ORIGINAL RESEARCH
published: 17 March 2016
doi: 10.3389/fdigh.2016.00003

Toward an Ontopedia for Historical Hebrew Manuscripts

Maayan Zhitomirsky-Geffet and Gila Prebor*

Department of Information Science, Bar-Ilan University, Ramat-Gan, Israel

Historical handwritten Hebrew manuscripts are one of the most unique and authentic witnesses of Jewish culture and thought that survived through the centuries. In order to enable a systematic research of the knowledge embedded in the manuscripts, there is a need for a formal conceptual data model with a high level of semantic granularity, an ontology. We propose to build a dynamic web-based framework that will allow scholars to create, enrich, and consult an "ontopedia" (ontology-based encyclopedia) of Hebrew manuscripts. The framework is based on an ontology especially designed and implemented for this domain and goals. We view a manuscript as a "living entity" and propose to design a new ontological data model of the narrative for a manuscript, stages/milestones in its biography (creation, copying, and acquisition). A sequence of events and places constitutes a timeline of history against which manuscripts, people, and their relationships can be placed. A large-scale automated reasoning based on the ontology will also enable us to construct a semantically rich social network of people and manuscripts, and to compare the effect of time and place on the manuscripts' qualitative characteristics and quantitative distribution.

Keywords: ontology for historical manuscripts, handwritten manuscripts, Hebrew historical manuscripts, semantic web, manuscript biographic ontology

OPEN ACCESS

Edited by:
Peter A. Stokes,
King's College London, UK

Reviewed by:
Marilena Maniaci,
Università degli studi di Cassino e del Lazio Meridionale, Italy
Elli Angelopoulou,
University of Erlangen-Nuremberg, Germany

****Correspondence:***
Maayan Zhitomirsky-Geffet
maayan.zhitomirsky-geffet@biu.ac.il

Specialty section:
This article was submitted to Digital Paleography and Book History, a section of the journal Frontiers in Digital Humanities

Received: *30 August 2015*
Accepted: *26 February 2016*
Published: *17 March 2016*

Citation:
Zhitomirsky-Geffet M and Prebor G (2016) Toward an Ontopedia for Historical Hebrew Manuscripts. Front. Digit. Humanit. 3:3. doi: 10.3389/fdigh.2016.00003

INTRODUCTION

Digitization of national cultural heritage is a rapidly expanding field essential for preserving and maintaining historical data and leveraging its future research. For thousands of years, even after the introduction of printing, Jews used handwritten texts for different purposes. Historical handwritten Hebrew manuscripts are one of the most unique and authentic witnesses of Jewish culture and thought that survived through the centuries. Scholars from various fields increasingly study these manuscripts to reveal historical, linguistic, religious, philosophical, and social aspects of Jewish life in different times and places. These manuscripts shed light on the intellectual, religious, and everyday life of Jews throughout the ages.

Most of the works on digitization of handwritten manuscripts, and in particular on Hebrew manuscripts, tackle the problem of their text transcription usually by introducing various image processing techniques (Wolf et al., 2011). In this study, we address a different challenge presented by these manuscripts: the organization and accessibility of the semantic knowledge imbedded in them.

Currently, the only available digital representation of these manuscripts' metadata is library catalogs. These catalogs are accompanied with search options to retrieve records by a limited number of parameters, such as author, title, date, and subject, while most of the data still remain unsearchable and thus undiscovered. Many important research questions cannot be answered by searching the existing manuscript catalogs, e.g., How many works on specified subjects were composed in a certain period in different countries? What people were involved in their creation and distribution, and whether and how they are related to each other? What historical events could influence these people and their works? Are these original works or copies of older works from other manuscripts or even from printed books? Hence, to enable a systematic research of the knowledge embedded in the manuscripts, there is a need for a formal conceptual data model with a high level of semantic granularity, an ontology, which reflects the various cultural riches stored in Hebrew manuscripts. To the best of our knowledge, there is no formal ontology for the realm of historical handwritten Hebrew manuscripts.

Hence, in this research, our goal is to design an ontological model to reflect all the cultural riches stored in historical Hebrew manuscripts. At the start, we focus on the post-medieval period (sixteenth century and later), because tens of thousands of works that belong to this period are under-explored in the research literature. Most research into Hebrew manuscripts that has been done to date has focused on manuscripts written until 1540 (Prebor, 2015).

The underlying philosophical approach is to view a manuscript as a "living entity" and develop a data model of its life story. Therefore, in this paper, we propose a new ontological model based on events. This is since one of the main and unique characteristics of Hebrew manuscripts is the large amount of changes and transmissions in their biography, which are most naturally represented by events. The events are stages/milestones in the biography of a manuscript, such as creation, copying, acquisition, printing, storing, citing, and censoring. Each event provides a framework that binds together all other objects involved in it and thus reflects the influence and interactions between manuscripts, people, organizations, places, and historical periods. This is the main conceptual difference between our model and the existing ontological models for historical manuscripts [e.g., the Europeana Data Model for manuscript representation: http://dm2e.eu and the ontology for Henri III fine rolls (Vieira and Ciula, 2007)].

The innovation in our approach also pertains to making these intellectual aspects of Hebrew manuscripts available as an online resource for research. The kind and depth of analysis goes far beyond current cataloging practice in its immediate utility for scholarly research with the global accessibility of online resources, and finally to the particular corpus selected for an initial population of the system. The ontology will help convert and extend the data of catalog records representing the manuscripts into linked data by linking them to similar concepts from the external ontologies and vocabularies on the semantic web (LOD[1]). In addition, we perform an analysis of the results applied to the subset of manuscript catalog records and present a methodology for their conversion to ontological statements (triples) in the proposed model. Thus, we show how manuscripts and also people, events, subjects, and places related to them can be classified and interlinked. Consequently, the cultural heritage contained in the manuscripts will become easily accessible and searchable online. This will further enable a large-scale quantitative examination of various aspects of the manuscripts revealing Jewish life and culture in this period. Eventually, we aim to develop interactive tools to dynamically construct the "ontopedia" for the post-medieval manuscripts by querying the constructed ontology. The term "ontopedia" was previously used by a Spanish project (Gamallo, 2014).[2] This project aimed to automatically construct an online encyclopedia by means of open data extraction, and in particular, automatic extraction of RDF-style triples, as facts, from different web sources, such as Wikipedia. Although the goals and methodology of the Spanish project were different from ours, the term "ontopedia" seems to fit well in our framework, in the sense of an online encyclopedia constructed of triples (as facts) found on the web. As a result, a framework for building a dynamic web-based encyclopedia (ontopedia) for historical Hebrew manuscripts based on a rich ontology will be constructed. We expect that the developed methodology and ontological model might be in large parts more generic than just for Hebrew manuscripts and be reusable for manuscripts in general, thus linking the semantic data in the Hebrew and European cultural history sphere in a meaningful way.

Related Work

Hebrew Historical Manuscript Research

The largest collection of Hebrew manuscripts metadata is offered by the catalog of the Department of Manuscripts and Institute of Microfilmed Hebrew Manuscripts (IMHM) in the National Library of Israel.[3] IMHM was founded in 1950 by the government of Israel and has undertaken the task of collecting microfilm copies of all Hebrew manuscripts or manuscripts written in Hebrew characters extant in public and private collections in the world. Today, most of the major Hebrew manuscripts collections are represented in IMHM (Richler, 2014, pp. 103–104). The Hebrew Paleography Project, a joint French–Israeli project founded in 1965 whose goal is to establish a historical typology of medieval Hebrew book production and consumption, has only addressed the dated manuscripts written until 1540 (5% of them).A summation of this decades-long project and a historical and comparative typology of Hebrew medieval codices can be found in two works that have been written lately by the two founders of the project (Sirat, 2013; Beit-Arié, 2014). The codicological database of the Hebrew Paleography Project and the Israeli Academy of Sciences and Humanities SfarData databases will be integrated into the website of the National Library of Israel (SfarData, n.d.).[4] The physical features of the dated manuscripts (parchment, quires, and writing), given now in Sfar-Data, are the tools we use to date, localize, and study the texts of the medieval non-dated manuscripts (Beit-Arié, 2014).

As opposed to what was covered in the Hebrew Paleography Project, in this article, we focus on the other part of the Hebrew manuscript collection, the later Hebrew manuscripts, created after 1540. Our research concentrates mainly on quantification, classification, and analysis of the text genres, which have survived in different areas and at different times, as well as on a

[1] http://linkeddata.org/

[2] http://fegalaz.usc.es/ontopediaweb/

[3] http://web.nli.org.il

[4] http://sfardata.nli.org.il/sfardatanew/home.aspx

detailed survey of the historical, social, and cultural context of works rather than on the codicological and paleographic aspects. These manuscripts open a new world that has hitherto not been systematically researched as a corpus. To this end, we propose to use innovative semantic web technologies rather than traditional databases.

Semantic Web

The Semantic web is the vision of the web's inventor Tim Berners-Lee (Berners-Lee et al., 2001). The concept behind the Semantic web is that while online information is accessible, its meaning remains incomprehensible to search engines. Computers can easily locate *search words*, but they do not understand the context within which they appear. The vision of the Semantic web is that the web will transform from a collection of documents comprehensible only to humans to a database, which computers will also be able "to understand," i.e., to process information based on formalized semantics.

Ontologies form the base of the Semantic web's realization. An ontology is a formal vocabulary, a rich semantic model of shared knowledge, which comprises a set of concepts, their definitions, and semantic inter-relationships (Uschold and Gruninger, 1996). Using an ontology, software applications and automatic agents could communicate with each other, to share and exchange information, and perform complex tasks together without human intervention. The W3C organization developed standards for the formal definition of ontologies, such as the RDF/RDFS (Brickley and Guha, 1999) and OWL (McGuinness and van Harmelen, 2004) XML-based languages. The building blocks of these languages are statements or facts on the domain of knowledge. Every statement is a triple of the form: *subject–predicate–object*, which expresses the semantic relationship (predicate) between two concepts (subject and object). Both concepts and relationships among them are defined and uniquely identified by some URI (namespaces). The concepts can be abstract classes or specific objects and then logical inference rules and restraints can be applied to them to induce new relationships that are not explicitly encoded. Furthermore, ontological concepts and relationships (called properties) can be added as components of meaning to web documents thus making them comprehensible both to computer programs and human users. For example, it is possible to identify the author of the web document or whether the document refers to a geographic location. Ontologies are usually built by experts in the specific fields (Sarasua et al., 2012), who might make use of corpus-based, automated-statistical aids, which can propose words similar in meaning (Lin, 1998; Zhitomirsky-Geffet and Dagan, 2009). Afterward, these words are examined by experts who then arrange them in a precise fashion and build the ontology.

