# Mapping Manuscript Migrations on the Semantic Web: A Semantic Portal and Linked Open Data Service for Premodern Manuscript Research

Eero Hyvönen[1,2(✉)], Esko Ikkala[1], Mikko Koho[1,2], Jouni Tuominen[1,2], Toby Burrows[3], Lynn Ransom[4], and Hanno Wijsman[5]

[1] Semantic Computing Research Group (SeCo), Aalto University, Espoo, Finland
eero.hyvonen@aalto.fi
[2] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Helsinki, Finland
mikko.koho@helsinki.fi
[3] Oxford e-Research Centre, University of Oxford, Oxford, UK
[4] Schoenberg Institute for Manuscript Studies, University of Pennsylvania, Philadelphia, USA
[5] Institut de recherche et d'histoire des textes, Aubervilliers, France

**Abstract.** This paper presents the *Mapping Manuscript Migrations* (MMM) system in use for modeling, aggregating, publishing, and studying heterogeneous, distributed premodern manuscript databases on the Semantic Web. A general "Sampo model" is applied to publishing and using linked data in Digital Humanities (DH) research and to creating the MMM system that includes a semantic portal and a Linked Open Data (LOD) service. The idea is to provide the manuscript data publishers with a novel collaborative way to enrich their contents with related data of the other providers and by reasoning. For the end user, the MMM Portal facilitates semantic faceted search and exploration of the data, integrated seamlessly with data analytic tools for solving research problems in manuscript studies. In addition, the SPARQL endpoint of the LOD service can be used with external tools for customized use in DH research and applications. The MMM services are available online, based on metadata of over 220 000 manuscripts from the Schoenberg Database of Manuscripts of the Schoenberg Institute for Manuscript Studies (University of Pennsylvania), the Medieval Manuscripts in Oxford Libraries, and Bibale of Institut de recherche et d'histoire des textes in Paris. Evaluation of the MMM Portal suggests that the system is useful in manuscript studies and outperforms current systems online in searching, exploring, and analyzing data.

**Keywords:** Manuscripts · Semantic portals · Linked data · Digital humanities

## 1 Introduction

The study of premodern manuscripts, i.e., books and documents produced before the age of print, is an essential element in understanding our shared intellectual and cultural heritages across time and geographies [6]. Manuscripts, unlike printed books, are unique witness to the times in which they were produced. Whereas a printed copy of a text exists in multiple identical copies, the textual contents of premodern manuscripts reflect specific circumstances of production that cannot be reproduced in other copies

A. Hotho et al. (Eds.): ISWC 2021, LNCS 12922, pp. 615–630, 2021.
https://doi.org/10.1007/978-3-030-88361-4_36

of the same text or textual groupings. Over the centuries, manuscripts have been bought and sold, stolen and lost, and broken up and rebound. Hundreds of thousands of European premodern manuscripts have survived until the present day.

Consider. e.g., the Christian Bible, repeatedly copied, translated, revised, and disseminated in a variety of formats until the 13th century when the book started to look something like the modern standardized Bible with chapter and verse divisions contained in a single volume in two-column format in a hand-holdable size. This process began with manuscripts resembling the Dead Sea Scrolls. Another example is Marco Polo's (1254–1324) original text *The Travels of Marco Polo* that he dictated in a prison to a fellow inmate. The original copy of his words has not been found, but a total of about 150 copies in various languages and produced at different times throughout the Middle Ages are known to exist in different collections.

Over the last twenty years there has been a proliferation of digital data relating to premodern manuscripts, including catalogues, specialist databases, and numerous collections of digital images[1]. The databases may contain metadata about the manuscripts, and also transliterated texts extracted from them, possibly with translations. However, there is little in the way of having a coherent, *interoperable digital infrastructure for the manuscript data* for Digital Humanities (DH) research [9,23]. As a result, cross-collection discovery and analysis requires the time-consuming exploration of numerous disparate resources. To mitigate this problem, this paper introduces the Mapping Manuscript Migrations (MMM) system, an outcome of the MMM project[2] [4]. MMM is a data publishing framework including a semantic portal demonstrator and a Linked Open Data (LOD) service for manuscript studies. The model supports several user groups: 1) The data publishers are provided with a collaborative model for harmonizing, enriching, and publishing their content in a shared knowledge graph hosted by a LOD server. 2) Collection managers and curators are facilitated with a semantic portal for accessing the enriched collections in order to develop and maintain their own collections. 3) Manuscript researchers are provided with a semantic portal for exploring, visualizing, and analyzing the data with seamlessly integrated data-analytic tooling without technical expertise. The researchers can also use the SPARQL endpoint and other APIs of the framework directly for custom-made analyses. 4) The APIs can be used by system developers for creating new applications on top of the data service. The MMM Portal[3] and LOD service[4] are in pilot use on the Semantic Web since 2020.

In the following, we first introduce the data and data model of MMM. After this the "Sampo model" for publishing and using data in DH is presented and applied to the MMM case study to create the MMM Portal and data service. Using the MMM Portal and the LOD service for studying the manuscripts are discussed with examples, including a presentation of the implementation. Finally, evaluation of the usability of the portal is discussed, contributions of the paper are summarized in relation to related works, and lessons learned are summarized. This paper concerns the MMM system from a LOD publishing and portal design perspectives, complementing our earlier papers on

[1] Using, e.g., IIIF: https://iiif.io.

[2] https://seco.cs.aalto.fi/projects/mmm/.

[3] https://mappingmanuscriptmigrations.org.

[4] https://www.ldf.fi/dataset/mmm.

the MMM project in general [4], on MMM data modeling and data transformations [18], tooling for implementing the portal interface [17], and on evaluating the system with end users [5].

## 2 Modeling Manuscript Data

**MMM Data.** The MMM knowledge graph (KG) aggregates data from three databases in which different data models and data base systems were used. Furthermore, the data contained in the databases was different in nature, including, e.g., both records of manuscripts and observations about them, such as transfers of custody in auctions.

1. **Bibale**[5]. The Bibale data comes from the Institute for Research and History of Texts (IRHT). The 55 000 Bibale database records belong to one of eight object types: manuscripts, works, persons, bindings, collections, ownership marks, texts, and sources.
2. **Schoenberg Database of Manuscripts (SDBM)**[6]. Entries in the SDBM use up to 36 fields to record data from observations of manuscripts found in published and unpublished sources. The data is in a MySQL relational database and contains over 250 000 records focusing on provenance-related manuscript histories.
3. **Medieval Manuscripts in Oxford Libraries (MMOL)**[7]. The MMOL dataset in MMM covers 10 272 manuscripts represented in TEI format[8].

