Ontology-Driven Linked Data for Hebrew Manuscripts

Semantic Web
XX(X):1–16

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Alexander Goldberg[1], Gila Prebor[2], Avshalom Elmalech[3]

Abstract
This paper presents the Hebrew Manuscripts Ontology (HMO), developed within the Mapping Hebrew Manuscripts (MHM) project, as an evaluated domain ontology and transformation workflow for representing Hebrew manuscripts as Linked Open Data. The problem addressed is that rich Hebrew-manuscript metadata in the National Library of Israel remain largely locked in MARC records, limiting entity-level querying, cross-record reasoning, and future interoperability with external knowledge graphs. HMO responds with a domain-specific model aligned with CIDOC CRM and LRMoo that combines three elements: structural granularity through Bibliographic Unit, Codicological Unit, and Paleographical Unit (BU-CU-PU); an event-centric representation of production, transfer, and ownership; and an explicit epistemological layer for source attribution and status, with certainty support defined at schema level. The paper evaluates this contribution at three levels: six fully instantiated pilot manuscripts chosen to cover key structural edge cases; schema-level validation over 37 SPARQL checks executed on the released ontology files and controlled vocabularies; and a larger-scale feasibility run of the conversion workflow over 10,000 catalog records. Within the pilot, HMO supports research questions that are difficult to ask in MARC alone, such as retrieving manuscripts with multiple hands, identifying codices with more than ten codicological units, tracing transfer-of-custody chains, and distinguishing textual witnesses from bibliographic works. The current release follows a MARC-only population policy for external identifiers: Wikidata, VIAF, GeoNames, and owl:sameAs slots are defined at schema level but are not yet populated in the pilot RDF, so interoperability is demonstrated here as alignment readiness rather than completed entity linking. The ontology (OWL/TTL), SHACL shapes, pilot RDF, crosswalk, validation materials, and conversion pipeline are included in the repository materials accompanying this paper for direct inspection. The contribution claimed here is therefore an evaluated and inspectable semantic framework, not a claim that full corpus-scale reconciliation or community-wide uptake has already been achieved.

Keywords
Hebrew manuscripts, linked open data, ontology engineering, CIDOC CRM, LRMoo, digital humanities

[1]Department of Information Science and Applied Artificial Intelligence, Bar-Ilan University, Ramat-Gan 5290002, Israel
[2]Department of Information Science and Applied Artificial Intelligence, Bar-Ilan University, Ramat-Gan 5290002, Israel
[3]Department of Information Science and Applied Artificial Intelligence, Bar-Ilan University, Ramat-Gan 5290002, Israel

Corresponding author:
Alexander Goldberg, Department of Information Science and Applied Artificial Intelligence, Bar-Ilan University, Ramat-Gan 5290002, Israel.
Email: golbera3@biu.ac.il

Introduction

Research Context

The geographic and institutional dispersion of Hebrew manuscripts—across national, academic, and ecclesiastical libraries, archives, and museums—has created over the years a situation of "cataloging fragmentation," in which essential information is scattered among different descriptive systems operating according to diverse professional standards and traditions. Against this backdrop, the Hebrew Manuscripts catalog of the National Library of Israel in Jerusalem occupies a unique position, aggregating comprehensive metadata on a substantial portion of the world's major collections and serving in practice as a global information hub in the field. Nevertheless, although the catalog reflects decades of investment in bibliographic, structural, and philological processing, the traditional data structures on which it relies—notably the MARC format—limit the ability to exploit the research potential of the information it contains, due in part to field inconsistency, gaps in authority files, and variation in name and work forms (Prebor et al., 2020b; Zhitomirsky-Geffet et al., 2020).

In the digital humanities, this limitation is becoming increasingly critical. The shift from textual record databases to conceptual models and linked data graphs is now widely perceived as a central paradigm for organising cultural heritage information, enabling researchers to pose questions that draw on the heterogeneity of sources, connectivity between entities, and cross-collection analysis at scale (Bermès, 2015; Berners-Lee et al., 2001; Dunsire, 2012; Landis, 2019). Libraries and cultural heritage initiatives have begun converting traditional catalogs to RDF structures and LOD graphs, adopting models such as CIDOC CRM and FRBRoo/LRMoo, thereby creating infrastructure for the unification, cross-referencing, and analysis of bibliographic and historical data from diverse sources (Bekiari et al., 2015; Hastings, 2015; Riva et al., 2017). Within this context arises the need for a dedicated model for Hebrew manuscripts that connects local cataloging tradition with international standards of knowledge representation, and leverages existing metadata for digital humanities research. This need stands at the centre of the Mapping Hebrew Manuscripts (MHM) project, whose principal output is the Hebrew Manuscripts Ontology (HMO). Throughout this paper, MHM refers to the overarching research project, while HMO denotes the ontology itself.

Problem Statement

Despite the quantitative and qualitative richness of metadata on Hebrew manuscripts in the National Library catalog system, the representation of information relies largely on the MARC format and the linear "catalog record" paradigm. This paradigm limits the ability to decompose the record into distinct entities and relationships, to link between different manifestations of the same entities in external repositories, and to execute complex queries that cross record and collection boundaries. The result is a gap between the research potential embedded in the metadata and the practical ability to exploit it within the digital humanities.

The literature explicitly identifies structural and substantive problems in MARC data: inconsistency in field and subfield definitions, multiplicity of homonyms in personal and place names, gaps in authority files, and considerable variation in the written forms of names and works (Weitz et al., 2016; Zeng, 2019). These problems complicate unambiguous identification of entities, their linking to global identifiers such as VIAF or Wikidata, and the performance of large-scale comparative analyses based on reliable and uniform data (Zhitomirsky-Geffet et al., 2020). In this situation, even when abundant data exist on scribes, manuscript owners, geographic centres, or textual traditions, they remain effectively "locked" within free-text fields and difficult to convert to semantic representation in a graph.

In parallel, the development of Linked Open Data infrastructures and their implementation in cultural heritage projects emphasises that the transition to semantic graphs is not merely a technical operation of format change, but involves the redefinition of conceptual models—what counts as an "entity," how events are represented, and what the relationship is between facts and interpretations and context-dependencies (Bermès, 2015; Berners-Lee et al., 2001; Dunsire, 2012; Landis, 2019). Although diverse ontological models for manuscripts have been developed over recent decades (as detailed in the literature review), the unique needs of the Hebrew corpus still require a corpus-specific model that can accommodate its philological and codicological complexity.

Hence the central research problem: How can a systematic, comprehensive, and interoperable ontological framework be designed and applied for Hebrew manuscripts—within the Mapping Hebrew Manuscripts (MHM) project—that serves as a semantic bridge between existing MARC catalogs and linked data environments? This paper presents HMO as a research contribution that combines ontology design, a documented MARC-to-RDF workflow, and an inspectable open evidence package. The goal is not to claim a completed linked-data publication pipeline for the whole corpus, but to show that a released and evaluated semantic framework can already make manuscript metadata more queryable, better structured, and more ready for future reconciliation with external knowledge graphs.

Research Objectives and Questions

The central aim of this paper is to present HMO as an integrative model and evaluated workflow for representing Hebrew manuscripts in a Linked Open Data environment, bridging the cataloging tradition of the National Library with international standards (CIDOC CRM, LRMoo). The paper pursues three objectives: (1) developing a multi-layer ontological model integrating cataloging-bibliographic, philological, and epistemological layers; (2) implementing a systematic MARC-to-HMO transformation workflow with explicit alignment points to Wikidata and VIAF; and (3) evaluating the contribution at three clearly delimited levels: pilot-scale instantiated case studies, schema/validation support, and larger-scale workflow feasibility.

