Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI
A. Salatino et al. (Eds.)

doi:10.3233/SSW240032

An Ontological Framework for Integrating the Heterogeneous Medieval Manuscript Resources: A Case Study of Progetto Irnerio and Mosaico

Faria FEROOZ [a,1], Monica PALMIRANI [a]
[a] *CIRSFID-ALMA-AI, University of Bologna, Itlay*
ORCiD ID: Faria Ferooz https://orcid.org/0000-0002-8150-7852, Monica Palmirani https://orcid.org/0000-0002-8557-8084

Abstract.

”Purpose: This study aimed to improve the organization and integration of heterogeneous medieval manuscript data across the Mosaico and Progetto Irneiro platforms. It addresses the challenges, such as the need for more standardization in data formats, metadata schemas, and inconsistent data quality, by developing a new ontology that supports the multifaceted analysis of medieval manuscripts. This analysis includes factors such as the historical context, physical characteristics, textual information, and artistic features.

Methodology: The approach began with analyzing metadata on two platforms. The MeLOn methodology is used to develop the Medieval Manuscript Data Integration Ontology (MMDIO), extending the MeMO ontology with elements from other ontologies and creating new data classes and relationships. The ontology is visualized with Graffoo and modeled in Protégé. Data was extracted according to this schema, converted into RDF triples, and tested in GraphDB. LODE is used to publish the ontology. This process enhanced data integration for Mosaico and Irnerio.

Findings: The use of MMDIO substantially enhanced the organization, accessibility, and uniformity of metadata formats on both platforms. However, it can now handle complex queries and integrate multiple types of manuscript data to facilitate a more comprehensive and organized approach in medieval manuscript research.

Value: The proposed MMDIO framework advances digital humanities research by increasing data semantic richness and interoperability, establishing the foundation for future research in cultural heritage preservation. Moreover, it demonstrates the value of adapted frameworks for handling complex data environments in particular research fields.”

Keywords. Data Integration, Digital Resource Collection, Medieval Manuscript Data Integration Ontology (MMDIO), MeLOn Methodology, Heterogeneous Platforms, Semantic Richness

[1]Corresponding Author: Faria Ferooz, faria.ferooz2@unibo.it.

1. Introduction

Medieval manuscripts are a valuable part of our cultural heritage because of their diversity and richness, and preserving them is essential to maintaining the richness and complexity of our global heritage [1]. These manuscripts were handwritten in Europe during the Middle Ages; they served a variety of purposes, such as religious, literary, scientific, and legal texts[2]. Thus, as significant cultural treasures, manuscripts provide insights into the artistic, spiritual, and legal practices of their time, offering a glimpse of the intellectual, social, and political history of the Middle Ages [3,4,5,6].

Considering their importance, academics from diverse fields study medieval manuscripts to evaluate their material, artistic, historical, and textual features. Iconographers investigate decorative elements, such as illumination and initials, to learn about medieval art, while textual scholars focus on the emergence of texts. Historians examine glosses and marginalia for medieval interpretations, whereas paleographers and codicologists concentrate on the physical characteristics of the codex. Furthermore, Legal experts also record the formation of legal theory from the Roman era to modern foundations [7,8,9].

The development of digital technologies has significantly improved the study of medieval manuscripts by making these materials globally accessible for analysis [10,11]. The way academics engage with medieval manuscripts has been completely transformed by digital libraries, which present previously unexplored possibilities for interdisciplinary study and novel data-driven strategies. A huge amount of information about medieval manuscripts has been made available by these advancements, including digitized images, transcriptions, metadata, and academic publications [12,13,14,15].

However, several specific challenges are associated with integrating heterogeneous medieval manuscript data. These include a lack of uniformity in the data formats and metadata schemas, as well as inconsistent data quality. This study intends to deal with a large collection of medieval manuscript data from two different platforms, Progetto Irnerio and Mosaico to meet the requirements of different types of scholars. These manuscripts are related to the fact that they were created at different times and from various points of view and perspectives. They could be more complete or better organized, making the meaning of data clearer in the absence of interpretation. To address these issues, these two platforms are thoroughly examined to identify modeling flaws and relationships between various elements.

Building on this foundational analysis, the MeLOn Methodology is applied to develop the Medieval Manuscript Data Integration Ontology(MMDIO), which is a single unified ontological framework that connects disparate datasets related to medieval manuscripts. The MMDIO ontology extends the existing Medieval Manuscript Ontology (MeMO), introduces new categories and relationships, and incorporates elements from the existing ontologies.

By streamlining the structure and ensuring consistency among various data sources, this process enables the effective utilization of MMDIO to organize and preserve the vast amount of knowledge and perspectives present in medieval manuscripts. The integrated MMDIO framework application to real-world scenarios on heterogeneous platforms is thoroughly examined. Nevertheless, this evaluation demonstrates the utilization of ontology to enhance medieval manuscript analysis in various contexts as well as enabling researchers to conduct advanced searches across multiple data sources.

The study is appropriately structured to allow for comprehensive understanding. Section 2 explores the unified ontological frameworks in medieval manuscripts specifically their role in managing and integrating data from heterogeneous data sources. The methodology is covered in detail in Section 3 focuses on platform analysis and metadata extraction, the use of MeLOn Methodology, and the MMDIO ontology structure. Section 4 evaluates the practical applications of use cases, highlighting their common characteristics and distinct focuses, and the results of competency questions and analysis. Lastly, Section 5 presents a concise summary of findings and recommendations for future research directions.

2. Literature Review

This review examines the practical applications of ontological frameworks in the area of medieval manuscripts, specifically in the context of their ability to handle and incorporate data from various sources. However, the focus is on examining the existing literature to gain an understanding of the current methodologies and their limitations in data integration. The results of this study can provide valuable guidance for developing a new ontology to enhance data interoperability and management across a wide range of datasets.