Each of the source datasets 1–3 has its own preconditions and goals, and thus follows its own data modeling conventions. Therefore, a unified data model for harmonizing the datasets was needed as well as a pipeline for transforming the datasets into the harmonized model including aligning the data values used in the metadata elements, such as historical people and places. For this purpose a set of shared ontologies was selected, such as the Getty Thesaurus of Geographic Names[9] (TGN), and both automatic and semi-automatic tools were used in the data transformation process.

A major challenge for the data harmonization was that the databases contain data that is semantically different in nature. Bibale and MMOL contain traditional metadata about the manuscripts, e.g., who is the author, when the text was written, and the shelf mark of the document. In contrast, SDBM focuses on provenance metadata about the object, e.g., who has owned the manuscript in different times, where has it been, and what has happened to it during the centuries. Actually, the fundamental question “what is a manuscript” is not easy to answer based on the entries in different databases. The different concepts related to a manuscripts as physical units (e.g., manuscript group, volume, item, part, fragment) are inconsistently used or missing, often even within a single database. Creating a comprehensive model covering all this variation of information is a challenge from a data modeling perspective.

[5] The current web service is described in http://bibale.irht.cnrs.fr.

[6] See https://sdbm.library.upenn.edu for details about the SDBM data and the web service.

[7] See https://medieval.bodleian.ox.ac.uk for a catalogue of Western manuscripts at the Bodleian Libraries and selected Oxford colleges.

[8] https://tei-c.org.

[9] https://www.getty.edu/research/tools/vocabularies/tgn/.

**MMM Data Model.** When dealing with premodern manuscripts, it is important to be able to make the distinction between the abstract "distinct intellectual or artistic creation" behind a manuscript (*work*, in terms of the Functional Requirements for Bibliographic Records (FRBR) model[10] [20, 27]), "the specific intellectual or artistic form that a work takes each time it is realized" (*expression* in FRBR), say a translation, and the "the physical embodiment of an expression of a work" (manifestation in FRBR). The manifestation represents all the physical objects that bear the same characteristics, in respect to both intellectual content and physical form, i.e., *items* in FRBR terminology.

The harmonizing MMM data model as well as the data harmonization pipeline is presented in detail in [18]. The model is a result of thorough discussions between manuscript researchers and computer scientists in the MMM project and is based on FRRBoo and CIDOC CRM[11]. The final model has 16 main classes for describing manuscripts and related intellectual property, seven classes for describing collections, and nine classes for representing transactions and manuscript observations with some 40 properties in between. For the purposes of this paper, focusing on using the MMM Portal on top of the data service, it is sufficient to consider the classes represented in Table 1, based on the Erlangen CIDOC CRM[12] and FRBRoo[13] namespaces. This is because the MMM Portal is based on searching instances of these classes and on performing data-analyses on subsets of the instances of these classes. These instances are characterized in terms of sets of properties whose values are represented as facets, such as places in a meronymy, in the faceted search engines of the MMM Portal. Table 2 summarizes the facet properties pertaining to the classes of Table 1. The most complex class is Manuscript whose instances may have 22 different properties.

**Table 1.** Main classes of the MMM knowledge graph whose instances are searched for in the MMM Portal.

| Class | # of inst. | URI | Meaning |
|---|---|---|---|
| Manuscript | 222 605 | frbroo:F4_Manifestation_Singleton | Physical manuscript objects |
| Work | 435 428 | frbroo:F1_Work | Intellectual manuscript contents |
| Event | 937 158 | crm:E5_Event | Events related to the manuscripts |
| Actor | 56 685 | crm:E39_Actor | People and institutions |
| Place | 5077 | crm:E52_Place | Places related to manuscripts and actors |

## 3 Application of the Sampo Model to the MMM System

The Sampo model [15] is a consolidated set of principles listed is Table 3 for collaborative publishing and using of LOD on the Semantic Web. The model has been developed gradually and tested in a dozen of online cultural heritage "Sampo" portals in 2002–2021

[10] https://www.ifla.org/publications/functional-requirements-for-bibliographic-records.

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

[12] crm = http://erlangen-crm.org/current/.

[13] frbroo = http://erlangen-crm.org/efrbroo/.

**Table 2.** Properties and property paths for the main classes of the MMM Portal in Table 1 that are used as facets in the MMM Portal.

| Class | # | Properties (facets) |
|---|---|---|
| Manuscript | 22 | Manuscript, Author, Work, Production place, Production data, Note, Language, Owner, Collection, Transfer of custody place, Transfer of custody date, Last known location, Material, Height, Width, Folios, Lines, Columns, Miniatures, Decorated initials, Historiated initials, Source |
| Work | 6 | Title, Possible author, Language, Manuscript production date, Collection, Source |
| Event | 5 | Type, Manuscript/Collection, Date, Place, Source |
| Actor | 6 | Name, Type, Birth/formation date, Death/dissolution date, Activity location, Source |
| Place | 3 | Name, Parent place, Source |

**Table 3.** Sampo model principles P1–P6

| |
|---|
| P1. Support collaborative data creation and publishing |
| P2. Use a shared open ontology infrastructure |
| P3. Provide multiple perspectives to the same data |
| P4. Standardize portal usage by a simple filter-analyze two-step cycle |
| P5. Support data analysis and knowledge discovery in addition to data exploration |
| P6. Make clear distinction between the LOD service and the user interface (UI) |

that have had up to millions of end users[14]. The model is based on standards and best practices of W3C for Linked Data publishing [11,12] supporting FAIR principles[15].

The Sampo model concerns only publishing data, not issues of maintaining linked data. It is assumed that there is a separate pipeline that creates the linked data in a SPARQL endpoint. This section shows how the principles P1–P6 were applied to the MMM system.

**P1. Support Collaborative Data Creation and Publishing.** The Sampo model is based on the idea of collaborative content creation, where data is aggregated, harmonized, and interlinked from multiple data silos in a global data service, based on a shared ontology infrastructure. The local data is enriched with each other by linking and by reasoning, based on Semantic Web standards[16]. This is arguably a win-win model for data publishers to join and, especially, for the end users of the enriched data.