These objectives give rise to three research questions, each answered explicitly in the evaluation section. First, to what extent can HMO capture the material and textual complexity of Hebrew manuscripts while overcoming the granularity limitations of MARC records? Second, what practical gain for research queries is achieved when manuscript metadata are transformed from MARC records into an HMO entity-event graph? Third, what degree of interoperability is already implemented in the present release, and what degree is currently defined only as future-ready schema alignment?

More precisely, the paper makes four bounded contributions. First, it introduces a corpus-specific ontology design for Hebrew manuscripts that combines CIDOC CRM and LRMoo with BU-CU-PU structural granularity, text-tradition entities, and epistemic qualification. Second, it documents a MARC-to-RDF transformation workflow that makes these modelling decisions operational. Third, it evaluates the resulting framework through a pilot focused on modelling adequacy and query usefulness, together with released validation materials for schema support and workflow feasibility. Fourth, it presents the ontology, pilot data, crosswalk, validation outputs, and replication guidance as reviewable artifacts. The contribution is therefore not "six manuscripts alone," but the combined package of ontology design, transformation workflow, inspectable query demonstrations, and released review materials. Correspondingly, the paper does not claim completed entity-level reconciliation against external authority graphs, corpus-wide scholarly validation, or broad community uptake beyond the current project context.

Background

Hebrew Manuscript Research and Metadata

In recent decades, a fundamental shift has occurred in the perception of manuscripts as sources for historical and cultural research: from mere "text carriers" they have been recognised as multi-layered material witnesses, combining textual traditions with evidence of writing, reading, and dissemination practices, and with broader institutional and social connections. In philological and codicological research, material and structural characteristics—type of material, volume format, quire division, ownership marks, and colophons—are now perceived as an integral part of the manuscript's "biography" and as a basis for identifying copying centres, scribe networks, and transfer routes between communities and regions (Beit-Arié, 2022; Burrows, 2018; Sirat, 2002).

In the Hebrew manuscript domain it has been repeatedly emphasised that these materials constitute one of the key sources for reconstructing Jewish cultural heritage and characterising the intellectual, religious, and daily life patterns of Jewish communities across the diaspora. The historical dispersion of Hebrew manuscripts, resulting from forced migrations, scholarly movements, and book trade, has meant that material evidence is now scattered across numerous national, academic, and ecclesiastical libraries, archives, and museums—a situation requiring descriptive and cataloging infrastructures capable of unifying heterogeneous information from diverse sources. In this context, the catalog of the Manuscripts Department and the Institute for Microfilmed Hebrew Manuscripts at the National Library of Israel in Jerusalem serves as the largest and most focused metadata repository for Hebrew manuscripts, representing, by estimates, approximately 95% of the world's major collections from the early centuries through the twentieth century (Beit-Arié, 2022; Richler, 2014; Sirat, 2002).

Empirical studies have demonstrated that catalog metadata can serve as a basis for quantitative research on copying patterns, scribe networks, and ownership chains. However, these studies also revealed substantial limitations of MARC data: lack of field uniformity, variant name forms, and partial authority files. Of more than 44,000 persons in the Hebrew manuscript catalog, only about a quarter were identified in authority files and VIAF (Zhitomirsky-Geffet et al., 2020). These findings place Hebrew manuscript metadata at a crossroads between rich research potential and representational limitations, constituting the point of departure for developing ontological models that leverage existing metadata while overcoming the formal constraints of MARC.

Conceptual Models for Cultural Heritage

Discussions in recent decades on knowledge organisation in cultural heritage institutions have led to the consolidation of general conceptual models seeking to overcome the fragmentation between museum, library, and archival descriptive traditions. The prominent model in this context is CIDOC CRM (ISO standard 21127), which presents a framework for describing objects and agents with emphasis on events as a central mechanism for representing historical dynamics (Bekiari et al., 2021; Doerr, 2003). The model enables consistent description of

"what happened" to objects throughout their life cycles and is based on the principle of monotonic reasoning, allowing extension without compromising the validity of existing representation.

In parallel, the library community developed the FRBR model and its object-oriented formulation FRBRoo (Bekiari et al., 2015), later updated as LRMoo (Aalberg et al., 2024), which define an abstract entity hierarchy (Work, Expression, Manifestation, Item). FRBRoo introduced the F4 Manifestation Singleton class to describe a unique physical copy in which work and expression merge (Le Boeuf, 2012). Although F4 was deprecated in LRMoo, HMO retains it as a pragmatic modelling choice, since it offers a direct and convenient mapping for representing individual manuscripts.

The synergy between CIDOC CRM and LRMoo enables the combination of an event-centric approach with bibliographic entity structures. This combination provides infrastructure for separating representation levels—textual idea, textual tradition, physical copies, and events along the timeline. However, as noted in recent reviews (Ferooz, 2025, pp. 79–81), a gap still exists regarding the semantic depth (granularity) required for describing specific corpora. These models, often designed for Latin manuscripts, do not fully address the codicological and cultural complexity of Hebrew manuscripts. At this point the HMO ontology builds on existing infrastructures while embedding unique semantics that enable the transition from data management to research-driven tools.

LOD in Digital Humanities and Manuscript Research

The Linked Open Data (LOD) framework has established itself over the past two decades as a central paradigm for representing structured data on the web, using URIs and RDF as the basis for semantic connectivity between heterogeneous repositories (Berners-Lee et al., 2001; Dunsire, 2012). In cultural heritage institutions, adoption of these principles serves as an engine for releasing data from libraries and archives beyond internal catalog system boundaries, and for assembling shared graphs enabling cross-referencing queries (Landis, 2019).

Library projects such as LIBRIS, the British National Bibliography, and the Library of Congress have demonstrated how crosswalk mapping from MARC to CIDOC CRM and LRMoo dismantles "data silos" and enables cross-repository graphs (Hastings, 2015). In the manuscript domain, Mapping Manuscript Migrations (MMM) aggregates data from major repositories for provenance research (Burrows, 2018), and MMDIO proposes semantic integration through polymorphic knowledge graphs (Ferooz, 2025). At the centre of the LOD ecosystem, Wikidata serves as a hub for linking identifiers (VIAF, GND), enabling complex queries not possible in internal systems (Evans, 2019), though posing challenges of data reconciliation (Ullah et al., 2018).

Among recent manuscript-focused LOD efforts, Mapping Manuscript Migrations (MMM) aggregates provenance data across major European repositories using CIDOC CRM and FRBRoo as base models, with its semantic emphasis placed on cross-repository migration patterns rather than on the codicological-philological description of any single corpus (Burrows, 2018, 2022; Koho et al., 2021). MMDIO subsequently proposes a polymorphic knowledge-graph approach to semantic integration across heterogeneous manuscript collections, but does not commit to a fixed structural-granularity model for the codex itself (Ferooz, 2025). HMO occupies a complementary niche: rather than aggregation infrastructure across collections, it provides a corpus-specific conceptual model that combines codicological-philological granularity (BU-CU-PU, TextTradition, TransmissionWitness) with an explicit epistemological layer for attribution and certainty—features motivated by the specific evidentiary practice of Hebrew-manuscript scholarship and not addressed jointly by the systems above.

