genres or regions would necessarily involve explicit interpretative decisions about how to normalize or preserve scribal difference. In our case, the relative stability of our corpus allowed us to encode what we

52 Murel and Smith, "Computational Bibliography."

saw directly in the manuscripts, and then, during post-transcription analysis, to decide whether to retain or to exclude the more unstable forms.

It is clear that no diplomatic transcription can fully capture the complexity of the handwritten page. The instability of medieval orthography and punctuation helps explain why modern readers and critics often prefer the normalized clarity of critical editions. Abbreviation practices and orthographic variation are only one set of the many challenges that confront an editor, but they illustrate how creating ground truth from scribal copies inevitably parallels the decisions made in textual criticism. Transcribing for a machine to learn introduces its own set of ambiguities, ones that bring us face to face with the interpretive labour embedded in every editorial act. For example, it is relatively easy to understand a graphetically unclear abbreviation by knowing the underlying Biblical text and understanding its grammar. While transcribing based on prior knowledge of the text seems to resolve the question of the instability and adds more consistency to the ground truth, it is, in fact, a form of editorial choice. The critical reflexivity which we spoke of above as a necessary approach to all computational humanities extends to the specific details of transcription that allows us to move between writing systems—medieval and digital. The editorial work of ground truth creation ends up embedding itself in models used later to automate transcription, and any critical decision-making in ground truth necessarily creates bias, albeit subtle, that perpetuates itself through any analytical process.

Despite the rigour that a term like ground truth suggests, HTR models remain inherently imperfect, and achieving a character error rate (CER) of 4 to 5 per cent is often considered a success. These error rates are typically calculated by comparing model output to ground truth data, but they fail to capture a range of other challenges that may compromise transcription quality. A newly encountered or particularly difficult scribal hand, folds or damage in the parchment, holes in the manuscript, or image degradation introduced by digitization—from original object to microfilm to digital surrogate—can all introduce noise that affects the model's performance. In practice, working with HTR models means accepting that transcriptions will always involve a degree of error. A model trained on a particular manuscript or scribal style may not extend well to others, and the ability to produce high-quality transcriptions often falters when applied beyond the specific parameters of the training data. Developing scholarly models of scribal practice for a given genre, period, or language is therefore as much an interpretive art as it is a technical process. But how we approach the imperfection of those transcriptions determines whether we treat error as a limitation or as a site of critical insight.

HTR systems are examples of AI infrastructure for which having simply more data does not necessarily make better models.[53] Setting up an HTR pipeline requires significant expert human knowledge that contextualizes the specific documents in question, aligns them with the digitized images and then chooses how to encode the variance one finds in handcrafted book objects. And yet as anyone who is trained in codicology and palaeography can attest, when we open a manuscript for the first time (or access a digitized copy), all kinds of details jump out at us. It is not possible to capture all of that detail in diplomatic transcription. The best possible model of Paris bible scribal practices would include a manual transcription of the many thousands of extant Paris bible manuscripts and take into account every character and every stroke of a pen on the parchment. And yet such a model is an absurd proposition, even computationally. Luckily, our ability to use the HTR-created imperfect transcriptions of manuscripts is facilitated by the exigencies of the analytical methods we employ. Some statistical approaches to authorship attribution have been demonstrated to be successful on typeset texts containing high error rates, up to forty per cent, with some research carried out on HTR-created text.[54] Knowing how accurate the output of the computer needs to be, or how many different samples one needs to include in ground truth to mitigate bias, are both judgement calls in research, often made in the face of the pressures of time or funding, or both.

In sum, modelling scribal practices through computer-assisted transcription involves balancing three key factors: the accessibility of manuscripts, the aspects of handwriting that are most practical to capture using available technologies, and the features deemed most relevant to the research questions at hand. We must focus on this delicate balance as we approach computational analysis with text transcribed from manuscript: from a close examination of the source materials used to train the HTR model, through to a critical evaluation of the automated transcription and any necessary fine-tuning, and to the computational analysis of the transcriptions. Indeed, ground truth is a problematic concept for medieval textuality. While the term designates reference data obtained through empirical examination and verified by human intelligence, and created as a benchmark against which the reliability of computational predictions or automated processes, the *mouvance* of the hand-copied text in medieval manuscripts constantly requires

53 Klein et al., "Provocations from the Humanities" address this question, building on boyd and Crawford, "Critical Questions," 662–79 evoking the provocation "bigger models are not better models."

54 Eder, "Mind Your Corpus"; Franzini et al., "Attributing Authorship."

interpretation and inference. This is not to say that ground truth for HTR cannot be trained for medieval manuscripts—it has been with relative success—but we must be careful in the assumptions that we make while crossing the writing system divide. Retraining, and the human labour involved in creating ground truth, help mitigate bias; as more manuscripted objects are processed, however, it becomes increasingly clear that continual adjustment and refinement are necessary, and that some of the specificity of the individual manuscript will be lost.

Automating Material Philology: Generating Data about Scribes with HTR

Transcription Choices for Computational Research

As outlined in Chapter 1, institutional collections and scholarly corpora pursue different goals in their cataloguing and transcription practices. This divergence is particularly evident in how each approaches the representation of manuscript features. Institutional collections, driven by a desire for access and discoverability, often prioritize standardization over the preservation of the unique, material characteristics of individual manuscripts. In doing so, they might choose to suppress the "manuscripted"—the specific, sometimes idiosyncratic, details of script, layout, and orthography—in favour of searchability, keyword spotting and/or general user usability. In a discovery interface visibility would most likely prevail over philological detail. In contrast, the goals of computational philology—such as modelling scribal practices or analyzing textual variation—require fidelity to such details, often calling for more granular, diplomatic transcriptions. These distinct objectives exist in tension. Moreover, the absence of a widely adopted standard for encoding such variations complicates the matter further.[55] For some, the mere availability of data is considered more important than its abstraction: "something that exists in reality is better than something that exists in the mind."[56] Nonetheless, for both cultural institutions and researchers, HTR technologies offer new and powerful ways to support visibility, at the same time that they raise questions about the prioritization of accessibility or scholarly detail. The resulting data of large scale institutional HTR transcriptions could, in many cases, be difficult to use within specialized research projects.

55 Vander Meulen and Tanselle, "A System of Manuscript Transcription."

56 Gibbs, "New Textual Traditions."

Scholars have definitely begun to see possibilities in using HTR with medieval manuscripts. It is now becoming possible to research the linguistic substrata of hand-copied texts and to assess the input of scribes in the creation and transmission of texts; but to accomplish this goal, specialized workflows are necessary and access to data across different archives is essential. As HTR technologies become increasingly integrated into manuscript studies, it is important to reflect on the new possibilities they offer for analyzing manuscripts at scale, but also how they challenge traditional methods of transcription and editorial decision-making. In particular, the process of making decisions about transcription for HTR—deciding what textual features to preserve or normalize—forces scholars to answer questions about the kinds of distinctions we want to make visible in our data, and how these choices shape both the transcription process and the research that ensues.

Transcription has long been a cornerstone of digital projects in medieval studies, but it has also acted as a rate determining step, limiting the pace and scope of research. We mentioned above the kinds of features encoded by the long-standing Canterbury Tales Project. To our knowledge, the project has not yet employed HTR technologies in its transcription workflow. Manually transcribing the diverse and complex letterforms and abbreviations found across the manuscript tradition proved both time-consuming and prone to human error.[57] In recent years, however, HTR technologies have been increasingly integrated into project pipelines, significantly improving both the speed and accuracy of transcription and offering new levels of precision and flexibility. With these tools now available, projects like the Canterbury Tales Project might have adopted a transcription strategy focused on the graphic level—that is, emphasizing the detailed representation of medieval letterforms and abbreviations, prioritizing detailed character representation over graphemic simplification.

The field of contemporary medieval studies is marked by a wave of experimental and innovative research leveraging HTR technologies to capture more detailed textual content of manuscripts. A number of scholars have recently designed their own transcription schemas and used HTR to carry out highly diplomatic transcriptions. Allow us to offer two illustrative examples and some thoughts from our own project. Haverals and Kestemont used HTR to produce a "hyper-diplomatic" transcription recording "the exact glyphs used on the page, including punctuation, abbreviations,

57 Robinson and Solopova, "Guidelines for Transcription."

and marginal notes" in their computational study of the Herne Charterhouse Scribal Community.[58] In a later study, they deliberately centred their analysis on scribal practices using many of the features of diplomatic transcription discussed above.[59] Although they restored word boundaries where words had been divided across line breaks with hyphens and did not distinguish between different heights of capital letters, they did retain characteristic features of medieval scribal writing, including superscript letters, macrons and other abbreviation marks, apostrophes, punctuation, and capitalization.

Schoen and Saretto have critically examined the differences between the fully diplomatic transcriptions required for computational analysis and the semi-diplomatic transcriptions more commonly produced by medievalists for human reading. The authors contend that technical constraints—such as the reliance on diplomatic transcriptions in OCR workflows—should prompt medievalists to reconsider how manuscripts are transformed across media. They argue that automatic transcription ought to be understood as one stage in the evolving remediation history of a manuscript.[60] We agree with this point. Furthermore, they underscore the challenges posed by palaeographic conventions, particularly the expectation to expand abbreviations, noting that "there is no prescribed convention for rendering abbreviations into Unicode characters." Automated transcription workflows, they argue, inevitably regularize the variability present in manuscript sources. It is neither practical, nor computationally feasible, to develop models capable of recognizing and reproducing the full range of rare or idiosyncratic abbreviation forms; some standardization in the transcription methodology is therefore unavoidable. Exactly where scholars will draw the line is an ongoing question of debate.

Standardization became a practical necessity as we developed transcription guidelines based for tracing scribal behaviour in our own corpus of Paris bibles. Following some observations in the Canterbury Tales Project, we chose to establish a single set of guidelines for the entire project, rather than manuscript-specific rules. As Bitner and Dase note, this approach "clearly benefits the economy of effort" by reducing confusion among transcribers, minimizing errors, and streamlining the training process.[61] Yet, as others

58 Haverals and Kestemont, "Silent Voices."

59 Haverals and Kestemont, "From Exemplar to Copy."

60 Schoen and Saretto, "Optical Character Recognition." We concur with the principle of this position in Guéville and Wrisley, "Everyone Leaves a Trace."

61 Bitner and Dase, "A Macron Signifying Nothing."

have emphasized, transcription is inherently interpretive, with objectivity in transcription remaining impossible to achieve. Our guidelines thus prioritize flexibility alongside standardization, recognizing that the complexity and inconsistency of scribal practices mean that exhaustive, rigid rules are counterproductive. Instead, we value the opportunity to work with a number of informed transcribers, working collaboratively, to make adaptive decisions in unique circumstances.[62] We view each automated transcription we make not as a definitive product (as one might create the definitive critical edition of a text) but as one stage in an ongoing process.

The idea of creating a universal, consistent transcription standard for medieval manuscripts, such as the CATMUS guidelines, has an understandable appeal—especially for cultural institutions dealing with large collections—but it is hard for us to imagine for precise genre-specific analyses.[63] Whereas we have other long-standing guidelines for encoding texts in digital humanities, such as the TEI XML guidelines for structure and content, standardized transcription for HTR does not enjoy the same level of community consensus. At the time we write, it is too early to know what global community adoption of such guidelines will be, particularly since universal standards pose a number of practical and epistemological questions. The most current version of CATMUS guidelines, for instance, vary from the practices we have adopted in the Paris Bible Project. In the cases of some allographs, the differences between our practices and the current CATMUS guidelines come down to a minor choice of different Unicode codepoints (for example, "ꝛ," Latin Small Letter Rum Rotunda, Unicode codepoint U+A75D versus ʀ, Latin Letter Small Capital Rum, Unicode codepoint U+A776). Other choices in CATMUS unfortunately choose to collapse fundamental layers of data reflecting scribal behaviour that we do not: i/j, u/v, normal d and insular ꝺ, normal r and rotund ꝛ, normal s and long ſ. How we represent abbreviations differs as well. All this scribal detail is useful to, even constitutive of, some research projects such as the Paris Bible Project.[64]

As work in the field of computational philology has suggested, rather than striving for a single, translinguistic standard that could merge all existing practices in medieval studies, it would be more productive to adopt a

62 Computational tools have been created to assess how uniformly different transcribers have applied any particular project guidelines. See, for example, Choco-Mufin, *PyPI: chocomufin*.

63 Pinche et al., "CATMuS-Medieval."

64 We illustrate the importance of special letterforms in this chapter in Figures 2 and 3.

modular approach.[65] Camps and colleagues have argued, for example, that instead of designing a unified pipeline that would subject every edition to the same stages (hyper-diplomatic, normalized, lemmatized with POS tagging, critical text), it might be better to focus on flexible pathways that better suit the wide range of goals. Approaches such as generic model creation—not designed to be used for direct application on digitized materials from the archives—are nonetheless useful starting points for fine-tuning project-specific models.[66] The same can be said of commons-based approaches such as HTR-United for the sharing and documentation of project data.[67] We discuss such infrastructures for data sharing in medieval studies in more depth in Chapter 4.

While the development of HTR "super models" are likely imminent for working with the Latin language (they have already been created for Dutch, German, Spanish, and other languages) we must acknowledge that the pursuit of ever more capable AI models and the needs of scholars in manuscript studies can sometimes be at odds with each other. The drive toward high-performing AI does not necessarily align with the nuanced, context-sensitive work based on fine-grained features of interest to computational philologists. After all, the way corpora are digitized, transcribed, and curated ultimately shapes the ways that models are created, the kinds of transcriptions that can be made and the interpretations that can be drawn. Unlike the vast datasets that drive innovation in commercial artificial intelligence, digitized manuscript archives are finite, fragmentary, and deeply contextual. In the study of manuscripts, in our opinion, quality, representativeness, and transparency outweigh sheer quantity. These issues challenge us to reconsider assumptions about training data size and standardization that are common outside the pre-modern humanities, and instead embrace a model of scholarly engagement that prioritizes interpretability, transparency, and the plurality of medieval textual cultures.

When HTR Models Encounter New and Different Data

As we demonstrated above, handwritten text recognition (HTR) models are powerful tools for transcribing medieval manuscripts, yet their use and development involve significant challenges. Critical issues that affect the quality of HTR-created text include the concepts of underfitting or

65 Camps et al., "Data Diversity."

66 Aguilar and Jolivet, "Handwritten Text Recognition," 12.

67 Chagué and Clérice, "HTR-United."

overfitting, by which we mean error which occurs in transcription when a model has not learned enough from the training data to perform accurately, or conversely becomes too narrowly specialized on the particularities of its training data. These issues are particularly problematic in manuscript studies, where high variability in handwriting, abbreviations, and orthographic conventions within and between manuscripts can make generalization difficult. For example, models trained on thirteenth-century French manuscripts written in a Latin gothic hand may fail when directly applied to other types of handwriting from other time periods, either because the texts are in another language, or because the script is different.

An interesting test case to illustrate these problems can be found in texts with distinct typographical norms, such as fifteenth-century incunabula which closely imitated the manuscripts they were based on and retained many medieval features, such as abbreviations, letterforms, and the characteristic two-column page layout. Editors of first editions and incunabula in the fifteenth century were often produced in versions that were much closer to the original manuscripts than modern editions. Striving to replicate the visual quality of manuscripts, these early printers preserved many other features as well, including running titles, rubrics, and medieval notational norms. Special letter types were created to include these abbreviations, such as the macron or others (ꝛ; ꝰ). These editions also retained distinctions like the normal and long "s" (s/ſ), normal and insular "d" (d/ꝺ), and normal and rotund "r" (r/ꝛ), distinctions often lost in modern manuscript editions. Abbreviations were not eliminated with the shift to print, but other characters took their place, reducing the palaeographic variance of glyphs.

In January 2023, we organized a "correct-a-thon" event with students from the Rare Book and Digital Humanities Master at the Université Marie-et-Louis-Pasteur (formerly Université de Franche-Comté) in Besançon, France, to learn about the possibilities and challenges of HTR in the context of digital humanities education. We will discuss this initiative more extensively in Chapter 4, but observations made by two of the participants are worth summarizing here.[68] Unlike their peers, they worked on a digitized Gutenberg bible, Beinecke, Zzi 56.[69] Faller and Rodriguez hypothesized that abbreviations in the Gutenberg bible would follow a consistent pattern and deployment scheme across the text, given that the entire Bible was

68 Faller and Rodriguez, "Paris Bible Correct-a-thon."

69 This incunable Beinecke, ZZi 56, is one of the forty-nine documented partial or complete copies of the Gutenberg bible that still exist today according to the Gutenberg Bible Census. See https://clausenbooks.com/gutenbergcensus.htm.

Table 3. The number of abbreviations and abbreviated words used in sample lines of text from the incunable Beinecke ZZi 56, compared with the number of words found in those lines. Table adapted from Faller and Rodriguez, "Paris Bible Correct-a-thon." Document in the public domain. Courtesy of the Beinecke Rare Book and Manuscript Library, Yale University.

Text in the Bible	No. of abbreviations	No. of words
mirabile quā ꝓduxerāt aque ī ſpecie	0	7
ſpecies ſuas. Factūq; ē ita. Et fecit de	8	6
fructū ⁊ habēs unūqdq; ſemen ſecūm	6	6
et terram. Terra autem erat īanis et	4	8

presumed to be printed using a fixed set of type characters. This fixed set of type characters would naturally reduce internal variance when compared to the variations introduced by hand, especially if multiple scribes contributed to the manuscript's creation.

Their hypothesis was found to be largely accurate; however, they also observed that the HTR model trained on manuscripts tended to repeat certain errors in the transcription when faced with specific type combinations. The model used to transcribe Beinecke, Zzi 56 was one trained on thirteenth-century handwritten Paris bibles, primarily LAD, MS 2013.051, and was therefore not entirely adapted to the transcription of incunabula, leading to some unexpected and recurring errors during the transcription process. They highlighted the fact that the abbreviations are not uniformly applied across samples of text and the use of abbreviations in Latin manuscripts and printed texts appears to be more a question of when they are applied rather than how they are applied (Table 3). Overall, there are numerous exceptions and cases where an abbreviation is inserted, complicating efforts to define a clear pattern or explain their recurrence.

Many questions arise with this example of the incunable bible: Do incunabula repeat the patterns of exemplar manuscript copies, or are the patterns determined by the typesetter?[70] Are Gutenberg bibles consistent in

70 Although the exemplar used to print the Gutenberg bible is currently unknown, studying potential patterns found other known incunabulum/manuscript pairs might offer insight into the way abbreviations were used. One such example might shed light on the fate of abbreviations and letterforms with the passage to print: Beinecke, MS 321 and Beinecke, Zi +4243. This pair contains the text of Poggio Bracciolini, *Historia Florentina*, translated into Italian by his son Jacopo di Poggio.

their abbreviation patterns? Does the way the Gutenberg bible is printed correspond to a specific manuscript or is it an idealized version thereof? In this context, Aguilar and Jolivet's focus on the "demand for ground-truth data aligned with specific needs" and "pre-trained models that can be adapted to unique requirements" are particularly salient.[71] Although the correct-a-thon was designed as a first engagement with issues with HTR, a more in-depth study of Latin incunabula would take the model trained on other bibles and retrain it for their specificities. This example underscores the importance of context-specific modelling and reinforces the need for critical reflection on the assumptions we make when transferring HTR tools across textual traditions.

Towards a Data-Centred Scribal Profile

What Was a Scribal Profile?

In this chapter, we have argued that digitized images of Latin biblical manuscripts offer an exceptional opportunity for modelling the use of specific glyphs and for developing custom HTR models that preserve graphetic features unique to individual scribes. While attention to abbreviations and letterforms has long preoccupied some editors of medieval texts, what is novel today is the ability to automate this work and to generate transcribed data at a scale that permits statistical analysis—making it possible to revise the idea of the scribal profile. Like the notion of diplomatic transcription discussed earlier in this chapter, the concept of scribal profiles remains somewhat ambiguous and, in our view, requires reconsideration, particularly in the context of large-scale copying traditions such as the Paris bible. In this final section, we revisit the idea of the scribal profile, reviewing features of the copyist's craft identified in previous scholarship, and propose a data-centred approach to scribal profiling as a complement to traditional palaeography. This method allows us to analyze transcriptions computationally, situating our work at the intersection of quantitative codicological methods and digital philology.

The study of scribal handwriting has traditionally concentrated on the palaeographical analysis of individual letters, emphasizing specific details, shapes, and graphetic variations that may assist both in distinguishing

This manuscript was used as the printer's copy for the first edition published by Jacobus Rubeus in Venice on March 8, 1476. For a comparison of manuscript and incunabula, see Meyers, "The Transition From Pen to Press."

71 Aguilar and Jolivet, "Handwritten Text Recognition," 1.

individual hands—that is, the handwriting of a particular copyist or group of copyists—and in characterizing broader script types associated with particular genres, regions, or periods.[72] Palaeographers commonly assess how letters are drawn, the presence or absence of identifiable palaeographic features (such as variations in the shape or size of letters, the use of ligatures, abbreviation systems, ascenders and descenders, punctuation marks, or diacritics), as well as spelling conventions and patterns of abbreviation.[73] These analyses may also extend to structural features such as spacing practices, page layout, or word separation. In addition to these specific elements, scholars sometimes refer to more subjective dimensions of a scribal hand's overall appearance or feel—an intuitive perception shaped by experience. From this perspective, a scribal profile may be defined as a composite of visual and measurable features consistently observed in a particular hand, with representative examples drawn from specific manuscripts and organized into typologies that facilitate comparison, attribution, and the identification of previously unclassified samples.

Advancements in the field of digital palaeography have opened the door to the study of historical scripts and the creation of digital resources for the identification, description and documentation of specific scribal hands.[74] DigiPal, the "Digital Resource and Database for Palaeography, Manuscript Studies and Diplomatic" is a research project developed at King's College, London that focused on texts produced in England during the eleventh century, with the aim to unite digital catalogues, descriptions of handwriting, and images of documents and their constituent letterforms.[75] The Late Medieval English Scribes project, on the other hand, focused on scribal hands, identified or unidentified, from manuscripts of five English authors: Geoffrey Chaucer, John Gower, John Trevisa, William Langland, and Thomas Hoccleve.[76] In this project, the scribal profile is made up of sample images for eight letters (a, d, g, h, r, s, w, and y) for each scribe identified and descriptions based on the specific shapes of the letters. They also add any feature that they deem relevant to identify a scribal hand, such as the use of

72 Derolez, *The Palaeography of Gothic Manuscript Books*.

73 For a discussion of seven aspects of a medieval hand forms, angle of writing, ductus, module, weight, writing support, internal characteristics, see Mallon, *Paléographie romaine*.

74 Ciula, "Digital Palaeography."

75 DigiPal.

76 Mooney, Horobin, and Stubbs, "Late Medieval English Scribes."

punctuation. While acknowledging the innovation of such projects, we can note their limitations to a specific geographic or authorial scope as well as their methodological choice.

Beyond the creation of traditional palaeographical catalogues, recent research increasingly employs computational methods to classify scripts for dating, localization, and scribal identification.[77] Projects such as ORIFLAMMS and its continuation, ECMEN, adopt a longitudinal perspective, combining letterform analysis with computer vision to study the evolution of medieval scripts across languages, regions, and time periods.[78] The ClaMM dataset (Classification of Medieval Handwritings in Latin Script), comprising over 8000 tagged images, has served as a key reference corpus.[79] Computational approaches have also been applied to features such as letter shapes,[80] script types,[81] width, and ink-tracing directionality to improve character recognition and hand identification.[82] As early as 2011, Stutzmann advocated treating texts as images to capture palaeographic variance beyond what transcription alone allows.[83] Studies have focused on cursive scripts, particularly chancery charters, where individual scribal traits are more evident. In these cases, features like the long s (ſ) or insular d (ð) aid in script classification.[84] However, such classification has largely depended on manual encoding, manageable for charters, but less so for large manuscript corpora, where the scale and complexity render manual methods impractical. Such research into large corpora, although it is difficult to scale, would lend itself to the integration of automated approaches—particularly those leveraging machine learning and pattern recognition—but the specificity of medieval handwriting demands ongoing scholarly oversight to ensure interpretive accuracy and contextual awareness.

---

77 Aussems and Brink, "Digital Palaeography"; Smit, "The Death of the Palaeographer?"

78 Stutzmann et al., "Les abréviations."

79 Cloppet et al., "ICFHR2016 Competition"; Cloppet et al., "ICDAR2017 Competition."

80 Ciula, "Digital Palaeography."

81 Hassner et al., "Digital Palaeography"; Kestemont et al., "Artificial Paleography."

82 Brink et al., "How Much Handwritten Text Is Needed," 1–4; Bulacu and Schomaker, "Text-Independent Writer Identification," 701–17; Aussems and Brink, "Digital Palaeography"; Stokes, "Computer-Aided Palaeography."

83 Stutzmann, "Nouvelles technologies," 217–23.

84 Stutzmann, "Paléographie statistique."

Scripts, Hands, Scribes

Current script classifications, including the widely-used system by Derolez—still regarded by some as the most precise for Gothic scripts—can fall short of effectively distinguishing between different script types and styles, particularly during transitional periods, or fully explaining historical developments.[85] Dating or situating manuscripts by script alone can be problematic since "several script families and script types are used contemporaneously, so that two samples of handwriting from the same date need not be similar or belong to the same category."[86] Davis offers a useful alternate definition of a hand as a compromise between

> an internalized model hand, acquired by practice, imitation, and, to a small extent creativity ... and the exigencies of the pen used, the writing material on which the text is to be written, the writing medium (ink, for instance), the writing surface (a desk, perhaps), on occasion the writing environment ..., the architecture of the hand, and the neurophysiological characteristics of the writer. [87]

A scribal profile cannot, it seems, be reduced to essential qualities, but needs to take into consideration a combination of factors specific to book culture: learned unconscious habits, other conscious ones along with a number of material aspects of the copying process with which a medieval scribe had to compromise: the specific characteristics of parchment, the variable size of the justified block on the folio and anticipating the spaces in which rubrication and illumination would take place, and so forth. Fully grasping what these factors were at the time of the copying of the manuscript is perhaps an impossible task. On the other hand, the study of versions of handwritten text from manuscripts does offer us a window into a deeper understanding.

Some scholars have argued that abbreviations can be used for identifying change of scribe in a text or regional characteristics and to identify scribal profiles and individual hands.[88] For instance, a half century ago McIntosh characterized a scribal profile as a "suitably organized inventory of a selection of a scribe's usages drawn up from the observation of the treatment of a number of items in a single piece of text written in one hand." He theorized a "unique linguistic profile" (LP) encompassing spellings and grammatical forms alongside a "graphetic profile" (GP), that focuses

85 Derolez, *The Palaeography of Gothic Manuscript Books*.

86 Stutzmann, "Clustering of Medieval Scripts."

87 Davis, "The Practice of Handwriting Identification."

88 Kestemont, "A Computational Analysis of the Scribal Profiles."

on "linguistically sub-systemic … phenomena."[89] While traditional palaeographical clues are included in the graphetic profile, features like abbreviations and unique letterforms are incorporated within the linguistic profile. Interestingly, he argued that only the linguistic profile can reveal the "likelihood that two texts in different modes [i.e., written with what seems to be different hands] are in reality both the work of one man," since a scribe might adapt their writing style.

Moreover, medieval scribes are generally believed to have been able to handle multiple scribal styles and scholars have illustrated that a scribal hand could evolve over time on account of age or illness, for example.[90] That is to say that focusing only on the palaeographic features has the potential of being misleading for scribal identification: a traditional palaeographical analysis of the hands is not enough for identifying scribal nuances. In fact, McIntosh highlighted the correlation between linguistic profiles and geographical position, arguing that the analysis of linguistic profiles might allow scholars to differentiate scribes within the same *scriptorium*. It has been claimed that

> [t]he ability to model scribal characteristics using purely linguistic means is … highly promising, especially for scribal studies that span different script registers. Mere paleographic inspection can be problematic in such multimodal cases, where profound differences in the shapes of glyphs and characters have been attested. Across different national traditions in philology, scholars have argued that a purely linguistic approach is a relevant complement to traditional paleography in this context. A scribe's idiosyncratic use of certain spellings, for instance, is not very likely to change if the scribe switches to another paleographic style.[91]

Modelling Scribal Behaviour

The research we have carried out in the Paris Bible Project highlights the significant variation found in manuscripts, including differences in word order, interpolations, orthography, and abbreviations, has helped us to articulate a data-centred, transcription-based approach that goes beyond the concepts of scribal profile, scribal syntax, or the linguistic profile and graphetic profile of McIntosh mentioned above. We do this through a

89 McIntosh, "Scribal Profiles from Middle English Texts."

90 Beach, *Women as Scribes*.

91 Haverals and Kestemont, "Silent Voices," 191, drawing on the work of Stokes, "Scribal Attribution" and Aussems, "Christine de Pizan."

Table 4. Three glyphs found in Beinecke, MS 387, their probable corresponding characters and sample Unicode codepoints we use to transcribe them. Manuscript in the public domain. Data by authors. Courtesy of the Beinecke Rare Book and Manuscript Library, Yale University.

Glyph found in manuscript	Corresponding characters	Unicode codepoint
	Often standing for the letters m or n	0304
	Usually replacing i, ibi, ihi, ri	0365
	Tironian et	204A

process of *computational modellin*g. Automatic transcription of many different Paris bibles from different times and locations with a custom transcription scheme is an example of such modelling. Modelling humanities data it has been argued is made up of "explicit explanatory, exploratory and empirical strategies of inquiry" that allow one both to create data from objects of humanistic study—in our case, hand-made and hand-copied documents—and to interact with these objects iteratively and creatively to explore specific research questions.[92]

As discussed above, our primary method for modelling scribal behaviour is HTR, which enables analysis at the level of individual characters. Central to this approach are the choices made in designing the transcription schema: whereas normalization tends to efface the distinctive traces of individual scribes, non-normalized transcription preserves them. By customizing a character set attuned to features typical of medieval scribal behaviour, we aim to capture both linguistic and graphetic dimensions, following McIntosh's distinction. Distinguishing among variant forms of letters such as s, r, and d thus retains palaeographic information that might otherwise be lost. HTR systems do not support infinitely extensible character sets, but they do allow for targeted inclusion of the most frequent glyphs, and in our Paris Bible corpus the number of variant letters is both relatively limited and well suited to these technical parameters. While Unicode modelling inevitably flattens many of the finer distinctions prized in palaeographic analysis, it also enables macro-level perspectives, revealing scribal patterns that would be burdensome to track manually. The wager of the Paris Bible Project, in

92 Ciula et al., "Models and Modelling."

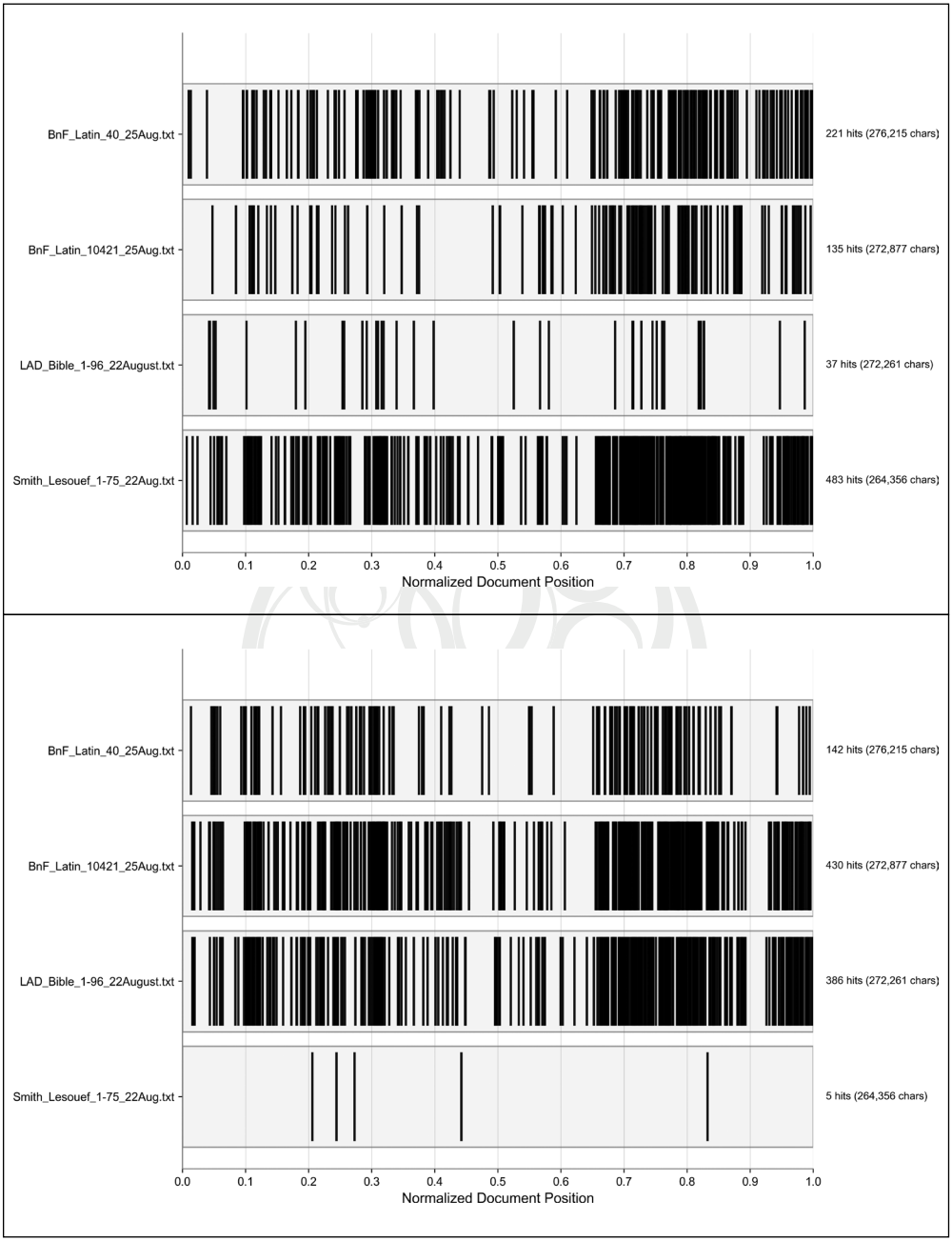


Figure 2. Density plots illustrating the frequency of use of *ɔn** (visualized in the top four lines) opposed to *ɔomin** (visualized in the bottom four lines) in BnF, MS latin 40; BnF, MS latin 10421; LAD, MS 2013.051; and BnF, MS Smith-Lesouëf 19, respectively. Data by authors. Visualization created in MatPlotLib and Python by authors.

