

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

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 Foy, “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 IIRF 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 “finger-print” of the scribes of the tradition, while also paying careful attention to material details across this corpus.

¹⁸ See Boillet et al., “HORAÉ”; 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 Lemerrier 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 Lemerrier 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 Lemerrier 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 Derolez⁵²), 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, Fiddymment 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”; Reinhout, “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 Fiddymment 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, Fiddymment 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 Fiddymment 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
Cologne, Fondation Martin Bodmer (hereafter Cologne), 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

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 (Cologne, MS Cod. 28, Göttingen 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?”