For Hebrew manuscripts, the LOD environment provides infrastructure for connecting catalog metadata to global repositories, but requires a domain model translating philological-codicological complexity into entities and relationships. HMO is positioned at this intersection: through its three-layer architecture (cataloging, philological, epistemological) and BU-CU-PU structure, it enables the transition from static "records" to dynamic "entities and relationships," leveraging the existing catalog for SPARQL queries, network analysis, and large-scale research in the digital humanities.

The HMO Model: Methodology, Architecture and Implementation

Data Source, Sample Design, and Evaluation Strategy

The ontology relies on the Hebrew Manuscripts catalog of the National Library of Israel in Jerusalem, representing approximately 95% of the world's major collections (Beit-Arié, 2022; Richler, 2014). The catalog records, based on the MARC format, reflect decades of philological and codicological processing, but suffer from inconsistency and lack of links to global identifiers (Zhitomirsky-Geffet et al., 2020). This situation makes the catalog both a challenge and an opportunity: it provides a rich descriptive base, but one that requires semantic decomposition and careful mapping to formal entities and relationships before it can support graph-based research.

Evaluation in this paper combines three complementary levels. First, a six-manuscript pilot was fully instantiated in RDF as a coverage-oriented adequacy design rather than as a statistically representative sample. The six cases were selected to cover the main structural and scholarly edge cases that motivated the ontology: a unified codex (Barberini Or. 82), a complex multi-unit codex (Vatican Ebr. 44), a large anthology (Parma 3122), a provenance-focused case (Huntington 115), a text-tradition case (Oppenheimer 129), and a compact manuscript

with multiple hands and transfer events (Jerusalem 8210). Second, the ontology schema was checked through Protégé-based ontology inspection together with a released SPARQL validation suite. Third, the conversion workflow was exercised in a larger internal feasibility run on 10,000 catalog records, reported separately as a SHACL-based validation report and used here only to characterize pipeline behavior beyond the six detailed pilot cases. These three levels are treated differently throughout the paper: the pilot supports the research-use claims, the validation suite supports schema-support claims, and the 10,000-record run supports feasibility claims about workflow behavior at larger scale.

Model Architecture and Development Principles

HMO development was based on selective adaptation of CIDOC CRM and LRMoo standards to the unique needs of the Hebrew corpus. The model is built in a three-layer architecture enabling transition between different description levels:

Cataloging-bibliographic layer: Implements the Work-Expression-Manifestation structure of LRMoo, with each manuscript mapped to F4 Manifestation Singleton.

Philological layer: Focuses on modelling text traditions (TextTradition) and transmission witnesses (TransmissionWitness), enabling definition of research relationships between different copies of the same work.

Epistemological layer: Describes the source of information and epistemological status, with certainty support defined at schema level, an essential aspect for documenting research hypotheses regarding scribes or copying places.

For representing physical structure, the model uses composition relationships to represent the BU-CU-PU structure (Bibliographic Unit, Codicological Unit, Paleographical Unit). This approach enables flexibility in describing complex volumes (assembled codices) containing independent units linked by is_composed_of relationships, without subordination to a rigid subclass hierarchy.

Semantic Mapping: MARC-HMO-Wikidata Crosswalk

A central component of HMO is building a systematic crosswalk that decomposes the MARC record into entities and relationships in the graph. The mapping was based on the "minimal duplication" principle, selecting fields with clear research contribution (such as 245 for title, 340 for material, 561 for provenance, and 700 for authors/scribes).

For global interoperability, the HMO schema was aligned to equivalent properties in Wikidata where appropriate (e.g., P8189 for National Library identifier, P186 for material). In complex structures such as anthologies (field 505), the model creates separate Work and Expression entities for each composition, linked to the manuscript through semantic part-whole relationships. The mapping is not merely technical format conversion, but a conceptual decision defining which textual data become entities in the graph and which are candidates for future reconciliation against external authority repositories.

Technical Implementation and Validation Workflow

Model implementation was carried out through a pipeline that converted MARC records to RDF format. The model was developed and inspected in the Protégé environment using Protégé-aligned OWL QA checks. The resulting graphs were then queried with SPARQL for the research-use scenarios reported below.

At the schema level, properties for external identifiers (viaf_id, wikidata_id, owl:sameAs) are defined, preparing the infrastructure for future reconciliation workflows. In the current pilot, however, a MARC-only policy is maintained: these fields remain unpopulated, marking a direction for future enrichment rather than an accomplished step. Consistent with this policy, the accompanying SHACL shapes intentionally over-constrain some parts of the model—requiring, for example, cidoc-crm:E56_Language instances rather than literal language tags—so that closed-world conformance reporting can expose unfinished authority enrichment. The repository materials include SPARQL validation outputs confirming that the ontology supports the required structures, while the SHACL conformance materials expose both enrichment gaps and source-data quality issues for inspection. Figure 1 illustrates the overall conversion pipeline.

To make the paper inspectable under SWJ's open-review expectations, the repository materials are organized around a minimal reviewer workflow. A reviewer can inspect the ontology and controlled vocabularies directly, load the pilot Turtle graphs into a local RDF environment, run the included sample SPARQL queries, and compare the observed outputs with the validation reports cited in this paper. All empirical claims in the main text are restricted to results that can be checked from those materials.

The HMO Ontology: Model Overview

The Three-Layer Architecture

The HMO ontology is built as a three-layer architecture (Figure 2), distinguishing between three complementary description levels of manuscripts: a cataloging-bibliographic layer, a philological-textual layer, and an

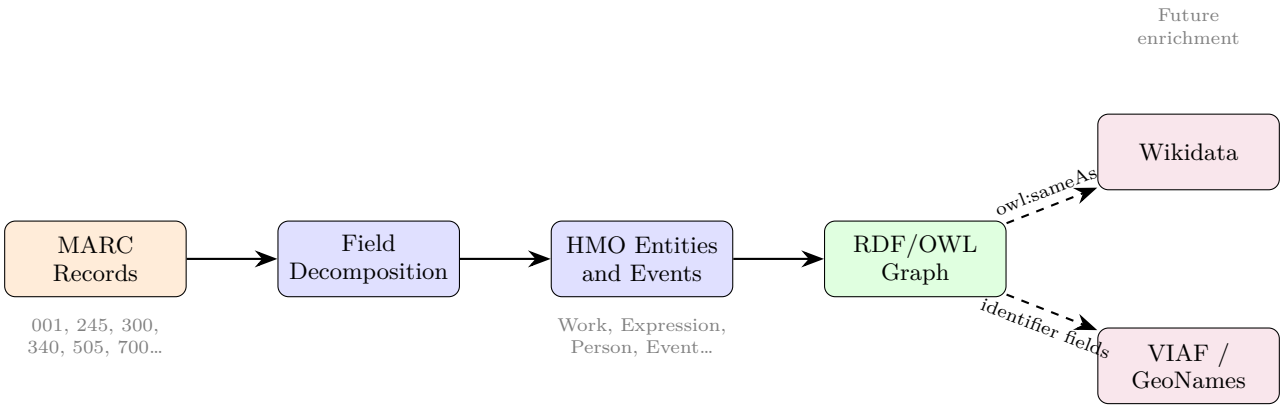

Figure 1. Figure 1. The MARC-to-HMO conversion pipeline. Solid arrows represent the current implementation; dashed arrows indicate schema-level LOD integration infrastructure defined for future enrichment.

epistemological layer. This distinction is designed to enable, on the one hand, compatibility with international standards of bibliographic description and knowledge representation, and on the other, to capture the unique research complexity of Hebrew manuscripts, including multi-textuality, dynamics of textual traditions, and gaps between certain and conjectured information.