One such project is Europeana, a platform that offers access to cultural heritage resources from 36 European countries. These resources encompass a wide range of materials, such as books, documents, artifacts, and audiovisual content. The Europeana data model (EDM), was designed to address the challenge of integrating the diverse range of objects and metadata schemas by accommodating the variety of objects and metadata schemas that were provided. The DM2E model is an adapted iteration of EDM specifically designed for manuscripts. It integrates metadata and objects from various origins. It was created using the bottom-up approach, which facilitated the incorporation of detailed data mapping while also using pre-existing ontologies like Dublin Core, Bibo, FaBio, or the OAI-ORE specification [16]. Although Europeana successfully incorporates a wide array of data from various sources, it often lacks the level of detail required for medieval manuscripts.

Another important project is the Mapping Manuscript Migration (MMM) developed for analyzing and integrating various pre-modern manuscript databases on the semantic web. They utilize an integrated ontology to track the provenance and history of manuscripts across diverse collections and borders. The unified data model is based on the FRBRoo and CIDOC CRM ontologies by converting data from three distinct databases into Linked Open Data (LOD) [17]. The MMM's specialized historical focus might restrict its adaptability in comprehensively handling additional significant elements of manuscript data from diverse sources.

Biblissima is also one of the notable digital humanities projects that provide a single point of access to multiple partner databases containing historical information on manuscripts and early printed books. The resources comprise a variety of metadata relevant to different fields of research, such as codicology, iconography, and so on. Due to the diversity of database systems and data types adopted by the partners, it was difficult to standardize the data and achieve image interoperability. To encounter these challenges, a single unified ontology is developed that is compliant with CIDOC CRM and FRBRoo, a common XML model, and a thesaurus to ensure interoperability among the partner

resources [18]. Biblissima continues to face major difficulties in the interoperability of data and uniformity across its partner databases, complicating the seamless integration of various data types and undermining its effectiveness on heterogeneous platforms.

This article addresses the challenges associated with transforming a collection of digitized manuscripts, such as museums, archives, and libraries, into a knowledge base that is easily searchable using Semantic Web technologies, in particular challenges, such as the presence of multiple languages, difficult-to-read handwriting, and historical information. As a result, different workflow approaches have been developed for transforming digitized manuscripts into a searchable database after investigating various manuscript enrichment workflows. The use of the Web Annotation Data Model is recommended for adding annotations to digital images on the web [19]. However, the study focuses on the limited record of provenance information and the limited use of Linked Data technologies in current systems, which restricts their effectiveness in managing heterogeneous data in medieval manuscripts.

The Progetto Irnerio is a valuable digital catalog containing a significant amount of legal, cultural, and archival metadata about Roman and Canon Law that was created by lawyers in Bologna between the XII and XVI centuries. The archive has been fully digitized, and over 138,000 digital reproductions are now accessible via the Progetto Irnerio catalog for remote consultation and documentation. The Medieval Manuscripts Ontology (MeMO) was created using the Simplified Agile Methodology for Ontology Development (SMOAD)in six iterations to facilitate the representation, identification, analysis, and retrieval of manuscript information. MeMO models the data related to medieval texts organized by Progetto Irnerio at the Royal College of Spain in Bologna, Italy. It enables clear semantics and easily manageable queries that maintain the flow of the historical narrative, the context, and the description of the collection's structure. MeMO makes use of several ontologies, including CITO, DCTerms, FaBio, FOAF, FRBRcore, Literal, TI, and TVC [20]. The MeMO ontology, which is particularly designed for medieval manuscript data present on Irnerio. However, it does not accommodate the broad requirements of diverse data integration across different platforms, but it offers a flexible framework that can be extended to accommodate these needs.

Mosaico, mainly serves as an online resource for accessing and viewing a collection of medieval manuscripts. The platform's goal is to effectively preserve and handle digital resources while providing multiform access to them from three distinct historical perspectives: Roman, medieval, and contemporary. The project has two main repositories: the XML repository, which contains manuscript descriptions, and the image repository, which includes high-resolution images of the manuscripts. The project gives researchers access to a significant collection of digitized manuscripts and scholarly resources that are managed in such a way that they can be explored and analyzed [21]. Despite aggregating various historical information, it does not use a formal ontology, which limits its semantic data operations.

These gaps highlight the need for a new ontology designed specifically for the Mosaico and Progetto Irnerio platforms to address current limitations in managing complex medieval manuscript information. Existing systems need more depth to handle the diverse aspects of historical data, textual features, material composition, and artistic attributes effectively. By adapting and improving the MeMO ontology, which is specifically designed for medieval manuscripts, this approach can be broadened to incorporate more diverse datasets, further enriching the platform's capabilities.

3. Methodology

3.1. Analysis of Archives Metadata

The process of metadata analysis begins with identifying relevant data sources and gathering the required metadata. This task signifies an extensive review of data models and schemas to identify and refine key metadata elements that are crucial for resolving data inconsistencies and standardizing format. This step sets a solid foundation for ontology development.

Progetto Irnerio platform while rich in data, suffers from challenges such as non-standard data formats, inconsistent metadata schema, and data quality concerns. For instance, the *century* field belongs to codices[2] that inappropriately combine spatial and temporal data, which are rectified by reformatting to standard date intervals, such as *start date* and *end date*. Furthermore, for some cases, *place of origin* is determined along with the *current location*, which is 'The Royal College of Spain in Bologna'. The *description* field is thoroughly assessed, and the following metadata elements are identified: decorative elements (*initial* and *rubrication*), *style of script*, and the *number of writing hands*. This analysis also uncovered gaps in the existing MeMO ontology, such as missing *textual roles* (*follows* and *notes*), and incorrect entries, like `memo:TextualMetadataInTime`, must be corrected by replacing them with a new *Textual Metadata* element to more accurately hold textual content. The metadata related to the *codex* includes *bibliographic references*, which are presented inadequately. Additionally, it requires the inclusion of other metadata, such as the manuscript's *language*, which is typically in 'Latin'.

