Semantic Modelling for Historical Library Reconstruction: Designing the *Fürstinnenbibliotheken* Ontology

Svenja Wagner[1,*,†] and Joëlle Weis[1,†]

[1] *Trier University, Universitätsring 15, 54296 Trier, Germany*

Abstract

The reconstruction of historical libraries requires the integration of heterogeneous, incomplete, and often ambiguous sources. This article presents the *Fürstinnenbibliotheken* Ontology (FBO), a semantic data model developed for the long-term project Princesses' Libraries and Knowledge Practices in the German-speaking World of the Eighteenth Century. Building upon established standards such as LRMoo, T-PRO, and Linked Open Data principles, the ontology combines interoperability with the flexibility needed for historical research. Implemented in Wikibase, it serves as the semantic backbone of the entire research data lifecycle, supporting data modelling, validation, reconciliation, and reuse while explicitly accommodating uncertainty and evolving historical evidence.

Keywords

library studies, Wikibase, princesses, collections

1. Introduction

Although the reading and collecting practices of early modern women have attracted increasing scholarly attention over the past two decades, research remains largely fragmented (for an overview on the state of the art see [1]). While numerous case studies have demonstrated the intellectual significance of women's libraries, systematic and comparative datasets are still scarce. Apart from the pioneering work of the MEDIATE18 project on women's private libraries [2], few attempts have been made to investigate these collections using data-driven or large-scale digital approaches. In German scholarship in particular, research has so far relied predominantly on analogue methods and individual case studies, making it difficult to assess the broader structures, networks, and patterns that shaped women's book ownership and knowledge practices.

The twelve-year research project *Fürstinnenbibliotheken und Wissenspraktiken im deutschsprachigen Raum des 18. Jahrhunderts* (english translation: Princesses' Libraries and Knowledge Practices in the German-speaking World of the Eighteenth Century) seeks to address this gap. Rather than examining individual collections in isolation, the project adopts a comparative perspective, reconstructing the intellectual worlds created through books, manuscripts, correspondence, and reading practices across multiple courts. The project asks how princely women acquired books, organised their collections, and used them in everyday life. It further explores the role of these libraries as interfaces between domestic space and wider scholarly, political, and commercial networks, highlighting the contribution of aristocratic women to the circulation of knowledge during the early modern period. These questions shift attention from libraries as static collections towards libraries as dynamic environments shaped by practices of acquisition, classification, reading, annotation, and exchange.

SWODCH'26: International Workshop on Semantic Web and Ontology Design for Cultural Heritage, Workshop days of ISWC 2026, Bari, Italy

wagnersv@uni-trier.de (S. Wagner[*,†]); weis@uni-trier.de (J. Weis[†])

0009-0009-3891-6333 (S. Wagner[*,†]); 0000-0002-0080-4362 (J. Weis[†])

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[*] Corresponding author.

[†] These authors contributed equally.

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-09

CEUR-WS.org/Vol-4320/paper8.pdf

A central challenge lies in the fragmented survival of these libraries. While some collections have remained largely intact, many were dispersed through inheritance, sales, or institutional integration. Their reconstruction therefore requires the combination of heterogeneous sources, including manuscript catalogues, surviving books, ownership marks, correspondence, estate inventories, wills, and archival documentation. These sources differ considerably in structure, completeness, and reliability, making traditional approaches to library reconstruction both labour-intensive and difficult to scale. At the same time, the increasing availability of digitised catalogues, provenance databases, and archival materials offers unprecedented opportunities for reconnecting dispersed evidence across institutions.

The *Fürstinnenbibliotheken* project therefore combines historical research with digital methods in order to reconstruct, analyse, and compare dispersed collections at scale. Rather than treating catalogues simply as inventories, the project models libraries as complex and evolving knowledge spaces whose material organisation, intellectual structure, and social use can be studied computationally alongside traditional source criticism. Digital reconstruction not only enables the identification of surviving copies and provenance relationships but also allows the visualisation of collecting patterns, linguistic preferences, subject classifications, acquisition strategies, and transregional networks of exchange. In doing so, digital methods become more than technical tools: they provide new ways of asking historical questions about women's intellectual agency and the material infrastructures of knowledge.

