A smart data case study using Wikidata to expand access and discovery in the Schoenberg Database of Manuscripts[1]

L.P. Coladangelo[a*] and Lynn Ransom[b]
[a]*College of Communication and Information, Kent State University, 800 E. Summit St., Kent, OH 44242, USA*
[b]*Schoenberg Institute for Manuscript Studies, Penn Libraries, University of Pennsylvania, 3420 Walnut Street, Philadelphia, PA 19104, USA*

Abstract. This case study explored the results and lessons learned from the initial contribution of over 10,000 name identifiers to Wikidata and considered the use of Wikidata for enhancement of data related to premodern manuscripts. Wikidata, as a Linked Open Data (LOD) repository and hub, was used in the semantic enrichment of a particular dataset from the Schoenberg Database of Manuscripts (SDBM) Name Authority, yielding unique insights only possible from linking data from Wikidata and the SDBM. Mapping named entity metadata related to premodern manuscripts from one context to another was also explored, with a particular emphasis on determining property alignments between the linked data models of the SDBM and Wikidata. This resulted in a workflow model for LOD management and enhancement of name authority data in library, archive, and museum (LAM) contexts to encourage the manuscript studies community to contribute further data to Wikidata. This research demonstrates how the application of smart data principles to an existing dataset can address knowledge gaps related to people traditionally underrepresented in the digital record and opens new possibilities for access and discovery.

Keywords: Semantic enrichment, Wikidata, smart data, digital humanities, name authorities

1. Introduction

While the possibilities of Linked Open Data to broaden access to and discovery of our material past have long been recognized in the cultural heritage sector [12], institutions and resource developers are still grappling with how best to create and implement Linked Open Data (LOD) strategies and practices that are inclusive and sustainable. In the last several years, project developers have been increasingly turning to Wikidata, which offers a free and open-access platform for contributing data to the world's largest, community-maintained knowledge graph with relatively few barriers. Among these projects is the Schoenberg Database of Manuscripts (SDBM; https://sdbm.library.upenn.edu/), an open access repository of data related to premodern manuscripts. The SDBM manages an internal Name Authority of over 50,000 names of people and institutions involved in the production and trade of premodern manuscripts from medieval times to the present day, many of which are not recorded anywhere else in a digital context. In an effort to leverage the value of this unique resource for manuscript studies and to "smarten up" the dataset to uncover lost voices in the history of premodern manuscript studies, the SDBM team undertook an exploratory exercise to contribute SDBM Names to Wikidata. The result is a case study of and an ecological model for LOD management and enhancement of name authority creation in library, archive, and museum (LAM) contexts that we hope will encourage the manuscript studies community and other related disciplines to contribute further data to enrich the digital record. Our case study demonstrates how the application of smart data principles to an existing dataset addresses knowledge gaps relating to people traditionally underrepresented in the digital record and opens up new possibilities for access and discovery.

[1] This research was supported by the Institute of Museum and Library Services (IMLS), LB21 LEADING project: RE-246450-OLS-20
* Corresponding author. E-mail: lcoladan@kent.edu

2. Background on the Schoenberg Database of Manuscripts and the SDBM Data Model

The SDBM was created in 1997 by the American manuscript collector Lawrence J. Schoenberg to track the value and movement of manuscripts across time. Since then, it has become the largest, freelyaccessible resource for manuscript data in the world with a growing number of over 278,000 entries documenting the existence and characteristics of manuscripts through observations recorded in various sources: auction and sale catalogs (dating to as early as the fifteenth century), institutional catalogs, inventories, online sources and personal observations made by members of the SDBM user community. The SDBM has become an important resource especially for provenance studies for its documentation of sellers, buyers, owners, and other figures involved in the historic movements of premodern manuscripts.

The collected observations that form the foundation of the SDBM are recorded as entries. To create an entry, the manuscript description and provenance information provided by a source is parsed into structured fields including author, title, date, place of production, scribe, artist, former owner, and physical details such as dimensions, number of folios, decoration, etc. Because a manuscript can appear in sources multiple times as it moves from one sale to the next or one collection to another, the matching entries can be linked to form "manuscript records" that simply aggregate the linked entries and contain no other data. Entries are further linked via internal name and place authorities so that all entries associated with a specific place or a person or institution can be accessed from the authority record. Also included when available are links to external authorities: the Virtual International Authority File (VIAF) for names and the Getty Thesaurus of Geographic Names (TGN) and GeoNames for places. As a result of these linkages, the SDBM Name and Place Authorities are powerful tools for access and discovery within and beyond the SDBM's data model.

The underlying data model[2] for the SDBM uses a local ontology to support publication of the crowdsourced data in the SDBM as LOD [5]. Conceptually, the SDBM is a database of observations about manuscripts, which are structured as entities known as Entries. Entries are derived from Sources, which can be auction or sale catalogs, published and unpublished resources, or even personal observations. Entries presumed to be about the same manuscript object are connected to an entity known as a Manuscript Record, which acts as a linking node for displaying composite provenance information from linked Entries. Name and Place Authority Records standardize information for people, organizations, and locations found in Entries and Sources in order to optimize search results.