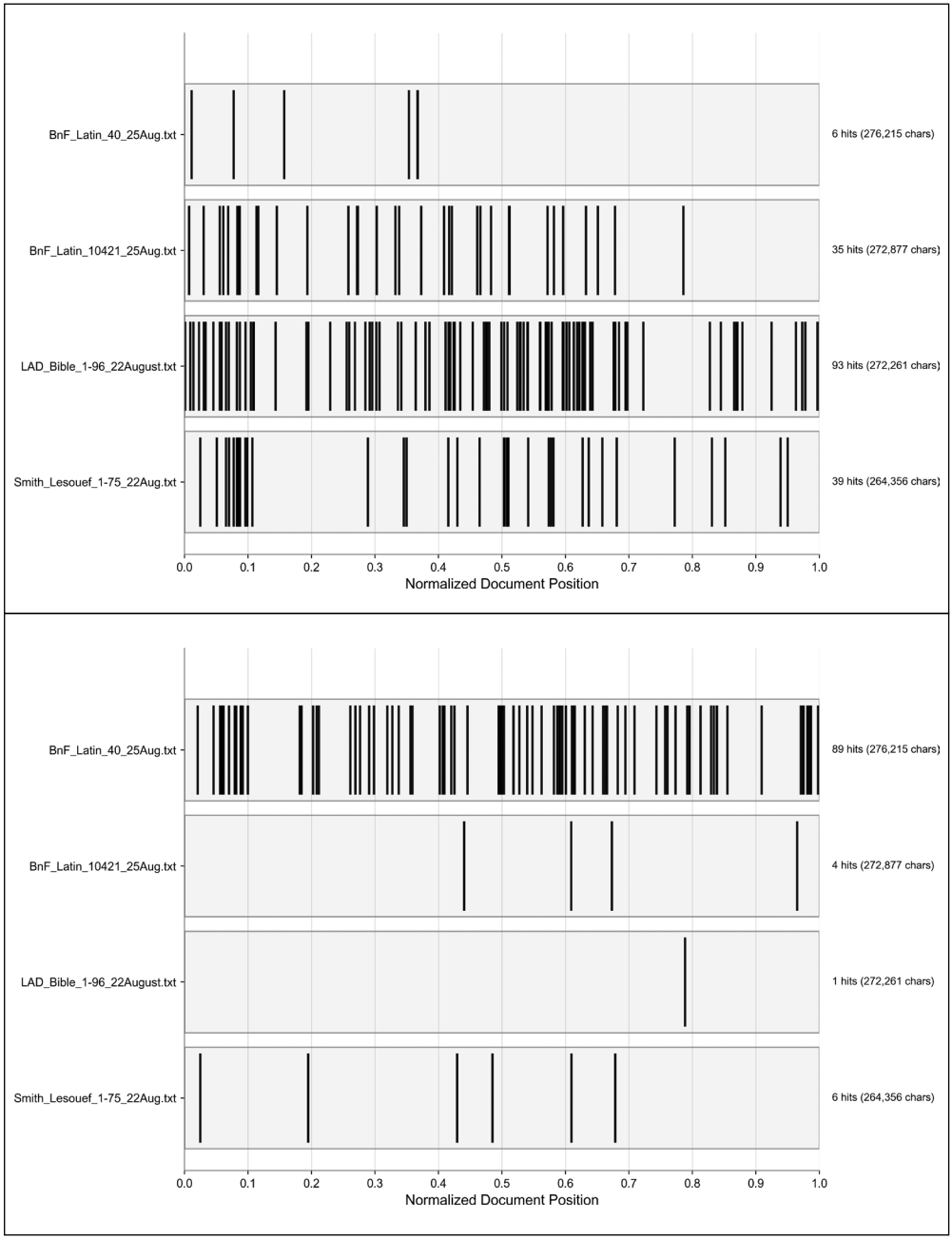


Figure 3. Density plots illustrating the frequency of use of *pɔɛ** (visualized in the top four lines) opposed to *pre** (visualized in the bottom four lines) in BnF, MSS latin 40 and latin 10421; LAD, MS 2013.051; and BnF, MS Smith-Lesouëf 19, respectively. Data by authors. Visualization created in MatPlotLib and Python by authors.

short, is that by modelling traces of scribal practice through automation, we can detect patterns in transcription that illuminate the habits of the artisans who fashioned these books.

Two simple examples demonstrate how careful attention to modelling scribal practices can produce scaled results across a manuscript transcription. For instance, as visualized in Figure 2, the scribe of Paris, BnF, MS Smith-Lesouëf 19 consistently uses the abbreviated form of *domino* (*ꝺn̄**) with only a few instances of the extended form (*ꝺomin̄**), both of which begin with the insular d which we have encoded as Latin small letter insular D (U+A77A). The use of the asterisk here is known as a wild card, allowing for all inflected forms of the Latin word to be captured. In contrast, the scribes of LAD, MS 2013.051 and BnF, MS Latin 10421 favour the extended version. Meanwhile, BnF, MS Latin 40 contains both forms in nearly equal numbers, though their distribution varies across the text.

Another common form of scribal paleographic variance in the Latin bibles we have looked at can be demonstrated by the alternation of the letter r and the rotund r encoded here as Latin small letter R rotunda (U+A75B). Figure 3 illustrates the use of the string often corresponding to the prefix *pre-* with and without the rotund r variant. Whereas no manuscript uses one of the forms exclusively, there are clear patterns; BnF, MS latin 40 prefers the form *ꝑꝛe** and the other manuscripts prefer the form *pre**.

Many questions come to mind when we see features of scribal behaviour visualized in such a way. Do such patterns help us to understand where the manuscripts were produced or by whom? Does the concentration of a given feature map onto a hand in the manuscript? Do the patterns help us with the dating or localization of the manuscripts? From a colophon, we know that BnF, MS Smith-Lesouëf 19 was copied by a scribe naming himself Arnulphus de Camphaing who was active around 1260 in Cambrai or Tournai.[93] LAD, MS 2013.051 is a two-volume bible produced by unnamed scribes and is thought to have been produced in the Rouennais around 1250.[94] The part of BnF, MS latin 10421 consisting of a Paris bible is dated and localized with less

93 The same scribe, Arnulphus de Camphaing, is associated with a manuscript of Prosper of Aquitaine, BL Add. MS 78830. Although we have not yet compared the features of these two manuscripts, the linkage provided by the named scribe does warrant further investigation to address the question of whether two different texts copied by the same scribe would contain similar linguistic and graphetic features. The main challenge of doing so is that, at the time we are writing, we have not found a digitized version of the London manuscript.

94 Guéville, "Les manuscrits médiévaux occidentaux."

precision: thirteenth-century Paris. Finally, BnF, MS latin 40 has a named scribe, Johensis, and has been dated to the third quarter of the thirteenth century and localized to Naples, Italy.

Given the level of detail available for certain manuscripts—such as the identification of individual scribes, the presence of multiple hands, and relatively precise information about the date or place of production—it is tempting to formulate hypotheses about correlations between such metadata and specific linguistic features, such as the frequency of *ꝺn** versus *ꝺomin** or of *p2e** and *pre**. For instance, might the preference for *ꝺomin** in LAD, MS 2013.051 and BnF, MS latin 10421—especially when the former is more precisely dated and localized—offer a way to refine the uncertain dating of the latter? Similarly, do bibles in the Paris style copied in Naples, particularly those produced for the court of Manfred, exhibit a distinctive pattern or signature style in alternating these two abbreviation forms? While such interpretive moves might be appealing in a world in which metadata about manuscripts is highly uncertain, drawing broad conclusions from limited—even if repeated—transcriptional features carries significant risk, as manuscripts are complex artifacts possessing multiple layers of meaning and production. Moreover, traditional methods of dating and localization have often relied heavily on decorative or iconographic evidence—such as illumination and ornamentation—instead of on linguistic or scribal data, complicating any attempt to correlate textual features with provenance.

Conclusion

In this chapter, we have examined recent advances in HTR and their implications for the transcription of medieval manuscripts, positioning transcription as a form of big data within medieval studies. We have explored a spectrum of transcription practices from earlier analogue and digital humanities editorial projects, from normalized to non-normalized approaches, and have reflected on how the concept of ground truth—borrowed from machine learning—might be problematic for medieval textuality. We also considered how HTR systems perform when they are provided with manuscripts they have never encountered before, and we proposed that character-level modelling of scribes may offer a scalable and data-centred alternative to more traditional, interpretive notions of scribal behaviour. Non-normalized transcriptions allow for a productive blending of material philology and computational humanities which seems to be growing in popularity among medievalists today, both preserving an interest in one of the key elements of the

handmade manuscript—scribal variance—while at the same time leveraging distant forms of analysis.

The integration of HTR into the study of scribal behaviour in medieval manuscripts exposes a number of tensions between the manual and the automated—a tension that encapsulates many of the methodological and ethical stakes of computational approaches in the humanities. While the practice of identifying scribal profiles, particularly in Paris bibles, builds on earlier work by quantitative codicologists,[95] what is novel in the current moment is the ability to scale these analyses across larger corpora through automation. This scalability opens up exciting possibilities for engaging with complex textual data using methods shared with computational social science, machine learning, and data science. However, the pursuit of better performing transcription models under the sign of automation and innovation in AI runs the risk of obscuring what is learned through the slow, iterative, and often meticulous labour of human transcription. The foundational act of creating ground truth—developing transcriptions that will serve as training data for HTR—is deeply interpretive and dependent on scholarly expertise, especially in highly specialized applications such as medieval manuscripts. The same could be said of the correction of outputs and model retraining.

If ground truth, created through careful philological interpretation, remains central to the development of bespoke models, we do not fully agree, however, that the entire trace of interpretative work involved in model training completely disappears. Scholars in digital art history have called attention to what they call the "epistemic entanglement" between AI models and the cultural materials they analyze—arguing that models not only process their data, but also reflect the assumptions, values, and priorities embedded in their training.[96] That same entanglement exists in HTR-enabled computational manuscript studies and it merits further discussion. The features that the creators of ground truth choose to include in our training data—that is, those we deem worthy of modelling—shape the possibilities of future analysis and interpretation. This is one reason why, along with Aguilar and Jolivet, we favour the creation of smaller, bespoke models, tailored to specific research questions and material traditions, rather than large-scale generalization for medieval studies. We also suspect that moving forward medievalists will need to devise new ways of exploring data provenance in models, assessing the degree to which the models they reuse bear

95 Ruzzier, *Entre université et ordres mendiants*.

96 Impett and Offert, "There is a Digital Art History."

the imprint of the decisions of others who created them as well as acknowledging the scholarly labour that they embody.[97]

Ground truth, at the moment we are writing, is not directly reusable across genres or scripts; rather, it requires careful fine-tuning, to use a term from machine learning, by which is meant reshaping in light of project choices and the specificities of the archive at hand. The scribal modelling process, then, is not a fixed method, but a general research framework—a flexible container of the visual and linguistic particularities of different manuscript traditions. The customization that is required in AI-based transcription practices foregrounds the role of scholarly oversight needed at every stage of computational humanities research. Editorial authority is not surpassed by computational rigour, but is retained in the selection, inclusion, and exclusion of machine-generated outputs, underscoring the fundamentally interpretive dimension of our work.

Character-level modelling enabled by HTR technologies offers a new lens for capturing scribal behaviour, yet it also illustrates the ongoing need for careful scholarly intervention when moving between manuscripts or between research projects. Ground truth data must be crafted with attention to the specific features and contexts of the manuscripts under study. Modelling at this level is not a standardized process. Different researchers may choose to encode different sets of features, opt for alternate Unicode codepoints for certain characters, or prioritize different kinds of variation, depending on project-specific research questions. This flexibility of the approach is both a strength and a liability. On the one hand, it allows for humanities research questions to drive automation. On the other hand, it resists the impulse to create broadly applicable, general solutions. The authority of the editorial scholar is not displaced by the model, but repositioned: they must now evaluate the outputs of machine learning systems, making informed decisions about what should be retained, revised, or discarded.

Finally, such work has broader implications about the future of manuscript studies and the training of scholars within it. Do computational techniques forecast the "death of the palaeographer?"[98] Most likely not. Do computational techniques suggest the end of manual transcription? In part, yes. Traditional competencies—such as palaeography, codicological description, and manuscript handling—will remain essential, but must now be complemented by new proficiencies in data curation, model evaluation, and

97 Romein et al., "Exploring Data Provenance."

98 Smit, "The Death of the Palaeographer?"

computational thinking.[99] Medievalists must learn to see transcription in the age of AI as a critical activity and as part of a longer research process of computational philology: how to design ground truth datasets, to assess the interpretive consequences of encoding choices, to understand the assumptions embedded in machine-generated outputs and to adapt models from other contexts to one's own. Indeed, as AI-based technologies reshape the way we access and analyze medieval manuscripts, they require a recalibration of the skills expected of established scholars and aspiring medievalists alike.

A reconfiguration of skills also carries epistemological implications: they position the scholar not merely as a transcriber or interpreter, but as an active participant in shaping research techniques for how computational models encounter and analyze historical texts. In this sense, something as foundational as ground truth creation is not only interpretive, but borders on the infrastructural, influencing how future research will be able to operate across corpora and platforms. Academic training thus needs to emphasize methodological reflexivity and ethical awareness, training scholars to navigate the balance between automation and interpretation. Future pedagogy in manuscript studies must cultivate a capacity to ask critical questions about tool use, model design, and the nature of textual evidence in a computational age. Computational manuscript studies need to be infused, in other words, with the lessons of critical AI.[100] In doing so, it will prepare scholars not only to work with HTR—after all, we do not know that HTR will exist in its current form in a decade—but to shape future methods in ways that remain accountable to the humanistic values at the heart of the discipline.

99 Berry and Fagerjord, "On the Way to Computational Thinking."

100 Impett, "Digital Art History as Critical AI."

Chapter 3

ASSESSING MANUSCRIPT CO-CREATION USING COMPUTATIONAL METHODS

THE CORPUS OF available Paris bibles is rich in possibilities for research into scribal practices, not only on account of the number of extant, digitized copies, but also for existing evidence and scholarly literature about their fabrication. Studies of the *pecia* system have brought to light the social and economic system by which the need for books, especially in medieval university towns such as Paris or Bologna, was met by the controlled circulation of *pecia* exemplars, that is, parts of manuscripts loaned out to a distributed group of actors who cooperated in its copying.[1] Along with the professionalization of the copyist, scholars have emphasized the resources that emerged from the religious orders with this scaled up process of textual criticism: the *correctoria* (a list of critical notes on the biblical text), the *distinctiones* (verbal concordances) and the like.[2] Despite this substantial body of criticism about the production of bibles in the context of the medieval university, little is known about the copyists themselves and the ways in which they worked.

This chapter looks at claims that are made about manuscript objects such as Paris bibles and their scribes in the scholarly and reference literature, and considers possible ways that we might go about assessing such claims through a combination of computational methods. We explore a few of the ways that medieval manuscripts—once they have been digitized and then transcribed with the methods described in Chapter 2—can be studied as data. We apply such an approach to a focused set of twelve Paris Bibles, using computational modelling to examine features such as abbreviation practices, graphetic variation, and distinctive letterforms. While studies suggest promising avenues for analysis, considerable work remains to be done across different languages, genres, and manuscript traditions in order to assess both the broader applicability and the methodological limitations of the approach. That is to say, we do not assume that such modelling will be effective for all manuscript corpora. What is clear, however, is that the utility of this method is contingent upon the presence of certain

1 Rouse and Rouse, *Manuscripts and Their Makers*.

2 Dahan, "Paris Bibles and Scholarship."

DOI: 10.17302/mfsk15868

scribal features—most likely a degree of abbreviation, variation in spelling, or the use of distinctive allographetic forms—and at the least several dozens of folios of transcription from a variety of manuscripts. These are conditions that, we imagine, apply to many textual situations in medieval studies. The precise threshold of such variation required for meaningful analysis remains uncertain. Yet, given the growing adoption of HTR in the production of machine-readable transcriptions, we anticipate that further applications of this method will shed light on these questions, refining our understanding of the conditions under which computational modelling of scribal practices is most productive. As we have explained in the previous chapter, we use a non-normalizing transcription schema within state-of-the-art HTR technology that not only allows much more text to be transcribed than ever before by a human hand, but also allows certain graphetic features of medieval writing found in manuscript to be recorded on demand, with significant (but not perfect) accuracy.

As we also discussed in Chapter 2, contemporary medieval studies have adopted different approaches to the graphetic features found in manuscript, with editorial practices largely favouring their normalization, but with some computer philological approaches favouring non-normalization and at other times attempting a combination of both. We believe that there is value in employing some well-established methods in computational textual creation and analysis, to assess claims that have been made about codex creation, either by scribes who indicate their participation through including a colophon, or by contemporary book historians and expert cataloguers who indicate the number of hands they can identify in a manuscript. This assessment might be to formulate new hypotheses about how scribes operated when copying manuscripts, but it could equally help to confirm existing hypotheses with new forms of evidence.[3]

Automatic transcription of Paris bibles using a pre-defined scheme and the analysis of these transcriptions are an example of *computational modelling*: that is, these two methods are "explicit explanatory, exploratory and empirical strategies of inquiry" that allow us both to create data from objects of humanistic study—in our case, hand-made and hand-copied documents—and to interact with these objects iteratively and creatively to explore specific research questions in mind.[4] As we have explained previously, the most important features in our data that we capture are scribal abbreviations

3 Eve, *The Digital Humanities and Literary Studies*, 131.

4 Flanders and Jannidis, *The Shape of Data*; Ciula et al., "Models and Modelling."

Table 5. List of the computational experiments in Chapter 3, organized by their appearance in the chapter. Data by authors.

No.	Shelfmark	Dating	Size (mm)	Digitization
1	Cambridge, Corpus Christi College (hereafter CCC), MS 49	thirteenth	345 × 225	Parker on the Web
2	Philadelphia, University of Pennsylvania (hereafter UPenn), MS Codex 236	ca. 1220	218 × 148	OPenn
	Paris, Bibliothèque Mazarine (hereafter Mazarine) MS 6	thirteenth	392 × 265	BNIF
3	BnF, MS latin 40	third quarter, thirteenth	255 × 180	Gallica
	BnF, MS latin 10428	third quarter, thirteenth	240 × 165	Gallica
	Città del Vaticano, Biblioteca Apostolica Vaticana (hereafter BAV), MS Vat. lat. 36	1250–1258	269 × 182	DigiVatLib
4	Girona, Arxiu Capitular de la Catedral, MS 52	fourteenth	430 × ?	HMML
	Lisboa, Biblioteca nacional de Portugal (hereafter BnP), MS IL 93	1251–1300	324 × 217	BnP
	BnF, MS latin 179	end of thirteenth to fourteenth	235 × 150	authors
	BnF, MS latin 211	end of thirteenth to fourteenth	165 × 110	authors
	BnF, MS latin 15477	1251–1275	286 × 196	authors
	Philadelphia, Free Library of Philadelphia (hereafter Free Library), MS Lewis E242, a.k.a the "Patou bible"	ca. 1250	175 × 120	Internet Archive
	Sarnen, Kollegiumsbibliothek, Stiftsarchiv Muri-Gries (hereafter Sarnen KB), MS Cod. membr. 16.	before 1267	250 × 170	HMML

and letterforms, and in this chapter we explain how we use these features to nuance what we believe to be true about Paris bible manuscripts as well as to open new evidence-based pathways for exploration of them.

Table 5 lists the dozen Paris bibles we have used in this chapter to create both partial and full automated transcriptions with HTR. These manuscripts are not the most famous exemplars of the Paris bible, nor do they originate from specific decades of the thirteenth or fourteenth centuries. Instead, we have chosen them, for three main reasons. First and foremost, for questions

Figure 4. Four images from a single manuscript, Mazarine, MS 6, visualized in several panes of Mirador, one of the viewers for the International Image Interoperability Framework (IIIF), comparing suspected shifts in scribe: from the OT (at left, fol. 5v), the end of the OT and colophon (centre top, fol. 429r), the beginning of the book of Matthew (upper right, fol. 430r) and the Interpretation of Hebrew Names (lower right, fol. 523v). Images courtesy of the Institut de France.

of access. The availability of digitized images of the manuscripts—either from official online repositories or via self-digitization—is a precondition for computational transcription and analysis. This criterion of selection, of course, limits us in terms of the issues that we might want to investigate. If we do not have access to specific manuscripts, if we cannot travel to see them or if the library does not allow us to take images for private research purposes, computational analysis is effectively off limits. Second, our choice of manuscripts has been made on the basis of existing material evidence documented by scholars about their fabrication, such as the number of hands, the presence of a colophon, and the attribution to a group of manuscript makers. Third, where we saw that other cognate manuscripts would be interesting for comparison, and we were able to visit the manuscript room, we digitized some limited samples of the manuscript ourselves on the fly using a regular smartphone camera.

Computational Experiments Based on Single Manuscripts or Groups of Manuscripts

Against the backdrop of the mass copying of the Paris bibles, what we know about the scribes themselves and how they worked is not much. In isolated cases we do have some names of scribes or dates associated with codices.[5] A thirteenth-century Latin bible, Mazarine, MS 6, contains a colophon at the end of the OT (fol. 429r): "Explicit Vetus Testamentum per Johannem de Cristemanneford scriptum, cui Deus reddat premium" ("Here ends the Old Testament, copied by John of Cristemanneford, whom God will reward.") Such colophons tend to be recorded as part of the metadata of manuscripts in collections, along with incipits or explicits, as an identifying or descriptive marker of the document.

Much has been made of the ability to carry out "dynamic" or synoptic readings of manuscripts, by which is understood the ability to deliver, display, and compare digital images of these manuscripts in display formats such as the International Image Interoperability Framework (IIIF) for human reading and examination.[6] A dynamic view in IIIF would certainly facilitate traditional palaeographic analysis of the manuscript (instead of seeing the manuscript in person or on microfilm) and a scholar might notice, for example, the ways in which the hand seems to change at the beginning of the book of Matthew (see Figure 4). Quantitative codicologists with access to digitized manuscripts might go a step further to compare manually the kinds of abbreviations used by different hands on a few pages of the New Testament (hereafter NT) with those found a few pages back. In the case of a manuscript of 549 folios, such as those making up Mazarine, MS 6, much more can be done than laying out multiple views of the document for a researcher to investigate visually. New computational methods are made possible by full digitization and access to high quality scans available in open access.

It is worth pausing for a moment to wonder what kinds of questions might be investigated in the case where we have material evidence such as a colophon. In the case of Mazarine, MS 6 with the colophon at the end of the OT, we might notice that the hand appears to change in the NT, and our investigation could end there. Alternatively, since the hand of the same scribe might also vary somewhat according to different factors such as fatigue, reaching the end of a quire, or changing from the hair to the flesh side of the parchment, we could pursue a deeper investigation looking at

5 Ruzzier, *Entre université et ordres mendiants*, 283.

6 Nichols, "Dynamic Reading of Medieval Manuscripts."

the hand over many folios, or at extended linguistic features in the hand to examine if there is a change in the way that the text is copied.[7] If we confirm that the scribe, or scribes, of Mazarine, MS 6 who took on the NT and the Interpretation of the Hebrew Names are in fact different, we could go on to establish what orthographic and abbreviation patterns are typical. We illustrated such analysis in Figures 2 and 3 in Chapter 2. Such investigation, although it involves tiring attention to a significant amount of detail, can nonetheless pave the way for other avenues of research. Did John of Cristemanneford copy the whole section on which he signed his name? If not, why might he have put the colophon on fol. 429v? Was he copying the text faithfully—abbreviations, letterforms and all—from another exemplar? Or does he alter the wording here and there? In this manuscript, is this the case of a single scribe writing in different hands? Or of multiple scribes converging on similar versions of a similar hand? We enumerate this list of rhetorical questions not to suggest that they are all appropriate for Mazarine, MS 6, but instead to suggest the kinds of analysis one might do based on manuscript transcriptions. The need for computational tools and machine learning arises in these scenarios when capturing such detail falls beyond the threshold of human attention to detail, that is, when there are simply too many details to record and to analyze.

Drawing on Chapter 2, in which we argue for the importance both of creating transcriptions from the complex and deeply layered manuscript traditions and interpreting this material in dialogue with humanistic knowledge, this chapter approaches the dozen or so Paris bibles listed in Table 5 through the lens of the *computational experiment*. By this we mean the process of modelling historical copying practices in manuscript, assembling evidence from digitized manuscripts, and assessing them using computational methods. We saw one such experiment in Chapter 1, with the example of prologues collected from manuscripts of differing geographical production. We have mentioned how Paris bibles provide an excellent corpus given its size and geographic scope. It is important to state, however, that our intention here is not to apply a one-size-fits-all approach to the entire Paris bible corpus. Instead, we home in on an accessible selection of them, in order to explore to what extent computational methods can confirm, revise, or nuance hypotheses we have (or others have had) about manuscripts from the thirteenth or fourteenth century. We address questions about the trustworthiness of colophons and how manuscript

7 Guéville and Wrisley, "Transcribing Medieval Manuscripts."

historians draw connections between manuscripts that they believe have been created by the same hands or in the same groups. In what follows, we offer a series of case studies that demonstrate both the possibilities and potential shortcomings of computational analysis based on HTR-created transcription. The interpretation of such models is not automatic, nor can it be far from humanistic knowledge, but it instead pays careful attention to what the transcription is designed to do, the assumptions that it makes and the limitations of method.

We feature four different scenarios identified from the Paris bible corpus in which we believe computational analysis of single manuscripts or groups of manuscripts might shed new light on our understanding of these book objects. We proceed in order of increasing complexity of these scenarios—not because the manuscripts themselves are materially more complex—but rather because the assertions that critics have made about them depend on different forms of evidence of varying complexity, ranging from small material details in individual manuscripts to larger claims about groups of copyists working together. In the first experiment, we analyze CCC, MS 49 in order to map changes of hands within a single manuscript and assess if methods such as computational stylistics (also called stylometry) are able to confirm what is otherwise visible to the human eye. The second experiment focuses on the two manuscripts UPenn, MS Codex 236 and Mazarine, MS 6 to assess if claims made in a colophon or in a catalogue description are computationally detectable and if we can perhaps confirm, nuance or disprove them. The third experiment analyzes three manuscripts claimed to have been produced (that is illustrated) by the Master of the Bible of Manfred. According to the colophons in two of the manuscripts, BnF, MS latin 40 and Città del Vaticano, Biblioteca Apostolica Vaticana (hereafter BAV), MS Vat. lat. 36 were copied by the scribe, Johensis, while the scribe of the third manuscript, BnF, MS latin 10428, is unknown. We explore the possible attribution of the scribe to the same scribe as the previous two, Johensis. Our fourth and last experiment focuses on attribution claims to a group first identified by Branner, the so-called Atelier of Johannes Grusch, looking at a number of manuscripts attributed to that circle in order to understand better if only the illuminators, or perhaps also the scribes, worked together.[8]

8 Branner, "The Johannes Grusch Atelier."

Stylometry with Medieval Texts: Challenges and Opportunities

Before we turn to our four case studies, it would be salutary to mention one approach in computational analysis commonly applied to medieval texts, stylometry. Stylometry—the quantitative analysis of countable features in a text—is, in fact, an umbrella term for a variety of related methods for the statistical analysis of textual features that has become important in recent decades.[9] Here is not the place to elaborate on all of the details of such methods, but it is helpful to highlight how they have been applied, particularly given the complexities of medieval textuality.

Many medieval works are unattributed to an author, a scribe, or a group of writers, and questions of authenticity and attribution loom over the scholarly literature. We know that the input of scribes in different textual traditions has been fundamental to, and sometimes constitutive of, the traditions themselves. Classic approaches to the puzzles of authorship in medieval texts have used vocabulary, grammar, rhyming patterns or material clues in manuscripts. With the rise of contemporary computational approaches, it has not become simpler to solve mysteries in book or literary history, but such approaches have certainly shifted focus with respect to the kinds of evidence, the means by which such questions can be asked, and the ways that new hypotheses can be proposed. Robust case studies exist in the critical literature of stylometry applied to medieval texts addressing questions of author, scribe, genre, dialect, and dating.[10] So far, in our assessment, there is no one-size-fits-all solution for how to handle the specifics of different textual scenarios, but we feel strongly that method of transcription has an intimate relationship with the kinds of research questions we ask, and like Bode, we feel that critical corpus construction in tandem with knowledge of the materiality of the sources is essential for critical computational analysis.[11]

Although there have been many studies in the stylometry of medieval texts, three particular examples might serve as critical reference points for our discussion here. Edlich-Much and Edlich-Much used function words and lemmatized tokens from digitized normalized editions of Malory to study the influence of Old French and Middle English sources on the Middle English *Morte Darthur*, seeking to resolve questions about source adaptation

9 The elegant, yet simple definition of style as a set of countable features is provided by Herrmann et al., "Revisiting Style."

10 De Gussem, "Computational Stylistics."

11 Guéville and Wrisley, "Transcribing Medieval Manuscripts."

and editorial originality.[12] They contribute to a body of Malory scholarship about authorship, revision, and textual unity, with their approach foregrounding the intersection of adaptation and stylistic agency. On the other hand, Camps and colleagues created a diplomatic transcription of a manuscript compilation of unattributed saints' lives, BnF, français 412, engaging with the hypothesis of Paul Meyer about hagiographic "series," or distinct groups, perhaps reflective of common authorship or source traditions.[13] More recently, Vandyck and Kestemont created a hyper-diplomatic transcription preserving brevigraphs and abbreviations using the HTR platform Transkribus (similar to the method used in the Paris Bible Project) and carried out Term Frequency-Inverse Document Frequency (TF-IDF) weighted analysis using character bigrams to establish the chronological order of manuscripts created by the Speculum scribe at the well-documented scriptorium, the Herne Charterhouse.[14]

Of these three approaches, the one that bears the least resemblance to ours is that of Edlich-Muth and Edlich-Muth, not because of the vernacular context of the source material, but on account of its reuse of critical editions and focus on content analysis as a means of accessing influence. Similar to Camps and colleagues, we also adopted a multi-step pipeline using HTR, but we eschew lemmatization and normalization. Their research workflows produce imperfect transcriptions like ours do, but we do not depend on questions of authorship or editorial groupings per se for our source material. Our method is the closest to that adopted by Vandyck and Kestemont, although we are working with a corpus which is simultaneously much larger and far less documented from the perspective of scribal attribution. Our method emphasizes, with Vandyck and Kestemont, three main aspects: scribal profiling over content analysis, a reliance on sufficient HTR quality, and character-level modelling.

Viewed together—and alongside other studies not cited here—these three contributions illustrate a spectrum of stylometric approaches, each attuned to the particular historical and material conditions of medieval textuality. They demonstrate how stylometry can enrich the study of unattributed corpora by detecting patterns suggestive of shared authorship or scribal networks. The scholarly humanistic and computational techniques offer new ways to trace individual or collective participation in textual

12 Edlich-Muth and Edlich-Muth, "A Computational Approach to Source Adaptation."

13 Camps, et al., "Noisy Medieval Data."

14 Vandyck and Kestemont, "Abbreviation Application."

production. Although they differ in transcription strategy, degrees of literary-historical certainty, and disciplinary historiographies, all three share a commitment to integrating computational modelling and interpretive inquiry. Crucially, the accelerating development of HTR for medieval manuscripts is without a doubt transforming the field. It allows researchers to work with transcriptions drawn directly from manuscript sources, capturing not only the authorial layer—typically accessed through normalized texts—but also the scribal and transmissional layers, through character-level modelling of individual documents. This dual focus expands the analytical horizon, enabling scholars to study computationally both the creation and the transmission of medieval texts.

Experiment 1: Mapping Hands Visible to the Eye to Scribal Behaviour

Manuscript catalogues are full of examples of descriptions of codices in which multiple hands have been identified. Usually, such descriptions are limited to a brief mention of the use of several hands, mention of a particular number of hands, or perhaps an indication of the specific folios on which a hand changes. It is unusual for such an observation to be accompanied by a folio-by-folio description of the hands, mapped onto observations about the hand changes and collation, however. The work of palaeographic analysis and collation is time-consuming, after all, and also highly dependent on both expertise and the material state of the binding of manuscripts.

According to the early twentieth-century Cambridge cataloguer James, there were three scribes involved in the copying of an English bible in the Paris bible style dated to ca. 1270–1280 from their collection, CCC, MS 49. In his description in the college's manuscript catalogue, James remarked that one hand wrote the "Pentateuch (?), Psalms, Maccabees etc.," another can be seen in "Proverbs etc.," and a third is visible in the Prophets and Epistles.[15] Such a terse and approximate description is not to be blamed, as the short catalogue descriptions were probably not designed to provide more detailed information. These comments by James encouraged us to examine the manuscript more closely, however, and the existence of a high-quality digital reproduction of the manuscript made available by the Parker Library on the Web provided us with the perfect opportunity to carry out our own palaeographical and codicological analysis. We were able to expand and

15 James, *A Descriptive Catalogue*, 98–100.

Table 6. Data from our visual analysis of the hand changes in CCC, MS 49, an English bible in the Paris bible style dated to ca. 1270–1280. Numbers in parentheses are implied based on the prose description provided by James. Data by authors.

Span of biblical books	Begin fol.	End fol.	James's judgement	Our judgement	Other naming	Comments
Letter of Jerome to Damasus–Letter of Jerome to Paulin	1r	3v	(1)	1	1.1	
Genesis–Deuteronomy	5r	72v	1	1	1.1	
Prologue to Joshua and Judges–Psalms	72v	215v	(1)	1	1.1	
Prologue to Proverbs–Proverbs	215v	223v	2	2	2.1	Hand changes at 216r
Ecclesiastes–Ecclesiasticus (or Sirach)	223v	246v	(2)	2	2.1	
Prologue to Isaiah–Isaiah	247r	247r	3	1	1.2	Hand changes at 247r
Prologue to Jeremiah–Ezekiel	263v	303v	3	2	2.2	Hand changes at 257v
Prologue to Daniel–Malachi	304r	331r	3	3	3.1	Hand changes at 288v
Prologue to 1 Maccabees–2 Maccabees	331r	349v	1	1	1.3	
Prologue to Gospels–Revelation (or Apocalypse of John)	349v	444r	–	3	3.2	No mention in James

nuance James's analysis of hands, gathering more granular information about their changes as well as the collation of the quires and the books, visualized in Table 6.

We also transcribed the whole manuscript using the HTR platform Transkribus, using an early version of the Latin model we trained at the beginning of our research in 2021. The resulting transcription amounts to approximately 700 thousand Latin tokens. The method we chose to analyze the full transcription of CCC, MS 49 is known as "rolling stylometry," which combines supervised machine learning classification with sequential analysis. Rolling stylometry looks across texts, in the words of the creator of the R package that features it, as "a set of linearly sliced chunks, in order to test their stylistic consistency."[16] It has been used primarily to

16 Eder, "Rolling Stylometry."

study authorial attribution in texts and to analyze co-authorship, as well as the role played by re-writing, translating and editing.[17] Rolling stylometry works by moving across the text and comparing the most frequent words in any given slice to a list of most frequent words of a portion of text that can be with certainty associated with a specific author (or in our case, hand). Since our transcription schema has preserved elements of medieval special letterforms, brevigraphs, and abbreviations, our experiment expands the notion of rolling stylometry as a practice used most often with printed texts to consider the scribal contribution of a manuscript. By stylistic consistency in the context of the Paris bibles, we do not mean only the specific word choices that occur from book to book of the biblical text—although such detail is definitely something that the algorithm could detect—but we consider an expanded notion of style that includes the features we have trained the HTR system to recognize in manuscript. Our hope, in short, was that a sequential analysis of the full transcription of the manuscript might detect the same shifts that a visual palaeographic analysis was able to identify. This method is interesting, of course, not because we are simply confirming human observation with an identical computational analysis of features, but because we are looking at a subset of features visible from the same hand-copied document.

