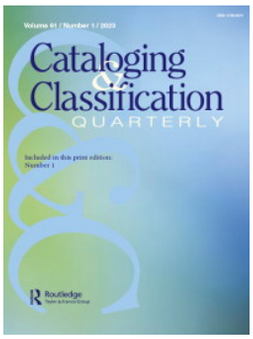

Routledge
Taylor & Francis Group

Cataloging & Classification Quarterly

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/wccq20

Entity Management Using Wikidata for Cultural Heritage Information

Lihong Zhu, Amanda Xu, Sai Deng, Greta Heng & Xiaoli Li

To cite this article: Lihong Zhu, Amanda Xu, Sai Deng, Greta Heng & Xiaoli Li (2023) Entity Management Using Wikidata for Cultural Heritage Information, Cataloging & Classification Quarterly, 61:1, 20-46, DOI: 10.1080/01639374.2023.2188338

To link to this article: https://doi.org/10.1080/01639374.2023.2188338

Published online: 17 Mar 2023.

CATALOGING & CLASSIFICATION QUARTERLY
2023, VOL. 61, NO. 1, 20–46
https://doi.org/10.1080/01639374.2023.2188338

Routledge
Taylor & Francis Group

Entity Management Using Wikidata for Cultural Heritage Information

Lihong Zhu[a], Amanda Xu[b], Sai Deng[c], Greta Heng[d], and Xiaoli Li[e]

[a]Washington State University, Pullman, WA, USA; [b]National Agricultural Library, Beltsville, MD, USA; [c]University of Central Florida, Orlando, FL, USA; [d]San Diego State University, San Diego, CA, USA; [e]University of California, Davis, CA, USA

ABSTRACT

Entity management in a Linked Open Data (LOD) environment is a process of associating a unique, persistent, and dereferenceable Uniform Resource Identifier (URI) with a single entity. It allows data from various sources to be reused and connected to the Web. It can help improve data quality and enable more efficient workflows. This article describes a semi-automated entity management project conducted by the "Wikidata: WikiProject Chinese Culture and Heritage Group," explores the challenges and opportunities in describing Chinese women poets and historical places in Wikidata, the largest crowdsourcing LOD platform in the world, and discusses lessons learned and future opportunities.

ARTICLE HISTORY
Received September 2022
Revised December 2022
Accepted March 2023

KEYWORDS
Wikidata; cultural heritage information; entity management; Linked Open Data (LOD); authority control

Introduction

Traditionally, library data was designed for use and consumption by humans and has been rarely integrated with the Semantic Web. This has caused low interoperability between library data and external sources, and thus hindered the dissemination of library resources in many sectors, especially for the cultural heritage (CH) sector. Due to the tradition of applying local cataloging standards and creating local repositories in the CH sector, data in the CH sector has been largely locked down inside individual institutions. Unless users take extra effort to visit institutional databases and websites, CH data is difficult to access and retrieve. With the trend of interdisciplinary engagement and collaboration among different institutions, the need for data sharing and breaking out of data silos is rapidly growing in the CH sector.

In order to improve the discoverability of their resources, libraries have been exploring and experimenting with options of opening up their

CONTACT Lihong Zhu lzhu2@wsu.edu Washington State University Libraries P.O. Box 645610, Pullman, WA 99164-5610, USA.

bibliographic silos by adopting Linked Open Data (LOD). For cataloging and metadata professionals, one big change brought by LOD is the shift from string to things–to transition "authority control work from a focus on creating authorized headings (i.e., "strings" of text) to minting identifiers (or URIs or "things")."[1] This change means that the focus of creating access points or headings for names and subjects should be shifted from creation of unique text strings to creation of a representation of the entity themselves (work, person, corporate body, place, event, etc.) using a Uniform Resource Identifier (URI) for a single entity. As OCLC mentioned in their Shared Entity Management Infrastructure project update in 2020, compared with traditional cataloging practice, entity management can link resources either inside or outside of libraries, provide richer context by connecting materials and collections, have higher data quality, allow machine to manipulate and use data, and ensure consistent and efficient metadata workflows.[2]

For catalogers working in the CH sector, they usually need to do entity management for local, regional, or "lesser-known" entities. While there are some LOD vocabularies and data models that are explicitly designed for certain areas in the CH sector, such as the Getty Thesaurus of Geographic Names Online[3] and the Europeana Data Model,[4] applying entity management in the CH sector is still challenging. As CH data can be interpreted differently in different cultures, a fundamental semantic problem is faced by CH researchers, which is how to make the heterogeneous contents/entities semantically interoperable and harmoniously presented across cultural boundaries.[5] To address this issue, one possible solution discussed in this article is to use Wikidata as a flexible entity management platform.