As linked and linkable data, the SDBM authorities provide a unique resource to build connections in the networks underlying the production and trade of premodern manuscripts. This is especially true of the Name Authority, which currently contains over 54,000 names. In addition to links to associated Entries and to associated VIAF identifiers when available, the SDBM Name Authority records also contain structured data relating to the person or institution identified, including a unique identifier, a human-readable label, dates of existence (expressed as "start" and "end" dates), roles (such as author, scribe, or provenance agent) expressed in related Entries and Sources, and links to associated places in the SDBM Place Authority. All of the classes of entities, properties, and value data types used to describe Name Authority entities, as well as other entities in the local ontology, are specified in the data model[3] and its entity-relationship diagram.[4]

For many of these names, the SDBM is the only digital record attesting to the existence of the person or institution identified in the record. Over half of the SDBM names cannot be associated with existing names in VIAF. Even though many medieval authors and artists tend to be well represented in the national and international authority files, there are still many who are not, and the majority of scribes and former owners mentioned in catalog entries generally do not rise to the level of significance in traditional and current manuscript cataloging practices to be assigned an authority record. This is especially true for women and other traditionally underrepresented groups. Contributing SDBM Names to Wikidata thus offers an opportunity to enhance and expand discovery of these persons and institutions and their networks that would otherwise remain hidden in a linked open data environment. It also affords the SDBM the opportunity to further develop knowledge about these actors that is not currently possible in the existing data model by leveraging the expansive Wikidata properties to complete a richer, more nuanced, and "smart-

[2] https://sdbm.library.upenn.edu/static/docs/SDBM_data_explanation2019.pdf

[3] https://sdbm.library.upenn.edu/pages/SPARQL%20Data%20Model

[4] https://sdbm.library.upenn.edu/static/docs/erd.pdf

er" data profile in Wikidata for these otherwise unacknowledged and underrepresented actors in the wider landscape of our shared cultural and intellectual heritages.

3. Literature review

This section reviews the research on smart data, semantic enrichment, and the infrastructure provided by Wikidata as a LOD repository and authority hub to assist development of smart data and enrichment projects.

3.1. Big data, smart data, and semantic enrichment

Big data is defined by the obstacles and opportunities presented in the contemporary proliferation of large, heterogeneous, rapidly accumulating datasets used in computational and analytic research [1,13]. In a big data environment, the benefits of searching, aggregating, connecting, and discovering patterns across large amounts of data are coupled with difficulties in managing data created and published quickly, in many forms, and in varying degrees of quality. Scholars have characterized these dimensions of big data using various lists of "V" terms, enumerating concepts like volume, velocity, variety, and veracity. These Vs represent advantages and disadvantages which must be considered and balanced, such as with voluminous datasets which may be unstructured or quickly proliferating data which may not have well-documented origins. The challenge for big data research is to take advantage of its positive aspects while mitigating its negatives.

Smart data, particularly in the context of the humanities, has been advanced as one such solution. Defined as data which affords insights when analyzed at any level [34], small or large, it is often the product of expert labor and technological resources that make it structured (i.e., machine-processable), explicit (i.e., well described through metadata), enriched (e.g., marked up, annotated, contextualized), and clean (i.e., corrected, standardized, and/or verified) [21]. Smart data research projects in the humanities allow researchers to use existing well-structured and well-sourced small datasets of smart data to make inferences about large-scale patterns or to help answer complex questions. This work can also focus on improving the quality, interoperability, and reusability of data in an effort to increase its value, such as structuring and linking unstructured information to help uncover previously hidden networks of activity in cultural heritage practices [16,17]. One type of big data which has been the focus of this type of smart data work has been the vast amounts of metadata produced from LAM communities [35]. Stored and maintained in places like library catalogs and archival records, this data has the advantages of being authoritative, standardized, and well-described, rendering it valuable and immediately useful for projects and applications in need of smart data.

Semantic enrichment is one such LOD-powered smart data strategy, which enhances the value of data using semantic techniques [29]. Semantic enrichment covers any number of approaches which augment datasets and create contextual associations between source data (to be enriched) and target data (resources used for enrichment) [11], which have the potential to help users gain insights from use and exploration of the enriched data. In the context of linked data, semantic enrichment often involves using semantic relationships (or properties) from linked data models, standards, or ontologies to align or reconcile data values and construct and publish links between datasets [22]. In the past decade, semantic enrichment of cultural heritage data, especially linked data, has been a burgeoning area of research for digital humanists [3,7,8,25,26,29], including various enrichment projects involving LAM data [34] and name authorities [4]. Projects of this nature provide reciprocal value to both source and target data by allowing users to meaningfully query across linked databases simultaneously.

While links between related datasets are valuable, such projects often benefit from using hub databases or repositories of structured data which cover large subject areas or multiple disciplines and domains. The scope, variety, and types of data available in such systems enhance the amount and diversity of contextual information to which datasets can be linked, and thus enriched. Although a number of resources exist within the ecosystem of LOD repositories and databases, Wikidata boasts many advantages which have made it increasingly popular [24] in support of semantic enrichment efforts, particularly in application to library [28] and digital humanities [36] projects.

3.2. Wikidata as a LOD repository for smart data applications

Wikidata serves as a database of structured, linked data on the backend of Wikimedia applications such

as Wikipedia. Wikidata is a multilingual, sociotechnical, and collaboratively edited system [18] for storing LOD. In this way, it functions as an open-source knowledge base leveraged by various projects in different domains and disciplines. The data model for Wikidata is based on subject-property-value statements built upon the RDF data standard. The RDFstructured database, or RDF graph, that makes up Wikidata includes two types of structured entities: items and properties. Items, identified by a string of numbers prefixed by the letter "Q," act as subjects and objects in the RDF-structured triples, or as nodes in an RDF graph. Properties, with prefix "P" before an identifying string of numbers, act as predicates or arcs, to define relationships between subjects and objects. Constraints or rules on the use of properties help to determine the valid range, types, or amount of values that can be used to express the objects of statements. Although data values inconsistent with property constraints can still be expressed as the objects of statements, such values are flagged with warnings to notify users that they may be erroneous or invalid given how a property was defined to be used.

Wikidata also allows refinement of RDF statements to provide additional contextual or provenance information through the use of references, rankings, and qualifiers. References are used to express the relationship of a given source to a particular value by documenting the source used to support the assertion made in the statement. Ranks are used to annotate statements to allow differentiation between multiple values of a statement, which can be useful when different values of a statement may represent temporal or conditional aspects, historical contexts, alternative perspectives, or controversies. Qualifiers are any other refinements, annotations, or contextualized information expressed about a statement value. The use of qualifiers enhances the validity and sourcing of statements and helps to express more granular structured data than may be possible at the statement level.