In recent years, ontologies in various fields of knowledge were developed and published on the web by groups, such as universities (protégé ontology library[5]), W3C social initiatives (FOAF[6] and SIOC[7]), public projects [DBPedia – ontology derived from the Wikipedia data,[8] Auer et al. (2007)], and government ministries (SemanticGov[9]). The concepts and relationships of these ontologies are interlinked and thus constitute linked data.[10] The linked data of RDF or OWL triples (statements) across ontologies can be effectively retrieved by semantic search engines (e.g., Virtuoso Universal Server[11] and Sesame[12]) by the means of SPARQL queries.[13] Hence, the Semantic web is the next step in the achieving the organization, management, and retrieval of the enormous amount of information on the internet.

Humanities and the Semantic Web

Nowadays, libraries are one of the main institutions, which produce digital information. This includes bibliographic records, authority files, and concept schemes. This information is currently stored in databases that have, for the most part, a web interface. However, these databases are not deeply integrated with other web sources. In the current situation, library standards, such as MARC or the information retrieval protocol Z39.50, are planned for use by librarians. However, the librarian community and the Semantic web community use different terminology for the same information concepts. To bridge this terminology gap, several libraries have recently taken the initiative to convert their catalogs to RDF-based triples and to linked data. Dunsire (2012) proposes a straightforward methodology for this purpose. For example, the British National Bibliography has been published as linked data by the British Library.[14] The Library of Congress[15] and the Library of Stanford University have also announced that linked data has been included in their roadmap.[16]

In the field of cultural heritage, two fundamental ontologies were recently developed by two large research groups: CIDOC-CRM, a result of a 10-year project[17] (Doerr, 2003) and Europeana Data Model (Winer, 2011, 2014). They are being successfully employed and extended by many national projects for cultural heritage digitization, e.g., the British Museum data collection, which organized the Museum's collection using the CIDOC-CRM for harmonizing with other cultural heritage data,[18] building an ontology for Dante's work (Tavoni et al., 2014), for the Henry III Fine Rolls (Ciula et al., 2008), for Spanish poetry (González-Blanco et al., 2014), and for the Canadian Writing Research Collaboratory (Brown et al., 2015).

However, there are still many fields without ontologies. This is especially the case in Israel and in the Hebrew language, including many fields of Jewish cultural and historical heritage (Judaism, Zionism, Jewish literature, and Jewish folklore). In particular, in this research, we focus on the domain of Jewish

[5] http://protegewiki.stanford.edu/wiki/Protege_Ontology_Library

[6] http://xmlns.com/foaf/0.1/

[7] http://www.w3.org/Submission/sioc-spec/

[8] http://dbpedia.org/About

[9] data.gov/semantic

[10] http://www.w3.org/standards/semanticweb/data

[11] http://virtuoso.openlinksw.com/

[12] http://rdf4j.org/

[13] http://www.w3.org/TR/rdf-sparql-query/

[14] http://www.bl.uk/bibliographic/datafree.html

[15] http://www.loc.gov/bibframe/

[16] http://dataliberate.com//wp-content/uploads/2012/01/LDWTechDraft_ver1.0final_111230.pdf

[17] http://www.cidoc-crm.org/

[18] http://collection.britishmuseum.org

historical manuscripts. The data on these manuscripts exist on the website of the National Library in Jerusalem, the catalogs of other European libraries,[19] and on various Jewish sites, such as Judaica Europeana[20] and the Jewish Agency site.[21] But these websites are not inter-connected, do not use a shared ontology, and the catalogs are not organized as linked data. This lack of website connectivity limits information retrieval and research to specific types of queries and specific isolated manuscript repositories.

MATERIALS AND METHODS

In this section, we first describe the corpus of post-medieval Hebrew manuscripts selected as a case study in our analysis. Further, we present the proposed design of the ontological data model that captures rich complex semantic relationships of this corpus.

Post-Medieval Manuscript Corpus as a Case Study

The estimated number of Hebrew manuscripts that have survived is 70,000–80,000 volumes, about half of which are post-medieval, dating from the seventeenth to the twentieth centuries (Sirat, 2002, p. 8; Beit-Arié, 2014, p. 53; Richler, 2010–2011, p. 14). Although the data model developed in this work is generic and can be used for Hebrew manuscripts from different periods, in the in-depth analysis, we chose to concentrate on the later post-medieval manuscripts. The number of dated post-medieval manuscripts is much larger than one of the medieval manuscripts. In this study, we particularly focus on the semantic dimension of the manuscripts and design an ontology that captures semantic rather than physical features of the manuscripts. But in a later stage of the project, the data model can be enriched to include paleographic and codicological data as well.

This work is the first one which may give us an idea of the content of this large field, which is almost totally ignored by historians of Jewish History. An examination of these manuscripts will reveal evidence of Jewish life in this period with its culture of reading and of learning. Manuscripts containing unknown works and content as yet unpublished will shed light on the society in which they were produced. They include information on the people who produced them, their families, and their surroundings.

Much has been written concerning manuscripts as historical sources, and, just as medieval manuscripts are useful in this case, so might be manuscripts of later periods (Sirat, 2002, pp. 234–256; Pasternak, 2009, pp. 18–21). Such historical information can be provided by colophons. In later periods, information can be gleaned from manuscripts with title pages (Rigler, 1995; Beit-Arié, 2014, pp. 127–130), indications of owners, from family records, incipits, marginal glosses, and even from erased passages.

We chose the manuscript collection of the Séminaire Israélite de France (an institute of higher education dedicated to Jewish and secular learning, also known as the École Rabbinique, that has trained rabbis, cantors, and Hebrew teachers for France and for Jewish communities in French-speaking countries) as our core corpus, since most of the collection (about 90%) consists of late manuscripts from the seventeenth to twentieth centuries. Another reason for our choice of corpus is the fact that this collection covers a range of topics and genres, including almost all religious topics of Hebrew literature, such as commentaries on the Bible, Jewish prayer, Talmud, Halakha (Jewish law), rabbinic literature, Kabbalah and poetry, homiletics, history, and philosophy. This collection contains about 200 works that have been previously investigated in the framework of a project involving the cataloging of Hebrew manuscripts in France, including a new catalog of the Paris Rabbinical Seminary's manuscript collection (Prebor, 2015). We only investigate dated manuscripts.

To further illustrate and justify our choice of corpus and show its importance, we selected manuscripts of three different authors: Moše Lifšiṣ (Ms. Paris – Ecole Rabbinque 48–53), ʻEzra Sayyag (Ms. Paris – Ecole Rabbinque 54), Avraham Śimḥa Katzenelbogen (Ms. Paris – Ecole Rabbinque 86–95), and two maḥzor prayer books (Ms. 24–25). These manuscripts originate from heterogeneous geographical regions (Germany, Syria, and Russia, respectively) and belong to diverse genres and subjects.

From the different fields in manuscript records, we can learn about their historical and social context and about people and events related to them. Thus, we found that Moše Lifšiṣ served as a cantor, scribe, rabbinical judge, and teacher in the Jewish community of Fuerth. His manuscripts were written in the first decades of the seventeenth century and include novella on the Talmud and the Midrašim, commentaries to the Pentateuch, Haftarot, and The Five Scrolls, together with homilies related to the Jewish festivals placed between the weekly Torah portions. These works shed light on the style of learning and thought prevalent in that period, and reveal those commentaries which were available to the author. His son, Shlomo, owned them and planned to print the manuscripts, but he did not print them. The manuscript "Peri Eṣ-Hadar" by ʻEzra Sayyag is probably an autograph. It includes commentaries and homilies on the Pentateuch and The Five Scrolls, and was written in Aleppo about the turn of the nineteenth century. The manuscript was copied in order to be printed. A note in the manuscript sheds further light on the preparation of this work for printing added by Eliyahu Sasson, the proprietor of the printing house in Aleppo (Yaari, 1936, part 1, pp. 33–34), who probably sold the manuscript to the Séminaire Israélite de France. From this note, we learn about the biography of the manuscript. However, this work was also never printed. Despite the prominence of the Jewish community of Aleppo in Syria among Oriental Jews and the productivity of their rabbinic scholars, the religious literature of this community is not well known. This is due to the loss, for various reasons, of many manuscripts and books produced by that community (Harel, 1997, pp. 20–25). Other manuscripts of historic importance are the series of manuscripts by Rabbi Avraham Śimḥa Katzenelbogen who served as rabbi of Yekaterinoslav, then part of the Russian Empire. He lived a difficult life during the course of which, by his own testimony, he wrote more than 60 books. Only 10 manuscripts survived, and, as far as we know, all these

[19] http://web.nli.org.il/
[20] http://www.judaica-europeana.eu/
[21] https://web.archive.org/web/20130306195553/http:/ejewish.info/

exist only in the Seminary's collection. None of these manuscripts have ever been published. The manuscripts deal with a variety of topics: homilies and a collection of Aggadic literature, two works in defense of monarchy, a rabbinic and Kabbalistic encyclopedia, as well as ethical literature. In some of the manuscripts, the author added an appendix in which he recounts his life story, the story of his family, and some historical events that occurred during his lifetime, for example, the Emancipation reform of 1861 promulgated by Tsar Alexander II. Some of his manuscripts are dedicated to famous personalities, such as Sir Moses Montefiore and the Rothschild family.

Currently, most of the above biographic details are incorporated and mixed together in the "notes" field of the catalog and consequently remain non-searchable for users. Thus, the purpose of this study is to make all the above data searchable by direct queries. For example, many important research questions cannot be answered by searching the existing manuscript catalogs, e.g., How many works on specified subjects were composed in this period in different countries? What people were involved in their creation and distribution, and whether and how they are related to each other? What historical events could influence these people and their works? Are these original works or copies of older works from other manuscripts or even from printed books?

This can be achieved by organizing all the above entities (people, places, manuscripts, historical events, and periods) and their semantic relationships into a single systemized data model, an ontology. The proposed ontology will include major key concepts (classes) related to the historical manuscripts and their historical, cultural, and social context. The ontology will help convert and extend the data of catalog records representing the manuscripts into linked data by linking them to similar concepts from the external ontologies and vocabularies on the semantic web (LOD[1]).

Design of the Ontological Data Model for Historical Handwritten Manuscripts

In this section, we present the design of a bilingual Hebrew–English ontology for historical Hebrew manuscripts. One characterizing aspect that distinguishes Hebrew manuscripts from other national manuscripts and cultural heritage objects is a large number of events/changes, i.e., frequent moving over time between different places and people. This is due to the history of the Jewish people, who were spread over the world, repeatedly exiled, and had to move again and again throughout the centuries, and their manuscripts sometimes had moved with them or were passed to others. A manuscript which illustrates this aspect is MS hebr. 7 of the Bibliothèque Nationale (**Figure 1**). The manuscript, which contains the Bible with vocalization, cantillation marks, and masoretic notation, was copied by Solomon ben Raphael at Perpignan in 1299 for his personal use. In 1405, it was sold at Camerino. There it was resold by Uzziel ben Abraham to the physician Menahem ben Yehiel of Ferrara in 1431. Menahem's son Judah sold it to Qalonymos ben Yehiel of Lacavilla in 1451. The manuscript's binding was probably made for Qalonymos' son Shabbetai. Shabbetai's widow Ferna had the manuscript sold by her agent Joseph of Ascoli to Isaac ben Abraham Finzi in 1510. Nothing is known of the manuscript's owners until the year 1620 when it was presented to the Oratory in Paris by Achille de Harlay de Sancy, the former French ambassador to the Ottoman Empire. In the wake of the nationalization of clerical property, it became part of the collection of the Bibliothèque Nationale (Sirat, 2002, p. 253). In 2010, the manuscript was digitized and is part of Gallica, the digital library of the Bibliothèque Nationale de France, and its partners.[22]

The existing ontological schemes for historical manuscript representation [such as the Europeana dm2e model and Vieira and Ciula (2007)] take an object/property-based approach, where a manuscript object can be directly linked through its properties to multiple agents, places, events, and periods by different individual/separated properties. As a result, for example, a list of geographical places related to the manuscript can be extracted from the ontology, but the information on the time periods or agents (e.g., people) related to each of these places is not explicitly encoded.