Upon examination of CCC, MS 49, we concurred with James that three hands (1, 2, and 3) were involved, but as detailed in Table 6, the hands are interwoven across the manuscript as we encounter it bound today. We label each of these hands according to our visual detection of their appearance in the codex: 1.1, 1.2, 1.3, 2.1, 2.2, 3.1, and 3.2 (where 1.3 is the equivalent of the third occurrence of hand 1). It is important to note that the hands did not participate equally in the copying of CCC, MS 49, nor are each of the segments of equal length.

Our first question was whether rolling stylometry could, via linguistic features including medieval ones, recover the same hand shifts that we observed by eye. We applied a standard rolling-classification procedure known as "rolling delta" from the Stylo package in R that assigns labels to sequential chunks across CCC, MS 49. In the top plot of Figure 5, the candidate classes we used included Hand 1 (full contribution), Hand 2.1 (a subset of Hand 2), and Hand 3 (full contribution). Here we were testing

17 Notable examples from medieval French literature are provided for Chrétien de Troyes, as well as for Clément Marot's controversial 1526 edition of the *Roman de la Rose*. See, respectively, Reilly and Dillon, "Virtuous Circles of Authorship Attribution," and Eder, "Rolling Stylometry."

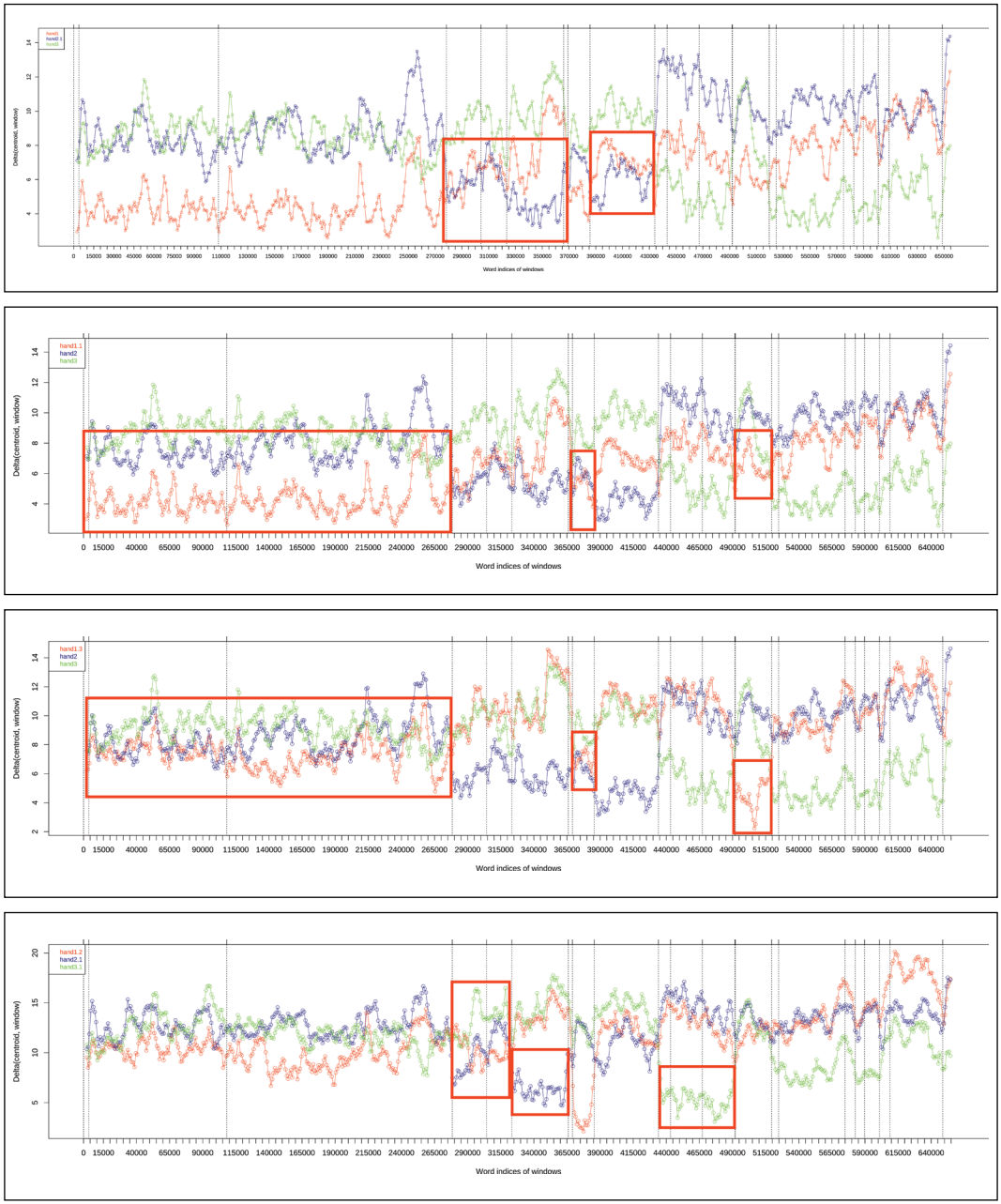


Figure 5. Rolling stylometry analysis of the different hands identified in the manuscript CCC, MS 49. The plot first from the top visualizes Hand 1 (red), 2.1 (dark blue) and 3 (green). The plot second from the top visualizes Hand 1.1 (red), 2 (dark blue) and 3 (green). The third plot from the top visualizes Hand 1.3 (red), 2 (dark blue) and 3 (green). The last plot visualizes Hand 1.2 (red), 2.1 (dark blue) and 3.1 (green). Transcription in Transkribus. Visualization created in R and Stylo by authors (rolling.delta function, 500 MWF, Classic Delta, Slice length 5000 with overlap of 1000, Milestones represent changes in hand, section and quire), adapted from graphs previously published in Guéville and Wrisley, "Transcribing Medieval Manuscripts."

whether using a small portion Hand 2.1 would generalize to the remainder of Hand 2 (Hand 2.2). The lowest curve at any position in the plots indicates the most likely class. The outcome of this first run of Stylo was encouraging: the rolling analysis reproduces our visual segmentation, even with a smaller portion of a given hand. The top panel clearly distinguishes the hands and their boundaries, supporting the view that abbreviations and medieval letterforms yield a stable stylometric signal characteristic of individual scribes.

We then wanted to test how much data was necessary to obtain such results. In the first plot, larger samples made attribution relatively straightforward. To test data limitations, we compared two subsets of Hand 1: a longer one (Hand 1.1, about 45 per cent of the manuscript) and a shorter one (Hand 1.3), in second and the third plot from the top. Using 1.1 as reference, the model easily recognized other Hand 1 sections. But using the shorter 1.3 as reference made recognition of a similar-length segment (Hand 1.2) more difficult. Finally, we tested whether even shorter samples—Hand 1.2, Hand 2.1, and Hand 3.1—could still detect hand shifts in the bottom plot. While results were less distinct than with longer samples, meaningful patterns still emerged. Signals sometimes overlapped (e.g., Hand 1.1 and 2.1), but with appropriate parameters, stylometry proved capable of identifying scribal identity. Importantly, we also discovered that relatively small text samples can suffice to track significant shifts across a manuscript. This finding is a pleasant surprise for our character modelling method, given that many who work with data will state that a significant amount of data is required to achieve high quality results in computational analysis, perhaps more than the typical humanist would have.

The results of this first experiment echo other findings from stylometric studies in a medieval context, where subtle layers of scribal or authorial intervention can be computationally distinguished if the appropriate features are modelled. It suggests that stylometry—when applied carefully to whole codices—can complement traditional palaeographic analysis, offering a reproducible way to detect shifts in scribal hands. While larger samples naturally yield clearer separations—we did automate the transcription of the entire manuscript, after all—our study suggests that even modest amounts of carefully transcribed and sampled material might preserve enough scribal signal to map individual activity across a codex.

Experiment 2: Mapping Multiple Hands to Different Signals

Mentions of multiple scribal hands in manuscript catalogues, as we saw in the previous example, are relatively common, but it is much rarer to find colophons that provide the name of a scribe or a date by which a manuscript can be situated in time or space. Ruzzier has compiled an important handlist of such manuscripts, which provides a valuable starting point for identifying candidate bibles for further analysis.[18] However, with a few additional examples we have identified to expand that list, such attributed bibles remain rare: Ruzzier's appendix lists only thirty-six manuscripts, and currently only about half of those manuscripts are accessible in digitized form. What makes a colophon naming a scribe particularly significant for computational analysis is the opportunity it offers us to study the manuscript forensically, as we did in the previous example with the anonymous (but visible) hands, testing where patterns of scribal behaviour might be detected by computational methods. A named individual offers something rarer: a direct link between a physical document and the historical record, grounding the manuscript in a particular context. In a largely anonymous world of medieval textual production, we might consider a named manuscript as an anchor, by which we mean a precious data point of certainty, providing a fixed point from which to map broader patterns of scribal practice, manuscript circulation, or historical agency.

In this section we turn to two manuscripts of particular interest to us: UPenn, MS Codex 236, a French bible attributed to Paris in the 1220s, and Mazarine, MS 6, likely to be French and dating from the thirteenth century. Although no direct connection between these two manuscripts has previously been asserted, they form an interesting pair for comparative analysis, given the claims made about their production. UPenn, MS Codex 236 does not contain a colophon, and its scribe(s) is/are unknown. However, a precise collation exists in the catalogue, alongside an unnamed manuscript cataloguer's assertion that the manuscript seems to be the product of "probably more than one hand, though it appears to be the work of a single scriptorium."[19] We interpret this comment to suggest a visual and stylistic consistency across the manuscript, in both decoration and handwriting. Our own careful visual inspection confirmed this impression of uniformity.

This uniformity raises an important question. Whereas we know something about the production of bibles in some urban contexts, we do not know

18 Ruzzier, *Entre université et ordres mendiants*, 283.

19 University of Pennsylvania Libraries, "Biblia sacra manuscripta."

everything about the genre. What was the nature of the coordinated effort to copy UPenn, MS Codex 236? Was this bible copied by a single, yet unattributed scribe, or could it be the product of a particularly well-coordinated collaborative effort—one subtle enough that it might only be detectable through computational methods? If the latter, might collation patterns offer additional evidence about how labour was divided during the manuscript's production? At the same time, we can also ask whether the scribal profile detected through computational analysis reflects unconscious, individual habits or whether it could have been consciously adopted—or imitated—by expert scribes contributing to the collective effort of producing a codex. These possibilities frame our investigation into the internal structures of the manuscript and the methods best suited to uncovering them.

The second manuscript, Mazarine, MS 6, with which we began this chapter, presents a different problem. At the end of the OT (fol. 429r), a colophon identifies a certain Johannes de Cristemanneford as the scribe. But does this attribution mean Johannes copied the whole Old Testament, only a section of the OT, or the entire manuscript? Can a named scribe anchor the manuscript securely to a specific time and place, or was Johannes more akin to a manager or overseer of its production? Similarly, if a date is included in a colophon, as in the case of Dole, Médiathèque de l'hôtel Dieu, MS 5, does it indicate the actual date of completion of copying, or a more arbitrary or ceremonial moment? The question we pursue in this experiment is whether colophons can be trusted as reliable witnesses to the creation of a manuscript. As Cohen has suggested, "it is uncertain, however, whether the scribe is to be believed in all cases and whether the meaning of his words is being interpreted correctly."[20] Scribes could claim credit for work they only partially completed, and modern scholars may sometimes overinterpret the limited evidence a colophon provides. Through computational analysis, we aim to test what the manuscript itself reveals, to see how the colophon fits within the broader production context, to uncover hidden forms of scribal labour, and to explore whether a single name may in fact conceal a collaborative effort.

Using a recent HTR model trained on the variety of manuscripts detailed in Table 4, we transcribed the two manuscripts in their entirety. In this experiment, we change methods from sequential stylometric analysis in favour of a method for text classification known as Term Frequency-Inverse Document Frequency (TF-IDF), also used by Vandyck and

20 Cohen, "Can Colophons Be Trusted?."

Kestemont, as explained above. The transcription was exported page by page, following the organization of the digitized manuscripts in Transkribus, producing many hundred files per manuscript. TF-IDF is a well-known method for textual corpus analysis that quantifies the significance of words, characters, or n-grams within each file relative to their occurrence across the full corpus, with frequent local terms and rare global terms receiving higher weight.[21] After building the TF-IDF matrices, we applied Principal Component Analysis (PCA) to reduce the dimensionality of the dataset. In the case of this experiment, PCA identifies the two principal components—that is, the axes along which variation is strongest—and projects the complex data into two dimensions for visualization. These visualizations allow us to explore clusters of folios that share similar patterns of abbreviated words or distinctive scribal features, suggesting underlying structures that would otherwise be impossible to observe across large amounts of raw manuscript transcription.

We use this method for two main reasons: once transcriptions are available, it is relatively easy to apply and does not rely on extensive manual tagging. As discussed earlier in relation to stylometry, such an approach typically requires prior knowledge of a corpus, and some form of certainty, such as visible scribal hands, or a colophon, or the claim by a scholar that a specific text can be attributed to a particular person. A stylometric experiment is then structured around confirming or refuting an attribution based on that data. As in the section above about CCC, MS 49, multiple hands are compared to see which one predominates, using samples of known attribution to classify unknown sections. In this section with UPenn, MS Codex 236 and Mazarine, MS 6, our goal varies slightly. With the former, we have no indication where one hand begins or ends; with the latter, we have a colophon. Moreover, following Vandyck and Kestemont, we restrict analysis to character n-grams containing abbreviations or distinctive letterforms. Brevigraphs, they argue, function much like frequent words in authorship attribution, serving as distinctive markers of scribal practice, that is, they "are distinctive choices made by the scribes themselves, they are relatively content independent and they are spread evenly throughout the entire corpus."[22]

21 More specifically, we implement the TfidfVectorizer of the skikit-learn library, using character 4-grams using the top 500 features and full TF-IDF weighting to emphasize distinctive terms. TfidfVectorizer converts all text to lowercase, but no stopword removal is implemented.

22 Vandyck and Kestemont, "Abbreviation Application."

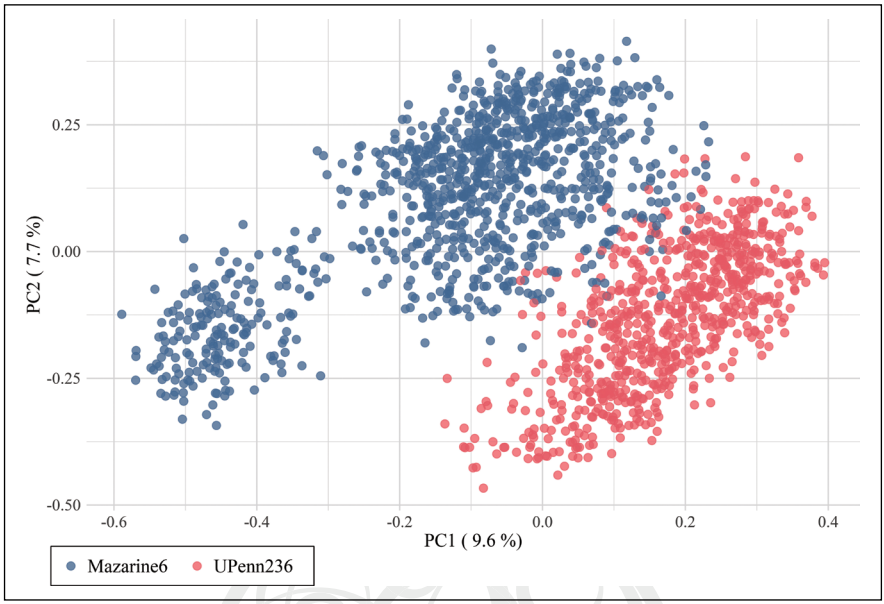


Figure 6. A 2D Principal Component Analysis (PCA) plot of a TF-IDF analysis of HTR-created transcriptions of two manuscripts, filtered for character 4-grams containing medieval letterforms, brevigraphs and abbreviations, Mazarine, MS 6 (in blue) and UPenn, MS Codex 236 (in pink). Transcriptions created in Transkribus. Visualization by authors, created in Python using TfidfVectorizer from scikit-learn and R using FactomineR and Plotly. Code adapted from Vierthaler, "NYU Abu Dhabi Stylometry," and from Stutzmann, Tensmeyer and Christlein, "Writer Identification and Script Classification." Eigenvalues: PC1=0.045 and PC2=0.036.

Adopting their approach, we processed our transcriptions to retain only character 4-grams that contain abbreviations or unique script features.[23]

TF-IDF analysis of transcriptions made from UPenn, MS Codex 236 did not result in clearly distinctive clusters belonging to parts of the manuscript, suggesting to us a higher degree of uniformity in the scribal profile. Although the divide between OT and NT chunks commonly appears when analyzing the text from the automated transcription, it is not as visible in the PCA in Figure 6. Several hypotheses may explain this finding. One possibility is that the manuscript was copied by a single scribe, making it difficult to distinguish multiple scribal habits. Alternatively, the manuscript may have been produced by a group of scribes with a strong house style, by which

23 We did not find that limiting our language sample to abbreviations, brevigraphs, and words containing special letterforms was effective in reducing the full textual signal. This finding can most likely be attributed to some abbreviations and brevigraphs corresponding to a uniquely NT lexicon.

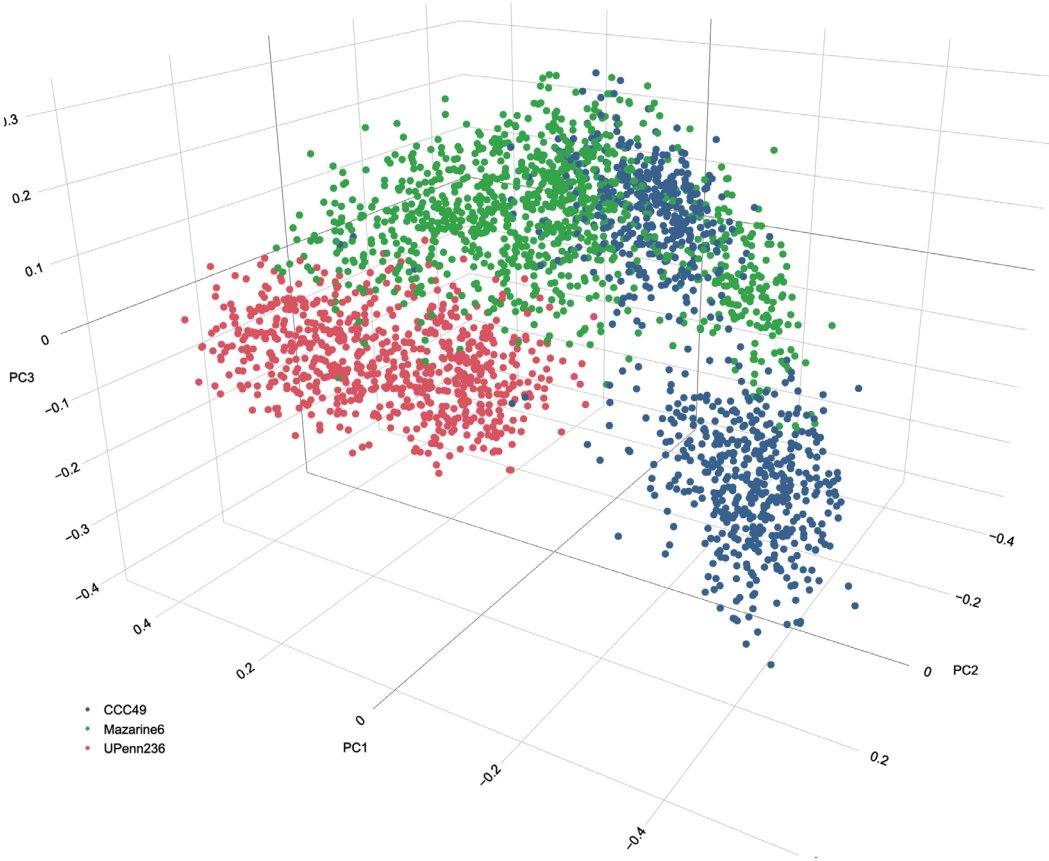


Figure 7. A 3D Principal Component Analysis (PCA) plot of a TF-IDF weighted analysis of HTR-created transcriptions of three manuscripts, filtered for character 4-grams containing medieval letterforms, brevigraphs, and abbreviations, Mazarine, MS 6 (in blue), UPenn, MS Codex 236 (in pink), and CCC MS 49 (in green). Transcriptions created in Transkribus. Visualization by authors, created in Python using TfidfVectorizer from scikit-learn and R using FactomineR and Plotly. Code adapted from Vierthaler, "NYU Abu Dhabi Stylometry," and from Stutzmann, Tensmeyer and Christlein, "Writer Identification and Script Classification." Eigenvalues: PC1=0.055, PC2=0.036, PC3=0.023.

we do not mean necessarily that they were working within the same physical space, but rather that they had an agreed upon copying style to which they conformed. Finally, they may have been copying faithfully from another unknown, yet uniform, exemplar. Our current results support the cataloguer's assertion that the manuscript was copied with a degree of regularity and the participation of multiple scribes cannot be definitively ruled out. If there were several scribes, a significant effort to converge on one style appears to be present. Future research using alternative techniques may be able to help provide different answers.

In the case of Mazarine, MS 6, the PCA of the TF-IDF analysis clearly distinguishes between the OT and the NT, in contrast to the previous example where the two sections are nearly indistinguishable. This difference is so pronounced that it suggests the NT was written by an entirely different scribe, or different scribes, distinct from John of Cristemanneford. And such evidence may confirm the reliability of the colophon in this instance: John links himself only to the OT and there is no proof at any point that he may have copied or been involved in any way with the production of the NT. The reliability of colophons depends both on their placement within the codex and the nature of the information they provide. Our positive results in the case of Mazarine, MS 6, should still be considered as hypothetical, and we caution against overgeneralizing information from colophons. While these can provide useful information, the results of our experiment reflect only one example and the way that we have modelled the entire manuscript. The method we outline here, instead of adding certainty to the analysis of manuscripts, can be useful for adding layers of new information for scholarly decision-making.

To explore our results with these two manuscripts further, we decided to evaluate them together with the manuscript from the first section of this chapter: CCC, MS 49. To reiterate, there is nothing that directly links these three manuscripts together, and we expect them to exhibit different features. Their distinctiveness is visualized in Figure 7. Interestingly, this three-dimensional analysis yielded similar results: all folios from UPenn, MS Codex 236 remain grouped together (in pink) whereas both Mazarine, MS 6 (in blue) and CCC, MS 49 (in green) are split in two. In CCC, MS 49, while one group seems to contain exclusively folios from the OT, the second group contains folios from both Testaments. This finding is also an interesting departure from the results from the first experiment where we could distinguish between three hands. Here, only two are visible: what corresponds to hand 1 and hand 3 in the first experiment. One possible explanation for the absence of hand 2 could be that this specific scribe copied only small portions of text, about sixty folios at the end of the OT. It may also be that his scribal signal or his abbreviation pattern may be similar to either of the other scribes. Additional analysis would be required to understand why we obtained these results and why the second hand is not as clearly visible as the others.

Experiment 3: Three Bibles by the Same Scribe as the Master of the Bible of Manfred?

The first two experiments in this chapter illustrated how the method of modelling scribal behaviour in transcription is able to confirm, or nuance, what the trained palaeographer is able to distinguish in manuscript. For us, these experiments serve as a proof of concept: the method is robust and can be applied to cases in which expert opinion is not settled. In the next two computational experiments carried out in this chapter, we draw attention to the contexts of manuscript production. In this experiment, we assess a series of manuscripts created in the milieu of Manfred (1232–1266), son of Emperor Frederick II, Prince of Taranto and then King of Sicily starting in 1258. An unnamed illuminator, likely active in Naples around the third quarter of the thirteenth century, who decorated a bible for Manfred has been named the "Master of the Bible of Manfred." The manuscript in question is BAV, MS Vat. lat. 36 and it contains a colophon identifying him as the patron, and a certain Johensis as the scribe on fol. 494v: "Princeps Mainfride regali styrpe create, Accipe quod scripsit Johensis scriptor, et ipsum Dignaris solita letificare manu." (Prince Manfred, born of royal lineage, receive what Johensis the scribe has written, and deign to make him happy with your customary hand).

Aspiring to the patronage of his father, Manfred commissioned several sumptuous manuscripts, among which BAV, MS Vat. lat. 36 is the best known. The Master of the Bible of Manfred is believed to have had an atelier and to have worked both for the imperial court and for a clientele of scholars and wealthy students.[24] Given the historical details that we have around the creation of these manuscripts, the case of the Master of the Bible of Manfred raises a number of questions: what was the relation of the master illuminator to other illuminators working around him? With what copyists, parchmenters, and binders did illuminators like the Master of the Bible of Manfred work? Did they have a working relationship over time? Our method of detecting similar features in the scribal contribution to manuscript creation cannot answer all of these questions, some being better left to art historians, but it can detect the appearance of a scribal hand in more than one manuscript.

Scholars have identified a number of other manuscripts, including many bibles, likely produced by the Master of the Bible of Manfred or his group. In total, the manuscripts which have been proposed as belonging to his artistic direction are Palermo, Biblioteca Nazionale, MS I. C. 13; BnF, MS latin 40;

24 Toubert, "Trois nouvelles bibles."

Table 7. List of the pages that we automatically transcribed from three manuscripts using Transkribus for Experiment 3 in this chapter: BnF, MS latin 40; BAV, MS Vat. lat. 36; and BnF, MS latin 10428. Data by authors.

Manuscripts	Book of Vulgate	Fols.
BnF, MS latin 40	Exodus	23r–25r
	Numbers	40v–42v
	Josue	69r–71r
	Matthew	352r–354r
	Mark	366r–368r
BnF, MS latin 10428	Exodus	21r–23r
	Numbers	41r–43r
	Josue	63r–65r
	Matthew	289v–291v
	Mark	300r–302r
BAV, MS Vat. lat. 36	Exodus	24v–26v
	Numbers	31v–33v
	Josue	81v–83v
	Matthew	386v–388v
	Mark	400r–402r

London, British Library (hereafter BL) MS Add. 31830;[25] Turin, Biblioteca Nazionale, MS E IV 14;[26] BnF, MS latin 10428; Bourges, Bibliothèque municipale, MS 5; and BnF, MS latin 217.[27] Tourbert has argued that some manuscripts were painted by the master himself, rather than by his entourage (BnF, MS latin 40, BAV, MS Vat. lat. 36, and BnF, MS latin 10428), although BnF, MS latin 40 probably had two helpers, perhaps apprentices or associates. BnF, MS latin 10428 is less lavish in appearance than BAV, MS Vat. lat. 36 and is of a considerably smaller size, details which distinguish it from the courtly works created for Manfred and which have led to the suggestion of

25 Daneu-Lattanzi, *Una Bibbia prossima alla Bibbia di Manfredi*; Daneu-Lattanzi, "Ancora sulla scuola miniaturistica," 105–62; Daneu-Lattanzi, *I manoscritti ed incunaboli miniati*; Daneu-Lattanzi, *Lineamenti di storia.*

26 Pettenati, "Un'altra 'Bibbia di Manfredi,'" 7–15.

27 Toubert, "Trois nouvelles bibles."

another client.[28] BnF, MS latin 10428 falls chronologically between BAV, MS Vat. lat. 36 and the slightly later Bible, BnF, MS latin 40.[29] But do they form a tight unit of manuscripts created under the supervision of the master and copied by the same hand?

The reconstruction of the manuscript illumination around Manfred draws on the collective expertise of generations of art historians, but what makes this scenario interesting to our work in computational textual analysis is that both BnF, MS latin 40 and BAV, MS Vat. lat. 36 contain colophons attributing them to the same scribal name: Johensis. Were we to be able to confirm the scribal contribution of Johensis in these two manuscripts, we would be in an interesting position to help add to the knowledge about the creation of the codices. While Johensis explicitly identifies himself in colophons within BnF, MS latin 40 ("Explicit expliceat ludere scriptor eat. Johensis" (fol. 432) and BAV, MS Vat. lat. 36 (cited above), BnF, MS latin 10428 lacks such a colophon and its scribal attribution remains uncertain. There are, of course, other manuscripts mentioned above; however, the convergence of scholarly hypothesis in Tourbet's strong assertion about the Master's direct involvement in these three manuscripts with the fact that these three manuscripts are already digitized provide an ideal scenario for computational analysis. Their existence in digitized format makes it possible to design a computational experiment that centres the role of the copyist, or the copying hands (rather than the artist's hands) in such a co-produced object. For these reasons, in the third experiment, we explore the possibility that the medieval scribe Johensis mentioned twice may be the same person and that he may have also been responsible for copying BnF, MS latin 10428.

To conduct this analysis, we used a selection of transcriptions extracted from these three manuscripts. The folios we used for our sample are listed in Table 7.[30] Our sampling method included twenty-five pages per manuscript, extracted as sequential excerpts of five pages from specific books of the Vulgate (Exodus, Numbers, Josue, Matthew, and Mark). Prologue pages were omitted to maintain consistency in lexicon.

We applied the same method of filtering the HTR-created text for character 4-grams, including special letterforms, brevigraphs, and abbreviations. As in the case of Mazarine, MS 6, from our second experiment, the

28 Avril et al., *Dix siècles d'enluminure italienne*, 53.

29 Avril et al., *Dix siècles d'enluminure italienne,* 54.

30 Since HTR can be a computationally demanding process, our choice to sample texts came from a dual interest in the sustainability of resources and material reasons. We discuss these reasons in the last section of this chapter.

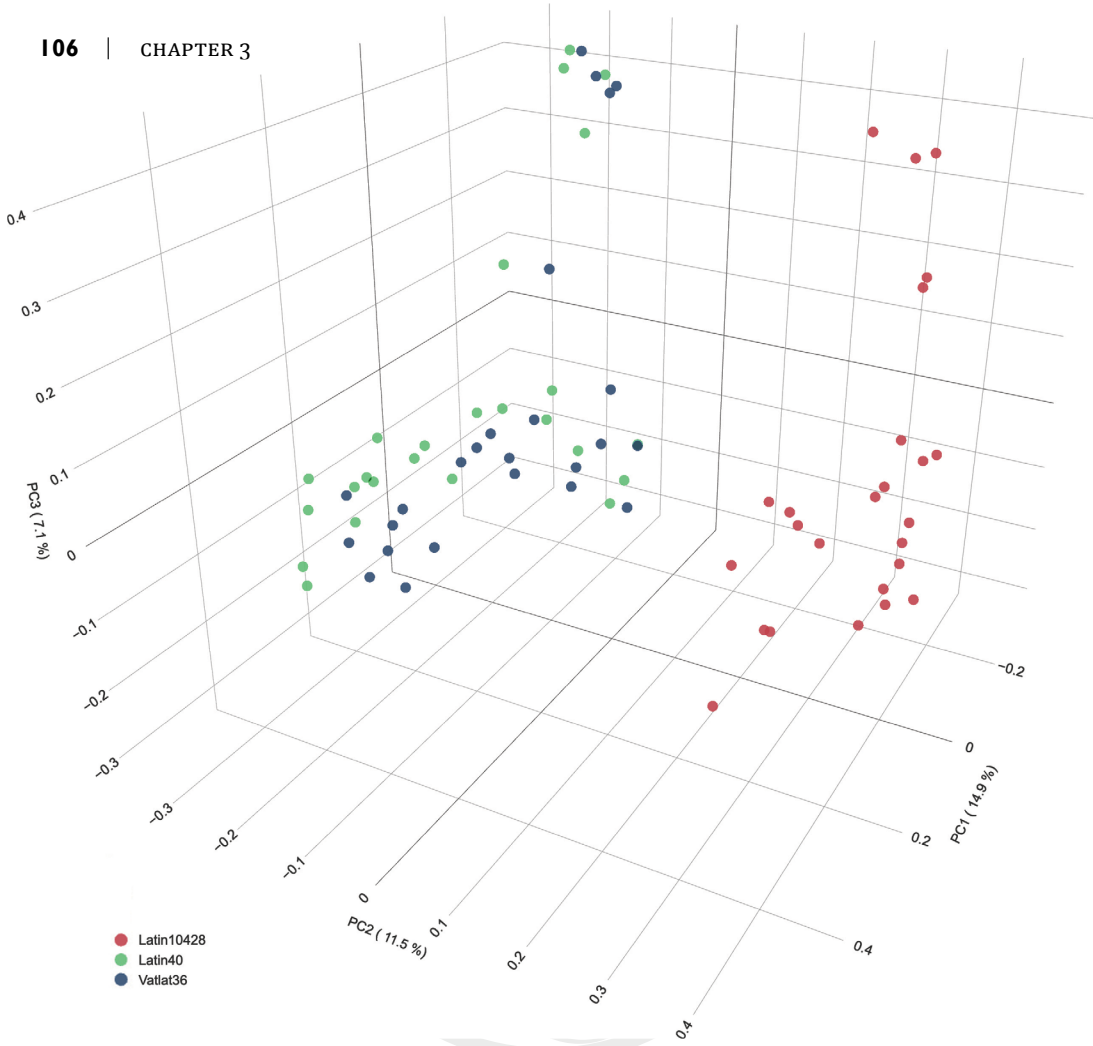


Figure 8. A 3D Principal Component Analysis visualization of a TF-IDF weighted analysis of character 4-grams containing special letterforms and abbreviations filtered from a sample of pages transcribed from BnF, MS latin 40 (green), BAV, MS Vat. lat. 36 (blue), and BnF, MS latin 10428 (red), as detailed in Table 7. Transcriptions created in Transkribus. Visualization by authors, created in Python using TfidfVectorizer from scikit-learn and R using Plotly. Code adapted from Vierthaler, "NYU Abu Dhabi Stylometry," and from Stutzmann, Tensmeyer and Christlein, "Writer Identification and Script Classification." Eigenvalues: PC1=0.078 PC2=0.06 PC3=0.037.

differentiation between the OT and NT books remained clear when working with filtered 4-grams. A second component of the PCA analysis also proved effective in distinguishing between individual scribes. Figure 8 illustrates the results of the TF-IDF analysis of the seventy-five pages sampled from each manuscript, but reduced to three dimensions in this case. The analysis clearly suggests that the scribe of BnF, MS latin 40 and BAV, MS Vat.

lat. 36—both identified as Johensis in two colophons—are indeed the same person. Furthermore, based on the clustering of data points, we can now assert, not only that the colophons with the name Johensis seem to point to the same scribe, but also that the scribe of BnF, MS latin 10428 appears not to be Johensis.

Although BnF, MS latin 10428 was decorated by the Master of the Bible of Manfred, its material and scribal features differ significantly from the others illuminated in the same atelier. It is significantly smaller, less lavish, and now we know it was clearly produced by a different scribe, perhaps pointing to a different patron. This case highlights how our method can be used with a group of manuscripts for which there are a number of existing theories, but also raises important questions about how medieval "ateliers" operated and the dynamics between scribes and illuminators. It suggests that scribes and illuminators did not always work in fixed partnerships; instead, illuminators might collaborate with different scribes depending on the nature of the commission, the intended patron, or the resources available. It would be interesting to compare with samples digitized from the other manuscripts evoked at the beginning of this section—which would require considerable effort on behalf of the libraries or scholars to visit them in person—but provided the analysis we have been able to carry out so far, the work of the Master of the Bible of Manfred across different manuscripts thus reflects a more fluid and project-based model of production over a rigidly organized workshop of long-term partnerships.