While Wikidata is an open platform and collaboratively edited, the infrastructure of references, ranking annotations, and qualifiers do allow users to have some assurances of transparency and accountability as these mechanisms can be used to verify or contextualize the quality of crowdsourced data. Furthermore, Wikidata items and properties, formatted and presented in the user interface as web pages, each have an audit trail of changes and revisions. Such changes are logged through the edit history of the Wikidata page for a given entity as well as listed by each individual user on the special page logging a user's contributions. If concerns about abuse or vandalism arose, such documentation would provide the necessary evidence that other members of the Wikidata community would need to investigate if information had been unduly edited or tampered with. Additionally, the collective and technical aspects of the system allow the Wikidata community to respond quickly and with relatively low access barriers to update and correct information. Social mechanisms such as thanking users for edits, item talk pages, and wiki-based discussions projects and proposals emphasize data stewardship, collaboration, and mutual reinforcement of best practices.

In these ways, Wikidata represents and enacts the beneficial Vs of big data in support of smart data projects. As of May 2023, Wikidata boasted a volume of over 103 million items created and edited with tremendous velocity in the amount of 1.9 billion edits since the project launched [31]. The variety of data can be seen by the many subject areas supported by various WikiProjects covering a wide range of domains and topics. Veracity of data is maintained and the risks of variability are mitigated by the data model and its associated affordances for structuring statements, flagging or warning users of inconsistencies in data values relative to property constraints, and representing qualifying and contextual information. Finally, its value is evidenced through increasing use in projects to enhance research data and cultural heritage metadata [2,9,10,20] as well as acting as a LOD hub for digital curation projects [27]. When taken together, all of these factors make Wikidata an appealing choice for semantic enrichment projects relying on development and integration of smart data.

3.3. Wikidata as an authority linking hub

The ability to use Wikidata for semantic enrichment comes in large part from the extensive integration of contextual information represented as structured metadata. In addition to statements which express descriptive information, Wikidata items also include statements which link items to external resources. For instance, a person found in Wikidata can be described through properties which express information like birth and death dates, occupations, affiliations with other people and organizations, and places with which they were associated. Properties for external IDs, on the other hand, structure statements which link identifiers found in outside data-

bases, vocabularies, and authority files to the same entities as represented in Wikidata. Additionally, a number of WikiProjects are dedicated to increasing the quality and quantity of items in Wikidata based on the use of external IDs for outside authorities and databases, such as those projects for Authority Control [30] and Biographical Identifiers [32].

Numerous institutions and organizations have Wikidata properties for their authority file IDs, such as the national libraries of Germany (P227), France (P268), Japan (P349), Spain (P950), Norway (P1015), Greece (P3348), Belarus (P3390), Lebanon (P7026), New Zealand (P7682), Slovakia (P7700), and Israel (P8189); the Library of Congress (P244) in the United States; and the name authorities IDs of the Russian State Library (RSL) (P947), the National Library of Australia (P1315), the Integral Information System of the National Library (NEKTÁR) of Hungary (P3133), the Hong Kong Chinese Authority (Name) (HKCAN) database (P5909), the Shanghai Library Name Authority Database (P6702), the Biographical Database of the Academia Sinica in Taiwan (P6705), the Inter-university Consortium for Political and Social Research (ICPSR) Personal Names Authority List (P10328), and the French-language name authority for Libraries and Archives Canada, the Canadiana Name Authority (P8179).

In terms of manuscript-related projects, external ID properties have enabled linking of Wikidata items to authority records from the National Library of Wales (P2966), Italian libraries with collections represented in the database Manus Online (P8975), Medieval Manuscripts in Oxford Libraries (P9017; P9018; P9019), Bodleian Archives & Manuscripts (P9594), and the Hill Museum & Manuscript Library (P9943).

Wikidata has also been leveraged in a number of research projects to enhance or expand the quality and usability of existing authority files and datasets as well as to improve discoverability of information through LOD. For instance, Wikidata has been used as an authority linking hub for data about economists from the Integrated Authority Files (GND) (P227) and the Research Papers for Economics (RePEc) (P2428) database, using VIAF IDs as one method to align entities [15]. The European Holocaust Research Infrastructure (EHRI) utilized Wikidata to both contribute data to Wikidata as well as to integrate data from Wikidata in its authority files for the improvement of both resources [6]. Recognizing the usefulness of Wikidata as an identifier hub, the Library of Congress begun to include links to Wikidata in its authority records [23]. Finally, the use of Wikidata as an open and publicly accessible database has been advanced as a strategy to enhance the visibility of Black artists in a project to align and enrich Wikidata with information from the Philadelphia Museum of Art (PMA) using PMA entity IDs (P8317) [23].

4. Methods

This section describes the process of extracting a test dataset of named entities from the SDBM and linking them to Wikidata items in order to semantically enrich the SDBM data with additional contextual information[5].

4.1. Extraction of the test dataset and the use of OpenRefine

Because the SDBM is composed of RDF-structured data, we were able to use the RDF query language SPARQL (SPARQL Protocol and RDF Query Language) to extract a test dataset of entities from the SDBM Name Authority. The SPARQL query (Appendix A) applied in the Yet Another SPARQL GUI (YASGUI)[6] client was asked to return structured information for named entities also represented by a VIAF ID in the database. These criteria ensured that the test dataset included verified names in another authority file as well as external identifiers to assist alignment of entities between the SDBM and Wikidata.

The SPARQL query also did not include deleted or deprecated names. Because of the crowdsourced nature of the SDBM, system administrators merge duplicates into a preferred record when found and hide deprecated records from the public interface to prevent their future use. For the purposes of data provenance, the hidden deprecated name records are retained in the database, making it possible they would appear in the query results if not filtered out.