The Mosaico digital library, which contains numerous medieval manuscripts, presents challenges due to its unstructured and fragmented presentation. This lack of standardized data structure complicates the process of discovering relationships within the data, which is crucial for accurate and meaningful metadata representation. A particular focus in the Mosaico digital library is the *book* 'Authenicum' for which metadata are carefully recorded. The *book* consists of multiple codices, each of which aggregates manuscripts. The *book* has specific *sections*, which include *bibliographic references* and are linked to *author biographies*. In modeling the metadata for these codices, crucial metadata identified include *codex identifier*, *place of origin*, *current location*, *century*, *foliation*, *size*, *measurement unit*, *material* composition, *style of script*, *number of writing hands*, and decorative elements, such as *initial* and *rubrication*. The *century* metadata for the *codex* and *manuscript* is identified and converted into specific *start date* and *end date* intervals.

Furthermore, the *current location* for the manuscripts present under codices is identified. Meticulous attention is also paid to the additional *descriptions* under 'Glosses' and *bibliographic references* in codices. While glosses themselves are annotations within the manuscripts, *gloss* information is identified in the manuscript, and metadata is created to describe these glosses in detail. This includes details, such as *title*, *folio number*, *type* (either *ordinary* or *extravagant*), and *placement* (*marginal* or *interlinear*). The incorporation of additional metadata, including the *language* of the manuscript, is also necessary.

[2]The MeMO term 'codex' refers to a tangible object, not 'the Codex' (i.e., the Codex Justinianeus) or the concept of the manuscript (a handwritten text). A codex is a collection of manuscripts without definite authorship or title.

This structured approach to metadata categorization significantly enhances data retrieval capabilities and supports more robust data analysis. The process ensures that the new ontology precisely and accurately represents the extensive and detailed information characteristic of medieval manuscripts.

3.2. MeLOn Methodology

The MeLOn (Methodology for Creating Legal Ontology) is used to create the ontology that helps legal specialists overcome challenges in modeling different domains [22]. To adapt this methodology for medieval manuscript studies, it included specific historical and philological terms, developed use-case scenarios relevant to manuscript research, and used data integration techniques to handle the challenges associated with heterogeneous medieval manuscript resources. The effective implementation of this interdisciplinary strategy involved the development of various legal ontologies, enabling a comprehensive framework that promotes the connection between disparate manuscript collections while also improving digital accessibility.

1. **Describe the goal of the ontology**. The main goal of this ontology is to provide a coherent framework that connects the multifaceted aspects of medieval manuscripts for organization and analysis. This framework attempts to improve and harmonize metadata across different collections, therefore facilitating a more thorough academic study. It specifically addresses the need for data consistency, accessibility, and integrative research capabilities within the field of medieval studies.
   The following **research questions** are designed to guide the ontology development process.
   (i) What specific methods does the ontology employ to standardize and integrate data from various manuscript sources into a unified framework?
   (ii) How does the ontology enable detailed and multifaceted research queries that combine textual, historical, artistic, and physical data analysis of manuscripts?
   (iii) In what specific ways does the ontology connect medieval manuscript data to external academic databases, enhancing interdisciplinary research and collaboration?

   The Subsequent **use cases** are defined to illustrate the functionality and practical impacts of ontology.
   (i) Improving medieval manuscript accessibility and analysis: Emphasizing the 'Authenticum' collection in the Mosaico digital library address fragmented and unclear manuscript presentation.
   (ii) Revolutionizing manuscript research on the Irnerio platform: Enhancing the platform data presentation and accessibility.
   (iii) Cross-platform integration for a unified research framework across Mosaico and Irnerio.

   The practical use case scenarios are discussed in detail in Section 4.
2. **Evaluation Indicators**. To evaluate the use cases according to the criteria outlines in **(Step 1)** of this methodology. These evaluation parameters include factors such as coherence, completeness, efficiency, usability, and agreement. These matrices ensure that the ontology not only adheres to academic standards but is also highly reliable and broadly employed.

3. **State of the Art-Ontology**. A comprehensive analysis of state-of-the-art ontologies has been conducted to enhance the representation and analysis of medieval manuscripts. The MeMO ontology has been enriched to more comprehensively detail manuscript properties and relationships, incorporating refined concepts from related ontologies such as FaBiO, FOAF, FRBR, C4O, DoCO, DEO, BiRO, DCTerms, CITO, ARCO, CORE, TI, and TVC to develop a robust semantic framework.
4. **List all the relevant terminology to produce a glossary**. The glossary is produced using the most relevant terms based on their frequency of occurrence and semantic significance in the medieval manuscript domain from the Mosaico and Irnerio Platforms and definitions are provided from authoritative text, which ensures historical accuracy and relevance.
5. **Use usable tools**. The conceptual tables are used for concepts, object properties, data properties, and ontology restrictions to generate a knowledge base for the ontology.
6. **UML Modeling**. A UML diagram, created using the Graffoo tool [23], visually represents the ontological framework. Graffoo helps graphically represent relationships and hierarchies within our ontology, simplifying complex concepts and providing a clear roadmap for further development.
7. **OWL MODELING**. The UML diagram is transformed into a functional ontology using Protégé [24], a popular ontology editor, and converted into the Web Ontology Language (OWL) for use in semantic web environments. As part of this process, data is carefully extracted from both the Mosaico and Irnerio platforms. Our data collection for the ontology involved different approaches for the Mosaico and Irnerio platforms. For Mosaico, all data from 28 manuscript collections under 'Authenticum' were manually gathered, including comprehensive metadata about the collections and associated metadata related to both the 'Authenticum' book and other associated books, due to the lack of structured data interfaces.
   In contrast, the Irnerio platform combined manual and automated processing for two codices, 'Codex-006' and 'Codex-282'. Notably, tasks involving textual metadata and folio details, etc, were facilitated using the capabilities of Python's Natural Language Toolkit (NLTK). The data is then transformed and mapped into the ontology using RDF (Resource Description Framework) triples through Python code, ensuring accurate integration and semantic consistency. The ontology schema, RDF database, and other related materials are publicly available on the GitHub repository(`https://github.com/irnerio-mosaico-opendata/mmdio`).
8. **Test**. Ontology is then employed in real-world use cases to evaluate its usefulness, effectiveness, and applicability. These tests use GraphDB [3] to ensure robust evaluation by modeling practical scenarios with extracted data and investigating how the ontology handles complex queries and data relationships. Moreover, changes were made based on user feedback, improving end users' accuracy and usability.
9. **Refine and optimize**. The test outcomes are used to enhance the performance, usability, and coverage of the ontology through refinement.
10. **Evaluate the ontology**. The ontology is re-evaluated using Onto Clean [25] in the context of assessment indicators and a set of SPARQL queries to ensure that it adheres to the specified goals and academic standards.