2. Project Data and Challenges

2.1. The Nature of the Project Data

One of the principal challenges of the project lies in the heterogeneous nature of its underlying data. This heterogeneity is not merely the result of bringing together material from numerous archives and libraries but is inherent in the historical sources themselves. Unlike modern library catalogues, eighteenth-century documentation was created without shared standards for description, identification, or data management. Each princess – and often each librarian or secretary – developed individual practices of recording, organising, and documenting books and related information. Consequently, the surviving evidence reflects highly diverse archival and bibliographical traditions.

The project therefore operates with data that varies considerably in complexity, completeness, and quality. Some collections are preserved through several successive manuscript catalogues, allowing changes over time to be reconstructed in detail. Others survive only in fragments, such as ownership marks in individual books, wills, correspondence, inventories, or isolated archival references. Even within a single collection, different source types often complement one another while simultaneously introducing contradictions or ambiguities. Information may be incomplete, duplicated, inconsistent, or entirely absent. Bibliographical descriptions range from highly detailed entries to abbreviated titles that can only be identified through contextual interpretation.

This diversity is further reinforced by the variety of source types integrated into the project. The reconstruction of a single library may involve manuscript catalogues, surviving books, provenance marks, correspondence, estate inventories, archival records, portraits, architectural plans, and other contextual materials. These sources differ fundamentally in their structure, granularity, and evidential value. Rather than representing a uniform dataset, they constitute a network of partially overlapping information that must be reconciled while preserving the uncertainty inherent in the historical record.

These characteristics have important methodological consequences. For our project, a Linked Open Data approach offers the most promising way of addressing them. Rather than forcing historical evidence into a rigid and fully standardised schema, the project requires a flexible data model that can represent incomplete, contradictory, and evolving information. At the same time, interoperability remains essential. The data is therefore modelled with established standards in

mind and published as Linked Open Data in accordance with the FAIR principles. FAIR means Findable, Accessible, Interoperable and Reproducible [3]. This approach combines semantic standardisation with the flexibility needed to accommodate the diversity of early modern sources, while enabling sustainable reuse and integration with external authority files, provenance databases, and digital library infrastructures.

2.2. Semantic Foundations

Rather than developing an entirely new ontology, the *Fürstinnenbibliotheken* project builds upon established standards and data models wherever possible. The aim was to maximise interoperability while limiting project-specific extensions to those areas in which existing models proved insufficient for representing the particularities of early modern princely libraries.

At the conceptual level, the project follows the object-oriented library reference model LRMoo [4], adopting the WEMI distinction between Work, Expression, Manifestation, and Item as the foundation for modelling bibliographical entities. This allows the project to distinguish consistently between intellectual works, their various editions, and the individual copies owned, annotated, exchanged, or inherited by the princesses. Since research questions primarily concern the histories of individual copies and collections rather than bibliographical descriptions alone, the item level plays a particularly important role.

The project further builds upon the LibReTo data model developed in previous research on historical library reconstruction [5]. LibReTo provides the basic conceptual framework for representing historical collections, ownership relations, catalogues, and provenance information. Within the *Fürstinnenbibliotheken* project, the model has been further extended to accommodate the specific requirements of dispersed princely libraries, including multiple catalogues, changing collection structures, uncertain provenance, and the integration of heterogeneous archival evidence.

Wherever possible, existing controlled vocabularies and authority files are employed instead of project-specific identifiers. Persons, places, institutions, languages, and subject terms are linked to established authority resources such as the Getty vocabularies, GND, Wikidata, and ISO language codes. Provenance information – one of our main areas of interest – is defined in accordance with T-PRO, a thesaurus of provenance vocabulary [6].

The systematic reuse of external authority resources not only improves data consistency but also enables interoperability with library catalogues, digital editions, provenance databases, and other Linked Open Data resources. By combining internationally established ontologies with domain-specific extensions, the project develops a semantic data model that remains both expressive enough for historical research and compatible with the wider cultural heritage ecosystem, while retaining the flexibility necessary for modelling the fragmentary and heterogeneous evidence characteristic of early modern library history.

