Wikibase Model for Premodern Manuscript Metadata Harmonization, Linked Data Integration, and Discovery

MIKKO KOHO, Semantic Computing Research Group, Aalto University, Finland
L. P. COLADANGELO, College of Communication and Information, Kent State University, USA
LYNN RANSOM and DOUG EMERY, Schoenberg Institute for Manuscript Studies, University of Pennsylvania, USA

To facilitate discovery of premodern manuscripts in U.S. memory institutions, Digital Scriptorium, a growing consortium of over 35 institutional members representing American libraries, museums, and other cultural heritage institutions, has developed a digital platform for an online national union catalog. The platform will allow low-barrier and efficient collection, aggregation, and enrichment of member metadata and sustainably publish it as Linked Open Data. This article describes the methods and principles behind the data model development and the decision to use Wikibase. The results of the prototype implementation and testing phase demonstrate the practicality and sustainability of Digital Scriptorium's approach to building an online national union catalog based on Linked Open Data technologies and practices.

CCS Concepts: • **Applied computing** → **Arts and humanities;** • **Human-centered computing** → *Interactive systems and tools;* • **Information systems** → **Network data models**; *Resource Description Framework (RDF)*; Digital libraries and archives;

Additional Key Words and Phrases: Linked Open Data, data modeling, Wikibase, data interoperability, premodern manuscripts, digital humanities, cultural heritage, semantic web

ACM Reference format:
Mikko Koho, L. P. Coladangelo, Lynn Ransom, and Doug Emery. 2023. Wikibase Model for Premodern Manuscript Metadata Harmonization, Linked Data Integration, and Discovery. *ACM J. Comput. Cult. Herit.* 16, 3, Article 56 (August 2023), 25 pages.
https://doi.org/10.1145/3594723

1 INTRODUCTION

For historians of medieval and early modern cultures who rely on manuscript evidence as their primary and even secondary sources of information, the problems associated with directing researchers to these sources have long plagued librarians and curators responsible for such collections. As is commonly understood, premodern manuscripts, that is, handwritten volumes produced before, and in many cases, co-existing in relation to print culture, are by nature unique objects. Each one bears its own story and associated cast of characters: authors, translators, compilers, scribes, artists, former owners, binders, and so on. Unlike printed books that are produced identically en masse for a more or less unknown readership, each premodern manuscript is made for a specific

Authors' addresses: M. Koho, Department of Computer Science, P.O. Box 15400, FI-00076 Aalto, Finland; email: mikko.koho@aalto.fi; L. P. Coladangelo, College of Communication and Information, Kent State University, 800 E. Summit St., Kent, OH 44242; email: lcoladan@kent.edu; L. Ransom, Schoenberg Institute for Manuscript Studies, University of Pennsylvania Libraries, 3420 Walnut St, Philadelphia, PA 19104; email: lransom@upenn.edu; D. Emery, University of Pennsylvania Libraries, 3420 Walnut St, Philadelphia PA 19104; email: emeryr@upenn.edu.

1556-4673/2023/08-ART56
https://doi.org/10.1145/3594723

and known reader or group of readers with different needs and expectations of what that manuscript should contain or look like. Copied texts will always differ from the original source either through intentional edits or unintended mistakes in copying, that then get repeated in successive copies so that copies several generations removed may differ quite dramatically from their source, yet still be considered to be the same work. Multiple texts bound together in one volume will rarely appear again together or in the same order as another volume. Copies of the same work can range from lavishly illuminated works of art intended for aristocratic audiences to working copies made by nonprofessional scribes for personal use. Further, the lives these books led after production can reveal much about changing patterns of readership and reception over time. For the modern scholar, each unique feature of a manuscript can provide further insight into why, when, where, and for whom the texts and images contained in each manuscript were produced, read, studied, and copied. Thus, every manuscript as a unique witness to a given work has the potential to reveal new insights into our shared cultural and intellectual heritages.

For premodern manuscript research, one cannot assume that one copy of a text reflects the same historical and functional context as another copy. Findability–or discoverability–of multiple witnesses of the same work or expression of a work (such as a translation or rescension of a work) is essential to this research and depends on published manuscript descriptions. Prior to the digital age, manuscript research was primarily conducted using print catalogs of collections and institutions. Methods of catalog description could range in scope from simple finding aids of one-line descriptions typical of early modern cataloging practices, noting only the title, author, and date and place of production, to more modern complex and detailed descriptions of the content and physical features of a manuscript, such as the type of script, the page layout, dimensions, binding structures and paratextual elements such as glosses and other forms of marginalia, as well as detailed provenance histories. With each feature being the equivalent of a data point, approaches to the data modeling underlying traditional manuscript description in print catalogs are as variable as the number of institutions producing the catalogs. As premodern manuscript description has moved into digital form over the last decades, these circumstances have not changed, and the lack of one guiding standard for publishing manuscript metadata that can be agreed upon across institutions continues to make discoverability a challenge.

There have been many efforts in the last twenty years to find practical solutions for aggregated searching across institutions and across platforms in the cultural heritage sector [2, 5, 22, 23, 34]. Among these is **Digital Scriptorium** (**DS**; https://digital-scriptorium.org/), a growing consortium of over 35 institutional members representing American libraries, museums, and other cultural heritage institutions. Started in 1997 as a joint project between the Bancroft Library at the University of California, Berkeley, and Columbia University, DS's primary mission has been to digitize, describe, and publish metadata related to western medieval and Renaissance manuscript codices and fragments in American collections via its digital platform, ultimately becoming an online national union catalog to facilitate research and discovery across collections. Since its inception, DS has been supported by annual member dues set by a weighted fee schedule to ensure that institutions at all levels can participate.[1]

In its first decade, DS achieved early success with an innovative relational model first expressed in an Access database, later converted to a WebGenDB content management platform [18] and hosted by the University of California, Berkeley. As of 2019, DS had cataloged 8,352 manuscripts among its 35 member institutions. However, challenges associated with both a heavy, complicated, and increasingly outdated technical infrastructure and workflow difficulties in uploading member institutions' data threatened DS's existence. In 2017, the University of California, Berkeley, made the decision to discontinue support for WebGenDB forcing the consortium to undertake a radical redevelopment and reconceptualization of its technical platform. The new platform would need to mitigate the previously mentioned high-barrier challenges to participation and be prepared to scale up to accommodate as yet uncataloged manuscripts and manuscripts from other manuscript traditions outside of

[1]Member institutions that fall below a calculated threshold of resources can join for free as non-voting members: https://digital-scriptorium.org/about/membership-information/.

Europe, including the Arabic, Jewish, Asian, African, and American traditions. The estimated number of uncatalogued manuscript objects in U.S. collections, including codices, rolls and fragments, has been determined to be as high as 100,000 [11]. The new platform would also need to respond to an evolving digital landscape that provided opportunities for aggregating, sharing, and linking data that were not readily available when DS was first formed while maintaining a long-term sustainability plan.

Known as DS 2.0, the redevelopment project has resulted in a digital platform based on ***Linked Open Data (LOD)*** practices and technologies that can efficiently collect, aggregate, and enrich member metadata and sustainably publish it as LOD.[2] DS 2.0 uses *Wikibase*, a free, open-source, and community-maintained software suite for building knowledge bases that can be published as LOD.[3] For structuring linked data, Wikibase is based on subject-property-value statements, which are built on an RDF semantic triple framework. Wikibase also drives Wikidata [36], the world's largest, openly accessible knowledge graph.[4] Wikimedia Deutschland has recently launched Wikibase.cloud, a cloud-based service platform that hosts instances of Wikibase, providing a free hosting service for organizations that do not otherwise have the technical support and infrastructure to support their own instance.[5] These features made Wikibase the obvious solution as the underlying technical platform for the DS 2.0 project.

DS 2.0 provides a sustainable infrastructure to build an online national union catalog. A project at this scale faces a number of challenges that are endemic to cultural heritage institutions: inconsistent metadata standards for describing cultural heritage objects, unevenly resourced institutions in terms of funding and expertise, and simple organizational challenges relating to costs, staffing, and workflow. To meet these challenges, the project team developed a set of principles at the outset that formed the foundation of both the data model and a series of straightforward workflows that aim to minimize the work burden placed upon DS member institutions and lower barriers to participation. The choice to use Wikibase aligned with these principles but also presented further challenges to the design of the data model. In the following, we will describe how these principles guided the design of the new data model and led us to choosing Wikibase as the most sustainable and practical solution for data collection and publication, despite certain limitations, to meet the challenges of building an online national union catalog in a LOD environment.

This article is organized as follows. The next Section presents the outset of the data modeling work for aggregating manuscript metadata and the design principles guiding the work. Section 3 discusses the choice and use of Wikibase as the technical platform. Modelling manuscript metadata on a general level is discussed in Section 4 and the devised DS 2.0 data model is presented in detail in Section 5. Section 6 gives an overview of the prototype Wikibase system and an evaluation of the system in relation to the data model and data aggregation is given in Section 7. Finally, the conclusion is provided in Section 8.

2 DESIGN PRINCIPLES UNDERLYING THE DS 2.0 DATA MODEL

Data modeling in manuscript cataloging projects, whether digital or analog, tends to prioritize original and expert bibliographic description to create and collect information about a manuscript object. The original Digital Scriptorium [18] is an example of such modeling. In this model, the catalog record ideally would present the most complete information known about a manuscript based on the expertise and knowledge of a cataloger, often relying on discursive, unstructured scholarly analysis to explain, for example, an argument for the date or place of a production if internal evidence did not exist. For a national union catalog, this model produces practical inefficiencies in terms of cost and workflow that made such an effort difficult to sustain.

[2]The DS 2.0 redevelopment project was supported by a National Leadership for Libraries Planning Grant funded by the Institute for Museum and Library Services (LGLG-246396-OLS-20), awarded in 2020.
[3]Wikibase/What is Wikibase: https://www.mediawiki.org/wiki/Wikibase/What_is_Wikibase.
[4]Wikidata.org: https://www.wikidata.org/wiki/Wikidata:Main_Page.
[5]The service is currently in beta and is invitation only. For information and updates, see https://www.wikibase.cloud/.

The history of Digital Scriptorium briefly outlined in the Introduction provides some context for the DS 2.0 project team's approach to designing DS 2.0. While Berkeley's decision to discontinue support of the outdated software underlying the previous platform may have been the catalyst for redevelopment, other factors were just as pressing. Many member institutions found it difficult to produce and maintain current metadata in DS. The requirements and expectations for DS data entry required a high level of expertise and often additional staffing to produce a sufficient level of description. These pressures proved to be significant barriers for many member institutions to create and maintain DS records. DS also required high resolution digital images, which many institutions could not afford to produce.

Further complicating participation efforts was the fact that in the last twenty years, many institutions were able to digitize and publish manuscript images and descriptive metadata in institutional catalogues, typically as MARC- or TEI-based records, or in publicly available repositories, such as the University of Pennsylvania's OPenn repository of open access data (https://openn.library.upenn.edu/), rendering the DS records of those institutions redundant and in most cases outdated. There was little incentive for catalogers to update a DS record after updating a corresponding institutional record as new knowledge about a manuscript became available. For those institutions maintaining their own separate digital catalogs, the DS record was an inferior duplicate record with no connection to, and often in conflict with, the institutional record. As a result, many institutional contributions to DS were, by 2019, completely out of date and/or incomplete.