In other words, the scheme does not capture when each of the places in the list was involved in the manuscript's biography or what people were involved and what actions were performed with the manuscript at each of these places (as illustrated in **Figure 1**). To overcome this problem of missing essential semantic links in data representation and to provide a more complete semantic representation of the manuscript biography, we propose an *event*-based ontological model. This model includes events in a manuscript life story (e.g., creation, printing, acquisition, copying, storing, citing, censoring, dedication, and more). Each event creates a complete semantic framework, which encapsulates all the related to the event data and interlinks all the other related objects, such as a manuscript (involved object), a person (involved agent), a time period, and a place (as shown in **Figure 2**).

The event-based approach was also utilized in modeling the provenance of digital objects in the CRMdig ontology (Theodoridou et al., 2010[23,24]), an extension of the CIDOC-CRM model, and for numerous heritage applications. For example, Mulholland et al. (2014) employ the event-based approach to assist the authors and readers of museum stories to better understand and explore the surrounding context. Many relationships can exist between the concepts (people, places, and museum objects) mentioned in museum stories. Unmentioned national or international events may have influenced what happened in the story. Hyvönen et al. (2014) develop an event-based approach to publishing life stories (biographies) as Linked Data, because data in biographies are fundamentally based on life events and their sequences. This approach also allows for data enriching. For example, metadata about a painting by an artist tells that there should be the corresponding painting event in his/her biography that may be missing.

The ontology also includes major key concepts (classes) related to the manuscripts and their historical, cultural, and social context. The ontological data model is constructed in the spirit of the "Livre et Société" approach (Baruchson-Arbib, 1993; Elyada, 1999), which

[22] http://gallica.bnf.fr/ark:/12148/btv1b9002997b/f1.item

[23] http://www.ics.forth.gr/isl/CRMext/CRMdig/docs/CRMdig3.2.pdf

[24] https://www.usenix.org/legacy/event/tapp11/tech/final_files/Doerr.pdf

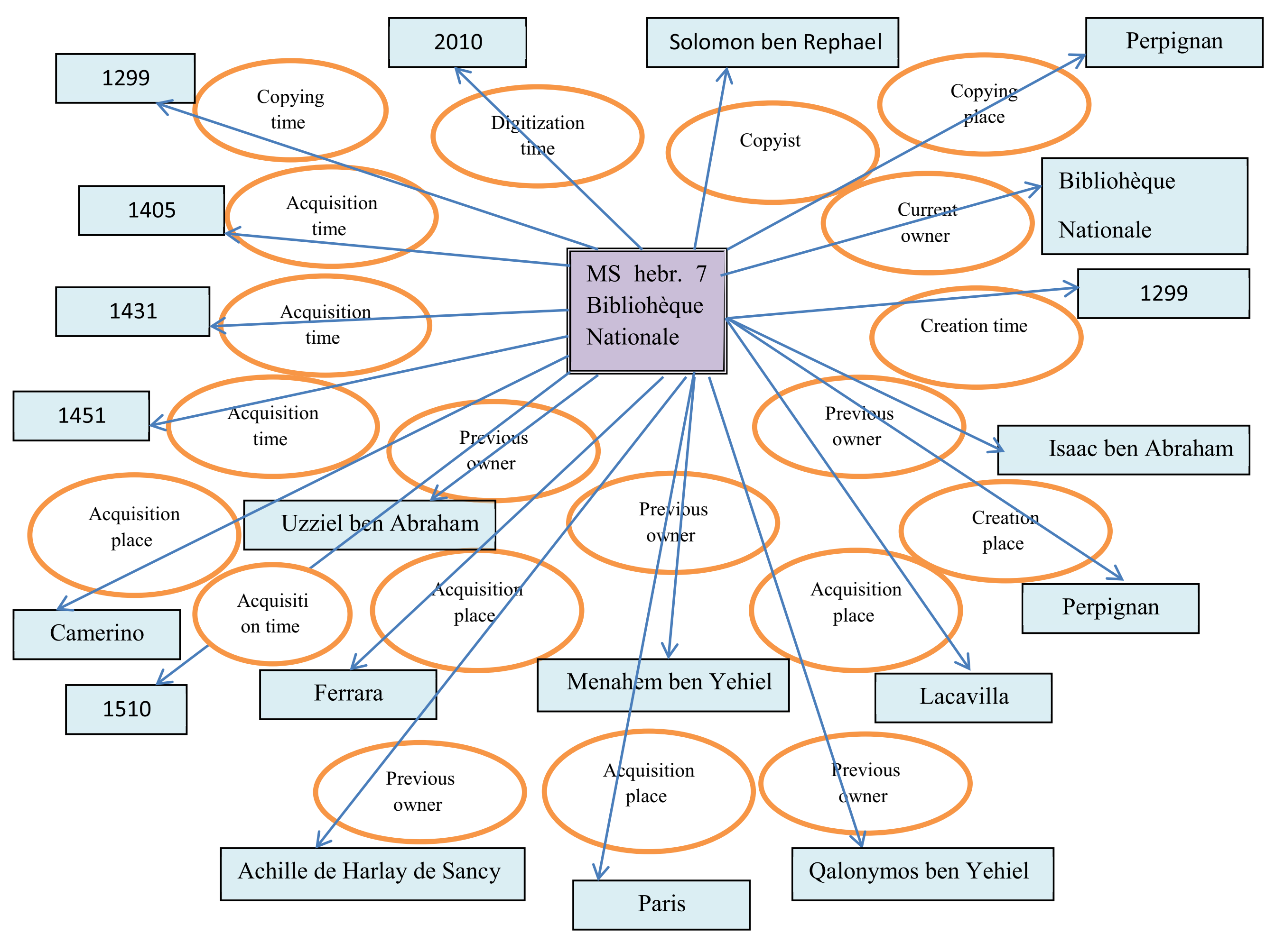


FIGURE 1 | **The object/property-based data model for the manuscript hebr. 7 of the Bibliothèque Nationale**.

views a book as an important source for understanding historical processes and social and cultural changes. Thus, different parts of manuscripts, such as colophons, incipits, title pages, and family records, can serve as historical sources (Sirat, 2002, pp. 234–256; Pasternak, 2009, pp. 18–21; Beit-Arié, 2014, pp. 127–130).

We assume that many of the classes and properties in our model might use classes and properties in existing ontologies, which reflect some related aspects. To test this assumption, we analyzed several relevant ontologies to learn whether and how they can be reused, linked as matching by Simple Knowledge Organization System (SKOS) relations (Miles et al., 2005), inherited, modified, or adapted in our model. These ontologies can be divided into the following domains:

1. General-purpose ontologies/data models related to cultural heritage, documents, and manuscripts, e.g., EDM (DM2E),[25] CIDOC-CRM (Doerr, 2003), and SKOS (Miles et al., 2005).
2. Ontologies related to bibliographic data, e.g., BIBO (D'Arcus and Giasson, 2009), CITO (Peroni and Shotton, 2012), FRBR (Tillett, 2005), FRBRoo,[26] and Dublin core (ISO 15836:2009[27]).
3. Ontologies related to people and organizations, e.g., FOAF,[28] BIO,[29] EAC-CFM,[30] VIVO.[31]
4. Geographic ontologies, e.g., the Geonames ontology and W3CBasic Geo Vocabulary (GWS84).[32]
5. Event/time models, e.g., BIO, CIDOC-CRM, W3C Time[33] (Hobbs and Pan, 2005), LODE (Shaw et al., 2009), and SEM (Hage et al., 2011).

[25] http://dm2e.eu/files/DM2E_Model_V1.2.pdf

[26] http://www.cidoc-crm.org/frbr_inro.html

[27] http://dublincore.org/

[28] http://xmlns.com/foaf/0.1/

[29] http://purl.org/vocab/bio/0.1/

[30] http://archivi.ibc.regione.emilia-romagna.it/ontology/reference_document/referencedocument.html

[31] https://wiki.duraspace.org/display/VIVO/VIVO-ISF+ontology+documentation

[32] http://www.w3.org/2003/01/geo/#vocabulary

[33] http://www.w3.org/TR/owl-time/

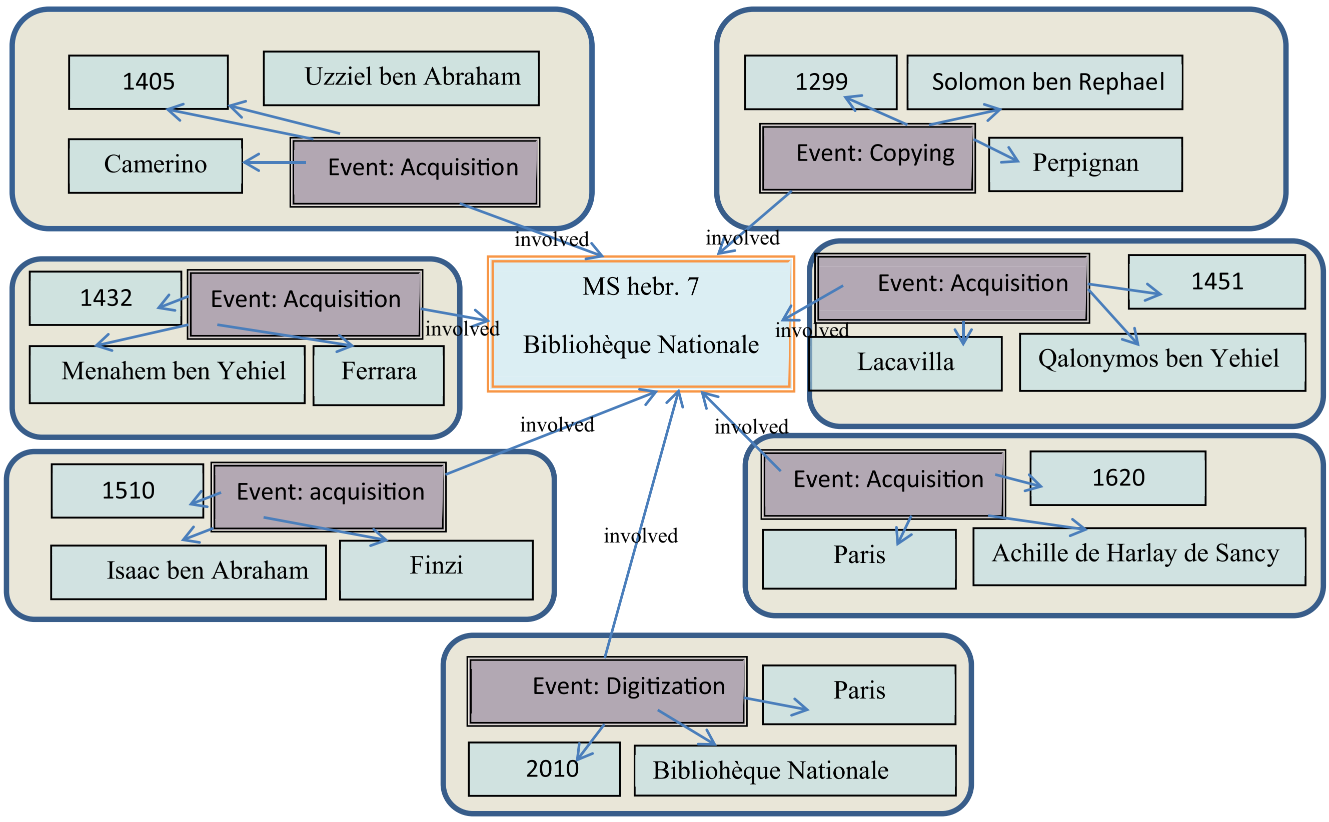


FIGURE 2 | **The event-based data model for the manuscript hebr. 7 of the Bibliothèque Nationale**.