2.3. Technical Infrastructure: Wikibase as Semantic Research Infrastructure

The semantic data model is implemented within Wikibase [7], an open-source knowledge graph platform originally developed as the software underlying Wikidata. Over the past decade, Wikibase has become one of the most widely adopted infrastructures for managing structured cultural heritage data due to its native support for RDF, Linked Open Data publication, collaborative editing, and semantic querying (See for example [8]). Its graph-based architecture is particularly well suited to the requirements of historical library reconstruction. Unlike relational databases, Wikibase does not require a rigid predefined schema but represents knowledge as interconnected entities linked through semantic statements. This flexibility allows the project to model incomplete, uncertain, or evolving historical evidence without forcing it into artificial structures.

The same flexibility, however, also presents challenges. Since Wikibase intentionally imposes only minimal structural constraints, semantically identical information can easily be represented in different ways by different contributors. Such inconsistencies reduce interoperability and

complicate automated analysis. For this reason, the project complements the technical infrastructure with a carefully documented semantic data model, modelling guidelines, controlled vocabularies, and validation procedures that ensure consistent data creation across the entire project.

The implementation is designed as a modular semantic infrastructure. The underlying RDF specification can be extended whenever new source types or research questions emerge without requiring fundamental changes to the existing data. This modularity allows the project to evolve continuously while maintaining semantic consistency and long-term sustainability.

Beyond data storage, Wikibase supports the complete research data lifecycle. It provides collaborative editing and version control, enables publication as Linked Open Data, offers a SPARQL endpoint for semantic querying, and supports federated queries that dynamically connect project data with external knowledge graphs such as Wikidata or authority databases. Comprehensive documentation accompanies the implementation and provides modelling rules, examples, and technical specifications, ensuring that both project members and external researchers can understand, reuse, and further develop the data in accordance with the FAIR principles.

