An Information System for Biblical Manuscripts Paratexts: Modeling, Implementation, and Future Directions

Andrea Brunello[1], Emanuela Colombi[1], Matteo Raffin[2] and Nicola Saccomanno[2,*]

[1]Department of Humanities and Cultural Heritage, University of Udine, Palazzo Caiselli, Vicolo Florio 2, 33100 Udine, Italy
[2]Department of Mathematics, Computer Science, and Physics, University of Udine, Via delle Scienze 206, 33100 Udine, Italy

Abstract

Paratexts—such as prologues, summaries, prefaces, and annotations—shape the presentation, interpretation, and transmission of texts across audiences and periods. Their study provides critical insights into the historical, philological, and socio-cultural dimensions of manuscript production, use, and dissemination. Yet, a comprehensive analysis of Latin biblical paratexts remains lacking despite notable efforts on specific subsets, such as Marilena Maniaci's researches on Atlantic Bibles and Chiara Ruzzier's studies on 13th-century portable Bibles. This article takes part at addressing such a gap presenting an information system for managing paratexts in medieval Latin biblical manuscripts. Our contribution is twofold: (1) we propose a conceptual model of the domain of medieval Latin biblical manuscripts paratexts to standardize the field and support future research; and (2) we implement such a model through a relational database, which acts as the core of an information system for documenting and analyzing paratexts. Its open access prototype, already available, facilitates data organization and analysis, enabling prospective advanced applications, including artificial intelligence techniques.

Keywords

Biblical manuscripts, Paratexts, Conceptual modeling, Relational databases

1. Introduction

When we think of the Bible, we often envision The Book for excellence, shaped by centuries of interpretation as a singular manifestation of divine word. However, its etymology, derived from the Greek *ta biblia* ("the books"), points to a plural textuality, while the history of biblical translations–from Hebrew (for what we call the Old Testament) into Greek, and subsequently into other ancient languages, such as Latin, and eventually into various modern languages–reveals transformations that challenge this presumed unity. These transformations result from the adaptation of the text to new historical and cultural contexts, as well as misunderstandings or textual innovations linked to the manuscript copying process.

In this work, we address the definition and the current and prospective development of an information system, with a relational database at its core, designed for the management of these textual discontinuities within Latin biblical manuscripts, focusing on two key aspects: *(i)* the structure of their content and *(ii)* specific types of paratexts.

As for the first point, the structural analysis of manuscripts reveals that not all exemplars contain the same biblical books in the same order, as the official Catholic canon was only definitively established at the Council of Trent in 1546. The manuscript tradition under investigation reflects the tension between a "short" canon, aligned with the Hebrew Bible and supported by Jerome in line with his preference for the Hebraica veritas, and a "long" canon, corresponding to the Greek Septuagint and earlier Latin translations. By the 9th century, surviving manuscript evidence shows a dual development: the gradual establishment of Jerome's revision and his canonical vision, and the rise of single–or two–volume Bibles (*pandectae*), requiring preliminary decisions about content and order. The inclusion of

IRCDL 2025: 21st Conference on Information and Research Science Connecting to Digital and Library Science, February 20-21 2025, Udine, Italy

*Corresponding author.
andrea.brunello@uniud.it (A. Brunello); emanuela.colombi@uniud.it (E. Colombi); raffin.matteo@spes.uniud.it (M. Raffin); nicola.saccomanno@uniud.it (N. Saccomanno)
0000-0003-2063-218X (A. Brunello); 0000-0002-0384-6664 (E. Colombi); 0000-0001-5916-3195 (N. Saccomanno)

variable deuterocanonical books led to hybrid editorial solutions, traceable in manuscripts from earlier and later periods, which the database seeks to document. This historical-theological aspect is closely tied to a codicological one: in Late Antiquity and the early Middle Ages, the Bible was primarily a bibliotheca–a collection of approximately ten codices, often with stable but somewhat variable contents. The juxtaposition or overlapping of transmission lines frequently caused absences or duplications, particularly for books with unstable placement, and these phenomena were exacerbated by the need to recover content from lost or damaged volumes. Over time, the pandectae gradually organized this diversity into a unified format, although not without challenges. For these reasons, as we will see, our proposed database records the books present in each exemplar and their order, enabling analyses of the transformations in the biblical canon as reflected in the manuscript evidence.

The second type of discontinuity addressed in the database concerns the paratexts of Latin biblical manuscripts, which can be broadly defined as elements that accompany the biblical text without strictly belonging to it, including prologues, chapter headings, initial and final titles, running headers, marginal numbering, and indications of book lengths. These elements are particularly suited to tracing the discontinuities of interest, as they tend to persist like "fossils" from one copy to another, even when they prove inconsistent or inadequate. While Gérard Genette's seminal works in the late 20th century defined paratexts as "thresholds" to the text characterized by authorial intention and responsibility [1, 2, 3], the study of medieval manuscript traditions requires adapting Genette's approach to reflect the material uniqueness of each exemplar. This point has been highlighted by the reflections of Patrick Andrist on the terminology and ontology of paratexts in manuscript traditions [4, 5, 6], which emerged from the developments of the ERC-funded Paratexts of the Bible project, dedicated to Greek biblical manuscripts [7, 8, 6]. Paratexts mediate the text for diverse audiences across time, offering insights into the cultural and social roles of the Bible, the interaction between the implied reader (*lector in fabula*) and the actual reader, and the contexts of manuscript production, transmission, and use.

This area of research, though increasingly prominent in recent decades, lacks a comprehensive study or multidisciplinary reflection for Latin biblical paratexts. Foundational contributions have emerged from the extensive research of Pierre-Maurice Bogaert [9, 10, 11, 12, 13, 14], the studies by Marilena Maniaci and Roberta Casavecchia on Atlantic and Beneventan Bibles [15, 16, 17, 18, 19, 20, 21], as well as from the research of María Adelaida Andrés Sanz on Spanish Bibles [22] and Chiara Ruzzier on 13th-century "Bibles portatives" [23], while the transmission of the Latin New Testament, including its paratextual aspects, has been explored by Hugh Houghton [24, 25]. However, a holistic study of the historical, philological, and socio-cultural significance of Latin biblical paratexts remains an important gap, particularly with regard to the books of the Old Testament. In this regard, the paratexts to the Octateuch in the Tours, Atlantic, and Beneventan Bibles are the subject of a recently funded PRIN PNRR project (see the acknowledgments section) to which the present research also belongs ([26]).