The cataloging-bibliographic layer represents the manuscript as Work, Expression, and Manifestation Singleton, maintaining compatibility with MARC-based catalogs. The philological layer adds TextTradition and TransmissionWitness entities, enabling representation of which text tradition a copy belongs to and its relationships to other witnesses. The epistemological layer addresses the source of knowledge through AttributionSource and EpistemologicalStatus, while leaving certainty fields available for later enrichment.

The three layers form a system of parallel, interconnected descriptions: each manuscript is simultaneously a bibliographic item, a textual witness, and a focus of knowledge claims with explicit source and status qualifiers (Figure 2). This architecture enables researchers to choose the appropriate resolution level for their question, harnessing established models (CIDOC CRM, LRMoo) alongside the flexibility required for the complex reality of Hebrew manuscripts.

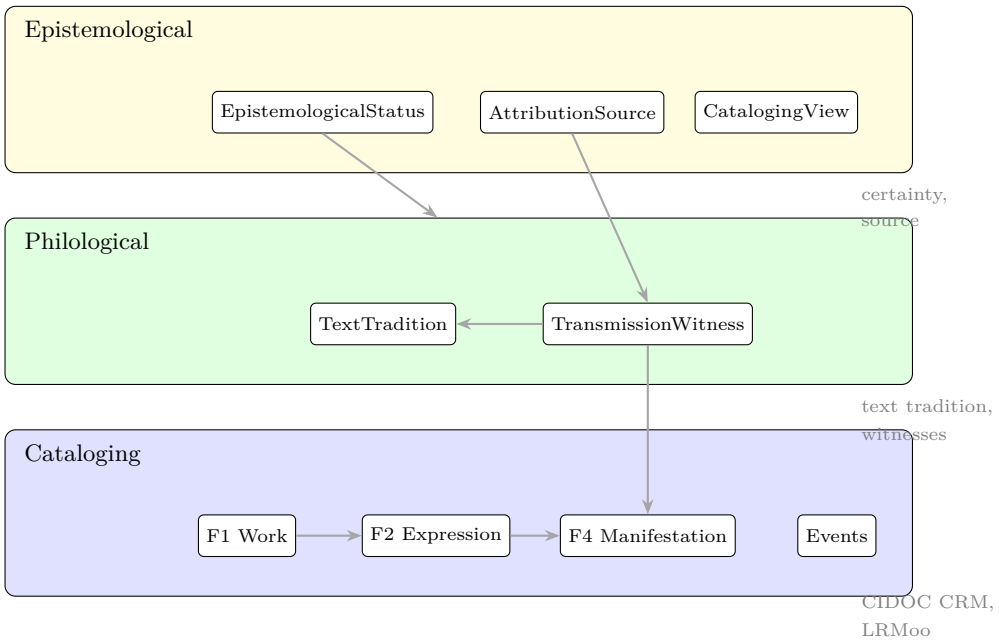

Figure 2. Figure 2. The HMO three-layer architecture. Each manuscript is simultaneously represented at the cataloging-bibliographic level (Work, Expression, Manifestation Singleton), the philological level (TextTradition, TransmissionWitness), and the epistemological level (AttributionSource, EpistemologicalStatus). Arrows indicate cross-layer relationships.

Physical and Codicological Structure: BU-CU-PU

One of the central challenges in describing manuscripts is accurate representation of their internal structure—unified volumes versus complex volumes, quires bound together at a later stage, and multiple hands within the

same work. For this purpose HMO adopts the methodology formulated by Beit-Arié, distinguishing between three physical-codicological description levels (Figure 3): Bibliographic Unit (BU), Codicological Unit (CU), and Paleographical Unit (PU) (Beit-Arié, 2022). A Bibliographic Unit (BU) is the entire physical volume bound under a single shelf mark; a Codicological Unit (CU) is a distinct part created at a specific time and place; a Paleographical Unit (PU) is a part written by a specific scribe's hand.

Rather than implementing BU-CU-PU as a rigid subclass hierarchy, the model uses composition relationships (is_composed_of, forms_part_of) linking separate classes. This modular approach enables flexible representation: a simple manuscript as a single CU, or a complex volume as a BU linked to multiple CUs and PUs. Queries can target each unit level independently, supporting different types of codicological and paleographic analysis.

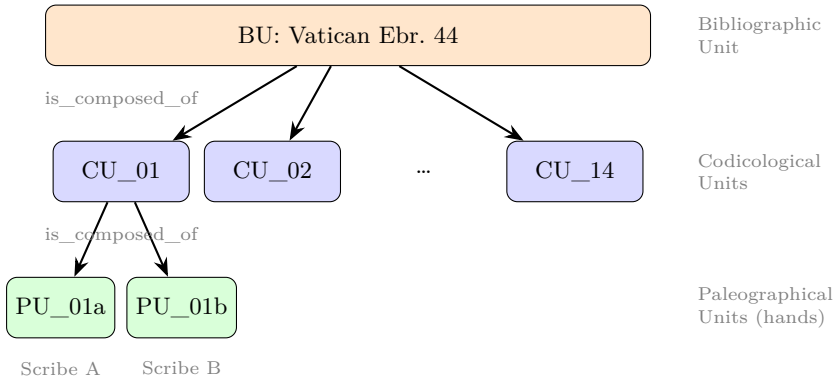

Figure 3. Figure 3. The BU-CU-PU composition model, illustrated with Vatican Ebr. 44. The Bibliographic Unit (BU) is composed of fourteen thematic Codicological Units (CU_01–CU_14) together with a base CU_main, each potentially containing multiple Paleographical Units (PUs) representing distinct scribal hands.

Event-Centric Modelling

Following CIDOC CRM principles, HMO adopts an event-based approach: rather than viewing the manuscript as a static object, the model describes its "life biography" as a sequence of events linking objects, people, places, and times (Bekiari et al., 2021; Doerr, 2003). Production events (E12 Production, F28 Expression Creation) link to Person (author, scribe), Place, and TimeSpan. Transfer events (E8 Acquisition, E10 Transfer of Custody) describe ownership chains. Digital-access links connect physical manuscripts to digital representations, and reference events capture mentions in catalogs and research (Bekiari et al., 2021; Burrows, 2018, 2022). This event model enables queries such as "all manuscripts copied in Candia in the fifteenth century and later acquired in Rome," connecting bibliographic, philological, and epistemological information within a single queryable graph.

Integration with Linked Open Data

HMO implements three complementary LOD integration mechanisms. First, the schema defines properties for external identifiers (VIAF, Wikidata QID, GeoNames); local identifiers such as NLI ID (P8189) and Sfardata ID are extracted from MARC. Second, formal owl:equivalentProperty declarations link HMO properties to Wikidata counterparts (e.g., has_number_of_folios ≡ P792, has_material ≡ P186, has_script_type ≡ P6573), preparing the model for future cross-graph querying (Evans, 2019; Hastings, 2015). Third, owl:sameAs links are defined for future entity-level linking to Wikidata and other repositories (Burrows, 2018, 2022; Koho et al., 2021).

In the current pilot, a MARC-only policy is maintained: external identifier fields (viaf_id, wikidata_id, owl:sameAs) are defined at the schema level but remain unpopulated, marking a direction for future enrichment rather than an accomplished step.