[3] https://graphdb.ontotext.com/documentation/10.1/graphdb-workbench.html

11. **Publish the document**. The ontology is published and fully documented using LODE [26], which extracts and presents the OWL ontology structures in a human-readable HTML format.
12. **Collect Feedback to the community**. The feedback is gathered from the community, and the ontology is refined based on domain experts' observations, maintaining its significance and applicability to real-world scenarios.

3.3. MMDIO Ontology

The Medieval Manuscript Data Integration Ontology (MMDIO) is an OWL2 DL ontology that extends the MeMO ontology by broadening its conceptual framework. It implements the FRBR [4] principles using the FaBiO (FRBR-aligned Bibliographic Ontology (FaBiO) approach which reinterprets FRBR entities to better align with the user's perspective, distinguishing between the intellectual content of works (expressions) and their physical manifestations. The MMDIO ontology enhances the categorization of intellectual and physical content, introducing new classes and properties inspired by FaBiO and closely aligned with the FRBR framework to strengthen the link between digital resources and their physical counterparts. This enhancement enables deeper digital collaboration with medieval manuscripts across diverse data sources such as the Progetto Irnerio and Mosaico platforms.

The MeMO ontology used by Progetto Irnerio focuses on the manuscript as the main entity, while the Mosaico portal consists of a collection of digital books along with manuscripts and codices. To integrate these two systems, the MMDIO ontology considers different perspectives of each platform. Thus, in the MMDIO ontology, `memo:Manuscript` is characterized as a subclass of `fabio:Book`. The Graffoo diagrams in **Figures**[1,2, 3] displays the present version of MMDIO. These diagrams are detailed in the following subsections, where each figure is aligned with the respective layer of the ontology. The prefixes and associated base URIs of the reused models in MMDIO are detailed in **Table 1**.

3.3.1. Expression Layer Enhancement Under `fabio:Expression`

This subsection explores the improvements made under `fabio:Expression`, emphasizing the organization and accessibility of digitized book and manuscript content and their scholarly context in digital resources. **Figures**[1,2] are included due to their relevance to the overall context and integration of the expression layer. The expression layer is enriched with classes that specifically deal with intellectual content and abstract representations of work, making them vital for academic analysis and digital interaction. **New classes** include `mmdio:DigitalResourceCollection`, which organizes the digital representations of medieval manuscripts from various sources. The `mmdio:ManuscriptCollection` allows for the grouping of the manuscripts within digital resources, thereby improving navigational structures. The `mmdio:GlossType` and `mmdio:GlossPlacement`;

[4] A FRBR Work is the conceptual essence of a creative or intellectual resource that is not restricted to a specific format. Each Work can be expressed in multiple ways, known as expressions, each of which is a distinct realization of the work in terms of content. These Expressions are further manifested through one or more manifestations, each of which has a distinct physical or digital format that embodies one or more expressions. Finally, the FRBR Item is the tangible instance of a Manifestation, with each Item representing a distinct instance of a manifestation that is physically located and identifiable [27].

specify the types and locations of annotations within texts, which improves textual analysis and understanding. Despite being new, `mmdio:TextualMetadata` is not included under `fabio:Expression` because its purpose is to provide metadata that enables textual content without explicitly expressing creative or intellectual expressions. **Existing Classes** are categorized under `fabio:Expression` like `fabio:Book`, `memo:Manuscript`, `biro:BibliographicList`, `biro:BibliographicReference`, `memo:Text`, `memo:Gloss`, `doco:Section`, and `memo:CriticalEditionVolume`. They are used to capture a wide variety of scholarly content, enabling comprehensive academic discussions and analyses.

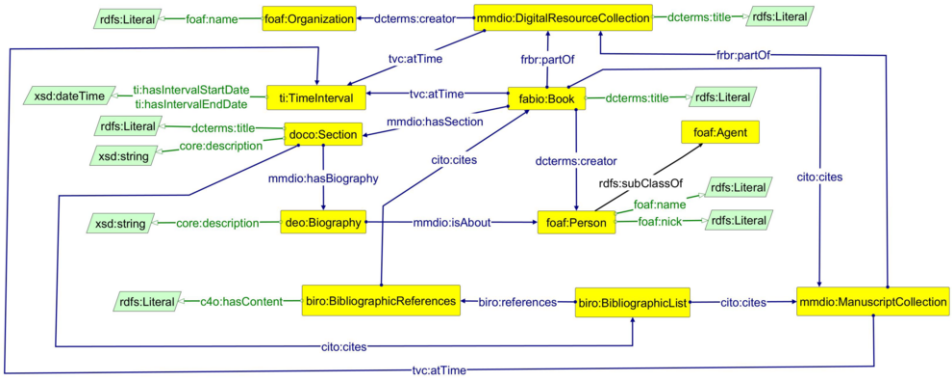

Figure 1. The Graffoo diagram presents digitized book information and scholarly context in digital resource collection.