This article seeks to address this gap by presenting a twofold contribution. First, we propose a thorough modeling of the domain of medieval Latin biblical manuscripts paratexts, aiming to standardize the field and establish a foundational framework for future studies. Second, we actualize this modeling through an information system based on a relational database, specifically designed to document and analyze these paratexts. The database, which is already accessible in its prototype form, serves as a dynamic tool for organizing and querying data, laying the groundwork for future expansions and applications, including the usage of artificial intelligence techniques, such as Large Language Models (LLMs) [27], to facilitate the interaction, and advanced graphical user interfaces to allow both information retrieval and the addition of new material. The long-term objective is for the information system to become a key reference in the literature for this domain, potentially integrated with other sources (e.g., [28]).

In the literature, similar works to ours, though narrower in scope or focusing on a different domain, include the aforementioned database Paratexts of the Bible [7], which contains information on texts and paratexts found in Biblical manuscripts written in Greek. Currently, the dataset is primarily focused on Gospel books and is linked with a significant interoperability initiative to the Greek manuscript database Pinakes [29]. It is enriched with an extensive set of metadata, and the queries are based on the structure of Pinakes, but further search options are under development. Moreover, it will be essential to

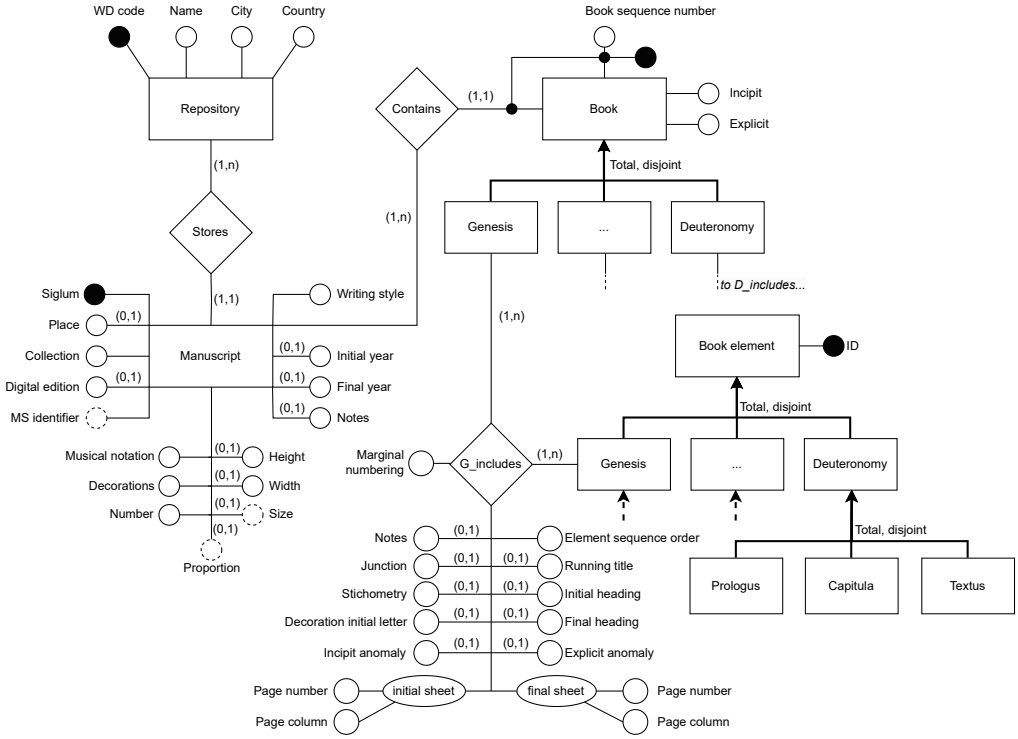

Figure 1: Conceptual Entity-Relationship diagram. For the sake of simplicity, we omitted to report the full list of book (element) types and the complete set of relationships linking them to their respective book elements.

establish connections with complementary projects, including, in particular, the tool under development for Biblical manuscripts at Ludwig-Maximilians-Universität in Munich [30]. Additional insights are also offered by the PASSIM database, recently published as the result of the ERC project, led by Shari Boodts at the Radboud University of Nijmegen, on the manuscript tradition of medieval homiliaries [31]. While these represent a different manuscript typology, the database serves as an interesting model for managing fluid and complex textual content. It also demonstrates the potential of using new data analysis tools to identify similarities and divergences in the organisation of textual units, offering a distinct perspective and a tailored infrastructure to advance research in this evolving area.

The paper is structured as follows: in Section 2, we present a conceptual modeling of the domain of medieval Latin biblical manuscripts; building on this foundational standpoint, Section 3 describes the design of a relational database, which serves as the core of an information system for managing these paratexts; Section 4 introduces the prototypical implementation of the database and provides usage examples to illustrate its core functionalities and to demonstrate how it can support research in the considered field. Finally, we conclude by evaluating the outcomes of this work and exploring directions for future developments.

2. Modeling Biblical manuscripts paratexts

The primary objective of our conceptual modeling is to document not only the biblical books contained in the analyzed manuscripts but also the sequence in which they are arranged, along with the accompanying prefatory material. These considerations, which form a central focus of our research, have guided the development of an entity-relationship (ER) diagram [32] centered on individual textual and paratextual units, shown in Figure 1. This approach allows for a finer level of granularity, beyond the level of the biblical book, enabling us to trace discontinuities and analyze the diverse ways these units are combined in individual manuscripts. Additionally, the model facilitates tracking of similarity relationships between manuscripts based on the presence, absence, and sequence of textual and paratextual units. In the following, we present the main elements composing the diagram. Note that, for the sake of brevity,

the diagram does not depict all entities and relationships. Instead, we present representative examples where the context makes them clear. For instance, we have omitted the full list of book (element) types and the complete set of relationships linking them to their respective book elements. Another consideration is that the ER diagram can be easily extended to accommodate additional information not currently tracked, should it become relevant (e.g., details about manuscript authors). Finally, although we are aware of the framework provided by the IFLA Library Reference Model [33], we used specific and immediately understandable domain names instead of the generic entity and relationship names provided by IFLA standards, though future reconciliations between the names we used and IFLA ones remain possible.