Experiment 4: Attributing Hands to a Scribal "Atelier"

From the previous example for which the context of manuscript production is not only specific—the patronage of the son of Frederick II of Sicily—but also well documented, we move to another example, perhaps more complex and significantly more protracted in time than that of the Italian example discussed above: that of the so-called Johannes Grusch atelier. The notion of there being an "atelier" associated with a certain style of Parisian manuscript painting seems to have been asserted for the first time in the landmark book about thirteenth-century illumination by Branner.[31] Branner's research was carried out around the same time that Toubert made her argument about the atelier of the Master of the Bible of Manfred. He comes up with the name "the Johannes Grusch atelier" and identifies it as a decades-long group of

31 Branner, *Manuscript Painting in Paris*.

artists and decorators of manuscripts, spanning the period generally corresponding to the reign of King Louis IX.

The historical record indeed contains the name of a scribe "Joanne Grusch," a scribe of a thirteenth-century bible in the Paris style, now held in a Swiss collection. His name is included in the colophon to Sarnen, Kollegiumsbibliothek, Stiftsarchiv Muri-Gries (hereafter Sarnen KB), MS Cod. membr. 16 containing a *terminus ante quem*: "Sacra Biblia scripta a Fr. Joanne Grusch O.S.B. Mon. Morenj. qui anno 1267 die 24 Martii obiit (ex Necrologio Murenji) annum agens 49" (This Holy Bible was copied by Frater Johannes Grusch O.S.B. of Kloster Muri (?) who died at the age of 49, on March, 24 1267). An important distinction must be made, however, in Branner's attribution of an atelier named after Grusch. It was not, in fact, based on the identification of Grusch's contribution in the copying of other manuscripts he attributes to the atelier, but based on his scholarly judgement that the illumination of Sarnen KB, Cod. membr. 16 bears a resemblance to the illumination of the other manuscripts. Branner asserts a similar logic in labelling an unnamed illuminator after a named scribe in an earlier article about BL MS Royal I. D. i.[32] In it, he explains that the artist of the latter is known as the William of Devon Painter because he illuminated a bible copied by the eponymous scribe. The process of identifying the illuminator(s) and naming them was a metonymic one: "For lack of a better term I shall call the Parisian atelier that of the Johannes Grusch painter, Johannes Grusch being, like William of Devon, the scribe of one of the manuscripts."[33] And such names persisted.

In *Manuscript Painting in Paris During the Reign of St. Louis* other ateliers, shops, and groups besides the Johannes Grusch Atelier are enumerated in Appendix V, the "Working Lists of Manuscripts": the Almagest Shop, the Alexander Shop, the Blanche Shop, the Atelier of the Vienna Moralized Bibles, the Toledo and Oxford Moralized Bible Ateliers, the Amiens Atelier, the Guines Atelier, the Pierre de Bar Atelier, the Gautier Lebaube Atelier, the Soissons Atelier, the Duprat Atelier, the Mathurin Atelier, the Bari Atelier, etc. Branner's persistent metonymic naming practice, associating groups of manuscripts with a scribe, a place, another person, the name of a work or manuscript, corresponds to a grouping of similar production according to known qualities or styles in others. Indeed, giving name to stylistic

32 Branner, "The Johannes Grusch Atelier," 24.

33 Branner, "The Johannes Grusch Atelier," 25.

groupings had important implications for the ways in which illuminated manuscripts from Paris have been received since the late 1970s.

Kidd has argued that Branner's terminology equating "ateliers" with "styles" is "highly problematic," mentioning that there has also been no monograph-length scholarly work published since to revise Branner's take on manuscript illumination at the time of Louis IX.[34] Branner's interchangeable use of the terms "style" and "atelier" has led to widespread confusion, suggesting that manuscript production was collocated in physical space, in a much more situated condition of manuscript creation than was probably the case. The Parisian book trade would most likely have been much larger an operation than the patronage circles of Manfred. In Paris they depended on networks of specialized skills of production by lay people: binders, parchmenters, scribes and illuminators.[35] Kidd contends that many illuminators likely worked independently, in similar, but distinct styles, rather than within shared paintshops (Branner's term). Kidd also contests the uncritical echoes of Branner's style labels in subsequent scholarship and dealer catalogues.[36] The Grusch group, perhaps because of its identification with the name of a scribe, has led to confusion of the role and perhaps over-attribution, lending prestige to manuscripts or manuscript fragments for sale. With Kidd, we believe that a cautious, evidence-based approach to manuscript attribution that avoids the perpetuation of outdated or oversimplified frameworks, is a way forward.

For a convenient shorthand in this chapter, let us call the network in which Grusch was an actor, the "Grusch network" for short. At the time we are writing, some of the manuscripts identified as associated with the Grusch network are available in digitized form, and the methods we have developed in previous examples provide us with the opportunity of thinking about these manuscripts through the optic of scribal contribution. The six manuscripts we used in this experiment include Sarnen KB, MS Cod. membr. 16; BnF, MS latin 15477; BnP, MS IL 93; Free Library, MS Lewis E242; BnF, MS latin 179; and BnF, MS latin 211. In reality, many more manuscripts were listed in Branner's Appendix, yet we chose to focus only on the Paris bible genre for the relative thematic coherence it provides us. That left us with six manuscripts to which we have access. It is worth mentioning—as is not atypical of much scholarship in the humanities—that Branner was not

34 Kidd, "Introduction," 9.

35 Rouse and Rouse, *Manuscripts and Their Makers*.

36 Kidd, "Introduction," 12.

uniformly certain about the attribution to his category of the Grusch atelier. In fact, his definition of the atelier is broad: a group of illuminators and copyists working over many decades to produce a number of styles.

Compared with the previous three sections of this chapter, we faced significant material challenges in collecting the data for this computational experiment. Of the six Paris bibles mentioned in Branner's handlist of the Johannes Grusch atelier, only three of them were digitized and only two are openly available online. The original manuscript signed "Joanne Grusch" was made available to us by a digitized microfilm from the Hill Manuscript Museum and Library. The other three (BnF, MS latin 179; BnF, MS latin 211 and BnF, MS latin 15477) are undigitized, but readily accessible in the manuscript consultation room at the BnF, where photography for private research purposes is thankfully permitted.

Following the sampling methodology detailed in the third experiment of this chapter, we assembled 125 pages of transcription drawn from six manuscripts, totalling 750 pages. The selection was in part shaped by the desire to have a sampling of text from the whole Bible, from both the OT and NT, but also was also dictated by the material conditions of the objects under study: many thirteenth-century bibles cannot be fully opened without risking damage due to their stiff and compact bindings. Consequently, the choice of books—Exodus, Numbers, Joshua, Matthew, and Mark—was determined pragmatically, based on the sections most accessible with a simple, non-invasive camera setup. Rather than waiting for large-scale institutional digitization efforts, we adopted an agile, on-demand digitization practice, enabling scholarship to proceed flexibly within the constraints posed by manuscript materiality. For us, this approach represents a carefully considered compromise between respecting the integrity of the manuscript as a physical artifact and still pursuing the pathways provided by computational analysis.

Here we sought to explore the extent to which the scribal profile of the identified copyist of Sarnen KB, MS Cod. membr. 16 resembles those found in other manuscripts identified by previous scholarship as associated with the same "Grusch network" of production. In this network, the roles of scribes, decorators, patrons, and intermediaries overlapped and shifted, prompting us to ask whether certain actors worked together repeatedly across multiple projects. Using a TF-IDF-weighted computational analysis of our transcriptions, we generated a three-dimensional principal component analysis (PCA) to visualize similarities in scribal practices (Figure 9). The results provided one strong conclusion that we did not set out to understand from the beginning: BnF, MS latin 179 and Paris, BnF, MS latin 211 exhibit very

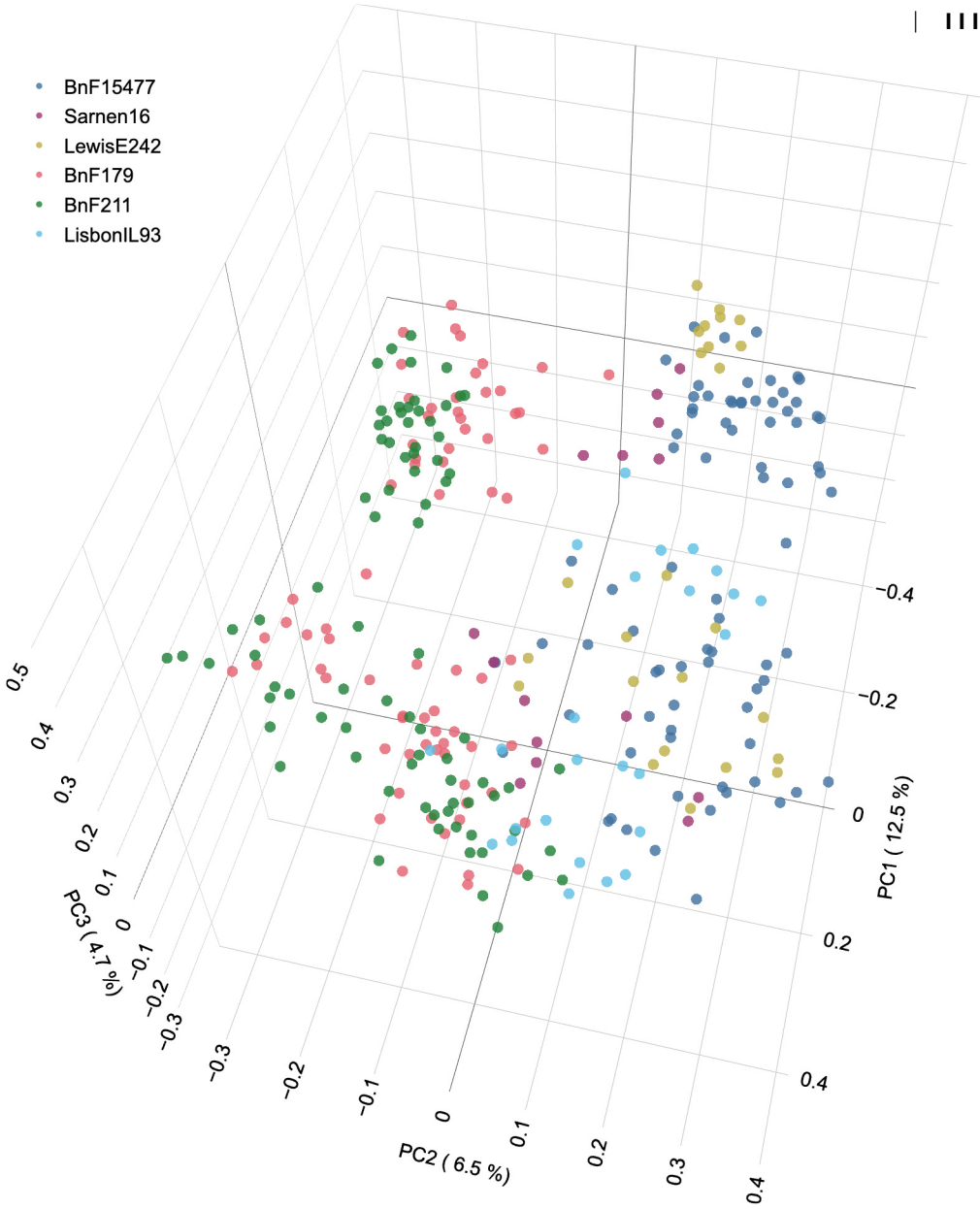


Figure 9. A 3D Principal Component Analysis (PCA) of a TF-IDF weighted analysis of HTR-created transcriptions of six manuscripts, filtered for character 4-grams containing medieval letterforms, brevigraphs, and abbreviations filtered, Sarnen KB, MS Cod. membr. 16; BnF, MS latin 15477; Free Library, MS Lewis E242; BnP, MS IL 93; BnF MS latin 179; and BnF MS latin 211. Transcriptions created in Transkribus. Visualization by authors, created in Python using TfidfVectorizer from scikit-learn and R using Plotly. Code adapted from Vierthaler, "NYU Abu Dhabi Stylometry," and from Stutzmann, Tensmeyer and Christlein, "Writer Identification and Script Classification." Eigenvalues: PC1=0.068, PC2=0.035, PC3=0.026.

different scribal behaviour than the other six manuscripts. Not only do they suggest that not all manuscripts previously grouped under a single "atelier" share the same copying practices, but the data suggest that the copyist(s) of these two manuscripts are indeed the same, or follow very similar scribal conventions.

The relationships among the other four manuscripts are more nuanced. Free Library, MS Lewis E 242 shares a scribal profile more closely aligned with that of Sarnen KB, MS Cod. Membr. 16 and BnF, MS latin 15477, although the latter without such a clear overlap. For what concerns the OT, there could be a relation between the sampled sections of BnP, MS IL 93 and Sarnen KB, MS Cod. Membr. 16, but the NT segment of the former is definitely not the scribal profile of Grusch in the latter. While more comprehensive transcription would be necessary to pursue a chronostylometric analysis capable of testing hypotheses about temporal phases of activity, these preliminary findings already complicate assumptions about stable workshop models. They point instead to a more flexible and dynamic configuration of collaboration, where groups of actors might come together for particular commissions, but not continue to collaborate after that.

Scholars in the field of book history have argued that *librarii* of the thirteenth and fourteenth centuries served as central coordinators of manuscript production, orchestrating the efforts of various craftsmen to deliver finished volumes. Nevertheless, data concerning the frequency with which copyists were subcontracted by booksellers remains incidental. If, as Rouse and Rouse and Kidd contend, manuscript production operated through a flexible and decentralized system involving many semi-independent workers contracted by booksellers, then the systematic examination of a broader set of manuscripts—particularly those with known origins and dates—may reveal previously unrecognized affinities between works, especially among manuscripts lacking explicit attribution. It is important to underscore that for the last two case studies discussed here—the networks surrounding Grusch and Manfred—other scribes may well have participated regularly in manuscript copying activities. In this chapter, we have deliberately focused on relatively small, curated sets of manuscripts in order to deepen our analysis of named individuals within the archival record; indeed much more research needs to be carried out.

In Table 8, we summarize, based on our analysis, the early, middle, and late periods into which Branner categorized manuscripts of the so-called Grusch atelier. In the fourth column, we present our own assessment, derived from the TF-IDF analysis discussed above. In the rightmost column, we further propose what we term a possible "companion scribal

Table 8. An overview of previous scholarly attributions of scribal identity (notably from Branner) along with our assessment of the identification with Grusch as copyist for six manuscripts: Sarnen KB, MS Cod. membr. 16; BnF, MS latin 15477; Free Library, MS Lewis E242; BnP, MS IL 93; and BnF, MSS latin 179 and latin 211. The last column offers a possible companion style, based on the data analysis of the scribal profile found in Figure 9. Data by authors.

Shelfmark	Digitization	Scholarly attribution of illumination group	Our attribution to Grusch copyist	Possible companion scribal style
Sarnen KB, MS Cod. membr. 16	HMML	Late Grusch atelier (Branner, *Manuscript Painting*)–Colophon	—	BnF, MS latin 15477 Free Library, MS Lewis E242 BnP, MS IL 93 - OT
BnF, MS latin 15477	authors	Late Grusch atelier (Branner, *Manuscript Painting*)	Perhaps	Sarnen KB, MS Cod. membr. 16 Free Library, MS Lewis E242 BnP, MS IL 93 - OT
BnP, MS IL 93	BnP	"Perhaps" Grusch Atelier (de Sousa, *Sacra Pagina*)	OT perhaps; NT no	NT is another scribe
Free Library, MS Lewis E242	archive.org	Early Grusch atelier (Branner, *Manuscript Painting*)	Yes	Sarnen KB, MS Cod. membr. 16 BnF, MS latin 15477 BnP, MS IL 93 - OT
BnF, MS latin 179	authors	Middle Grusch Atelier (Branner, *Manuscript Painting*)	No	BnF, MS latin 211
BnF, MS latin 211	authors	Late Grusch atelier (Branner, "Johannes Grusch Atelier" and *Manuscript Painting*)	No	BnF, MS latin 179

style," identifying manuscripts that exhibit notable affinities according to the exploratory data analysis. This column gestures toward the unconnected mass of medieval manuscripts alluded to earlier—those works whose relationships have remained largely unexplored and for which we are likely never to have attributions to name. Although these suggestions remain provisional, one potential trajectory for computational humanities approaches

to medieval manuscript studies may lie in the development of metrics that, while distinct from traditional palaeographic methodologies, offer alternative perspectives for exploring co-production networks. Although such distant metrics would require further articulation and validation, it is conceivable that by integrating multiple elements of manuscript production—such as scribal profiling, illumination characteristics, and material features—scholars might be able to construct innovative new layers of data. These, in turn, could substantially enrich the interpretive frameworks available to book historians seeking to reconstruct the labour structures of medieval manuscript creation.

Conclusion

In this chapter, we have returned to the broader question of metadata and description in both manuscript cataloguing and scholarship—particularly the way that forms of attribution stem from decorative, script, or material impressions left in manuscripts. Our analyses have demonstrated that evidence gleaned from computational modelling sometimes confirms the material record and the intuitions of book historians, but also offers critical refinements that go beyond what is visible to the naked eye. Our study demonstrates a proof of concept for a method of studying manuscripts for which material evidence already provides an answer, but can add layers of confirmation to palaeographic and codicological arguments. It is also a computational method that can be used exploratively to suggest relationships between manuscripts for which we do not yet have robust data.

The historiography of manuscript studies has often privileged illumination and decoration as the main method for understanding the handmade codex, where it comes from and how it was put together, leaving the figure of the scribe relatively underexamined. Yet, as our experiments in modelling scribal practice suggest, understanding the copyist(s)—especially in traditions as variably transmitted as Paris bibles—is challenging, but still possible. Each of the four experiments we carried out illustrates different aspects of this potential. With CCC, MS 49, computational results aligned closely with traditional manual analysis, but perhaps reached their limits with short textual samples of one of the hands. With our experiment using Mazarine, MS 6, and UPenn, MS Codex 236, we were able to confirm the colophon of the former, but by contrast, revealed that adjusting feature sets—such as emphasizing character 4-grams—allow other structural signals to disappear or to surface. In the third and the fourth experiment, computational analysis takes us beyond confirming what has been assumed by scholars to be true

to nuancing our stance on scribes. In our investigation of the scribe Johensis, a partial transcription provided strong confirmation of attribution in two manuscripts and rejection of attribution in another, suggesting that computational methods may provide the opportunity to uncover relationships between art historical evidence, patronage networks, and scribal practice through sampling methods that respect the physical integrity of the codex and are more economical than full transcription. Moreover, the analysis of the six manuscripts attributed to the so-called Grusch atelier demonstrated the serendipitous nature of computational scholarship: while drawing into question the basic assumptions of previous scholarly categories, we are still able to surface patterns such as the likely participation of Grusch—or scribes imitating Grusch—in three out of six manuscripts studied.

Beyond the mass of detail that automated transcription creates, the integration of computational methods into the study of scribal trends radically reshapes our epistemic relationship to evidence from manuscripts. HTR-based transcriptions offer a different way of seeing, closer to the way that a medieval reader would have experienced a manuscript, but also one that allows for features across documents to be aggregated. The abstraction inherent in modelling does not necessarily flatten the manuscript totally, but exchanges tactile, up-close experience of the book object and the possibility of capturing larger patterns rooted nonetheless in the materiality of these objects. We have mentioned Andler's notion of the double enigma suggesting that artificial intelligence in this context has a tendency to promise interpretive insight. What HTR systems surface are sequences of characters without understanding their linguistic, cultural, or intentional significance. Any application of AI in manuscript studies must be accompanied by a critical reflexivity concerning the gap between what we model with these systems and what we claim they reveal. Such reflexivity must come through scholarly judgement and contextual sensitivity.

Looking ahead, several avenues for future research present themselves. A truly large-scale analysis across a broader corpus could test the scalability of these methods, particularly with more advanced machine learning tools. Expanding computational inquiry to non-textual features—such as stylometry of images, analyses of colours, decorative motifs, or iconographic microvariation—could forge new methods that bridge text-based and art-historical manuscript studies. Such work remains heavily dependent on the availability and quality of digitized resources, however. The manufacture of thirteenth- and fourteenth-century bibles, while moving toward standardization, remained deeply artisanal. The handcrafted nature of these objects—evident in scribal idiosyncrasies, variations in layout, and inconsistencies in

textual execution—offers critical points of entry for computational inquiry. At the same time, it poses significant obstacles to the "perfect" digitization that HTR and modern machine learning methods often presume. The mass transcription and computational study of these manuscripts expose both the potential and the limits of digital tools, foregrounding the complex interplay between human craftsmanship and computational abstraction in the analysis of medieval book production. The sheer number of surviving thirteenth- and fourteenth-century bibles—often described as the first mass-produced book in the West—further amplifies these complexities, since the tension between standardization and individual variation becomes a defining element in the study of their digital avatars. It is precisely within these frictions, between the scale of production and the handmade nature of each object, that some of the most inventive and creative moments of the Paris Bible Project have emerged, inviting new methodological approaches that acknowledge both human and machine in the reconstruction of medieval textual cultures.

That is to say, that when we speak of a large manuscript tradition of a couple thousand manuscripts, in reality, this tradition is not in the truest sense of the term accessible and open, the way that a national corpus of thousands of nineteenth-century novels might be considered to be open, so that we can present results in a reproducible fashion as is the case with more and more research in the computational humanities. For now, HTR-created custom transcriptions allow us to assess some of the claims made about manuscripts and from these preliminary results to create and refine workflows that will allow us—hopefully—to collect much more data for an inquiry into the larger corpus. As methods in the age of artificial intelligence evolve—and they are doing so rapidly in the second half of the 2020s—the methods that we use here may indeed be eclipsed by forthcoming technologies, allowing the means of data collection and capture, as well as the analytical frameworks for studying these data, to be refined. Only time will tell how rapidly the computational study of manuscripts will advance. One point is clear: it will be difficult to include undigitized manuscripts, or digitized copies of manuscripts that sit in dim or dark archives, in our future analyses.

Chapter 4

TOWARDS A FUTURE OF COLLABORATIVE MEDIEVAL STUDIES

IN THE FIRST three chapters of this book, we have examined both the scholarly motivations and the practical challenges of investigating a corpus of medieval manuscripts. In our case, it meant the involvement of digital and computational methods at many steps of the research process: reusing manuscript metadata, tracking down digitized manuscripts or in some cases on the fly digitization, AI model training to transcribe old handwriting, and computational analysis of the transcriptions. Throughout, we have invited our readers to consider the collective thinking and expanded skill set that has made it possible to assemble and analyze such a variable and geographically dispersed corpus. Mature systems for HTR and open-source toolkits in R and Python have freed researchers from building computational systems entirely from scratch, making it possible to focus more directly on the complexities of humanities data.[1] The significance of such tools lies in the ways they have reshaped how scholarship is both conceived and practised. By rendering new kinds of textual evidence visible at scale, digital and computational methods encourage iterative, team-based research, inviting a myriad of new questions that engage with larger trends in contemporary digital culture and the evolving ways evidence is mobilized and interpreted in academic inquiry. Medieval studies do not require computation in order to be interdisciplinary, of course, but it is increasingly obvious that platform-based, data-driven methods are shaping the conditions of possibility for transnational and funded, cross-disciplinary collaboration. Platforms not only enable the sharing and analysis of data across projects, but also set the terms for access and participation, influencing what kinds of questions can be asked and what forms of knowledge are most visible in scholarship.

In Chapter 1 we discussed the variance and disparateness of the corpus, and in the second, the intricacies of implementing a custom transcription system that preserves medieval abbreviations and letterforms and allows us to model how, and to what extent, scribes left a mark on the texts they copied.

1 Van Erp et al., "The Future of Digital Humanities Research."

DOI: 10.17302/mfsk15869

In Chapter 3 we demonstrated how straightforward statistical approaches can be brought to bear on computer-generated transcriptions, analyzing the resultant transcriptions on a subset of Paris bibles produced in a historically documented setting in order to detect more precise scribal participation in the creation of these book objects. In this last chapter, we shift perspective, broadening our focus to reflect more generally on the role of collaboration in already interdisciplinary fields like medieval studies, with a particular focus on the study of manuscript cultures and the possibilities such collaboration might enable. We consider both the past of medieval studies and contemporary trends, arguing that computational medieval studies can benefit from critical reflection on both co-authorship practices and the evolving frameworks of the academic commons. Such reflection is not only retrospective, but also foundational, as we look forward: it allows us to envision more sustainable, equitable, and impactful forms of research and pedagogy, while actively valourizing the labour, infrastructures, and collaborative cultures that make such work possible—particularly within an academic system still shaped by persistent logics of individualism and prestige. A collaborative and interdisciplinary model of research in medieval studies in dialogue with contemporary research data cultures can foster broader impact within and beyond the field, extending its reach and influence for scholars, especially those who are committed to building an innovative, but also more inclusive, scholarly community.[2] Such inclusivity and influence do not arise naturally from data cultures, however, and we must adopt ethical practices that sustain the scholarly commons.[3]

In the first part of the chapter, we take a look back at publication cultures of a century of medieval studies, approaching them through a critical approach to the framework known as the "science of science" (SciSci). Using bibliometric data from 38,000 papers in medieval studies (most of them produced in a pre-digital, pre-computational context), we demonstrate that although co-authored, article-length publications remain relatively rare, they have gradually increased across subfields. Moreover, we demonstrate that such co-authored work tends to attract higher citation rates, registering visibly as a scholarly presence, but also necessitates different forms of

2 A portion of this chapter was presented at the 2025 conference of the Association of the Computers and the Humanities (ACH): Guéville and Wrisley, "Co-Authorship, Collaboration, and Community." We would like to acknowledge the thoughtful comments of our five anonymous reviewers that helped us to reshape our argument.

3 Bowker and Star, *Sorting Things Out*; Risam, *New Digital Worlds*; D'Ignazio and Klein, *Data Feminism*.

graduate training in the humanities. Building on this long scholarly view, we fast-forward to the 2020s, and consider how digital platforms have opened new possibilities—not for publication, but for collective data creation. The move from our bibliometric study of co-authorship to collective data creation highlights a broader shift visible in digital scholarship: from analyzing the products of scholarship to examining and enriching the processes that generate them. Scholarly collaboration, in this sense, is not confined to article-length co-authorship, but also extends into the design, execution, and afterlife of data creation. Since our own computational approaches to medieval manuscripts have mostly centred on the transcription of textual content, we focus in particular on the rise of community-based training, emphasizing that its value depends not only on careful design and theoretical grounding, but also on the pedagogical framing of computational reuse of data. After considering the emergence of crowd-transcription projects in the early 2020s, we close the chapter by turning to the ways digital and computational humanists have elaborated the notion of an academic commons for sharing data and building upon the work of others.

The idea of a commons has particular resonance for the computational humanities, serving to consolidate shared resources for new forms of comparative analysis, reducing duplication of effort, and opening scholarship to a wider range of contributors. Its allure, however, comes bound to significant challenges, not the least of which is its integration into the pedagogy of medieval studies. To speak of a commons is also to raise questions of governance, credit, preservation, and sustainability, as well as to acknowledge the unequal conditions in which different scholars and institutions can contribute. In the age of AI, when data circulates rapidly across disciplines, industries and borders, a critical approach to commons-building is required—one that both accounts for the diverse audiences of interdisciplinary research and remains vigilant to the ways the expanding economy of generative AI threatens to appropriate and monetize shared, open scholarly data.

The dynamics we discuss in this chapter should be considered within the broader economy of scholarship, and digital scholarship in particular, both of which run the risk of commodifying their outputs while concealing the collaborative processes and labour that make them run. Despite the reliance of computational projects on teams of contributors—annotators, coders, designers, student assistants, librarians, volunteers—academic systems still can overemphasize, and reward, individual achievement and authorship. This tension reveals a deeper contradiction in the digital humanities: while it often champions openness, collaboration, and innovation, it can still be embedded in institutional cultures and technological systems that

undervalue, or even conceal, collective knowledge-making. When it does acknowledge collective labour, it can do so by consolidating existing research paradigms rather than offering innovative directions for reshaping the field altogether. We conclude this chapter, asking whether the computational humanities—a wave of research that turns to algorithmic analysis as both a tool and an object of reflection in humanities disciplines—will continue to champion such values in particular in medieval studies, and if so, how.

Working Together in Diverse Teams: Insights from the Science of Science

Over the past decade, a notable shift has taken place in academic research: collaboration and interdisciplinarity have been recast not simply as methodological choices, but as strategic imperatives for producing high-impact scholarship and a requirement for funding in many fields including the humanities. Collaboration is framed in some fields as an essential ingredient for generating high-quality outputs, reinforcing the notion that scientific value is increasingly being tethered to collective, cross-disciplinary efforts. But has this been the case for medieval studies? Emerging from this context is the interdisciplinary field known as the science of science (SciSci), which uses large scale data to examine the structures and incentive systems that shape the larger research ecosystem in the contemporary university.[4] SciSci investigates the idea of scientific impact, the interplay between productivity and creativity, the conditions that make collaborations effective, and the influence of both failure and success on a researcher's trajectory. By analyzing key metrics, the field theorizes fundamental mechanisms driving scientific progress, notably how current research structures may privilege accumulation over innovation and stability over disruption. A widely cited 2023 study analyzing forty-five million publications and almost four million patents across six decades argued that, contrary to the idealized narrative of science as a forward-moving engine of breakthroughs, contemporary scholarship is increasingly unlikely to produce disruptive work that reshapes a field's trajectory across all fields of knowledge.[5] This thesis raised uncomfortable questions about the structural conditions—bureaucratic, evaluative, and institutional—that govern modern research and that might inadvertently suppress transformative scholarship.

4 Fortunato et al., "Science of Science"; Wang and Barabási, *The Science of Science*.

5 Park et al., "Papers and Patents," 138–44.

On the other hand, some studies in SciSci have suggested that, although team-based knowledge production has become a default, special conditions are required for teams to destabilize entrenched ideas and open up new problem spaces. In an analysis of nearly twenty million papers over five decades, Wuchty and colleagues revealed that teams are required for complex, interdisciplinary and data-intensive work, and that their collaborative publications receive higher citation counts, a trend consistent across sciences, engineering, social sciences, arts, and humanities, yet such large teams tend to be risk-adverse, focusing more on incremental expansion of established knowledge and less on resetting the research landscape.[6] Another study argued that the most field-shifting research blends conventional knowledge with unexpected, atypical combinations of ideas: the highest-impact papers not having greatest novelty, but blending novelty and otherwise conventional combinations of prior work.[7] While such combinations are rare, teams are 37.7 per cent more likely than solo authors to introduce them, highlighting the role of collaboration in shifting a field. Although collaboration poses considerable challenges, a recent study demonstrates that what is known as "long-distance interdisciplinarity"—not the geographic distance of the researchers, but rather the integration of conceptually distant subfields—is associated with higher scientific impact.[8] These findings, based on bibliometric and network analysis, suggest that engaging with knowledge beyond one's primary discipline not only enhances the visibility and influence of research outcomes, but also opens up new research directions. A 2019 study nuanced these results, revealing that each additional discipline added to a research team increases research impact by approximately 20 per cent.[9]

Moreover, studies have suggested that the size and composition of research teams play a crucial role in determining the nature and impact of their contributions. A 2014 study examined articles published between 1900 and 2011 using three indicators: number of authors, institutional addresses, and countries represented.[10] In fact, as collaboration has steadily increased over the past century, increasingly large and diverse teams are required to achieve the same impact. In 2019, Wu, Wang, and Evans' analysis of over

6 Wuchty et al., "The Increasing Dominance of Teams," 1036–39.

7 Uzzi et al., "Atypical Combinations," 468–72.

8 Larivière et al., "Long-Distance Interdisciplinarity."

9 Okamura, "Interdisciplinarity Revisited," 1–9.

10 Larivière et al., "Team Size Matters," 1323–32.

sixty-five million papers, patents, and software products from 1954 to 2014, concurred that larger, more diverse teams are instrumental in advancing scientific and technological discoveries.[11] They noted, however, along with Uzzi and colleagues that smaller teams are more likely to introduce truly disruptive ideas and novel opportunities. This SciSci literature on team-based knowledge production points to the necessity of creating specific conditions for research, accommodating a wide range of disciplinary and technical expertise, not only for getting the research done, but also for fostering the methodological diversity that can challenge existing paradigms, and sustain collaboration over the long term.[12]

Bibliometrics, Collaboration, and Medieval Studies

One may reasonably ask why we have made this digression into the domain of the science of science (SciSci), and whether it bears any relevance to the field of medieval studies. Do interdisciplinary teams of two to four persons have the potential for disrupting the field? How might we go about finding out? In fact, the discipline of medieval studies has already been the subject of some bibliometric analysis in a few of its sub-fields. Nielsen looked at three decades of scholarship on the Crusades, using data from International Medieval Bibliography (IMB) and Bibliographie de Civilisation Médiévale (BCM), noting a general decline in the quantity of Crusades studies since 2001. Kaya used data from the Web of Science to investigate bibliometrics of Constantinople studies with the observation that keywords have a significant impact on how one does bibliometrics in transnational communities in the humanities and social sciences.[13] A more wide-ranging analysis was carried out by Hérubel who conducted a study of the titles, keywords, and disciplinary affiliations of more than eleven thousand papers presented at a key conference in medieval studies, the International Medieval Congress (IMC), between 1997 and 2002.

Hérubel argues that medieval studies are not confined to a single academic domain, but instead are an inherently heterogeneous field. Not only are medievalists required to master multiple languages and methodologies, he claimed, but research in medieval studies is particularly well-placed to

11 Wu et al., "Large Teams Develop," 378–82.

12 Siemens, "It's a Team if you Use 'Reply All.'"

13 Nielsen, "Research Output in Medieval and Crusade Studies"; and Kaya, "Bibliometric Analysis of Constantinople Studies."

explore combinations of methodologies. Hérubel's study identified thematic bundles that recur in this corpus of conference presentations, underscoring the difficulty of using single discipline-based bibliometrics to assess the field.[14] What is striking for our purposes, however, is that key metrics from the literature of SciSci discussed above—such as number of authors, academic rank, institutional and geographic affiliations, canonical versus innovative citation—are absent from Hérubel's analysis. His study has a different purpose in mind: conceptualizing interdisciplinarity and multidisciplinarity primarily as a feature of the research content, rather than as markers of social practices of research through citation. Hérubel's essay appeared in 2005, only eleven years after the annual IMC conference began at the University of Leeds. The present moment offers far greater scope for such inquiry. The IMC, together with a wide range of professional conferences and bibliographies across the world, represents a rich landscape through which future bibliometric analyses might generate deeper insights into the intellectual organization, collaborative patterns, and shifting priorities of the field.

In the rest of this section of the chapter, we turn to our own exercise of bibliometric analysis, that of prominent journals in medieval studies available via Publish or Perish—the popular software developed by an academic from Middlesex University.[15] Our analysis of over 38,000 articles published across thirty-seven journals in medieval studies (largely in English and French) offers insight into the historical role of collaboration within the field as well as indicators of recent trends.[16] Publish or Perish is a software application that retrieves citation metrics, including the number of papers, total citations and the h-index, from the Google Scholar API. Importantly, the software allowed us to explore key features emphasized in our discussion of SciSci above that were missing from Hérubel, specifically the number of co-authors on papers as a general index of collaboration in the field and the number of citations of these papers. As we will explain below, compared to other fields of the contemporary academy, medieval studies have been late to adopt large-scale collaborative research models, but the situation seems to be changing, albeit slowly.