Evaluation

Coverage and Modelling Adequacy

The six-manuscript pilot was designed as a coverage-oriented adequacy test of whether HMO can represent the range of structures and scholarly situations that recur in Hebrew-manuscript cataloging, not as a claim of statistical representativeness for the full corpus. Table 1 summarizes the instantiated cases. The two primary cases are Barberini Or. 82 and Vatican Ebr. 44, because together they test the two structural extremes addressed by the ontology: a unified codex and a complex assembled codex. The four additional manuscripts were selected to probe distinct challenge types that those extremes do not capture on their own: anthology ordering, provenance chains, text-tradition entities, and multiple-hand scenarios.

Table 1. Pilot coverage and modelling adequacy across the six instantiated manuscripts.

Manuscript	Primary modelling challenge	Modeled successfully	Key HMO structures exercised
Barberini Or. 82	Unified codex with ten textual entries in one physical volume	Yes	1 BU, 1 CU, 1 PU; 10 Work–Expression pairs; TransmissionWitness; E12 Production; explicit attribution-source and epistemological-status framing.
Vatican Ebr. 44	Complex assembled codex with many thematic units and multiple hands	Yes	1 BU; 15 CUs; 3 PUs; anthology ordering; one production event with multiple scribes/script types; transfer and custody chain.
Parma 3122	Large anthology with ordered textual parts and folio ranges	Yes	1 BU; 16 CUs; has_anthology_position; ordered expressions; folio-based decomposition.
Huntington 115	Provenance-rich manuscript	Yes	Transfer_of_Custody event; provenance extraction from MARC 561-like data.
Oppenheimer 129	Closely related textual witnesses	Yes	5 CUs; 5 works; 5 TransmissionWitness entities linked to 5 TextTradition entities.
Jerusalem 8210	Compact manuscript with multiple hands and repeated transfer events	Yes	has_multiple_hands; 4 Transfer_of_Custody events; compact unit structure with ownership history.

Taken together, the pilot shows that HMO can represent both simple and highly composite manuscript structures within one model. It also shows that the ontology is not limited to physical decomposition alone: the pilot instantiates philological entities such as TextTradition and TransmissionWitness, and it records epistemological qualifiers at the level of attribution source and epistemological status. For the bounded claims of this paper, the relevance of the pilot lies in this deliberate coverage of structural and scholarly edge cases rather than in sample size alone. This answers the first research question in the affirmative for the pilot scope: HMO captures the material and textual complexity of Hebrew manuscripts more explicitly than MARC records can, while retaining a traceable mapping back to the original catalog fields.

Query Usefulness: What Becomes Easier than in MARC

The practical value of HMO lies not only in ontology design but in the kinds of manuscript-research questions that can be asked once catalog records are transformed into an entity-event graph. Table 2 lists representative competency queries supported by the six-manuscript pilot. These queries are formulated at the level of scholarly use, rather than as infrastructure demonstrations only.

Table 2. Representative research queries that are difficult in MARC but directly supported in HMO.

Research question	HMO graph pattern	Pilot output	Why MARC is insufficient
Which manuscripts in the pilot contain multiple hands?	BU/CU/PU structure + has_multiple_hands	Hunt. 115; Vatican Ebr. 44; Jerusalem 8210	In MARC this information is scattered across notes and cannot be queried as an explicit structural relation.
Which codices contain more than ten codicological units?	BU linked to multiple CUs via is_composed_of	Vatican Ebr. 44; Parma 3122	MARC records do not expose codicological units as countable entities.

Table 2. (continued)

Research question	HMO graph pattern	Pilot output	Why MARC is insufficient
Which witnesses belong to a shared textual tradition?	TransmissionWitness ↔ TextTradition links	Shared traditions across Parma 3122 and Vatican Ebr. 44	MARC titles and notes do not cleanly separate work, expression, and witness levels.
Which manuscripts show repeated transfer-of-custody sequences?	Event-centric provenance using E8/E10 relations	Jerusalem 8210; Vatican Ebr. 44	Provenance notes in MARC are usually free text and difficult to normalize into event sequences.
Which anthologies preserve ordered internal textual positions?	Expression + AnthologyPosition + folio ranges	Oppenheimer 129; Parma 3122; Vatican Ebr. 44	MARC field 505 can list contents, but not as reusable ordered entities with explicit graph relations.
Which manuscripts copied in a given place and period are also tied to named scribes?	E12 Production + Person + Place + TimeSpan	Barberini Or. 82 (Candia, 1407, named scribe)	MARC can store these data in dispersed fields, but not as one joinable event pattern across manuscripts.

The value of these queries is not only technical. Each query corresponds to a humanities-relevant research task: identifying anthology structure, reconstructing provenance, distinguishing witnesses from works, or studying scribal practices. The graph therefore improves practical queryability rather than merely providing a new serialization of existing records. This answers the second research question within the pilot scope: the HMO transformation yields an inspectable gain for manuscript research queries.

Three compact examples illustrate the evidential status of this claim. First, the multiple-hands query returns Hunt. 115, Vatican Ebr. 44, and Jerusalem 8210 because the pilot instantiates an explicit has_multiple_hands marker and, in the Vatican case, PU-level differentiation; in MARC, the same information remains dispersed across descriptive notes. Second, the codicological-unit count query returns Vatican Ebr. 44 and Parma 3122 as codices with more than ten CUs because is_composed_of links make those units countable entities; MARC exposes no directly countable CU layer. Third, a repeated-transfer query retrieves Jerusalem 8210 and Vatican Ebr. 44 as manuscripts with multi-step transfer-of-custody sequences because ownership events are instantiated as E8/E10 patterns rather than left in free-text provenance notes. These are still pilot-scale demonstrations, but they show that the gain is operational and inspectable rather than hypothetical.

To keep this claim reproducible rather than impressionistic, the repository materials include executable SPARQL competency queries together with the pilot RDF graphs on which these examples are based. We therefore treat query usefulness here as an inspectable demonstration that specific manuscript distinctions become operational in graph form, not as a benchmark-style claim about large-scale retrieval performance across the entire catalog.

Validation, Quality, and Released Evidence

Taken together, the released evidence forms a three-step ladder: instantiated pilot cases for modelling adequacy and query usefulness, a validation suite for schema support, and a larger feasibility run for workflow behavior at scale. Validation was performed at both schema and data-support levels. At the schema-support level, the primary finding is that the released SPARQL validation suite executes all 37 checks successfully over the ontology files and controlled vocabularies, covering ten groups of expert concerns: certainty/confidence, variants, anthology and multi-volume structure, textual overlap, scribal interventions, lost manuscripts, text tradition, fine-grained textual location, canonical hierarchy links, and foreign-unit marking. Within that set, the results report 14 PASS, 23 NO DATA, and 0 ERROR over a tested graph of 2,155 triples. The important point for review is not the raw shorthand alone, but that the full suite runs without schema failure and exposes the intended classes and properties for inspection.

At the instance-population level, the 23 NO DATA outcomes are expected because those queries target optional patterns that are intentionally defined in the schema but not yet populated in the current pilot release. Quality control also included ontology inspection in Protégé, OWL QA review, and SHACL-based conformance reporting. The released conformance materials indicate a mixed picture: some non-conformance results arise from intentional authority-enrichment gaps under a closed-world validation strategy, such as requiring controlled language entities where the current sample still contains simpler population patterns, while others expose source-data quality

problems such as malformed URLs or manuscripts lacking linked Expression instances. These materials are therefore better interpreted as a transparent roadmap for enrichment and cleanup than as evidence of broken model structure. Accordingly, the validation evidence should be read as evidence of schema support, inspectability, and workflow transparency, not as evidence that every optional modelling pattern has already been populated in the pilot data.