Launched on October 29, 2012, Wikidata is a free, open, collaborative, and multilingual knowledge base of structured linked data, serving as a hub for linking resources. The Wikidata data model is matched to Resource Description Framework (RDF), "as most data is encoded via an item (rdf:subject), a property (rdf:predicate) and a value for that property (rdf:object)."[6] In Wikidata, an item is notable if and only if it meets at least one of these three criteria: (1) The item should contain at least one valid link to a page on a Wikimedia Project; (2) The item should represent an instance of a clearly identifiable conceptual or material entity; (3) The item should also fulfill a structural need.[7] Each item is uniquely identified by a Wikidata identifier (starting with a Q prefix). The data for each item is added mostly in the form of statements in RDF format.[8] Items can and should be linked to each other. A link to another entity is usually added as a "property" by clicking on "edit" in the "statements" section. Each property is uniquely identified by a Wikidata identifier (starting with a P prefix).[9]

Since 2012, more and more users have been experimenting with using Wikidata as an entity management platform to create Wikidata items for different entities including persons, corporate bodies, conferences, works, events, concepts, and places. As the largest crowdsourcing LOD platform in the world, Wikidata, now containing over 99 million items,[10] can be a powerful tool for the CH sector to do entity management in a LOD environment. The possibility of linking and leveraging library data using Wikidata and its semantic ontology support in the creation of rich semantic metadata in a multilingual knowledge base has made Wikidata one of the feasible solutions to address library data silo issues and the semantic problems in entity management in the CH sector. As stated in the 2016 IFLA report, Wikidata has the potential to take advantage of linked data by using different ontologies and taxonomies in different languages to support researchers around the world.[11] In addition, the built-in SPARQL query service in Wikidata provides a SPARQL endpoint, including a Graphical User Interface (GUI) which allows people to query against the Wikidata set, and the query results can be rendered for visualization in meaningful ways. OCLC conducted linked data surveys in 2014, 2015 and 2018. According to its 2018 survey, the biggest change was the surge in consuming Wikidata, four times more than that in the linked data implementations in 2015.[12]

Inspired by the PCC (Program for Cooperative Cataloging) Wikidata Pilot Project[13] and LD4 Wikidata Affinity Group Calls,[14] in October 2020, several Chinese American librarians from different institutions formed the "Wikidata: WikiProject Chinese Culture and Heritage Group" in hopes of promoting Chinese cultural heritage, expanding their horizons in library LOD implementation and seeking collaboration opportunities. The Group has undertaken multiple projects and presented their initial findings about characteristics of Wikidata related to Chinese cultural heritage, challenges in creating and enhancing Wikidata items, data models for Chinese women poets and historical places, and different visualizations using Wikidata SPARQL Query Service. Driven by the passion for Chinese culture, the Group aimed to contribute more Wikidata items related to Chinese cultural heritage. The Group has been working on three related Wikidata projects:

1. As the initial learning experience, the Group chose to start with the Great Prose Masters of the Tang and Song dynasties (唐宋八大家) project due to its small scale. The Great Prose Masters of the Tang and Song dynasties refer to the eight most renowned prose writers during the Tang and Song dynasties: Han Yu (768–824), Liu Zongyuan (773–819), Ouyang Xiu (1007–1072), Su Xun (1009–1066), Su Shi (1037–1101), Su Zhe (1039–1112), Wang Anshi (1021–1086), and Zeng Gong (1019–1083).

2. Realizing that Wikidata had limited information about Chinese women poets, the Group decided to focus on creating and enhancing Wikidata items for Chinese women poets as its second project. This will help the diversification of Wikidata in terms of cultural heritage. The Group also wanted to contribute to WikiProject "Women in Red" which aims to "create and improve Wikidata items related to women, women's works, and women's issues."[15]
3. The third project was to create and enhance Wikidata items for Chinese historical places that were the birthplaces for Chinese women poets.

Through those three projects, the Group focused on person and place entities, and investigated the following research questions:

1. Is Wikidata an effective platform for entity management of Chinese cultural heritage information?
2. What are the challenges and opportunities for presenting Chinese prose masters, poets and historical places in Wikidata?
3. What benefits does Wikidata bring in the discovery and dissemination of Chinese cultural heritage information?

This article will present the Group's research findings, discuss challenges and lessons learned, and explore future opportunities.