Therefore, first, we conducted a comparative analysis of the numerous existing ontologies that are relevant to our ontological model for historical Hebrew manuscripts. In particular, in **Table 1**, we compared the classes corresponding to the main entities in the manuscript realm: "Manuscript," "Agent," and "Event," in DM2E (namespace denoted as "ens:"), FOAF, BIO, BIBO, SEM, LODE, CIDOC-CRM, and FRBRoo semantic models. The table presents the definitions of the classes related to each of the three classes specified above.

Based on the analysis in **Table 1**, we defined an upper generic ontological model for historical manuscripts as shown in **Figure 3**.

As indicated, this is an event-based model with the following new central classes: manuscript biographic event, Historical manuscript, Manuscript agent, and Historical figure. Next, we describe these (and other related) classes in more detail and show their relationships with classes from the existing ontologies as presented in **Table 1** above. We also discuss the reasons for our decisions on modeling these classes.

"Manuscript biographic event" is a subclass of "lode:Event" and "crm:E7_Activity" and inherits all their properties. This is the central class of the ontology, which makes the connections between all the main entities, such as manuscripts, people, time, and place. The main properties of "Manuscript biographic event" comprise "lode:involved agent" (links to persons and organizations), "lode:involved (object)" (links to a manuscript), "lode:atPlace" (with values of "E53:Place"), "crm:occurs before/after" (to connect between events on the timeline), "crm:was influenced by," and "crm:was intended use of." Many times the exact dates are unknown; hence, there is a need in definition for time span or interval expressed by the property "lode:atTime" (with values of "crm:E52_Time_Span"). The subclasses of "Manuscript biographic event" are classes corresponding to specific manuscript life events, such as "Manuscript Creation" (also a subclass of "E65_Creation" with a "brought into existence" property), "Acquisition" (also a subclass of "E8_Acquisition" with "transferred title to/from/of" properties), "Copying," "Printing," "Censoring," "Digitization," and "Dedication." Note that CIDOC-CRM "E5_Event" and SEM "Event" classes fit agent-oriented event setting, but they do not include a property for the involved object, which is required in the case of manuscript-oriented events. Therefore, we use a more specific "crm:E7_Activity" class as a superclass for the new defined class. "lode:Event" is a subclass of crm:E2_Temporal Entity. It collects and simplifies the main CIDOC properties related to events and unlike CIDOC distinguishes between the factual information and the interpretation of the events. This distinction is important in the realm of historical manuscripts for data consistency reasons, where diverse and even contradictory opinions might exist with regard to dates, places, people, and influences. Thus, an instance (object) of an

TABLE 1 | Analysis of the exclusive properties of classes related to Manuscript, Agent, and Event classes in different existing ontologies.

Main property names	Appear in ontology	Appear in class	Class meaning definition and comments
Manuscript-related classes analysis			
has component, refers to, is about (is subject of)	CIDOC-CRM	E73_Information Object	This class comprises identifiable immaterial items, such as texts and multimedia objects, that have an objectively recognizable structure and are documented as single units
created by, has former keeper, has current keeper, present at	FRBRoo	F4_Manifestation Singleton	This class is suitable for describing a cultural heritage object and its physical aspects
reviewOf, editor list and author list	BIBO	Manuscript	An unpublished document, which may also be submitted to a publisher for publication, suitable for description of modern scientific documents and manuscripts. This class is a subclass of Document class and inherits its properties
topic and primary topic	FOAF	Document	foaf:Document is equivalent to the class Document in BIBO ontology
Has-former/current-owner, consists-of, has-section, has-condition, has-former/current-location	DM2E	Manuscript	Like FRBRoo Manifestation Singleton this describes physical aspects of manuscripts. Manuscript class in DM2E is a subclass of Physical Thing
incipit, explicit, origin, modeOfAcquisition, mentioned, refersTo, isRelatedTo, wasTaughtBy, honoree, patron, owner, previousOwner, copyist, has-met, nextInSequence (relates manuscript parts)	DM2E	Provided cultural heritage object (CHO)	This class comprises the Cultural Heritage objects that Europeana collects descriptions about In DM2E model, the class Provided CHO can be of type ens:Manuscript
Agent-related classes analysis			
knows, based_near	FOAF	Person	Most of the existing ontologies use this class
Publications	FOAF	Agent	Class of things that do stuff (not necessarily people)
Has-current/former residence, participated in event, possesses	CIDOC-CRM	E39_Actor	This class comprises people, either individually or in groups, who have the potential to perform intentional actions for which they can be held responsible
Sub-properties of foaf:knows: spouseOf, parentOf, collaboratesWith, closeFriendOf	BIO, EAC-CPF	Person	BIO ontology reuses the classes Agent and Person of FOAF as well
took_part, witnessed, principal (main figure), lived_when	BIO	Agent	A person, organization, or group that plays a role in an event
has-met, studentOf, influencedBy	DM2E (EDM)	Agent	DM2E is equivalent to CIDOC-CRM E39_Actor class, Person from FOAF is linked as a subclass to ens:Agent and extended with additional properties
Event-related classes analysis			
had_participant (participated_in), occurred_in_the_presence_of (was_present_at); has time-span, starts, finishes, occurs during, occurs before (from E2_Temporal Entity); took place at, falls within (from E4_Period)	CIDOC-CRM	E5_Event	This class comprises actions intentionally carried out by instances of E39_Actor that result in changes of state in the cultural, social, or physical systems documented. This notion includes complex, composite, and long-lasting action, such as the building of a settlement or a war, as well as simple, short-lived actions, such as the opening of a door. This is a subclass of E4_Period, which is a subclass of E2_Temporal Entity
carried out by (actor in the role of), was influenced by, used specific object, was motivated by, was intended use of, had specific purpose, and used specific technique	CIDOC-CRM	E7_Activity	This class comprises actions intentionally carried out by instances of E39 Actor that result in changes of state in the cultural, social, or physical systems documented. E7_Activity is a subclass of E5_Event
Place, Date, Agent, Position, Event Interval, Preceding Event, Following Event, Employer, State, Parent, Principal, Partner, Witness, Spectator, and Organization	BIO	Event	An event is an occurrence that brings about a change in the state of affairs for one or more people and/or other agents. Events are assumed to occur over a period of time and may not have precise start and end points This class is intended to describe biographical events, i.e., events in the life of a person
atPlace, atTime (E52_time span, at most one for an event), involvedAgent, inSpace (e.g., a geospatial point or region, at most one for an event), involved (object)	LODE	Event	"Something that happened," as might be reported in a news article or explained by an historian. LODE event model only allows for expressing characteristics about factual events for which a stable consensus has been reached. This is a subclass of CIDOC's E2_Temporal Entity class, which generalizes and simplifies CIDOC's properties related to events
hasPlace, hasTime, has Actor, constraints on the properties, e.g., Role (the specific role of the agent in the event), and View (which allows for co-existence of conflicting statements supported by diverse authorities)	SEM	Event	Events are things that happen. This comprises everything from historical events to web site sessions and mythical journeys. SEM model includes a few properties for encoding uncertain time intervals

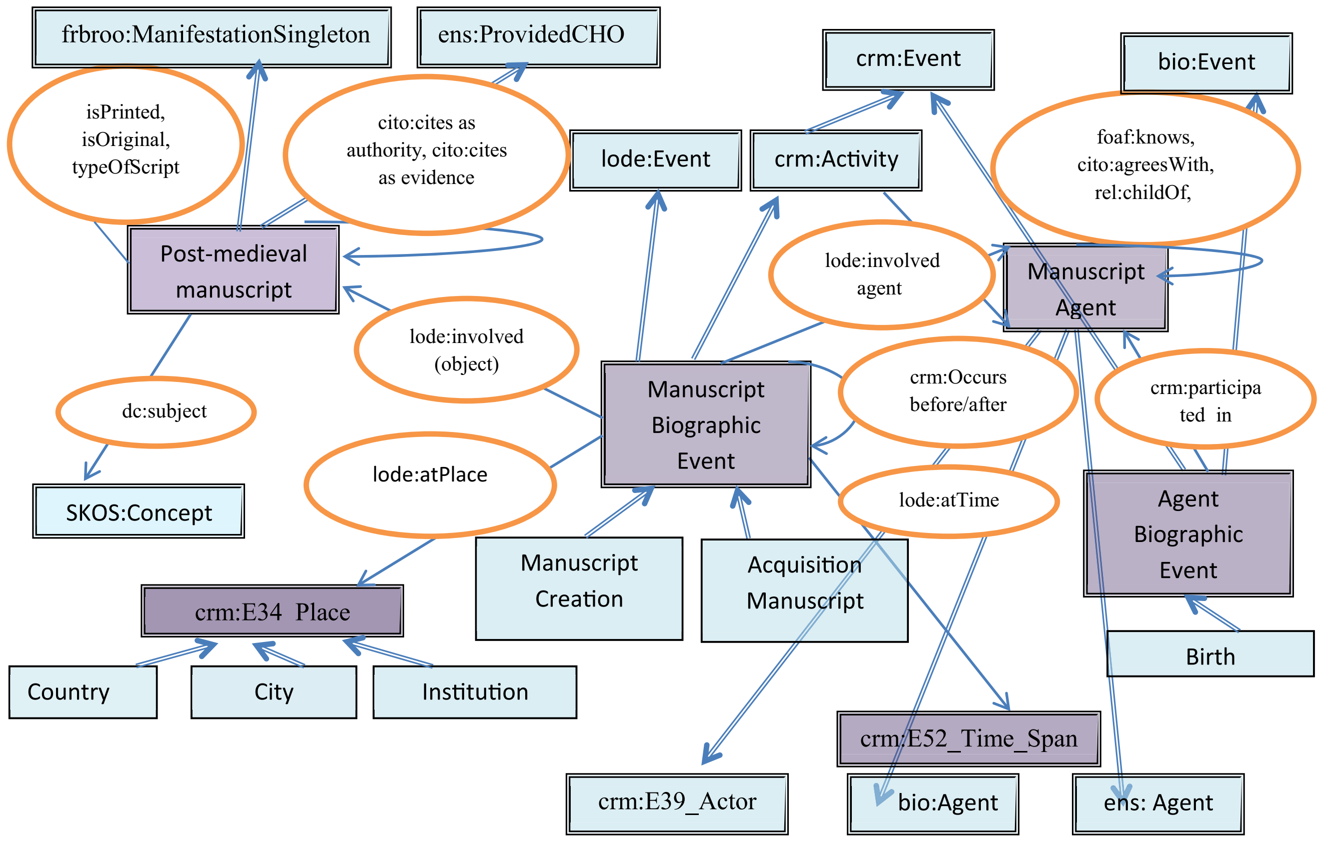


FIGURE 3 | The proposed conceptual model for the upper level historical manuscript ontology. The square boxes represent the top-level classes and their subclasses, the properties are displayed by single-line arrows, while the names of properties appear near the corresponding arrows. Child–parent arrows represent rdfs:subClassOf relationships.