At the workflow-feasibility level, the conversion pipeline was also exercised on a separate 10,000-record feasibility set. That larger run generated 469,654 triples and produced one SHACL error caused by a malformed digital URL together with 306 warnings, most of them involving manuscripts that lacked linked Expression instances because the source MARC records were themselves incomplete. We do not treat that 10,000-record run as a substitute for the detailed pilot evaluation, and we do not use it to claim large-scale scholarly validation. Its role in this paper is narrower: it provides evidence that the transformation workflow can run at substantially larger scale and that its main observed issues are data-quality problems in the source records rather than ontology-schema failures.

The paper's evidence base is intentionally inspectable by reviewers. The repository materials include the ontology serializations, controlled vocabularies, crosswalk materials, pilot RDF, validation report, sample SPARQL queries, and implementation materials. In SWJ terms, this directly addresses inspectability and reusability: the model is not only described in prose but provided in a form that can be examined, queried, and reused. The larger feasibility run provides preliminary evidence of practical utility, even though broader community uptake remains future work.

Interoperability Status: Implemented vs. Future-Ready

The interoperability contribution of HMO is staged. What is already implemented in the present release is schema-level alignment: local manuscript identifiers are extracted from MARC, HMO properties are formally aligned to selected Wikidata properties through owl:equivalentProperty, and slots for VIAF, Wikidata, GeoNames, and owl:sameAs are built into the ontology and conversion workflow. What is not yet implemented in the pilot RDF is populated entity-level external linking. This distinction is important because it means that the current paper demonstrates interoperability readiness and a clear reconciliation path, but not yet a completed cross-graph linked-data publication.

This distinction helps locate the present contribution precisely. HMO contributes an evaluated domain model and workflow that make future reconciliation possible without claiming that reconciliation has already been completed. For SWJ review purposes, the paper therefore presents a conservative and verifiable interoperability claim: HMO currently defines and tests the semantic infrastructure required for future linking to external authority and knowledge-graph ecosystems.

Table 3. Implemented in the present release vs. future-ready infrastructure.

Aspect	Implemented and evidenced here	Defined for future enrichment, not yet populated in pilot RDF
Ontology and workflow materials	OWL/TTL ontology, controlled vocabularies, SHACL shapes, pilot RDF, crosswalk, sample queries, and replication materials are present in the repository materials used for this paper	Ongoing version expansion beyond the current repository snapshot.
Pilot manuscript modelling	Six instantiated manuscripts covering unified codices, complex codices, anthologies, multiple hands, provenance, and text tradition	Broader corpus-wide population beyond the pilot.
Validation evidence	37 SPARQL validation checks execute successfully over released ontology files; SHACL conformance materials are released	Additional instance-level passes for currently unpopulated optional patterns.
Interoperability	Schema-level alignment points to Wikidata/VIAF/GeoNames and explicit slots for owl:sameAs	Populated entity-level authority reconciliation and external links in pilot or corpus-scale RDF.

Table 3. (continued)

Aspect	Implemented and evidenced here	Defined for future enrichment, not yet populated in pilot RDF
Epistemic modelling	Source-attribution and epistemological-status mechanisms are implemented and instantiated in pilot data; certainty fields are defined at schema level but unpopulated in the pilot	Broad, systematic population of certainty and attribution fields across the dataset.

Comparison with Prior Work

To clarify the novelty claim, Table 4 compares HMO with the prior systems most relevant to this paper's scope. The point of the comparison is not to rank systems feature by feature, but to clarify the different burden of contribution carried by HMO relative to adjacent manuscript-LOD efforts. The comparison is therefore limited to features discussed explicitly in the respective publications and to the aspect most relevant here: whether the system offers a manuscript model suited to Hebrew-manuscript research rather than a general aggregation framework alone. Only features explicitly described in the cited sources are counted as documented.

Cells marked "Doc." indicate that the capability is explicitly described in the cited source; "N/D" indicates that it is not documented in the cited publication used for comparison. These labels should not be read as claims about every later version or component of the broader project.

Table 4. Comparison of HMO with the most relevant prior approaches.

Approach	Primary focus	Corpus-specific model	BU-CU-PU granularity	Epistemic layer	Anthology ordering	Event provenance	Released evidence
MMM (Burrows, 2018, 2022; Koho et al., 2021)	Cross-repository migration and provenance aggregation	N/D	N/D	N/D	N/D	Doc.	Partial
MMDIO (Ferooz, 2025)	Polymorphic semantic integration across manuscript collections	N/D	Partial	N/D	Partial	Partial	Partial
HMO (this paper)	Hebrew-manuscript ontology + evaluated transformation workflow	Doc.	Doc.	Doc.	Doc.	Doc.	Doc.

The comparison clarifies that HMO addresses a different burden of contribution from MMM and MMDIO. Those projects are important comparators, but their center of gravity is cross-collection aggregation and semantic integration. HMO occupies a complementary niche: it provides a corpus-specific conceptual model for Hebrew manuscripts that combines codicological granularity, philological witness modeling, an explicit epistemological layer, and a released transformation-and-validation workflow in one inspectable package.

Discussion

HMO's Contribution to Manuscript Scholarship

The main scholarly contribution of HMO is that it turns manuscript descriptions from record-level text into queryable entities and relationships while preserving the distinctions manuscript scholars actually need. The BU-CU-PU model permits explicit representation of codicological granularity, while TextTradition and TransmissionWitness separate textual-witness analysis from bibliographic description. The epistemological layer

adds a further manuscript-scholarship contribution: dates, attributions, and identifications need not appear as flat facts only, but can be connected to sources and status categories, with certainty support defined in the schema for later enrichment. This combination responds directly to the needs of Hebrew-manuscript scholarship, where material evidence, textual transmission, and scholarly judgment frequently intersect (Beit-Arié, 2022; Prebor et al., 2020a; Richler, 2014; Sirat, 2002; Zhitomirsky-Geffet et al., 2020).

HMO's Contribution to Digital Humanities

For digital humanities, HMO's contribution is methodological as much as conceptual. The paper does not present an ontology in isolation; it presents a tested workflow that transforms MARC records into a graph that supports non-trivial research queries. The evaluation shows that the gain over MARC is practical rather than merely formal: the graph enables integrated questions about structure, provenance, textual transmission, and scribal activity that otherwise remain fragmented across notes, repeated fields, and local catalog conventions. The schema-level connection to Wikidata and related authority ecosystems further positions the ontology as a bridge between manuscript-specific description and future graph integration.

Relative to recent work on semantic integration for manuscript collections (Ferooz, 2025; Ferooz & Palmirani, 2024), HMO's novelty lies in the cumulative conjunction of features rather than in any one element taken alone: corpus-specific Hebrew-manuscript semantics, explicit BU-CU-PU granularity, a philological witness layer, an epistemological layer, and a released transformation-and-validation workflow organized for direct inspection in review.

Challenges and Limitations

HMO implementation highlights several challenges (summarized in Table 5). At the data level, MARC records suffer from field inconsistency, variant name forms, and gaps in authority-file links; only about a quarter of persons in the catalog were identified in global authority files (Zhitomirsky-Geffet et al., 2020). At the conceptual level, decisions about entity granularity (e.g., what constitutes a separate Work) require philological examination and are not always amenable to uniform enforcement. Domain extensions must remain monotonic to preserve compatibility with CIDOC CRM and LRMoo (Bekiari et al., 2015, 2021; Doerr, 2003). Building and maintaining a full RDF graph for an extensive catalog also requires institutional commitment, ongoing curation, and systematic authority reconciliation.