Literature review

Linked open data

Linked Data is a set of best practices for publishing and connecting structured data on the Semantic Web. To support Linked Data, we need four essential technologies: (1) Uniform Resource Identifiers (URIs); (2) Hypertext Transfer Protocol (HTTP); (3) Resource Description Framework (RDF); (4) RDF Query Language (SPARQL) and SPARQL Protocol. It is also critical to observe the four key linked data principles: (1) Use URIs as names for things; (2) Use HTTP URIs so that people can look up those names; (3) When someone looks up a URI, provide useful RDF information using the standards (RDF, SPARQL); (4) Include links to other URIs so that users can discover more things. Linked Open Data (LOD) is linked data released under an open license.[16]

In order to increase the value and discoverability of library data, libraries have been exploring options for making their data available and useful outside of the data silos of the library world through the adoption of LOD. There are already some projects in the CH sector that took advantage of LOD and Semantic Web technologies, such as CulturaSampo,[17]

DigiCULT[18] and CASPAR.[19] Sturgeon introduced a crowdsourced system in 2021 to efficiently extract historical data from the widely used full-text digital library of classical Chinese works,[20] available from the Chinese Text Project.[21] This system uses semantic annotations of entity reference within texts, and a knowledge graph recording data about these entities and their relationships to one another. Nonetheless, in 2012, Vavliakis, Karagiannis, and Mitkas[22] noted that the mainstream use of LOD and Semantic Web technologies had not yet been achieved in the CH sector. The massively heterogeneous yet semantically interlinked CH resources still posed challenges for librarians and archivists wanting to adopt LOD and Semantic Web technologies. Both Makela, Hyvonen & Ruotsalo[23] and Oomen & Aroyo[24] have noted personalized ontology issues. Data coming from different sources can be interpreted in various ways, which requires semantic inference to maintain or resolve conflicting information, and the creation of open and clear reviewing procedures and annotation rules. In addition, LOD and Semantic Web technologies still have much room for improvement in such areas as dealing with complex underlying knowledge, deriving fuzzy inference rules and probabilistic description logics, reasoning over dynamically evolving data, and providing scalable and robust solutions.[25] Other challenges of implementing LOD and Semantic Web technologies discussed in the CH sector include copyright and ownership issues, diversity and equity concerns, and linguistic techniques.

Entity management in linked open data environment

Entity management or resources management is not a new concept in libraries. The long-existing bibliographic and authority records can be considered as resource descriptions or entity descriptions. Catalogers have been managing entities using text strings and flat structured data in the bibliographic records. Functional Requirements for Bibliographic Records (FRBR), which was developed in the 1990s, is an entity-relationship model of metadata for information objects and resources in libraries. This model includes four-level entities (work, expression, manifestation, and item) to describe library resources and define the relationships between them. Since the launch of FRBR, the entity-relationship concept was expanded from bibliographic data to authority data (Functional Requirements for Authority Data (FRAD)) in 2009 and subject authority data (Functional Requirements for Subject Authority Data (FRSAD)) in 2010. The birth of FRBR introduced a shift of cataloging focus from the construction of flat and aggregated data to the creation of entity-focused, interlinked, and disaggregated data. A book then is not just a book. It becomes a complex set of entities that reflect the meaning, expression, and physicality of a resource.[26]

In the LOD environment, entity management is a process of associating a unique, persistent, and dereferenceable URI with a single entity. To transform traditional library data to LOD, it is crucial for libraries to begin the shift from authority control based on strings to entity management using URIs. Godby & Smith-Yoshimura[27] recommended that the libraries should make progress on three goals: (1) Define entities and relationships that are important to the library community; (2) Use the best features of MARC and other library standards to help with the transformation; and (3) Replace text with identifiers that may originate from librarianship but conform to Linked Data conventions. In 2020, OCLC was awarded a grant from Andrew W. Mellon Foundation to develop a shared entity management infrastructure called "WorldCat Entities" to support Linked Data management. Before putting linked data into everyday cataloging practice, libraries need to build a foundation of reliable and persistent identifiers and metadata for entities.[28] Although WorldCat Entities has not yet provided tools to upload data to Wikidata or vice versa, it has already incorporated Wikidata identifiers when appropriate into its entities.