Manuscript. It represents the physical manuscript, which is stored in one and only one repository and may contain one or more biblical books. In the case of fragmentary manuscripts preserved in different locations, the distinct units are recorded separately, highlighting their connections. The *MS Identifier* is a derived value which combines the repository *WD code* (see the paragraph containing the description of the entity *Repository*), *Collection* (defaulting to "MS" if the shelfmark consists only of a number), and *Number*. The *Number* works with the collection to uniquely identify manuscripts, and *Digital edition* links to an online digital reproduction, if available (optional). A manuscript is identified by the *Siglum*, a unique identifier within our model and internal to the project, as no universal list of biblical manuscript sigla currently exists. Siglum is essentially a shortened version of the attribute *MS identifier*. The list of manuscripts considered and their assigned sigla will be shared alongside the publication of the research outputs, making them universally identifiable to the scholarly community. A manuscript is described by several attributes. The *Place* refers to where the manuscript was likely written or its earliest traceable location (optional). The *Initial year* and *Final year* indicate its production time range, while *Decorations* and *Musical notation* are true/false attributes indicating the presence of decorative elements or musical notation, respectively. *Writing style* specifies the script style; other optional attributes record the *Width* and *Height* of each manuscript, with the *Size* (the sum of width and height) and the *Proportion* (i.e., the relationship between width and height) calculated automatically. The latter is expressed as a decimal number, increasing as the page's shape approaches a perfect square (with a proportion of 1). These attributes are designed to provide an immediate visual impression of the manuscript and to facilitate quantitative codicological research. The attribute *Notes* provides additional details (optional).

Repository. It is the entity that represents the current physical location of the manuscript, typically a library or a conservation institution. Its key is the *WD code*, which consists of a unique alphanumeric code extracted from the Wikidata portal [34], enabling information interoperability. The other attributes are *Name*, *City*, and *Country*, which respectively represent the name of the institution, the city, and the country where it is located. A repository may store one or more manuscripts (relationship *Stores*).

Book. It represents a specific, "physical" biblical book within a manuscript. The modelling of this entity has proven to be crucial and has materialized in the concept of a book as a "container" encompassing both the biblical text itself and the accompanying paratexts, such as prologues and summaries. This approach enables a more comprehensive and historically accurate perspective on the manuscript: what are commonly referred to as biblical books are, in fact, the biblical texts, to which each exemplar adds further materials (elements) that belong to it and shape its interpretation. Each book belongs to one and only one predefined kind, representing its title (e.g., Genesis), following the form established in the critical edition of the Vulgate edited by Robert Weber and Roger Gryson [35]. Uniquely identifying a book requires knowing both the manuscript in which it is contained and its sequence within that manuscript (attribute *Book sequence number*). A preliminary note concerns the current limitation of the census to the books of the Old Testament: the database is, however, designed to accommodate future expansions to include New Testament books.

Book element. It represents a generic, "abstract" component of a biblical book, categorized into one of three types: prologue, summary, or text. Unlike "physical" book elements, which may include paratextual or decorative attributes (e.g., the specific decorations or headings), the book element is defined solely by its textual content as standardised by the reference repertoires or edition. This abstraction allows for a clear distinction between the abstract, conventional textual content of a book element and its various physical representations (see relationship *Includes)*, which may differ in their physical characteristics, such as text anomalies or varying paratexts, but share the same textual content identity. Each book element is uniquely identified by an *ID* based on its type and established domain references. For prologues, the *ID* corresponds to the numbering system of Friedrich Stegmüller's *Repertorium Biblicum* [36]. For example, two distinct prologues for Genesis would each have a unique *ID*, distinguishing them. Conversely, a prologue with the same *ID* appearing in multiple books represents the same shared textual content. Summaries (*capitula*) follow the classification system proposed by Donatien de Bruyne [37], where *ID*s consist of series defined by letters or sigla, to which we have added the abbreviation (as defined in the Weber-Gryson edition) of the associated biblical book (e.g., *A_Gn* for Genesis summaries in series *A*). This structure facilitates vertical searches across books, enabling the identification of series spanning multiple biblical books, while simultaneously allowing for the recognition of the type of summary prefixed to the same biblical book in different exemplars. The respective incipits and explicits have also been added as attributes for each element, using the standardized forms published in the Weber-Gryson edition for biblical texts, Stegmüller's repertoire for prologues, and de Bruyne's edition for summaries.[1] It was also necessary to establish predefined conventions for prologues and summaries: when these refer to groups of books (e.g., the prologue to the Pentateuch or to the collection of the Twelve Prophets), the element is "linked" to the first book of the group (in these cases, Genesis and Hosea, respectively).

Includes. These relationships (one for each kind of *Book*/*Book element*) track the many-to-many association between a specific manuscript's book and its elements. We have defined that a book can contain multiple (typically, up to five) prologues, at most one summary, and exactly one text. Books may also exist as texts without prologues and/or summaries. While the *book element* entity represents an abstract component of a book, defined by its type and textual content (via the *ID* attribute), linking a *book element* to a *book* "materializes" it. This connection captures the attributes of the specific physical instance of the book element as it appears in a particular manuscript's book. These attributes include the *Element sequence order*, which specifies the order of the element within the book. The text always appears as the last element, while prologues and summaries can be arranged in any order. The *Initial sheet* and *Final sheet* indicate where the element begins and ends in the manuscript, defined by page numbers and column markers (e.g., "ra" for recto-column a, "vb" for verso-column b). The *Initial heading* and *Final heading* represent the opening and closing headings of the element, with manuscript's abbreviations expanded using mixed-case letters to improve searchability. The primary goal is to facilitate the identification of discontinuities that have been largely underexplored but can provide valuable insights into relationships between codices. Moreover, headings tend to crystallize during transmission, sometimes offering crucial information about earlier layers and the convergence of different traditions in the titles used to identify both biblical texts and paratexts (e.g., prologus, praefatio, argumentum, capitula, brevis, capitulatio, etc.). These attributes can also be marked as "om." if they have been omitted for undetermined reasons, or as "om. lac." in cases where the omission is due to physical damage (material lacuna). The *Running title* is an optional attribute that records the possible book title as written in the upper margin. Similarly, the *Decoration initial letter* is an optional attribute indicating the presence of a decorated initial letter, which may also be marked as "om." and/or "lac." if missing or damaged. *Stichometry* is another optional attribute that records, in Arabic numerals, the number of lines declared at the end of the text in certain manuscripts; the original form in Roman numerals is instead recorded as part of the final heading. Additional optional attributes include *Incipit anomaly*

[1] An additional enhancement that could provide a valuable service to database users would be the inclusion of the full text of prologues and summaries, though this would require verification regarding reproduction rights.