Three limitations are especially important for interpreting the present contribution. First, the evaluation remains pilot-scale at the level of fully instantiated case studies, even though the broader conversion workflow has been exercised on a larger set of records. Second, external authority reconciliation is not yet populated in the pilot RDF, so interoperability is demonstrated here as schema readiness rather than completed linking. Third, attribution-source and epistemological-status mechanisms are populated in the current sample data, whereas certainty fields remain implemented at schema level but unpopulated in the pilot. Fourth, we do not yet provide a public SPARQL endpoint; reviewers and future users inspect the released Turtle dumps and sample queries instead. These limitations constrain what can be claimed today, but they do not undermine the central result that the ontology and workflow already support more explicit and research-useful representation than MARC alone.

Table 5. Summary of modelling challenges, systematic solutions, and implementation status in HMO.

Identified Challenge	Systematic Solution	Implementation Status
Confusion between codicological unit and content unit	Conceptual separation between structural and textual layers	Implemented: explicit separation between BU/CU/PU and Work/Expression; flexible composition model instead of rigid hierarchy.
Certainty levels and distinction between fact and researcher determination	Epistemological layer with attribution source and certainty level	Partially implemented: has_epistemological_status, has_attribution_source exist; has_certainty/certainty_percentage fields defined in schema, not yet populated in sample.
Expansion of event space	Adoption of full event model per CIDOC CRM	Implemented: active event model with E12 Production, E8 Acquisition, E10 Transfer_of_Custody and links to manuscripts.

Table 5. (continued)

Identified Challenge	Systematic Solution	Implementation Status
Multi-volume mechanism (Vat. 44)	Grouping volumes in MultiVolumeSet with reversible relationships	Implemented: MultiVolumeSet, is_volume_of, has_volume with bidirectional consistency in data sample.
Intentional anthology structure (Parma 3122)	Canonical pattern based on AnthologyPosition	Implemented: has_anthology_position → AnthologyPosition → anthology_order.
Limitations of rigid hierarchy (Oppenheimer)	Transition to flexible composition relationships instead of subclass hierarchy	Implemented: representation based on forms_part_of/is_composed_of enables multi-unit complexity without breaking model.
Overlap between hands and textual units (Huntington 115)	Definition of overlap and duplication relationships at schema level	Implemented in schema; has_textual_overlap_with, duplicates_part_of, overlap_folio_range exist; currently 0 instances in data sample.
Margins and relation to main text (London Sifrei Toviah)	Definition of marginalia relationships at schema level	Implemented in schema; has_marginalia/is_marginalia_of exist; currently 0 instances in data sample.
Foreign units that are not core (Jerusalem 8210)	Marking foreign additions through dedicated unit type	Implemented in schema; is_foreign_addition and UnitStatusType exist; currently 0 instances in data sample.
Lost manuscripts and copy fidelity (London 9599)	LostManuscript model and is_copy_of_lost relationship	Implemented in schema; is_copy_of_lost with LostManuscript entity exists; currently 0 instances in data sample.
Canonical hierarchies (Bible/Mishnah/Mishneh Torah/Shulchan Aruch)	Definition of hierarchy types and canonical coverage relationships	Implemented in schema; CanonicalHierarchyType and canonical_hierarchy/covers_canonical_range relationships exist; currently 0 instances in data sample.

Conclusions and Future Work

Summary of Main Contributions

The HMO ontology offers an evaluated framework for representing Hebrew manuscripts as entity-event graphs rather than as record-bound catalog descriptions. It combines the CIDOC CRM and LRMoo super-models with domain concepts such as BU-CU-PU, TextTradition, TransmissionWitness, and explicit epistemological qualification. Within the scope tested here, the released evidence supports three linked claims: a coverage-oriented six-manuscript pilot shows modelling adequacy and query usefulness across the key structural and scholarly edge cases targeted by the paper, the validation suite shows schema support for the required structures and makes optional-yet-unpopulated patterns inspectable, and the 10,000-record run shows workflow feasibility beyond the pilot. Together, these results show that the model captures both simple and composite manuscripts, supports research questions that are difficult to formulate in MARC alone, and releases its evidence base in reusable open form.

HMO therefore makes two linked contributions. First, it provides a manuscript-domain ontology tailored to Hebrew materials and the scholarly problems they raise. Second, it provides a documented transformation and validation workflow showing how an established catalog tradition can be brought into a Linked Open Data environment without overstating what has already been achieved. The paper's claim is therefore not built on six manuscripts alone, but on the integrated package of corpus-specific modelling, operational transformation design, inspectable query demonstrations, and released validation materials. The present release demonstrates evaluated

modeling adequacy, schema support for the semantic framework, and interoperability readiness at schema level; future releases can build on this base to add authority reconciliation, larger-scale population, and richer public query services.

Future Directions

Going forward, one of the central challenges and opportunities is extending HMO implementation beyond the pilot described here to as full coverage as possible of the National Library Hebrew Manuscripts catalog and related collections. Such extension would enable larger-scale testing, expose additional metadata inconsistencies, and improve the ontology in continued dialogue with researchers and catalogers. In parallel, external access layers can be developed—visual search interfaces, network-analysis dashboards, geographic maps, and timelines—that mediate between the graph and digital-humanities researchers who are not experts in RDF or SPARQL.

A further direction is deepening integration with international Linked Open Data projects in the manuscript domain, including initiatives focusing on premodern manuscript metadata and their integration into Wikidata and other graphs. Integrating HMO with graphs such as Mapping Manuscript Migrations, and with similar projects in other language communities, would enable for the first time systematic comparison of copying patterns, provenance, and scribe networks in Hebrew and non-Hebrew collections within a single conceptual framework. Finally, there is room to develop theoretical research lines examining how ontological models such as HMO affect the ways manuscripts are represented and understood in general—for example, in the tension between "new philology" and older catalog traditions—and how similar principles can be applied to other corpora in the digital humanities.

Funding

This research was supported by the Israeli Ministry of Innovation, Science and Technology (Grant No. 1001706678).

Acknowledgments

The authors wish to thank Prof. Moshe Lavee and Dr. Eliezer Baumgarten for their valuable contributions to this research.

Resource Availability

To support direct inspection during review, the repository materials accompanying this paper currently include the following:

- Ontology and shapes. The repository contains the HMO ontology and related schema materials, including ontology/hebrew-manuscripts.ttl, ontology/controlled-vocabularies.ttl, and ontology/shacl-shapes.ttl.
- Pilot data. The six-manuscript pilot graphs discussed in this paper are present as Turtle files under data/output/, including the merged inspection graph data/output/test_manuscripts_merged.ttl.
- Validation materials. The repository contains the schema-validation summary in docs/sparql-validation-results.md, the Protégé-oriented QA summary in docs/PROTEGE_QA_VERIFICATION_REPORT.md, and the 10,000-record feasibility validation report in data/marc-subset-10k.validation-report.md.
- Queries and verification scripts. Sample competency and verification queries are included in docs/, together with executable local verification scripts (docs/run_verify_claims.py and docs/run_sparql_verification.py) that reproduce the pilot-level checks reported in this paper.
- Crosswalk and implementation materials. The repository contains MARC-to-HMO mapping documentation in mappings/ and the conversion/validation code in scripts/ and converter/.
- Local inspection workflow. Reviewers can inspect the ontology and controlled vocabularies, load the pilot Turtle graphs into a local RDF environment, execute the included sample queries, and compare the outputs with the validation reports cited in this paper. We do not currently provide a public SPARQL endpoint.