event represents a single interpretation of its properties. In order to support the diversity of interpretations, we create a different event instance for every interpretation. We further adopt the CRMinf ontology (Doerr et al., 2011) for belief and uncertainty modeling. To this end, every event instance can be viewed as an "I2_Proposition set," which can be assigned an "I6_Belief Value" of a particular "E39_Actor." To describe historical events, we use "crm:E5_Event" class, which is a subclass of "crm:E4_Period." "Manuscript biographic event" is linked with historical events by properties: "crm:falls within" and "crm:occurs before/during/after." Finally, to describe people's biographic events, we define a new class "Agent biographic event" as a subclass of "crm:Event" and "bio:Event," and its subclasses (such as "Birth," "Marriage," and "Death") are also subclasses of the corresponding specific classes in BIO ontology (bio:Birth, bio:Marriage, and bio:Death).

"Historical manuscript" is a subclass of frbroo: ManifestationSingleton and ens:ProvidedCHO and inherits all their properties. It has two subclasses "Medieval manuscript" and "Post-medieval Manuscript." The "Medieval manuscript" class contains more information on codicological features of the manuscripts than the "Post-medieval manuscript" class, which concentrates mostly on the semantic characteristics and relationships. As indicated in **Table 1**, the "ens:ProvidedCHO" class also includes all the relevant properties from CIDOC-CRM "E73_Information object" and "foaf:Document." However, some of the properties are not sufficiently specific and semantically rich for our purposes. For example, the properties "crm:refers to" (cites a person or another manuscript) and "ens:mentioned" in our model have been refined with properties from the CITO ontology, such as "cites_as_evidence," "is_cited_as_authority_by," "is_criticized_by," and "is_cited_by", to capture the sentiment of the citation, which plays an important role in Jewish literature. Moreover, properties directly linking the manuscript to its current and previous owners and other agents are redundant in our case, since we take the event-centered approach as explained in the previous paragraph, which provides the complete information on the event rather than just the details of its agent. In addition, new datatype properties should be added, such as "wasPrinted" (was the manuscript ever printed?), "isOriginal" (is the manuscript an original work or a copy of another manuscript or printed work?), and "typeOfScript" [there are three types of writing each of them could be written in different formats according to geocultural areas (Sirat, 2002, pp. 182–203; Beit-Arié, 2014, pp. 389–449)] to enable easy retrieval by these parameters.

Two new classes for describing people are defined: "Manuscript agent" and "Historical figure." "Manuscript agent" is a class to describe people involved in the manuscript life cycle. It is defined as a subclass of "crm:E39_Actor," "ens:Agent," and "bio:Agent." In particular, it inherits "crm:has residence property," the properties for personal and professional relationships with other people, e.g., "foaf:knows" and its sub-properties from the Relationship vocabulary[34] (e.g., "rel:studentOf," "rel:parentOf," and "rel:collaboratesWith"), and from the Europeana Data Model (e.g., "ens:influencedBy"). It also uses CITO ontology citation act properties (e.g., "cito:agreesWith" and "cito:critiques"). In addition, a person is linked with various types of events, such as manuscript-related events, biographic, and historical events. We also define a class "Historical figure" for people mentioned in the manuscripts as a subclass of "crm:E21_Person" and "bio:Person." Among others, this class inherits the "bio:lived_when" property, which is used to bind an historical figure with the corresponding historical events. In our event-based model, "Manuscript agent" is only linked to the "Historical manuscript" class indirectly *via* the "Manuscript biographic event" class, which holds the place, time, and other data on the manuscript life event. This structure provides our model with the desired flexibility to enable insertion of new entities as more manuscript data are analyzed and processed. Also, this event-based approach provides a more precise and complete encoding in the setting, where several owners (or other agents) exist for a given manuscript at different times and places than using frbroo:former_keeper, ens:has-former-owner, or dcterms:writtenAt properties. However, "Historical figure" is connected with the "Historical manuscript" class directly by the "ens:mentioned_in" property and its more specific properties, such as ens:honoree (when the historical figure is mentioned as a honoree of the manuscript). Names of people can change over time; thus, we need to offer a property for "historical name" similar to Geonames's "alternate name" property.

"crm:E34_Place" was used to determine the geographical entities. Among others, this class includes the "crm:is_identified_by" property, the values of which are geographic coordinates of the place required to enable GIS-based representation of the data on e-maps as in GeoNames and VIAF online search systems. Names of places can change during time; thus, we need to offer a property for "historical name" similar to Geonames's "alternate name" property.

"SKOS:Concept" class is used for describing different topics in the content of the manuscript, which are defined as instances of this class, such as books of the Bible, Talmud, Halacha (Jewish law), Kabbalah, Commentaries, and more.

As can be noticed, the property sets in some of the existing ontological classes are overlapping, such as "crm:participated in" and "bio:took part" (in event). Hence, inherited properties with identical semantics will be denoted by the owl:sameAs construct. The ontological model will be iteratively updated and extended until the model eventually converges. The namespace for the proposed ontology is http://www.ontology.org.il/HebrewManuscripts#, for both the ontology definition and the individuals' instances. The namespace prefix used is "hmo." The model is implemented in RDF/S ontology language and can be queried using SPARQL endpoints.

This model will enable an in-depth research of the cultural heritage reflected by the manuscripts at different levels and perspectives. Distant currently unrevealed relations between objects will be discovered through the ontology. Moreover, the ontology will directly interlink semantically related objects, such as members of the same family, parts of the same manuscript, and cities of the same country. This is another type of new information that will be contributed by the structure of the ontology, which is missing in the catalog. The proposed model makes it possible to follow the history of individual manuscripts and answer complex queries, e.g., to extract all the events of censoring in Italy in the seventeenth century or to extract all the people involved in acquisition events of any manuscript in Russia in the nineteenth century. Such queries are useful in error identification and correction, for instance, to reveal "holes" in the biography of manuscripts, i.e., periods when there is no available information on the manuscript or overlapping periods, contradictory information, missing parts, and events that were intended but never executed.

The proposed ontology utilizes existing ontologies for the domain of cultural heritage, the manuscripts' catalog records (in particular, from the catalog of the IMHM in the National Library of Israel) and also knowledge accumulated by state-of-the-art research on the historical Hebrew manuscripts.

Once the basic upper level ontology has been constructed, we will analyze the catalog to extract the specific data/instances to populate the lower level ontology and to further verify and extend the upper level ontological model.

Decomposition Analysis of the Catalog Records

The objective of this analysis was to determine a methodology for converting information encoded in catalog records to the ontological data. To this end, the work on the core corpus concentrated on the identification of different types of fields in the catalog records and their decomposition into atomic data units. These atomic units (entities) were further examined in order to incorporate them in the proposed ontology. As a result, we determined the following types of fields:

(1) Simple consistent fields comprising only one atomic entity that can be directly converted into ontological properties; their values were defined as classes in the ontology;
(2) Complex fields that were divided into two or more distinct properties in the ontology;
(3) Inconsistent and general meaning fields in which content was represented by several different ontological properties in different classes;
(4) Missing fields extracted by means of traditional research that were added to the ontology (including the unique ones for Hebrew manuscripts).

RESULTS AND DISCUSSION

The goal of the decomposition analysis was to check the feasibility and validity of the proposed ontological model and methodology for catalog records analysis on real data. As mentioned earlier, we

[34] http://vocab.org/bio/

selected representative manuscripts from the manuscript collection of the Séminaire Israélite de France as a case study for the analysis. Next, we present an analysis of a representative sample of the records with varying complexity for the three geographically and topically diverse manuscript authors (presented in the previous section): Ezra Sayyag from Syria, Moshe Livshitz from Germany, and Simcha Avraham Katsnelbogen from Russia. The catalog of the Department of Manuscripts and the IMHM of the National Library of Israel is built according to the AACR2 cataloging rules and encoded in MARC format. An example of a catalog record is displayed in **Figure 4**.

Next, we aim to encode all the above information in the form of ontological statements (RDF-style triples) of the form: *class 1–property–class 2*. All the names (of persons, places, manuscripts, events, and other objects) were submitted as queries to a semantic search engine, Virtuoso server,[35] to look them up in external ontologies, such as Freebase (Bollacker et al., 2008), DBPedia (Auer et al., 2007), and Opencyc (Foxvog, 2010; Lenat et al., 2010[36]), and in the authority files, such as VIAF,[37] LCNS,[38] GETTY,[39] GND,[40] and GeoNames.[41] Once found (as shown in the last column of **Table 2**: links to authority files), they were linked directly by using the extracted URI or by owl:sameAs relation for similar terms in different vocabularies.