Wikidata

Wikidata has been a research topic in many disciplines since 2012. In recent years, Wikidata research has chiefly focused on empirical studies in the areas of knowledge organization, languages, collaboration, applications in natural language processing, data quality and validation, information retrieval, knowledge integration, and cross-domain studies.[29] Van Veen[30] proposed to "make a transition from a multitude of different identifiers to using a single, universal identifier for all relevant named entities, in the form of the Wikidata identifier." Biswas[31] explored the possibilities of Wikidata for serials cataloging community. Amaral et al.[32] assessed the quality of sources in Wikidata across languages. Kaffee et al.[33] noted that "Multilinguality is an important topic for knowledge bases, especially Wikidata, that was built to serve the multilingual requirements of an international community. Its labels are the way for humans to interact with the data." In late July 2020, OCLC Research Library Partnership convened a discussion on Wikidata, Wikibase and the library linked data ecosystem. One of the themes of this discussion was that "Entity management is seen as a pathway to engagement with and access to digital collections."[34] Stinson et al.[35] encouraged the GLAM (Galleries, Libraries, Archives, and Museums) community to start paying attention to Wikidata since "Wikidata can connect other databases and collections of information, allowing computers and software to see connections between hundreds of data sources."

In recent years, the GLAM community started to see more efforts to use Wikidata as a platform for entity management in order to facilitate and promote the creation, enhancement, and ingestion of cultural heritage data. The LD4 Wikidata Affinity Group was established in 2019 to provide a welcoming, collaborative, and supportive space to discuss Wikidata related topics with "the goal of understanding how the library can contribute to and leverage Wikidata as a platform for publishing, linking, and enriching library linked data."[36] In 2020, PCC started a Wikidata Pilot which aimed to enable pilot participating institutions to be part of the Wikidata community and learn how they can incorporate Wikidata into their authority work.[37] As part of the PCC Wikidata Pilot Project, Smithsonian Libraries and Archives established the Chinese Ancestor Portrait Project (CAPP) which created Wikidata items for 90 Chinese ancestor portraits from the collections of the Freer Gallery of Art and Arthur M. Sackler Gallery of the National Museum of Asian Art.[38] It was reported that, for many sitters depicted in the portraits, there were no Wikidata items, and if there were Wikidata items, the data was very sparse. The CAPP team saw the need to learn how Wikidata links different languages. Since Qing court culture included both Chinese and Manchu language names, the CAPP team explored the implications of adding different language values for Wikidata properties.[39] "Wikidata: WikiProject Historical Place" is a WikiProject focusing on discussing recommendations for describing changes of a place over time, which could help the CH sector in creating and enhancing Wikidata items for historical places.[40] "Wikidata: WikiProject CJKV Character" maintains Wikidata items, Wikidata properties, and Wikidata ontology related to CJKV (Chinese, Japanese, Korean, and Vietnamese) characters, which will help the CH sector for identifying CJKV characters.[41] In addition to Wikidata, the CH sector has been exploring the use of Wikimedia[42] for cultural heritage information. The Digital Public Library of America (DPLA)'s Wikimedia Project, launched in 2019, developed and operated "a single pipeline for the many diverse DPLA contributing institutions to contribute open access digital assets to Wikimedia Commons, which makes them available for inclusion in Wikipedia articles, allowing for increased discovery and use of these assets."[43]

Materials and methods

Databases/data sources/tools

Due to the notability[44] threshold in Wikidata**,** it is essential to identify authoritative sources for the Group's projects. Fortunately, there are a vast number of sources on Chinese cultural heritage for researchers and the public. Some of the notable databases include China Biographical Database

(CBDB),[45] Ming Qing Women's Writings (MQWW),[46] and China Historical Geographical Information System (CHGIS).[47] CBDB contained biographical information of about 515,488 individuals primarily from the 7th through the 19th century as of December 2021. The long-time goal of CBDB is to "systematically include all significant biographical material from China's historical record" and to "make the content available free of charge, without restriction, for academic use."[48] The CBDB data has been frequently updated, and new entries have been added for historical figures from various dynasties such as Tang, Five dynasties, Liao, Song, Jin, Yuan, Ming and Qing. MQWW, launched in 2005, is a digital archive and database dedicated to the digitization of collections of writings by women in late imperial China (1368–1911). As of December 2021, MQWW included 421 collections of poetry and other writings by about 5233 women poets and writers, and 2436 male writers who wrote paratexts (prefaces, biographies and postscripts, etc.) or compiled/edited some of the collections.[49] CHGIS contains a series of datasets on the administrative geography of Chinese history. It shows the changes over time for provinces, circuits, prefectures and counties. The CHGIS time series covers the dynastic period from 221 BCE to 1911 CE, and it also contains the nationwide data from 1820 to 1911. Using GIS software such as ArcGIS or MapInfo (or even GoogleEarth), CBDB output can be combined with CHGIS datasets.[50] CBDB has CHGIS ID, which can be used to link geographical names between the two databases. Those three databases have served as the primary sources for the Group's projects.