Information extracted from the database for each authority record included the SDBM identifier and human-readable controlled name or label, the corresponding VIAF ID number recorded in the SDBM, the name type (personal or corporate), and any associated start and end dates and places (if available). The targeted SPARQL query of authority records with VIAF IDs yielded 13,486 named entities. The isolated dataset of extracted names and associated

[5] Dataset CSVs and JSON recipe files can be found here: https://doi.org/10.5281/zenodo.8033928

[6] https://yasgui.triply.cc/

information was saved as a CSV file and loaded into OpenRefine[7] version 3.7.2, a free, open source, web-based application designed to clean and process data. In addition to being an open, widely used platform in data-related digital humanities projects, OpenRefine was chosen because it supports multiple data reconciliation services (e.g., Wikidata, VIAF, Getty Vocabularies) and an extension for automated batch editing of Wikidata items.

4.1.1. Advantages and limitations of using VIAF IDs for data reconciliation

While the use of VIAF IDs may represent a limitation in the wider applicability of this study for those datasets without VIAF IDs, this study provides further evidence of the value of Wikidata as an authority linking hub, as well as providing a workflow for future SDBM contributions to Wikidata.

In addition to the goal of leveraging structured information found only in Wikidata for named entities recorded in the SDBM, a guiding principle of this case study was to add SDBM IDs to Wikidata in a responsible way that would reduce the introduction of erroneous data to the Wikidata environment (i.e., adding SDBM IDs to the "wrong" entities in Wikidata, or duplicating existing entities). The presence of VIAF IDs served as a main data point that aided reconciliation and acted as a guardrail for uploading SDBM IDs to mitigate the possibility of contributing misleading or duplicative data.

While our approach focused on the use of VIAF IDs, the approach is generalizable for integrating any name authority data into Wikidata by making use of Wikidata as an authority linking hub. VIAF IDs were leveraged because they were present in SDBM, but any other authority IDs represented in both datasets could have been used in the same way. Additionally, as a future research trajectory, SDBM authority records without VIAF IDs could be reconciled against Wikidata items using names/labels and other data points. The generalizable value, we believe, is in 1) enriching Wikidata with additional external IDs, and thus contributing to its growing role as an authority linking hub and 2) leveraging data about named entities present in Wikidata, such as genders, occupations, and other associations, which are absent in both SDBM and VIAF.

4.2. Using Wikidata items, statements, and properties for semantic enrichment

Wikidata items are described through two different types of subject-property-value triples: statements and external IDs. Statements act as descriptive metadata about an item. Wikidata items for people, groups, and organizations tend to have statements regarding what type of entity they are (e.g., human, organization), birth or founding dates, death or ending dates, associated places, and other descriptive information. External IDs act as administrative metadata by relating the entity represented in Wikidata to its counterpart in other resources or authority files. These external resources include VIAF, the Library of Congress Name Authority File (LCNAF), the International Standard Name Identifier (ISNI) system, social media account names/handles, and identifiers found in national library authority files and other Open Linked Databases. External ID statements for a Wikidata item are structured using a specific property for that identifier system and the ID value from the authority.

Our enrichment strategy for the authority records was to link SDBM IDs to corresponding Wikidata items within Wikidata. To do this, a special property was needed so that Wikidata items could be described by an external ID for the SDBM. Unlike creation and edits of item pages, property creation is more tightly controlled by the Wikidata platform and its constituent user communities. Property creators are users empowered with the technical ability to create properties and are expected to act responsibly when participating in the process, including commenting on and responding to property proposals. Users can propose a property to be created in any of the property proposal topic groups, such as the group for authority control. The definition, use, and constraints of properties are to be included in the proposal, and members of the designated community discuss the proposal and formally comment on whether the property should be created. Once a consensus has been reached, and the property is considered non-duplicative and appropriate, a property creator formally creates the property, at which time it is assigned a P identifier and is available for structuring statements.

To make our data donation of SDBM IDs to Wikidata, we availed ourselves of the property proposal process. The first author reached out to the Wikidata authority control topic group and a proposal was submitted to create properties for both SDBM name

[7] https://github.com/OpenRefine

and place authority identifiers. The SDBM name (P9756) and place (P9757) ID properties were subsequently approved and created.

4.3. Data reconciliation and Wikidata editing through OpenRefine

Once the test dataset was extracted and loaded into OpenRefine and the properties to support the data donation were created, the Wikidata reconciliation service in OpenRefine was used to match named entities from the SDBM to Wikidata items. A reconciliation service is a web service which uses some text string or label identifying a piece of data to provide a ranked list of potential matches for that data based on given criteria. Matches that are ranked highly enough can be automatically matched, while less high ranking matches can be matched manually, making the entire reconciliation process semi-automated. This project used different combinations of data points by using the name in the SDBM as the main criteria, along with a VIAF ID number, a Wikidata item type (Q5 human for personal names and Q43229 organization for corporate names), and start and end dates. After attempting matching subsets of the test dataset (around 2000 names) using different combinations of the above criteria (Tables 1-10), it was found that the most workable strategy to perform reconciliation to match Wikidata items for both personal (n = 10962) and corporate (n = 2524) names in SDBM, as well as personal names with start and end dates (n = 9779) and corporate names with start and end dates (n = 392), was to use the SDBM name, the corresponding VIAF ID recorded in the SDBM, and the Wikidata item type. Any additional information, like dates, did not seem to improve matching and ranking. Using only names, item types, and VIAF ID also prevented data subsets from needing to be further segmented between those records with dates and those without.