Table 2 demonstrates the results of the decomposition analysis of the different types of fields in the selected manuscript records. **Table 2** shows the type of the field and to which ontological classes and properties its data might be converted, according to the basic model in **Figure 1**. Most of the catalog entities were

[35] http://virtuoso.openlinksw.com/

[36] http://www.cyc.com/platform/opencyc

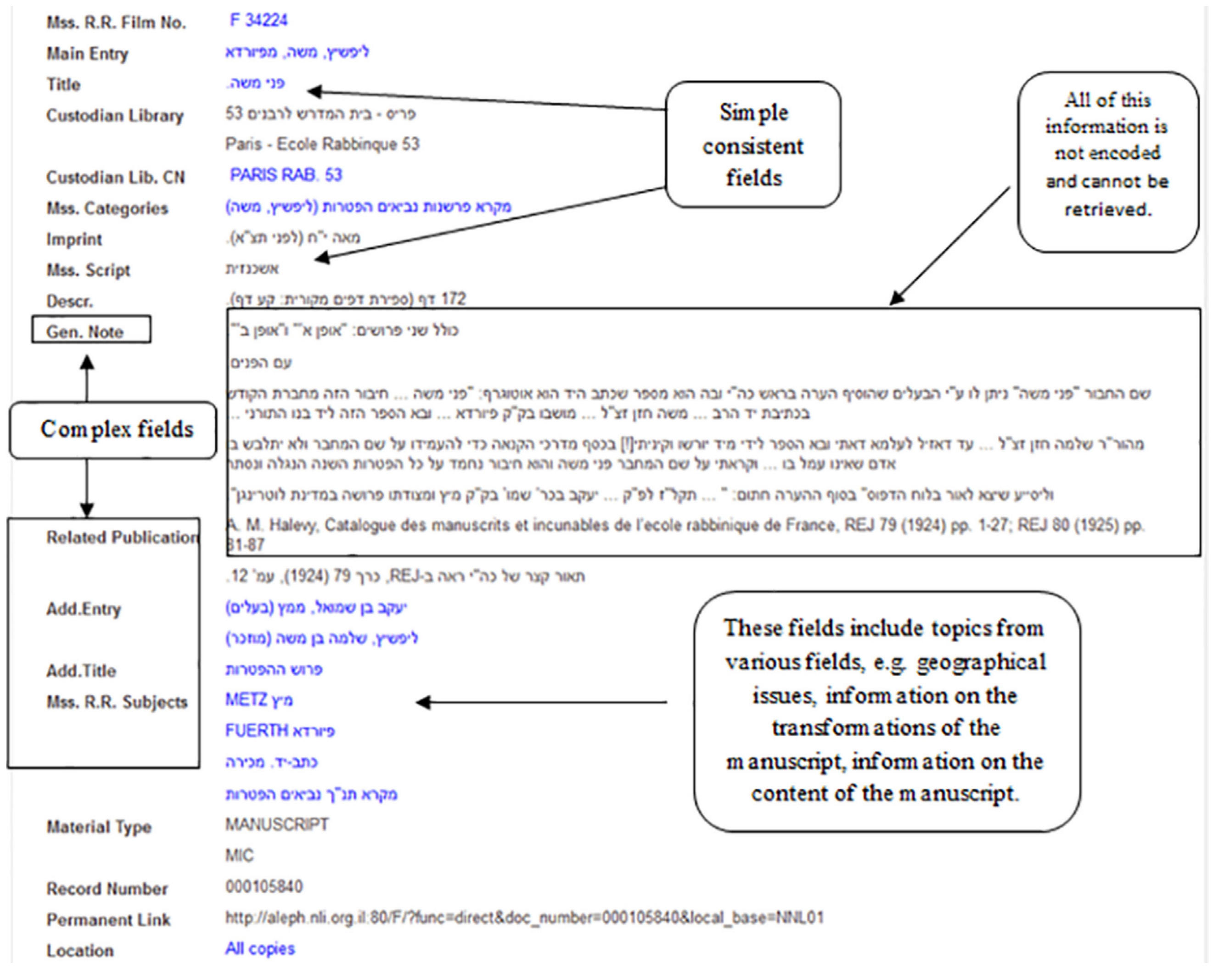
[37] http://viaf.org/

[38] http://authorities.loc.gov/

[39] http://www.getty.edu/research/tools/vocabularies/index.html

[40] http://en.wikipedia.org/wiki/Integrated_Authority_File

[41] http://www.geonames.org/ontology



Mss. R.R. Film No.	F 34224
Main Entry	ליפשיץ, משה, מפיורדא
Title	פני משה.
Custodian Library	פריס - בית המדרש לרבנים 53 Paris - Ecole Rabbinque 53
Custodian Lib. CN	PARIS RAB. 53
Mss. Categories	מקרא פרשנות נביאים הפטרות (ליפשיץ, משה)
Imprint	מאה י"ח (לפני תצ"א).
Mss. Script	אשכנזית
Descr.	172 דף (ספירת דפים מקורית: קע דף).
Gen. Note	כולל שני פרושים: "אופן א'" ו"אופן ב'" עם הפנים שם התבור "פני משה" ניתן לו ע"י הבעלים שהוסיף הערה בראש כה"י ובה הוא מספר שכתב היד הוא אוטוגרף: "פני משה ... חיבור הזה מתברת הקודש בכתיבת יד הרב ... משה חזן זצ"ל ... מושבו בק"ק פיורדא ... ובא הספר הזה ליד בנו התורני .. מהור"ר שלמה חזן זצ"ל ... עד דאזיל לעלמא דאתי ובא הספר לידי מיד יורשו וקניתי[!] בכסף מדרכי הקנאה כדי להעמידו על שם המחבר ולא יתלבש ב אדם שאינו עמל בו ... וקראתי על שם המחבר פני משה והוא חיבור נחמד על כל הפטרות השנה הנגלה ונסת וליס"ע שיצא לאור בלוח הדפוס" בסוף ההערה חתום: " ... תקל"ז לפ"ק ... יעקב בכר' שמו' בק"ק מ"ץ ומצודתו פרושה במדינת לוטרינגן"
Related Publication	A. M. Halevy, Catalogue des manuscrits et incunables de l'ecole rabbinique de France, REJ 79 (1924) pp. 1-27; REJ 80 (1925) pp. 81-87 תאור קצר של כה"י ראה ב-REJ, כרך 79 (1924), עמ' 12.
Add.Entry	יעקב בן שמואל, ממץ (בעלים) ליפשיץ, שלמה בן משה (מוזכר)
Add.Title	פרוש ההפטרות
Mss. R.R. Subjects	METZ מ"ץ FUERTH פיורדא כתב-יד. מכירה מקרא תנ"ך נביאים הפטרות
Material Type	MANUSCRIPT MIC
Record Number	000105840
Permanent Link	http://aleph.nli.org.il:80/F/?func=direct&doc_number=000105840&local_base=NNL01
Location	All copies

Simple consistent fields

All of this information is not encoded and cannot be retrieved.

Complex fields

These fields include topics from various fields, e.g. geographical issues, information on the transformations of the manuscript, information on the content of the manuscript.

FIGURE 4 | **An example of a National Library catalog record for a manuscript "Pnei Moše" by Moše Lifšiş.**

TABLE 2 | This table includes representative examples for analysis of different types of fields from the selected set of 20 manuscripts.

	The original record value	Type of field	Corresponding ontological triples of the form: class/ instance1–"property"–class/instance2	Links to authority files
Subject field tag 966, record no. 183145	Jewish preaching	Simple consistent field	Historical Manuscript:183145 – "dc:subject" – skos:Concept:Jewish preaching	Jewish preaching from LCSH: http://id.loc.gov/authorities/subjects/sh85106181
Author field tag 100, record no. 105840	Livshitz, Moshe from Fuerth	Complex field	1. ManuscriptAgent:Livshitz, Moshe – "lode:involved agent" – (Manuscript) Creation event 2. Historical Manuscript:105840 – "lode:involved" – Manuscript Creation event. 3. Manuscript Agent:Livshitz, Moshe – " lode:involved agent " – Birth 4. Birth – "lode:atPlace" – crm:Place:Fuerth	Fuerth from LCSH: http://id.loc.gov/authorities/names/n85127769.html Geonames: http://www.geonames.org/maps/google_49.476_10.989.html
Subject field tag 600, record no. 183154	Montefiore Moses, son of Joseph Eliahu, 1784–1885	Complex field. The person is mentioned in the manuscript as honoree.	1. Historical figure:Montefiore Moses – "ens:honoree" – Historical Manuscript:183154 2. Historical figure:Montefiore Moses – "rel:childOf" – Historical figure: Joseph Eliahu 3. Historical figure:Montefiore Moses – " lode:involved agent " – Birth 4. Historical figure:Montefiore Moses – " lode:involved agent " – Death 5. Birth – "lode:atTime" – crm:Time Span:1784 6. Death – "lode:atTime" – crm:Time Span:1885	Montefiore Moses, from LCNA: http://id.loc.gov/authorities/names/n83319120 VIAF: http://viaf.org/viaf/13112023
Subject field tag 967, record no. 105843	Fuerth Talmud commentaries (Livshitz, Moshe), print, Metz, homily	Complex inconsistent field. The author and location mappings are already shown above. The geographic location is doubled by field "Place of production"	1. Historical manuscript:105843 – "dc:subject" – skos:Concept:homily 2. Historical manuscript:105843 – "dc:subject" – skos:Concept:Talmud 3. Historical manuscript:105843 – "lode:involved_in" – Acquisition event 4. Acquisition event – "lode:atPlace" – crm:Place:Metz 5. Historical manuscript:105843 – "crm:intended for" – Printing event. [The information on printing is imprecise, the manuscript was prepared (crm:intended) for printing but was never actually printed.]	Metz, from Getty: http://www.getty.edu/vow/TGNFullDisplay?find=Metz&place=&nation=&prev_page=1&English=Y&subjectid=7008418 Geonames: http://www.geonames.org/maps/google_49.119_6.173.html homily from LCSH: http://id.loc.gov/authorities/subjects/sh85120261.html
Add entry field tag 700, record no. 105840	Jacob ben Shmuel from Metz, owner Livshitz, Shlomo, mentioned	Complex inconsistent field	1. Manuscript Agent:Jacob ben Shmuel – " lode:involved agent" – (Manuscript) Ownership event 2. Historical Manuscript: 105840- "lode:involved_in" – (Manuscript) Ownership event 3. Historical figure:Jacob ben Shmuel – "lode:involved agent" – Birth 4. Birth – "lode:atPlace" – crm:Place:Metz 5. Historical figure:Livshitz, Shlomo – "ens:mentioned_in" – Historical Manuscript: 105840	Jacob ben Shmuel – N/A Livshitz Shlomo from VIAF: http://viaf.org/viaf/313003475
General note field tag 500, record no. 183104	The rest of the manuscript including the Sukkot festival Mahzor is located in Amsterdam	Complex inconsistent field. It contains partial information on the other part of the manuscript	1. Historical Manuscript: 183104 – "ens:isNextInSequence" 2. Historical Manuscript: missing no. (of the 2nd part of the manuscript) 3. Historical Manuscript: 183104 – "dc:subject" – skos:Concept:Sukkot festival prayer (Mahzor) 4. Historical Manuscript: missing no. – "lode:involved_in" – Storage event 5. Storage event – "lode:atPlace" – gn:city:Amsterdam	Amsterdam from Getty: http://www.getty.edu/vow/TGNFullDisplay?find=Amsterdam&place=&nation=&prev_page=1&english=Y&subjectid=7006952 Geonames: http://www.geonames.org/2759794/amsterdam.html

*We show a possible mapping to ontological triples using the upper ontology model from **Figure 1** and linking to similar entities in the existing semantic web ontologies.*

found in the existing authority files as shown by the last column of **Table 2**.