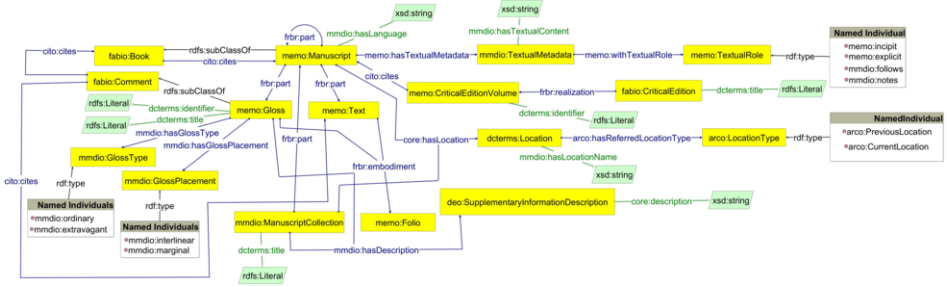

Figure 2. This figure illustrates the manuscript provenance and textual analysis

3.3.2. Manifestation Layer Enhancement Under `fabio:Manifestation`

In this subsection, enhancements under `Fabio:Manifestation` are explored, focusing on codex features and their representation in the MMDIO ontology. **Figure**[3] primarily represents the manifestation layer, including the expression embodiment into manifestation and codex as a manifestation. The manifestation layer deals with the physical components of manuscripts and associated materials, emphasizing those that are essential for

preservation and archival concerns. The following classes, `memo:Codex`, `memo:Folio`, `memo:Side`, and `memo:Binding`, are emphasized within the `fabio:Manifestation` category. This includes concentrating on the physical structure, binding, and foliation of manuscripts, which are essential for understanding the materiality of ancient records. A **new class,** `mmdio:Decoration`, has been introduced to capture the decorative elements of manuscripts, including illuminated initials, and rubrication, which are crucial for understanding the artistry of medieval texts.

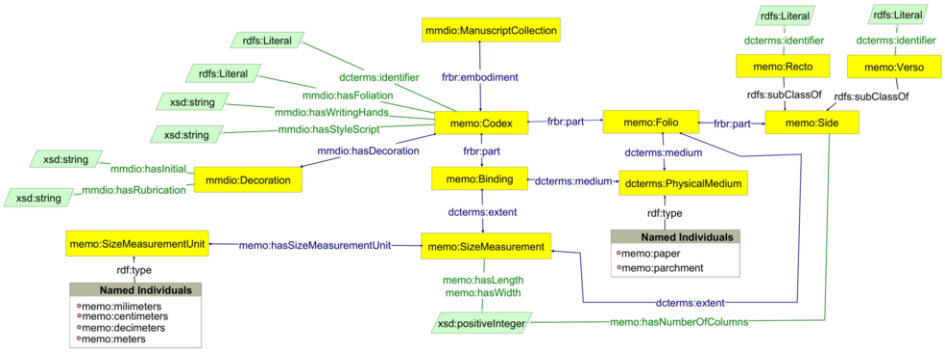

Figure 3. Representation of codex features in MMDIO ontology

3.3.3. Object Properties and Data Properties [5] Expansion

The ontology is enriched with new object properties, such as `mmdio:hasGlossType`, `mmdio:hasGlossPlacement`, `mmdio:hasDescription`, `mmdio:hasSection`, `mmdio:hasBiography`, and `mmdio:isAbout` along with data properties, such as `mmdio:hasTextualContent`, `mmdio:hasLocationName`, `mmdio:hasLanguage`, `mmdio:hasWritingHands`, `mmdio:hasStyleScript`, `mmdio:hasInitial`, and `mmdio:hasRubrication`. These properties substantially expand the ontology's descriptive capabilities, enabling a more comprehensive presentation of the textual, historical, and artistic aspects of the manuscripts.

3.3.4. Standardization through Controlled Value Lists

The ontology includes newly named individulas such as `mmdio:ordinary`, `mmdio:extravagant`, `mmdio:interlinear`, and `mmdio:marginal` are introduced as standard terms for describing gloss types and placements, to encourage consistency across digital annotations and descriptions. In addition, `mmdio:follows` and `mmdio:notes` are added as new textual roles.

By incorporating these crucial advancements, the MMDIO ontology not only bridges the digital and physical realms of medieval manuscript research but also promotes a more organized and accessible framework for managing heterogeneous data sources.

[5] In developing the MMDIO ontology, the aim was for maximum adaptability and scalability to facilitate the integration of data from heterogeneous platforms. As such, domain and range restrictions for the properties have not been specified. This design decision enhances the flexibility of the ontology, allowing it to be applied to a wide range of contexts and to evolve with emerging data integration needs.

Table 1. The prefixes and their associated base URIs for the models incorporated into MMDIO

Prefix	Base URI
mmdio	https://w3id.org/irnerio-mosaico/ontology/mmdio/
memo	https://w3id.org/irnerio/ontology/memo
fabio	http://purl.org/spar/fabio/
foaf	http://xmlns.com/foaf/0.1/
frbr	http://purl.org/vocab/frbr/core#
c4o	http://purl.org/spar/c4o/
arco	https://w3id.org/arco/ontology/location/
core	https://w3id.org/arco/ontology/core/
deo	http://purl.org/spar/deo/
doco	http://purl.org/spar/doco/
dcterms	http://purl.org/dc/terms/
cito	http://purl.org/spar/cito/
biro	http://purl.org/spar/biro/
literal	http://purl.org/spar/literal
owl	http://www.w3.org/2002/07/owl#
ti	http://www.ontologydesignpatterns.org/cp/owl/timeinterval.owl#
tvc	http://purl.org/spar/tvc
xsd	http://www.w3.org/2001/XMLSchema#

4. Application of Use Cases

The developed ontological framework has been used in several significant ways, especially through use cases that highlight how it can be practically applied to improve the accessibility, analysis, and management of data from medieval manuscripts. This section highlights the versatility and ability of the ontology to address difficult data integration problems.