At this time, the DS Board of Directors, in conversation with its membership, began a process of reviewing and reconceptualizing both the mission of DS and its technical infrastructure. Understanding and acknowledging the needs and limitations of member institutions at the outset was the first step in rethinking the data modeling and workflow processes of DS 2.0. A survey of collection managers at member and prospective member institutions provided insights that informed the data and workflow modeling.[6] The survey received a total thirty-five responses. Responses to the question "What are or would be the benefits in maintaining a DS membership for your institution?" showed that accessibility, visibility, discovery, and exposure were primary concerns; respondents noted "collaborating with colleagues," "finding connections in other collections," "learning from other libraries working on manuscript description and digitization," and "potential assistance in cataloging manuscripts and fragments in languages for which we lack expertise" as particular benefits. Responses to the question "What are or would be the challenges in maintaining a DS membership for your institution?" ranked financial concerns, staff resources, difficulty in updating records, and proprietary concerns as the top challenges. The survey also showed a need for DS to expand its geographic scope beyond only Western European materials as most collections held premodern manuscripts from Arabic, Asian, and/or African cultures as well as from the early Americas. A majority of institutions prioritized the use of LOD technologies, the ability to collaborate with international partners and projects, and an ambivalence about the necessity of hosting images in the DS 2.0 platform. The survey also revealed the many data formats institutions are using to collect and publish manuscript metadata: MARC, EAD, TEI, XML, MARC/XML, PDF, MODS, ContentDM, DACS, JSON, and simple spreadsheets.

The results of the survey highlighted three important needs among participating institutions: (1) lower technical and workflow barriers for data contribution, (2) an ability to work with a diversity of cataloging practices among member institutions, and (3) a greater return on the investment for participation. Based on this needs assessment, the project team developed a set of six principles to guide development to ensure success and circumscribe the project's scope:

- As a national union catalog, DS 2.0's primary function will be to enable researchers to find premodern manuscripts in U.S. collections, including non-European manuscripts.
- DS 2.0 would impose minimal standards for data entry.

[6]Survey results can be found online: https://digital-scriptorium.org/wp-content/uploads/2023/06/DS-Institutional-Survey-Results.pdf.

- Members will manage their own manuscript metadata in their institutional formats. DS 2.0 will take data in the form that members supply and will not correct or add to a member's metadata.
- DS 2.0 will not host images of manuscripts but will provide the ability to view images published following the standards of the International Image Interoperability Framework (IIIF).[7]
- DS 2.0 will transform and enhance member institutions' metadata by reconciling with external authorities and with in-house Name and Manuscript ID Authorities.
- DS 2.0 data will be open access, employ LOD technologies and practices, and be made available for reuse.

Adhering to these principles produced a light and lean data model and workflow and guided the choice of which technical platform to use. As the components of a finding aid, DS 2.0 records do not need full, expert-level descriptions but can function with only minimal amounts of data; links to the more developed institutional records can be supplied in the DS 2.0 record for easy access. Disposing of the image requirement removes a further barrier to entry for many institutions. The only required data for a manuscript is an acknowledgment that an institution owns it. Beyond that, members only need to contribute as much structured descriptive metadata as staffing, expertise, and resources allow, including links to the institution's catalog record; if available, then IIIF manifests can also be linked to the Wikibase record through a IIIF property available in Wikidata's property list. By prioritizing the institutional records as the data source, member institutions only need to maintain one record, a copy of which they agree to contribute to DS for reconciling in the DS 2.0 data model, to ensure consistency in representation with DS's record. With its primary responsibility to reconcile and harmonize member data with external and in-house authorities, DS 2.0's focus rests solely on increasing the value of membership by creating a distinct but connected, well-structured, aggregated, and semantically enriched dataset to be made available to the world on an open access platform for discovery and reuse and to facilitate external collaboration [9, 31, 38].

3 THE WIKIBASE SOLUTION

The design principles also guided the decision to choose Wikibase as the technical platform. Wikibase contains an extension for displaying IIIF images managed by member institutions, relieving DS of the costly burden of hosting images. As a shared platform with Wikidata, Wikibase instances can easily participate in the larger Wikimedia environment. Wikibase is increasingly being used by GLAM institutions and projects around the world as a low-cost, highly sustainable solution for creating knowledge bases in collaborative environments and has been noted for its facility in handling bibliographic data [21]. Further, Wikibase has recently implemented a cloud-based platform, Wikibase.cloud, that frees institutions from having to find and fund a local hosting service.[8] There are, of course, potential risks associated with using the Wikibase.cloud, such as the loss of direct control over the decision-making process in the technical development or life-cycle of the platform. A community-maintained resource like the Wikibase.cloud, in which multiple people and projects have a vested interest in supporting the infrastructure, is less likely to collapse than a self-hosted instance that depends on the variable needs and resources of an individual person or institution for continued support and development.

Overall a Wikibase.cloud instance aligns with DS 2.0 design principles in that it is a low-cost, low-barrier solution to creating and managing the DS technical platform and creates free, open access, semantically enriched data out of DS member data. While Wikibase has its limitations, as discussed below and by Bergamin [8], it nevertheless succeeds as a solution for projects in the chronically underfunded and under-resourced GLAM sector.

For projects seeking a low-cost solution to publish data as LOD, Wikibase provides an easy, high-return, out-of-the-box solution. The RDF graph that makes up a Wikibase instance includes items (which act as subjects and values) and properties, which act as predicates of a statement. Items are identified by a prefix "Q" followed by a string of numbers unique to the item; properties are similarly identified by a unique string of numbers preceded

[7] International Image Interoperability Framework (IIIF): https://iiif.io/.
[8] https://www.mediawiki.org/wiki/Wikibase/Wikibase.cloud.

by the letter "P." Items and properties are given both a preferred label and alternate labels (aliases) as well as a brief description. Both items and properties have their own pages in the Wikibase, connecting items to items through properties and connecting properties to related items, resulting in a linked data structure.

In a Wikibase, items can act as classes or individuals. In the DS 2.0 Wikibase, individuals of a class are identified through statements in which an item is an instance of (P16) a particular type of item, such as an individual place being an instance of a Place Authority File (Q16). Wikibase properties define the relationships between a Wikibase item as the subject of a statement and values that describe particular aspects of or information related to an item. For example, an instance of a DS 2.0 Record (the subject) representing a manuscript description can include a statement about information regarding the place of production (P27, "place of production as recorded") for a manuscript that is represented by a string value (e.g., "Germany") extracted from the metadata record for the described manuscript provided by the member institution. Values of properties may be of various data types, but valid data types for a property are defined by corresponding property constraints.

A special kind of element found in the Wikibase data infrastructure is a qualifier, of which DS 2.0 makes significant use. A qualifier is any refinement, annotation, or contextualized information about a statement that allows expression of more granular structured data than is possible with a property-value pair [36]. To extend the above example through its use in DS 2.0, a qualifier property (P28, "place in authority file") can be used to annotate the string value "Germany" in the place of production statement to link that unstructured string value to the individual Wikibase item Germany found in the DS 2.0 Place Authority. The Wikibase item for Germany is in turn linked to an external authority, creating a linked data connection between the string values found in manuscript records and corresponding structured values found in LOD vocabularies, thus semantically enriching values recorded in a DS 2.0 Record. Where multiple entities are present in a single string value, such as multiple place entities in a production place as recorded, the string value is qualified by each individual entity contained in the string through each corresponding authority file (using a qualifier property for an entity in an authority file). For example, the entire string "Amberg; Bavaria; Germany" is annotated/qualified as containing (and is therefore linked to) individual Place Authority Files for Amberg, Bavaria, and Germany. This allows us to represent the recorded value in its original and unaltered form while semantically annotating it and enriching it with related structured data.

For the reasons outlined above, and because its open access platform allows collaboration across institutions and organizations, a number of projects in the cultural heritage sector have turned to Wikibase to publish their structured, machine-readable data.[9] For many projects, such as *Enslaved: Peoples of the Historical Slave Trade*[10] [30] and the Black Bibliography Project,[11] Wikibase provides a collaborative space to collect and publish data relevant to a specific topic. Several institutions employ Wikibase to publish authority files, including the German National Library, which has pioneered the use of Wikibase for authority files [20], as well as the French consortium Biblissima+,[12] a multi-site digital infrastructure for researching the history of ancient and medieval texts in manuscript and printed book forms.

These projects provided DS 2.0 with models for collaboration-based data collection and authority file management in Wikibase. To our knowledge, DS 2.0 is the first project to use Wikibase to create a meta-catalog, that is a catalog that seeks to unite disparate catalogs in one aggregated environment, for unique cultural heritage objects without requiring member institutions to change their cataloging practices or technologies [35]. DS institutions contribute any structured metadata available to DS, which is then cross-walked into an "agnostic" spreadsheet for reconciliation and semantic enrichment using OpenRefine. The resulting dataset can then be transferred into the Wikibase. One further benefit of this process is that each manuscript receives a persistent

[9]Wikimedia Deutschland Showcase: https://wikiba.se/showcase/.
[10]Enslaved: Peoples of the Historical Slave Trade (Enslaved.org): https://enslaved.org/.
[11]Black Bibliography Project: https://blackbibliog.org/.
[12]Biblissima+: https://projet.biblissima.fr/en.

alphanumeric DS unique resource identifier (DS1, DS2, DS3, etc.) that can then be both returned to the holding member institution and contributed to Wikidata, thus creating a semantic bridge between the holding institution and Wikidata to connect a global community of researchers to each DS member institution. Once in Wikidata, DS manuscript records can potentially be linked to other similar collections already in Wikidata such as the Peniarth Manuscripts of the National Library of Wales [19]. The DS identifier then becomes a powerful tool and a model for a national system to create persistent unique identifiers for premodern manuscripts.

4 MODELLING PREMODERN MANUSCRIPT DESCRIPTIONS

As noted by Bair and Steuer [3], modeling manuscript metadata is challenged by many factors. Unlike printed books, every manuscript is a unique production and requires object-specific description elements related to both the textual contents—author, title, language—and the physical properties–date and place of production, script, materials, dimensions for both the object and page layout, decoration, folio count, binding, and so on. Also critical to manuscript description, though not specific to it, is provenance information that often cannot be documented beyond the manuscript itself–an inscription on a flyleaf, for example, noting possession by an otherwise unknown owner. Unlike printed books that typically come with shared and specific descriptive metadata across copies usually found within the pages of the book–publisher, editor, printer, date and place of publication, and so on—knowledge about the circumstances of a manuscript's production is often speculative and debatable. For example, the ability to date or localize a manuscript may be based on a comparison of its script or decoration to other manuscripts known to have been produced around certain times in certain places. Data elements are often ambiguous and fuzzy: "Northern France," "Persia," or "early fourteenth century," or "after 1400." Further complicating manuscript metadata modeling practices is the fact that there are no uniformly agreed upon standards for manuscript description, whether for print or digital cataloging systems. How an institution decides to catalog its manuscript collections–if it can invest in this work at all–often depends on the subject expertise of staff, the capacity for staff to dedicate time to manuscript description, and existing cataloging formats designed for printed books. As a result there are almost as many different ways to model manuscript metadata as there are institutions with manuscript collections to describe.

Demand for aggregated research resources, however, is driving research to define a more adaptable approach to data modeling to enable data-sharing, especially in LOD environments [6, 7, 26]. A major barrier to the development of aggregated research resources has been finding one solution to data modeling that will be accepted by all stakeholders, but a number of projects working cross-institutionally are implementing more simplified approaches to modeling based on structured data values common to all manuscript traditions that can easily be translated into, if not created as, LOD. For instance, the Al-Furqan Digital Library,[13] Biblissima,[14] e-codices,[15] Firhist,[16] Handschriftenportal,[17] ManusOnline,[18] Medieval Manuscripts in Dutch Collections (MMDC),[19] Medieval Manuscripts in Flemish Collections (MMFC),[20] and the Southeast Asia Digital Library[21] are developing platforms to provide a single access point for manuscripts cataloged and housed at institutions across national or international institutions. Underlying the platforms are standard data points for basic manuscript description also used to model DS data (e.g., author, title, artist, scribe, place and date of production, provenance history, subject, genre).