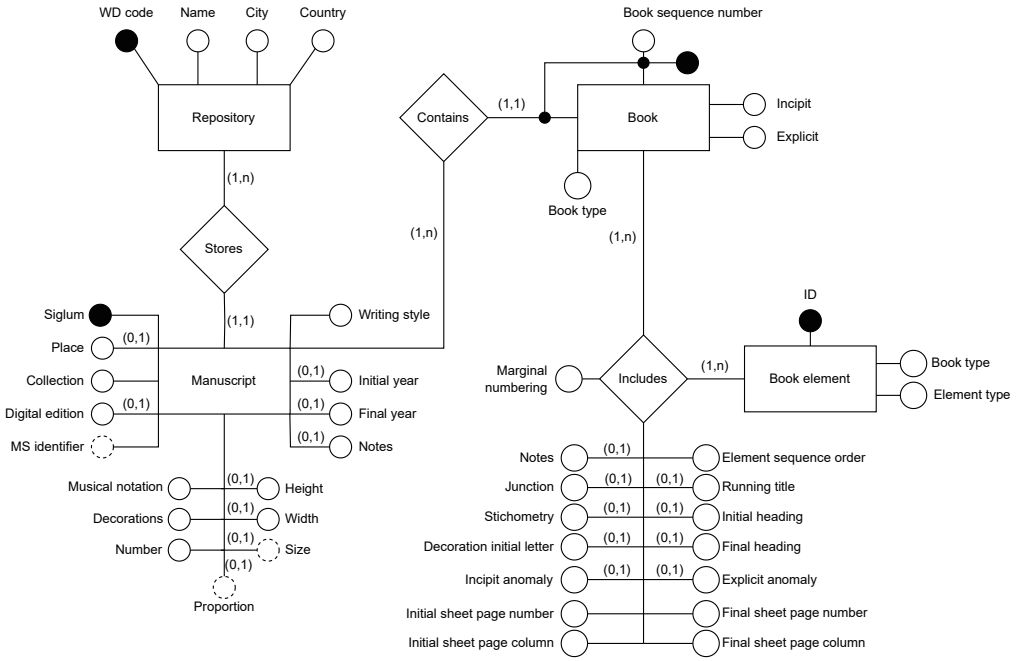

Figure 2: Restructured Entity-Relationship diagram.

and *Explicit anomaly*, which record divergences in the beginning or end of the element compared to the standard reference text recorded in *Book Element*. Priority is given to the separate recording of all units provided with an ID: for instance, in the not uncommon case where a prologue is composed of the consecutive transcription of multiple prologues, presented as a single text, each textual unit is recorded separately, with the *Notes* field specifying that there is no break between these elements. Other anomalies, such as a total number of chapters differing from that indicated for a specific summary in de Bruyne's edition, are documented in the *Notes* field. The attribute *Marginal numbering* records the presence of the marginal "capitulation" throughout the text and whether it corresponds to the initial summaries, with the letter K indicating the presence of chapter headings interspersed within the text as subtitles. The *Junction* attribute specifies whether the element ends at a gathering's junction, with possible values including (other than `nu11`): "(x)" caesura with blank spaces, "?" possible caesura, "(?)" possible caesura with blank spaces, "/" confirmed caesura that includes the following incipit, "/?" possible caesura that includes the following incipit, "(/)" caesura that includes the following incipit and blank spaces, "(/?)" possible caesura that includes the following incipit and blank spaces, and "B" anomalous blank spaces at the end of a text. Finally, the *Notes* attribute is an optional textual field for providing additional information about the element.

3. Relational database development

In this section, we present the structure of the relational database system that implements the ER diagram of Figure 1. We chose to rely on a relational database rather than a NoSQL solution, such as a graph database, because our data is highly structured and relational databases offer fine-grained control over data consistency. To define the database schema, we first restructured the ER diagram, removing elements that could not be directly mapped to a relational schema—specifically, generalizations and composite attributes in our case. The result is shown in Figure 2.

Notably, we replaced the composite attributes in all *Includes* relationships with their individual components. Next, we chose to retain the derived attributes *MS identifier*, *Size* and *Proportion* of the *Manuscript* entity. Finally, we removed the *Book* and *Book element* specializations, replacing them with