Table 1

SDBM Personal Names Reconciled using Item Type (by Confidence Score)

		Auto-match	100 and over	80 and over	60 and over	40 and over	20 and over
Number of Records	10962	629	833	2450	4748	4886	4899
Percent of Total Records	100%	6.47%	7.79%	22.91%	44.41%	45.7%	45.73%

Table 2

SDBM Personal Names Reconciled using Item Type & VIAF ID (by Confidence Score)

		Auto-match	100 and over	80 and over	60 and over	40 and over	20 and over
Number of Records	10962	8425	8467	8664	8815	8834	8838
Percent of Total Records	100%	78.8%	79.19%	81.03%	82.44%	82.62%	82.66%

Table 3

SDBM Personal Names with Start and End Dates Reconciled using Item Type & Dates of Birth and Death (by Confidence Score)

		Auto-match	100 and over	80 and over	60 and over	40 and over	20 and over
Number of Records	9779	1565	173	2320	3413	4102	4208
Percent of Total Sample	100%	16%	1.77%	23.72%	34.9%	41.95%	43.03%

Table 4

SDBM Personal Names with Start and End Dates Reconciled using Item Type & VIAF ID (by Confidence Score)

		Auto-match	100 and over	80 and over	60 and over	40 and over	20 and over
Number of Records	9779	7829	7842	7973	8091	8105	8108
Percent of Total Sample	100%	80.05%	80.19%	81.53%	82.74%	82.88%	82.91%

Table 5

SDBM Personal Names with Start and End Dates Reconciled using Item Type, VIAF ID, & Dates of Birth and Death (by Confidence Score)

		Auto-match	100 and over	80 and over	60 and over	40 and over	20 and over
Number of Records	9779	7799	7717	7802	7867	8028	8058
Percent of Total Sample	100%	79.75%	78.91%	79.78%	80.45%	82.09%	82.4%

Table 6

SDBM Corporate Names Reconciled using Item Type (by Confidence Score)

		Auto-match	100 and over	80 and over	60 and over	40 and over	20 and over
Number of Records	2524	739	883	1124	1332	1389	1415
Percent of Total Records	100%	29.28%	34.98%	44.53%	52.77%	55.03%	56.06%

Table 7

SDBM Corporate Names Reconciled using Item Type & VIAF ID (by Confidence Score)

		Auto-match	100 and over	80 and over	60 and over	40 and over	20 and over
Number of Records	2524	1224	1243	1360	1462	1492	1501
Percent of Total Sample	100%	48.49%	49.25%	53.88%	57.92%	59.11%	59.47%

Table 8

SDBM Corporate Names with Start and End Dates Reconciled using Item Type & Inception and End Time (by Confidence Score)

		Auto-match	100 and over	80 and over	60 and over	40 and over	20 and over
Number of Records	392	0*	0*	0*	23	135	153
Percent of Total Sample	100%	0%	0%	0%	5.87%	34.44%	39.03%

*highest confidence score was 78

Table 9

SDBM Corporate Names with Start and End Dates Reconciled using Item Type & VIAF ID (by Confidence Score)

		Auto-match	100 and over	80 and over	60 and over	40 and over	20 and over
Number of Records	392	162	162	172	185	191	191
Percent of Total Sample	100%	41.33%	41.33%	43.88%	47.19%	48.72%	48.72%

Table 10

SDBM Corporate Names with Start and End Dates Reconciled using Item Type, VIAF ID, & Inception and End Time (by Confidence Score)

		Auto-match	100 and over	80 and over	60 and over	40 and over	20 and over
Number of Records	392	128	128	128	138	179	189
Percent of Total Sample	100%	32.65%	32.65%	32.65%	35.2%	45.66%	48.21%

Reconciliation assisted by VIAF IDs was iterative due to the nature of IDs in the SDBM records. When a VIAF ID in the SDBM matched the same information for a Wikidata item, the matching score was high enough to automatically reconcile the data without manual confirmation. When inconsistencies occurred in either Wikidata or the SDBM (such as the same VIAF ID being assigned to more than one Wikidata item or an incorrect VIAF ID recorded in the SDBM), errors were corrected or verified before being reconciled manually. Where SDBM records contained deprecated VIAF IDs, OpenRefine was used to retrieve JSON data from VIAF to find the current IDs (using a general expression[8] to render VIAF IDs into URLs using the General Refine Expression Language). This was possible because old IDs/URIs in VIAF redirect to current records. Retrieved JSON

[8] "https://viaf.org/viaf/"+value +"/viaf.json"

data was parsed[9] in OpenRefine to isolate the current VIAF ID, which was then used as part of the criteria for reconciliation.

The Wikidata reconciliation service was run using CSV files with groups of approximately 2,000 names at a time. Once a CSV spreadsheet was reconciled and checked, an extension to take tabular data from OpenRefine and render it into Wikidata statements was used. This involved creation of a schema to align the QIDs for the Wikidata items matched in reconciliation as the statement subjects, the SDBM name ID property as the property in the statement triple, and the corresponding SDBM ID as the statement value. With the schema to structure statements, OpenRefine was then used to automate edits of Wikidata item pages with statements adding the SDBM IDs as external identifiers.

5. Results

This section describes the results of the data donation and reconciliation process and their evaluation through SPARQL queries. It also describes steps taken to initiate a community of practice around Wikidata as a LOD hub for manuscript scholars interested in name authorities as smart data.

5.1. Enriched dataset and SPARQL query exploration and evaluation

Out of the test dataset of 13,486 names with VIAF IDs, 10,370 SDBM name IDs were matched and added to Wikidata (Table 11). The result was an enriched dataset in which information from Wikidata could be extracted and presented alongside SDBM ID data. It also afforded the ability to query LOD in Wikidata about entities in the SDBM and to combine data from both datasets using their matching identifiers. To evaluate the possibilities of this expanded and unified access, the research team enlisted other staff at Penn Libraries as well as digital humanists with an interest in manuscript studies to explore the results of the data reconciliation and linking.

[9] Using the GREL: value.parseJson().redirect.directto

Table 11

SDBM Name IDs Reconciled and Contributed to Wikidata

	Records with VIAF IDs	Semi-Automated Match with Recorded VIAF IDs	Semi-Automated Match with Alternate VIAF IDs	SDBM IDs Uploaded to Wikidata
Personal Names	10962	8696	347	9043
Corporate Names	2524	1274	53	1327
Total	13486	9970	400	10370
Percent	100%	73.93%	2.97%	76.89%

A series of SPARQL queries (Appendix B) were generated for two types of questions. The first type were three data-related questions to examine the connections made between the two databases and the relationships of SDBM data to other external authorities. The first query generated a list of SDBM names for human entities, found in Wikidata through instances of (P31) the type human (Q5). The second query generated the converse: any SDBM-linked names which were not instances of Q5 human. This second query was also expanded to show what non-human names were instances of, such as abbeys (Q160742), private universities (Q902104), and national museums (Q17431399). The third query was flexibly constructed to discover which SDBM-linked names also appeared in another given authority file, such as the Library of Congress authorities (P244), the ISNI (P213) database, the Getty's Union List of Artist Names (ULAN) (P245), and the French (P268) and German (P227) national libraries, among others. Identifiers from those authorities were also included in the query. These queries could be constructed using entities and properties from Wikidata and generated results which could be examined by the team for accuracy.