[13] Al-Furqan Digital Library: https://digitallibrary.al-furqan.com/world_library.
[14] Biblissima: https://portail.biblissima.fr/.
[15] e-codices: https://www.e-codices.unifr.ch/.
[16] Firhist: https://www.fihrist.org.uk/.
[17] Handschriftenportal: https://handschriftenportal.de/.
[18] ManusOnline: https://manus.iccu.sbn.it/.
[19] Medieval Manuscripts in Dutch Collections: http://www.mmdc.nl.
[20] Medieval Manuscripts in Flemish Collections: https://mmfc.be/.
[21] Southeast Asia Digital Library: https://sea.lib.niu.edu/.

The application of semantic technologies to aggregating manuscript metadata has been further explored in projects such as Europeana, one of the first and most successful efforts to aggregate data across cultural heritage institutions using LOD practices and technologies, and the recently completed **Mapping Manuscript Migrations (MMM)** project [23, 25]. Both of these projects modeled how content-related but distinctly modeled datasets could be semantically linked based on a simplified structure of common data elements. The result for each was a new knowledge graph that allowed researchers to query multiple datasets without modifications having to be made to the original models.

There are two major approaches to modeling cultural heritage metadata as linked data [15, 17]:

- Object-centric data models, such as the *DCMI Metadata Terms*,[22] focus on describing individual resources (e.g., cultural heritage objects) of interest with a set of shared metadata description. This approach has been used in many manuscript databases, including the original Digital Scriptorium and Europeana.
- Event-centric data models, such as *CIDOC Conceptual Reference Model (CRM)* [16], attempt to model the events relating to the history of the objects of interest, such as their creation and possible changes of ownership, and so on. This approach has been used in, e.g., MMM, the IMAGO project [4], and could also be used with Europeana in addition to the object-centric model [27]. These also make use of the CIDOC CRM compatible versions of ***Functional Requirements for Bibliographic Records (FRBR)*** and ***IFLA Library Reference Model (LRM)*** [37], FRBRoo and LRMoo, respectively.

The metadata stored in cultural heritage institutions, e.g., in collection management systems or a library catalog systems, is typically object-centric, requiring much less effort to convert to another object-centric data model than an event-centric one [15].

5 DS 2.0 DATA MODEL

Guided by the design principles outlined in Section 2, the DS 2.0 data model responds to the workflow problem of collecting and sharing member data from different sources by providing a shared data infrastructure within which each member's data can be aligned. The scope of the data model is defined by what is commonly available in member institutions' manuscript metadata and what is seen as relevant for interested parties to find manuscripts. Discoverability of manuscripts based on structured metadata is the main goal in this process. Europeana's and MMM's approach to respecting the original format of the source data and transforming the salient data points into RDF to build the aggregated knowledge graph thus set the example for the DS 2.0 data and workflow modeling.

The data model creation commenced in two stages: first, as the creation of a more generic linked data-based data model for manuscript metadata in the scope of DS and then, as the choice of Wikibase as the platform was confirmed and the prototype development started, the data model was adapted to be used with the underlying data model of Wikibase and developed further.

In both stages the data model was devised in an iterative process with an interdisciplinary team with expertise in several key areas: linked data, manuscript metadata practices, DS source data, the DS organizations and their capabilities and resources, and data models for cultural heritage.

The DS 2.0 data model consisting of classes and properties is shown in Figure 1. The colored rectangles correspond to classes of items, whereas rectangles with white background are primitive Wikibase data types. The green color rectangles are the main manuscript items, of which *Manuscript* corresponds to the physical manuscript item and *DS 2.0 Record* depicts a metadata record about the manuscript. The yellow classes refer to classes of which instances are shared between several manuscript records. Instances of the blue-colored classes are resources in external authority files, for which local instances are created. The solid directed arrows correspond to properties used for items and dashed arrows mark properties that are used as qualifiers in statements.

[22]DCMI Metadata Terms: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

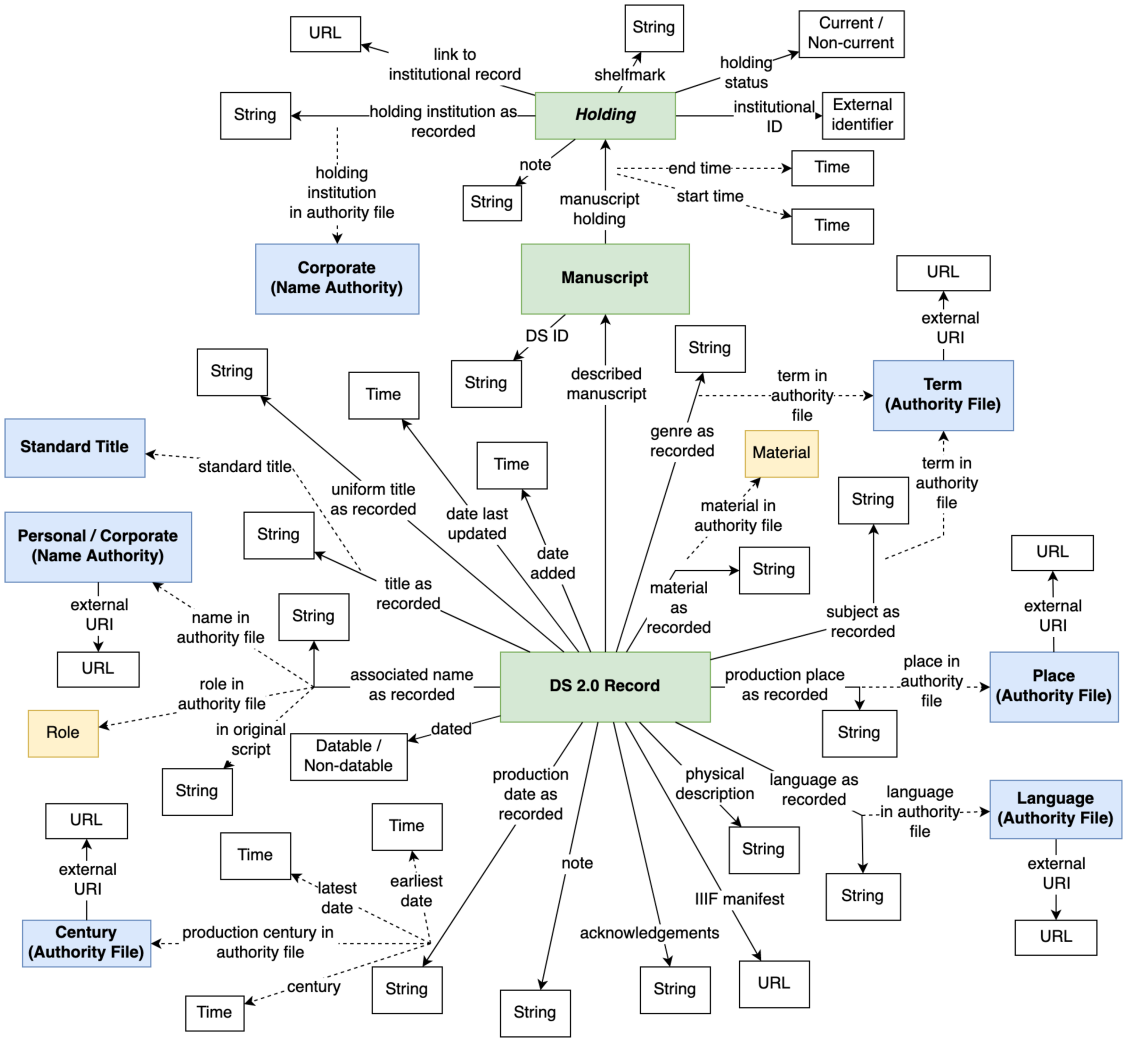


Fig. 1. DS 2.0 wikibase data model as a diagram, showing the classes in green, blue, and yellow rectangles and wikibase data types and items in white rectangles. Properties used in statements about the class instances are shown as arrows, with dashed arrows from the statements represent the use of qualifiers.

The separation between the physical manuscript and its metadata record(s) is conceptually similar to that used in, e.g., Europeana Data Model [17]. This is different from FRBRoo and the MMM data model [25], in which the manuscript and its metadata description are not separate entities. Contrary to the model in FRBRoo, IFLA LRM, or BIBFrame, the DS 2.0 data model does not model the intellectual contents of the manuscripts as separate entities.

The DS 2.0 data model is object-centric, with the main manuscript metadata record (class *DS 2.0 Record* as the focal point, to which most of the metadata is attached to in a star-like manner. This is conceptually different from, e.g., the MMM data model, which is event-based and aims to capture a variety of information not directly related to the manuscript, such as the provenance-related manuscript observations in an attempt to depict a history of the manuscript from creation to the most recent observation.

For the *DS 2.0 Record* all immediate values originate from the source metadata record whereas most of the qualifiers include an interpretation of the contents of an original string value that link the statement to an authority file item.

The *Holding* item depicts the manuscript holding by a DS member institution. In the rare case that a manuscript would be moved from one institution to another, a new *Holding* instance would be created for the new holding and the old one marked as non-current.

The data model has a class *Standard Title* (Q6) for conventional, non-authoritative titles like *Book of hours*, *Metaphysics*, and *De rerum natura*. Standard Title is primarily a functional category for use in SPARQL[23] queries and in our search interface as facets. While users have a clear understanding of these titles, titles as recorded in sources vary widely. Searching for all entities with a particular title requires a user both to know and then search by *all* possible variations of the title as recorded by using alternate versions of the title, the title in alternate languages, in alternate spellings, or with different letter cases, which requires special handling in SPARQL queries, because the language is case-sensitive. For example, books of hours can have the supplied titles *Book of Hours*, *Book of hours* ("hours" with lower case "h"), *Horae*, and *Book of Hours (Use of Rome)*, along with many other variations. Neither Uniform Title authorities nor Topical and Form term authorities offer appropriate terms that address this problem. Title authorities like Faceted Application of Subject Terminology (FAST)[24] have titles that are often too specific; for example, *Book of hours (Catholic Church)*,[25] *Book of hours (HM 1173)*,[26] and *Hours of Simon de Varie*,[27] Additionally the FAST and Library of Congress Uniform Title designates a work.[28]

The DS data model does not have a concept of work, because DS source records often lack the concept themselves. Manuscripts frequently contain multiple titles by multiple authors, and some sources list titles and authors separately without coordinating them. This is the approach DS takes, listing titles and authors as separate DS 2.0 Record properties. Form and topical terms authorities do not suffice either. They designate either the type or the subject of a work rather than its title; for example, "Books of hours,"[29] rather than the title *Book of Hours*. Further, topical and form authority files lack terms for specific, non-generic titles, like *De Rerum Natura*, *Shāhnāmah*, and *The City of God*. The Standard Titles list will be added to based on "title as recorded" values and, where provided in MARC records, Uniform Titles. A large number of manuscripts in DS will be represented by a relatively small number of titles for common religious and liturgical texts, and popular philosophical and literary works. To partially address the limitations of a lack of representation for works in member data, standard titles and authors can be combined in faceted search or in SPARQL query parameters to find manuscripts that may contain the texts of specific works.

Furthermore, although the previous METS-based schema for DS data offered the ability to model the presence and order of works in parts or folios of the manuscript objects in which they appeared, the process for inputting such detailed metadata proved too cumbersome and inconsistent across contributing institutions to be feasibly applied. This meant that data quality suffered as institutions could not sustain such intensive cataloging practices to consistently and equally contribute granular data of that type to DS. The expectation that DS records would need to contain that level of granularity also proved unnecessary as the redevelopment of the database envisioned DS as a discovery portal that would assist users toward more detailed institutional records, not as an exhaustive republication of those catalog records.