3. The *Fürstinnenbibliotheken* Ontology

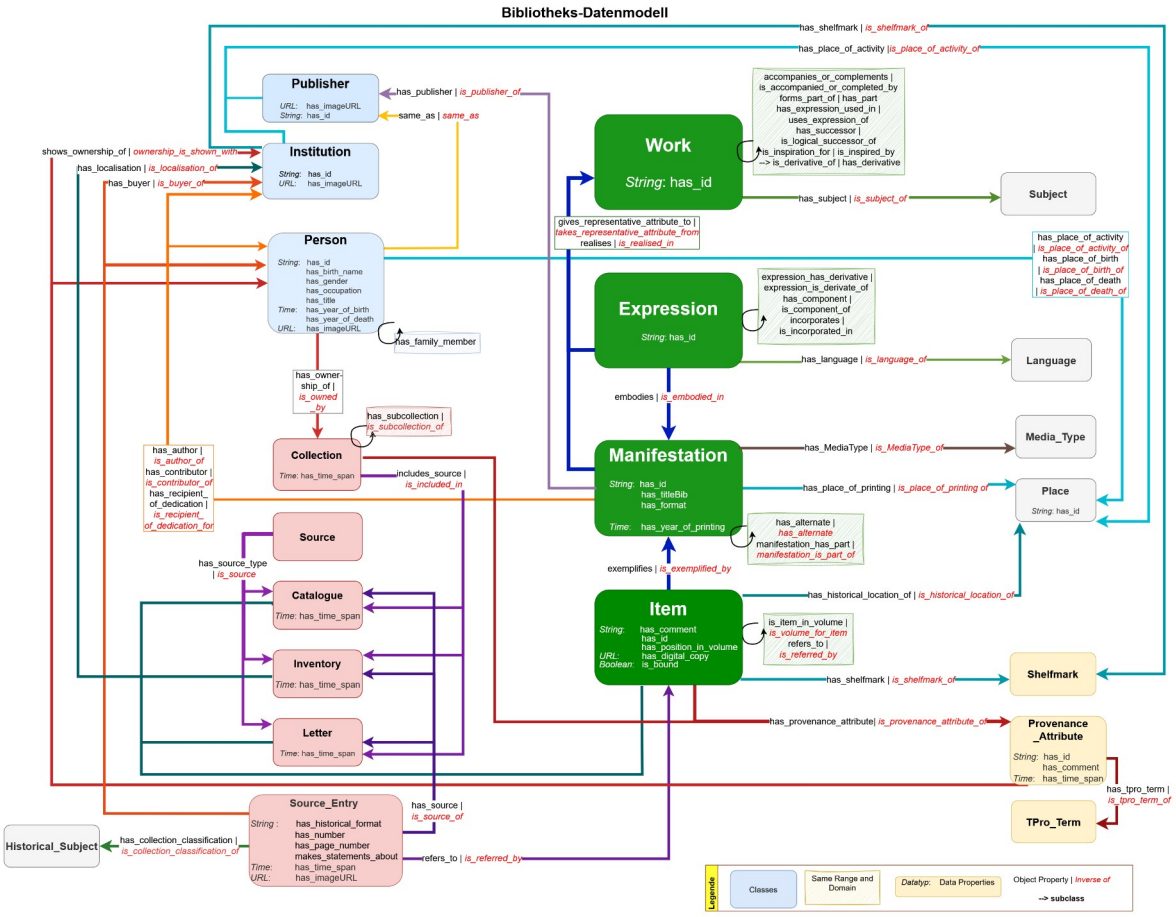


Figure 1: The *Fürstinnenbibliotheken* ontology (as of July 2026). Underscores are replaced by spaces inside the Wikibase and all further usage of the ontology. They result from the initial Resource Description Framework.

The *Fürstinnenbibliotheken* ontology (FBO), as shown in Figure 1, was developed to balance two partly conflicting objectives. On the one hand, it is tailored to the specific requirements of

reconstructing dispersed historical libraries; on the other hand, it aims to remain sufficiently generic to enable interoperability with existing Linked Open Data resources. Rather than introducing project-specific concepts wherever possible, the model reuses established ontologies and authority files, extending them only where necessary. This approach keeps the ontology closely aligned with the project's research questions while ensuring that its concepts remain understandable and reusable beyond the project itself.

To facilitate semantic interoperability, each central class is linked to corresponding concepts in external ontologies through semantic mappings. Besides standard RDF relations such as *is instance of* and *is subclass of*, the ontology documents explicit correspondences with external conceptual models, thereby enabling straightforward integration into wider semantic ecosystems. The model is organised around four central domains: books, actors, sources, and provenance evidence. These represent the project's primary research objects, while additional entities – including places, languages, institutions, and subject categories – provide contextual information shared across all domains.

The bibliographical component follows the WEMI structure of LRMoo [9] while adapting it pragmatically to the project's requirements. The distinction between *Work*, *Expression*, *Manifestation*, and *Item* is maintained throughout the model, as it forms the basis for distinguishing intellectual content from individual surviving copies. However, the complete complexity of LRMoo has deliberately not been implemented. Components that are irrelevant for the project's research questions – such as the detailed modelling of publication events – have been simplified in favour of a more practical representation. Instead, information such as publication year, place of printing, printer, language, and physical characteristics is represented directly. This balance between conceptual rigour and practical usability proved essential given both the heterogeneous source material and the large number of bibliographical records to be processed.

One of the most distinctive components of the FBO concerns the modelling of provenance and evidence. Rather than treating provenance simply as ownership history, the model records a wide range of physical traces – including ownership inscriptions, bookplates, seals, annotations, shelfmarks, bindings, and other material features – that provide evidence for ownership, circulation, and knowledge practices. Each provenance feature is assigned a corresponding T-PRO concept wherever appropriate, thereby embedding the observations within an internationally recognised provenance vocabulary. Importantly, the model distinguishes between the observation itself and the historical interpretation derived from it. A seal, for example, constitutes stronger evidence for ownership than a handwritten date, whereas a shelfmark primarily documents the location of a copy within a modern holding institution. By explicitly modelling these observations rather than their interpretation, the ontology preserves the evidential basis of historical argumentation while allowing researchers to draw their own conclusions.

Historical sources constitute a second central pillar of the model. Manuscript catalogues, inventories, correspondence, wills, and other archival documents provide the evidence from which collections are reconstructed. These source types are therefore modelled explicitly and linked both to the collections they describe and to the individual bibliographical entities they document. Multiple independent sources may thus refer to the same book, while individual sources frequently document several books simultaneously. This many-to-many relationship allows the reconstruction process itself to remain transparent and traceable.