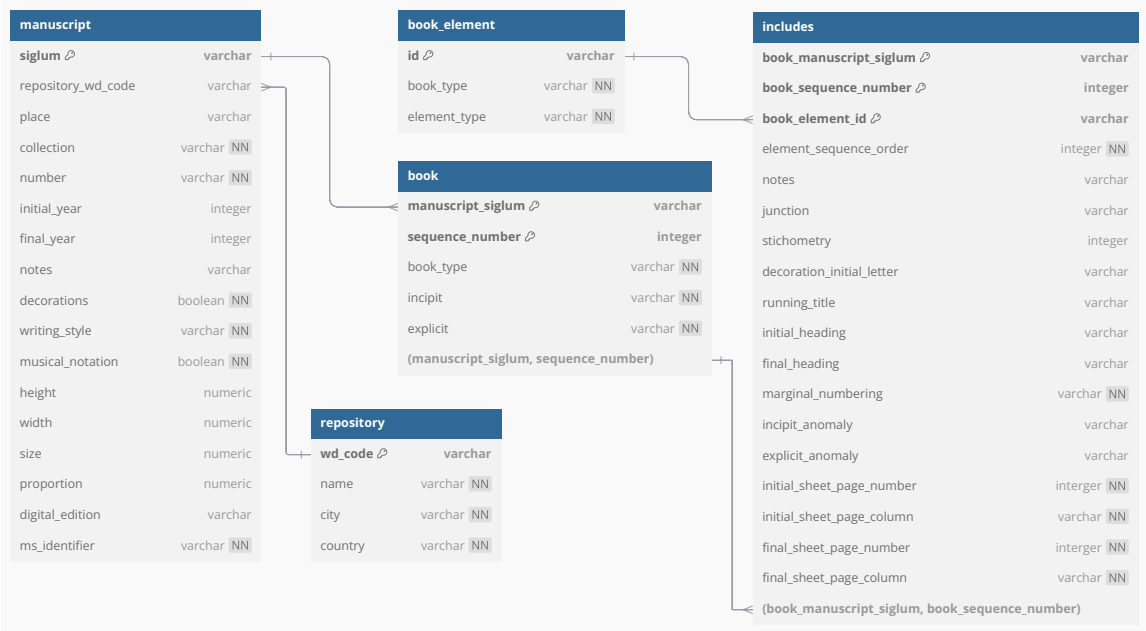

Figure 3: Logical relational schema.

attributes within the parent entities to preserve their distinctions.[2] This adjustment also allowed us to consolidate all the book-specific includes relationships into a single *Includes* relationship between *Book* and *Book element*. Upon closer examination of the data at our disposal, this proved to be indeed the correct choice, as there are cases where a book element inherently associated with a specific book type (e.g., a Genesis Prologue) was attached to a different book type (e.g., Deuteronomy) due to errors in the manuscript's original assembly or its subsequent preservation. While the original ER diagram represents the ideal scenario, our adjustment accommodates such discrepancies and allows to keep track of them.

Finally, according to a set of well-established mapping rules [32], we derived the database logical schema from the restructured diagram shown in Figure 2. The resulting relational schema, presented in Figure 3, illustrates the tables, attributes (along with their data types), and relationships that form the backbone of the database. This schema not only serves as a bridge between the conceptual design and the physical database implementation but also provides the database user with a clear map to guide their interaction with the system, ensuring a better understanding of its structure and functionality.

4. Prototype of the system

A prototypical version of the relational database, implemented in PostgreSQL [38], and which serves as the core of the prospective information system, is already freely accessible.[3]

The source code of our implementation, including the SQL Data Definition Language (DDL) for deploying the database, the raw data underlying the database instance (which is continuously expanding), the database data import script, and the definitions of several useful SQL queries, will be made available on the project's GitHub page [40] upon paper acceptance.

As of the date of article submission, the database–continuously growing alongside the raw data–contains 23 repositories, 48 manuscripts, 1303 books, 251 book elements, and 2567 instances of the

[2]Specifically, for the *Book element*, in addition to the type of the book (e.g., Genesis), encoded by the *Book type* attribute, we include the *Element type* attribute, which can take the values T (text), P (prologue), and C (summary).

[3]The system is accessible at http://158.110.146.222:8080/. Upon connecting, users are presented with a pgAdmin [39] web server interface that prompts for login credentials (username = `tester_biblical@ai4ch.uniud.it`, password = `UXftJGM5eNMdPGZ`). A read-only user account grants privileges to perform select operations on the `public` schema of the `biblical` database.

Includes relationship.

4.1. Exemplary interactions

The database supports a wide range of interactions, from basic queries to more advanced use cases. Below, we report some notable examples of SQL queries that can be directly run against the prototypical online implementation of the database. SQL user defined functions will be implemented to simplify user interaction.

Determine all summaries of Genesis. This type of query allows the identification of the various types of introductory paratexts (prologues and chapter headings) associated with the same biblical book, allowing for an examination of the diversity of editorial arrangements across different manuscripts.

Query & results: All summaries of Genesis

```
select id as book_element_id
from book_element
where book_type='Genesis' and element_type='C';
```

book_element_id
A_Gn
B_Gn
C_Gn
…

Determine all summaries from series A. Such queries allow for the cross-sectional verification of the presence of capitula across the various biblical books identified by De Bruyne [37] with the same letter. The search can be restricted to a single biblical book by specifying its reference abbreviation in the query (e.g., _Gn) according to the conventions of the Weber-Gryson edition [35], or by filtering on the attribute *book_type*.

Query & results: All summaries from series A

```
select id as book_element_id, book_type
from book_element
where element_type='C' and id like 'A\_%';
```

book_element_id	book_type
A_1 Mcc	1 Macchabeorum
A_1 Sm_1 Rg	1 Samuhel_1 Regum
A_2 Mcc	2 Macchabeorum
…	…

Determine which prologues are associated with books named Genesis, and their respective frequencies. This query provides a ranked view of the distribution of paratexts (prologues or chapter headings) for a specific biblical book, emphasizing the relative prevalence of distinct editorial choices within the manuscript tradition.

Query & results: Prologues associated with books named Genesis

```
with
  gbooks as (
    select *
    from book
    where book.book_type = 'Genesis'
  ),
  tmp_res as (
    select
      book_element.id as prologue_id
    , count(*) as asbsolute_occurrences
    from gbooks
         join includes on (gbooks.manuscript_siglum = includes.book_manuscript_siglum
                           and gbooks.sequence_number = includes.book_sequence_number)
         join book_element on includes.book_element_id = book_element.id
    where book_element.element_type = 'P'
    group by book_element.id
  )
select
   tmp_res.*
,  round(tmp_res.asbsolute_occurrences/(select sum(asbsolute_occurrences)
                                        from tmp_res),2) as rel_occurences
from tmp_res
order by rel_occurences desc;
```

prologue_id	asbsolute_occurrences	rel_occurrences
285	27	0.59
284	14	0.30
290	2	0.04
…	…	…

Determine the initial and final headings of the book Canticum canticorum for all manuscripts in which it is present. This type of query allows for a synoptic visualization of the initial and final headings of a specific biblical book, enabling the identification of discontinuities and potential affinities, particularly in the case of more elaborate headings than the standard formula (incipit liber…explicit liber). These may contain valuable information; for instance, benevolent formulas (such as Deo gratias or Amen) could point to an earlier exemplar where such formulas marked the beginning or end of an independent volume, which was later incorporated into a larger or differently composed collection. The same query can be adapted to analyze the initial and final headings of both prologues and capitula.