Research stages

The research of the Group was conducted in three stages. In Stage 1, the Group created data models, and manually enhanced Wikidata items. As a learning initiative, the Group first manually enhanced Wikidata items for the Eight Great Prose Masters of the Tang and Song dynasties due to its small scale. They then decided to focus on creating and enhancing Wikidata items for Chinese women poets due to their limited presence in Wikidata. They also discussed how to tackle the problems of semantic interoperability for CH resources in Wikidata. In Stage 2, they utilized external data from CBDB and MQWW to enhance the existing Wikidata items for Chinese women poets. Because of the presence of CBDB IDs in both MQWW entries and Wikidata items, they were able to address the reconciliation problems that are common for name string matching projects and develop a batch processing workflow. In Stage 3, they explored the creation of Wikidata items for Chinese historical places that were the birthplaces of the women poets that they worked on in Stage 2. A preliminary data model for describing Chinese historical places in Wikidata

was developed. For both Stage 2 and Stage 3, they used Wikibot,[51] a Wikidata API editing software, to automate the processes of creating and enriching Wikidata items for persons and places. They developed Python scripts[52] using Pywikibot,[53] a Python library to interact with Wikidata via a Wikibot.

Findings

Exploring data model for Chinese women poets

The Group's WikiProject data models cover information about which Wikidata properties to use to describe a Wikidata item. To ensure the consistency of data and facilitate batch processing, the Group devised a data model for describing Chinese women poets by selecting appropriate Wikidata properties and deciding on their appropriate property values. In order to create the data model, they studied both Wikidata properties that were pertinent to their topic (Chinese, women, poets) and Wikidata item pages related to Chinese cultural heritage. Based on their study, they categorized the properties in their data model into three groups: (1) Core properties–basic and required; (2) Constant properties–a subset of Core properties with constant values; and (3) Extended properties–optional (added if the information is available). See Appendix 1 for the data model for Chinese women poets. As Appendix 2 shows, this data model includes many data elements used by CBDB and MQWW, which made it easier for the Group to facilitate the reconciliation process that is discussed in the next section.

Using external data to enhance existing Wikidata

Wikidata items can be created/enhanced manually or through a batch process using reliable external data. The Group selected MQWW and CBDB as reliable external data for its projects because of those two databases' rich biographic information created and maintained by reputable universities. MQWW was designed and implemented by the McGill Library Digital Initiatives team, and it features a link for each writer to CBDB hosted at Harvard University.[54] The development of CBDB is a joint project of Fairbank Center for Chinese Studies at Harvard University (费正清中国研究中心), Institute of History and Philology of Academia Sinica (中研院历史语言研究所), and Center for Research on Ancient Chinese History at Peking University (北京大学中国古代史研究中心).[55] Before starting its projects, the Group secured permissions from both MQWW and CBDB project teams to contribute their data to Wikidata. The Group also studied the Wikidata: Data

Import Guide[56] and followed its instructions. The four steps used by the Group to enhance existing Wikidata items for Chinese women poets are outlined below:

At the first step, data was downloaded from MQWW (7,466 poets) and converted to Excel format–this created List A. The file of List A included the following information about each poet: name, gender, ethnic group, marital status, geographic location, and biographies.

The second step was to find which of those 7,466 poets in MQWW already had Wikidata items (Wikidata Q numbers). The Group first tried to use the OpenRefine reconciliation process following the OpenRefine User Manual.[57] They loaded List A into OpenRefine, and ran the reconciliation process. The process produced 4,283 Wikidata item matches and a list of possible matches for the remaining 3,183 poets. They sampled about 100 Wikidata items that were matched and found some false results. Two possible factors might have contributed to this mismatch: 1) Many Wikidata items were brief and did not have sufficient data; 2) Chinese names in Romanized forms were not unique enough to distinguish one person from another. As a result, they decided to take a different approach which is described as the following: First, data was downloaded from CBDB (which contains poet IDs from MQWW)–this created List B. Second, they used a Wikidata SPARQL query to find all Wikidata items that have the value "female" in P21 (sex or gender) and also have CBDB ID (P497)–this created List C. Third, they compared List A and List B, and picked Wikidata items that have Women poet IDs from MQWW in the spreadsheets–this created List D. Fourth, they compared List D and List C, and picked Wikidata items that have Women poet IDs from MQWW in the spreadsheets–this created List E which was a list of those Chinese women poets in MQWW that already had Wikidata items (Wikidata Q numbers).