Finally, the actor model connects collections, books, provenance evidence, and historical sources. Persons and institutions may fulfil multiple roles simultaneously – as owners, authors, editors, printers, booksellers, correspondents, family members, or holding institutions – and are therefore modelled independently from their individual functions. Wherever possible, actors are linked to external authority files such as GND or Wikidata, while the project itself records only those attributes necessary for reconstructing historical library networks. Although the overall structure of the FBO appears relatively straightforward, many modelling decisions required careful balancing between conceptual completeness and practical usability. The historical sources frequently resist clean categorisation, and early attempts at excessive

simplification often proved incompatible with the diversity of the surviving evidence. The final model therefore deliberately combines semantic precision with pragmatic flexibility, allowing uncertainty, ambiguity, and incomplete information to remain visible rather than forcing historical evidence into artificially rigid structures.

3.1. Key Design Principles of the Ontology

From the outset, the *Fürstinnenbibliotheken* ontology was conceived as a hybrid ontology that combines formal semantic modelling with the practical requirements of historical research. Rather than striving for complete conceptual abstraction, the ontology was designed to support concrete research questions while remaining interoperable with existing semantic standards. Three design principles guided its development: semantic consistency, contextual scalability, and the explicit representation of uncertainty.

3.1.1. Semantic Consistency

One of the ontology's primary objectives is to ensure semantic consistency across a large and collaboratively created dataset. Besides the formal semantics inherited from RDF and OWL – such as object property characteristics and class hierarchies – the ontology incorporates a number of project-specific modelling rules.

A central example is the implementation of the WEMI hierarchy. Since *Work*, *Expression*, *Manifestation*, and *Item* form a conceptual inheritance chain, information assigned to higher abstraction levels is assumed to remain valid for their descendants whenever appropriate. A subject assigned to *Work*, for example, also applies to all associated Expressions, Manifestations, and Items. This principle considerably reduces redundant data while maintaining semantic precision. Because Wikibase itself does not implement such inheritance mechanisms,[1] these relationships are explicitly defined within the project's semantic framework and are taken into account during data processing and querying.

Beyond formally implementable rules, the ontology also defines modelling conventions for individual properties. For example, the property *has shelfmark* (connecting *Item* and *Shelfmark*) always refers to the present holding institution, whereas *has author* (linking a *Person* as author to *Manifestation*) is restricted to persons contributing intellectually to the content of a work or manifestation. More general modelling guidelines are documented separately in the project's technical documentation, allowing the ontology itself to remain concise while ensuring consistent data creation across the project.

3.1.2. Contextual Scalability

A second guiding principle is scalability through semantic interconnection. Rather than functioning as an isolated knowledge graph, the ontology was designed to become part of the wider Linked Open Data ecosystem. Most central entities are therefore linked to external authority files and ontologies, allowing automatic interlinking with other Wikibases, library catalogues, and semantic datasets.

These semantic mappings considerably expand the contextual information available for individual entities. They also enable federated queries that combine project data with external knowledge graphs without requiring data duplication. Whenever automatic reconciliation is not possible, researchers can establish additional links manually. Newly identified correspondences are then incorporated into the project dataset, continuously enriching both the local knowledge graph and its connections to the wider semantic web.

[1] (In the documantation of Wikidata is written: „If you are familiar with object-oriented programming, you might expect that instances of a class inherit the statements of a class. This is generally not the case“ [10]).

3.1.3. Representing Historical Uncertainty

Perhaps the greatest challenge in modelling historical library data is the representation of uncertainty. Historical evidence is frequently incomplete, ambiguous, or contradictory, and the ontology deliberately avoids creating an illusion of certainty where none exists. Rather than interpreting missing information as negative evidence, the ontology distinguishes explicitly between confirmed statements and unknown information. The absence of a statement therefore never implies that a characteristic was absent historically. This distinction is particularly important when analysing incomplete historical collections, where gaps in the documentation are often the result of transmission rather than historical reality.