The second group of SPARQL queries involved research-related questions regarding attributes of the SDBM-linked entities. Query 4 produced a graph database of the children (P40) and grandchildren (i.e., children of children) of SDBM names. Query 5 generated a list of students (P802). Queries 6 and 7 looked at the intersection of groups of people in the SDBM by overlapping attributes: people described as having the occupation (P106) of lawyers (Q40348) who were also described as collectors (Q3243461) (query 6, assumed for our purposes to include manuscript collectors) and women (Q6581072) described as collectors (query 7). Query 8 identified members of (P463) an organization, specifically the Royal Society (Q123885), and query 9 examined SDBM names (both people and organizations) associated with (P611 property for belonging to a religious order) the Franciscans. All of these queries could be expressed in SPARQL and successfully displayed results for the consulting scholars.

One of the concerns that arose from queries 6-9 was the dearth of names present in the results. After examining the original SDBM test dataset, it was found that while the results were limited, they were accurate given the constraints of the dataset and the databases. These constraints were: 1) names must be present in both the SDBM and Wikidata, and thus linked in Wikidata; 2) names must have had a VIAF ID (i.e., be from the test dataset); 3) the properties and values for retrieved statements must be used correctly; and 4) the subject-property-value statements being queried must be present for corresponding Wikidata items to be retrieved. So, if a known female collector from the SDBM is not linked to Wikidata and not described as either a woman or a collector in Wikidata, the name would not be retrieved. These concerns drove the decision to expand the practical scope of the case study to encourage improvement of manuscript trade-related data in Wikidata.

Finally, query 10 asked to find the gender affiliation of monastic institutions in the SDBM. This query was the only one which could not be answered due to the nature of the Wikidata data model and the properties available to describe monasteries (Q44613) and religious orders (Q2061186). Wikidata did not have a way of describing statements about gender for these entities, rendering the query unsolvable. Instead, Wikidata expresses gender for religious personnel (Q69944008) and various related concepts, like women's monasticism (Q19677296), in ways not directly useful to answering this question. This revealed limitations in the descriptions of these entities in Wikidata which the team hoped to explore at a later date to potentially improve description of religious institutions and gender.

5.2. Toward cultivating a community of LOD practices for premodern manuscripts through Wikidata

The dearth of specific results from some queries indicated to the team that Wikidata could be improved as a hub in a LOD ecosystem through crowdsourced scholarly contributions. The SDBM Name Authority, by design, limits the amount of information about named entities which can be contributed, because of the nature and scope of the database reflected in its data model. The nature of the SDBM infrastructure means that semantic enrichment for the types of data scholars were interested in (e.g., gender, occupation, association) is only currently possible unidirectionally through the contribution of SDBM IDs (and other information in the SDBM) to Wikidata, as the SDBM does not support integration of additional properties to describe name class entities.

This means that scholars are currently able to query data about SDBM-represented named entities in Wikidata for which IDs have been contributed but are limited in querying SDBM data in concert with Wikidata data. This limitation has spurred thoughts about what a future redevelopment of the SDBM might look like to better integrate Wikidata into its infrastructure.

In the meantime, administrators of the SDBM can avail themselves of data present in Wikidata such as VIAF IDs or date and place data present for SDBM-aligned Wikidata items to update or correct errors in SDBM Authority records where possible. Alignment of the SDBM with Wikidata also can help alert administrators to the presence of duplicate records in the SDBM, where reconciliation matches more than one SDBM Authority record with a single Wikidata item.

Despite its limitations, the SDBM contains a wealth of data in the form of names and associated information not found in any other database or resource, particularly regarding individuals traditionally underrepresented in manuscript studies. These two factors encouraged our advancement of Wikidata as the central linking infrastructure in a web of data about people and organizations involved in manuscript production and trade. First, the successful results of the donation to Wikidata revealed the potential for further exposure and access of the SDBM to a larger audience. Wikidata is growing as a LOD hub as more projects connect to it, putting the SDBM in direct contact with many other LOD databases and repositories [27]. Second, when scholars have data beyond the scope of the SDBM, they are now empowered to contribute to Wikidata: by creating Wikidata items, linking SDBM IDs to existing items, and providing additional information derived from their scholarly work.

To that end, the team developed a series of recommendations and tools to help scholars structure their research information for input into Wikidata. Workshops and training sessions were held to show how Wikidata items can be created and data can be input for items, both through manual entries and through automated tools like OpenRefine. To improve data in Wikidata items for SDBM-linked names, metadata maps[10] were developed to crosswalk data structured in the SDBM data model with corresponding Wikidata properties and expected types for values. This would allow the authoritative data in SDBM to be properly structured using the Wikidata data model. To improve the representation of underrepresented people and organizations related to manuscripts in Wikidata, an application profile was developed using the Wikidata properties and expected value types that would match information manuscript scholars would wish to add to Wikidata. Although work has been done on how to represent manuscripts using Wikidata [19,33], and lists of general properties applied to people and organizations exist, a schema of Wikidata properties and value ranges for representing metadata about these entities has not been isolated for use by specific domain needs, such as manuscript studies. The ongoing plan is to encourage collaboration using these methods to help humanities scholars familiarize themselves with LOD structures of Wikidata and to work with and within the Wikidata community to represent their data in a LOD environment. This would include additional property creation requests, refinements in definitions for entities as applied to manuscript studies, and the creation of new Wikidata items for manuscript-related entities and concepts.