Figure 1 depicts the overall publication model of the MMM system. The three datasets are transformed (T1–T3 in the figure) into the unified harmonizing data model used in the MMM Linked Data Service that is depicted in the middle of the figure. The data service can be used in external applications via the SPARQL endpoint (on the left), and the data is also documented and can be studied using publishing tools (on the right).

**P2. Use a Shared Open Ontology Infrastructure.** In MMM the key idea is to enrich data from the three databases with each other, as the same manuscripts, persons, places,

[14] See https://seco.cs.aalto.fi/applications/sampo/ for more info about the Sampo portals.

[15] https://www.go-fair.org/fair-principles/.

[16] https://www.w3.org/standards/semanticweb/.

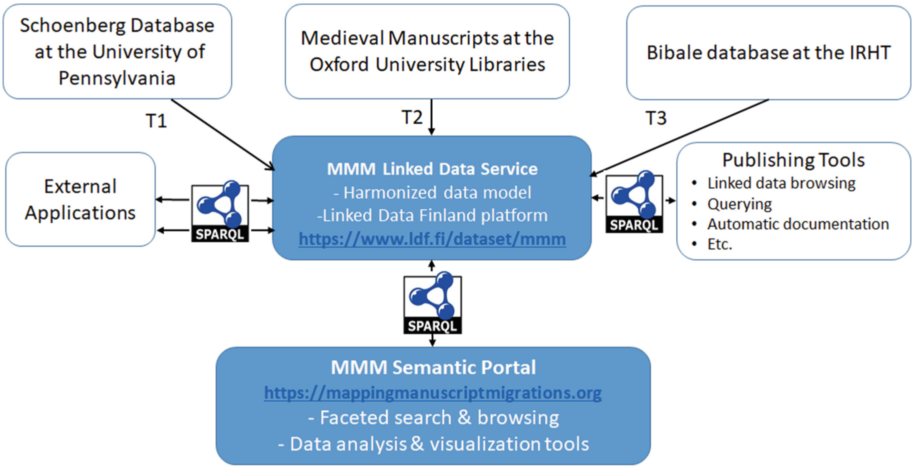

Schoenberg Database at the University of Pennsylvania
Medieval Manuscripts at the Oxford University Libraries
Bibale database at the IRHT
T1
T2
T3
External Applications
SPARQL
MMM Linked Data Service
- Harmonized data model
-Linked Data Finland platform
https://www.ldf.fi/dataset/mmm
SPARQL
Publishing Tools
• Linked data browsing
• Querying
• Automatic documentation
• Etc.
SPARQL
MMM Semantic Portal
https://mappingmanuscriptmigrations.org
- Faceted search & browsing
- Data analysis & visualization tools

**Fig. 1.** MMM publishing model

and other entities can be mentioned in all of them. The key elements of the underlying ontology infrastructure are the data model of Sect. 2 and a set of domain ontologies, such as TGN and an ontology of historical people, that are used for populating the instances of the data model classes.

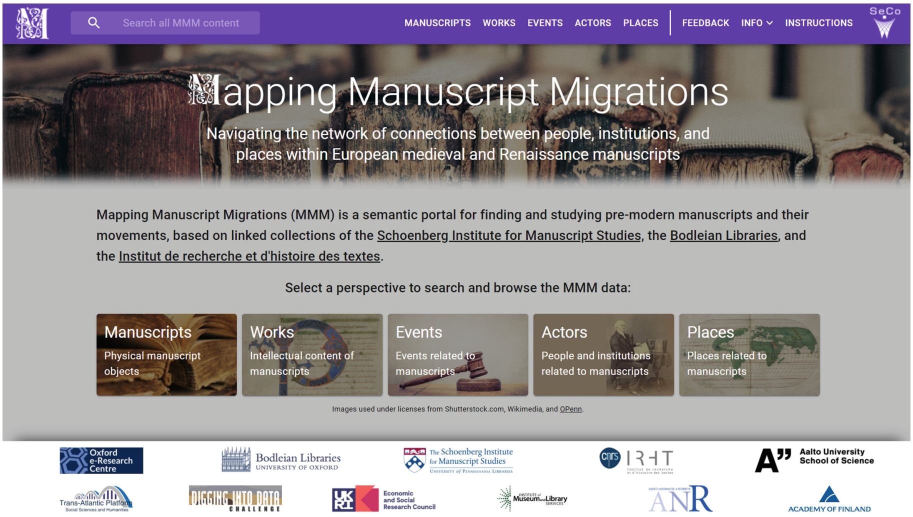
**P3. Provide Multiple Perspectives to the Same Data.** Sampo model fosters the idea that on top of the LOD service different *application perspectives* to the data can be created by re-using the data service, without modifying the data, which is typically costly. In each perspective, the result set can be studied through a set of visualizations, e.g., as a table, using a chart, or on maps. Furthermore, each instance to be searched for in the perspectives has a homepage aggregating data about it with the possibility of providing visualizations of the individual and its relations.

The perspectives are provided on the landing page of the Sampo system, and enrich each other by data linking. By selecting one of them the corresponding application is opened. The landing page of the MMM Portal depicted in Fig. 2 offering five perspectives for digging into the data: Manuscripts, Works, Events, Actors, and Places.

**P4. Standardize Portal Usage by a Simple Filter-Analyze Two-Step Cycle.** The application perspectives can be used by a two-step cycle for research: Firstly, the focus of interest, the target group, is filtered out using faceted semantic search [30,31]. Secondly, the target group is visualized or analyzed by using ready-to-use DH tools of the application perspectives. This idea was inspired by the research method used in prosopographical research [32][17].

In the MMM Portal each application perspective enables the user to filter out instances of the core class of the perspective (cf. Table 1). After this, the filtered

[17] Prosopography is a method that is used to study groups of people through their biographical data. The goal of prosopography is to find connections, trends, and patterns from these groups.



Mapping Manuscript Migrations

Navigating the network of connections between people, institutions, and places within European medieval and Renaissance manuscripts

Mapping Manuscript Migrations (MMM) is a semantic portal for finding and studying pre-modern manuscripts and their movements, based on linked collections of the Schoenberg Institute for Manuscript Studies, the Bodleian Libraries, and the Institut de recherche et d'histoire des textes.

Select a perspective to search and browse the MMM data:

Manuscripts
Physical manuscript objects

Works
Intellectual content of manuscripts

Events
Events related to manuscripts

Actors
People and institutions related to manuscripts

Places
Places related to manuscripts

Images used under licenses from Shutterstock.com, Wikimedia, and OPenn.