The same principle applies to bibliographical entities. Historical sources frequently provide only fragmentary descriptions of books – for example, mentioning an author, title, or format without supplying sufficient information to identify a specific surviving edition. Instead of postponing modelling until complete identification becomes possible, the project adopts a pragmatic strategy: every historically attested book receives its own WEMI representation based on the available evidence. This allows additional information to be incorporated over time and facilitates later reconciliation with other records as further evidence becomes available. By explicitly modelling uncertainty rather than concealing it, the ontology preserves the evidential character of historical research while simultaneously enabling computational analysis. It thus supports both scholarly transparency and the gradual refinement of the dataset as new knowledge emerges.

3.2. Selected Modelling Challenges

Designing the *Fürstinnenbibliotheken* ontology was not primarily a technical challenge, but a conceptual one. Most difficulties arose from the heterogeneous and often ambiguous nature of the historical evidence rather than from limitations of the underlying technologies. Throughout the development process, the guiding principle was therefore to balance conceptual rigour with practical applicability, favouring transparent and reproducible modelling decisions over theoretical completeness.

The most significant challenge concerned the implementation of the WEMI model. Although LRMoo provides a well-established conceptual framework for bibliographical description, the reality of eighteenth-century collections frequently resists such clear categorisation. Composite volumes, rebound books, collections of separately printed works, and manuscripts assembled by their owners often blur the boundaries between individual items and larger bibliographical units. Moreover, historical catalogues frequently provide only indirect evidence for these relationships. In many cases, it remains uncertain whether several works were physically bound together, belonged to a historical collection, or merely appeared together in a catalogue. Rather than attempting to eliminate these ambiguities, the ontology was designed to accommodate them explicitly while documenting the assumptions underlying each modelling decision.

A second challenge concerns the representation of bibliographical entities that cannot yet be identified with certainty. Historical sources frequently refer to books only indirectly, for example through abbreviated titles, physical descriptions, shelf locations, or mentions embedded in inventories, correspondence, or wills. These descriptions can, for example, read like this: “the large books with engravings, among them the Theatres d'Italie, de Savoie et Piemont, Vitruve Danois (etc). 31 vol.”[2] In many cases, these references are thus insufficient to identify a specific edition or even to determine whether two historical records refer to the same surviving copy. Nevertheless, such entries constitute valuable historical evidence and should not be excluded simply because they cannot yet be reconciled with modern bibliographical records. Rather than postponing modelling until complete identification becomes possible, the project adopts a pragmatic and iterative strategy. As mentioned earlier, every historically attested book still receives a complete WEMI representation based on the information available at the time of data entry, even if some levels of

[2] English translation by the authors (Original: “Darinnen sind die großen Bücher mit Kupferstüssn, worunter die Theatres d'Italie, de Savoie et Piemont, Vitruve Danois (etc). 31. Bd.“ [11])

the bibliographical hierarchy remain hypothetical. This allows additional information to be incorporated incrementally as new sources are examined or external catalogues become available, making reconciliation an ongoing rather than a prerequisite process.

This strategy also required a practical solution for the identification and labelling of incompletely described entities. Wikibase requires every entity to have a label upon creation, yet historical sources often do not provide enough information to construct meaningful bibliographical titles. Instead of relying on manually created placeholder labels, the project employs a semi-automated inheritance strategy. If no bibliographical title can be established, the Item initially inherits its label directly from the historical source entry. The corresponding Manifestation, Expression, and Work entities subsequently inherit this contextual label unless more specific information becomes available. This creates a coherent internal structure while allowing the complete upload process to be fully automated at this stage.

Automation, however, is deliberately balanced with scholarly control. While automatically generated labels considerably reduce manual effort and provide a consistent basis for further processing, bibliographical contextualisation remains a semi-automatic task. Future developments will therefore provide researchers with automatically generated reconciliation suggestions based on semantic similarity and bibliographical metadata, leaving the time-consuming search process to the machine while reserving the final decision for the human expert. Researchers remain free to reject proposed matches, improve labels, add aliases, or provide additional contextual information, which in turn enhances the quality of future reconciliation processes. In this way, automation supports rather than replaces scholarly judgement, allowing the dataset to evolve continuously as new evidence becomes available.