As indicated, only a minority of the examined fields contained atomic and consistent information, such as the field 966 (subject). The majority of the fields were complex, i.e., a single field that contains a few different types of ontological data. For instance, the author field (tag 100) typically comprised the name of the manuscript author and his place of birth. Many of these complex fields were also inconsistent and included different types of ontological entities in different records. For example, the subject field (tag 967) which might contain information on the author's name, place of manuscript acquisition, subjects of the manuscript, and its printing event. The definition of some other fields is not specific enough and is inconsistent, such as the general note (500) and the add entry (700) fields, which contain additional information on persons, places, and events in involved with the manuscripts. It can also be noticed that some data are duplicated in the different fields of the catalog.

Consequently, the aforementioned limits and flaws in the structure of the catalog records make it hard or even impossible to search for a large part of the data contained in them, which are important for researchers and learners. In addition, the inter-relationships of the entities listed above are not encoded in the catalog (e.g., relationships between different people, places, manuscripts, and events). These limitations are resolved by the proposed ontology. In the ontology, all these complex fields were split into several ontological properties.

Based on the aforementioned decomposition analysis and the types of fields that are revealed, semi-automatic techniques of information extraction can be further employed to populate the ontology with specific instances [for such techniques, see Mohit (2014)].

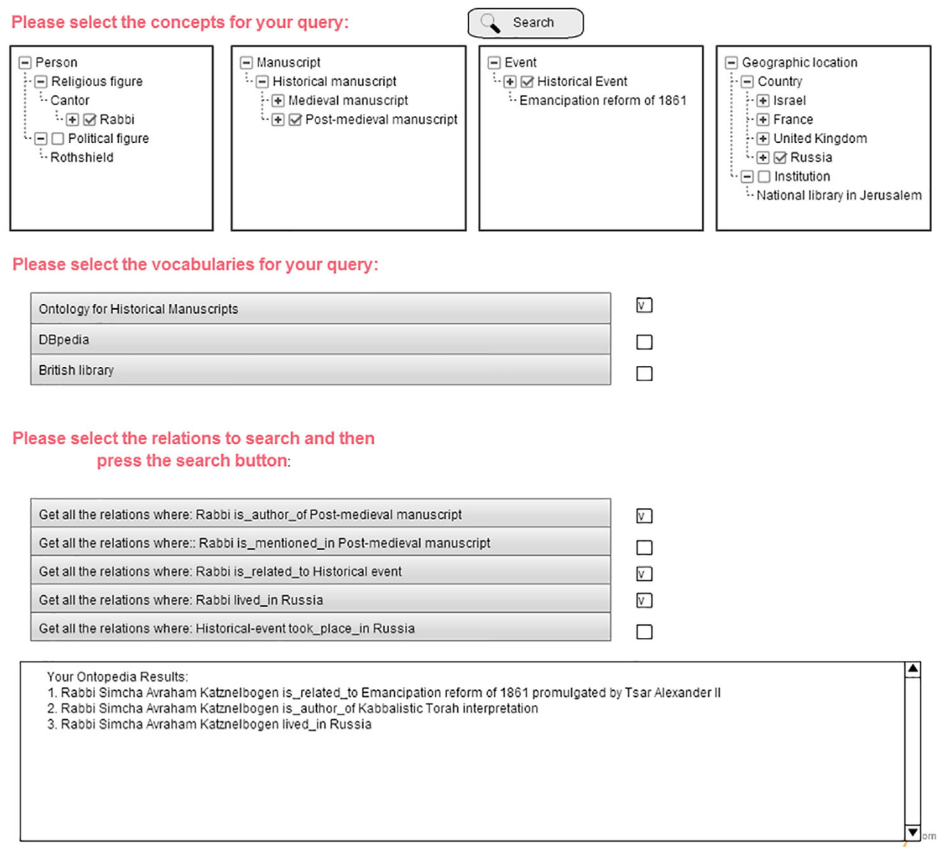
Searching the Catalog vs. Querying the Ontology

Due to the flaws listed above in catalog encoding (complex fields styled as free text with many entities of different types mixed together), many of the entities and their relationships cannot be directly searched and retrieved by the catalog search engine. The catalog search engine returns a list of manuscript records by specified authors, location, or time periods, rather than providing direct answers to data-oriented queries. In addition, the current structure of the records does not provide links to relevant information found on the internet. Without employing semantic web technology, the search can only be performed in the current catalog records with no possibility to search other related manuscript catalogs and repositories on the web.

Unlike the catalog search which merely returns a list of manuscripts wherein the keywords of the query term appear, the ontology-based search will support complex queries and will return direct answers. Hence, research can be performed on relationships between people and events directly without searching the manuscripts in which they are mentioned. Examples of such direct answers could be a list of cities where manuscripts were written and the percentage of manuscripts from a selected city in the entire corpus. Retrieval of names of copyists will be possible through the corresponding ontological relationship (property) between persons and manuscripts. Retrieval of authors who participated in some event in a specific region (rather than a city) will be possible through the ontological links between persons and events, events, and places, and through the link between cities, countries, and regions, as opposed to the original records which only contain information on cities. In addition, the ontology will enable the user to receive answers, which are not explicitly encoded in the catalog records, such as masters and their students, manuscripts that were never printed, original manuscripts and their copies, censored manuscripts, parts of the same manuscript currently stored at separate geographic locations, family relations, manuscript owners, and the members of their families, through the corresponding ontological properties.

To this end, we developed a prototype of user interface that supports the following types of SPARQL endpoints automatically generated by pressing the search button: (1) queries by individual classes of the ontology (limited by some user-selected properties), e.g., all the persons who were honorees of some manuscripts, or manuscripts that were never printed, or historical events in Russia; (2) queries on two or more related classes, e.g., all the authors of manuscripts and their manuscripts that were printed and were originally written in Russia during the Emancipation reform period. **Figure 5** displays a prototype user interface of the search system based on the constructed ontology. As demonstrated in **Figure 5**, users are shown diverse classes and their inter-relationships (properties) and can choose at least two classes and one property to construct a query.

To demonstrate the strength of the rich semantics encoded in the ontology, **Figure 5** shows an example of a search for the query: "All the authors who were Rabbis and were related to some historical event in Russia" and its results. First, the interface displays the top classes of the ontology. Once the user choses (checks) the class/es, the related classes and subclasses are shown, and then the user can continue to browse the class hierarchy and modify her/his choice. Then, the user also selects the vocabularies in which to search. The system decomposes the query into ontological entities and applies reasoning of the ontology to retrieve triples relevant to the query, e.g., "*Rabbi is a (indirect) subclass of Manuscript agent*," "*Manuscript agents involved in Manuscript Biographic events*," "*involved in Agent biographic events that took place in Russia*," and also "*involved in Historical event that took place in Russia*." Using ontological reasoning, these triples are then generalized to corresponding facts that are displayed to the user.shown as relations in **Figure 5**, e.g., "*Rabbis who lived in Russia*," "*Rabbis who were authors of Historical manuscript*," and "*Rabbis who were related to Historical event*." Finally, out of all the found relations that have been found, the user has to select only the relations relevant to her/his search. If a user does not select specific relations, all the instances with all the displayed relations will be retrieved. Ontological reasoning rules (Noy et al., 2001) are applied for specific instances of the classes (e.g., specific places and persons) to induce that, for example, Dnepropertrovsk is a city in Russia. Therefore, if a person lived there, then it follows that he also lived in Russia and this should be retrieved as part of the given query.



Please select the concepts for your query:

⊟ Person
⊟ Religious figure
Cantor
⊞ ☑ Rabbi
⊟ □ Political figure
Rothshield

⊟ Manuscript
⊟ Historical manuscript
⊞ Medieval manuscript
⊞ ☑ Post-medieval manuscript

⊟ Event
⊞ ☑ Historical Event
Emancipation reform of 1861

⊟ Geographic location
⊟ Country
⊞ Israel
⊞ France
⊞ United Kingdom
⊞ ☑ Russia
⊟ □ Institution
National library in Jerusalem

Please select the vocabularies for your query:

Ontology for Historical Manuscripts ☑
DBpedia □
British library □

Please select the relations to search and then press the search button:

Get all the relations where: Rabbi is_author_of Post-medieval manuscript ☑
Get all the relations where:: Rabbi is_mentioned_in Post-medieval manuscript □
Get all the relations where: Rabbi is_related_to Historical event ☑
Get all the relations where: Rabbi lived_in Russia ☑
Get all the relations where: Historical-event took_place_in Russia □

Your Ontopedia Results:
1. Rabbi Simcha Avraham Katznelbogen is_related_to Emancipation reform of 1861 promulgated by Tsar Alexander II
2. Rabbi Simcha Avraham Katznelbogen is_author_of Kabbalistic Torah interpretation
3. Rabbi Simcha Avraham Katznelbogen lived_in Russia

FIGURE 5 | **A prototype of the user interface for query construction for selected ontological relations by browsing the ontological hierarchy/taxonomy**.

As a result, the proposed biographic ontology for historical Hebrew manuscripts will provide a basis for characterization of the entire corpus and its constituent parts. All the retrieved answers (a list of classes and instances for the first type of queries above or a list of triples for the second type of queries) are displayed to the user and stored in a digital library as part of the constructed ontopedia. Furthermore, the characterization of Hebrew manuscripts by selected parameters may be visualized by a chart with a map of the statistical distribution according to questions that will be asked about the manuscripts. The results for queries related to some periods or places will be represented as points on the timeline or geographic map, respectively. Clicking at such a point will display all the information on the corresponding object, such as person, event, or manuscript details and characteristics according to the composed query.

CONCLUSION

In this paper, we presented a generic event-based ontological model for historical Hebrew manuscripts. The ontology was built as an extension to the existing ontologies in the field of cultural heritage and facilitates their classes and properties. We used the catalog of the National Library in Jerusalem as a source of data on the manuscripts. A systematic analysis of the structure and content of a representative subset of the catalog records was performed, and a methodology for conversion of these records to the ontological data was proposed.

The resulting event-based ontological model will enable large-scale analysis and qualitative and quantitative research of the manuscripts and their cultural, social, and historical context. The results of our project will greatly contribute to the study of Hebrew manuscripts and cultural heritage. It will enable posing queries and cross-referencing data from various vocabularies in the semantic web. A sequence of events and places related to the manuscripts will constitute a timeline of history against which manuscripts, people, and their relationships can be placed. A social network of people associated with these manuscripts can be constructed. A large-scale automated reasoning will also enable researchers to compare the effect of time and place on the manuscripts' qualitative characteristics and quantitative distribution. In addition, the ontology has the potential to become a highly valuable resource for the scholarly research community and educated amateurs of the subject matter, representing rich scholarly semantics and high quality of data.

However, this research was based on a limited amount of dated manuscripts and to the specified period of time (post-medieval – sixteenth century and later). The proposed ontology model is not intended to be complete; instead, the model will be flexible to be further extended and updated as more manuscripts are analyzed and inserted in the ontology.

Thus, in future research to refine and extend the ontology, we will later expand the manuscript collection to additional 2,000 dated manuscripts from the same period from different collections, such as the National Library of Israel, the Bibliothèque Nationale de France, and the Ets Haim library in Amsterdam. To this end, we will use the state-of-the-art named entity extraction systems (Ben Mordecai and Elhadad, 2005) for semi-automatic conversion of the corresponding catalog records to ontological triples. RDF/OWL triple search engines (e.g., Open Link Virtuoso) available on the web will be utilized to link the extracted data to other existing vocabularies and authority files on the web.