Use of Generative AI

Generative AI tools were used only for assistive tasks during manuscript preparation, specifically literature search support, LaTeX formatting assistance, and language editing suggestions. They were not used as a source of scholarly claims, analysis, interpretation, or authorship. All manuscript content, argumentation, and verification were reviewed and revised by the authors, who take full responsibility for the accuracy and integrity of the submitted work.

Conflict of Interest

The authors declare no conflicts of interest.

References

Aalberg, T., Riva, P., & Zumer, M. (2024). Lrmoo: Object-oriented definition and mapping from the IFLA library reference model (tech. rep.). IFLA Bibliographic Conceptual Models Review Group and ICOM CIDOC CRM SIG. Retrieved February 6, 2026, from https://repository.ifla.org/items/94aedb49-2d6e-4a6d-9974-f33abb7e3c0e

Beit-Arié, M. (2022). Hebrew codicology. Israel Academy of Sciences; Humanities. https://doi.org/10.25592/uhhfdm.9349

Bekiari, C., Bruseker, G., Doerr, M., Ore, C.-E., Stead, S., & Velios, A. (2021). Definition of the CIDOC conceptual reference model, version 7.1.1 (tech. rep.). ICOM/CIDOC CRM Special Interest Group. Retrieved February 6, 2026, from https://cidoc-crm.org/sites/default/files/cidoc_crm_v.7.1.1_0.pdf

Bekiari, C., Doerr, M., Le Boeuf, P., & Riva, P. (2015). Definition of FRBRoo: A conceptual model for bibliographic information in object-oriented formalism, version 2.4 (tech. rep.). International Working Group on FRBR and CIDOC CRM Harmonisation. Retrieved February 6, 2026, from https://www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf

Bermès, E. (2015). Following the user's flow in the digital pompidou. In H. F. Cervone & L. G. Svensson (Eds.), Linked data and user interaction (pp. 19–30). De Gruyter Saur. https://doi.org/10.1515/9783110317008-004

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific American, 284(5), 34–43. https://doi.org/10.1038/scientificamerican0501-34

Burrows, T. (2018). Connecting medieval and renaissance manuscript collections. Open Library of Humanities, 4(2), 1–23. https://doi.org/10.16995/olh.269

Burrows, T. (2022). Linked open data and medieval studies: Some lessons from the mapping manuscript migrations project. International Journal of Humanities and Arts Computing, 16(1), 64–77. https://doi.org/10.3366/ijhac.2022.0277

Doerr, M. (2003). The cidoc conceptual reference model: An ontological approach to semantic interoperability of metadata. AI Magazine, 24(3), 75–92. https://doi.org/10.1609/aimag.v24i3.1720

Dunsire, G. (2012). Linked data for manuscripts in the Semantic Web (Pre-print) (Presented at the Summer School in the Study of Historical Manuscripts, Zadar, Croatia, 28 September 2011). Self-published. Edinburgh. Retrieved February 6, 2026, from http://www.gordondunsire.com/pubs/docs/LinkedDataForManuscripts.pdf

Evans, J. (2019). Treasured manuscript collection gets the wikidata treatment. Retrieved February 6, 2026, from https://web.archive.org/web/2024/https://blog.library.wales/treasured-manuscript-collection-gets-the-wikidata-treatment/

Ferooz, F. (2025). Legal knowledge modelling of medieval manuscript of history of law: Linked open data of cultural heritage [Doctoral dissertation]. University of Luxembourg and University of Bologna. Retrieved April 15, 2026, from https://orbilu.uni.lu/handle/10993/66022

Ferooz, F., & Palmirani, M. (2024). An ontological framework for integrating the heterogeneous medieval manuscript resources: A case study of progetto irnerio and mosaico. In A. Salatino, M. Alam, F. Ongenae, S. Vahdati, A.-L. Gentile, T. Pellegrini, & S. Jiang (Eds.), Knowledge graphs in the age of language models and neuro-symbolic ai (pp. 403–419, Vol. 60). IOS Press. https://doi.org/10.3233/SSW240032

Hastings, R. (2015). Feature: Linked data in libraries: Status and future direction. Computers in Libraries, 35(9), 12–16. Retrieved April 15, 2026, from https://www.infotoday.com/cilmag/nov15/Hastings--Linked-Data-in-Libraries.shtml

Koho, M., Burrows, T., Hyvonen, E., Ikkala, E., Page, K., Ransom, L., Tuominen, J., Emery, D., Fraas, M., Heller, B., Lewis, D., Morrison, A., Porte, G., Thomson, E., Velios, A., & Wijsman, H. (2021). Harmonizing and publishing heterogeneous premodern manuscript metadata as linked open data. Journal of the Association for Information Science and Technology, 73(2), 240–257. https://doi.org/10.1002/asi.24499

Landis, C. (2019). Linked open data in libraries. In K. J. Varnum (Ed.), New top technologies every librarian needs to know (pp. 3–15). American Library Association.

Le Boeuf, P. (2012). Modeling rare and unique documents using frbroo/cidoc crm. Journal of Archival Organization, 10(2), 96–106. https://doi.org/10.1080/15332748.2012.709164

Prebor, G., Zhitomirsky-Geffet, M., & Miller, Y. (2020a). A multi-dimensional ontology-based analysis of the censorship of hebrew manuscripts. Digital Humanities Quarterly, 14(1). Retrieved February 6, 2026, from https://www.digitalhumanities.org/dhq/vol/14/1/000442/000442.html

Prebor, G., Zhitomirsky-Geffet, M., & Miller, Y. (2020b). A new analytic framework for prediction of migration patterns and locations of historical manuscripts based on their script types. Digital Scholarship in the Humanities, 35(2), 441–458. https://doi.org/10.1093/llc/fqz038

Richler, B. (2014). Guide to hebrew manuscript collections (2nd ed.). Israel Academy of Sciences; Humanities.

Riva, P., Le Boeuf, P., & Zumer, M. (2017). Ifla library reference model: A conceptual model for bibliographic information. IFLA.

Sirat, C. (2002). Hebrew manuscripts of the middle ages. Cambridge University Press.
Ullah, I., Khusro, S., Ullah, A., & Naeem, M. (2018). An overview of the current state of linked and open data in cataloging. Information Technology and Libraries, 37(4), 47–80. https://doi.org/10.6017/ital.v37i4.10432
Weitz, J., Toves, J., Vizine-Goetz, D., Naught, N., & Bremer, R. (2016). Mining MARC's hidden treasures: Initial investigations into how notes of the past might shape our future. Journal of Library Metadata, 16(3-4), 166–180. https://doi.org/10.1080/19386389.2016.1262653
Zeng, M. L. (2019). Semantic enrichment for enhancing LAM data and supporting digital humanities. El profesional de la información, 28(1), e280103. https://doi.org/10.3145/epi.2019.ene.03
Zhitomirsky-Geffet, M., Prebor, G., & Miller, I. (2020). Ontology-based analysis of the large collection of historical hebrew manuscripts. Digital Scholarship in the Humanities, 35(3), 688–719. https://doi.org/10.1093/llc/fqz058

Supporting Information

Additional supporting information can be found in the Supporting Information document accompanying this article. The Supporting Information includes: (A) detailed MARC-HMO-Wikidata crosswalk tables; (B) full class inventories for all HMO, LRMoo, and CIDOC CRM classes used in the model; (C) complete object property and data property tables with domain, range, and Wikidata equivalences; and (D) mapping pattern examples with illustrative SPARQL queries.