The modelling of provenance information presented different challenges. Here, the existence of the T-PRO vocabulary provided a solid conceptual basis, allowing provenance observations to be represented using an established semantic framework. The principal question was therefore not how to classify provenance features but how to distinguish between observations and their historical interpretation. Physical traces such as ownership inscriptions, seals, annotations, bindings, or shelfmarks provide different degrees of evidence for ownership and use. Instead of encoding historical conclusions directly, the ontology records these material observations and links them to appropriate provenance concepts, thereby preserving the evidential basis from which historical arguments can subsequently be developed.

Taken together, these examples illustrate that the ontology was deliberately designed as a pragmatic research instrument rather than as a theoretically exhaustive representation of bibliographical reality. Wherever historical complexity conflicted with conceptual elegance, priority was given to transparency, reproducibility, and the possibility of gradual refinement as new evidence becomes available.

3.3. Ontology-Driven Workflows

Within the *Fürstinnenbibliotheken* project, the ontology serves not merely as a conceptual model but as the central specification for the entire research data lifecycle. Rather than existing as separate documentation, it provides a machine-readable description of the project's semantic structure from which multiple technical components are derived.

The ontology forms the basis for data creation, validation, transformation, publication, and exploration. Definitions of classes and properties, multilingual labels, descriptions, cardinality constraints, and modelling conventions are maintained in a single semantic specification, ensuring consistency throughout the project. This "single source of truth" considerably reduces redundancy and simplifies long-term maintenance, since modifications of the ontology automatically propagate to downstream workflows.

This ontology-driven approach also enables a high degree of automation. During data preparation, semantic information is used to generate upload structures, suggest entity mappings, and perform consistency checks before data is imported into the Wikibase. Once published, the

same semantic structure supports querying through SPARQL, visualisation of relationships, and the integration of external Linked Open Data resources. Researchers therefore interact with a coherent semantic environment throughout the complete workflow rather than with a series of disconnected technical tools. At the same time, the ontology remains sufficiently flexible to accommodate the iterative nature of historical research. As new source types or modelling requirements emerge, individual components can be extended without affecting the overall structure. This modularity not only facilitates further development during the project but also ensures that the infrastructure remains maintainable beyond its funded lifetime.

Ultimately, the ontology is intended not only to describe the project's data but to actively support their creation, validation, and reuse. It thus becomes the semantic backbone of the research infrastructure, linking conceptual modelling with practical data management.

4. Outlook

The current implementation demonstrates that an ontology-driven approach can provide a robust foundation for the reconstruction of historical libraries. Future development will focus on transforming the ontology from a descriptive semantic framework into an active component of the research infrastructure that assists users throughout the complete data lifecycle.
A first priority is the development of ontology-driven data entry interfaces. While the current workflow already supports highly automated data preparation and upload, contributors are still required to understand parts of the underlying data model. Future project-specific input forms will therefore expose only context-dependent fields and automatically enforce the semantic constraints defined by the ontology. Mandatory properties, value restrictions, and modelling rules will be handled transparently in the backend, allowing researchers to focus on the historical sources rather than on the technical implementation. The objective is to lower the threshold for high-quality data creation while preserving semantic consistency across the entire dataset.

A second area of development concerns ontology-based validation. Since the ontology already represents the project's semantic ground truth, it naturally provides the basis for automated consistency checks. Existing technologies such as WshEx [12] offer promising approaches for validating RDF graphs, although they currently do not cover all aspects required by the project – for example, descriptive documentation and some project-specific semantic constraints. Extending these validation mechanisms will allow structural errors, missing mandatory information, and modelling inconsistencies to be detected automatically during data preparation and upload. Because the ontology itself is already maintained as a machine-readable RDF specification, these validation procedures can be generated directly from the semantic model, avoiding redundant rule definitions and ensuring that the validation logic evolves together with the ontology.