In the third step, upon further analyzing List E, the Group came up with two more lists: (i) List F–a list of entries that were matched on both Wikidata Q numbers and the name labels from different sources. The entries in List F did not need manual review and were used for batch updating the Wikidata items using Wikibots. (ii) List G–a list of entries that were matched only on poet IDs but not on name labels from different sources. The Group manually reviewed List G. If the entries were actually accurate matches, they were also used for batch updating the Wikidata items using Wikibots.

The fourth step was to use Wikibot to batch update Wikidata items by adding new statements and the constant data outlined in the data model. For details, please see Appendixes 1–3. In the future, the Group will continue exploring different approaches to create, enhance and manage

Wikidata for persons, historical places and notable works both manually and via batch loads.

Describing Chinese historical places in Wikidata

Since CBDB documents the majority of women poets' birthplaces and links them to the geographic database CHGIS, the Group decided to take advantage of the geographic information in CBDB and CHGIS. They chose a small sample of 23 poets' birthplaces that had no Wikidata items, and created Wikidata items for them. See Appendix 3 for a primitive data model for Chinese historical places. This model includes basic description (English and Chinese label, simplified and traditional Chinese label) and core statements (instance of (P31), country (P17), official name (P1448), native label (P1705), coordinate location (P625), located in time zone (P421), and CHGIS ID (P4711)). In Wikidata, instance of (P31) is usually used to describe the administrative level for places. For example, Berlin (Q64) is an instance of the federal capital (Q257391). Champaign (Q577964) is an instance of a city of the United States (Q1093829). Bianliang (Q7223839) is an instance of a capital (Q5119) and also a historical administrative division (Q19832712). Following the existing examples in Wikidata, the Group used P31 (instance of) to indicate the administrative level of those places. They deliberately chose to include only basic descriptions and core properties at first in this preliminary data model; they decided not to include extended statements because they had questions on how to describe places with the same name but different boundaries in various time periods, and how to describe places whose administrative level is unique to China and hard to translate. The concern about those two questions is discussed later in this article.

To collect poets' birthplaces data, the Group used the CBDB API to retrieve the birthplaces of poets who had entries in both CBDB and Wikidata, and selected places that had CHGIS ID and contained coordinates. They used the CHGIS API to gather more information about those places, and reconciled them against Wikidata items using OpenRefine. They then manually verified the matches. For the 23 non-matched places which did not have existing Wikidata items, they used Pywikibot to create new Wikidata items in a batch following the data model. Since CHGIS has place names and historical administrative units for the Chinese dynasties, the Group referred to CHGIS when recording place names. As a literatus could serve as an official and hold positions in various places, several properties were used to record this information, such as place of birth (P19), place of death (P20), residence (P551) and place of burial (P119).

Using Wikidata SPARQL query for discovery and presentation

SPARQL is a RDF query language, that is, a semantic query language for querying data in RDF triple stores.[58] Wikidata has provided a SPARQL endpoint including a powerful Web-GUI since September 2015. Any kind of data can be extracted using SPARQL with a query composed of logical combinations of triples in RDF stores. Anyone can edit and submit Wikidata SPARQL queries to the query engine in Wikidata SPARQL Query Service GUI[59] and retrieve result sets which can be rendered in different views including HTML table view, image grid view, timeline view, map view, and graph view.

Below is a Wikidata SPARQL query that the Group used to retrieve Wikidata items for women poets writing in Chinese whose Wikidata entries also contain their pictures. (The query was run on March 10, 2023):

```
#find poets; female; writing in Chinese
SELECT DISTINCT? person? personLabel? pic? birthPlace? coordinates? birthDate
WHERE
{
?person wdt:P106 wd:Q49757 . # person whose occupation is poet
?person wdt:P21 wd:Q6581072 . # person whose gender is female
?person wdt:P1412 wd:Q7850 . # person who writes in Chinese
?person wdt:P19? birthPlace . # person's birth place
?birthPlace wdt:P625? coordinates . # GPS coordinates of the birth place
?person wdt:P569? birthDate . # person's birth date
?person wdt:P18? pic . # person's picture
SERVICE wikibase:label {bd:serviceParam wikibase:language "en"}
}
```

The query above contains the SELECT clause which lists variables the Group wanted to retrieve uniquely, e.g., the person, person label, picture, birthplace, coordinates, and birth date. The variables start with a question mark.