[23]SPARQL Protocol and RDF Query Language used to query and manipulate RDF graph data. A full specification can be found here: http://www.w3.org/TR/sparql11-overview/.
[24]Faceted Application of Subject Terminology (FAST): https://www.oclc.org/research/areas/data-science/fast.html.
[25]Book of hours (Catholic Church): http://id.worldcat.org/fast/1365973/.
[26]Book of hours (HM 1173): http://id.worldcat.org/fast/1374414/.
[27]Hours of Simon de Varie: http://id.worldcat.org/fast/1381061/.
[28]See the MARC 21 documentation for Uniform Title, fields 130 (https://www.loc.gov/marc/bibliographic/bd130.html) and 240 (https://www.loc.gov/marc/bibliographic/bd240.html).
[29]"Books of hours" topical term in FAST, http://id.worldcat.org/fast/836484/.

As the sources generally have only textual representations of key metadata entities like actors and places, the general approach is to create a statement for the *DS 2.0 Record* that contains the textual representation as a value with datatype "String." The interpretation of the string value is then linked from this statement to the authority file instance using a qualifier. In some cases further qualifiers are added to depict possible uncertainty in the interpretation as a string value. In the case of actor names related to the manuscript metadata record, how the actor is related to the manuscript (property *role*) is stored.

A ***Digital Scriptorium 2.0 identifier (DS ID)*** is generated for each physical manuscript (class *Manuscript*) during first ingestion of its metadata. The DS ID is intended to serve as a unique identifier for the manuscripts in both DS 2.0 and other systems to identify the manuscripts. Once generated, DS IDs can be included in institutional metadata records to connect the manuscript and its record back to the catalog representation of the object. A long term plan for DS IDs includes contributing these identifiers to Wikidata in an effort to link DS data about the same manuscript objects to other linked data in external knowledge bases.

Interoperability of the data model is currently taken into account by being aware of other relevant data models and ontologies and considering conformance to them on a conceptual level. Semantic mapping to external relevant schemas and ontologies like CIDOC CRM, DCMI Metadata Terms, "**Digitised Manuscripts to Europeana**" **(DM2E)** model [2], Europeana Data Model [17], FRBRoo [28], and MMM data model [25] was initially considered. But as Wikibase does not currently allow for the reuse of external RDF vocabularies [14], this is not yet implemented. Mapping to external vocabularies is currently a planned feature for Wikibase[30] and additionally an extension[31] is currently being developed to this end [29].

The data model is implemented completely on Wikibase, and it does not reuse any Wikidata items or properties. In some cases an existing Wikidata property could have been considered, but the lack of creating semantic mapping to external properties hinders the reuse of also Wikidata properties. It is possible to copy items and properties from Wikidata and create a custom property with data type "external identifier" for marking the connection between DS 2.0 and Wikidata. However, the External identifier datatype is basically a string, which in Wikidata is used for local identifiers, not URIs or IRIs. It is possible to use property specific formatter URLs to format External identifiers as URLs on the item pages, but these are commonly used to create URLs for web pages, not resource IRIs, which for example in the case of Wikidata are different.

5.1 Classes

The classes of the DS 2.0 data model are shown in Table 1, including the class names, short descriptions, and Wikibase QIDs. There are two main classes of the data model: the *DS 2.0 Record* and *Holding*. The *DS 2.0 Record* contains the main metadata about a manuscript that the institution identifed in *Holding* supplies. These two classes are linked together with an instance of the *Manuscript* class, which corresponds to a physical manuscript, in between the *DS 2.0 Record* and *Holding* instances.

The classes with "Name Authority" and "Authority File" in the class name serve as local authority files in the knowledge graph. Authority file items are created for actors, actor roles, terms, languages, centuries, place names, and material types. These are all linked to the external authority files or vocabularies.

The classes are implemented as Wikibase items and a specific "instance of" property, *P16*, was created that corresponds to creating an instance of a class as with the *P31* of Wikidata or *rdf:type*. Alignments of the DS 2.0 data model and existing ontologies and metadata models have been considered, which could be implemented as, for example, RDF Schema.[32] subclass (*rdfs:subClassOf*) or *owl:sameAs* relations, depending on the case. The namespace prefixes used in this article are shown in Table 2.

[30] Wikibase/DataModel: https://www.mediawiki.org/wiki/Wikibase/DataModel.
[31] Wikibase RDF Extension: https://github.com/ProfessionalWiki/WikibaseRDF.
[32] RDF Schema 1.1: https://www.w3.org/TR/rdf-schema/.

Table 1. DS 2.0 Data Model Classes and Their Descriptions

Class Name	Description	QID
Manuscript	A manuscript	Q1
Holding	Holding of a manuscript in an organization etc.	Q2
DS 2.0 Record	A DS 2.0 metadata record	Q3
Standard Title	A standard (non-authoritative) string title to facilitate search (e.g., Bible, City of God, Metaphysics)	Q6
Actor (Name Authority)	An actor instance	Q7
Personal (Name Authority)	A person instance	Q8
Corporate (Name Authority)	A corporate entity instance	Q9
Role (Authority File)	Role of an actor related to a manuscript	Q10
Term (Authority File)	An authorized subject, named subject, genre, form, or other term	Q11
Language (Authority File)	Language in which the text of the manuscript is written	Q12
Century (Authority File)	Century	Q13
Place (Authority File)	A place	Q16
Material (Authority File)	Physical medium of manuscript (parchment, paper, etc.)	Q17

Table 2. Namespace Prefixes Used in This Article

Prefix	URI	Name
bd:	*http://www.bigdata.com/rdf#*	Wikibase: Blazegraph Database
owl:	*http://www.w3.org/2002/07/owl#*	Web Ontology Language (OWL)
p:	*http://www.wikidata.org/prop/*	Wikibase: Links entity to statement
ps:	*http://www.wikidata.org/prop/statement/*	Wikibase: Links value to statement
pq:	*http://www.wikidata.org/prop/qualifier/*	Wikibase: Links qualifier to statement node
rdf:	*http://www.w3.org/1999/02/22-rdf-syntax-ns#*	Resource Description Framework (RDF)
rdfs:	*http://www.w3.org/2000/01/rdf-schema#*	RDF Schema
wdt:	*http://www.wikidata.org/prop/direct/*	Wikibase: Truthy assertions about the data, links entity to value directly.
wikibase:	*http://wikiba.se/ontology#*	Wikibase ontology
xsd:	*http://www.w3.org/2001/XMLSchema#*	XML Schema

5.2 Properties

The data model properties are shown as arrows in Figure 1. Properties that are meant to be used as qualifiers are shown with dashed arrows and other properties with solid arrows. The properties for the manuscript metadata contained in a DS 2.0 record and related authority files is described in Table 3. Table 4 describes properties used for instances of *Manuscript* and *Holding*.

There are several properties ending with "as recorded," which capture the string value of certain fields of the source metadata, e.g., *material as recorded* and *language as recorded*. Where possible, these values are then mapped to instances contained within a local authority file by using a qualifier with the statement and specific properties like *material in authority file* and *language in authority file*. The instances of the local authority files are largely based on existing authority files, to which they are linked. This enables retrieving DS 2.0 records based on the linked items in the authority files.

This approach of using qualifiers for storing important information allows to both keep the connection between multiple string values for a single "as recorded" property and the authority file instances to which each of them is linked. Alternatives for using the qualifiers would be to either create separate properties for linking

Table 3. DS 2.0 Data Model Properties Used in the DS 2.0 Records and Related Authority Files

Property Name	Description	Data Type	PID
described manuscript	The described physical manuscript	*Manuscript*	P3
title as recorded	Title as given in the source metadata record	string	P10
- standard title	The assigned standard title	string	P11
uniform title as recorded	Uniform title as recorded in the source record	string	P12
associated name as recorded	Associated actor, such as author, artist, or scribe	string	P14
- in original script	Name/title in original script	string	P13
- name in authority file	Actor in authority file	*Actor*	P17
- role in authority file	Actor role	string	P15
genre as recorded	Description of what the manuscript is (esp. useful for liturgical books)	string	P18
- term in authority file	An authorized subject, named subject, genre, form, or other term	*Term*	P20
subject as recorded	What the works/texts in the manuscript are about	string	P19
- term in authority file	An authorized subject, named subject, genre, form, or other term	*Term*	P20
language as recorded	Language of the text of the manuscript	string	P21
- language in authority file	Language in authority file	Language	P22
production date as recorded	Date when the manuscript was produced as given on the source metadata record	string	P23
- earliest date	Earliest date at which the production could have happened	time	P37
- latest date	Latest date at which the production could have happened	time	P36
- century	Century of production	time	P25
- production century in authority file	Century of production in authority file	*Century*	P24
dated	Is the manuscript dated or datable	item	P26
production place as recorded	Location where the manuscript was produced	string	P27
- place in authority file	Place in authority file	*Place*	P28
physical description	Description of physical characteristics	string	P29
material as recorded	Manuscript material as given on the source metadata record	string	P30
- material in authority file	Physical medium (parchment, paper, etc.)	*Material*	P31
note	Miscellaneous note	string	P32
acknowledgements	Acknowledgements for 3rd party catalogers	string	P33
date added	Timestamp of DS record creation	time	P34
date last updated	Timestamp of last DS record update	time	P35
external URI	External URI describing the same entity	URL	P44
IIIF manifest	IIIF Presentation API document URL reference	URL	P45

Properties that are meant to be used as qualifiers are shown in italic under the properties for which they are intended to be used. The PID column depicts the property identifier in the DS 2.0 Wikibase prototype.

from the DS 2.0 record directly to both the "as recorded" and "in authority file" and lose the connection between each string value and linked item, or create artificial items that link the DS 2.0 record with a string value and the corresponding authority file instance. The latter would go against the Wikidata data modeling philosophy where the items are considered to have the same level of value as the subjects of Wikipedia pages [36].

6 DS 2.0 PROTOTYPE

A prototype Wikibase service was developed, which uses the data model and ingests data provided by the member institutions. The prototype uses a schema, which contains the items and properties of the data model as an

Table 4. DS 2.0 Data Model Properties Used for the Instances of the Classes Manuscript and Holding

Property Name	Description	Data Type	PID
DS ID	Digital Scriptorium 2.0 identifier	string	P1
holding institution as recorded	Name of the holding institution in the source metadata record	string	P5
- holding institution in authority file	Holding institution in authority file	*Corporate*	P4
holding status	The status of the holding (current or non-current)	Current/Non-cur.	P6
institutional ID	Institutional identifier for the manuscript	external identifier	P7
shelfmark	The institutional call number assigned to a manuscript	string	P8
link to institutional record	Link to catalog record or digital facsimile	URL	P9
manuscript holding	Holding of a manuscript	*Holding*	P2
- start time	time an item begins to exist or a statement starts being valid	time	P38
- end time	time an item ceases to exist or a statement stops being valid	time	P39
note	Miscellaneous note	string	P32

Properties that are meant to be used as qualifiers are printed in italic.

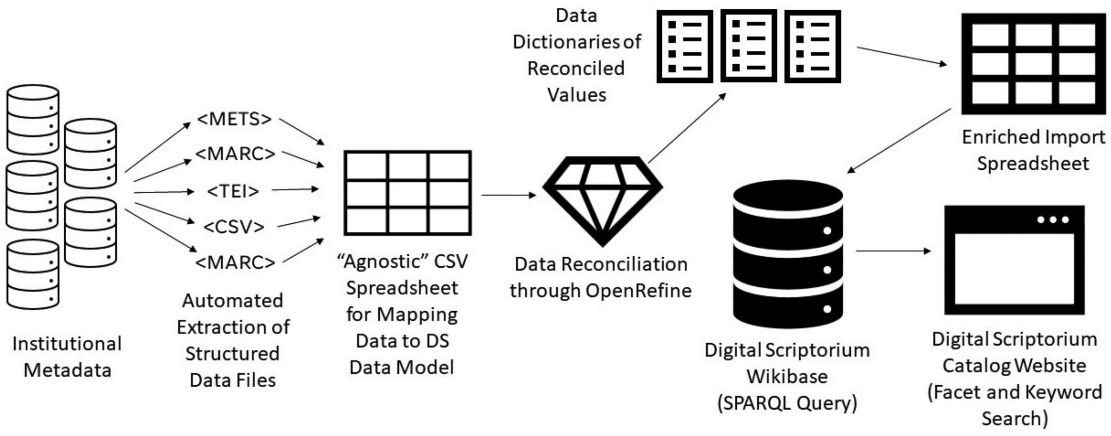


Fig. 2. Overall workflow for data extraction, aggregation, enrichment, and publication in DS.