Query & results: Initial and final headings of the book `Canticum canticorum`

```
select
    I.book_manuscript_siglum as manuscript_siglum
,   I.initial_heading
,   I.final_heading
from includes I
       join book B on (I.book_manuscript_siglum=B.manuscript_siglum
                       and I.book_sequence_number=B.sequence_number)
where B.book_type='Canticum canticorum';
```

manuscript_siglum	initial_heading	final_heading
Am5	INCIPIUNT CANTICA …	EXPLICIUNT CANTICA …
Amt	INCIPIT LIBER…	EXPLICIT LIBER …
An2	INCIPIUNT CANTICA …	EXPLICIUNT CANTICA …
…	…	…

Determine the relative order in which the books '1 Ezras,' '2 Ezras (Neemia),' 'Iudith,' 'Hester,' and 'Tobias' are presented in the manuscripts, considering only the manuscripts that contain at least one of these books. This more complex type of query addresses one of the fundamental research questions outlined earlier: the relationship between the canon expressed by individual biblical manuscripts and their arrangement, including codicological aspects, whether they are bibliothecae or pandectae, incorporating earlier partial collections. This aligns with reflections on the 'modular' structure of Atlantic Bibles discussed by [15] (pp. 54–56) and the requirements outlined by [30].

The example provided focuses on a group of biblical books characterized by significant instability in their presence and relative order. This instability can be traced back to Jerome's editorial project for the Vulgate and his adherence to the Hebraica veritas. Tobit, Judith, and parts of Esther belong, in fact, to the so-called deuterocanonical books, included in the Greek Septuagint translation but absent from the Hebrew canon. However, Jerome agreed to translate them, indirectly validating their inclusion in Vulgate manuscripts. Nonetheless, their sequence oscillates [41], which also impacts the complex dossier of the book of Ezra [42, 43].

Query & results: Books relative order

```
select
   B.manuscript_siglum as man_siglum
,  row_number() over (partition by B.manuscript_siglum
                      order by B.sequence_number) as rel_order
,  B.book_type
,  min(initial_sheet_page_number) as i_sheet
,  min(initial_sheet_page_column) as i_column
,  max(final_sheet_page_number) as f_sheet
,  max(final_sheet_page_column) as f_column
from book B
      join includes on (B.manuscript_siglum =includes.book_manuscript_siglum
                        and B.sequence_number =  includes.book_sequence_number)
where B.book_type in ('1 Ezras', '2 Ezras (Neemia)', 'Iudith', 'Hester', 'Tobias')
group by B.manuscript_siglum, B.sequence_number;
```

man_siglum	rel_order	book_type	i_sheet	i_column	f_sheet	f_column
Amt	1	Tobias	701	va	708	vb
Amt	2	Iudith	708	ra	720	rb
Amt	3	Hester	720	ra	730	rb
…	…	…	…	…	…	…

Determine the difference between two manuscripts in terms of book ordering, considering only the books they have in common. We focus on a specific pair of manuscripts, identified by "Amt" and "Sg1." First, we retrieve all books they have in common, beginning with those in "Amt." For each shared book, we calculate its relative order of appearance within only the shared ones, and we also report its complete sequence within the manuscript. We then repeat this procedure for the "Sg1"

manuscript. Next, we pinpoint all cases where the two manuscripts present different books occupying the same relative position in their respective sequences. In each such case, we list the book as it appears in "Amt" and provide its full order of appearance, comparing it to its full order in "Sg1." On the same row, we also record the counterpart book that "Sg1" places in the same relative position, along with the location of that book in the "Amt" manuscript. For the two manuscripts considered, from the query we obtain 17 rows over a total number of shared books of 32. Note that, starting from a similar query, it is possible to calculate an "index of diversity" between manuscripts, for instance, drawing inspiration from the Kendall tau rank distance [44], which represents the number of element swaps needed to transform one list into another.

Query & results: Manuscripts difference in terms of book ordering

```
with
  manuscript_1 as (
    select
      book_type
    , row_number() over (partition by manuscript_siglum
                         order by sequence_number) as rel_order
    , sequence_number as book_order
    from book
    where manuscript_siglum = 'Amt'
          and book_type IN (select book_type
                            from book
                            where manuscript_siglum = 'Sg1')
  ),
  manuscript_2 as (
    select
      book_type
    , row_number() over (partition by manuscript_siglum
                         order by sequence_number) as rel_order
    , sequence_number as book_order
    from book
    where manuscript_siglum = 'Sg1'
          and book_type IN (select book_type
                            from book
                            where manuscript_siglum = 'Amt')
  )
select
   m1.book_type as b1_m1
,  m1.book_order as b1_m1_order
,  m1b.book_order as b1_m2_order
,  m2.book_type as b2_m2
,  m2.book_order as b2_m2_order
,  m2b.book_order as b2_m1_order
from manuscript_1 m1
        join manuscript_2 m2 on (m2.rel_order = m1.rel_order
                                 and m1.book_type < m2.book_type)
        join manuscript_2 m1b on m1b.book_type = m1.book_type
        join manuscript_1 m2b on m2b.book_type = m2.book_type;
```

b1_m1	b1_m1_order	b1_m2_order	b2_m2	b2_m2_order	b2_m1_order
Ionas	31	22	Psalmi	31	15
Micha	32	23	Psalmus CLI	32	16
Naum	33	24	Proverbia	33	17
…	…	…	…	…	…

4.2. Broader research questions and interactions supported by the system

The previous exemplary queries illustrate how the system enables users to efficiently extract information about texts and paratexts. This functionality supports diverse research endeavors, including the study of transformations in the biblical canon, which are evident, for example, both in the selection of books included in a manuscript and in the order in which those books are arranged. Furthermore, the system can provide support for the automated content similarity evaluations across a large number of manuscripts, providing new opportunities to investigate the historical and codicological factors that shape the arrangement of books within exemplars. In the following, we outline more in detail some research questions and interactions that can be effectively addressed using the system.