Oxford e-Research Centre
Bodleian Libraries UNIVERSITY OF OXFORD
The Schoenberg Institute for Manuscript Studies University of Pennsylvania Libraries
cnrs
IRHT
Aalto University School of Science
Trans-Atlantic Platform Social Sciences and Humanities
DIGGING INTO DATA CHALLENGE
UK RI Economic and Social Research Council
Institute of Museum and Library Services
ANR
ACADEMY OF FINLAND

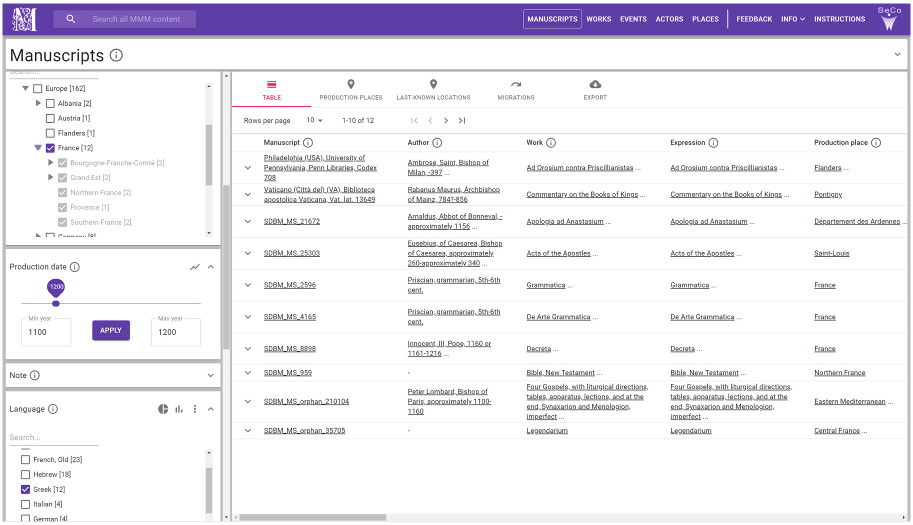
**Fig. 2.** MMM Portal landing page

instances can be explored and browsed for close reading, or data-analytic tools can be applied to the filtered result set for distant reading [24,28].

The facets in each perspective are the same as the properties of the corresponding classes in Table 2. For example, Fig. 3 depicts the Manuscripts perspective that the user has selected on the landing page. The user has made three clicks on the facets on the left: *Place of production* = France; *Production date* = 1100–1200; *Language* = Greek. The 12 results found are shown as a table on the right, paginated in groups of ten manuscripts. The table columns correspond to the facets and the metadata involved. Notice that some facets, such as *Place of production* based on the Getty Thesaurus of Geographical Names (TGN), are hierarchical. By selecting *France*, all provinces, cities villages etc. within France are automatically included in the search—the user does not need to know more about the placenames in France. This is arguable useful even if the semantic problems of representing historical places are challenging in many ways due to, e.g., temporal changes [16,29]. In our case study, Bibale data was originally based on references to the contemporary GeoNames[18] gazetteer, but in SDBM and MMOL data TGN was already used as the main place authority. To align the gazetteers, a mapping from GeoNames to TGN was created as there was none available.

**P5. Support Data Analysis and Knowledge Discovery in Addition to Data Exploration.** The model aims, as discussed in [14], not only at data publishing with search and data exploration [22], but also to data analysis and knowledge discovery with seamlessly integrated tooling for finding, analysing, and even solving research problems in interactive ways, based on AI techniques.

[18] https://geonames.org.



☐ Europe [162]
☐ Albania [2]
☐ Austria [1]
☐ Flanders [1]
☑ France [12]
☑ Bourgogne-Franche-Comté [2]
☑ Grand Est [2]
☑ Northern France [2]
☑ Provence [1]
☑ Southern France [2]
☐ Germany [8]

Production date
1200
Min year 1100
Max year 1200

☐ French, Old [23]
☐ Hebrew [18]
☑ Greek [12]
☐ Italian [4]
☐ German [4]

Rows per page 10 1-10 of 12

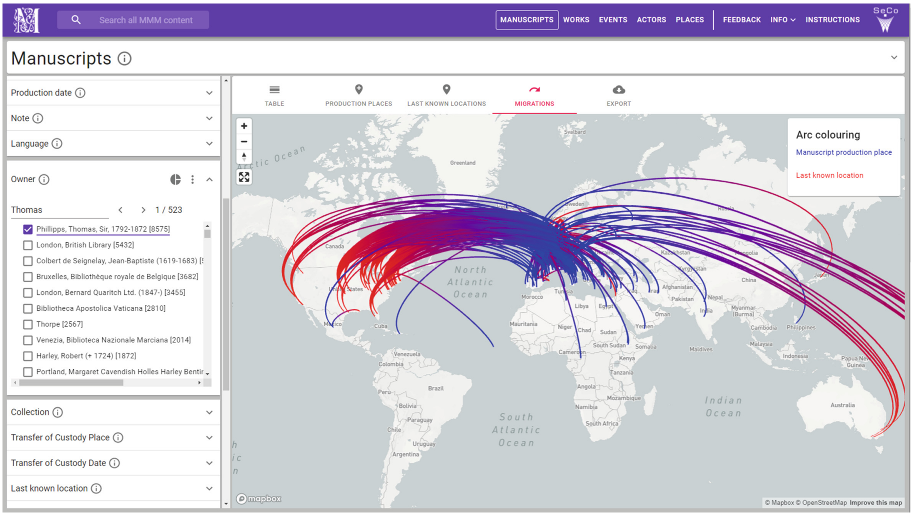
| Manuscript | Author | Work | Expression | Production place |
|---|---|---|---|---|
| Philadelphia (USA), University of Pennsylvania, Penn Libraries, Codex 708 | Ambrose, Saint, Bishop of Milan, -397 ... | Ad Orosium contra Priscillianistas ... | Ad Orosium contra Priscillianistas ... | Flanders ... |
| Vaticano (Città del) (VA), Biblioteca apostolica Vaticana, Vat. lat. 13649 | Rabanus Maurus, Archbishop of Mainz, 784?-856 | Commentary on the Books of Kings ... | Commentary on the Books of Kings ... | Pontigny |
| SDBM_MS_21672 | Arnaldus, Abbot of Bonneval, -approximately 1156 ... | Apologia ad Anastasium ... | Apologia ad Anastasium ... | Département des Ardennes ... |
| SDBM_MS_25303 | Eusebius, of Caesarea, Bishop of Caesarea, approximately 260-approximately 340 ... | Acts of the Apostles ... | Acts of the Apostles ... | Saint-Louis |
| SDBM_MS_2596 | Priscian, grammarian, 5th-6th cent. | Grammatica ... | Grammatica ... | France |
| SDBM_MS_4165 | Priscian, grammarian, 5th-6th cent. | De Arte Grammatica ... | De Arte Grammatica ... | France |
| SDBM_MS_8898 | Innocent, III, Pope, 1160 or 1161-1216 ... | Decreta ... | Decreta ... | France |
| SDBM_MS_959 | - | Bible, New Testament ... | Bible, New Testament ... | Northern France |
| SDBM_MS_orphan_210104 | Peter Lombard, Bishop of Paris, approximately 1100-1160 | Four Gospels, with liturgical directions, tables, apparatus, lections, and at the end, Synaxarion and Menologion, imperfect ... | Four Gospels, with liturgical directions, tables, apparatus, lections, and at the end, Synaxarion and Menologion, imperfect ... | Eastern Mediterranean ... |
| SDBM_MS_orphan_35705 | - | Legendarium | Legendarium | Central France ... |