XML file that is then imported into Wikibase in the prototype setup phase. This allows for easily recreating the prototype when needed.

6.1 Data Workflow

Once manuscript metadata were extracted or received from member institutions, the data were compiled and saved in a single CSV file (Figure 2 illustrates the overall process from data extraction to publication). As part of the extraction process, date information was structured to support future import into the Wikibase using the time datatype as well as URIs for century values derived from the **Art and Architecture Thesaurus (AAT)**.[33] Two workflows for data enrichment were tested and documented: one in which data transformation took place in a single CSV file through a series of multi-layered steps across all metadata elements and one in which metadata were split into separate CSV files that would be enriched by element type. It was found that this second workflow was easier to document and implement as a workable, streamlined process. The result was generation

[33] AAT: https://www.getty.edu/research/tools/vocabularies/aat/.

of separate CSV files for names, production places, genres, subjects, languages, and material types as described in manuscript metadata records, with a single CSV spreadsheet of all records that could be used for reference to answer questions about values found in the separated data.

Data enrichment for the manuscript records involved entity reconciliation for metadata values. The process was designed to take the individual CSV files for each metadata element type to create a metadata repository, in which existing string values in metadata records were matched to their counterparts in linked data authorities/vocabularies and LOD repositories. CSV files were loaded into OpenRefine (version 3.6.2) and reconciled using either the built-in reconciliation service (Wikidata) or one supported or developed by third parties for the necessary vocabularies (i.e., Getty Vocabularies and FAST reconciliation services). As a free, open source, web application for data cleaning and processing, OpenRefine was chosen because of its support for multiple data reconciliation services [12, 13], as well as its successful use in projects to transform metadata into linked data [32] and to facilitate data curation and semantic enrichment [1, 33, 38]. JSON recipes resulting from data processing and reconciliation in OpenRefine were extracted and uploaded to a GitHub repository. The resulting repository provided documentation so that OpenRefine workflows could be revised as they were implemented and tested against multiple iterations of the prototype data. Once the workflow was finalized, the resulting CSV files of string values and structured values were also uploaded into a GitHub repository and used to enrich the manuscript description CSV that is imported into the Wikibase instance.

6.2 Data Ingestion

Data ingestion begins with term reconciliation and the conversion of source data to the DS import CSV format. The input for the process is a set of source data from a member institution. For the prototype routines were developed for sample records in MARC XML, TEI XML, and METS XML format. We plan soon to extended the scripts to work with other source formats, beginning with CSV. The scripts are written in Ruby and are hosted on GitHub as DS-Convert.[34] These scripts extract values for reconciliation and generate the manuscript CSV for import into Wikibase. This is a two-step process. In the first step the reconciliation values are extracted into separate CSVs for each term type: support materials, languages, names, places, subject, named subjects, genres, and titles. These are enriched using the process described above, and the newly reconciled values are added to the reconciled terms CSVs in the "ds-data"[35] GitHub repository. In the second step, a DS-Convert script is used to extract manuscript descriptions from the same source data set and write it out as an import CSV. The newly updated reconciliation CSVs from the first step are used to add reconciled values to records in the import CSV. The format of the import CSV is the same regardless of the source type; however, the richness of the import CSV data (that is, the amount of information included for each manuscript) depends on the source records.

The Wikibase ingestion process takes as input the import CSV from the reconciliation and conversion process just described. The data import pipeline is built with Python using the WikibaseIntegrator[36] library. Import is also a two-step process. In the first step, the CSV is inspected for each authority value type in turn: support materials, languages, names, places, subjects, named subjects, genres, and titles. New values are added to their respective Authority Files. In the second step, the import script loops over the rows of the CSV file and either creates or updates manuscript records, based on whether a DS ID has been given for the manuscript in the CSV. Interlinked entities are created for the physical manuscript (class *Manuscript*, Q1), DS 2.0 metadata record (class *DS 2.0 Record*, Q3), and the holding of the manuscript in an institution (class *Holding*, Q2). The *Manuscript* instance contains only the DS ID and links to the other two instances. If a Holding item with the same institutional ID exists, then the existing Holding is left unchanged or a new Holding is created if the shelfmark, institutional ID, or the link to institutional record has changed. When a manuscript is updated, that is, when an existing DS ID is provided, all existing DS 2.0 Record properties are removed and replaced by the new description.

[34]DS-Convert: https://github.com/DigitalScriptorium/ds-convert.
[35]DS-Data reconciled terms directory: https://github.com/DigitalScriptorium/ds-data/tree/main/terms/reconciled.
[36]WikibaseIntegrator: https://github.com/LeMyst/WikibaseIntegrator.

6.3 Vocabularies

As part of the work to aggregate and link heterogeneous data originating from different GLAM sources and metadata records, metadata values (as string values) were matched and linked to their equivalents in authority files and controlled vocabularies. After reviewing the nature of the metadata records, the various standards used by member institutions, and the overall goals of DS 2.0, an authority plan was developed to connect string values as recorded in metadata records to the following LOD KOS:

- material types: a controlled pick list from AAT
- languages: Wikidata
- place names: **Thesaurus of Geographic Names (TGN)**[37]
- names of actors: Wikidata
- actor roles: a controlled pick list consisting of author, artist, scribe, former owner
- genre/form terms: AAT or FAST
- subject terms: FAST

The use of authority files and controlled vocabularies was guided by several aspects related to respect for member metadata practices while facilitating interoperability and reuse of heterogeneous data found in manuscript records. The first consideration was that any KOS used for linking string values would have to be available and published as LOD. In some cases, such as certain genre/form standards used in manuscript description, the lack of a **Linked Open Vocabulary (LOV)** for these KOS meant that string values from those vocabularies would not be linked to an identifier or URI from that authority (as one does not exist). The second consideration involved an endeavor to use the same or closely related KOS used by member institutions in representing their data as linked data. For instance, where genre vocabularies were known through MARC record codes, string values for AAT-derived terms were linked to AAT identifiers. Conversely, when the **Rare Book Manuscripts Section (RBMS)** Controlled Vocabularies are officially integrated and published as linked data under the RBMS Controlled Vocabulary for Rare Materials Cataloging (https://id.loc.gov/vocabulary/rbmscv.html), those identifiers can be used in the future to link RBMS-derived terms.

Where designations or codes about the vocabularies used in metadata records were not given, the third consideration examined LOV authorities in common use by the GLAM community for the purposes of creating linked data. For example, name authorities for places of manuscript production were not coded in metadata records, but matching identifiers from TGN were used to represent place names. Although linking people and organizations to Wikidata allows us to leverage the external authority files that are linked to their corresponding Wikidata entities, linking to multiple vocabularies for places (or other entities) would have been prohibitive and unsustainable for our current workflows, as places in the dataset could not be reliably reconciled with entities in Wikidata. Much like in the MMM project [25] TGN was chosen for places, because the source data (i.e., manuscript records) matched nearly all places found in TGN, which is itself widely adopted, global in scope, and supports temporal geographic information such as historical place names.

Fourth, leveraging the value of linked data to facilitate browsing and faceted search guided the decision to link precoordinated strings to identifiers for their various linked data components. This issue was brought to bear when considering the use of subject headings and genre/form terms from vocabularies maintained by the **Library of Congress (LC)**. Choosing a vocabulary in which concepts would be represented individually, within their own facets/categories, and without subordination to other topics [24] guided the decision. Rather than linking terms constructed and available as linked data at different levels of precoordination as presented in the LC linked data service, the choice was made to reconcile individual components of strings to their equivalents in FAST, which is maintained by OCLC but based on LC vocabularies. The use of FAST transformed LC-derived subject and genre string values into facet-friendly linked data.

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

The need to overcome perceived deficiencies in existing authorities, particularly name authority files, guided the fifth consideration to reconcile actors (personal and corporate names) to Wikidata. Although some authors and artists of premodern manuscripts described in metadata records may be represented in the **Library of Congress Name Authority File (LCNAF)** and the **Virtual International Authority File (VIAF)**, many other names associated with manuscript culture are not well represented in existing name authorities. Underrepresented actors include scribes and former owners of manuscripts as well as actors involved in manuscript production outside of Europe. These names are not often found in existing bibliographic resources, because they exist beyond the scope of these authorities [10]. Although one remedy to this shortcoming would be to attempt to contribute underrepresented names to traditional bibliographic authority files, such authorities often have institutional barriers to entry that Wikidata, as an open, collaboratively edited system, does not. To meet the immediate and ongoing needs of the DS 2.0 project, it was determined that Wikidata possessed greater flexibility and represented a more inclusive universe of entities to which actors in the universe of premodern manuscripts could be linked.

Furthermore, for names not already present in Wikidata, items could be quickly created. Thus, names represented only in DS 2.0 records could be contributed/added to Wikidata and linked back to their corresponding entry/item in the DS 2.0 Wikibase. In this way, DS 2.0 would take advantage of Wikidata as both a LOD repository for adding names and as a LOD hub for linking DS 2.0 to the larger Wikidata ecosystem. The ability to link DS 2.0 records to the wealth of linked data present in Wikidata, including descriptive information and links to other authorities, could then help advance computational and linked data research related to actors involved in the universe of premodern manuscripts.

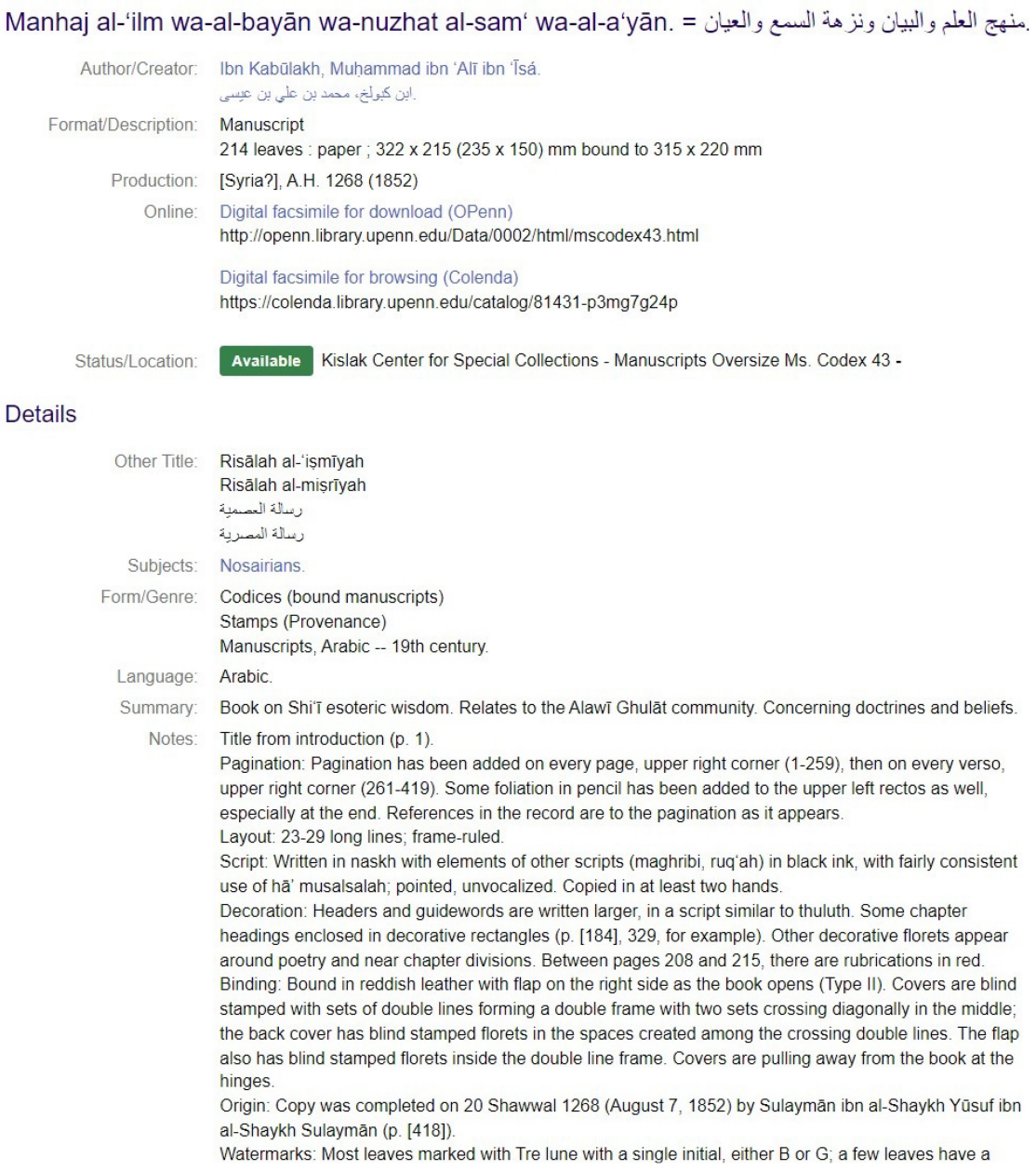
Last, the need for a manageable data processing workflow related to entity reconciliation and metadata enrichment was balanced against the heterogeneous nature of the data. Part of the workflow involved investigating reconciliation services through OpenRefine that would reliably match string values to their linked data equivalents in the authorities and vocabularies described above. With the parameters and considerations for the chosen vocabularies and standards in mind, reconciliation of nearly all entities could be accomplished through OpenRefine and services for Wikidata, the Getty Vocabularies (i.e., AAT and TGN), and FAST. Even when data reconciliation was not supported directly by a reconciliation service, such as with ISO language codes, an existing service (i.e., Wikidata) was used to leverage data that was retrieved using reconciliation services through OpenRefine. Consolidating efforts with the use of just three reconciliation services in one interface helped to streamline the process while still addressing the goal of representing as much manuscript metadata as possible with linked data authorities and vocabularies.