Before looking at the data, we would like to clarify two main points about our use of bibliometric software. First, we are not equating citations

14 Hérubel, "Disciplinary Affiliations and Subject Dispersion."

15 Harzing, "Publish or Perish."

16 Impact factors could not be incorporated into our analysis since many journals in the humanities are not indexed in the Web of Science's Journal Citation Reports (JCR), the standard source for official impact factors.

of an article with the quality of research contained therein, but rather we see them as one way of estimating the visibility and impact of a research article. Second, we do not adopt the use of bibliometrics to evaluate the productivity and citations of individual scholars as they might be used in the context of a performance or promotion review. It has been argued, in fact, that bibliometrics in the humanities are highly problematic since they "do not adequately cover the non-uniform nature of humanities," leading some to adopt so-called altmetrics.[17] Furthermore, humanities scholars tend to be critical of citation counts and bibliometrics and the fact that they are not as effectively used by funding bodies to evaluate research quality in the humanities as they are in STEM contexts.[18] Nonetheless, we do believe that bibliometric data retrieved from Publish or Perish does offer interesting insights into general, longitudinal evolution with respect to co-authorship and the visibility and quantity of citation of co-authored research in English and French.

Among the twenty overall most cited papers from the sample of 38,000 articles in medieval studies, only one has been co-authored—a 2002 article from the *Journal of Medieval and Early Modern Studies*.[19] This finding alone suggests that the most cited work in medieval studies has traditionally been produced by individual scholars rather than by collaborative research teams. However, when adjusting for the age of a publication using a citations-per-year metric, a modest increase in collaborative scholarship emerges. Among the twenty most cited papers by annual rate, four have been multi-authored.[20] While this detail suggests a gradual shift toward collaborative modes of research, single-authored scholarship still dominates. Notably, eleven papers appear on both citation lists, underscoring that high-impact contributions—whether judged by total citations or annual citation averages—remain primarily the product of individuals or small research teams. This dual presence suggests continuity in what constitutes scholarly influence within the field, despite emerging signs of collaborative engagement.

Nonetheless, when examining patterns more granularly through statistical regression, different trends and historical shifts emerge. In Figure 10,

17 Hammarfelt, "Four Claims."

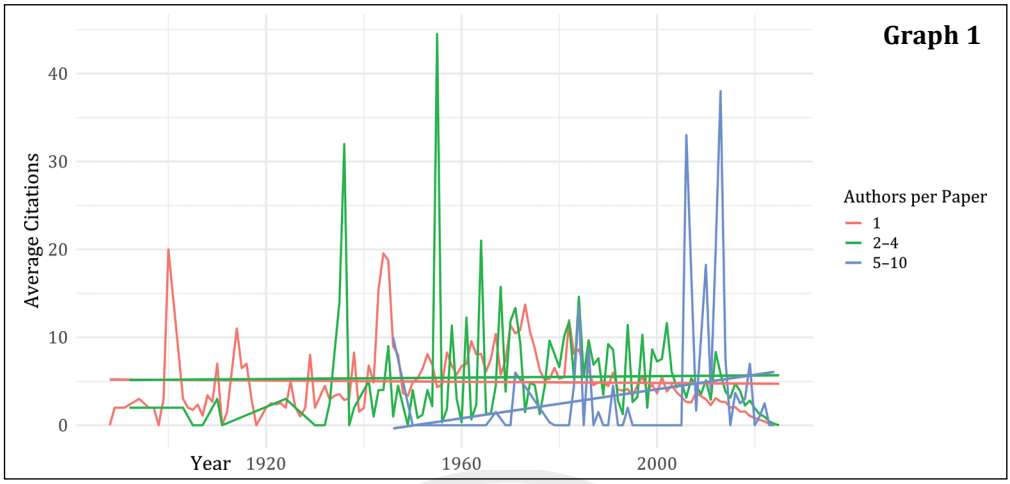
18 Hammarfelt and Haddow, "Conflicting Measures."

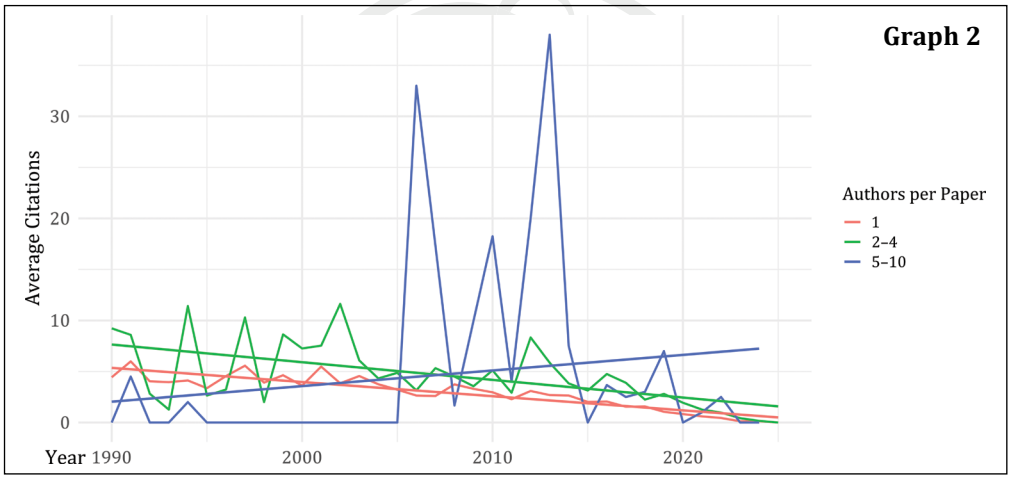
19 Ashley and Plesch, "The Cultural Processes."

20 Arnold and Goodson, "Resounding Community"; Price et al., "Polygyny, Concubinage, and the Social Lives of Women"; Ashley and Plesch, "The Cultural Processes"; Rambaran-Olm et al., "Medieval Studies."

Table 9. Data from thirty-seven journals in medieval studies.
Data retrieved by authors using the software Publish or Perish.

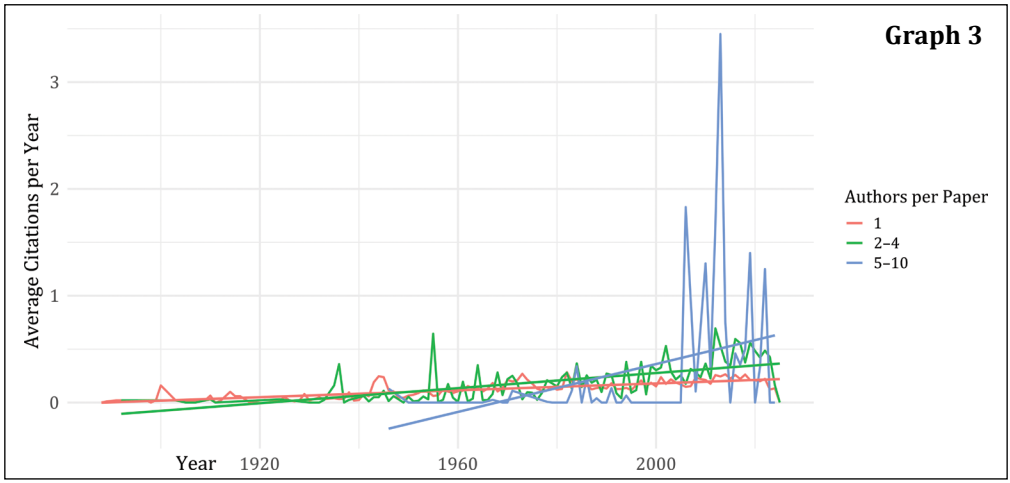
Journal Title	Dates	Percentage of multi-authored papers	First appearance of multi-authored paper
Bibliothèque de l'École des chartes	2008–2024	0.31	2010
Cahiers de civilisation médiévale	1958–2024	0.18	1986
Digital Medievalist	2003–2024	13.27	2008
Digital Philology	2012–2024	4.17	2013
Essays in Medieval Studies	2001–2024	0.33	2018
Gesta	1963–2024	1.89	1978
Hortus artium medievalium	1995–2024	6.39	1996
Journal of Early Books Society	1997–2024	0.56	2008
Journal of Australian Early Medieval Association	2005–2024	0	–
Journal of Medieval and Early Modern Studies	1970–2024	0.27	2008
Journal of Medieval Iberian Studies	2009–2024	3.66	2015
Journal of Medieval Monastic Studies	2012–2024	0	–
Journal of Medieval Military History	2004–2024	0.73	2011
Journal of Medieval Religious Cultures	1974–2024	0.39	2024
Le Moyen Âge	1888–2024	1.96	1951
Medieval Encounters	1995–2024	1.37	1996
Medieval English Theatre	1979–2024	1.62	1983
Medieval Feminist Forum	1984–2024	0.51	1996
Medievalia humanistica	1943–2024	0	–
Medieval Low Countries	2014–2024	4.76	2017
Medieval Sermon Studies	1991–2024	0	–
Medieval Worlds	2015–2024	3.55	2019
Medium ævum	1932–2024	0.06	1958
Mirator	2007–2024	3.9	2010
New Medieval Literature	1998–2024	0.53	2002
Nottingham Medieval Studies	1957–2024	0.67	1990
Peregrinations	2002–2024	0.31	2009
Peritia	1980–2024	1.04	1984
Perspectives médiévales	1975–2024	0.78	2007
postmedieval	2010–2024	2.46	2011
Revue d'histoire des textes	1971–2024	1.58	1971
Revue Mabillon	1905–2024	1.45	2000
Studies in Iconography	1975–2024	0.47	1981
Scriptorium	1946–2024	1.79	1946
Traditio	1943–2024	1.48	1947
Viator	1971–2024	0.43	1986
Viking and Medieval Scandinavia	2005–2024	2.59	2006





the first graph tracks citation counts by authorship size across time and reveals that single-authored and small-team papers published during the mid-twentieth century were generally more highly cited. However, this pattern begins to shift in later decades, with papers authored by larger teams (five to ten authors) with a noticeable increase in citation impact. This trend is captured by the upward trajectory of the blue regression line, while the citation impact of single-authored papers remains relatively flat and even exhibits a slight downward trajectory, *suggesting stagnation in their visibility and influence over time*.

Again in Figure 10, Graph 2 covers a shorter time span (1990–2024), and the trajectory of the average number of authors per paper and its relation to citation impact is made clearer. Here, a linear increase in co-authorship



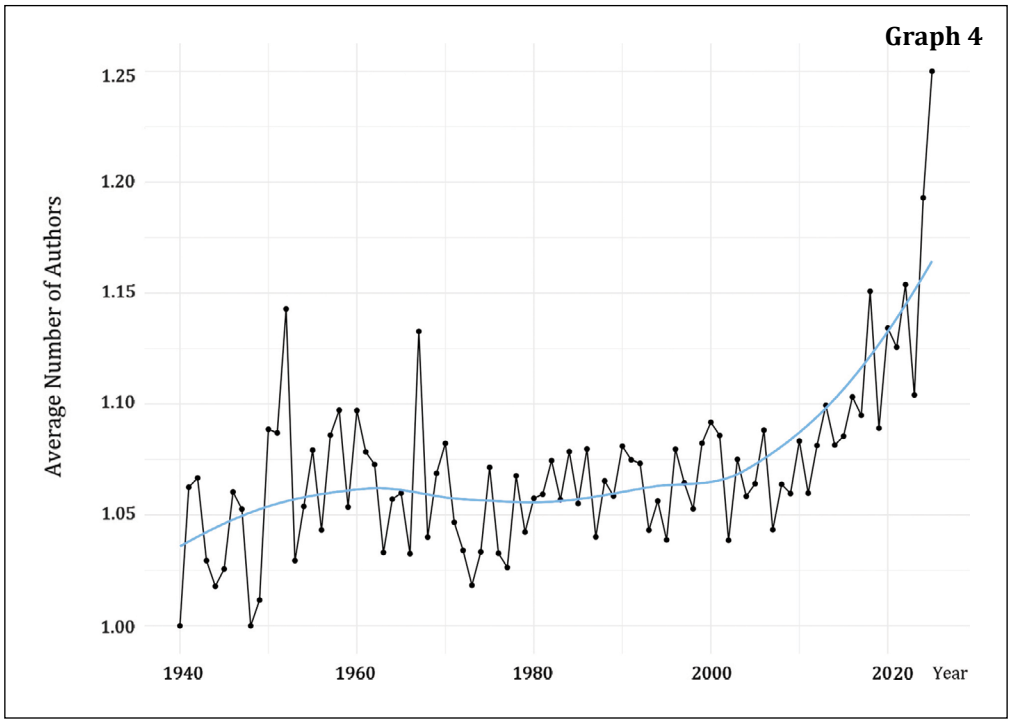


Figure 10. Graph 1 illustrates the impact of co-authorship on citation counts (total citations) between the 1900s and 2024. Graph 2 illustrates the impact of co-authorship (the number of authors per paper) on citation counts (total citations) between 1990 and 2024. Graph 3 illustrates the impact of co-authorship (the number of authors per paper) on citation counts per year between the 1930s and 2024. Graph 4 illustrates the evolution of the number of authors between 1940 and 2024. Data retrieved using the software Publish or Perish by authors. Visualization in R by authors.

becomes evident starting in the early 2000s. The red trend line representing single-author papers remains stable, suggesting little change in citation. By contrast, papers with two to four authors represented by the green line exhibit a modest increase in citation impact and generally outperform single-author publications. Most notably, papers with five to ten authors represented by the blue line illustrate a fluctuating but overall increasing trend, with significant citation peaks in 2006, around 2010–2015 and then in the 2020s. This higher variance suggests that while not all highly co-authored papers are equally impactful, the most cited works in recent years are increasingly produced by significantly larger groups of co-authors. In summary, these patterns point to a gradual, but clear, shift toward collaborative publication in the field in what could be considered mainstream, non-computational journals.

It is important to note, however, that the first two graphs represent the average total number of citations for all papers published in a given year. This approach inherently privileges older publications, which have had more time to accumulate citations, over more recent works—potentially obscuring the impact of newer research. To correct for temporal bias, dividing a paper's total citations by the number of years since its publication allows for a more equitable comparison across decades.[21] Graph 3 of Figure 10 presents citation impact normalized by time, displaying the average number of citations per year for papers published between the 1930s and 2024. While the earlier paper may have achieved greater cumulative recognition, the latter exhibits higher relative impact within a shorter timeframe. This normalization reveals that more recent, co-authored papers often achieve significant citation momentum early on, and may in time surpass the total impact of older, single-authored works.

Graph 3 of Figure 10 reveals a gradual increase in average citations per year across all authorship categories, reflecting a general upward trend in scholarly impact over time. Notably, papers authored by larger teams—specifically those with five to ten authors, represented by the blue line—exhibit a steeper trajectory and more pronounced citation spikes compared to single-author (red line) and small-team (two to four authors, green line) papers. While single-authored publications dominated in earlier decades, the prevalence and citation impact of multi-authored papers have grown significantly in recent years. In particular, works produced by teams of five

21 For example, a paper published in 1940 with forty citations would yield an annual citation rate of approximately 0.47, whereas a 2020 publication with fifteen citations would result in an average of three citations per year.

to ten authors demonstrate a marked increase in annual citation rates, indicating a greater visibility in the field. These patterns suggest a correlation between the number of authors and citation impact: papers involving larger research teams are, on average, more frequently cited.

In one last analysis, we applied a linear regression model to examine the relationship between the number of authors on a paper (AuthorCount), the year of publication (Year), and their combined effect on average citations per year in order to assess whether the influence of co-authorship on citation rates has changed over time. The results, visualized in Graph 4 of Figure 10, reveal a statistically significant negative main effect for AuthorCount (Estimate = -6.205, $p < 0.001$), suggesting that, when holding the publication year constant, papers with more authors historically tend to receive fewer citations per year. However, the interaction between AuthorCount and Year is both positive and highly significant (Estimate = 0.0031, $p < 0.001$), indicating that this negative relationship diminishes over time. In other words, while multi-authored papers in earlier decades were associated with, on average, lower citation rates, more recent co-authored works tend to attract higher annual citations—potentially reflecting evolving norms around collaboration and its perceived value in the field. Despite the statistical significance of these findings, this analysis accounts for only a small proportion of the variance in citation impact, suggesting that other factors—such as disciplinary subfield, journal reputation, topical relevance, or methodological novelty—play a far more substantial role in determining how frequently a paper is cited than authorship count alone.

In summary, looking at some 38,000 articles in medieval studies journals has demonstrated that while single-authored papers still dominate, there is a clear, gradual shift toward co-authorship, implying increasing forms of collaboration in research. Since the early 2000s, multi-authored articles—especially those with five to ten authors—have gained greater visibility and citation impact. It is important to underscore that correlation does not imply causation: co-authorship alone does not guarantee greater visibility or influence. Importantly, not only do we not know whether co-authorship on these papers stems from collaboration between the subdisciplines of the humanities or between the humanities and the social sciences or sciences, but we are unable to assess from the numbers alone how pathbreaking these papers have been. Suffice it to say that these trends point to a slowly shifting landscape in which collaborative research with larger numbers of authors is increasingly associated with higher citation in journal articles, even in medieval studies. Traditionally dominated by individual scholarship and monographic production, the field now appears to be aligning—albeit

gradually—with broader academic patterns observed in the sciences and social sciences, where collaboration has long been a driver of research visibility and influence. The rise of multi-authored, highly cited papers might, at least partially, be due to the development of computational, digital humanities, and large-scale collaborative projects. However, since such practices do not go back many decades in any field, the question is a relatively new one, and it is hard to judge.

While the preceding analysis offers provocative insights into a changing publication culture, it is important to acknowledge some of the significant limitations in studying a dataset of citation metrics that may underrepresent the prevalence of collaborative scholarship in medieval studies. There are many factors we are unable to examine, such as the academic rank and scholarly status of the authors, acceptance rates, or impact factor of the journals, as such data are not readily available. Furthermore, as we do not have millions of papers at our disposal, understanding the "disruption index" in the context of medieval studies seems problematic. Moreover, it must be acknowledged that the retrieval of data about scholarly papers using software and the Google API in only English and French could provide us with an inadequate view of journal publication. Perhaps more importantly, our choice to focus on core subject-area journals listed in Table 9 ignores some of the most important work of interdisciplinary medievalists already engaged in collaborative research who publish their work in digital or computational venues, leading us to underestimate how much co-authorship the field exhibits.[22] Such venues often emphasize and welcome collaboration, with research teams drawing on diverse expertise in computational methods, data science, manuscript studies, and beyond. For this reason, the exclusion of these publications from our dataset certainly skews our findings, understating the extent and influence of co-authored work within the broader contours of the field. Future work in the bibliometrics of medieval studies will need to broaden our approach beyond core subject-area journals, incorporating publications from interdisciplinary digital venues, conference proceedings, edited collections, project-based outputs, and conferences, perhaps combining them with altmetrics and qualitative approaches such as ethnographic, oral histories of interdisciplinary endeavours. Such inclusivity would provide a fuller picture of how collaborative and computational research is reshaping medieval studies.

22 The obvious outlier in our dataset is the journal *Digital Medievalist* which has at least three times more co-authored publications than some of the mainstream journals.

Platform-Based Co-Creation of Knowledge in Medieval Studies

In the previous section, we underscored how bibliometrics of mainstream medieval studies journals over a century reveal a gradual trend toward increased numbers of co-authored publications. By focusing on journal articles, however, we run the risk of obscuring a far wider terrain of collaboration that has long preceded formal publication in our field and in many others. The intellectual labour of data creation through building corpora, transcribing manuscripts, and curating datasets—essential for fields such as bibliography, digital editing, material philology, social history and historical geography—has for a long time remained structurally invisible in the profession's reward systems. The expanding body of digital scholarship in medieval studies makes this disjuncture ever more visible with time. Knowledge in the humanities is increasingly generated by many hands before it is ever formalized as argument in print, as the example of the creation of training data for applications of AI to text or image corpora makes clear.

In recent decades, with the availability of digitized surrogates, the physical encounter with medieval archives—a privileged, and historically central, mode of engagement for understanding the materiality and tactile dimensions of the archival object at the heart of medieval studies[23]—has increasingly been supplanted by mediated, screen-based interactions with digital avatars. Direct physical engagement allows scholars to assess aspects of archival objects often lost in reproduction: parchment, ink, marginal interventions, catchwords, collation, or details of mise en page; yet our time and frequency of access to them is also significantly limited. At the same time, the ways medievalists engage with their sources have been radically reshaped by processes beyond, and on top of, digitization: e-books, online repositories, the Internet Archive, bibliographic databases, and the International Image Interoperability Framework (IIIF).[24] Such mediated engagement has made it possible to consult geographically dispersed archives in digitized form, to compare textual versions across collections, and to participate in collaborative platforms that enable annotation, enrichment, and increasingly large-scale computational analyses. While the growing availability of digital facsimiles has undeniably expanded access—decoupling some forms of analysis from physical archives and inviting broader participation

23 Gilissen, *Prolégomènes à la codicologie*; Van Lit, "The Digital Materiality."

24 Francomano and Bamford, "Whose Digital Middle Ages?," 15–27; Hanneken, "What to Think About," 256–84; Vandendorpe, *Du papyrus à l'hypertexte*.

across geographies—it also reconfigures our relationship to our object of study, often privileging a flattened visual representation over embodied, tactile modes of engagement. In fact, high-resolution digital reproductions allow for forms of prolonged, daily engagement with archives that would be impossible in a reading room, creating opportunities for slow, detailed study, for teaching, and for collaborative analysis at a distance. Yet such sustained encounters, mediated through images, remain qualitatively different from the experience of handling the object itself.

The challenge (and promise) for computational medieval studies lies in bringing these two modes together: the irreplaceable insights of expert physical contact, and the expanded analytical capacities afforded by digital surrogates. By developing innovative methods inclusive of materiality while also exploiting the possibilities of textual, visual, and computational recombination the field will move forward—yet such methods also deepen the "computationality" of medieval studies,[25] where AI-driven mediation increasingly structures how knowledge is produced and discovered. This shift holds immense promise, but also raises urgent concerns: namely that algorithmic biases, opaque infrastructures, and commercial logics may shape what is visible, searchable, or valued, subtly reconfiguring the very terms of scholarly inquiry. Academic research, and the graduate education that underpins the profession of medieval studies, will not only need to innovate technically but, along with the computational humanities in general, remain critically attuned to the risks of such work. By being at the table, we believe, medievalists are more likely to influence decisions, and to model ethical and community-sustaining methods, than if they are blissfully absent.

Over the past decade or so, the notion of academic crowdsourcing—a research method that leverages public participation in, or contributions to, project activities—has gained traction as a strategy for expanding access and activating collections.[26] During and after the Covid-19 pandemic, a number of transcription challenges with medieval documents (also sometimes labelled a "transcrib-a-thon" or "transcriboquest") sprung up, due in large part to the leadership of European research teams and digitally-minded North American medievalists such as Laura Morreale. To date, some of these

25 Berry, "The 'Computational Turn,'" 1–22.

26 Hedges and Dunn, "Introduction," in *Academic Crowdsourcing*; Dunn and Hedges, "From the Wisdom of Crowds"; Ridge, "Crowdsourcing in Cultural Heritage."

included the Image du Monde Challenges,[27] *La Sfera* Challenges,[28] Houghton MS lat. 5: The Transcription Challenge,[29] Early Irish Hands TranSCRIPTion Challenge,[30] and the Saint Dunstan Transcription Challenge[31] and the different iterations of the Transcriboquest.[32] These events were often collocated with academic conferences or designed as summer schools. For instance, "Transcribing the *Pelerinage de Damoiselle Sapience*" happened during the Thirteenth Annual (Virtual) Schoenberg Symposium on Manuscript Studies in the Digital Age in 2020.[33]

Rather than serving merely as training in digital platforms or data collection, these gatherings foster shared experimentation and collaboration—central elements in the reinvention of medieval studies. They bring together diverse communities of scholars at different career stages, demonstrating how distributed scholarly labour, when carefully focused on specialized tasks, can generate first drafts of concrete editorial outputs and stimulate entirely new research trajectories. Such initiatives also reveal how platform-based approaches may facilitate time-bound collaborations that serve as models for larger, asynchronous modes of participatory knowledge production. They align with contemporary reinventions of collective intelligence as a dynamic process that is both technical and social.

---

27 The Image du Monde I took place from September 25 to October 9, 2020, with five teams working on five manuscripts containing the eponymous text by Gauthier de Metz. The Image du Monde II challenge was a reprise and took place from January 8–22, 2021, with teams working on the same five manuscripts.

28 *La Sfera* Challenge I took place between May 22 and June 5, 2020, transcribing three witnesses of one text, Goro Dati's fifteenth-century geographic treatise, *La Sfera*. *La Sfera* Challenge II which took place between July 17 and July 31, 2020, was a reprise of *La Sfera* I and included five teams. A spin-off of the transcription challenge involved geospatial analysis of the text. See Agostini and Beneš, "A Geospatial *La Sfera*."

29 This transcription challenge took place between March 23 and March 27, 2022, in hybrid format. The participants were students in Harvard's Latin Palaeography and Manuscript Culture course (Medieval Studies 202) who worked with a team of online volunteers to create a complete transcription of Cambridge, MA, Harvard University, Houghton Library, MS Lat. 5.

30 Volmering, "Early Irish Hands TranSCRIPTion Challenge | FromThePage."

31 The Saint Dunstan Transcription Challenge took place virtually, between July 21 and August 4, 2023. The volunteers transcribed a saint's legend in Middle English.

32 Biblissima+, "TranscriboQuest Summer School."

33 Over the three-day conference, participants produced a collaborative scholarly micro-edition of *Le Pelerinage de Damoiselle Sapience*, using UPenn, MS Codex 660. See Morreale et al., "Transcribing 'Le Pèlerinage de Damoiselle Sapience.'"

While some of these collaborative undertakings have matured into sustained projects, editions, or published articles, many others have not seen their data repurposed for broader computational research.[34] Despite popular narratives that platforms naturally enable collective work, the reality is more complex: collaboration is never an entirely open field in which all contributions are interchangeable. The outcome of many hands often depends on highly specific expertise, contextual judgement, and multi-step reasoning—whether in classifying galaxies, extracting information from complex documents, identifying ornithological species, or, as in our case, correcting draft ground truth for HTR from medieval manuscripts. Seemingly straightforward tasks such as transcription can entail significant learning curves, requiring participants to apply many specialized skills consistently that necessarily restrict participation to a smaller, trained, or highly motivated group.

In the Paris Bible Project we have tried to model what a future medieval studies might look like by experimenting with collaborative infrastructures that test how data models for highly specialized research questions can be implemented. We believe that their true value lies in broadening the research to include as many perspectives as possible. Many eyes on a body of material expose patterns, anomalies, and interpretive questions that no single scholar could identify alone, while also providing feedback that sharpens the design of models and methods. For this potential to be realized, however, significant community input and careful orchestration remain essential. Tasks must be clearly defined, supported by precise guidance, and moderated in ways that ensure quality and comparability across contributions.[35] What might appear to be processes that can run on their own, in fact, require sustained attention, with significant labour required for communication, design, and review. The challenge is not simply to collect more data, but to channel diverse forms of expertise and perception into productive feedback loops for project development. It requires providing robust quality controls, with coordination, planning, and documentation that sometimes demand as much effort as the scholarly analysis itself. Framed in this way, the scale of such initiatives is not measured only by output volume, but by the depth of insight that many sets of eyes can bring to complex materials—insight that, when sustained by dedicated infrastructure and a core team, can refine a data model, advance a specialized project, and open new research pathways.

34 The resulting transcriptions from the Image du Monde II challenge were subsequently studied by Guéville and Wrisley, "Everyone Leaves a Trace."

35 Blickhan et al., "Individual vs. Collaborative Methods."

Experimenting with Academic Co-Creation of Text

The academic co-creation of text, especially from medieval documents, is particularly intricate, requiring contextual expertise and prior experience. Within this landscape, event-based initiatives have emerged as important entry points for computational methods in medieval studies. For organizers, such events offer opportunities to activate particular collections, to prototype new workflows, or to build capacity within the broader scholarly community. For participants, however, the motivations for participation are often multiple and layered: they may join a crowdsourcing event to learn new technical skills, to connect with colleagues, or to experience how computation can intersect with traditional philological practices. Others may be drawn by the chance to make a meaningful contribution to collective projects, to find purpose and stimulation in uncertain times, or to participate in an intellectually and emotionally engaging form of research.[36] For us, the possibility of participants working in groups toward a well-framed research goal aligns well with principles of research-led teaching. Tasks such as transcription can support learning since they can draw on each other's strengths and experiences—linguistic, palaeographical, or technical. Focusing on methodology over data output allows for greater flexibility and more sustainable contributions from participants with varying levels of expertise. An example of a curricular embedding of a computational humanities experiment with HTR was our own "Paris Bible Correct-a-thon" organized at the Université Marie-et-Louis Pasteur in Besançon, France, in January 2023.[37]

Our event took a slightly different approach than the transcription challenges mentioned above, emphasizing critical AI perspectives. Instead of transcribing from scratch, we focused on crowd-correction of transcriptions previously generated through an HTR workflow, evaluating the automated outputs, revising them as ground truth to fine-tune HTR models.[38] The correct-a-thon spanned approximately one week and was carried out in collaboration with the Rare Book and Digital Humanities Master's program,

36 Hossain, "Users' Motivation to Participate in Online Crowdsourcing Platforms."

37 We would like to acknowledge and thank all the participants in the challenge organized in January 2023. Their names are listed in the Acknowledgements section of this book.

38 An asynchronous, public challenge took place in 2022 to encourage the automated correction of errors in HTR output from Byzantine manuscripts and papyri. The target audience was the NLP and AI community. According to the organizers some 271 submissions were received, "yielding only a handful of promising approaches." See Pavlopoulos et al., "Challenging Error Correction," 1.

and with the participation of colleagues from local universities and GLAM institutions. We framed the "correct-a-thon"—including a hands-on encounter with a few Latin bibles housed at the Bibliothèque municipale de Besançon—as a master's-level experiential learning exercise in book history.[39] The Paris Bible correct-a-thon worked with six manuscripts: Aarau KB, MS MsWettF 11; Besançon, Bibliothèque municipale, MS 4, Besançon, Bibliothèque municipale, MS 8; Beinecke, MS 1100; Palo Alto, Stanford University Libraries, MS 23 and one incunable, Beinecke, ZZi 56. Beyond the immediate pedagogical aims of transcription of medieval Latin, however, the event also pointed toward a broader trajectory for the future of academic work. As AI systems increasingly mediate scholarly access to, and processing of, cultural collections, expert opinion and critical AI perspectives remain indispensable for ensuring that outputs from processes such as HTR are accurate and meaningfully integrated into research. In this sense, graduate education takes on a new responsibility: preparing students not only to master technical workflows but to act as informed, critical mediators between automated processes and the cultural record. Such training positions the next generation of medievalists to take an active role in critiquing and shaping new professional methodologies, rather than simply inheriting ones of the past.

Our correct-a-thon had four primary learning objectives: (1) training in how to integrate digital tools into research in book history; (2) critical examination of AI tools in research, by learning to elicit specific outputs from them, as well as to evaluate those outputs; (3) acquisition of specific digital skills (using a text editor and a versioning system, training and correcting HTR models); and (4) participation in the enrichment and creation of a dataset. Two additional, more individual, objectives for us as organizers included (1) providing a model for collaborative public knowledge creation between students, researchers, and library professionals and (2) revising and expanding our project's abbreviation transcription guidelines through working with other manuscripts.[40] The event attracted considerable interest from students, with two master's students ultimately choosing to complete their second-year internships with our team. From the outset, the event prioritized a pedagogical focus on critical AI perspectives and collaborative,

39 See Guéville and Wrisley, "Correct-A-Thon."

40 Some of the specific learning objectives of the Paris Bible included: learning to use a text editor, using a versioning system such as GitHub, training and correcting HTR models, reflecting critically on HTR output, authoring reflection writing in Markdown, creating an ORCID account and using it when publishing one's data and working with computational notebooks.

experiential learning. In the end, the Paris Bible Project was able to add a half a dozen fully transcribed pages into the current training data for our HTR models. Considering the amount of post-correction required and the effort to organize and plan the event, however, a large significant increase in ground truth for the Paris Bible Project cannot be said to have been an outcome of the event.

While we organized the event as an intensive one-week exercise, the same activities could easily, and perhaps more productively, have been extended across an entire semester, allowing for deeper engagement not only with technical and collaborative skills but also with additional discussions of readings in critical AI, open standards in the GLAM sector, current debates in platform governance, and comparative experience working across HTR platforms. We underscore the importance of time in grappling with these complex issues: when framed pedagogically, introducing students or early career researchers to HTR workflows is not simply a matter of following consistent guidelines for transcribing medieval manuscripts. Rather, it opens up broader questions about the long-term scholarly implications of AI in research and pedagogy. The use of AI-based automation is not a one-off technical fix, but an ongoing methodological shift that demands sustained reflection in the academy—something that a week of practical exercises alone cannot adequately convey.

The iterative cycle of increasing the quality of HTR models—through phases of ground truth creation, model training, automated transcription, output correction, and model retraining—is slow and often opaque to non-specialists, demanding sustained commitment before tangible improvements are observed. The addition of the specificities of pre-modern manuscript context makes the tasks even more complex. Moreover, when only limited time or capacity is available for correction and retraining, the resulting outputs may not exhibit substantial gains in performance, running the risk of distancing participants from the payoff of their efforts. In this sense, the pace of HTR-driven research is markedly slow—frequently outpacing the shorter, goal-oriented time bound nature typical of an academic term or public engagement initiative. This mismatch underscores a broader tension in the integration of AI into humanities education (and into medieval studies) in our opinion: while the promise of human contribution to the improvement of applied AI processes is real, the reality of time-bound teaching is plagued by the potential of delayed returns and a lack of appreciation of the larger picture.

Academic co-creation within the humanities, be it inside or outside the curriculum, raises questions about accessibility, scope, and commitment.

While academic crowdsourcing in the humanities traditionally has been perceived as a matter of time- and space-bound events, modern digital platforms have expanded its potential, making it more accessible to a wider audience, including non-specialists, as well as to asynchronous application. The task of engaging in academic co-creation, however, often requires significant training and documentation, particularly for complex tasks, such as those involved in humanities research. The commitment of participants, whether they are simply contributing on a casual basis or engaging in more sustained research efforts, plays an important role in the success of such efforts. The concept of "citizen researchers" highlights how such crowdsourcing efforts can bring in the general public, making scholarly endeavours more inclusive.[41]

Moreover, the balance between machine learning and human effort is another significant consideration: how much human input is truly enough to complement the capabilities of algorithms, and how can co-creation of data adapt to meet the demands of complex, evolving research questions? Finally, ethics, paid labour, and derived value must be carefully considered, as participation in these initiatives often offers emotional rewards and meaningful engagement, rather than the highest forms of academic recognition such as peer-reviewed publications.[42] Sustainable collaborative research projects in medieval studies will probably look less like mass collaboration of untrained participants and more like federated networks of small, highly skilled contributors, coordinated through shared infrastructures and open standards. The thoughtful design of such data includes carefully documented metadata that makes all of the transcription choices and specificities visible.