**Fig. 3.** Manuscripts perspective in the MMM Portal

In MMM, reasoning is to used to enrich the data by rules based on SPARQL CONSTRUCT and SPARQL path expressions in a pre-processing phase. For example, reasoning was used for determining the last known locations of the manuscripts based on provenance data. On the data analysis and knowledge discovery side, it is possible to create alternative data analytic visualizations, represented as separate tabs, for the result set in addition to the table view illustrated in Fig. 3. For example, in the case of the Manuscripts perspective, there are the following tabs available in addition to the default TABLE view: 1) PRODUCTION PLACES tab shows the results on a map based on their place of production. 2) LAST KNOWN LOCATIONS tab shows the last known location of the manuscripts in the same vein. 3) MIGRATIONS tab shows how the filtered manuscripts have migrated from the place of production to the last know location. This is illustrated in Fig. 4 for the 8575 manuscripts owned by the well-known collector Sir Thomas Phillipps (1792–1872). This visualization is an answer to one of the original research question in manuscript studies set when starting the MMM project [3].

In addition to analyzing and visualizing the results on tabs, the facets provide buttons for visualizing the statistics of the results along the facet dimensions. For example, the *Production date* facet provides a button for showing the Phillipps manuscripts distribution along a timeline and the *Owner* facet a button for visualizing the distribution of former and current owners of the manuscripts in the result set as a pie chart.

In addition to studying result sets, each instance in the result set is associated with an information "homepage" that contains an aggregated description on the instance and how it is related to other instances. For example, Sir Thomas Phillipps can be found as a person instance in the Actors perspective with the following metadata fields on his homepage as a table: full name, birth and death dates, locations of activities, works created by the person (none in the case of Thomas Phillipps), manuscripts related to



Production date

Note

Language

Owner

Thomas 1 / 523

☑ Phillipps, Thomas, Sir, 1792-1872 [8575]
☐ London, British Library [5432]
☐ Colbert de Seignelay, Jean-Baptiste (1619-1683) [5
☐ Bruxelles, Bibliothèque royale de Belgique [3682]
☐ London, Bernard Quaritch Ltd. (1847-) [3455]
☐ Bibliotheca Apostolica Vaticana [2810]
☐ Thorpe [2567]
☐ Venezia, Biblioteca Nazionale Marciana [2014]
☐ Harley, Robert (+ 1724) [1872]
☐ Portland, Margaret Cavendish Holles Harley Bentir

Collection

Transfer of Custody Place

Transfer of Custody Date

Last known location

**Fig. 4.** Migrations of manuscripts owned by Sir Thomas Phillipps (1792–1872) from the place of production (blue end of an arc) to the last known location (red end of the arc) (Color figure online)

the person, and roles of the person in the data (collection owner, manuscript owner, and selling agent for Sir Thomas). Also the URI and the class of the instance are shown.

**P6. Make Clear Distinction Between the LOD Service and the User Interface (UI).** The architecture in Fig. 1 makes a clear distinction between the MMM Linked Data Service and the user interface, i.e., MMM Semantic Portal, based on only the standard SPARQL API. The MMM knowledge graph is available on the Linked Data Finland (LDF) platform [13], providing a home page for the dataset and its graphs[19], and a public SPARQL endpoint[20]. The homepage provides information, such as schema documentation automatically generated by the platform (using the LODE service[21] [25]), sample SPARQL queries, and metadata using *SPARQL Service Description*[22] and *Vocabulary of Interlinked Datasets (VoID)*[23]. The LDF platform also provides the user with dereferencing of URIs for both human users and machines, and a generic RDF browser for technical users, which opens when a URI is visited directly with a web browser. The data is also available as a data dump on the Zenodo repository[24] with a canonical citation [19].

[19] The home page of the KG: https://www.ldf.fi/dataset/mmm.
[20] The public SPARQL endpoint: http://ldf.fi/mmm/sparql.
[21] https://essepuntato.it/lode/.
[22] https://www.w3.org/TR/sparql11-service-description/.
[23] https://www.w3.org/TR/void/.
[24] https://zenodo.org/record/4440464.

## 4 Using the Data Service

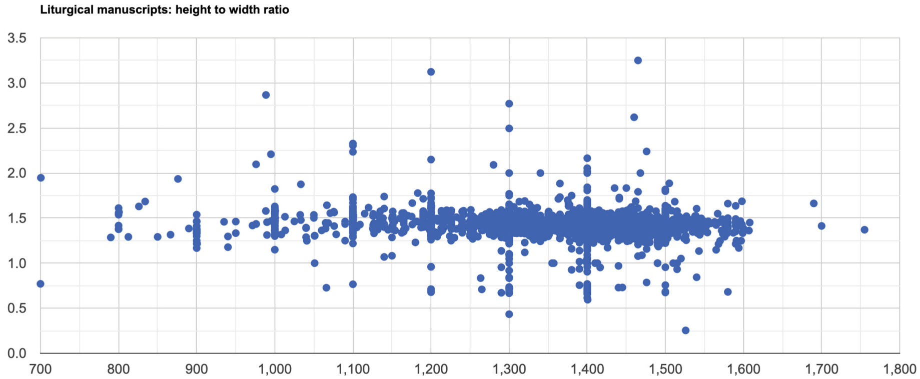


**Fig. 5.** Visualization of height to width ratios of liturgical manuscripts through SPARQL

In addition to using the MMM Portal, the MMM LOD service can be used directly via the SPARQL endpoint. This expands the possibilities for running complex research questions against the data. For example, the question "What are the ratios of height to width in liturgical manuscripts[25] produced between 700AD and 1800AD?" can be addressed through a SPARQL query[26], but not through the MMM Portal interface. The ratios calculated for 4 030 liturgical manuscripts are shown in Fig. 5 where the x-axis represents the year of production 700–1800 and y-axis the ratio. Most manuscripts have a ratio between 1.25 and 1.6, while ratios of less than 1.0 are only found for a small number of manuscripts which are wider than they are tall. The types of manuscripts covered are missals, breviaries, antiphonals, and graduals.