Tracking textual and paratextual features: Documenting the sequence and location of (para)textual elements within individual manuscripts and recording the presence or absence of decorative elements, such as initials or other embellishments. This includes examining initial and final headings for patterns in textual transmission, including formulaic incipit and explicit expressions. **Stichometric and quantitative textual analysis:** The inclusion of data such as the recording of *stichometry*, presence of *Decoration initial letter* (as a boolean attribute), and *Junction* (with symbols differentiated by the degree of certainty and the presence of blank spaces) facilitates the adoption of computational processing and quantitative analysis. **Cross-manuscript comparison:** Facilitating the discovery of relationships between (groups of) manuscripts through shared features in headings, capitula, and structural design, to study the divergence and convergence in traditions. Advanced tools, including tokenization and semantic comparison of headings and textual strings, may also enable similarity computation, for instance by means of machine-learning models or string-embedding techniques. **Full-text search and annotations:** Supporting full-text search capabilities to identify anomalies in headings, numbering, or other textual elements. **Textual transmission and scribal practices:** Investigating how formulaic incipit and explicit expressions evolve across manuscript traditions and how scribes exercise autonomy in reproducing or adapting paratextual elements, offering insights into the historical dynamics of manuscript preparation and adaptation. **Codicological studies:** Studying structural features, such as junctions between textual elements and manuscript gatherings, as well as the size and proportions of manuscripts, provides valuable insights into codex production across different times and regions. This approach supports the identification of distinctive features within various manuscript traditions and facilitates the application of quantitative codicological analysis.

5. Conclusions and future developments

In this work, we presented the conceptualization of the domain of biblical manuscripts paratexts, followed by the design and development of a relational database system for managing such kind of data. The conceptualization effort not only guided the system's development but also contributes to the standardization of the field, laying a robust foundation for future studies. While this is an ongoing project, a prototype of the relational database—intended to serve as the core of a comprehensive information system—is already freely accessible online.

As for future work, in addition to the research avenues already discussed, plans include the development of a graphical user interface (GUI) for interacting with the database. This interface will be further enhanced by incorporating artificial intelligence techniques, such as large language models (LLMs), to enable natural interaction with the stored data. For instance, users could query the system using natural language or engage in conversational interactions to explore the data more intuitively, leveraging text-to-SQL systems [45].

Acknowledgments

The research is part of the DOBiPS – Data Oriented Biblical Paratexts Studies project, awarded for the 2023–2025 biennium to the research units of the Universities of Udine (P.I. E. Colombi) and Cassino

(P.I. R. Casavecchia) under the competitive PRIN PNRR call – National Recovery and Resilience Plan, Mission 4 Education and Research, funded by the European Union Next-GenerationEU (protocol no. P2022ZW4AW). Nicola Saccomanno also acknowledges the support from the Interconnected Nord-Est Innovation Ecosystem (iNEST), which received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.5 – D.D. 1058 23/06/2022, ECS00000043).

References

[1] G. Genette, Introduction à l'architexte, Paris: Seuil, 1979.
[2] G. Genette, Palimpsestes. La littérature au second degré, Paris: Seuil, 1982.
[3] G. Genette, Seuils, Paris: Seuil, 1987.
[4] P. Andrist, Toward a definition of paratexts and paratextuality: The case of ancient Greek manuscripts, De Gruyter, Berlin, Boston, 2018, pp. 130–150. doi:10.1515/9783110603477-010.
[5] P. Andrist, The limits of paratexts/paracontents in manuscripts: Revisiting old questions and posing new ones, COMSt Bulletin 8 (2022) 215–233.
[6] P. Andrist, Asymmetrical descriptions of biblical manuscripts: A key to the success of the Paratexts of the Bible (ParaTexBib) project and its database, Bibliothek und Wissenschaft 55 (2022) 63–78.
[7] M. Wallraff, et al., Manuscripta Biblica, 2020. URL: https://www.manuscripta-biblica.org, accessed: 2024-11-15.
[8] P. Andrist, M. Wallraff, Paratexts of the Bible: A new research project on Greek textual transmission, Early Christianity (2015) 237–243.
[9] P.-M. Bogaert, Les particularités éditoriales des Bibles comme exégèse implicite ou proposée. Les sommaires ou capitula donatistes, in: I. Iudaicum (Ed.), Lectures bibliques. Colloque du 11 nov. 1980, Publications de l'Institutum Iudaicum, Bruxelles, 1982, pp. 7–21.
[10] P.-M. Bogaert, La Bible latine des origines au moyen âge. Aperçu historique, état des questions, Revue Théologique de Louvain 19 (1988) 137–159; 276–314.
[11] P.-M. Bogaert, Aux origines de la fixation du canon. Scriptoria, listes et titres. Le Vaticanus et la stichométrie de Mommsen, in: J.-M. Auwers, H. D. Jonge (Eds.), The Biblical Canons, Louvain, 2003, pp. 153–176.
[12] P.-M. Bogaert, The Latin Bible, c. 600 to c. 900, in: R. Marsden, E. A. Matter (Eds.), The New Cambridge History of the Bible, Volume 2, Cambridge University Press, Cambridge, 2012, pp. 69–92.
[13] P.-M. Bogaert, The Latin Bible, in: J. C. Paget, J. Schaper (Eds.), The New Cambridge History of the Bible, Volume 1, Cambridge University Press, Cambridge, 2013, pp. 505–526.
[14] P.-M. Bogaert, Entre canon(s) et textes bibliques. Que traduire?, Recherches de Science Religieuse 106 (2018) 53–71.
[15] M. Maniaci, La struttura delle Bibbie atlantiche, in: M. Maniaci, G. Orofino (Eds.), Le Bibbie atlantiche. Il Libro delle Scritture tra monumentalità e rappresentazione, Catalogo della mostra, Milano, 2000, pp. 47–60. P. 47.
[16] M. Maniaci, Written evidence in the Italian Giant Bibles: Around and beyond the sacred text, in: L. I. Lied, M. Maniaci (Eds.), Tracing Annotations and Annotation Practices in Late Antique and Medieval Biblical Manuscripts, De Gruyter, Berlin, Boston, 2018, pp. 85–100. URL: https://doi.org/10.1515/9783110603477-006. doi:10.1515/9783110603477-006.
[17] M. Maniaci, The structure of Atlantic Bibles, Trends in Statistical Codicology (2021). URL: https://api.semanticscholar.org/CorpusID:243901180.
[18] M. Maniaci, Chapter lists in Giant and Beneventan Bibles: Some preliminary remarks, in: T. Bernheimer, R. Vollandt (Eds.), Synopses and Lists. Textual Practices in the Pre-Modern World, Cambridge, 2023, pp. 295–297.
[19] R. Casavecchia, M. Maniaci, G. Orofino, Montecassino e la Bibbia. Forme, contenuti, decorazione, in: La Bibbia a Montecassino / The Bible at Montecassino, Turnhout, 2021, pp. 12–68.