The next few lines in the query are the WHERE clause that contains restrictions on the variables in the form of triples. There are 7 triples (subject, predicate and object) in the WHERE clause for this query. For example, in the first triple, there is Person (subject), wdt:P106 for occupation (predicate) and wd:Q49757 for poet (object)–this defines what data criteria to be used for retrieval, that is, person whose occupation is poet. For fixed values in the triples in WHERE clause, items are prefixed with "wd:" and properties with "wdt:"

In English, the second triple is to find a person who is a woman. The third triple is to find a person who writes in Chinese. The fourth

triple is to find a person who has a birth place. The fifth triple is to find geographic coordinates of the birthplace. The sixth triple is to find a person with a birth date. The seventh triple is to find a person who has a picture in their Wikidata item. The last line in the query is the SERVICE snippet which is the default statement for using the service in English.[60]

Using this Wikidata SPARQL query, 40 results were retrieved that meet the search criteria for women poets writing in Chinese, including their names, their birthplaces, geographic coordinates of their birth places, their birth dates, and their pictures. The number in the result will grow since the Group will create or enhance more Wikidata items for the persons, birthplaces, and notable works of women poets writing in Chinese.

The Wikidata SPARQL Query Service GUI renders the query result in the default HTML table view and displays the results as a table of values. The user can also choose other views, such as, the ImageGrid view which displays images in the result as a grid, the TimeLine view which displays results having dates, the Map view which displays coordinate points, and the Graph view which displays result as connected graph.[61]

See Figure 1 for the image grid view of the query result for women poets writing in Chinese with pictures in Wikidata.

See Figure 2 for the timeline view of the query results for women poets writing in Chinese with pictures in Wikidata.

See Figure 3 for the map view of the query result for women poets writing in Chinese with pictures in Wikidata.

See Figure 4 for the graph view of the query result for women poets writing in Chinese with pictures in Wikidata.

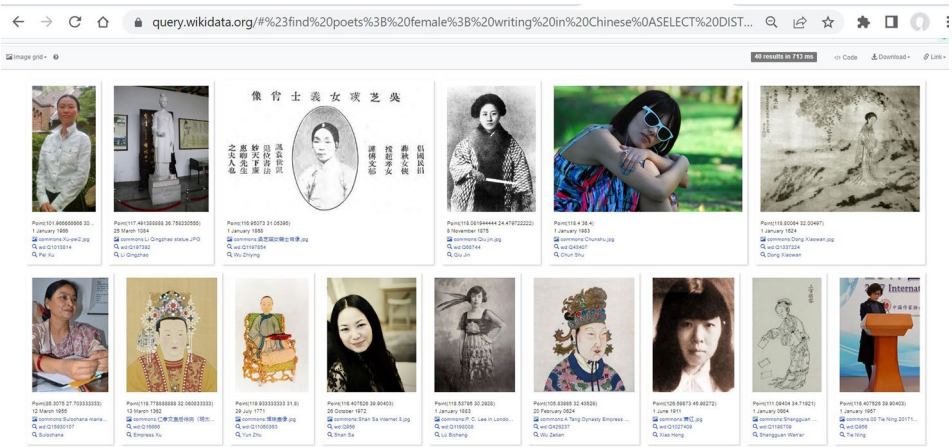

Figure 1. The image grid view of the query result for women poets writing in Chinese with pictures in Wikidata.

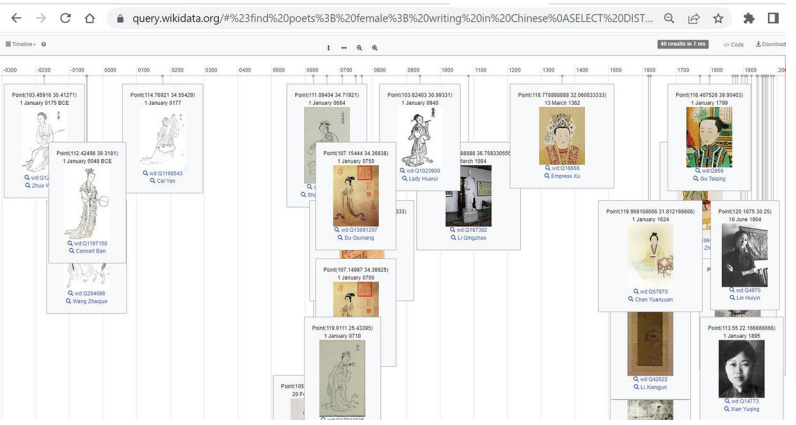

Figure 2. The timeline view of the query results for women poets writing in Chinese with pictures in Wikidata.