Manuscripts often have production dates in the form of an estimated range, such as "1200–1300", since the exact date is unknown. This query uses the earliest date in the range. It also averages the dates when a manuscript has more than one estimated production range, usually because of differences between the source datasets or because of multiple records for the same manuscript in the Schoenberg Database. Averaging is also used when a manuscript has more than one set of height and width measurements, for similar reasons. The query can also be adjusted to show ratios for each specific sub-type of liturgical manuscripts, as well as for other types of manuscripts.

To illustrate how the SPARQL endpoint is programmatically used by the MMM Portal for implementing faceted search coupled with data analytic tools, the relatively short query[27] for creating the migrations visualization in Fig. 4 is listed below:

[25] Liturgical manuscripts are retrieved using string comparison on the work labels as there is no classification of manuscript types in the data sources.

[26] The SPARQL query can be seen and run at: https://api.triplydb.com/s/czV6XZJx8.

[27] The SPARQL query can be seen and run at: https://api.triplydb.com/s/91ZiMF51i.

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX crm: <http://erlangen-crm.org/current/>
PREFIX mmm-schema: <http://ldf.fi/schema/mmm/>
PREFIX mmm-actor: <http://ldf.fi/mmm/actor/>

SELECT DISTINCT
?arc_id ?from_id ?from_prefLabel ?from_lat ?from_long
?to_id ?to_prefLabel ?to_lat ?to_long
(COUNT(DISTINCT ?manuscript) as ?instanceCount)
WHERE {
  ?manuscript crm:P51_has_former_or_current_owner
      mmm-actor:bodley_person_73979081 ; # Sir Thomas Phillipps
      ^crm:P108_has_produced/crm:P7_took_place_at ?from_id ;
      mmm-schema:last_known_location ?to_id .
  ?from_id skos:prefLabel ?from_prefLabel ;
           geo:lat ?from_lat ;
           geo:long ?from_long .
  ?to_id skos:prefLabel ?to_prefLabel ;
         geo:lat ?to_lat ;
         geo:long ?to_long .
  BIND(IRI(CONCAT(STR(?from_id), "-", REPLACE(STR(?to_id),
    "http://ldf.fi/mmm/place/", ""))) as ?arc_id)
  FILTER(?from_id != ?to_id) # ignore manuscripts that have stayed put
}
GROUP BY ?arc_id ?from_id ?from_prefLabel ?from_lat ?from_long
?to_id ?to_prefLabel ?to_lat ?to_long
ORDER BY desc(?instanceCount)
```

The query fetches all unique arcs from place of production to last known location, and counts how many manuscripts have travelled that route. The number of manuscripts is used for scaling the width of the arcs in the interactive visualization. Manuscripts are limited to those owned by Sir Thomas Phillipps at some point of time. Here the benefits of the LOD approach implemented in the MMM data conversion pipeline can be clearly seen: the Bibale[28], SDBM[29], and MMOL[30] records for Sir Thomas have been merged into one MMM record[31], and all references in the data have been corrected to point to this unified record.

Due to missing data, only place of production and last known location are used in the query. If there were more complete and harmonized data in the source databases about the locations and dates of the manuscripts throughout their histories, the query could be expanded for visualizing the full details of the movement of a limited group of individual manuscripts as series of arcs numbered in chronological order.

Experiences in using the MMM data service by SPARQL are discussed in more depth in [2].

## 5 Implementation: MMM Portal and Data Service

**MMM Portal.** The user interface of the MMM Portal is implemented as a web-based application[32], written purely in JavaScript. The general architecture, provided

[28] http://bibale.irht.cnrs.fr/933.

[29] https://sdbm.library.upenn.edu/names/7182.

[30] https://medieval.bodleian.ox.ac.uk/catalog/person_73979081.

[31] http://ldf.fi/mmm/actor/bodley_person_73979081.

[32] https://github.com/SemanticComputing/mmm-web-app.

by the Sampo-UI framework [17], is presented in Fig. 6. The application consists of a NodeJS[33] backend build with Express framework[34] (top right) and a client based on React[35] and Redux[36] (top left). The client makes use of base maps from external map services (bottom left). The MMM Data Service is shown on the bottom right corner.

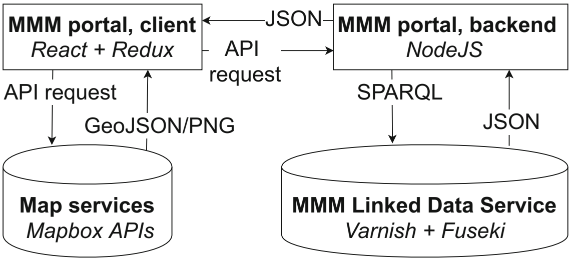


**Fig. 6.** MMM Portal architecture

**MMM Linked Data Service.** The MMM knowledge graph is published on the Linked Data Finland platform, which is powered by a combination of Fuseki SPARQL server[37] for storing the primary data[38] and a Varnish Cache web application accelerator[39] for routing URIs, content negotiation, and caching.

**Deployment.** The portal and the data service implementation are based on a microservice architecture, using Docker containers[40]. Each individual component (MMM Portal, Varnish, Fuseki) is run in its own dedicated container, making the deployment of the services easy due to installation of software dependencies in isolated environments, enhancing the portability of the services.

Currently, we use as an underlying technical infrastructure a combination of the OpenShift container cloud[41] (MMM Portal, Varnish) and virtual machines on the OpenStack cloud platform[42] (Fuseki), provided by the CSC – IT Center for Science, Finland. By using containers, the services can be migrated to another computing environment in a straightforward way, and third parties can re-use and run the services on their own. The container architecture also allows for horizontal scaling for high availability, by starting new container replicas based on demand.

[33] https://nodejs.org/en/.
[34] https://expressjs.com.
[35] https://reactjs.org.
[36] https://redux.js.org.
[37] https://jena.apache.org/documentation/fuseki2/.
[38] https://github.com/mapping-manuscript-migrations/mmm-fuseki.
[39] https://varnish-cache.org.
[40] https://www.docker.com.
[41] https://www.openshift.com.
[42] https://www.openstack.org.