6. Conclusions and implications

This research demonstrated a successful test case for contributing and linking a small, smart dataset of LOD IDs to Wikidata to improve access and discov-

[10] https://docs.google.com/spreadsheets/d/1p0eYDsF84obhVGP4_JLFRrscdT9rMTfQuJWyGT11nxk/edit?usp=sharing
(please note that this resource is actively being used and updated and is not a static document)

ery to contextual information for premodern manuscript studies. By linking the well-structured and verified data about people and organizations in the SDBM and making it "smarter" through semantic enrichment, this research led to the development of a workflow to reconcile data values and edit Wikidata pages which can be used to make future contributions of SDBM Name Authority IDs. This work also confirmed the research benefits of Wikidata and integrated technologies like OpenRefine to smart data and semantic enrichment projects in the digital humanities.

While this case study only contributed names with existing digital records (i.e., VIAF), it nevertheless illustrates the value of continued contribution of names to a LOD repository as means for expanding access and discovery to name records underrepresented in traditional LAM name authorities. The contribution of SDBM names to Wikidata presented here can serve as a model for other related projects to perform the same work. The workflow is adaptable and the lessons learned could be applied to building a larger community of practice among project developers interested in creating and sharing metadata for names involved in the production and trade of premodern manuscripts.

7. References

[1] boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662-679. https://doi.org/10.1080/1369118X.2012.678878

[2] Candela, G., Escobar, P., Carrasco, R. C., & Marco-Such, M. (2019). A linked open data framework to enhance the discoverability and impact of culture heritage. Journal of Information Science, 45(6), 756-766. https://doi.org/10.1177/0165551518812658

[3] Chen, S.-J. (2019). Semantic enrichment of linked archival materials. Knowledge Organization, 46(7), 530-547. https://doi.org/10.5771/0943-7444-2019-7-530

[4] Chen, S.-J. (2019). Semantic enrichment of linked personal authority data: A case study of elites in late Imperial China. Knowledge Organization 46(8), 607-614. https://doi.org/10.5771/0943-7444-2019-8-607

[5] Coladangelo, L. P., Thomson, E., & Ransom, L. (2023). Leveraging the power of crowdsourcing and linked open data: Transformation of the Schoenberg Database of Manuscripts and the SDBM Name and Place Authorities. Journal of Library Metadata, 1-22. https://doi.org/10.1080/19386389.2023.2168120

[6] Cooey, N. (2019). Leveraging Wikidata to enhance authority records in the EHRI Portal. Journal of Library Metadata, 19(1-2), 83-98. https://doi.org/10.1080/19386389.2019.1589700

[7] Dragoni, M., Cabrio, E., Tonelli, S., & Villata, S. (2016). Enriching a small artwork collection through semantic linking. In H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, S. Ponzetto, & C. Lange (Eds.), The Semantic Web, Latest Advances and New Domains, ESWC 2016. (pp. 724-740). Springer. https://doi.org/10.1007/978-3-319-34129-3_44

[8] Gracy, K. F. (2018). Enriching and enhancing moving images with linked data: An exploration in the alignment of metadata models. Journal of Documentation, 74(2), 354-371. https://doi.org/10.1108/JD-07-2017-0106

[9] Höper, J., & Müller-Birn, C. (2018). Assisting in semantic enrichment of scholarly resources by connecting neonion and Wikidata. https://refubium.fu-berlin.de/handle/fub188/22790

[10] Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., & Keravuori, K. (2019). Linked data—A paradigm change for publishing and using biography collections on the Semantic Web. In Proceedings of the Third Conference on Biographical Data in a Digital World, 5-6 September 2019, Varna, Bulgaria. https://seco.cs.aalto.fi/publications/2019/hyvonen-et-al-bs-2019b.pdf

[11] Isaac, A., Manguinhas, H., Stiller, J., & Charles, V. (2015). Report on enrichment and evaluation. Europeana Task Force on Enrichment and Evaluation. http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Enrichment_Evaluation/FinalReport_EnrichmentEvaluation_102015.pdf

[12] Jones, E., & Seikel, M. (Eds.). (2016). Linked data for cultural heritage. ALA Editions.

[13] Larson, E. (2020). Big questions: Digital preservation of big data in government. The American Archivist, 83(1), 5-20. https://doi.org/10.17723/0360-9081-83.1.5

[14] Miller, M. (2019, May 22). Integrating Wikidata at the Library of Congress. The Signal: Digital Happenings at the Library of Congress. https://blogs.loc.gov/thesignal/2019/05/integrating-wikidata-at-the-library-of-congress/

[15] Neubert, J. (2017). Wikidata as a linking hub for knowledge organization systems? Integrating an authority mapping into Wikidata and learning lessons for KOS mappings. In Proceedings of the 17th European Networked Knowledge Organization Systems Workshop (pp. 14-25). https://ceur-ws.org/Vol-1937/paper2.pdf

[16] Pattuelli, M. C. (2012). Personal name vocabularies as linked open data: A case study of jazz artist names. Journal of Information Science, 38(6), 558-565. https://doi.org/10.1177/0165551512455989

[17] Pattuelli, M. C., Hwang, K., & Miller, M. (2017). Accidental discovery, intentional inquiry: Leveraging linked data to uncover the women of jazz. Digital Scholarship in the Humanities, 32(4), 918-924. https://doi.org/10.1093/llc/fqw047

[18] Piscopo, A., & Simperl, E. (2018). Who models the world? Collaborative ontology creation and user roles in Wikidata. Proceedings of the ACM on Human-Computer Interaction, 2, 1-18. https://dl.acm.org/doi/10.1145/3274410

[19] Poulter, M. L. (2021, October 14). Manuscripts on Wikidata: A state of the art?. Medium. https://medium.com/@infobomb/manuscripts-on-wikidatathe-state-of-the-art-7aeab63e0d56

