

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."

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 Manuscript 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*.