## 6 Discussion

**Evaluation.** Usability of the MMM Portal based on the Sampo model has been evaluated by researchers in manuscript studies in [5]. The overall conclusion of the evaluation report was that “the MMM portal is an excellent tool, and very easy to use”. However, the testers also made several suggestions for further development related to usability and noted that it is not easy to differentiate the challenges between the quality of the underlying data and portal design. According to [3] the evaluation showed that the portal performed significantly better than the original current interfaces and was capable of fully answering most of the original 25 research questions about manuscript history and provenance set in the beginning of the project. Also using the MMM Linked Data service has been deemed useful as discussed in [2]. The manuscript researchers now have a flexible way to access their enriched data and, for example, the researchers at the Schoenberg Institute started to arrange weekly “SPARQL Wednesdays” for learning more about the technology and the data. The ability to find interesting knowledge from the MMM Portal has been noted also by R. Engels in [7].

Thus far the MMM Portal has been used by 8400 distinct users from Hong Kong (18 %), US (17 %), UK (9 %), France (7 %), Italy (6 %), and from other countries (131 countries in total), according to Google Analytics.

**Related Work.** There are various online resources for studying manuscripts, in addition to the databases of our research, such as *e-codices – Virtual Manuscript Library of Switzerland*[43], *vHMML*[44] initiative of the Hill Museum & Manuscript Library, *METAscripta*[45], *Biblissima*[46] [8], and *Digital Scriptorium*[47]. These aggregate manuscript information from multiple sources and make the information accessible from a single user interface. Metadata about manuscripts is harmonized to some extent, for search purposes, but the provenance metadata is shallow or it doesn’t exist. Instead of metadata, many of these systems focus on delivering high quality images of manuscripts to manuscript scholars and other interested users. The *Digitized Medieval Manuscripts*[48] project is producing a map of manuscript repositories around the world.

Challenges in connecting data from manuscript collections are discussed in [1] along with an overview of existing quantitative research on aggregated manuscript data. There are some existing Semantic Web approaches for harmonizing manuscript collections, of which most are based on CIDOC CRM and FRBRoo. Modeling rare and unique documents like manuscripts using CIDOC CRM and FRBRoo has been studied in [20], and we have used the insights of the study to guide the modeling work. Zhitomirsky et al. [33] have modelled a catalog of post-medieval Hebrew manuscripts as Linked Data using, e.g. FRBRoo, and provided a decomposition analysis of the data, and built prototype user interfaces for the data.

[43] http://www.e-codices.unifr.ch/en.

[44] https://www.vhmml.org.

[45] https://metascripta.org.

[46] https://biblissima.fr.

[47] https://digital-scriptorium.org.

[48] https://digitizedmedievalmanuscripts.org.

The ideas behind the Sampo model have been explored and developed before in different contexts. For example, the notion of collaborative content creation by data linking is a fundamental idea behind the Linked Open Data Cloud movement[49] and has been developed also in various other settings, e.g., in ResearchSpace[50]. The idea of providing multiple analyses and visualizations to a set of filtered search results has been used in other portals, such as the ePistolarium[51] [26] for epistolary data, and using multiple perspectives have been studied as an approach in decision making [21]. Faceted search [10,30,31], also know as "view-based search" and "dynamic ontologies", is a well-known paradigm for explorative search and browsing [22] in computer science and information retrieval, based on S. R. Ranganathan's original ideas of faceted classification in Library Science in the 1930's. The two step usage model is used in prosopographical research [32] (without the faceted search component). The novelty of the Sampo model lies in combining several ideas and operationalizing them for developing applications in Digital Humanities, and for delivering the solutions related to user interfaces for re-use in the open source Sampo-UI framework [17].

**Lessons Learned.** The premodern manuscript data turned out in many ways more challenging from a data modeling and technical perspectives than expected. Defining the very concept of "the manuscript" itself raised many ontological modeling questions, since manuscripts can be just fragments of a whole, can be separated into parts, copied, annotated, and united to others over time. Also the data from three sources was very heterogeneous and represented both documents and their provenance. A major goal of the MMM project was to map manuscript migrations in spatio-temporal spaces using maps and timelines, but references to locations in many cases are missing, the mentions refer to historical places that may not exist on modern maps or may have changed over hundreds of years of history, and initially many placenames mentioned were not even geocoded. The data are often incomplete, uncertain, and imprecise in many ways. The amount of data is also large, hundreds of thousands of records, which set efficiency challenges for the technical solutions.

The project started by creating a list of Digital Humanities research questions relating to manuscript histories, and continued by trying to figure out what kind of data model and data are needed to solve them. In spite of the challenges related to the data, the Linked Open Data approach and Sampo model turned out to be successful in the helping the researchers in solving their research question, and managed in our mind to set a new norm for the state-of-the-art for supporting DH research in manuscript studies for further research.

**Acknowledgements.** Thanks to Antoine Brix, Doug Emery, Mitch Fraas, Benjamin Heller, David Lewis, Synnøve Myking, Kevin Page, Pierre-Louis Pinault, Guillaume Porte, Emma Thomson, Athanasios Velios, Pip Willcox and other members of the project team for discussions and collaboration. Our work was funded by the Trans-Atlantic Platform under its Digging into Data Challenge https://diggingintodata.org for 2017–2019. The project was led by the University of Oxford, in partnership with the University of Pennsylvania, Aalto University and

[49] https://lod-cloud.net.

[50] https://www.researchspace.org.

[51] http://ckcc.huygens.knaw.nl.