[20] Röpert, D., Reimeier, F., Holetschek, J., & Güntsch, A. (2019). Semantic annotation of botanical collection data. Biodiversity Information Science and Standards, 3: e36187. https://doi.org/10.3897/biss.3.36187

[21] Schöch, C. (2013). Big? Smart? Clean? Messy? Data in the humanities. Journal of Digital Humanities, 2(3). http://journalofdigitalhumanities.org/2-3/big-smart-cleanmessy-data-in-the-humanities/

[22] Simou, N., Chortaras, A., Stamou, G., & Kollias, S. (2017). Enriching and publishing cultural heritage as linked open data. In M. Ioannides, N. Magnenat-Thalmann, & G. Papagiannakis (Eds.), Mixed reality and gamification for cultural heritage (pp. 201-223). Springer. https://doi.org/10.1007/978-3-319-49607-8_7

[23] Smith, S. (2015, December 15). Designing a Wikidata project: Mapping Black Philly art. Scholars Studio Blog. https://sites.temple.edu/tudsc/2021/12/15/designing-wikidata/

[24] Smith-Yoshimura, K. (2018, August 6). The rise of Wikidata as a linked data source. Hanging Together: The OCLC Research Blog. http://hangingtogether.org/?p=6775

[25] Szekely, P., Knoblock, C. A., Yang, F., Fink, E. E., Gupta, S., Allen, R., & Goodlander, G. (2014). Publishing the data of the Smithsonian American Art Museum to the linked data cloud. International Journal of Humanities and Arts Computing, 8, 152-166. https://doi.org/10.3366/ijhac.2014.0104

[26] Tan, X., Luo, X., Wang, X., Wang, H., & Hou, X. (2021). Representation and display of digital images of cultural heritage: A semantic enrichment approach. Knowledge Organization, 48(3), 231-247. https://doi.org/10.5771/0943-7444-2021-3-231

[27] Thalhath, N., Nagamori, M., Sakaguchi, T., & Sugimoto, S. (2021). Wikidata centric vocabularies and URIs for linking data in Semantic Web driven digital curation. In E. Garoufallou & M.-A. Ovalle-Perandones (Eds.), Metadata and Semantic Research, 14th International Conference, MTSR 2020, Madrid, Spain, December 2-4, 2020, revised selected papers (pp. 336-344). Springer. https://doi.org/10.1007/978-3-030-71903-6_31

[28] Tharani, K. (2021). Much more than a mere technology: A systematic review of Wikidata in libraries. The Journal of Academic Librarianship, 47(2), 102326. https://doi.org/10.1016/j.acalib.2021.102326

[29] Wang, X., Tan, X., Gui, H., & Song, N. (2021). A semantic enrichment approach to linking and enhancing Dunhuang cultural heritage data. In K. Golub & Y.-H. Liu (Eds.), Information and knowledge organisation in digital humanities (pp. 87-105). Routledge. https://doi.org/10.4324/9781003131816-5

[30] Wikidata (2022, September 29). WikiProject Authority control. Retrieved May 24, 2023 from https://www.wikidata.org/wiki/Wikidata:WikiProject_Authority_control

[31] Wikidata. (2023, April 16). Statistics. Retrieved May 24, 2023 from https://www.wikidata.org/wiki/Wikidata:Statistics

[32] Wikidata. (2023, April 19). WikiProject Biographical Identifiers. Retrieved May 24, 2023 from https://www.wikidata.org/wiki/Wikidata:WikiProject_Biographical_Identifiers

[33] Wikidata. (2023, May 10). WikiProject Books. Retrieved May 24, 2023 from https://www.wikidata.org/wiki/Wikidata:WikiProject_Books

[34] Zeng, M. L. (2017). Smart data for digital humanities. Journal of Data and Information Science, 2(1), 1-12. https://doi.org/10.1515/jdis-2017-0001

[35] Zeng, M. L. (2019). Semantic enrichment for enhancing LAM data and supporting digital humanities. El profesional de la información, 28(1), e280103. https://doi.org/10.3145/epi.2019.ene.03

[36] Zhao, F. (2022). A systematic review of Wikidata in Digital Humanities projects. Digital Scholarship in the Humanities, fqac083. https://doi.org/10.1093/llc/fqac083

Appendix A: SDBM SPARQL Query for Dataset

Specified endpoint: https://sdbm.library.upenn.edu/sparql/sdbm/query

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX sdbm: <https://sdbm.library.upenn.edu/>

SELECT ?entity ?sdbmID ?name ?viafID ?startDate ?endDate ?type
WHERE {
  ?entity sdbm:names_id ?sdbmID .
  ?entity sdbm:names_name ?name .
  ?entity sdbm:names_viaf_id ?viafID .
  OPTIONAL {?entity sdbm:names_startdate ?startDate .}
  OPTIONAL {?entity sdbm:names_enddate ?endDate .}
  ?entity sdbm:names_subtype ?type .
  FILTER NOT EXISTS {?entity sdbm:names_deleted "true"^^xsd:boolean}
}
```

Appendix B: Wikidata SPARQL Query Links

Query 1

Human entities in SDBM linked to Wikidata (included with gender and occupation, etc.): https://w.wiki/598p

Query 2

Non-human entities in SDBM linked to Wikidata (with type): https://w.wiki/598r

Query 3

Names represented in other authorities (flexible/editable SPARQL query): https://w.wiki/598s

Query 4

Family relationships (example SPARQL query for children and grandchildren): https://w.wiki/598t

Query 5

Student relationships (example SPARQL query for students of SDBM linked names): https://w.wiki/598v

Query 6

Collectors and lawyers: https://w.wiki/598w

Query 7

Female collectors: https://w.wiki/598x

Query 8

Members of the Royal Society: https://w.wiki/598o

Query 9

Names associated with Franciscan Order: https://w.wiki/598y

Query 10

Monastic institutions example queries: monasteries: https://w.wiki/598$; religious orders: https://w.wiki/5992