7 PROTOTYPE EVALUATION

Ensuring that metadata from member institutions is accurately mapped to the DS 2.0 data model and represented in the Wikibase instance is an important part of providing access and discovery to users for the aggregated, cross-institutional dataset. As an example of how member metadata has been represented in the DS 2.0 Wikibase, Figures 3, 4, and 5 show a manuscript record from an institutional catalog from the user interface of the University of Pennsylvania Libraries online catalog, the MARC record from the staff view, and how the extracted metadata has been presented and enriched as linked data as a DS 2.0 Record, respectively.[38]

Functional evaluation of the prototype was also done through basic SPARQL queries of the structured data loaded into the DS 2.0 Wikibase instance. SPARQL queries were conducted through the Wikibase query service. A grid of target data points was developed to track the individual structured values that would be the object of each example query. The queries were developed based on the metadata elements found in the DS 2.0 Record that were part of semantic enrichment. SPARQL queries were conducted to ensure that values in

[38]Please note that this prototype record is no longer accessible. To see the current structure of DS records, please visit https://catalog.digitalscriptorium.org.



Manhaj al-ʻilm wa-al-bayān wa-nuzhat al-samʻ wa-al-aʻyān. = منهج العلم والبيان ونزهة السمع والعيان.

Author/Creator: Ibn Kabūlakh, Muḥammad ibn ʻAlī ibn ʻĪsá.
ابن كبولخ، محمد بن علي بن عيسى.

Format/Description: Manuscript
214 leaves : paper ; 322 x 215 (235 x 150) mm bound to 315 x 220 mm

Production: [Syria?], A.H. 1268 (1852)

Online: Digital facsimile for download (OPenn)
http://openn.library.upenn.edu/Data/0002/html/mscodex43.html

Digital facsimile for browsing (Colenda)
https://colenda.library.upenn.edu/catalog/81431-p3mg7g24p

Status/Location: Available Kislak Center for Special Collections - Manuscripts Oversize Ms. Codex 43 -

Details

Other Title: Risālah al-ʻiṣmīyah
Risālah al-miṣrīyah
رسالة العصمية
رسالة المصرية

Subjects: Nosairians.

Form/Genre: Codices (bound manuscripts)
Stamps (Provenance)
Manuscripts, Arabic -- 19th century.

Language: Arabic.

Summary: Book on Shiʻī esoteric wisdom. Relates to the Alawī Ghulāt community. Concerning doctrines and beliefs.

Notes: Title from introduction (p. 1).
Pagination: Pagination has been added on every page, upper right corner (1-259), then on every verso, upper right corner (261-419). Some foliation in pencil has been added to the upper left rectos as well, especially at the end. References in the record are to the pagination as it appears.
Layout: 23-29 long lines; frame-ruled.
Script: Written in naskh with elements of other scripts (maghribi, ruqʻah) in black ink, with fairly consistent use of hā' musalsalah; pointed, unvocalized. Copied in at least two hands.
Decoration: Headers and guidewords are written larger, in a script similar to thuluth. Some chapter headings enclosed in decorative rectangles (p. [184], 329, for example). Other decorative florets appear around poetry and near chapter divisions. Between pages 208 and 215, there are rubrications in red.
Binding: Bound in reddish leather with flap on the right side as the book opens (Type II). Covers are blind stamped with sets of double lines forming a double frame with two sets crossing diagonally in the middle; the back cover has blind stamped florets in the spaces created among the crossing double lines. The flap also has blind stamped florets inside the double line frame. Covers are pulling away from the book at the hinges.
Origin: Copy was completed on 20 Shawwal 1268 (August 7, 1852) by Sulaymān ibn al-Shaykh Yūsuf ibn al-Shaykh Sulaymān (p. [418]).
Watermarks: Most leaves marked with Tre lune with a single initial, either B or G; a few leaves have a

Fig. 3. Online view of catalog entry Ms. Codex 43. For full catalog record, see https://franklin.library.upenn.edu/catalog/FRANKLIN_9914696423503681.

the metadata descriptions that had been reconciled to LOD vocabularies were retrievable through their LOD values.

The target goals were designed to retrieve manuscript records with data values for a given metadata element (names, places, subjects, century, etc.) in two ways: (1) by specifying the desired value for the element in the

LEADER 06630ctm a2200697Ia 4500			
005 20220609153226.0			
008 940207s1852 xx 000 0 ara d			
001 9914696423503681			
035			**a|** (OCoLC)122545123
035			**a|** (OCoLC)ocn122545123
035			**a|** (PU)1469642-penndb-Voyager
040			**a|** PAU **b|** eng **e|** amremm **c|** PAU **d|** OCLCF **d|** OCLCO **d|** OCLCQ **d|** PAU
099		9	**a|** Oversize Ms. Codex 43
100	1		**6|** 880-01 **a|** Ibn Kabūlakh, Muḥammad ibn ʻAlī ibn ʻĪsá.
880	1		**6|** 100-01//r **a|** ابن كبولخ، محمد بن علي بن عيسى.
245	1	0	**6|** 880-02 **a|** Manhaj al-ʻilm wa-al-bayān wa-nuzhat al-samʻ wa-al-aʻyān.
880	1	0	**6|** 245-02//r **a|** منهج العلم والبيان ونزهة السمع والعيان.
246	3		**6|** 880-03 **a|** Risālah al-ʻiṣmīyah
880	3		**6|** 246-03//r **a|** رسالة العصمية
246	3		**6|** 880-04 **a|** Risālah al-miṣrīyah
880	3		**6|** 246-04//r **a|** رسالة المصرية
264		0	**a|** [Syria?], **c|** A.H. 1268 (1852)
300			**a|** 214 leaves : **b|** paper ; **c|** 322 x 215 (235 x 150) mm bound to 315 x 220 mm
336			**a|** text **b|** txt **2|** rdacontent
337			**a|** unmediated **b|** n **2|** rdamedia
338			**a|** volume **b|** nc **2|** rdacarrier
520			**a|** Book on Shiʻī esoteric wisdom. Relates to the Alawī Ghulāt community. Concerning doctrines and beliefs.
500			**a|** Title from introduction (p. 1).
500			**a|** Pagination: Pagination has been added on every page, upper right corner (1-259), then on every verso, upper right corner (261-419). Some foliation in pencil has been added to the upper left rectos as well, especially at the end. References in the record are to the pagination as it appears.
500			**a|** Layout: 23-29 long lines; frame-ruled.
500			**a|** Script: Written in naskh with elements of other scripts (maghribi, ruqʻah) in black ink, with fairly consistent use of hāʾ musalsalah; pointed, unvocalized. Copied in at least two hands.
500			**a|** Decoration: Headers and guidewords are written larger, in a script similar to thuluth. Some chapter headings enclosed in decorative rectangles (p. [184], 329, for example). Other decorative florets appear around poetry and near chapter divisions. Between pages 208 and 215, there are rubrications in red.

Fig. 4. MARC view of catalog entry Ms. Codex 43.

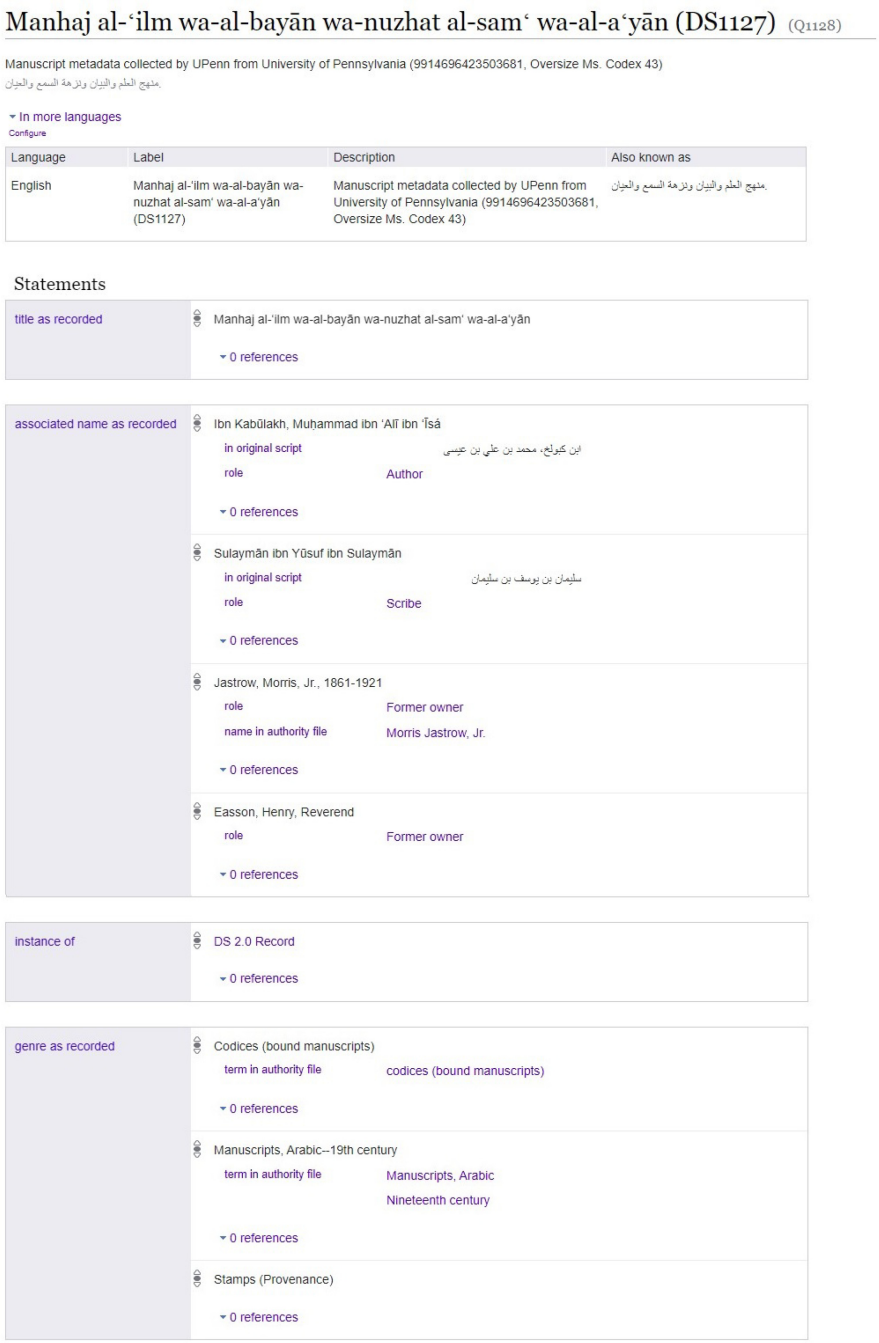
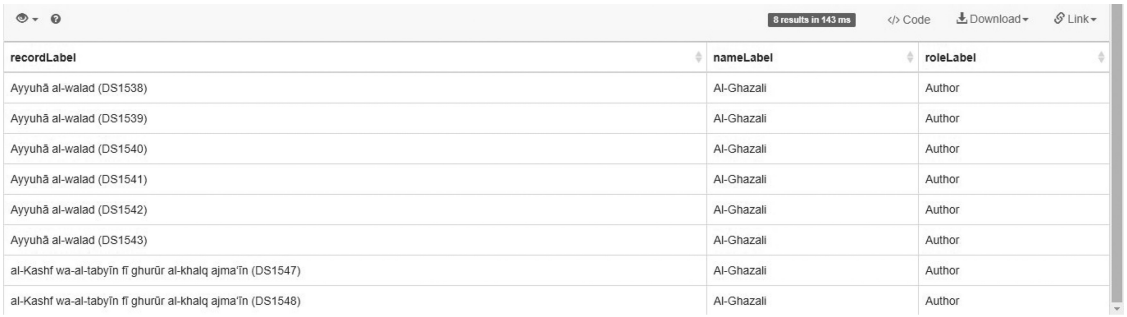