4.1. Integrated Approach for Shared Elements

This subsection outlines the main elements and features of the MMDIO framework, focusing on how it addresses data integration from heterogeneous platforms. It provides an overview of the shared elements applicable to Mosaico and Irnerio, setting the stage for understanding the specific use cases. **Use Cases 1** and **2** are based on a common ontological foundation to facilitate digital cataloging and improve accessibility to medieval manuscripts and codices. The common framework encompasses:

The top-level class `mmdio:DigitalResourceCollection` is central and represents the medieval manuscript information integrated from heterogeneous sources and which has the particular title `dcterms:title` such as Irnerio and Mosaico. The class `foaf:Orgaization` which represents the organization where data has been taken for the `mmdio:DigitalResourceCollection` by using the property `dcterms:creator` and the organization name is represented through `foaf:name`.

Manuscript Collection Aggregation: The `mmdio:ManuscriptCollection` class serves as an important aggregator of manuscripts, directly linked to the `memo:Manuscript` using the `frbr:part` relationship. A `memo:Manuscript` is a

handwritten composition that contains only one memo:Text (the core textual content) and zero or more instances of memo:Gloss (textual annotations or commentaries). The memo:Gloss is conceptualized as a rdfs:subClassOf of fabio:Comment, placed within the margins or between the lines of the manuscript text, enriching it with scholarly annotations. The fabio:Comment utilizes cito:cites object property to reference memo:Text. The frbr:part property describes the complex part-whole relationship between a memo:Manuscript and its components (i.e., *Text* and *Gloss*). The mmdio:hasLanguage data property is used to capture the language of particular memo:Manuscript and frbr:partOf object property is used to link mmdio:ManuscriptCollection to mmdio:DigitalResourceCollection.

Manuscript and Textual Metadata Relationship: The memo:Manuscript encompasses various aspects of textual data that is associated to mmdio:TextualMetadata entity using the mmdio:hasTextualMetadata object property. This connection makes it easier to associate textual metadata with a specific textual role using object property memo:withTextualRole and these roles include memo:incipit for beginnings or memo:explicit for endings. Furthermore, the mmdio:TextualMetadata directly captures the manuscript's textual content using the mmdio:hasTextualContent data property.

Bibliographic References: Each biro:BibliographicList includes one or more a biro:BibliographicReference through the biro:references object property and biro:BibliographicList cites mmdio:ManuscriptCollection using cito:cites signifies scholarly engagement. Further, each biro:BibliographicReference holds citations and annotations using c4o:hasContent data property. However, the BibliographicList and BibliographicReference are also under fabio:Expression, highlighting the scholarly context in which manuscripts are studied and cited.

Temporal and Spatial Context: The mmdio:ManuscriptCollection have precise spatial and temporal dimensions that provide a comprehensive historical and geographical context. The tvc:atTime property links individual manuscripts and the entire collection to ti:TimeInterval, defining the time frame boundaries with ti:hasIntervalStartDate and ti:hasIntervalEndDate. Spatially, the collection and individual manuscript collections and manuscripts are connected to dcterms:Location using core:hasLocation that provides location information, while mmdio:hasLocationName provides the specific geographical name. arco:hasReferredLocationType simplifies the location type by defining arco:PreviousLocation, and arco:CurrentLocation, providing comprehensive insights into the manuscripts' provenance and journey over time.

Codex and Manuscript Collection Linkage: The frbr:embodiment object property enables the mmdio:ManuscriptCollection to transform into the memo:Codex at the manifestation level. The mmdio:ManuscriptCollection class was essential as an intermediate entity to connect expressions with their manifestations. Entities such as biro:BibliographicList and fabio:Book (**from Use Case 1**) refer to the codices, but citations are typically confined to expression-level entities. Therefore, integrating mmdio:ManuscriptCollection as an intermediate entity is vital for maintaining semantic accuracy while linking expressions with their manifestations. In this framework, mmdio:ManuscriptCollection operates at the fabio:Expression level and is con-

sidered at the fabio:Manifestation level when embodied in memo:Codex, which includes the codex's expression-level details.

Codex Structure: The memo:Codex is made up of one or more memo:Folio instances, each of which is a part of the codex. The memo:Folio represents the individual leaves of the codex, which are further divided into memo:Recto and memo:Verso are rdfs:subClassOf memo:Side. This hierarchical structure not only provides a detailed representation of the manuscripts but also maintains their physical arrangement and organization. Each memo:Codex is uniquely identified by dcterms:identifier, enabling accurate referencing and access. The mmdio:hasFoliation property indicates the sequence and structure of the codex's leaves, while the script style is designated by mmdio:hasStyleScript, specifying the calligraphic hand. Additionally, mmdio:hasWritingHands identifies the scribes who contributed to the manuscript, shedding light on different scribal practices. Decorative features such as illuminated initials and rubrication are managed under the entity mmdio:Decoration, with mmdio:hasInitial and mmdio:hasRubrication capturing these artistic elements.

4.2. Use Case 1: Improving medieval manuscript accessibility and analysis: Emphasizing the 'Authenticum' collection in the Mosaico digital library to address fragmented and unclear manuscript presentation

This use case details the distinct elements of the Mosaico Platform, highlighting its unique data structures and metadata. The 'Authenticum' book in the Mosaico digital library represents a significant advancement in the digitization and accessibility of medieval manuscripts. An extensive structural reorganization strategy was proposed for the 'Authenticum' collection, which was originally compiled by 'Christina Vano' in 2010. This strategy improves the collection's digital presentation and organization by utilizing the rich semantic structure.**Use Case 1** uniquely focuses on:

Digital Book Integration: In the digital realm, each fabio:Book is a digitized version of a traditional physical book that is part of a broader mmdio:DigitalResourceCollection using frbr:partOf object property. This represents the book's inclusion in an extensive compilation of digital texts, with each book uniquely identified by its title (via the dcterms:title property) to ensure accurate recognition within the collection. The fabio:Book cites series of mmdio:ManuscriptCollection.