AUTHOR CONTRIBUTIONS

All authors listed have made substantial, direct, and intellectual contribution to the work and approved it for publication.

ACKNOWLEDGMENTS

This research was supported by the Israel Science Foundation (grant no. 342/15).

REFERENCES

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). DBpedia: a nucleus for a web of open data. In *Proceedings of the 8th International Semantic Web Conference (ISWC2007) Volume 4825 of Lecture Notes in Computer Science*, 722–735. Berlin: Springer.

Baruchson-Arbib, S. (1993). *Sefarim ṿe-ḳor'im: tarbut ha-ḳeri'ah shel Yehude Iṭalyah be-shilhe ha-Renesans*. Ramat-Gan: Universiṭat Bar-Ilan (In Hebrew).

Beit-Arié, M. (2014). *Hebrew Codicology: Historical and Comparative Typology of Hebrew Medieval Codices Based on the Documentation of the Extant Dated Manuscripts from a Quantitative Approach, Pre-Publication, Internet Version 0.3 (In Hebrew)*. Available at: http://web.nli.org.il/sites/NLI/Hebrew/collections/manuscripts/hebrewcodicology/Pages/default2.aspx

Ben Mordecai, N., and Elhadad, M. (2005). *Hebrew Named Entity Recognition*. Master's thesis, Department of Computer Science, Ben Gurion University of the Negev, Beer Sheva.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American* 284: 28–37. doi:10.1038/scientificamerican0501-34

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* 1247–1250, Vancouver: ACM.

Brickley, D., and Guha, R. (1999). *Resource Description Framework (RDF) Schema Specification. W3C Proposed Recommendation*. Available at: http://www.w3.org/TR/1999/PR-rdf-schema-19990303/

Brown, S., Antoniuk, J., Brundin, M., Simpson, J., Ilovan, M., and Warren, R. (2015). An entity-based approach to interoperability in the Canadian Writing Research Collaboratory. In *Proceedings of the Digital Humanities Conference*. Sydney.

Ciula, A., Spence, P., and Vieira, J.M. (2008). Expressing complex associations in medieval historical documents: the Henry III Fine Rolls Project. *Literary and Linguistic Computing* **23**(3): 311–325.

D'Arcus, B.R.U.C.E., and Giasson, F. (2009). *Bibliographic Ontology Specification*. Available at: http://bibliontology.com/specification

Doerr, M. (2003). The CIDOC conceptual reference module – an ontological approach to semantic interoperability of metadata. *AI Magazine* 24: 75–92. doi:10.1145/1921614.1921615

Doerr, M., Kritsotaki, A., and Boutsika, A. (2011). Factual argumentation – a core model for assertions making. *Journal on Computing and Cultural Heritage (JOCCH)* 3: 34.

Dunsire, G. (2012). *Linked Data for Manuscripts in the Semantic Web. Summer School in the Study of Historical Manuscripts*. Available at: http://www.gordondunsire.com/pubs/docs/LinkedDataForManuscripts.pdf

Elyada, O. (1999). The 'annales' school and the culture of the book. In *Literature and History*, Edited by R. Cohen and J. Mali, 299–323. Jerusalem: The Zalman Shazar Center for Jewish History (In Hebrew).

Foxvog, D. (2010). *Cyc. In Theory and Applications of Ontology: Computer Applications*. Netherlands: Springer. 259–78.

Gamallo, P. (2014). An overview of open information extraction, In *Proceedings of the Third Symposium on Languages, Applications and Technologies (SLATE-2014)*. 13–16, Bragança, Portugal.

González-Blanco, E., Seláf, L., Del Rio Riande, M.G., Martínez Cantón, C.I., and Martos Pérez, M.D. (2014). Building a metrical ontology as a model to link digital poetic repertoires. In *Digital Humanities Conference Lausanne*. Available at: http://dharchive.org/paper/DH2014/Paper-674.xml

Hage, W.R., van Malaisé, V., Segers, R., Hollink, L., and Schreiber, G. (2011). Design and use of the simple event model (SEM). *Journal of Web Semantics* 9: 128–36. doi:10.1016/j.websem.2011.03.003

Harel, Y. (1997). *The Books of Aleppo: The Rabbinic Literature of the Scholars of Aleppo*. Jerusalem: Ben Zvi Institute for the study of Jewish communities in the East (In Hebrew).

Hyvönen, E., Alonen, M., Ikkala, E., and Mäkelä, E. (2014). Semantic national biography: an event-based approach to publishing life stories as linked data. In *Proceedings of ISWC. 18. ISO 15836:2009 – Information and Documentation – The Dublin Core Metadata Element Set. Iso.org. 18 February 2009*. Available at: http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=52142

Lenat, D., Witbrock, M., Baxter, D., Blackstone, E., Deaton, C., Schneider, D., et al. (2010). Harnessing cyc to answer clinical researchers' ad hoc queries. *AI Magazine* 31: 13–32.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics*, Vol. 2, 768–774. Quebec: Association for Computational Linguistics.

McGuinness, D.L., and Van Harmelen, F. (2004). OWL web ontology language overview. *W3C Recommendation*, 10(10): 2004.

Miles, A., Matthews, B., Wilson, M., and Brickley, D. (2005). SKOS core: simple knowledge organisation for the web. In *International Conference on Dublin Core and Metadata Applications*. Madrid, Spain.

Mohit, B. (2014). Named entity recognition. In *Natural Language Processing of Semitic Languages*. Edited by I. Zitouni, 221–245. Berlin: Springer.

Mulholland, P., Wolff, A., Kilfeather, E., and McCarthy, E. (2014). Using event spaces, setting and theme to assist the interpretation and development of museum stories. In *19th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2014), 24-28 November 2014*, 320–332. Linkoping: Springer.

Noy, N.F., Sintek, M., Decker, S., Crubézy, M., Fergerson, R. W., and Musen, M.A. (2001). Creating semantic web contents with protege-2000. *IEEE Intelligent Systems* 16(2), 60–71.

Pan, F., and Hobbs, J.R. (2005). Temporal Aggregates in OWL-Time. In *Proceedings of the 18th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, 560–565. Clearwater Beach, Florida: AAAI Press.

Pasternak, N. (2009). *Together and Apart: Hebrew Manuscripts as Testimonies to Encounters of Jews and Christians in Fifteenth-Century Florence the Makings, the Clients, Censorship*. Doctoral dissertation, Hebrew University Jerusalem, Jerusalem (In Hebrew).

Peroni, S., and Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *Web Semantics* 17: 33–43. doi:10.1016/j.websem.2012.08.001

Prebor, G. (2015). The manuscript collection of the Séminaire Israélite de France. Towards a new catalogue. *La Bibliofilia* CXVII: 3–26.

Richler, B. (2014). Lezione Dottorale. *Guide to Hebrew Manuscript Collection, Second Revised Edition*. Jerusalem: Israel Academy of Sciences and Humanities.

Richler, B. (2010–2011). Lezione Dottorale. *Materia Giudaica*: XV-XVI: 14–6.

Rigler, M. (1995). *Colophons of Medieval Hebrew Manuscripts as Historical Sources*. Ph.D. dissertation, Hebrew University of Jerusalem, Jerusalem (In Hebrew).

Sarasua, C., Simperl, E., and Noy, N.F. (2012). Crowdmap: crowdsourcing ontology alignment with microtasks. In *The Semantic Web–ISWC 2012*, Edited by P. Cudré-Mauroux, J. Heflin, E.S. Tania, T. Jérôme Euzenat, M. Hauswirth, J.X. Parreira, J. Hendler, et al., 525–541. Berlin: Springer.

SfarData. (n.d.). *The Codicological Data-Base of the Hebrew Palaeography Project*. The Israel Academy of Sciences and Humanities. Available at: http://sfardata.nli.org.il/sfardatanew/home.aspx

Shaw, R., Troncy, R., and Hardman, L. (2009). LODE: linking open descriptions of events. In *The Semantic Web SE – 11*, Vol. 5926, Edited by A. Gómez-Pérez, Y. Yu, and Y. Ding, 153–167. Berlin: Springer.

Sirat, C. (2002). *Hebrew Manuscripts of the Middle Ages*. Cambridge; New York: Cambridge University Press.

Sirat, C. (2013). *Hebrew Manuscripts of the Middle Ages: An Introduction to Student (In Hebrew)*. Available at: http://hsf.bgu.ac.il/cjt/files/Sirat/sirat%20palographie/man2.pdf

Tavoni, M., Andriani, P., Bartalesi, V., Locuratolo, E., Meghini, C., and Versienti, L. (2014). Towards a semantic network of Dante's works and their contextual knowledge. In *Digital Humanities Conference Lausanne*. Available at: http://dharchive.org/paper/DH2014/Paper-417.xml

Theodoridou, M., Tzitzikas, Y., Doerr, M., Marketakis, Y., and Melessanakis, V. (2010). Modeling and querying provenance by extending CIDOC CRM. *Distributed and Parallel Databases* 27: 169–210. doi:10.1007/s10619-009-7059-2

Tillett, B. (2005). What is FRBR? A conceptual model for the bibliographic universe. *The Australian Library Journal* 54: 24–30. doi:10.1080/00049670.2005.10721710

Uschold, M., and Gruninger, M. (1996). Ontologies: principles, methods and applications. *The Knowledge Engineering Review* 11: 93–136. doi:10.1017/S0269888900007797

Vieira, J.M., and Ciula, A. (2007). Implementing an RDF/OWL ontology on Henry the III fine rolls. In *BT – Proceedings of the OWLED 2007 Workshop on OWL: Experiences and Directions*. Innsbruck. Available at: http://ceur-ws.org/Vol-258/paper06.pdf

Winer, D. (2011). Judaica Europeana D2.5: semantic interoperability report with representation of selected controlled vocabularies in RDF/SKOS. In *Semantic Interoperability Report with Representation of Selected Controlled Vocabularies in RDF/SKOS*. Available at: http://www.judaica-Europeana.eu/docs/D2-5_Semantic_interoperability_report.pdf

Winer, D. (2014). Judaica Europeana: an infrastructure for aggregating Jewish content. *Judaica Librarianship* 18: 88–115. doi:10.14263/2330-2976.1027

Wolf, L., Litwak, L., Dershowitz, N., Shweka, R., and Choueka, Y. (2011). Active clustering of document fragments using information derived from both images and catalogs. In *BT – IEEE International Conference on Computer Vision, ICCV 2011*. Barcelona.

Yaari, A. (1936). *Hebrew Printing in the East*. Jerusalem: Hebrew University Press (In Hebrew).

Zhitomirsky-Geffet, M., and Dagan, I. (2009). Bootstrapping distributional feature vector quality. *Computational Linguistics* 35: 435–61. doi:10.1162/coli.08-032-R1-06-96

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.