Fig. 5. Prototype Wikibase item for DS 2.0 record for Ms. Codex 43.



8 results in 143 ms Code Download Link

recordLabel	nameLabel	roleLabel
Ayyuhā al-walad (DS1538)	Al-Ghazali	Author
Ayyuhā al-walad (DS1539)	Al-Ghazali	Author
Ayyuhā al-walad (DS1540)	Al-Ghazali	Author
Ayyuhā al-walad (DS1541)	Al-Ghazali	Author
Ayyuhā al-walad (DS1542)	Al-Ghazali	Author
Ayyuhā al-walad (DS1543)	Al-Ghazali	Author
al-Kashf wa-al-tabyīn fī ghurūr al-khalq ajmaʻīn (DS1547)	Al-Ghazali	Author
al-Kashf wa-al-tabyīn fī ghurūr al-khalq ajmaʻīn (DS1548)	Al-Ghazali	Author

Fig. 6. SPARQL query results for manuscript records associated with a specific name.

Assise of all manner of breade (DS1107)	Schoenberg, Lawrence J	Former owner	Lawrence J. Schoenberg	Q107542788
Assise of all manner of breade (DS1107)	Phillipps, Thomas, Sir, 1792-1872	Former owner	Thomas Phillipps	Q2147709
Astrological material? (DS1956)	This collection of English binding fragments was given to NYU by Homer Lewis Bartlett (1858-1940), who graduated in 1884 from NYU with a degree in Science and Civil Engineering; he married Clarice Sherman Noble in 1892; they had five children; by 1902, he was second vice-president of the Eastern District Savings Bank. De Ricci states that the collection (of 101 fragments with 70...	Former owner		
Ayyuhā al-akh (DS1544)	Sabrī, ʻAbd al-Raḥmān ibn Aḥmad, d. 1726 or 7	Author		
Ayyuhā al-walad (DS1538)	Ghazzālī, 1058-1111	Author	Al-Ghazali	Q9546
Ayyuhā al-walad (DS1539)	Ghazzālī, 1058-1111	Author	Al-Ghazali	Q9546
Ayyuhā al-walad (DS1540)	Ghazzālī, 1058-1111	Author	Al-Ghazali	Q9546
Ayyuhā al-walad (DS1541)	Ghazzālī, 1058-1111	Author	Al-Ghazali	Q9546
Ayyuhā al-walad (DS1542)	Ghazzālī, 1058-1111	Author	Al-Ghazali	Q9546
Ayyuhā al-walad (DS1543)	Ghazzālī, 1058-1111	Author	Al-Ghazali	Q9546
Badāʼiʻ al-funūn (DS1228)	Smith, David Eugene, 1860-1944	Former owner	David Eugene Smith	Q1174369
Badīʻ al-ḥisāb (DS1227)	Rajab ʻAlī Beg	Author		

Fig. 7. Excerpt of SPARQL query results for manuscript records with any associated names.

record (to mimic results from a search function) and (2) by generating a list of all manuscripts with any value for that metadata element (mimicking results that would appear from a browsing function). For instance, an example query of the first type for names of associated actors was structured to return manuscript records associated with a specific name (Figure 6), and the second type of query would return any and all manuscript records with any and all associated names (Figure 7). The SPARQL queries for a specific name and for any names associated with a manuscript were as follows:

```
# find manuscript record associated with name ‘‘X’’
SELECT ?recordLabel ?nameLabel ?roleLabel
WHERE {
# bind query variables
  BIND(p:P14 AS ?nameAsRecordedStatement).
  BIND(ps:P14 As ?nameAsRecorded).
  BIND(pq:P17 AS ?hasName).
  BIND(pq:P15 AS ?hasRole).
  BIND(wdt:P42 AS ?hasQID).
```

```
# statement: manuscript record has statement for name
  ?record ?nameAsRecordedStatement ?nameStatement .
# statement: name statement has name object recorded as string value
  ?nameStatement ?nameAsRecorded ?nameString .
# statement: name statement has qualifier for structured/authority value
  ?nameStatement ?hasName ?name .
# statement: name statement has qualifier for role value
  ?nameStatement ?hasRole ?role .
# statement: authority value has QID
  ?name ?hasQID ‘‘Q9546’’ ^^ xsd:string .

SERVICE wikibase:label { bd:serviceParam wikibase:language ‘‘[AUTO_LANGUAGE]’’. }
}

# find manuscript records with associated names
SELECT DISTINCT ?recordLabel ?nameString ?roleLabel ?nameLabel ?QID
WHERE {
# bind query variables
  BIND(p:P14 AS ?nameAsRecordedStatement).
  BIND(ps:P14 As ?nameAsRecorded).
  BIND(pq:P17 AS ?hasName).
  BIND(pq:P15 AS ?hasRole).
  BIND(wdt:P42 AS ?hasQID).
# statement: manuscript record has statement for name
  ?record ?nameAsRecordedStatement ?nameStatement .
# statement: name statement has name object recorded as string value
  ?nameStatement ?nameAsRecorded ?nameString .
# statement: string statement has qualifier for possible role value
  ?nameStatement ?hasRole ?role .
# statement: string statement has qualifier for possible structured/authority value and QID
  OPTIONAL { { ?nameStatement ?hasName ?name . } OPTIONAL { ?name ?hasQID ?QID .} }
  SERVICE wikibase:label { bd:serviceParam wikibase:language ‘‘[AUTO_LANGUAGE]’’. }
  }
  ORDER by ?recordLabel ?nameLabel
```

Example queries of this nature were repeated for different types of structured data in manuscript records, including names of actors, places of production, subjects, genres, centuries, material types, languages, and holding institutions. Example queries were also written to find manuscript records containing information about dates, including records for manuscripts with a specified earliest year of production, latest year of production, production within a range of years, and records documenting manuscripts determined to be datable or non-datable by scholars. It should be noted that at this time, titles have not been semantically enriched, and SPARQL queries for titles do not use URIs or Wikidata IDs for retrieving manuscript records. Instead, a SPARQL query that matches or contains a string value for a recorded or uniform title can be used.

The example SPARQL queries described above for dataset exploration and prototype testing were uploaded to a Github repository (https://github.com/DigitalScriptorium/ds-testing/tree/main/sparql). With prototype data loaded into the Wikibase instance, data for each type of structured value tested by SPARQL queries was found to be returned successfully using the integrated SPARQL endpoint. As structured by the data model, items, datatypes, properties, and qualifiers functioned as intended in all 24 queries (see Table 5) used in testing. Data

Table 5. Example SPARQL Queries

Metadata Element Tested	First Example Query	Second Example Query
century	records with value for "seventeenth century" (AAT ID: 300404511)	all manuscript records sorted by century names
genre	records with value for "Qur'ans" (AAT ID: 300265128)	all manuscript records sorted by genre terms
holding institution	records with value for "University of Pennsylvania" (QID: Q49117)	all manuscript records sorted by holding institutions
language	records with value for "French" (ISO 639-3: fra)	all manuscript records sorted by languages
material	records with value for "parchment" (AAT ID: 300011851)	all manuscript records sorted by material types
name	records with value for "Al-Ghazali" (QID: Q9546)	all manuscript records sorted by associated names
place	records with value for "Italy" (TGN ID: 1000080)	all manuscript records sorted by places
subject	records with value for "Astronomy, Arab" (FAST ID: 819764)	all manuscript records sorted by subject terms
dated manuscripts	records with "datable" value for dated property	records with "non-datable" value for dated property
earliest date	records for manuscripts with an exact earliest production date of "1700"	records for manuscripts with an earliest production date after "1066"
latest date	records for manuscripts with an exact latest production date of "1699"	records for manuscripts with a latest production date before "1812"
date range	records for manuscripts with a production date range between "1245" and "1345"	records for manuscripts with an earliest production date before "1000" or a latest production date after "1900"

imported into the Wikibase instance were discoverable using these queries, showing that the data model and the Wikibase met basic functional requirements of discoverability as intended for finding DS 2.0 manuscript record data.

8 CONCLUSION

The DS 2.0 prototype demonstrated the functionality of the data model design and its alignment with the DS 2.0 principles outlined at the outset. Designed for implementation in Wikibase, the DS 2.0 data model facilitates an efficient workflow able to capture a sufficient amount of structured metadata from a variety of institutional records using a variety of formats. The prototype serves as a proof of concept for building a national union catalog using Wikibase. SPARQL queries designed to test the prototype also indicate how end users can flexibly query the dataset. While a more user-friendly, facet-based interface (currently in development) will ease inexperienced users into the process of querying the DS 2.0 dataset, more advanced users will be able to delve more deeply into complex queries via a SPARQL endpoint and leverage the connected and more expansive Wikidata knowledge graph for deeper and broader research questions related to premodern manuscript studies. The DS 2.0 model takes into account the social and practical implications of data collection that must be acknowledged and accounted for in a collaborative project where the collaborators are not necessarily equal in terms of funding, staffing, and expertise. The model supports a less-is-more approach to data collection, but at the same time

leverages LOD technologies so that the "less" produces multiple and enhanced pathways to "more." As DS 2.0 moves toward implementation and as with all digital projects, the greatest challenge still remains one of sustainability. However, the reduced workflow barriers for contributing data and the enhanced research value of enriched DS 2.0 data in the larger manuscript studies research community renders this challenge much less daunting for both the near- and long-term prospects of this platform.