Furthermore, fabio:Book is linked to its sections identified by the class doco:Section using the mmdio:hasSection object property. The title of each section is captured using dcterms:title, and the description is stored using core:description. Further, doco:Section is linked to the authors deo:Biography using mmdio:hasBiography object property and offers a comprehensive overview of their lives. The author's biography information is expressed using core:description, and its connection with the authors is determined using mmdio:isAbout, which links it directly to the foaf:Person. The structured method not only effectively organizes textual and contextual information, but it also improves the digital representation of historical literary works.

Authorship and Chronology: Each fabio:Book authorship is linked to foaf:Person, representing the author with the individual attributes, indicated by foaf:name. The historical significance of each fabio:Book is highlighted by

tvc:atTime, which associates the book with a ti:TimeInterval that defines the work's chronological relevance, as indicated by ti:hasIntervalStartDate and ti:hasIntervalEndDate, providing context regarding its original publication period or the span of its significance. The memo:Manuscript provides temporal context by employing tvc:atTime object property that is linked to ti:TimeInterval.

Gloss and Citations: Each gloss is represented by the class memo:Gloss and characterized by the properties such as dcterms:title and dcterms:identifier, mmdio:hasGlossType indicates the *gloss type* (either mmdio:ordinary or mmdio:extravagant), while mmdio:GlossPlacement specifies the placement (mmdio:interlinear or mmdio:marginal) of gloss. Furthermore, the biro:BibliographicReference is citing or referring to a book (represented by fabio:Book) using cito:cites. Specifically, this reference pertains to bibliographic references that cite ancient versions of the book, such as the 'Authenticum' book authored by 'Heimbach' in 1846.

Through these detailed ontological applications, the 'Authenticum' collection's enhancement strategy not only addresses digital accessibility and navigability challenges but also preserves and enriches the scholarly depth and historical context of the medieval manuscripts within the Mosaico digital library.

The following **competency questions** have been formulated to assess the effectiveness of the use case:

1. Retrieve the titles of all manuscript collections cited in the book 'Authenticum'.
2. Retrieve all the folios that contain a gloss including gloss title, identifier, type, and placement in the text.
3. Retrieve all manuscripts from the 'El Escorial, S.I.9' collection along with the bibliographic references and the book that cites this collection.

4.3. Revolutionizing manuscript research on Irnerio platform: Enhancing the platform data presentation and accessibility.

The Irnerio platform catalogs an extensive collection of codices, each containing manuscripts important for scholarly research. This use case demonstrates its specific data requirements and the challenges of integrating Irnerio data into the MMDIO framework. It emphasizes cataloging and systematic restructuring to make the Irnerio platform more navigable and user-friendly.

Use case 2 extends the framework to incorporate:

Critical Edition, Textual Metadata, and Recursive Relationship: Critical editions have been conceptually separated from their realizations in two specific and distinct layers, characterized by two classes: fabio:CriticalEdition and memo:CriticalEditionVolume. The fabio:CriticalEdition class is a rdfs:subClassOf of fabio:Work and describes the edition essence, independently from the revisions that can characterize it in time.

The textual roles are further enriched to incorporate textual metadata from Irnerio manuscripts such as mmdio:notes, and mmdio:follows) to improve the descriptive power and academic usefulness of metadata. Some manuscripts in the Irnerio collection have parts denoted within the manuscript itself. This structure encourages the use of a recursive relationship within memo:Manuscript, using the frbr:part object property. The validation of this use case is performed through the following **competency questions**.

1. What are the textual roles and content of the textual metadata associated with the manuscript titled 'Tres libri cum glossa Accursii'.
2. List all folios within the codex 'Codex-006', including the identifiers for each recto and verso side.
3. Retrieve the critical edition and critical edition volume that cite the 'Moralia sive Expositio in Iob' manuscript.

4.4. Use Case 3: Cross-platform integration for a unified research framework across Mosaico and Irnerio.

This use case addresses the specific integration of data between the Mosaico and Irnerio Platforms, aiming to establish a comprehensive data model that enables seamless integration and interaction while addressing data dispersion and organization issues. Specifically, the objective is to connect the 'Authenticum' book in Mosaico, which references manuscripts cataloged in Irnerio.

To illustrate a practical application: In the Mosaico portal's digital library, the digital version of a book identified as `fabio:Book` uses the `cito:cites` object property to link to an Irnerio portal manuscript (`memo:manuscript`). This integration uses ontology characteristics to improve the effectiveness and efficiency of medieval manuscript research, allowing scholars to access a larger and more interconnected dataset. The ontology facilitates comprehensive analyses and explorations by linking manuscripts across platforms, providing a cohesive academic research framework.

The effectiveness of data integration from heterogeneous platforms is measured through the following **competency question**.

1. Which books in the 'mosaico' digital collection cite manuscripts that are part of the 'irnerio' collection?

4.5. Evaluation and Results

The effectiveness of the ontology in specific use cases has been extensively evaluated using sample datasets from the Mosaico and Irnerio platforms. The ontology demonstrated the ability to integrate different data sources, standardize metadata, and validate its importance in medieval manuscript studies. The positive test outcomes demonstrate that the ontological framework can transform the field and is capable of widespread adoption, which is promising for future advances in manuscript data management.

5. Conclusion and Future Works

This study proposes an integrated ontology that effectively bridges the gap between digital technology and medieval manuscript research. The MMDIO ontology encourages scholarly research and public involvement by structuring metadata and improving accessibility across platforms, such as Mosaico and Irnerio. Furthermore, the modular architecture of the ontology enables future adaptation to similar systems, potentially making it a scalable solution for a wider range of digital humanities projects.

In the future, we plan to broaden the ontology to include additional use cases from the Mosaico Platform and create a knowledge graph to enrich the scholarly and cultural landscape.

References

[1] Borowiecki KJ, Forbes N, Fresa A. Cultural heritage in a changing world. Springer Nature; 2016, doi: 10.1007/978-3-319-29544-2

[2] De Hamel C. The European medieval book. In: The Book: A Global History. 2013:59-79.