Another important area is semantic reconciliation. Existing reconciliation tools such as OpenRefine [13] mainly rely on textual similarity and make only limited use of the semantic information contained in the data model. Future developments therefore aim to exploit the ontology itself during entity matching. The envisioned reconciliation workflow consists of three complementary stages. The first stage performs fully automated rule-based matching using deterministic semantic rules. Some of these procedures are already implemented within the upload workflow – for example, the automatic reuse of controlled vocabulary entries such as T-PRO concepts or the merging of bibliographical entities fulfilling clearly defined criteria. More sophisticated rules can combine several semantic attributes, such as publication place, publication year, printer, and authorship, to identify duplicate manifestations with high confidence.
The second stage introduces ontology-aware similarity matching for cases where deterministic rules are insufficient. Rather than comparing labels alone, future algorithms will take the complete semantic context of an entity into account, allowing more reliable identification of potential duplicates while generating ranked reconciliation suggestions. Since these matches remain probabilistic, the final decision will always remain with the human expert, who may accept, reject, or further refine the proposed correspondences.

Finally, reconciliation will increasingly include external authority resources. New entities will be compared automatically with authority files such as GND, Wikidata, and other Linked Open Data repositories in order to retrieve identifiers and additional metadata whenever appropriate. Initial implementations of this functionality already exist within the project, but future work aims to integrate these procedures into a generic ontology-driven matching agent capable of combining internal semantic reasoning with external knowledge graph enrichment. Together, ontology-based validation, semantic reconciliation, and guided data entry converge towards an integrated upload environment in which the ontology actively assists data creation rather than merely documenting it. The first implementation already enables a highly automated workflow that produces a largely consistent and semantically rich knowledge graph. Future developments will further reduce manual effort while preserving the transparency, flexibility, and scholarly control required for historical research.

Acknowledgements

The *Fürstinnenbibliotheken* project is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 534456186.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.5 in order to: Improve writing style; Grammar and spelling check. After using the tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Caren Reimann, Joëlle Weis, Fürstinnenbibliotheken, in: Sarah Gaber, Stefan Höppner, Stefanie Hundehege (Eds.), Provenienz. Materialgeschichte(n) der Literatur, Wallstein, Göttingen, 2024, pp. 251 – 260. DOI: 10.15499/kds-009-018.

[2] MEDIATE. Understanding the literary system of the 18th century. URL: https://mediate18.nl.

[3] Mark D. Wilkinson, Michel Dumontier, IJsbrand J. Aalbersberg et al., "The FAIR Guiding Principles for scientific data management and stewardship“, in: Sci Data 3 (2016). DOI: 10.1038/sdata.2016.18.

[4] LRMoo Library Reference Model. URL: https://cidoc-crm.org/lrmoo.

[5] Hartmut Beyer, LibReTo Ontologie – Rekonstruktion historischer Bibliotheken. URL: https://bibliotheksrekonstruktion.hab.de.

[6] T-PRO Thesaurus der Provenienzbegriffe. URL: https://provenienz.gbv.de/T-PRO_Thesaurus_der_Provenienzbegriffe.

[7] Wikibase. URL: https://wikiba.se.

[8] Alan Ang, Maxime Guénette, Sabine von Mering et al. (Eds.), "Wikidata Across the Humanities: Datasets, Methodologies, Reuse", in: Journal of Open Humanities. URL: https://openhumanitiesdata.metajnl.com/collections/wikidata_across_the_humanities.

[9] Dunsire, Gordon, "Building Resource Descriptions with WEMI“, in: Catalogue & Index 213 (2025). URL: https://journals.cilip.org.uk/catalogue-and-index/article/view/834.

[10] Wikidata:Data model, Inheritance. URL: https://www.wikidata.org/wiki/Wikidata:Data_model#Inheritance.

[11] Systematischer Katalog der Handbibliothek der Herzogin Philippine Charlotte (1754), Wolfenbüttel, Herzog August Bibliothek, BA I, 641. URL: http://diglib.hab.de/mss/ba-1-641/start.htm?image=00149.

[12] Labra Gayo, Jose Emilio, "WShEx: A Language to Describe and Validate Wikibase Entities.“, in: arXiv (2022). DOI: 10.48550/arXiv.2208.02697.

[13] Open Refine. URL: https://openrefine.org/.

A. Online Resources

The current Ontology as Resource Description Framework (RDF-XML) is available at: https://gitlab.rlp.net/fuerstinnenbibliotheken/fuerstinnenbibliotheken/-/blob/main/Datenmodell/Ontologie.rdf. The latest version of Figure 1 can also be viewed in our open GitLab repository.