[20] R. Casavecchia, Bibbia e paratesti a Montecassino: I capitula al libro della Genesi, Scripta. An International Journal of Codicology and Palaeography 16 (2023) 61–94.
[21] R. Casavecchia, M. Maniaci, Partial Bibles in southern Italy: The case of Montecassino, in: P. Andrist, E. Attia, M. Maniaci (Eds.), From the Thames to the Euphrates. Intersecting Perspectives on Greek, Latin and Hebrew Bibles / De la Tamise à l'Euphrate. Regards croisés sur les Bibles grecques, latines, et hébraïques, volume 9 of *Manuscripta Biblica*, De Gruyter, Berlin, 2023, pp. 83–102.
[22] A. Sanz, M. Adelaida, Les préfaces de la Bible latine dans le haut Moyen âge hispanique, Annuaire de l'École pratique des hautes études (EPHE), Section des sciences historiques et philologiques. Résumés des conférences et travaux (2019) 205–221.
[23] C. Ruzzier, Entre Université et ordres mendiants: La production des bibles portatives latines au XIIIe siècle, volume 8 of *Manuscripta Biblica*, De Gruyter, Berlin, 2022.
[24] H. A. Houghton, Chapter divisions, capitula lists, and the Old Latin versions of John, Revue bénédictine 121 (2011) 316–356.
[25] H. A. G. Houghton, The Latin New Testament: A Guide to its Early History, Texts, and Manuscripts, Oxford University Press, Oxford, 2016. URL: https://global.oup.com/academic/product/the-latin-new-testament-9780198744733. doi:10.1093/acprof:oso/9780198744733.001.0001.
[26] R. Casavecchia, E. Colombi, M. Maniaci, A. Peri, La ricerca del Progetto DOBiPS - Data Oriented Biblical Paratext Studies, Brepols Publishers, Paratext Studies series, 2025, p. in press.
[27] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao, Large language models: A survey, arXiv preprint arXiv:2402.06196 (2024).
[28] R. Astell, Biblissima, Digital Philology: A Journal of Medieval Cultures 10 (2021) 331–334.
[29] IRHT, Pinakes, 2024. URL: https://pinakes.irht.cnrs.fr/, accessed: 2024-11-20.
[30] P. Andrist, T. Englmeier, S. Dirkse, New digital strategies for creating and comparing the content structure of biblical manuscripts, Journal of Data Mining & Digital Humanities (2023). doi:10.46298/jdmdh.10981.
[31] S. Boodts, G. Schmidt, R. Macchioro, I. Denis, M. Rempt, E. Komen, T. Hermsen, PASSIM research tool, 2024. URL: https://passim.rich.ru.nl/, accessed: 2024-09-15.
[32] P. Atzeni, S. Ceri, S. Paraboschi, R. Torlone, Database systems: Concepts, languages & architectures, McGraw-Hill, 1999.
[33] P. Riva, P. Le Boeuf, M. Žumer, IFLA Library Reference Model: A Conceptual Model for Bibliographic Information, Technical Report, International Federation of Library Associations and Institutions (IFLA), 2018.
[34] Wikidata contributors, Wikidata: The free knowledge base, 2024. URL: https://www.wikidata.org/wiki/Wikidata:Main_Page, accessed: 2024-10-01.
[35] R. Weber, R. Gryson, Biblia sacra iuxta Vulgatam versionem. Editio quinta, Stuttgart: Deutsche Bibelgesellschaft, 2007.
[36] F. Stegmüller, Repertorium Biblicum Medii Aevi, 1. Initia biblica. Apocrypha. Prologi, Consejo Superior de Investigaciones Cientificas, Madrid, 1950.
[37] D. de Bruyne, Sommaires, divisions et rubriques de la Bible latine, A. Godenne, Namur, 1914. Reprinted as *Summaries, Divisions and Rubrics of the Latin Bible*, with introductions by Pierre-Maurice Bogaert and Thomas O'Loughlin, Turnhout: Brepols, 2015.
[38] P. G. D. Group, PostgreSQL, 2022. URL: https://www.postgresql.org/, accessed: 2024-11-01.
[39] pgAdmin Development Team, pgAdmin, 2022. URL: https://www.pgadmin.org/, accessed: 2024-10-03.
[40] A. Brunello, E. Colombi, M. Raffin, N. Saccomanno, GitHub page of the relational database for medieval Latin biblical manuscripts project, 2024. URL: https://github.com/dslab-uniud/Database-biblical-manuscripts, accessed: 2024-09-01.
[41] P.-M. Bogaert, Tobie, Esther et Judith dans la stichométrie de Mommsen, in: Miscellanea codicologica F. Masai dicata MCMLXXIX, volume II, Story-Scientia: Gand, 1979, pp. 545–550.
[42] P.-M. Bogaert, Les livres d'Esdras et leur numérotation dans l'histoire du canon de la Bible latine,

Revue Bénédictine 110 (2000) 5–26.
[43] M. Morard, Bibliotheca Sacra. les variations des livres d'Esdras dans la Bible latine, in: Sacra Pagina, IRHT-CNRS, 2024. Consultation du 21/11/2024. https://gloss-e.irht.cnrs.fr/php/page.php?id=182.
[44] R. Sedgewick, K. Wayne, Algorithms, 4th ed., Addison-Wesley Professional, 2011.
[45] G. Katsogiannis-Meimarakis, G. Koutrika, A survey on deep learning approaches for text-to-SQL, The VLDB Journal 32 (2023) 905–936.