[3] Kogman-Appel K. A Mahzor from Worms: Art and Religion in a Medieval Jewish Community. Harvard University Press; 2012.

[4] Steiner E, Barrington C, editors. The letter of the law: legal practice and literary production in medieval England. Ithaca (NY): Cornell University Press; 2002.

[5] Bromberg S. The Social Life of Illumination: Manuscripts, Images, and Communities in the Late Middle Ages. The Catholic Historical Review. 2014;100(4):813-814, doi: 10.1353/cat.2014.0226

[6] Keene BC, editor. Toward a Global Middle Ages: Encountering the World through Illuminated Manuscripts. Getty Publications; 2019.

[7] De Luca E, Loic E, Cavero AM. Intermediality in medieval Iberian manuscript cultures: methodological reflections on ongoing and future research. J Mediev Iber Stud. 2022;14(1):1-14, doi: 10.1080/17546559.2021.2021588

[8] Tourais A, Casanova C, Barreira CF. Filling the gap: new approaches to medieval bookbinding studies. J Mediev Iber Stud. 2022;14(1):109-126, doi: 10.1080/17546559.2021.2021589

[9] Córdoba de la Llave R. Interdisciplinary exploration of medieval technical manuscripts from the Iberian Peninsula. J Mediev Iber Stud. 2022;14(1):96-108, doi: 10.1080/17546559.2021.2019296

[10] Winkler A. Digitized medieval manuscripts in the classroom: a project in progress. Hist Teach. 2002;35(2):201-223, doi:10.2307/3054178

[11] Fischer F. Digital corpora and scholarly editions of Latin texts: features and requirements of textual criticism. Speculum. 2017;92(S1):S265-S287, doi: 10.1086/693823

[12] Zhitomirsky-Geffet M, Prebor G. Toward an ontopedia for historical Hebrew manuscripts. Frontiers in Digital Humanities. 2016;3:3, doi: 10.3389/fdigh.2016.00003

[13] Leydier Y, Lebourgeois F, Emptoz H. Text search for medieval manuscript images. Pattern Recognition. 2007 Dec;40(12):3552-3567, doi: 10.1016/j.patcog.2007.04.024

[14] Rose T. Technology's impact on the information-seeking behavior of art historians. Art Documentation: Journal of the Art Libraries Society of North America. 2002;21(2):35-42, doi: 10.1086/adx.21.2.27949206

[15] Freire N, Isaac A, Robson G, Howard JB, Manguinhas H. A survey of Web technology for metadata aggregation in cultural heritage. Information Services & Use. 2017;37(4):425-436, doi: 10.3233/ISU-170859

[16] Dröge E, Iwanowa J, Hennicke S. A specialisation of the Europeana Data Model for the representation of manuscripts: The DM2E model. LIBRARIES IN THE DIGITAL AGE. 2014;41.

[17] Burrows T, Emery D, Fraas AM, Hyvönen E, Ikkala E, Koho M, et al. Mapping manuscript migrations knowledge graph: data for tracing the history and provenance of medieval and renaissance manuscripts. Journal of Open Humanities Data. 2020, doi: 10.5334/johd.14

[18] Frunzeanu E, Robineau R, MacDonald E. Biblissima's Choices of Tools and Methodology for Interoperability Purposes = Biblissima: selección de herramientas y de metodología para fomentar la interoperabilidad. CIAN-Revista de Historia de las Universidades. 2016;19:115-132, doi: 10.20318/cian.2016.3146

[19] Stork L, Weber A, van den Herik J, Plaat A, Verbeek F, Wolstencroft K. From Historical Handwritten Manuscripts to Linked Data, doi: 10.1007/978-3-030-00066-0_34

[20] Barzaghi S, Palmirani M, Peroni S. Development of an ontology for modelling medieval manuscripts: the case of Progetto IRNERIO. Umanistica Digitale. 2020;(9):117-140, doi: 10.6092/issn.2532-8816/11187

[21] Palmirani M, Cervone L. A multi-layer digital library for mediaeval legal manuscripts. In: Digital Libraries and Archives: 8th Italian Research Conference, IRCDL 2012, Bari, Italy, February 9-10, 2012, Revised Selected Papers 8. Springer Berlin Heidelberg; 2013. p. 81-92, doi: 10.1007/978-3-642-35834-0_10

[22] Palmirani M, Martoni M, Rossi A, Bartolini C, Robaldo L. Pronto: Privacy ontology for legal compliance. In Proc. 18th Eur. Conf. Digital Government (ECDG); 2018 October; pp. 142-151, doi: 11585/648220

[23] Falco R, Gangemi A, Peroni S, Shotton D, Vitali F. Modelling OWL ontologies with Graffoo. In: The Semantic Web: ESWC 2014 Satellite Events: ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers 11. Springer International Publishing; 2014. p. 320-325, doi: 11585/570664

[24] Noy NF, Crubézy M, Fergerson RW, Knublauch H, Tu SW, Vendetti J, Musen MA. Protégé-2000: an open-source ontology-development and knowledge-acquisition environment. In: AMIA... annual symposium proceedings. AMIA Symposium; 2003 January; pp. 953-953.

[25] Mahlaza Z, Keet CM. 'OntoClean in OWL with a DL reasoner—A tutorial. Dept. Comput. Sci., Univ. Cape Town, Cape Town, South Africa; 2019 .

[26] Peroni S, Shotton D, Vitali F. The Live OWL Documentation Environment: A Tool for the Automatic Generation of Ontology Documentation. In: The Semantic Web: ESWC 2014 Satellite Events: ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers 11, doi: 10.1007/978-3-642-33876-2_35

[27] Layne SS. Functional Requirements for Bibliographic Records (FRBR). In: Encyclopedia of library and information sciences. 2010, doi: 10.1081/E-ELIS3-120043744