The Sustainability and Mutualization of Knowledge in Medieval Studies

Some three decades ago, Pierre Lévy introduced the notion of *collective intelligence*, which he described as an emancipatory, decentralized process of shared knowledge creation through human collaboration in online spaces—ultimately leading to the "mutual recognition and enrichment of people" (*la reconnaissance et enrichissement des personnes*).[43] His argument suggested that digital spaces could serve as opportunities for open

41 Heinisch et al., "Citizen Humanities."

42 Andersdotter and Nauwerck, "Secretaries at Work."

43 Lévy, *L'intelligence collective*, 29.

exchange, participatory knowledge-making, and collective growth. In some cases, such as Wikipedia, open-source software communities, or HTR platforms, this vision is partially realized today: users access and contribute to infrastructure and bases of information, collaboratively contributing to shaping knowledge systems that evolve over time.

A body of literature in media ecology, critical infrastructure, and critical AI studies has since raised questions about Lévy's vision and the governance of online platforms and the unequal power dynamics embedded in them. It has been pointed out that the mechanisms through which knowledge is produced, validated, and disseminated can be opaque and controlled by a narrow set of actors—corporate or state entities and technocratic elites.[44] Despite the frequent use of metaphors of hospitality to describe digital spaces—using terms such as access, hosting, homepage, privacy, or members—to soften unequal participation of different users, online spaces are critiqued instead for being structurally exclusionary.[45] After all, platform-based infrastructures do not simply facilitate the creation and dissemination of scholarship—they actively reshape how knowledge is produced, distributed, and valued. They also raise difficult questions: who participates, and under what conditions? What forms of labour remain unacknowledged, and what ethical stakes are attached to the creation and sharing of data?

The tension between Lévy's ideal of participation and mutual enrichment and the actual infrastructures underpinning contemporary digital knowledge becomes more pronounced in the context of infrastructures which operate using AI. Today's AI development is largely concentrated in the commercial technical sector and wealthy institutions that possess the human and computational infrastructure required to support such systems. Whereas we have demonstrated, in Chapter 3, some pathbreaking applications of AI-based transcription analyzed by computational techniques with potential for reshaping how we view scribal culture, there are potential costs to the collective use of expert training data for AI. Cultural expertise, it could be argued, is transformed into an extractable intellectual labour, with uncertain benefits for the communities that provide it. That is to say, critical perspectives such as data sovereignty and platform governance are also real issues in the digital humanities and the academic commons, where the potential reproduction of old inequalities and the risk of foreclosing

44 Crawford, *Atlas of AI*.

45 Casilli, "'Posthumani nihil a me alienum puto.'"

on modes of equitable collective inquiry are real.[46] If the ideals of collective intelligence are to remain relevant, they must be reconceptualized—not only in terms of technologically advanced infrastructure, but also with a redoubled focus on equity, governance, and access—acknowledging the structural realities that mediate who gets to participate in building, and deriving benefit from, collective knowledge.

Sustainability of Platforms for Knowledge Creation

As academic co-creation gains traction in pre-modern studies, questions of sustainability and scalability become more important. While a growing number of platforms such as FromThePage, Zooniverse, PyBossa, Scripto for Omeka, MicroPasts, Transkribus, and eScriptorium offer infrastructure for collaborative knowledge creation through tasks such as transcription or annotation, their long-term viability depends on consistent maintenance, institutional support and funding, and a sustained user base that, in some cases, underwrites the maintenance of such platforms through membership fees, rather than through public grants. Indeed, the sustainability of these platforms, and more generally any digital research project, cannot be taken for granted. Regular server maintenance, security, and user support all require ongoing investment and attention.

A lively and diverse global ecosystem of infrastructure supporting transcription and automated text recognition (ATR) exists today. These infrastructures differ significantly in their organization, governance structures, and adoption of the software-as-service model, with important implications for how data, rights, and access are managed across national and institutional boundaries. Software has an economy, whether its financial models are obvious to us or not. Transcription platforms such as FromThePage operate on a freemium model—where a basic version is offered for free, but users must pay to access other features, platform support or expanded usage. FromThePage has evolved from a specific digital project into a widely used subscription-based service maintained by developers with long-term experience working with the humanities and the GLAM sector. Transkribus, also initially a grant funded project, transitioned to a cooperative model with a member base made up of large GLAM sector institutions, libraries, and private members. It also offers limited free access to some elements of the basic software product while charging for full use, with profits being

46 Crawford, *Atlas of AI.*

reinvested into research capacity and infrastructure development.[47] Calfa, by contrast, functions as a nonprofit focused on AI and natural language processing (NLP) tools for Oriental languages, linking researchers with support from cultural and philanthropic institutions. Finally, eScriptorium, also originally developed as a part of a funded project, is an open-source ATR architecture oriented toward linguistic and scriptural inclusivity, providing its code base openly to the community of researchers, rather than a hosted application delivered over the internet.

These differences in the economy of software have important implications for the humanities and for geographic locations or institutions where infrastructural support is lacking. Moreover, while open-source frameworks regularly champion accessibility and transparency, in reality, they demand sustained investments of time, expertise, and funding to be deployed, localized, and maintained. Despite being called "open" or "FAIR" such frameworks can remain out of reach for isolated researchers or under-resourced academic communities, and invisibly so to those who promulgate their usage. EScriptorium, for instance, does not provide a centralized platform, but distributes its code and documentation, to promote access and collaboration.[48] Any potential user must have access to research computing resources and the technical know-how to maintain the software. This software's dependence on institutional resources creates a structural barrier: well-funded labs or universities can realistically host and sustain such tools, leaving smaller institutions and independent scholars at a disadvantage when they wish to participate in the broader conversations shaping the future of digital medieval studies, either experimenting with new workflows, or developing and training models tailored to their sources, as we did during the Paris Bible correct-a-thon. Whereas other initiatives rely on a paid or freemium software-as-service model, for many scholars the main barrier to eScriptorium at present remains access and scalability, requiring significant institutional or consortial resources for setup and maintenance.

In sum, the differences between these transcription infrastructures underscore how national funding regimes, institutional commitments, and technical philosophies shape not only the functionality and availability of tools, but also who can effectively use them, which languages and scripts are prioritized, and how scholarly labour is distributed and recognized. It is

47 Terras et al., "The Artificial Intelligence Cooperative."

48 Eleftheriadi, "Online Tools for Handwritten Text Recognition"; for a trenchant critique of the FAIR principles in a global context, see Rojas Castro, "Los principios FAIR."

easy to see how there are different forms of infrastructural inequality and exclusion inherent to each of these platforms: economic, technical, institutional, geographical or geopolitical. More generally, these issues raise important questions about the future of participatory and independent uses of AI in scholarly contexts, pushing the discussion beyond concerns of specialized knowledge and training toward broader questions of accessibility, equity, and the infrastructures that will shape medieval studies in the years to come.

Such differences in organizational models exhibit not only varied degrees of access, but also the inherent fragility of academic infrastructures that depend on external support—whether public, philanthropic, institutional, or commercial. As the broader landscape of digital infrastructure evolves, and does so especially within the larger context of a rapidly changing landscape of AI, these projects remain highly exposed to shifts in political will, funding priorities, or technological standards. While some platforms benefit from sustained investment or integration into stable research ecosystems, many others—even those rooted in long-running digital humanities communities—rely on more precarious frameworks: short-term grants or partnerships, volunteer labour, or the dedication of small, overstretched teams. Such vulnerability was starkly illustrated among federally funded digital scholarship initiatives, by the April 2025 decision of the US Department of Government Efficiency (DOGE) to cancel most grants administered by the National Endowment for the Humanities in the United States, abruptly jeopardizing the research labour underpinning them. One such project known to the North American medievalist community, is the *Middle English Text Series* (METS), which had been awarded a three-year grant to produce TEI-encoded editions of medieval texts. It was forced to seek emergency funding from its scholarly community. In a widely circulated appeal on the Digital Medievalist listserv, its directors called on supporters to sustain the project's mission, emphasizing that its future "depends on those who believe in the value of collaborative, open-access scholarship and in the importance of historical engagement and understanding."[49]

More broadly, the sustainability of such infrastructures is further complicated by the fact that academic communities and national research cultures operate with markedly different value systems of research support—some emphasizing open access, public investment and the concomitant longevity of research cooperation, others depending more heavily on commercial or institutional buy-in, or hybrid models. These divergences complicate

49 Hahn and Siebach-Larsen, email sent to the *Digital Medievalist* listserv, April 14, 2025.

transnational collaboration, particularly when technical platforms and human labour span national borders, yet lack shared models of recognition or support. For an interdisciplinary and geographically dispersed field like medieval studies, it is unclear what entity might be able to step in to fill the infrastructural gap. It seems more likely that initiatives are to remain embedded in local or national research ecosystems, each ushering forward different expectations of openness, labour, and maintenance. The result is unfortunately a patchwork of capacity and access—rich with scholarly possibility, yet persistently uneven and perennially vulnerable.

The Rise of Mutualization of Data in the Humanities

In the previous section, we discussed the different social, technical, and commercial models that exist for platforms that enable transcription and, more specifically, ATR systems. Barriers for entry exist in each of them, either through the cost of platform-as-service or through the maintenance and technical and infrastructural development of such platforms. Technical and human forms of infrastructure, however, are not the only requirements for working in the computational humanities; data relevant to the sources and research questions at hand are also fundamental. Applications of AI such as HTR or computer vision require a significant amount of data to be trained on. As we have mentioned before, the creation and sharing of datasets that work for large categories of manuscripts for which there are many digitized copies—in specific scripts, for certain chronological periods or document categories—that can be used as a baseline for further training and adaptation to specific case studies is certainly one way forward that the community of medievalists interested in machine learning have taken.[50] While the Transkribus and eScriptorium user communities have created a significant amount of data for Latin, French, and other European languages, and world languages such as Arabic, for medievalists interested in understudied research languages creating ground truth for HTR—from scratch or through correcting HTR outputs—is still onerous.

Many project pipelines using HTR continue to depend on manual labour, especially in the early stages of implementation when high-quality ground truth data must be produced from scratch or when existing segmentation and transcription models require careful adaptation for project-specific purposes. Indeed, the prospect of harnessing academic co-creation to generate this foundational data carries a compelling appeal: it offers the potential to

50 Aguilar and Jolivet, "Handwritten Text Recognition."

distribute the labour-intensive tasks of annotation and transcription across a wider network of participants, in the hope of accelerating progress while tapping into the collective attention, expertise, and goodwill of an interested group of specialists. Yet, in fields like medieval studies, where palaeographic variation and idiosyncratic scribal practices are the norm, the challenges of coordinating such a collective effort are substantial. Especially with transcription, carrying it out according to highly specialized diplomatic conventions—or correcting preliminary HTR output with a trained eye—can be daunting, particularly when medievalists are unfamiliar with the exacting standards required for machine-readable transcription. Transcription is not an exception in computational medieval studies, however, as there are many other tasks that would face similar challenges: annotating marginalia, aligning multilingual versions of texts, or generating datasets for the recognition of visual features in illumination.

In this section, we turn to alternative modes of sourcing and sharing data within the humanities. While institutional repositories and their commercial counterparts have long supported the archiving of research papers—and, to a limited extent, datasets—such platforms focus on showcasing finished published products and can suffer from limited visibility and discoverability, diminishing their utility for collaborative or interdisciplinary research. Furthermore, scalability and language present technical and political challenges to infrastructure. For example, France has developed the IR* HumaNum, a national research infrastructure that supports the social sciences and humanities (SSH) through digital services for research data. In practice, it provides digital infrastructure that enables SSH communities to develop, conduct, and preserve research programmes, data and tools, within an open science and data-sharing framework. But this structure, developed by the Ministère de l'enseignement supérieur et de la recherche, and operated by CNRS, the Campus Condorcet, and Aix-Marseille Université, is limited to French researchers and institutions. International projects are hosted only when the leading team is French. Another model of data and research paper sharing relevant to medieval studies has been proposed via the Open Science Foundation framework. In fact, there is an archive for papers in medieval studies based on OSF architecture—Bodoarxiv—although its adoption by researchers, particularly those in digital medieval studies, has been limited.[51]

In recent years, global researchers seeking to share or locate archival versions of datasets relevant to the humanities have turned to Zenodo, a

51 See BodoArXiv, "Welcome / About BodoArXiv," and BodoArXiv, *BodoArXiv: Open Repository for Medieval Studies*.

platform that has become central to open data practices across disciplines. Based in Switzerland and maintained by CERN (the European Organization for Nuclear Research), Zenodo has been supported through major European Union research frameworks, including Horizon 2020 and OpenAIRE-Nexus. It represents a model of publicly funded pan-European infrastructure that is globally accessible, free, and rooted in principles of openness and long-term access. Importantly, Zenodo not only hosts datasets but also provides useful services for data-intensive fields, such as the digital humanities. It assigns persistent digital object identifiers (DOIs), supports versioning of records, and offers integration with widely used platforms like GitHub, all features that ensure that datasets remain citable, traceable, and usable over time.

At the time we are writing, an example of the desire to share or "mutualize" project data within a user community is the HTR-United initiative. HTR mutualization refers to collective efforts to centralize and share ground truth datasets—high-quality, manually transcribed texts aligned with digitized images of manuscripts—used to train models for reading historical handwriting. HTR-United is not a repository in the conventional sense, like Zenodo or Bodoarxiv, but rather a metadata commons: a structured index of datasets with documentation on standards of creation.[52] Its goal is to standardize dataset descriptions using a shared schema, to provide guidelines, and to promote tools for quality control. Unlike repositories that focus primarily on reproducibility, HTR-United emphasizes the visibility of bespoke datasets and their potential reuse, ideally with editorial statements, lists of Unicode codepoints used, and transcription protocols that situate them in humanistic context. It draws attention to ground truth data available online for reuse for the purpose of model training and fine-tuning, and can be said to address the problem of reinventing the wheel, but its utility ultimately also depends on the alignment of goals, methods, and values across users. Beyond this model of cataloguing datasets, some researchers have begun depositing their data in Hugging Face, where projects such as the CATMuS medieval dataset offer string- and line-level transcriptions labelled with location and period. Compared to the DOI-based permanence of Zenodo or the versioning affordances of GitHub, Hugging Face offers an interactive and visible space for datasets ready for machine learning, making it well suited to model training and wider community engagement. At the same time, it lacks some of the long-term guarantees of persistence and citability that Zenodo provides, and its strong orientation toward machine learning may

52 Chagué and Clérice, "I'm Here to Fight for Ground Truth."

not always align with slower, humanistic priorities of documentation and contextualization.

Again, the infrastructure we have described here for mutualization of data reflects a shift in the textual computational humanities at the time we are writing: rather than aiming for universal models, it acknowledges the enduring value of finely crafted data and recognizes that openly available training corpora themselves are an important part of the scholarly process.[53] In recent years, we have witnessed an important shift from the collective creation of data to designing the means for data pooling, a subtle but meaningful shift in how collaboration is conceived. These infrastructures for research in 2025 may not exist in a few years, having been supplanted by others. Yet, they introduce new tensions. We have argued in Chapter 2 that ground truth, especially for medieval texts, is not as universal as it is purported to be. Within the same script or language group, scholars may model different textual phenomena, apply different normalization strategies, and ask different research questions. This fact means that ground truth must be continually reassessed and models adapted for the context of new projects.

Furthermore, there are salient critiques of data mutualization in the computational humanities that highlight how infrastructural disparity is not merely a logistical inconvenience, but rather a structural issue that shapes who is able to produce knowledge, how that knowledge circulates, and whose perspectives and data are ultimately centred—or excluded. While discourses around mutualization are often framed in the language of collaboration, openness, and equity, they fall short of genuine inclusivity.[54] Mutualization presumes that all contributors have comparable access to time, funding, institutional infrastructure, and technical expertise, and, perhaps more importantly, that their academic institutions reward (or require) their contributions to open access data sharing. In reality, disparities in digital access, administrative support, training, or reward systems for open access publication frequently make it complex for less-resourced individuals, institutions without relevant programmes, and scholars working in non-Western environments to produce work that meets the expectations of mutualization. These exclusions are not incidental—they shape the very contours of the field's development and perpetuate existing hierarchies of academic participation.[55]

53 Chagué and Clérice, "HTR-United."

54 Piron, "Postcolonial Open Access."

55 Risam, *New Digital Worlds*.

Indeed, open access and mutualization are not a one-size-fits-all approach to knowledge creation.

Such disconnects are evident in computational manuscript studies within Western scholarly environments, where the aspiration for open access to training data often collides with the legal, material, and institutional constraints that shape access to, and distribution of, primary sources. In our work on the Paris Bible Project—focused on a genre, examples of which are found in a wide range of global collections—we have sought to include manuscripts from diverse locales in order to challenge what we have termed collection bias. However, gaining access to these materials, particularly outside well-resourced European or North American institutions, has often been difficult, let alone the acquisition of digital images of undigitized manuscripts suitable for analysis. With some libraries and archives in Europe and North America, we encountered a willingness to permit informal photography of medieval manuscripts for research purposes—including for use in HTR model training—but under the condition that such images be used solely for private research and not be disseminated publicly. As a result, we hold numerous pages of carefully proofread ground truth data that cannot be shared alongside their corresponding images, the preferred method of sharing data according to HTR-United and the necessary condition for ground truth in Hugging Face. These datasets were crucial for the analysis presented in Chapter 3, yet we would likely face resistance in parts of the computational humanities community were we to publish them without their corresponding image data. These limitations underscore the frictions between the ideal of mutualized data, the notion of reproducibility that sits behind it, and the practical realities of working with fragile or institutionally restricted archival material.

Such examples underscore the considerable labour involved in producing a wide range of datasets stemming from computational humanities research—whether they be TEI-encoded editions, geospatial gazetteers, or annotated image sets—intended for reuse under explicit licensing frameworks. While well-resourced institutions may benefit from dedicated teams for metadata curation, digital preservation, and infrastructural support, and may have funding requirements to publish research data openly, these conditions are far from universal. Although they may be published openly, humanities datasets are not neutral, fungible objects: they are deeply embedded in the interpretive labour, scholarly judgement, and material and legal constraints of the environments in which they are created. While metadata helps make these datasets more discoverable and legible to others, it cannot fully convey the subtlety of the interpretive processes underlying their

construction. In practice, many humanities datasets—particularly those emerging from under-resourced or short-term projects—can be incomplete or undeveloped in nature, making them difficult to share or reuse in other contexts, even amongst seasoned digital humanists. Carefully curated and documented corpora, such as those indexed by HTR-United, must be reassessed and often restructured when taken up for new research questions, or grounded in alternative scribal conventions. The interpretive specificity of such humanities data can resist the ideals of seamless transfer and reproducibility, promulgated by more standardized scientific data creation processes. Moreover, the work required to align, translate, or reshape inherited data to new scholarly contexts is often underestimated. It is not merely a matter of format conversion, but of grappling with embedded assumptions, historical contexts, and methodological difference. These challenges have led some scholars to question whether the promise of mutualized data reuse in the humanities is truly worth the effort.[56] The labour needed to repurpose datasets often approaches that of creating them anew, raising difficult questions about how value and credit for such work are determined. Ultimately, while the mutualization of data offers real opportunities for scholarly exchange, it cannot be divorced from the structural conditions in which data is increasingly being produced, shared, and appropriated.

Of course, multidisciplinary research has its share of pitfalls and cautionary tales—imbalances of funding, differing reward structures and timelines for research—but we believe that there is untapped potential in developing collaborative research models that expand current systems of reward. This structural fragility is compounded by disciplinary and institutional disparities. Students and scholars at institutions without access to digital humanities training, including those with rich archival collections, may not develop the computational expertise necessary to contribute to or lead data-driven projects. This paradox produces uneven development in the field, where institutions with rich archives may unfortunately experience technical exclusion, and other institutions with smaller archives may excel in technical know-how, overemphasizing the artifacts they have at their disposal. Moreover, expanding access to collaborative networks, particularly for scholars working outside major research universities, has been shown in other fields to significantly benefit researchers' careers.[57] Such collaborative work is going on, as can be gleaned from publication in digital and computational

56 Camps et al., "Data Diversity."

57 Li et al., "Early Coauthorship."

humanities venues, but if (or when) more professional societies aligned with the complex field of medieval studies were to encourage collaborative work with non-humanities fields like Statistics, Computer Science, Biology, and Environmental Studies, leading to co-publication, the field may potentially experience shifts in research impact similar to those observed in the sciences and social sciences, leading to a renewed visibility and influence.[58] As the field continues to integrate digital methodologies and data-driven research with a critical eye to the specificities of humanities archives, the importance of co-authorship in medieval studies may indeed increase.

Credit taxonomies like TaDiRAH[59] and recent thinking about crediting data provenance[60] attempts to model ways that credit for the labour of collaborative work in computational humanities can be acknowledged. Yet, the logic of artificial intelligence—where data are aggregated, generalized, and reused—complicates traditional models of (co-)authorship and citation. If ground truth data is created by one group, refined by another, and used to train models elsewhere, how do we assign scholarly credit in ways that acknowledge layered intellectual contributions, rather than contributing to them being obscured? If mutualizing data leads to its extraction and aggregation, losing the trace of such scholarly labour, is this situation one that humanists want to promote? Openness is, indeed, not an easy fix.[61] These various challenges, along with the rise of synthetic data approaches for machine learning,[62] underscore the need for more context-sensitive and ethically grounded infrastructures and transparent discussions in the field about research workflows.[63]

---

58 In a North American context, many professional associations have disseminated guidelines for the evaluation of innovative forms of scholarship, particularly regarding digital humanities research: a few examples of this profession-created guidance include the Canadian Federation of the Humanities and Social Sciences (2002), American Historical Association (2015), College Art Association (2015), the American Council of Learned Societies (2021), and a recently updated report from the Modern Language Association Committee on Information Technology (2024). Digital scholarship presents specific challenges and opportunities for the humanities, as do interdisciplinary partnerships, and yet such statements on evaluating the value of the interdisciplinary are not as common within the humanities.

59 Borek et al., "Information Organization and Access."

60 Romein et al., "Exploring Data Provenance."

61 Klein et al., "Provocations from the Humanities," 10–11.

62 Offert and Bell, "Generative Digital Humanities."

63 An example, although not specifically from medieval studies, of collective reflection on workflows can be found in Baillot et al., eds., "Workflows."

Conclusion

We conclude this chapter by advancing the proposition that the future of collaborative medieval studies should not be oriented toward scale or speed alone, but rather toward the development of community-sustaining infrastructures that support ethical and inclusive collaborative research practices, framed by a postcolonial approach to information. They must be designed to mutualize interpretive work, acknowledge diverse forms of expertise, and support a wide variety of epistemologies. And if we are to adopt the notion of the commons as a guiding model—for shared resources, datasets, methods, and publications—what forms of sharing are most likely to foster an equitable, participatory future? While the creation of high-quality metadata and ground truth datasets for HTR represents a meaningful beginning, it cannot be the endpoint. What might come next is an interconnected set of community-governed environments, annotation spaces, and pedagogical frameworks that allow scholars at different levels of technical proficiency to contribute meaningfully to the collective endeavour. If we take the conclusions of the SciSci literature seriously, a vibrant commons inclusive of many different kinds of expertise will be required. In this vision, the commons has to be imagined not merely as an index of data, but a space of encounter, negotiation, and ongoing ethical engagement, and importantly, evolving as both technology and digital literacies evolve.

Our research on thirteenth- and fourteenth-century Latin bibles surfaces general ethical questions of all forms of research in the age of AI and automation that we believe should be a matter of debate in our professional circles. Creating reusable data in research is a trend in contemporary academic research that one finds even outside of largely computational projects and we do not expect for such a trend to wane. It is important, in our eyes, that contemporary research approach the question of the diversity of research outputs with care, and in particular, not producing outputs that only have relevance in a highly specialized, quantitative, computational sphere.[64] Rather, more inclusive and integrated methods for incorporating and validating theories that arise from computational work need to be devised. In the case of medieval studies, validating them across different scales, archives and collections seems to be particularly salient, as well as in combination with other (admittedly slower) forms of codicological analysis that the face-to-face encounter with the physical codex privileges. Bringing together computational analysis with more traditional methods is also

64 Joyeux-Prunel, "Digital Humanities in the Era of Digital Reproducibility," 41.

a way to bridge the gap between already strongly computational scholars and those learning about digital environments and methods. For researchers whose expertise lies outside the digital domain, a future of collaborative medieval studies requires infrastructure that prioritizes a *commons of interpretation*—where scholarly annotations, hypotheses, and close readings can be treated as shareable and citable outputs. Such spaces would foster knowledge-making across interpretive communities, in which contributions by non-digital scholars are not merely accommodated, but actively incorporated. Joyeux-Prunel has argued that, in the humanities, "issues of corpus, method, and interpretation cannot be separated, rendering a procedural definition of reproducibility impractical," and she has proposed a post-computational framework for the extension of the FAIR guidelines.[65]

For reasons that we have provided above about the relative size of the medieval studies community and its global distribution, it is unlikely that medievalists will take on these issues of infrastructural design and implementation by themselves. Instead, cooperation between the community and other actors and advocacy will be essential to make sure that the specificities of the discipline are included in the larger conversations. One domain, however, in which medieval studies will need to double down concerns pedagogy and the training of new generations of medievalists. In this chapter we have outlined one example of a critical student-centred training opportunity designed to introduce a new generation of students to the challenges and the promises of AI-centred data creation. It is perhaps in the domain of pedagogy that commons-based approaches can help cultivate a new generation of scholars equipped to bridge the various gaps between traditional research practices and the computational humanities.[66] The development of a pedagogical commons—comprising modular, open-access materials for a variety of historical and linguistic domains—would facilitate the integration of computational methods into medievalist training. Modelled on other commons such as the *Programming Historian* or the *Digital Orientalist*, medieval-specific contributions would emphasize interpretive engagement with issues of manuscript culture and medieval language. They would

65 Joyeux-Prunel, "Digital Humanities in the Era of Digital Reproducibility," 25.

66 There are venues for the discussion of pedagogy in the context of medieval studies, and ones with an explicitly digital orientation, such as the *Middle Ages for Educators* where some attention is placed on the acquisition of digital and computational skills, but where most of the content engages with how to teach the Middle Ages with digital resources that have already been made. See Middle Ages for Educators, *Middle Ages for Educators*.

provide a venue for methodological discussions relevant for existing scholarly communication venues for the debate about medieval cultural objects in the post-digital era.[67] Engaging junior and senior medievalists alike in the details of reconstruction of dispersed codices or in the collaborative annotation of under-documented texts, such a commons would acknowledge the computationality of the field and the mass of digitized objects we study, while facilitating a critical discussion of how new methods result in new knowledge and paths of interpretation.

It is also valuable to reflect on the larger picture of the kinds of training and support needed by new generations of medievalists to be able to take full advantage of future-forward infrastructures. Obviously, initiatives such as the Digital Medieval Studies Institute (DMSI) that took place for the first time as a collaboration between the Medieval Academy of America (MAA) and NYU's campus in Washington, DC, and that has continued collocated with the MAA conference and the International Medieval Congress in Leeds, UK, is an excellent example.[68] Training opportunities, however, vary widely depending on national contexts. In North America, students can pursue certificates alongside their main degrees (BA, MA, PhD), but there are few stand-alone degrees in digital humanities. In Europe, by contrast, there are numerous specialized master's programmes—for instance at the École des chartes, King's College London, or within several German and Scandinavian universities—as well as programmes focusing on the history of the book or manuscript studies that often integrate digital components, such as the host programme for our abovementioned Paris Bible Correct-a-thon. National systems that have invested in digital infrastructures also support a wealth of summer schools, organized by projects (e.g., Biblissima, ERC projects) or by universities in places such as Besançon, Oxford, Leipzig, Madrid, and Venice, as well as long-running programmes like the Digital Humanities Summer Institute (DHSI) in Canada. These initiatives demonstrate that there are many possible pathways into digital training, though future medievalists and early career researchers must often integrate themselves into broader, more technical networks that extend beyond medieval studies.

Disciplinary training alongside digital methods also impacts outputs produced by young scholars. According to the SciSci literature, we know that young scholars who work and publish with more experienced and established researchers, both in and outside their field, tend to develop more

67 Some of these journals include *Digital Philology*, *postmedieval*, *Fragmentarium*.

68 Medieval Academy of America, "Digital Medieval Studies Institute (DMSI) 2025."

impactful and disruptive work. While there is no magic formula for this relationship to develop, we believe that academic institutions and scholarly infrastructures need to create the context for serendipitous collaborations to develop. This context can be nurtured in multiple ways: beyond conferences and congresses, summer schools are a natural place where researchers at different career levels and from different contexts can meet and develop common scholarly interests. Mentoring initiatives within professional organizations are another way to create these relationships. More importantly perhaps, academic institutions ideally would dedicate some research time to experimental humanities in which they encourage their scholars and students to pursue tangential and collaborative projects, supported by the time, place and the funding, projects that do not result in canonical scholarly output. While there has been some advocacy work done on validating digital scholarship for the tenure-track faculty member, less attention has been paid to the young medievalist. Traditional scholarly outputs—PhD thesis, single-authored publications—are still the expected norm, particularly in the North American context. There, professional organizations and the entire scholarly worlds around medieval studies could step up and write guidelines for digital and computational work in the humanities, similar to the MLA guidelines about how to model future professional work.[69] Being explicit about the criteria for the formal evaluation of digital scholarship has the informal benefit of defining professional ideals for future scholars in training.

Professional societies and conferences with a pre-modern focus such as the Renaissance Society of America (RSA), the Medieval Academy of America (MAA), the "Co-operative for the Advancement of Research through a Medieval European Network" (CARMEN), International Medieval Congress (IMC), International Congress on Medieval Studies (ICMS) and others can lead this effort, dedicating sessions to collaborations between digital/computational humanities and non-humanities fields. Awards and recognition structures—such as prizes for collaborative projects that make data openly available or develop pedagogical toolkits—can shift incentive structures toward shared scholarly labour. Hackathons and collaborative data sprints, built around real project datasets some of which have been mentioned in this chapter, could be hosted alongside traditional paper panels, encouraging emerging scholars to engage hands-on with digital methods and to contribute to community-oriented infrastructure like the commons-based model we discussed above. Existing programmes, such as the digital humanities master

69 Modern Language Association, "Guidelines."

at the École nationale des chartes-PSL, already organize such hackathons. In a call for papers sent in September 2025, they solicit projects focusing on AI and data science applications in the social sciences and humanities. Each challenge will be based on a data corpus (with preference for open-access sources) and designed around tasks feasible within one week, such as classification, annotation, information extraction, visualization, or modelling.

Beyond recognizing technical sophistication, however, these societies could tap into public humanities approaches to help translate interdisciplinary and computational work into forms that are legible and valuable to broader audiences, including to the profession itself. As the number of co-authored and data-rich publications rises, the profession can promote formats that integrate visualization and digital narrative in ways that invite wider participation. Conference platforms should encourage scholars to present digital findings in conjunction with critical reflections on method and pedagogy. Professional organizations and the leadership within them can help ensure that new methods do not harden disciplinary boundaries, but instead expand the interpretive commons of the field. At the same time, embracing mixed publication pipelines—wherein graduate students and early career researchers could have code notebooks, datasets, and data papers peer-reviewed—would align scholarly communication with the realities of collaborative work. It has been shown that humanities research based on freely available and openly licensed datasets is more likely to be cited, leading to increased reproducibility rates and greater public confidence in the research.[70] If we extend this scenario beyond datasets to the intermediate materials of computational humanities, such as computational notebooks or HTR ground truth, we may have an environment in which more medievalist confidence in the research can be built. As the landscape of research in the humanities evolves—driven by computational tools, interdisciplinary collaboration, and open data practices—professional organizations, scholarly conferences, and mainstream journals (many of which we studied in this chapter) must take a more active role in shaping the conditions for inclusive, future-oriented scholarship. These organizations are uniquely positioned to support innovative practices and the ethical and pedagogical frameworks that sustain its presence in the contemporary university.

70 Colavizza et al., "The Citation Advantage," cited in Van Erp et al., "The Future of Digital Humanities Research."

CONCLUSION

IN THIS BOOK, we have argued for a new engagement with medieval manuscripts, one that discusses the promises and the limitations of computational approaches to textual variance, scribal attribution, and the materiality of the handwritten book. We used the large corpus of Paris bibles, a form of the medieval Latin Bible of the thirteenth and fourteenth century, as a case study to explore how scribal contribution to a manuscript can be captured by automated transcription, and to suggest that new and different questions can be advanced using digital and computational methodologies. We argue for innovative ways of transcribing and analyzing manuscript data and advocate for more collaborative partnerships in the humanities, whether through data co-creation, data sharing, or multidisciplinary research collaboration. Our project illustrates Bobley's four major transformations in medieval studies and in the humanities more broadly brought about by digital technologies: it sheds light on access to digitized medieval manuscripts in global archives, brings new computational methods to bear in the analysis of medieval sources, reconfigures interactions between scholars, libraries, and archives, and suggests generative pathways for multidisciplinary collaboration. At its core, the Paris Bible Project, and this book reflecting on it, argue that neither humans nor machines can interpret the vast medieval record alone, but rather a collective effort, accompanied by significant critical humility, is required.

In Chapter 1, we presented the form of the Paris bible, reflecting on the challenges of building a research corpus and highlighting limitations in terms of availability, quantity, and quality of digitized resources needed to carry out computational analysis. We approached two ideas from contemporary digital culture and their applicability in the study of the premodern humanities: the concepts of big data created from medieval sources and the idea of the pattern searching within those sources. Contemporary technologies such as HTR provide access to a significant amount of text that can be created in direct dialogue with book history. By introducing the idea of computational analysis through an experiment on the biblical prologues, we highlighted the potential of such methods to understand patterns in archival objects and explored the implications of working with digitized collections.

DOI: 10.17302/mfsk158610

We questioned how our access to digital surrogates not only reshapes what we can see in manuscripts, but also what new forms of interpretation we can forge from these data-rich collections.

For many researchers today, access to the physical object is no longer the fundamental condition of research on medieval manuscripts, but rather serves as a supplement to work done with digital copies. As we demonstrate throughout this book, this is certainly the case of the Paris Bible Project, where the dispersion of the genre makes it next to impossible for a group of researchers to access the corpus first-hand. Seen from this perspective, we have attempted to explore how digitization of our sources not only enables—but also reshapes—possibilities of scholarly interpretation. We concluded that even though we do not have access to everything that was produced in the Middle Ages—because so much has been lost, and because the many small cultural institutions holding manuscripts will not digitize them—what we already have is too much to process. New ways of managing and collectively studying digitized manuscripts in the computational age need to be developed and taught.

In Chapter 2, we made the case for human-curated, domain-specific datasets in computational humanities. Rather than pursuing unreasonable ideals of completeness or one-size-fits-all solutions, we argued that more bespoke HTR models grounded in material and digital philology allow for forms of ethical, scholarly analysis. Looking at practices in diplomatic and normalized transcriptions in the scholarly literature, we asked how we might approach producing data about scribal practices using recent technological advances to automate material philology with both caution and pragmatism. Our approach to transcription foregrounds scribal practices as complex, layered phenomena, requiring careful modelling of handwriting and textual features at the character level. We argued that machine learning has particular promise for the study of transcribed medieval texts and that the creation of ground truth datasets—far from being a neutral pre-processing step—is a deeply interpretive act that shapes the potential outcomes of scholarly research. Questioning the idea of the scribal profile that has existed for some time, we proposed an approach to modelling scribal behaviour at the intersection of quantitative codicological methods and digital philology. Such modelling requires defending the features that are most relevant to the research questions we have, rather than exhaustively describing these book-objects. This chapter also highlighted implications for a future of manuscript studies in which scholars will most likely complement traditional knowledge in language skills, palaeography, and codicology with competencies in data curation, model evaluation, and computational thinking.