REFERENCES

[1] Sotirios Angelis and Konstantinos Kotis. 2021. Generating and exploiting semantically enriched, integrated, linked and open museum data. In *Proceedings of the 14th International Conference on Metadata and Semantic Research (MTSR'20) (Communications in Computer and Information Science, Vol. 1355)*, Emmanouel Garoufallou and María-Antonia Ovalle-Perandones (Eds.). Springer, Cham, 367–379.

[2] Konstantin Baierer, Evelyn Dröge, Kai Eckert, Doron Goldfarb, Julia Iwanowa, Christian Morbidoni, and Dominique Ritze. 2017. DM2E: A linked data source of digitised manuscripts for the digital humanities. *Semantic Web* 8, 5 (2017), 733–745. https://doi.org/10.3233/SW-160234

[3] Sheila A. Bair and Susan M. B. Steuer. 2013. Developing a premodern manuscript application profile using dublin core. *J. Libr. Metadata* 13, 1 (2013), 1–16. https://doi.org/10.1080/19386389.2013.778725

[4] Valentina Bartalesi, Daniele Metilli, Nicolò Pratelli, and Paolo Pontari. 2021. Towards a knowledge base of medieval and renaissance geographical latin works: The IMAGO ontology. *Dig. Scholar. Human.* 37, 1 (July 2021), 34–50. https://doi.org/10.1093/llc/fqab060

[5] Valentina Bartalesi, Nicolò Pratelli, and Emanuele Lenzi. 2022. A knowledge base of medieval and renaissance geographic latin works. In *Proceedings of the 18th Italian Research Conference on Digital Libraries (CEUR Workshop Proceedings, Vol. 3160)*. 6 pages. Retrieved from https://ceur-ws.org/Vol-3160/short13.pdf.

[6] Sebastian Barzaghi, Monica Palmirani, and Silvio Peroni. 2020. Development of an ontology for modelling medieval manuscripts: The case of progetto IRNERIO. *Umanistica Digitale*9 (2020), 117–140.

[7] Anna Bellotto. 2020. Medieval manuscript descriptions and the semantic web: Analysing the impact of CIDOC CRM on Italian codicological-paleographical data.*Dig. Human. Quart.* 14, 1 (2020). Retrieved from https://www.proquest.com/scholarly-journals/medieval-manuscript-descripti ons-semantic-web/docview/2553808011/se-2.

[8] Giovanni Bergamin. 2022. Wikibase, or the search for the unicorn. *JLIS.it* 13, 3 (Sep.2022), 49–62. https://doi.org/10.36253/jlis.it-484

[9] Giovanni Bergamin and Cristian Bacchi. 2018. New ways of creating and sharing bibliographic information: An experiment of using the Wikibase data model for UNIMARC data. *JLIS.it* 9, 3 (2018), 35–74. https://doi.org/10.4403/jlis.it-12458

[10] Carlo Bianchini, Stefano Bargioni, and Camillo Carlo Pellizzari di San Girolamo. 2021. Beyond VIAF. *Info. Technol. Libr.* 40, 2 (June2021). https://doi.org/10.6017/ital.v40i2.12959

[11] Melissa Conway and Lisa Fagin Davis. 2015. Directory of collections in the United States and Canada with pre-1600 manuscript holdings. *Papers Bibliogr. Soc. Amer.* 109, 3 (2015), 273–420. https://doi.org/10.1086/682342

[12] Antonin Delpeuch. 2020. Running a reconciliation service for wikidata. In *Proceedings of the 1st Wikidata Workshop (Wikidata'20) (CEUR Workshop Proceedings, Vol. 2773)*, Lucie-Aimée Kaffee, Oana Tifrea-Marciuska, Elena Simperl, and Denny Vrandečić (Eds.). 8. Retrieved from http://ceur-ws.org/Vol-2773/paper-17.pdf.

[13] Antonin Delpeuch. 2020. A survey of openrefine reconciliation services. In *Proceedings of the 15th International Workshop on Ontology Matching Co-Located with the 19th International Semantic Web Conference (ISWC'20) (CEUR Workshop Proceedings, Vol. 2788)*, Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn (Eds.). 82–86. Retrieved from http://ceur-ws.org/Vol-2788/om2020_STpaper3.pdf.

[14] Dennis Diefenbach, Max De Wilde, and Samantha Alipio. 2021. Wikibase as an infrastructure for knowledge graphs: The EU knowledge graph. In *Proceedings of the International Symposium on the Semantic Web (ISWC'21)*, Andreas Hotho, Eva Blomqvist, Stefan Dietze, Achille Fokoue, Ying Ding, Payam Barnaghi, Armin Haller, Mauro Dragoni, and Harith Alani (Eds.). Springer International Publishing, Cham, 631–647.

[15] Chris Dijkshoorn, Lora Aroyo, Jacco Van Ossenbruggen, and Guus Schreiber. 2018. Modeling cultural heritage data for online publication. *Appl. Ontol.* 13, 4 (2018), 255–271.

[16] Martin Doerr. 2003. The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata. *AI Mag.* 24, 3 (2003), 75–75.

[17] Martin Doerr, Stefan Gradmann, Steffen Hennicke, Antoine Isaac, Carlo Meghini, and Herbert Van de Sompel. 2010. The Europeana data model (EDM). In *Proceedings of the 76th IFLA General Conference and Assembly of the World Library and Information Congress*. IFLA, Gothenburg, Sweden, 12.

[18] Consuelo Dutschke. 2007. Digital scriptorium: Ten years old. In *Conoscere il manoscritto: Esperienze, progetti, problemi. Dieci anni del progetto Codex in Toscana (Millennio medievale, Vol. 70)*. SISMEL: Edizioni del Galluzzo, Florence, 189–205.

[19] Jason Evans. 2019. *Treasured Manuscript Collection Gets the Wikidata Treatment*. The National Library of Wales. Retrieved from https://blog.library.wales/treasured-manuscript-collection-gets-the-wikid ata-treatment/.

[20] Barbara Fischer. 2022. Towards an open and collaborative authority control. *JLIS.it* 13, 1 (Jan.2022), 283–290. https://doi.org/10.4403/jlis.it-12767

[21] Jean Godby, Karen Smith-Yoshimura, Bruce Washburn, Kalan Knudson Davis, Christine Fernsebner Eslao, Steven Folsom, Xiaoli Li, Marc McGee, Karen Miller, Honor Moody, Craig Thomas, and Holly Tomren. 2019. *Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage.* OCLC Research, Dublin, OH. https://doi.org/10.25333/faq3-ax08

[22] Joy Humphrey. 2007. Manuscripts and metadata: Descriptive metadata in three manuscript catalogs: DigCIM, MALVINE, and digital scriptorium. *Catalog. Classif. Quart.* 45, 2 (2007), 19–39.

[23] Eero Hyvönen, Esko Ikkala, Mikko Koho, Jouni Tuominen, Toby Burrows, Lynn Ransom, and Hanno Wijsman. 2021. Mapping manuscript migrations on the semantic web: A semantic portal and linked open data service for premodern manuscript research. In *Proceedings of the International Symposium on the Semantic Web (ISWC'21) (Lecture Notes in Computer Science, Vol. 12922)*. Springer, Cham, 615–630. 10.1007/978-3-030-88361-4_36

[24] R. E. Johnson. 2016. Reconciling metadata to FAST linked data. Retrieved from https://remerjohnson.github.io/2016/02/16/reconciling-metadata-to-FAST-li nked-data.html.

[25] Mikko Koho, Toby Burrows, Eero Hyvönen, Esko Ikkala, Kevin Page, Lynn Ransom, Jouni Tuominen, Doug Emery, Mitch Fraas, Benjamin Heller, David Lewis, Andrew Morrison, Guillaume Porte, Emma Thomson, Athanasios Velios, and Hanno Wijsman. 2021. Harmonizing and publishing heterogeneous pre-modern manuscript metadata as linked open data. *J. Assoc. Info. Sci. Technol.* 73, 2 (May2021), 240–257. https://doi.org/10.1002/asi.24499

[26] Patrick Le Bœuf. 2012. Modeling rare and unique documents: Using FRBRoo/CIDOC CRM. *J. Arch. Organiz.* 10, 2 (2012), 96–106. https://doi.org/10.1080/15332748.2012.709164

[27] Carlo Meghini, Valentina Bartalesi, Daniele Metilli, and Filippo Benedetti. 2019. Introducing narratives in Europeana: A case study. *Int. J. Appl. Math. Comput. Sci.* 29, 1 (2019), 7–16. https://doi.org/10.2478/amcs-2019-0001

[28] Pat Riva, Martin Doerr, and Maja Žumer. 2009. FRBRoo: Enabling a common view of information from memory institutions. *Int. Catalog. Bibliogr. Control* 38, 2 (2009), 30–34. Retrieved from https://archive.ifla.org/IV/ifla74/papers/156-Riva_Doerr_Zumer-en.pdf.

[29] Lozana Rossenova, Paul Duchesne, and Ina Blümel. 2022. Wikidata and Wikibase as complementary research data management services for cultural heritage data. In *Proceedings of the 3rd Wikidata Workshop 2022 Co-located with the 21st International Semantic Web Conference (ISWC'22) (CEUR Workshop Proceedings)*, Lucie-Aimée Kaffee, Simon Razniewski, Gabriel Amaral, and Kholoud Saad Alghamdi (Eds.). Retrieved from https://ceur-ws.org/Vol-3262/paper15.pdf.

[30] Cogan Shimizu, Pascal Hitzler, Quinn Hirt, Dean Rehberger, Seila Gonzalez Estrecha, Catherine Foley, Alicia M. Sheill, Walter Hawthorne, Jeff Mixter, Ethan Watrall, et al. 2020. The enslaved ontology: Peoples of the historic slave trade. *J. Web Semant.* 63 (2020), 100567.

[31] Nikolaos Simou, Alexandros Chortaras, Giorgos Stamou, and Stefanos Kollias. 2017. Enriching and publishing cultural heritage as linked open data. In *Mixed Reality and Gamification for Cultural Heritage*, Marinos Ioannides, Nadia Magnenat-Thalmann, and George Papagiannakis (Eds.). Springer International Publishing, Cham, 201–223. https://doi.org/10.1007/978-3-319-49607-8_7

[32] Silvia B. Southwick. 2015. A guide for transforming digital collections metadata into linked data using open source technologies. *J. Libr. Metadata* 15, 1 (2015), 1–35. https://doi.org/10.1080/19386389.2015.1007009

[33] Nishad Thalhath, Mitsuharu Nagamori, Tetsuo Sakaguchi, and Shigeo Sugimoto. 2021. Wikidata centric vocabularies and URIs for linking data in semantic web driven digital curation. In *Proceedings of the 14th International Conference on Metadata and Semantic Research (MTSR'20) (Communications in Computer and Information Science, Vol. 1355)*, Emmanouel Garoufallou and María-Antonia Ovalle-Perandones (Eds.). Springer, Cham, 336–344.

[34] Bjarki Valtysson. 2012. EUROPEANA. *Info., Commun. Soc.* 15, 2 (2012), 151–170. 10.1080/1369118X.2011.586433

[35] Tobias Viegener. 2009. SwissBib: Ein metakatalog nextgeneration oder 2.0. *arbido* 3 (2009). Retrieved from https://arbido.ch/fr/edition-article/2009/digitale-dienstleistungen-als-h erausforderung-in-i-d/swissbib-ein-metakatalog-nextgeneration-oder-2-0.

[36] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[37] Maja Žumer and Pat Riva. 2017. IFLA LRM—Finally here. In *Advancing Metadata Practice: Quality, Openness, Interoperability: 2017 Proceedings of the International Conference on Dublin Core and Metadata Applications.* Dublin Core Metadata Initiative (DCMI), Washington, DC, 13–23.

[38] Marcia Lei Zeng. 2019. Semantic enrichment for enhancing LAM data and supporting digital humanities. Review article. *El profesional de la información* 28, 1 (2019), 1699–24077.

Received 1 June 2022; revised 20 December 2022; accepted 13 February 2023