Helsinki Centre for Digital Humanities (HELDIG) at the University of Helsinki, and the Institut de recherche et d'histoire des textes (IRHT). The technical development presented in this paper was funded mainly by the Academy of Finland. CSC – IT Center for Science provided computational resources for the project.

## References

1. Burrows, T.: Connecting medieval and renaissance manuscript collections. Open Libr. Humanit. **4**(2) (2018). https://doi.org/10.16995/olh.269
2. Burrows, T., Cleaver, L., Emery, D., Koho, M., Ransom, L., Thomson, E.: Using SPARQL to investigate the research potential of an aggregated linked open data dataset: the Mapping Manuscript Migrations project. DH Benelux (2021)
3. Burrows, T., et al.: Mapping Manuscript Migrations: Digging into data for researching the history and provenance of medieval and renaissance manuscripts (white paper), August 2020. https://diggingintodata.org/file/1281/download?token=x59u8fFQ
4. Burrows, T., Hyvönen, E., Ransom, L., Wijsman, H.: Mapping Manuscript Migrations. Digging into data for the history and provenance of medieval and renaissance manuscripts. Manuscript Studies J. Schoenberg Institute Manuscript Studies **3**(1), 249–252 (2018)
5. Burrows, T., Pinto, N.B., Cazals, M., Gaudin, A., Wijsman, H.: Evaluating a semantic portal for the "Mapping Manuscript Migrations" project. DigItalia **2**, 178–185 (2020). http://digitalia.sbn.it/article/view/2643
6. Clemens, R., Graham, T.: Introduction to Manuscript Studies. Cornell University Press, Ithaca (2007)
7. Engels, R.: Digital scholarship and medieval manuscripts: access, technologies and potential. In: Payer, B.A., Wall, A. (eds.) Illuminating Life: Manuscript Pages of the Middle Ages. The University of Guelph (2020)
8. Frunzeanu, E., Robineau, R., MacDonald, E.: Biblissima's choices of tools and methodology for interoperability purposes = Biblissima: selección de herramientas y de metodología para fomentar la interoperabilidad. CIAN-Revista de Historia de las Universidades **19**, 115–132 (2016). https://doi.org/10.20318/cian.2016.3146
9. Gardiner, E., Musto, R.G.: The Digital Humanities: A Primer for Students and Scholars. Cambridge University Press, New York (2015). https://doi.org/10.1017/CBO9781139003865
10. Hearst, M.: Design recommendations for hierarchical faceted search interfaces. In: ACM SIGIR Workshop on Faceted Search, Seattle, WA, pp. 1–5 (2006)
11. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space, 1st edn. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool (2011). http://linkeddatabook.com/editions/1.0/
12. Hyvönen, E.: Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, Palo Alto (2012)
13. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: a 7-star model and platform for publishing and re-using linked datasets. In: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8798, pp. 226–230. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11955-7_24
14. Hyvönen, E.: Using the semantic web in digital humanities: shift from data publishing to data-analysis and serendipitous knowledge discovery. Semantic Web **11**(1), 187–193 (2020). https://doi.org/10.3233/SW-190386
15. Hyvönen, E.: Digital humanities on the Semantic Web: Sampo model and portal series (2021, submitted). https://seco.cs.aalto.fi/publications/2021/hyvonen-sampo-model-2021.pdf

16. Hyvönen, E., Tuominen, J., Kauppinen, T., Väätäinen, J.: Representing and utilizing changing historical places as an ontology time series. In: Ashish, N., Sheth, A. (eds.) Geospatial Semantics and Semantic Web: Foundations, Algorithms, and Applications, pp. 1–25. Springer, Boston (2011). https://doi.org/10.1007/978-1-4419-9446-2_1
17. Ikkala, E., Hyvönen, E., Rantala, H., Koho, M.: Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. Semantic Web - Interoperability, Usability, Applicability (2021, in press). http://www.semantic-web-journal.net/
18. Koho, M., et al.: Harmonizing and publishing heterogeneous pre-modern manuscript metadata as linked open data. J. Assoc. Inf. Sci. Technol. (JASIST) 1–18 (2021). https://doi.org/10.1002/asi.24499
19. Koho, M., et al.: Mapping Manuscript Migrations knowledge graph (2021). https://doi.org/10.5281/zenodo.4440464
20. Le Boef, P.: Modeling rare and unique documents: using FRBRoo/CIDOC CRM. J. Arch. Organ. **10**(2), 96–106 (2012). https://doi.org/10.1080/15332748.2012.709164
21. Linstone, H.A.: Multiple perspectives: concept, applications, and user guidelines. Syst. Pract. **2**(3), 307–331 (1989). https://doi.org/10.1007/BF01059977
22. Marchionini, G.: Exploratory search: from finding to understanding. Commun. ACM **49**(4), 41–46 (2006). https://doi.org/10.1145/1121949.1121979
23. McCarty, W.: Humanities Computing. Palgrave, London (2005)
24. Moretti, F.: Distant Reading. Verso Books (2013)
25. Peroni, S., Shotton, D., Vitali, F.: The live OWL documentation environment: a tool for the automatic generation of ontology documentation. In: ten Teije, A., et al. (eds.) EKAW 2012. LNCS (LNAI), vol. 7603, pp. 398–412. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33876-2_35
26. Ravenek, W., van den Heuvel, C., Gerritsen, G.: The ePistolarium: Origins and techniques. In: van Hessen, A., Odijk, J. (eds.) CLARIN in the Low Countries, pp. 317–323. Ubiquity Press (2017). 10.5334/bbi
27. Riva, P., Doerr, M., Žumer, M.: FRBRoo: enabling a common view of information from memory institutions. Int. Cataloguing Bibliographic Control **38**(2), 30–34 (2009)
28. Shultz, K.: What is distant reading? New York Times, 24 June 2011
29. Southall, H., Mostern, R., Berman, M.L.: On historical gazetteers. Int. J. Humanit. Arts Comput. **5**(2), 127–145 (2011)
30. Tunkelang, D.: Faceted search. Synth. Lect. Inf. Concepts Retr. Serv. **1**(1), 1–80 (2009)
31. Tzitzikas, Y., Manolis, N., Papadakos, P.: Faceted exploration of RDF/S datasets: a survey. J. Intell. Inf. Syst. **48**(2), 329–364 (2017)
32. Verboven, K., Carlier, M., Dumolyn, J.: A short manual to the art of prosopography. In: Prosopography Approaches and Applications. A handbook, pp. 35–70. Unit for Prosopographical Research (Linacre College) (2007). 1854/8212
33. Zhitomirsky-Geffet, M., Prebor, G.: Toward an ontopedia for historical Hebrew manuscripts. Front. Digital Humanit. **3**, 3 (2016). https://doi.org/10.3389/fdigh.2016.00003