In Chapter 3, we designed four experiments aimed at putting the arguments developed in the previous chapters into practice. Using a dozen Paris bible manuscripts, our case studies examine scribal attribution, while underscoring both the potential and limitations of interpretation using computational methods. By demonstrating how scribes left a mark on the texts they copied, our analyses reveal how computational modelling sometimes confirms the scholarly record, but also offers critical refinements that go beyond long-standing arguments in the field. From comparing the hands in a single codex with what the trained palaeographer can see with the trained, but naked eye (Experiment 1) to testing attributions based on colophons or cataloguers' claims (Experiment 2), we demonstrated how computational models corroborate, refine, or challenge human expertise. Experiments 3 and 4 took the analysis a few steps further, by analyzing previous scholarly claims of artistic groups and circles of manuscript production, nuancing the way that we frame and analyze such production. We also introduced the idea of a companion scribal style for describing manuscripts exhibiting statistical affinities with each other. We argue, therefore, for the need for hybrid approaches in manuscript studies, in which statistical techniques support—but do not supplant—narrative modes of argumentation. The experiments we carried out highlighted that while computational methods can offer significant insight, they also foreground a tension in medieval studies: the more we lean into algorithmic abstraction, the more we can run the risk of distancing ourselves from the materiality of such objects. We argue, therefore, for the need for more widespread, hybrid approaches in manuscript studies, which blend statistical techniques and philological, palaeographical, and codicological expertise.

Finally, in Chapter 4, we shifted perspectives to invite our reader to consider the role of collaboration in medieval studies in both the past and the future, and to the possibilities such collaboration might enable. In the first half of the chapter, we looked back to the last century, through the lens of the science of science and bibliometrics in publications in medieval studies. We highlighted not only how multi-authored research in medieval studies is slowly on the rise, but also how varied teams of researchers tend to have increased citation and impact over time. We also highlighted the benefits of medievalists working together, but also the many possible limitations of such endeavours. Then, we examined recent trends in data co-creation and ways medievalists are working together. Collaborative research in computational medieval studies can expose tensions between technological ambition and disciplinary practices. Moreover, computational infrastructures are not neutral: their design, distribution of labour, and mechanisms of credit shape what kinds of collaboration are possible and whose expertise counts.

While the idea of collectively studying manuscripts dispersed across institutions and continents using their digital avatars is appealing, such visions run up against the material realities of access and infrastructure. Thus, computational approaches can easily fall short of the global, cross-collection approaches desired by serious historical scholarship. The organization of computational humanities research, at least as it is currently constituted as we are writing, tends to be organized within national, institutional, or collection-specific frameworks. The prospect of a future in medieval studies shaped by diverse participation in large-scale knowledge production across boundaries is compelling, but it is important to recognize that data co-creation and mutualization—as a methodological framework and in practice—pose significant challenges to equitable participation in the field. Since infrastructure may inadvertently re-inscribe the very exclusion that collaborative scholarship hopes to overcome, the mutualization of infrastructures for collaborative work cannot only consist in the pooling, or indexing, of data, but needs to strive for a more radical transformation of disciplinary cooperation. Rather than celebrating scale for its own sake, we call for models of cooperation that render visible the interpretive and technical labour sustaining them, and for societies and journals in medieval and pre-modern studies to champion innovative forms of shared, cross-disciplinary work. Ultimately, we conclude that medieval studies should consider collaboration in our fields more seriously as an opportunity to create better, more sustainable research.

We would be remiss to conclude a book that featured medieval Latin bibles without also pointing toward the possibility and potential of future research concerning the specific genre under study. We believe such computational research has significant potential in studying the textual micro-variance long observed in thirteenth- and fourteenth-century bibles across the thousands of copies and across different communities, the presence and use of fragments of the *Vetus Latina* among the text of Vulgate bibles, as well as spatio-temporal patterns of variability of prologues across the wider corpus as suggested in Chapter 1. As we have claimed at the beginning of this book, a comprehensive history of the Paris bible has yet to be written, but such a history would necessarily encompass a wide range of subjects that computational methods can address: theme and variation in iconographic programmes; textual transmission and variance; detailed analysis of the prologues, and scribal contribution to them beyond the general indexing created by Stegmüller; exploration of marginalia and decoration; computational approaches to quires, codicology and miniaturization; as well as the material conditions of manuscript production and the networks of

circulation that shaped medieval book culture. Following the lead of generations of research in the domain of quantitative codicology, such an account would necessarily foreground the manuscript not only as a carrier of text, but as a crafted and designed object. Our suggestions here for future directions of study for the Paris bible corpus may just have resonance for medievalists working in other genres, languages, and periods. We believe that there are myriad promising directions for manuscript studies to expand to incorporate computational workflows including, but not limited to, complex layout segmentation to study the "cognitive layout" of manuscript design; image-text alignment and visual stylometry; or generative or agent-based approaches that help us model complex textual transmission. More generally, integrating decorative and iconographic analysis with textual study in a computational way could also illuminate the dynamic relationship between image and text, opening pathways to a fuller understanding of manuscripts, including how they were made, used, and transmitted. Such complex research requires not only cross-collection access, but also creativity and cross-disciplinary cooperation.

What is certain is that the digital and computational turns have transformed what counts as evidence, method, and participation in medieval studies. Not all scholars will engage in large-scale data modelling or transcription, yet all are implicated in the shifting infrastructures and epistemologies that shape our field. The expansion of computational approaches demands vigilance: while they offer new precision in tracking textual variance, scribal identity, and material form, they also reconfigure the humanities themselves. Following Tilton, Mimno, and Johnson, we advocate a reflexive and participatory stance that models new forms of interpretative work, but also helps shape the tools, environments, and pedagogies that mediate our research. As technological landscapes shift, we humanists (and medievalists) must be much more involved in critiquing and shaping a research future, insisting that research environments not only reflect our technical needs, but also the humanistic values and scholarly commitments that we hold dear. What is at stake is not only how we study the medieval past, but how we imagine and sustain the humanities to come.

Finally, as a way of concluding this book, we would like to reflect on the collaborative work that made it possible, and to situate it in relation to the questions raised in Chapter 4. Although the research project that gave rise to this book may be considered by the authors' respective institutions as a testament to innovation or a reflection of scholarly values worthy of praise in today's scholarly landscape, it is important to clarify that it did not emerge from an institutional mandate, strategic initiative, or external funding agenda.

Rather, it was driven by the intellectual curiosity, sustained commitment, and collaborative energy of its two co-authors and their occasional collaborators. The work represents, therefore, not the fulfilment of an institutional vision, but the product of a self-directed and self-organized scholarly endeavour. We recognize, however, that such autonomy—while often idealized in academia—is not evenly accessible and would not have been possible without the scholarly freedom that both authors were afforded by their respective institutions. The ability to pursue independent, long-term research outside the constraints of project-based funding cycles is tied to structural privilege: to those with secure academic and professional positions, supportive infrastructures, and the temporal flexibility to engage in exploratory, often interdisciplinary work. When the project started, the independence we enjoyed was possible only because both authors had stable professional positions. Free from the constraints of funding cycles, tenure pressures, or grant deliverables, we were able to take risks and pursue paths that might otherwise have been closed. While this example may match the conditions that the SciSci literature would consider ideal for innovation, in reality, the project's autonomy reflects conditions that are rare, and we acknowledge the broader systemic inequities that would exclude others from participating fully in this mode of scholarship.

BIBLIOGRAPHY

Manuscripts and Incunabula

Aargau Kantonsbibliothek MS MSWettf 11.
Abu Dhabi, Louvre Abu Dhabi, MS 2013.051.
Besançon, Bibliothèque municipale, MS 4.
Besançon, Bibliothèque municipale, MS 8.
Budapest, Eötvös Loránd Tudományegyetem Könyvtára, MS Cod. Lat. 18.
Cambridge, Corpus Christi College, MS 49.
Cambridge, MA, Harvard University, Houghton Library, MS Lat. 5.
Città del Vaticano, Biblioteca Apostolica Vaticana, MS Vat. lat. 36.
Cologny, Fondation Martin Bodmer, MS Cod. 28.
Dole, Médiathèque de l'hôtel Dieu, MS 5.
Girona, Arxiu Capitular de la Catedral, MS 52.Göttweig, Benediktinerstift, MS Cod. 116.
Lisboa, Biblioteca nacional de Portugal, MS IL 93.
London, British Library, Additional MS 78830.
London, Lambeth Palace, MS 1362.
London, Lambeth Palace, MS 1364.
Montréal, Université du Québec à Montréal, MS without shelfmark.
New Haven, Yale University, Beinecke Library, MS 321.
New Haven, Yale University, Beinecke Library, MS 387 (the "Ruskin Bible.")
New Haven, Yale University, Beinecke Library, MS 433.
New Haven, Yale University, Beinecke Library, MS 1100.
New Haven, Yale University, Beinecke Library, Zi +4243.
New Haven, Yale University, Beinecke Library, ZZi 56.
Palo Alto, Stanford University Libraries, MS 23.
Paris, Bibliothèque nationale de France, MS français 412.
Paris, Bibliothèque nationale de France, MS latin 40.
Paris, Bibliothèque nationale de France, MS latin 179.
Paris, Bibliothèque nationale de France, MS latin 211.
Paris, Bibliothèque nationale de France, MS latin 10421.
Paris, Bibliothèque nationale de France, MS latin 10426.
Paris, Bibliothèque nationale de France, MS latin 10428.
Paris, Bibliothèque nationale de France, MS latin 11935.
Paris, Bibliothèque nationale de France, MS latin 15477.
Paris, Bibliothèque nationale de France, MS Smith-Lesouëf 19.
Paris, Institut de France, Bibliothèque Mazarine, MS 6.
Paris, Les Enluminures, MS TM 844.
Paris, Les Enluminures, MS TM 1226.
Paris, Les Enluminures, MS TM 1327.
Paris, Sotheby's, June 27, 2024, "Livres et Manuscrits," Lot 1.
Philadelphia, Free Library of Philadelphia, MS Lewis E242 (the "Patou Bible").
Philadelphia, University of Pennsylvania, MS Codex 236.
Philadelphia, University of Pennsylvania, MS Codex 660.
Sarnen, Kollegiumsbibliothek, Stiftsarchiv Muri-Gries, MS Cod. membr. 16.
Schaffhausen, Ministerialbibliothek, MS Min. 6.
St. Gallen, Kantonsbibliothek, MS VadSlg 332.

DOI: 10.17302/mfsk158711

Secondary Works

Agostini, Caterina, and Carrie Beneš. "A Geospatial *La Sfera*: Navigating the Renaissance in the Mediterranean." In *GeoHumanities '21: Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities* (November 2021), 22–27, ACM, 2021. https://dl.acm.org/doi/pdf/10.1145/3486187.3490207.

Aguilar, Sergio Torres, and Vincent Jolivet. "Handwritten Text Recognition for Documentary Medieval Manuscripts." In "Historical Documents and Automatic Text Recognition." Special issue, *Journal of Data Mining and Digital Humanities* (December 22, 2023): 1–13. https://doi.org/10.46298/jdmdh.10484.

Andersdotter, Karolina, and Malin Nauwerck. "Secretaries at Work: Accessing Astrid Lindgren's Stenographed Manuscripts Through Expert Crowdsourcing." *The 6th Digital Humanities in the Nordic and Baltic Countries Conference* (DHNB 2022, Uppsala, March 15–18, 2022), 9–22. CEUR, 2022. https://ceur-ws.org/Vol-3232/paper01.pdf.

Andler, Daniel. *Intelligence artificielle, intelligence humaine: La double énigme*. Gallimard, 2023.

Andrist, Patrick, Tobias Englmeier, and Saskia Dirkse. "New Digital Strategies for Creating and Comparing the Content Structure of Biblical Manuscripts." In "On the Way to the Future of Digital Manuscript Studies." Special issue, *Journal of Data Mining and Digital Humanities* (October 11, 2023): 1–19. https://doi.org/10.46298/jdmdh.10981.

Arnold, John Hugh, and Catherine Goodson. "Resounding Community: The History and Meaning of Medieval Church Bells." *Viator* 43, no. 1 (2012): 1–30.

Ashley, Kathleen M., and Véronique Plesch. "The Cultural Processes of 'Appropriation.'" *Journal of Medieval and Early Modern Studies* 32, no. 1 (2002): 1–15.

Aussems, Johannes Franciscus Alphonsus. "Christine de Pizan and the Scribal Fingerprint: A Quantitative Approach to Manuscript Studies." Master's thesis, Universiteit Utrecht, 2006.

Aussems, Mark, and Axel Brink. "Digital Palaeography." In *Kodikologie und Paläographie im digitalen Zeitalter: Codicology and Palaeography in the Digital Age*, vol. 2, edited by Franz Fischer, Christiane Fritze, George Vogeler and Pádraig Ó Macháin, 293–308. Books on Demand, 2009.

Avril, François, Yolanta Załuska, Marie-Thérèse Gousset, and Michel Pastoureau. *Dix siècles d'enluminure italienne: VIe–XVIe siècles.* Bibliothèque nationale, 1984.

Baillot, Anne, Françoise Gouzi, and Toma Tasovac, eds. "Workflows: Digital Methods for Reproducible Research Practices in the Arts and Humanities." Special issue, *Transformations: A DARIAH Journal* 1 (June 2025). https://transformations.episciences.org/volumes/965.

Beach, Alison Isdale. *Women as Scribes: Book Production and Monastic Reform in Twelfth-Century Bavaria*. Cambridge Studies in Palaeography and Codicology. Cambridge University Press, 2004.

Bender, Emily Menon, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜." *FAccT '21 Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. https://doi.org/10.1145/3442188.3445922.

Berry, David M. "The 'Computational Turn': Thinking About Digital Humanities." *Culture Machine* 12 (2011): 1–22.

Berry, David M., and Anders Fagerjord. "On the Way to Computational Thinking." In *Digital Humanities: Knowledge and Critique in the Digital Age*, 40–59. Polity, 2017.

Bitner, Kendall, and Kyle Dase. "A Macron Signifying Nothing: Revisiting *The Canterbury Tales Project* Transcription Guidelines." *Digital Medievalist*, Special Cluster 2 (December 22, 2021). https://doi.org/10.16995/dm.8068.

Blickhan, Samantha, Coleman Krawczyk, Daniel Hanson, Amy Boyer, Andrea Simenstad, and Victoria van Hyning. "Individual vs. Collaborative Methods of Crowdsourced Transcription." *Journal of Data Mining and Digital Humanities* (December 3, 2019): 1–33. https://doi.org/10.46298/jdmdh.5759.

Bobley, Brett. "Why the Digital Humanities?" Office of Digital Humanities, National Endowment for the Humanities. www.neh.gov/sites/default/files/inline-files/odh-resource_why_the_digital_humanities.pdf, accessed October 15, 2025.

Bod, Rens. "Modelling in the Humanities: Linking Patterns to Principles." In "Models and Modelling between Digital & Humanities–A Multidisciplinary Perspective." Special issue, *Historical Social Research / Historische Sozialforschung*, Supplement, 31 (2018): 78–95.

Bode, Katherine. *A World of Fiction: Digital Collections and the Future of Literary History*. University of Michigan Press, 2019.

Boillet, Mélodie, Marie-Laurence Bonhomme, Dominique Stutzmann, and Christopher Kermorvant. "HORAE: An Annotated Dataset of Books of Hours." In *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing*, 7–12. ACM: 2019.

Borek, Luise, Canan Hastik, Vera Khramova, Klaus Illmayer, and Jonathan David Geiger. "Information Organization and Access in Digital Humanities." In *Information between Data and Knowledge: Information Science and Its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021), Regensburg, Germany, March 8–10, 2021*, edited by Thomas Schmidt and Christian Wolff, 321–32. Schriften zur Informationswissenschaft 74. Hülsbusch, 2021.

Borges, Jorge Luis. *La biblioteca de Babel: El jardín de senderos que se bifurcan*. Sur, 1941.

Bourgain, Pascale. "Sur l'édition des textes littéraires latins médiévaux." *BECh* 150, no. 1 (1992): 5–49.

Bowker, Geoffrey C., and Susan Leigh Star. *Sorting Things Out: Classification and its Consequences.* MIT Press, 1999.

boyd, danah, and Kate Crawford. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication and Society* 15, no. 5 (2012): 662–79.

Bozzolo, Carla, Dominique Coq, Daniel Muzerelle, and Ezio Ornato. "Une machine au fonctionnement complexe: Le livre médiéval." In *Le texte et son inscription: Actes du colloque organisé par le Centre d'études de l'écriture de l'université Paris VII et le groupe Paragraphe de l'université Paris VIII (Paris, 13–15 juin 1984)*, edited by Paul Bady and Roger Laufer, 69–78. Centre national de la recherche scientifique, 1989.

Bozzolo, Carla, and Ezio Ornato. *Pour une histoire du livre manuscrit au Moyen Âge: Trois essais de codicologie quantitative.* Centre national de la recherche scientifique, 1983.

Branner, Robert. "The Johannes Grusch Atelier and the Continental Origins of the William of Devon Painter." *The Art Bulletin* 54, no. 1 (1972): 24–30.

Branner, Robert. *Manuscript Painting in Paris During the Reign of Saint Louis: A Study of Styles*. California Studies in the History of Art. University of California Press, 1977.

Brink, Axel, Marius Bulacu, and Lambert Schomaker. "How Much Handwritten Text Is Needed for Text-Independent Writer Verification and Identification." *19th International Conference on Pattern Recognition* (Tampa, December 8–11, 2008), 1–4. IEEE, 2008. https://doi.org/10.1109/icpr.2008.4761908.

Bulacu, Marius, and Lambert Schomaker. "Text-Independent Writer Identification and Verification Using Textural and Allographic Features." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, no. 4 (2007): 701–17.

Buringh, Eltjo. *Medieval Manuscript Production in the Latin West: Explorations with a Global Database*. Global Economic History Series 6. Brill, 2011.

Burrows, Toby, Doug Emery, Arthur Mitchell Fraas, et al. "A New Model for Manuscript Provenance Research: The Mapping Manuscript Migrations Project." *Manuscript Studies: A Journal of the Schoenberg Institute for Manuscript Studies* 6, no. 1 (2021): 131–44.

Camps, Jean-Baptiste, Thibault Clérice, and Ariane Pinche. "Noisy Medieval Data, From Digitized Manuscript to Stylometric Analysis: Evaluating Paul Meyer's Hagiographic Hypothesis." *Digital Scholarship in the Humanities* 36, Supplement 2 (2021): ii49–ii71.

Camps, Jean-Baptiste, Chahan Vidal-Gorène, Dominique Stutzmann, Marguerite Vernet, and Ariane Pinche. "Data Diversity in Handwritten Text Recognition: Challenge or Opportunity?" In *Digital Humanities 2022: Responding to Asian Diversity. Conference Abstracts (University of Tokyo, Japan, July 25–29, 2022)*, 160–65. DH2022 Local Organizing Committee, 2022. https://hal.science/hal-03916914.

Casilli, Antonio. "'Posthumani nihil a me alienum puto': Le discours de l'hospitalité dans la cyberculture." *Sociétés* 83, no. 1 (2004): 97–116.

Chagué, Alix, and Thibault Clérice. "I'm Here to Fight for Ground Truth: HTR-United, a Solution Towards a Common for HTR Training Data." In *Digital Humanities 2023: Collaboration as Opportunity* (University of Graz, Graz, Austria). DH2023 Local Organizing Committee, 2023. https://inria.hal.science/hal-04094233.

Cisne, John L. "How Science Survived: Medieval Manuscripts' 'Demography' and Classic Text's Extinction." *Science* 307 (2005): 1305–7.

Ciula, Arianna. "Digital Palaeography: Using the Digital Representation of Medieval Script to Support Palaeographic Analysis." *Digital Medievalist* 1 (2005). https://doi.org/10.16995/dm.4.

Ciula, Arianna, Øyvind Eide, Cristina Marras, and Patrick Sahle. "Models and Modelling between Digital and Humanities: A Multidisciplinary Perspective." In "Models and Modelling between Digital & Humanities–A Multidisciplinary Perspective." Special issue, *Historical Social Research / Historische Sozialforschung* 43, no. 4 (2018): 343–61.

Clanchy, Michael T. *From Memory to Written Record in England, 1066–1307*. Edward Arnold, 1979.

Clarkson, Christopher. "Rediscovering Parchment: The Nature of the Beast." *The Paper Conservator* 16, no. 1 (1992): 5–26.

Clérice, Thibault, Malamatenia Vlachou-Efstathiou, and Alix Chagué. "CREMMA Medii Aevi: Literary Manuscript Text Recognition in Latin." *Journal of Open Humanities Data* 9, no. 4 (2023): 1–19.

Cloppet, Florence, Véronique Eglin, Marlène Helias-Baron, Cuong Kieu, Nicole Vincent, and Dominique Stutzmann. "ICDAR2017 Competition on the Classification of Medieval Handwritings in Latin Script Dataset." In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017)*, 1371–76. IEEE, 2018.

Cloppet, Florence, Véronique Églin, Cuong Kieu, Dominique Stutzmann, and Nicole Vincent. "ICFHR2016 Competition on the Classification of Medieval Handwritings in Latin Script Dataset." *15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2017, 590–95. https://hal.science/hal-01403775.

Cohen, Daniel Jared, and Roy Rosenzweig. *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*. University of Pennsylvania Press, 2005.

Cohen, Evelyn M. "Can Colophons Be Trusted?: Insights from Decorated Hebrew Manuscripts Produced for Women in Renaissance Italy." In *The Hebrew Book in Early Modern Italy*, edited by Joseph R. Hacker and Adam Shear, 17–25. University of Pennsylvania Press, 2011.

Coladangelo, L. P., Emma Thomson, and Lynn Ransom. "Leveraging the Power of Crowdsourcing and Linked Open Data: Transformation of the Schoenberg Database of Manuscripts and the SDBM Name and Place Authorities." *Journal of Library Metadata* 23, nos. 1–2 (April 3, 2023): 1–22.

Crawford, Kate. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, 2021.

Cronk, Nicholas, and Glenn Roe. *Voltaire's Correspondence: Digital Readings*. Cambridge University Press, 2020.

Cugliana, Elisa, and Gioele Barabucci. "Signs of the Times: Medieval Punctuation, Diplomatic Encoding and Rendition." *Journal of the Text Encoding Initiative* 14 (2021): 1–32.

Dahan, Gilbert. "Paris Bibles and Scholarship." In *The Oxford Handbook of the Latin Bible*, edited by Hugh Alexander Gervase Houghton, 241–57. Oxford University Press, 2023.

Daneu-Lattanzi, Angela. "Ancora sulla scuola miniaturistica dell'Italia Meridionale Sveva." *La Bibliofilia* 66 (1964): 105–62.

Daneu-Lattanzi, Angela. *Una bibbia prossima alla Bibbia di Manfredi.* De Magistris, 1957.

Daneu-Lattanzi, Angela. *Lineamenti di storia della miniatura in Sicilia.* Olschki, 1966.

Daneu-Lattanzi, Angela. *I manoscritti ed incunaboli miniati della Sicilia* I. Accademia di scienze, lettere e arti di Palermo, 1965.

Davis, Tom. "The Practice of Handwriting Identification." *The Library* 8, no. 3 (September 1, 2007): 251–76.

D'Ignazio, Catherine, and Lauren Frederica Klein. *Data Feminism*. MIT Press, 2020.

D'Iorio, Paolo. "Qu'est-ce qu'une édition génétique numérique ?" *Genesis: Manuscrits – Recherche – Invention* 30 (2010): 49–53.

De Gussem, Jeroen. "Computational Stylistics and Medieval Texts." In *Routledge Resources Online: Medieval Studies*, edited by Hannele Klemettilä, Samu Niskanen, and James Willoughby, 1–12. Routledge, 2022.

De Sousa, Luís Correia. *Sacra pagina: Textos e imagens das Bíblias portáteis do século XIII pertencentes às coleções portuguesas*. Paulus, 2015.

Deploige, Jeroen, and Jeroen De Gussem. "Medieval Authorship and Canonicity in the Digital Age: An Introduction." *Interfaces: A Journal of Medieval European Literatures* 8 (2021): 113–24.

Derolez, Albert. *The Palaeography of Gothic Manuscript Books: From the Twelfth to the Early Sixteenth Century*. Cambridge University Press, 2006.

Destrez, Jean. *La Pecia dans les manuscrits universitaires du XIIIe et du XIVe siècle*. Vautrain, 1935.

Dunn, Stuart, and Mark Hedges. "From the Wisdom of Crowds to Going Viral: The Creation and Transmission of Knowledge in the Citizen Humanities." In *Citizen Inquiry: Synthesising Science and Inquiry Learning*, edited by Christothea Herodotou, Mike Sharples, and Eileen Scanlon, 25–41. Routledge, 2018.

Duranti, Luciana. *Diplomatics: New Uses for an Old Science*. Bloomsbury, 1998.

Eder, Maciej. "Mind Your Corpus: Systematic Errors in Authorship Attribution." *Literary and Linguistic Computing* 28, no. 4 (2013): 603–14.

Eder, Maciej. "Rolling Stylometry." *Digital Scholarship in the Humanities* 31, no. 3 (2016): 457–69.

Eder, Maciej, Jan Rybicki, and Mike Kestemont. "Stylometry with R: a package for computational text analysis." *R Journal* 8, no. 1 (2016): 107–21. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

Edlich-Muth, Christian, and Miriam Edlich-Muth. "A Computational Approach to Source Adaptation in Thomas Malory's Morte Darthur." *Digital Medievalist* 12, no. 1 (2019): 5. https://doi.org/10.16995/dm.86.

Eleftheriadi, Konstantina. "Online Tools for Handwritten Text Recognition: A Comparative Review of Transkribus and eScriptorium for Byzantine Paleography." *The Stoa: A Review for Digital Classics*. January 21, 2025. https://blog.stoa.org/archives/4308.

Les Enluminures. "The Bishop Carr Bible." *Les Enluminures*. www.textmanuscripts.com/medieval/the-bishop-carr-bible-261972, accessed October 15, 2025.

Les Enluminures. "Medieval Vulgate Bible," *Les Enluminures*. www.textmanuscripts.com/medieval/medieval-vulgate-bible-79779, accessed October 15, 2025.

Les Enluminures. "The Rugby-De Brailes Bible: In Latin, Illuminated Manuscript on Parchment." *Les Enluminures*. www.textmanuscripts.com/medieval/the-rugby-de-brailes-bible-195344, accessed October 15, 2025.

Eve, Martin Paul. *The Digital Humanities and Literary Studies.* The Literary Agenda. Oxford University Press, 2022.

Faller, Úna, and Diego Rodriguez. "Paris Bible Correct-a-thon Besançon: Beinecke ZZi 56." *Paris Bible Project* (blog). June 29, 2023. https://doi.org/10.5281/zenodo.8040632.

Fiddyment, Sarah, Bruce Holsinger, Chiara Ruzzier, et al. "Animal Origin of 13th-Century Uterine Vellum Revealed Using Noninvasive Peptide Fingerprinting." *Proceedings of the National Academy of Sciences* 112, no. 49 (2015): 15,066–71.

Flanders, Julia, and Fotis Jannidis, eds. *The Shape of Data in Digital Humanities: Modeling Texts and Text-Based Resources*. Routledge, 2018.

Fortunato, Santo, Carl T. Bergstrom, Katy Börner, et al. "Science of Science." *Science* 359, no. 6379 (2018): 1–7. https://doi.org/10.1126/science.aao0185.

Foys, Martin. "How I Learned to Stop Worrying and Love Big Data." *Burnable Books* (blog). January 16, 2013. https://web.archive.org/web/20160205024906/http://burnablebooks.com/foysonbigdata/, accessed October 15, 2025.

Francomano, Emily C., and Heather Bamford. "Whose Digital Middle Ages?: Accessibility in Digital Medieval Manuscript Culture." *Journal of Medieval Iberian Studies* 14, no. 1 (2022): 15–27.

Franzini, Greta, Mike Kestemont, Gabriela Rotari, et al. "Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm." *Frontiers in Digital Humanities* 5 (2018): 1–15.

Fridlund, Mats, Mila Oiva, and Petri Paju. *Digital History: Emergent Approaches Within the New Digital History*. Helsinki University Press, 2020.

Gallo, Federico. *Diplomatics: The Science of Reading Medieval Documents: A Handbook*. Milano University Press, 2024.

Ganz, David. "Carolingian Bibles." In *The New Cambridge History of the Bible*. Vol. 2, *From 600 to 1450*, edited by E. Ann Matter and Richard Marsden, 325–37. Cambridge University Press, 2012.

Gibbs, Frederick William. "New Textual Traditions from Community Transcription." *Digital Medievalist* 7 (2012). https://doi.org/10.16995/dm.39.

Gilissen, Léon. *Prolégomènes à la codicologie: Recherches sur la construction des cahiers et la mise en page des manuscrits médiévaux*. Story-Scientia, 1977.

Gillespie, Alexandra. "In Praise of Small Data" *Burnable Books* (blog). March 1, 2013. https://web.archive.org/web/20160712041841/http://burnablebooks.com/in-praise-of-small-data/, accessed October 15, 2025.

Gilsdorf, Sean, and Laura K. Morreale, eds. *Digital Medieval Studies: Experimentation and Innovation*. Arc Humanities, 2024.

Graham, Shawn, Ian Milligan, Scott Weingart, and Kim Martin. *Exploring Big Historical Data: The Historian's Macroscope.* 2nd ed. World Scientific, 2022.

Grésillon, Almuth. *Éléments de critique génétique*. Presses universitaires de France, 1994.

Guéville, Estelle. "Les manuscrits médiévaux occidentaux dans la collection du Louvre Abu Dhabi: 2009–2017." In *Le manuscrit médiéval: Texte, objet et outil de transmission*, vol. 1, 105–53. Pecia: Le livre et l'écrit 22. Brepols, 2019.

Guéville, Estelle, and David Joseph Wrisley. "Co-Authorship, Collaboration, and Community in Medieval Studies (1930–2025): Insights from the Science of Science." Paper presented at the *Association for Computers and the Humanities (ACH) 2025*, June 11, 2025.

Guéville, Estelle and David Joseph Wrisley. "Correct-A-Thon." *Paris Bible Project* (blog). January 2023. Accessed January 15, 2025. https://parisbible.github.io/correct-a-thon/, accessed October 15, 2025.

Guéville, Estelle, and David Joseph Wrisley. "Crowd Post-Correction of HTR Output in a Pedagogical Context." Paper presented at DH Benelux, June 30, 2023.

Guéville, Estelle, and David Joseph Wrisley. "Everyone Leaves a Trace: Exploring Transcriptions of Medieval Manuscripts with Computational Methods." *Digital Studies in Language and Literature* 1, nos. 1–2 (2024): 36–54.

Guéville, Estelle, and David Joseph Wrisley. "From Localization to Chronological and Geographical Prediction." Paper presented at International Medieval Congress (IMC 2022), July 7, 2022.

Guéville, Estelle, and David Joseph Wrisley. "Transcribing Medieval Manuscripts for Machine Learning." In "On the Way to the Future of Digital Manuscript Studies." Special issue, *Journal of Data Mining and Digital Humanities* (July 2, 2024). https://doi.org/10.46298/jdmdh.9805.

Hammarfelt, Björn. "Four Claims on Research Assessment and Metric Use in the Humanities." *Bulletin of the Association for Information Science and Technology* 43, no. 5 (2017): 33–38.

Hammarfelt, Björn, and Gaby Haddow. "Conflicting Measures and Values: How Humanities Scholars in Australia and Sweden Use and React to Bibliometric Indicators." *Journal of the Association for Information Science and Technology* 69, no. 7 (2018): 924–35.

Hanneken, Todd Russell. "What to Think about When Thinking About Digitization of Manuscripts." *Digital Philology: A Journal of Medieval Cultures* 12, no. 2 (2023): 256–84.

Hassner, Tal, Robert Sablatnig, Dominique Stutzmann, and Ségolène Tarte. "Digital Palaeography: New Machines and Old Texts (Dagstuhl Seminar 14302)." *Dagstuhl Reports* 4, no. 7 (2014): 112–33.

Haverals, Wouter, and Mike Kestemont. "From Exemplar to Copy: The Scribal Appropriation of a Hadewijch Manuscript Computationally Explored." In "On the Way to the Future of Digital Manuscript Studies." Special issue, *Journal of Data Mining and Digital Humanities* (April 20, 2023): 1–21. https://doi.org/10.46298/jdmdh.10206.

Haverals, Wouter, and Mike Kestemont. "Silent Voices: A Digital Study of the Herne Charterhouse Scribal Community (ca. 1350–1400)." *Queeste: tijdschrift over middeleeuwse letterkunde in de Nederlanden* 27, no. 2 (2020): 186–95.

Hedges, Mark, and Stuart Dunn. *Academic Crowdsourcing in the Humanities: Crowds, Communities and Co-Production*. Chandos Publishing, 2017.

Heinisch, Barbara, Kristin Oswald, Maike Weißpflug, Sally Shuttleworth, and Geoffrey Belknap. "Citizen Humanities." In *The Science of Citizen Science*, edited by Katrin Vohland et al., 97–118. Springer, 2021.

Herrmann, Julia Berenike, Karina van Dalen-Oskam, and Christof Schöch. "Revisiting Style: a Key Concept in Literary Studies." *Journal of Literary Theory* 9, no. 1 (2015): 25–52.

Herrnstein Smith, Barbara. "What Was 'Close Reading?': A Century of Method in Literary Studies." *minnesota review* 87 (2016): 57–75.

Hérubel, Jean-Pierre V. M. "Disciplinary Affiliations and Subject Dispersion in Medieval Studies." *Behavioral and Social Sciences Librarian* 23 (2005): 67–83.

Hinchen, Dan. "Transcription Challenge, Round Two." *The| Beehive*. www.masshist.org/beehiveblog/2016/06/transcription-challenge-round-2/.

Hodel, Tobias. "Das Ende der Edition?: Ein Blick auf geschichtswissenschaftliche Editionen mit Fokus auf die Frühe Neuzeit – eine Provokation." In *Post aus Nürnberg: Interdisziplinäre Forschungen zu den Briefbüchern des 15. Jahrhunderts*, 61–66. Nürnberger Forschungen 34. Schmidt, 2024.

Hodel, Tobias. "Supervised and Unsupervised: Approaches to Machine Learning for Textual Entities." In *Archives, Access and Artificial Intelligence: Working with Born-Digital and Digitized Archival Collections*, edited by Lise Jaillant, 157–77. Bielefeld University Press, 2022.

Hodel, Tobias, David Schoch, Christa Schneider, and Jake Purcell. "General Models for Handwritten Text Recognition: Feasibility and State-of-the-Art. German Kurrent as an Example." *Journal of Open Humanities Data* 7 (2021). https://doi.org/10.5334/johd.46.

Holsinger, Bruce. "The Googlization of … Paleography???" *Burnable Books* (blog). January 22, 2013. https://web.archive.org/web/20160416173246/http://burnablebooks.com/thegooglizationofpalaeography/, accessed October 15, 2025.

Hossain, Mokter. "Users' Motivation to Participate in Online Crowdsourcing Platforms." In 2012 *International Conference on Innovation Management and Technology Research* (ICIMTR2012), Malacca, Malaysia, May 21–22, 2012, 310–15. https://sci-hub.st/10.1109/ICIMTR.2012.6236409.

Impett, Leonardo. "Digital Art History as Critical AI." *The Art Bulletin* 106, no. 2 (2024): 11–14.

Impett, Leonardo, and Fabian Offert. "There is a Digital Art History." *Visual Resources* 38, no. 2 (2022): 186–209.

Jaillant, Lise. "Introduction." In *Archives, Access and Artificial Intelligence: Working with Born-Digital and Digitized Archival Collections*, edited by Lise Jaillant, 7–28. Bielefeld University Press, 2022.

James, Montague Rhodes. *A Descriptive Catalogue of the Manuscripts in the Library of Corpus Christi College Cambridge*. Vols. 1–2. Cambridge University Press, 1909–1912.

Joyeux-Prunel, Béatrice. "Digital Humanities in the Era of Digital Reproducibility: Towards a Fairest and Post-Computational Framework." *International Journal of Digital Humanities* 6 (2024): 23–43.

Julien, Octave. "Construction et composition des recueils français du XV^{e} siècle: Apports de la codicologie quantitative." *Babel: Littératures plurielles* 16 (2007): 13–30.

Kaplan, Frédéric. "A Map for Big Data Research in Digital Humanities." *Frontiers in Digital Humanities* 2 (2015): 1–7.

Kaya, Metin. "Bibliometric Analysis of Constantinople Studies Through Voswiever." *Kültür Araştırmaları Dergisi* 22 (2024): 308–18.

Ker, Neil Ripley. *Medieval Manuscripts in British Libraries*. Vol. 1. Clarendon, 1969.

Kestemont, Mike. "A Computational Analysis of the Scribal Profiles in Two of the Oldest Manuscripts of Hadewijch's Letters," *Scriptorium* 69, no. 2 (2015): 159–77.

Kestemont, Mike, Vincent Christlein, and Dominique Stutzmann. "Artificial Paleography: Computational Approaches to Identifying Script Types in Medieval Manuscripts." *Speculum* 92, suppl. 1 (2017): 86–109.

Kestemont, Mike, Folgert Karsdorp, Elisabeth de Bruijn, et al. "Forgotten Books: The Application of Unseen Species Models to the Survival of Culture." *Science* 375, no. 6582 (2022): 765–69.

Kestemont, Mike, Sara Moens, and Jeroen Deploige. "Collaborative Authorship in the Twelfth Century: A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux." *Digital Scholarship in the Humanities* 30, no. 2 (2015): 199–224.

Kidd, Peter. "Introduction." In *The McCarthy Collection*. Vol. 3, *French Miniatures*, 9–13. Ad Illisvm, 2021.

Kiessling, Benjamin, Robin Tissot, Peter Anthony Stokes, and Daniel Stökl Ben Ezra. "eScriptorium: An Open Source Platform for Historical Document Analysis." In *2019 International Conference on Document Analysis and Recognition Workshops* (ICDARW, Sydney, NSW, Australia), 19–24. IEEE, 2019. https://doi.org/10.1109/icdarw.2019.10032.

Kirmizialtin, Suphan, and David Joseph Wrisley. "Exploring Gulf Manumission Documents with Word Vectors." *Journal of Digital Islamicate Research* 2, nos. 1–2 (2024): 1–29.

Klein, Julie Thompson. *Interdisciplining Digital Humanities: Boundary Work in an Emerging Field*. Digital Humanities Series. University of Michigan Press, 2015.

Klein, Lauren Federica, Meredith Martin, André Brock, Maria Antoniak, Melanie Walsh, Jessica Marie Johnson, Lauren Tilton, and David Mimno. "Provocations from the Humanities for Generative AI Research." arXiv, February 26, 2025. https://doi.org/10.48550/arXiv.2502.19190.

Kline, Mary-Jo. *A Guide to Documentary Editing.* 2nd ed. Johns Hopkins University Press, 1998.

Koho, Mikko, Toby Burrows, Eero Hyvönen, et al. "Harmonizing and Publishing Heterogeneous Premodern Manuscript Metadata as Linked Open Data." *Journal of the Association for Information Science and Technology* 73, no. 2 (2022): 240–57.

Koschwitz, Eduard. *Commentar zu den ältesten französischen Sprachdenkmälern*. Henninger, 1886.

Koschwitz, Eduard. *Les plus anciens monuments de la langue française publiés pour les cours universitaires*. Henninger, 1879.

Lang, Andrew S. I. D., and Joshua Rio-Ross. "Using Amazon Mechanical Turk to Transcribe Historical Handwritten Documents." *Code4Lib Journal* 15 (October 31, 2011). https://journal.code4lib.org/articles/6004, accessed October 15, 2025.

Larivière, Vincent, Yves Gingras, Cassidy Rose Sugimoto, and Andrew Tsou. "Team Size Matters: Collaboration and Scientific Impact Since 1900." *Journal of the Association for Information Science and Technology* 66, no. 7 (2015): 1323–32.

Larivière, Vincent, Stefanie Haustein, and Katy Börner. "Long-Distance Interdisciplinarity Leads to Higher Scientific Impact." *PLOS ONE* 10, no. 3 (March 30, 2015): 1–15. https://doi.org/10.1371/journal.pone.0122565.

Laßwitz, Kurd. *Die Universalbibliothek: Erzählung*. JMB, 2010.

Lemercier, Claire, and Claire Zalc. *Quantitative Methods in the Humanities: An Introduction*. Translated by Arthur Goldhammer. University of Virginia Press, 2019.

Lévy, Pierre. *L'intelligence collective: Pour une anthropologie du cyberspace*. La Découverte, 1994.

Li, Charles. "Critical Diplomatic Editing: Applying Text-Critical Principles as Algorithms." In *Advances in Digital Scholarly Editing*, 305–10. Sidestone, 2017.

Li, Weihua, Tomaso Aste, Fabio Caccioli, and Giacomo Livan. "Early Coauthorship with Top Scientists Predicts Success in Academic Careers." *Nature Communications* 10, no. 1 (2019): 1–9.

Light, Laura. "The Thirteenth Century and the Paris Bible." In *The New Cambridge History of the Bible*. Vol. 2, *From 600 to 1450*, edited by E. Ann Matter and Richard Marsden, 380–91. Cambridge University Press, 2012.

Light, Laura. "What Was a Bible For?: Liturgical Texts in Thirteenth-Century Franciscan and Dominican Bibles." *Lusitania Sacra* 34 (2016): 165–82.

Linde, Cornelia. *How to Correct the Sacra Scriptura? Textual Criticism of the Bible Between the Twelfth and Fifteenth Centuries*. Society for the Study of Medieval Languages and Literature, 2015.

Magrini, Sabina. "Production and Use of Latin Bible Manuscripts in Italy During the Thirteenth and Fourteenth Centuries." *Manuscripta* 51, no. 2 (2007): 209–57.

Mallon, Jean. *Paléographie romaine*. Consejo Superior de Investigaciones Científicas / Instituto Antonio de Nebrija de Filología, 1952.

Maniaci, Marilena, ed. *Trends in Statistical Codicology*. Studies in Manuscript Cultures. De Gruyter, 2021.

Marciano, Richard. "Afterword: Towards a New Discipline of Computational Archival Science (CAS)." In *Archives, Access and Artificial Intelligence: Born-Digital and Digitized Archival Collections*, edited by Lise Jaillant, 205–18. Bielefeld University Press, 2022.

Massot, Marie-Laure, Arianna Sforzini, and Vincent Ventresque. "Transcribing Foucault's Handwriting With Transkribus." *Journal of Data Mining and Digital Humanities* (2019). https://doi.org/10.46298/jdmdh.5043.

McCarty, Willard. *Humanities Computing*. Palgrave Macmillan, 2005.

McGillivray, Murray. "Statistical Analysis of Digital Paleographic Data: What Can It Tell Us?" *Digital Studies / Le champ numérique* 11 (2005). https://doi.org/10.16995/dscn.248.

McGrady, Deborah. "Change in the Age of Big Data, or, How Nostalgia-Driven Studies May Be Our Future." *Burnable Books* (blog). January 27, 2013. https://web.archive.org/web/20160416173241/http://burnablebooks.com/change-in-the-age-of-big-data/, accessed October 15, 2025.

McIntosh, Angus. "Scribal Profiles from Middle English Texts." *Neuphilologische Mitteilungen* 76, no. 2 (1975): 218–35.

Meyers, C. "The Transition From Pen to Press," Master's of Fine Arts Thesis. Yale University, 1983.

Milligan, Ian. *History in the Age of Abundance?: How the Web Is Transforming Historical Research*. McGill-Queen's University Press, 2019.

Modern Language Association (MLA). "Guidelines for Evaluating Digital Scholarship." *Modern Language Association of America*, 2024. www.mla.org/About-Us/Governance/Committees/Committee-Listings/Professional-Issues/Committee-on-Information-Technology/Guidelines-for-Evaluating-Digital-Scholarship.

Morard, Martin. "A la recherche de la 'Lettre commune': La Bible latine du Moyen Âge tardif." *Sacra pagina*. Institut de recherche et d'histoire des textes / Centre national de la recherche scientifique, 2025. https://gloss-e.irht.cnrs.fr/php/page.php?id=35.

Morreale, Laura K., Gloria Sánchez Argüelles, Timothy Baldwin, et al. "Transcribing *Le Pèlerinage de Damoiselle Sapience*: Scholarly Editing Covid19-Style." *Digital Medievalist* 15, no. 1 (2022): 1–33.

Muehlberger, Guenter, Louise Seaward, Melissa Mhairi Terras, et al. "Transforming Scholarship in the Archives through Handwritten Text, Recognition: Transkribus as a Case Study." *Journal of Documentation* 75, no. 5 (2019): 954–76.

Murel, Jacob, and David Smith. "Computational Bibliography as Contextualization." Paper presented at the Renaissance Society of America Annual Meeting, March 21, 2025.

Muzerelle, Denis. "Pour revenir sur et à la 'taille' des manuscrits." *Gazette du livre médiéval* 50 (2007): 55–63.

Neddermeyer, Uwe. *Von der Handschrift zum gedruckten Buch: Schriftlichkeit und Leseinteresse im Mittelalter und in der frühen Neuzeit. Quantitative und qualitative Aspekte.* Vol. 1. Harrassowitz, 1998.

Nichols, Stephen George. "Dynamic Reading of Medieval Manuscripts." *Florilegium* 32 (2015): 19–57.

Nichols, Stephen George. *From Parchment to Cyberspace: Medieval Literature in the Digital Age*. Peter Lang, 2016.

Nichols, Stephen George. "It's the Manuscripts, Stupid!" *Burnable Books* (blog). January 30, 2013. https://web.archive.org/web/20160507033039/http://burnable2books.com/its-the-manuscripts-stupid/.

Nichols, Stephen George. "Why Material Philology?" *Zeitschrift für deutsche Philologie* 116 (1997): 10–30.

Nielsen, Torben Kjersgaard. "Research Output in Medieval and Crusade Studies 1981–2011: A Bibliometric Survey." *Crusades* 16 (2017): 147–64.

Nockels, Joseph, Paul Gooding, Sarah Ames, and Melissa Mhairi Terras. "Understanding the Application of Handwritten Text Recognition Technology in Heritage Contexts: A Systematic Review of Transkribus in Published Research." *Archival Science* 22, no. 3 (2022): 367–92.

Nockels, Joseph, Paul Gooding, and Melissa Mhairi Terras. "The Implications of Handwritten Text Recognition for Accessing the Past at Scale." *Journal of Documentation* 80, no. 7 (2024): 148–67.

Offert, Fabian, and Peter Bell. "Generative Digital Humanities." In *CHR 2020: Workshop on Computational Humanities Research, November 18–20, 2020, Amsterdam, The Netherlands*, 202–12. CEUR, 2020.

Okamura, Keisuke. "Interdisciplinarity Revisited: Evidence for Research Impact and Dynamism." *Palgrave Communications* 5, no. 1 (2019): 1–9.

Ornato, Ezio. "La codicologie quantitative: Outil privilégié de l'histoire du livre médiéval." *Gazette du livre médiéval* 6, no. 1 (1985): 7–13.

Ornato, Ezio. "Les conditions de production et de diffusion du livre médiéval (XIII^{e}–XV^{e} siècles)." *Publications de l'École française de Rome* 82 (1985): 57–84.

Padilla, Thomas. "On a Collections as Data Imperative," *UC Santa Barbara*. 2017. https://escholarship.org/uc/item/9881c8sv.

Park, Michael, Erin Leahey, and Russell J. Funk. "Papers and Patents Are Becoming Less Disruptive over Time." *Nature* 613, no. 7942 (2023): 138–44.

Pavlopoulos, John, Vasiliki Kougia, Esteban Garces Arias, et al. "Challenging Error Correction in Recognised Byzantine Greek." *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, 1–12. ACL, 2024. https://doi.org/10.18653/v1/2024.ml4al-1.1.

Pettenati, Silvia. "Un'altra 'Bibbia di Manfredi.'" *Prospettiva* 4 (1976): 7–15.

Pierazzo, Elena. "Modelling Digital Scholarly Editing: From Plato to Heraclitus." In *Digital Scholarly Editing: Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo, 41–58. Open Book, 2016.

Pierazzo, Elena. "A Rationale of Digital Documentary Editions." *Literary and Linguistic Computing* 26, no. 4 (2011): 463–77.

Pinche, Ariane, and Peter Anthony Stokes. "Historical Documents and Automatic Text Recognition: Introduction." In "Historical Documents and Automatic Text Recognition." Special issue, *Journal of Data Mining and Digital Humanities* (March 19, 2024). https://doi.org/10.46298/jdmdh.13247.

Piotrowski, Michael. *Natural Language Processing for Historical Texts*. Springer, 2012.

Piper, Andrew. *Enumerations: Data and Literary Study*. University of Chicago Press, 2018.

Piron, Florence. "Postcolonial Open Access." In *Open Divide: Critical Studies in Open Access*, edited by Ulrich Herb and Joachim Schöpfel, 117–28. Litwin, 2018.

Plachta, Bodo. *Editionswissenschaft: Eine Einführung in Methode und Praxis der Edition neuerer Texte*. Reclam, 1997.

Price, Neil, Mark Collard, and Ben Raffield. "Polygyny, Concubinage, and the Social Lives of Women in Viking-Age Scandinavia." *Viking and Medieval Scandinavia* 13 (2017): 165–209.

Rambaran-Olm, Mary, M. Breann Leake, and Micah James Goodrich. "Medieval Studies: The Stakes of the Field." *postmedieval: A Journal of Medieval Cultural Studies* 11 (2020): 356–70.

Ramsay, Stephen. "In Praise of Pattern." *TEXT Technology* 14, no. 2 (2005): 177–90.

Reilly, Brian J., and Moira R. Dillon. "Virtuous Circles of Authorship Attribution through Quantitative Analysis: Chrétien de Troyes's *Lancelot*." *Digital Philology: A Journal of Medieval Cultures* 2, no. 1 (2013): 60–85.

Reynhout, Lucien. "Codicologie quantitative et paradigmes scientifiques: Une typologie des formules latines des colophons de manuscrits occidentaux." *Gazette du livre médiéval* 39, no. 1 (2001): 1–9.

Ridge, Mia. "Crowdsourcing in Cultural Heritage: A Practical Guide to Designing and Running Successful Projects." In *Routledge International Handbook of Research Methods in Digital Humanities*, 363–83. Routledge, 2020.

Ridge, Mia. *Crowdsourcing Our Cultural Heritage*. Routledge, 2016.

Rigg, Arthur George. "The Editing of Medieval Latin Texts: A Response." *Studi medievali* 24, no. 3 (1983): 385–88.

Risam, Roopika. *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy.* Northwestern University Press, 2019.

Robinson, Peter. "What Text Really Is Not, and Why Editors Have to Learn to Swim." *Literary and Linguistic Computing* 24, no. 1 (1997): 41–52.

Robinson, Peter, and Elizabeth Solopova. "Guidelines for Transcription of the Manuscripts of the *Wife of Bath's Prologue*." In *The Canterbury Project Occasional Papers*, vol. 1, edited by Norman F. Blake and Peter Robinson, 19–52. Office for Humanities Communication, 1993.

Rojas Castro, Antonio. "Los principios FAIR y el Proyecto Humboldt Digital: Una confrontación." In *Humanidades digitales y patrimonio cultural: Proyectos y tendencias*, edited by Anna Peirats and José Antonio Calvo, 161–83. Tirant Lo Blanch, 2023.

Romein, Annemieke Christel, Tobias Hodel, Fieke Gordijn, et al. "Exploring Data Provenance in Handwritten Text Recognition Infrastructure: Sharing and Reusing Ground Truth Data, Referencing Models, and Acknowledging Contributions. Starting the Conversation on How We Could Get It Done." In "Historical Documents and Automatic Text Recognition." Special issue, *Journal of Data Mining and Digital Humanities* (March 18, 2024). https://doi.org/10.46298/jdmdh.10403.

Rosenzweig, Roy. "Scarcity or Abundance?: Preserving the Past in a Digital Era." *The American Historical Review* 108, no. 3 (2003): 735–62.

Rouse, Richard Hunter, and Mary Ames Rouse. "Book Production in Paris." In *The Oxford Handbook of Latin Palaeography*, edited by Frank T. Coulson and Robert G. Babcock, 812–22. Oxford University Press, 2020.

Rouse, Richard Hunter, and Mary Ames Rouse. *Manuscripts and Their Makers: Commercial Book Producers in Medieval Paris, 1200–1500*. Miller, 2000.

Russell, John E., and Merinda Kaye Hensley. "Beyond Buttonology: Digital Humanities, Digital Pedagogy, and the ACRL Framework." *College and Research Library News* 78, no. 11 (2017). https://crln.acrl.org/index.php/crlnews/article/view/16833/18427.

Ruzzier, Chiara. "Continuité et rupture dans la production des bibles au XIII^{e} siècle." In *Comment le livre s'est fait livre: La fabrication des manuscrits bibliques (IVe–XVe siècle)*, edited by Chiara Ruzzier and Xavier Hermand, 155–68. Bibliologia 40. Brepols, 2015.

Ruzzier, Chiara. *Entre université et ordres mendiants: La production des bibles portatives latines au XIIIe siècle*. De Gruyter, 2022.

Ruzzier, Chiara. "The Miniaturisation of Bible Manuscripts in the Thirteenth Century: A Comparative Study." In *Form and Function in the Late Medieval Bible*, edited by Eyal Poleg and Laura Light, 105–12. Brill, 2013.

Ruzzier, Chiara. "Qui lisait les bibles portatives fabriquées au XIII[e] siècle?" In *Lecteurs, lectures et groupes sociaux au Moyen Âge*, edited by Xavier Hermand, Etienne Renard, and Céline Van Hoorebeeck, 9–28. Texte, Codex et Contexte 17. Brepols, 2014.

Salmi, Hannu. *What Is Digital History?* Polity, 2021.

Schneiderman, Ben. *Human-Centered AI*. Cambridge University Press, 2022.

Schoen, Jenna, and Gianmarco E. Saretto. "Optical Character Recognition (OCR) and Medieval Manuscripts: Reconsidering Transcriptions in the Digital Age." *Digital Philology: A Journal of Medieval Cultures* 11, no. 1 (2022): 174–206.

Shillingsburg, Peter Leroy. *Scholarly Editing in the Computer Age: Theory and Practice*. University of Michigan Press, 1996.

Siemens, Lynne. "It's a Team if You Use 'Reply All': An Exploration of Research Teams in Digital Humanities Environments." *Linguistic and Literary Computing* 24, no. 2 (2009): 225–33.

Smit, Jinna. "The Death of the Palaeographer?: Experiences with the Groningen Intelligent Writer Identification System (GIWIS)." *Archiv für Diplomatik* 57 (2011): 413–26.

Smithies, James. "Software Intensive Humanities." In *The Digital Humanities and the Digital Modern*, 153–202. Palgrave Macmillan, 2017.

Somfai, Anna. "Medieval Manuscript Layouts: A Cognitive Journey through the Page." *The Vatican Library Review* 3 (2024): 1–35.

Somfai, Anna. "Visual Thinking: A Cognitive Reading of Codex Layouts." In *Visual Learning–A Year After*, edited by András Benedek and Kristóf Nyíri, 19–27. Budapest Visual Learning Lab, 2019.

Sotheby's. "Bible du XIII[e] siècle, Paris, vers 1240–1250." *Sotheby's*. www.sothebys.com/en/buy/auction/2024/livres-et-manuscrits-2/bible-du-xiiie-siecle-paris-vers-1240-1250-tres, accessed October 15, 2025.

Stegmüller, Friedrich [Fridericus]. *Repertorium biblicum medii aevi*. Vol. 1. Consejo Superior de Investigaciones Científicas, 1981.

Stinton, Timothy. "An Unrevolutionary Revolution: The Other 99%." *Burnable Books* (blog). January 18, 2013. https://web.archive.org/web/20160712042516/http://burnablebooks.com/unrevolutionaryrevolution/, accessed October 15, 2025.

Stokes, Peter Anthony. "Computer-Aided Palaeography, Present and Future." In *Kodikologie und Paläographie im digitalen Zeitalter 2: Codicology and Palaeography in the Digital Age 2*, edited by Franz Fischer, Christiane Fritze, and Georg Vogeler, 309–38. Books on Demand, 2009.

Stokes, Peter Anthony. "Scribal Attribution Across Multiple Scripts: A Digitally Aided Approach." *Speculum* 92(S1) (2017): 65–85.

Stokes, Peter Anthony, Benjamin Kiessling, Daniel Stökl Ben Ezra, Robin Tissot, and El Hassane Gargem. "The eScriptorium VRE for Manuscript Cultures." *Classics@* 18, no. 1 (2021). https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/.

Stubbs, Estelle Vivienne. "A Study of the Codicology of Four Early Manuscripts of the *Canterbury Tales*: Aberystwyth, National Library of Wales MS. Peniarth 3920 (Hengwrt), Oxford, Corpus Christi College, MS. 198 (Corpus), London, British Library MS. Harley 7334 (Harley 4), and California, San Marino, Huntington Library MS. El. 26 C 9 (Ellesmere)." PhD diss., University of Sheffield, 2006.

Stutzmann, Dominique. "Clustering of Medieval Scripts Through Computer Image Analysis: Towards an Evaluation Protocol." *Digital Medievalist* 10 (2016). https://doi.org/10.16995/dm.61.

Stutzmann, Dominique. "Nouvelles technologies au service de la codicologie et de la paléographie." *Scriptorium* 65, no. 1 (2011): 217–23.

Stutzmann, Dominique. "Paléographie statistique pour décrire, identifier, dater… normaliser pour coopérer et aller plus loin?" In *Kodikologie und Paläographie im digitalen Zeitalter 2: Codicology and Palaeography in the Digital Age 2*, edited by Franz Fischer, Christiane Fritze, and Georg Vogeler, 247–77. Books on Demand, 2011.

Stutzmann, Dominique. *Projet ANR-12-CORP-0010 Oriflamms: Compte-rendu de fin de projet*. 2016. http://oriflamms.hypotheses.org/files/2017/04/Oriflamms-Compte-rendu-final.pdf, accessed October 15, 2025.

Stutzmann, Dominique, Viola Mariotti, and Floriana Ceresato. "Les abréviations dans les manuscrits français du XIII^{e} siècle: Analyses statistiques." Paper presented at *L'emersione delle scritture volgari – L'émergence des écrits en langue vulgaire – The Rise of Vernacular Writing: La prospettiva paleografica – Le point de vue paléographique – The Palaeographical Perspective. XXI Convegno del Comité international de paléographie latine*. 2020. https://shs.hal.science/halshs-03560918v1.

Stutzmann, Dominique, Christopher Tensmeyer, and Vincent Christlein. "Writer Identification and Script Classification: Two Tasks for a Common Understanding of Cultural Heritage." *manuscript cultures* 15 (2020): 11–24.

Terras, Melissa Mhairi. "The Role of the Library When Computers Can Read: Critically Adopting Handwritten Text Recognition (HTR) Technologies to Support Research." In *The Rise of AI: Implications and Applications of Artificial Intelligence in Academic Libraries*, edited by Amanda Wheatley and Sandy Hervieux, 137–48. Association of College and Research Libraries, 2022.

Terras, Melissa Mhairi, Bettina Anzinger, Paul Gooding, Günter Mühlberger, Joe Nockels, Christel Annemieke Romein, Andy Stauder, and Florian Stauder. "The Artificial Intelligence Cooperative: READ-COOP, Transkribus, and the Benefits of Shared Community Infrastructure for Automated Text Recognition." *Open Research Europe* 5, no. 16 (2025). https://doi.org/10.12688/openreseurope.18747.1.

Thiel, Sonja, and Johannes Christian Bernhardt, eds. *AI in Museums: Reflections, Perspectives and Applications*. Transcript, 2023.

Thompson, Rodney Malcolm. "Technology of Production of the Manuscript Book: Parchment and Paper, Ruling and Ink." In *The Cambridge History of the Book in Britain*, vol. 2, edited by Nigel Morgan and Rodney Malcolm Thompson, 75–84. Cambridge University Press, 2008.

Thylstrup, Nanna Bonde. *The Politics of Mass Digitization*. MIT Press, 2019.

Tilton, Lauren, David Mimno, and Jessica Marie Johnson. "What Gets Counted: Computational Humanities Under Revision." In *Computational Humanities*, edited by Lauren Tilton, David Mimno, and Jessica Marie Johnson, vii–xviii. University of Minnesota Press, 2024.

Toubert, Hélène. "Trois nouvelles bibles du Maître de la Bible de Manfred et de son atelier." *Mélanges de l'École française de Rome: Moyen Âge* 89, no. 2 (1977): 743–68.

Treharne, Elaine. "Fleshing Out the Text: The Transcendent Manuscript in the Digital Age." *postmedieval: A Journal of Medieval Cultural Studies* 4 (2013): 465–78.

Treharne, Elaine. "'I Tag Bad,' or, Learning New Tricks in the Age of Big Data." *Burnable Books* (blog). February 4, 2013. https://web.archive.org/web/20160712041836/http://burnablebooks.com/i-tag-bad/, accessed October 15, 2025.

Treharne, Elaine. "More, True, Better: The Digital Book and Its Frameworks of Understanding." In *Perceptions of Medieval Manuscripts: The Phenomenal Book*, edited by Elaine Treharne, 15–30. Oxford University Press, 2021.

Treharne, Elaine, Benjamin Albritton, and Georgia Henley, eds. *Medieval Manuscripts in the Digital Age*. Routledge, 2020.

Underwood, Ted. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, 2019.

University of Pennsylvania Libraries. "Biblia sacra manuscripta (Ms. Codex 236)" [catalogue entry]. https://find.library.upenn.edu/catalog/9915517913503681, accessed October 15, 2025.

Uzzi, Brian, Satyam Mukherjee, Michael Stringer, and Ben Jones. "Atypical Combinations and Scientific Impact." *Science* 342, no. 6157 (2013): 468–72.

Van Dalen-Oskam, Karina. *The Riddle of Literary Quality: A Computational Approach*. Amsterdam University Press, 2023.

Vandendorpe, Christian. *Du papyrus à l'hypertexte: Essai sur les mutations du texte et de la lecture*. La Découverte, 1999.

Vander Meulen, David L., and George Thomas Tanselle. "A System of Manuscript Transcription." *Studies in Bibliography* 52 (1999): 201–12.

Vandyck, Caroline, and Mike Kestemont. "Abbreviation Application: A Stylochronometric Study of Abbreviations in the Oeuvre of Herne's Speculum Scribe." In *Computational Humanities Research Conference, December 4–6, 2024, Århus, Denmark,* 881–91. CEUR, 2024. https://ceur-ws.org/Vol-3834/paper15.pdf.

Van Erp, Marieke, Barbara McGillivray, and Tobias Blanke. "The Future of Digital Humanities Research: Alone You May Go Faster, but Together You'll Get Further." In *Computational Humanities*, edited by Lauren Tilton, David Mimno, and Jessica Marie Johnson, 233–46. University of Minnesota Press, 2024.

Van Lit, Cornelis L. W. "The Digital Materiality of Digitized Manuscripts." In *Among Digitized Manuscripts: Philology, Codicology, Paleography in a Digital World*, 51–72. Brill, 2019.

Victor, Benjamin. "Une Bible portative du XIII^{e} siècle dans les collections de l'UQAM." *Memini: Travaux et documents* 15 (2011): 23–38.

Volmering, Nicole. "Early Irish Hands TranSCRIPTion Challenge | FromThePage." https://fromthepage.com/nicolev/eih-transcription-challenge, accessed May 18, 2025.

Wang, Dashun, and Albert-László Barabási. *The Science of Science*. Cambridge University Press, 2021.

Warren, Michelle R. *Holy Digital Grail: A Medieval Book on the Internet*. Stanford University Press, 2022.

Whearty, Bridget. *Digital Codicology: Medieval Books and Modern Labor*. Stanford University Press, 2022.

Whearty, Bridget, and Dot Porter. "Not Just 'Can We?' but 'Should We?' and 'Why?': Understanding Digital Manuscripts as (Big) Data." *Digital Philology: A Journal of Medieval Cultures* 14, no. 1 (2025): 142–56.

Widner, Michael. "Toward Text-Mining the Middle Ages." In *The Routledge Research Companion to Digital Medieval Literature*, edited by Jennifer E. Boyle and Helen J. Burgess, 131–44. Routledge, 2017.
Wrisley, David Joseph. "Enacting Open Scholarship in Transnational Contexts." *POP! Public Open Participatory* 1 (October 31, 2019). https://doi.org/10.21810/pop.2019.002.
Wrisley, David Joseph. "Infrastructure as Privilege." *Historical Reflections / Réflexions historiques* 49, no. 3 (2023): 28–36.
Wrisley, David Joseph, Estelle Guéville, and Niccolò Acram Cappelletto. "Creating New Audiences for Digital Objects Through Museum-University Collaboration," *Museums in the MENA Journal* 3 (2022): 61–63.
Wu, Lingfei, Dashun Wang, and James A. Evans. "Large Teams Develop and Small Teams Disrupt Science and Technology." *Nature* 566, no. 7744 (February 2019): 378–82.
Wuchty, Stefan, Benjamin Felt Jones, and Brian Uzzi. "The Increasing Dominance of Teams in Production of Knowledge." *Science* 316, no. 5827 (May 18, 2007): 1036–39.
XP Method. "Architectures of Knowledge: Digital Methodologies Across the Humanities." Columbia University. https://xpmethod.columbia.edu/events/2016-07-04-AoK_Mumbai.html, accessed October 15, 2025.
Zaagsma, Gerben. "Digital History and the Politics of Digitization." *Digital Scholarship in the Humanities* 38, no. 2 (2023): 830–51.
Zumthor, Paul. "Intertextualité et mouvance." *Littérature* 41 (1981): 8–16.

Digital Resources

All digital resources were last accessed October 15, 2025.

Archives de littérature du Moyen Âge (ARLIMA). "Présentation." ARLIMA. www.arlima.net/presentation.html.
Biblioteca nacional digital. Biblioteca nacional de Portugal. https://bndigital.bnportugal.gov.pt/.
Bibliothèques numériques de l'Institut de France (BNIF). https://bibnum.institutdefrance.fr/.
Biblissima+. "TranscriboQuest Summer School." Biblissima+ (Projet Biblissima), September 2024. https://projet.biblissima.fr/en/news/transcriboquest-summer-school.
BodoArXiv. BodoArXiv: Open Repository for Medieval Studies. OSF. https://osf.io/preprints/bodoarxiv.
BodoArXiv. "Welcome / About BodoArXiv." BodoArXiv (WordPress). https://bodoarxiv.wordpress.com/.
Chagué, Alix and Thibault Clérice. HTR-United. https://htr-united.github.io/.
Choco-Mufin. PyPI: chocomufin. https://pypi.org/project/chocomufin/.
Cornell Lab of Ornithology. "Home." www.birds.cornell.edu/home/.
DigiPal: Digital Resource and Database of Manuscripts, Palaeography and Diplomatic. London, 2011–2014. www.digipal.eu/.
Digital Scriptorium. Digital Scriptorium Search Portal. https://search.digital-scriptorium.org/.

DigiVatLib. Digital Library Service of the Biblioteca Apostolica Vaticana. https://digi.vatlib.it/.
Gallica. Bibliothèque numérique de la Bibliothèque nationale de France. https://gallica.bnf.fr/.
Guéville, Estelle and David Joseph Wrisley. Paris Bible Project. https://parisbible.github.io/.
Gutenberg Bible Census. https://clausenbooks.com/gutenbergcensus.htm.
Handschriftenportal. https://handschriftenportal.de/.
Harzing, Anne-Wil. Publish or Perish [software]. 2007. https://harzing.com/resources/publish-or-perish.
Humanities and Design Lab—Stanford University, Palladio. https://hdlab.stanford.edu/palladio-app/.
Internet Archive. https://archive.org/.
Medieval Academy of America. "Digital Medieval Studies Institute (DMSI) 2025." The Medieval Academy Blog, January 30, 2025. www.themedievalacademyblog.org/digital-medieval-studies-institute-dmsi-2025/.
Middle Ages for Educators. *Middle Ages for Educators*. https://middleagesforeducators.princeton.edu/.
Mooney, Linne, Simon Horobin, and Estelle Stubbs. Late Medieval English Scribes. www.medievalscribes.com.
MUFI: The Medieval Unicode Font Initiative. https://mufi.info/
Open Greek and Latin Project. www.opengreekandlatin.org/.
Oriflamms. "Script Classification and Writer Identification: Two Tasks for a Common Understanding of Cultural Heritage." [dataset and code] https://github.com/oriflamms/Script-Classification-Writer-Identification/.
Parker on the Web. Manuscripts in the Parker Library at Corpus Christi College, Cambridge. https://parker.stanford.edu/parker.
Philobiblon. "Home." Universitat Pompeu Fabra. https://philobiblon.upf.edu/html/index_es.html.
Pinche, Ariane, Thibault Clérice, Alix Chagué, et al. "CATMuS-Medieval: Consistent Approaches to Transcribing ManuScripts. A Generalized Set of Guidelines and Models for Latin Scripts from the Middle Ages (8th–16th Century)." 2023. https://catmus-guidelines.github.io/.
Scribes of the Cairo Geniza. www.scribesofthecairogeniza.org/.
TAPAS Project. https://tapasproject.org/.
Teklia. The Horae Project. https://teklia.com/research/projects/horae/details/.
University of Pennsylvania. OPenn: Primary Digital Resources Available to Everyone. https://www.openn.library.upenn.edu/.
Vierthaler, Paul. "NYU Abu Dhabi Stylometry." https://github.com/vierth/nyuabudhabi.
Virtual Hill Museum and Manuscript Library. Online Resources for the Study of Manuscript Cultures. www.vhmml.org/.
Weil, Stefan. "HTR Model for German Manuscripts Trained from Several Datasets." Zenodo, 2023. https://doi.org/10.5281/zenodo.7933463.
Zooniverse. www.zooniverse.org/.

INDEX

DOI: 10.17302/mfsk158712