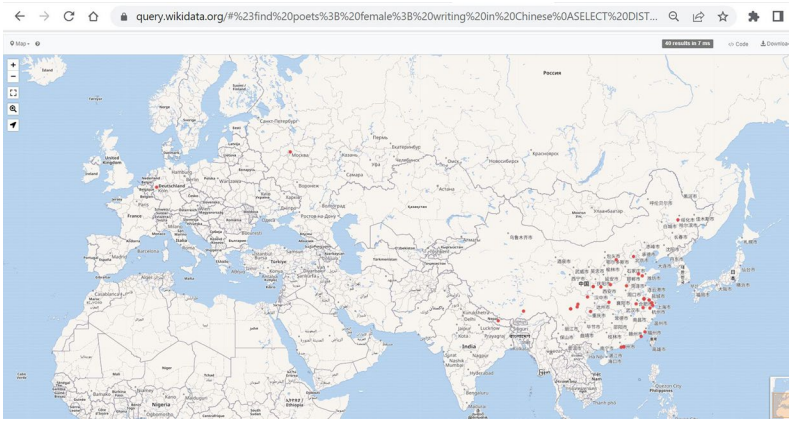

Figure 3. The map view of the query result for women poets writing in Chinese with pictures in Wikidata.

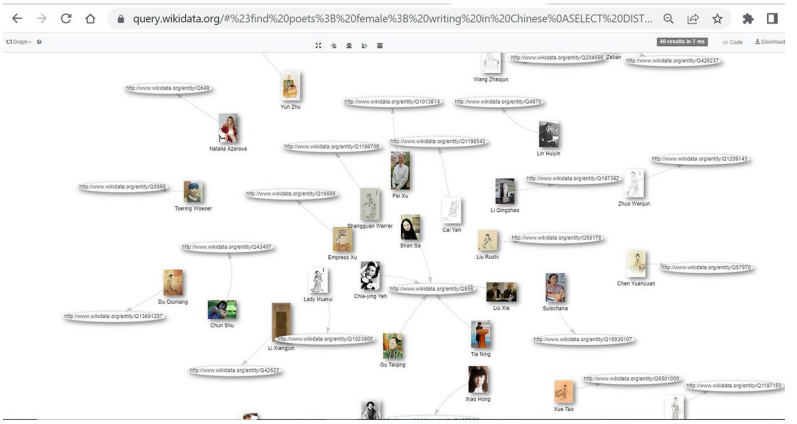

Figure 4. The graph view of the query result for women poets writing in Chinese with pictures in Wikidata.

Discussion

While working on these projects, the Group came across a number of challenges and are still working on solutions. They have also learned a few lessons.

Reconciliation challenges

Inaccurate information in Wikidata

While reconciling poets' names against Wikidata, it was found that a considerable number of poets had incorrect gender information. It was also noticed that some informations conflicted with each other. For example, Pei Shu (Q45625354)[62] was described as a Tang dynasty person but had both Tang (618–906) and Qing (1644–1912) dynasty values of property for country of citizenship (P27). Upon further research, it was verified that Pei Shu was a Chinese poet in the Tang dynasty.[63] Thus, the Qing dynasty should not be listed as a value of P27. This inaccurate information resulted in unexpected extra efforts in reconciliation.

Insufficient information in Wikidata

Insufficient information in Wikidata has increased the difficulty of reconciliation. Some poets' Wikidata entries had only minimal information and no CBDB ID. In some extreme cases, there were only poets' names and P31 (instance of) available in their Wikidata items. For historical places, the Group also came across Wikidata items with just one statement and one Chinese label. For example, Wikdiata item Q60992243[64] only has the Chinese label 海宁州 and the Google Knowledge Graph ID (P2671).

Name changes for historical places

Recording place names has posed reconciliation challenges since place names can change in history. For example, Hua Ting (华亭) was named Hua Ting, Huating Fu, Huating Xian, Huating Shangguan and Huating Xiaguan at different times. It can be difficult to know what the place was called during a person's lifetime, and if the place names in the resources are historical names or modern names.

Missing key elements used for reconciling historical places

Since Wikidata is a crowdsourcing knowledge base, historical places are defined and described in different ways. The Group has not found the key elements that can be used to reconcile Chinese historical places. While Wikidata statements for the Chinese women poets usually include P31 (instance of; value: human) and P21 (sex or gender; value:female), it is hard to find commonly used Wikidata properties for historical places. P31 (instance of) is often used to describe the administrative units, and the

value of this property varies. In addition, information, especially in P17 (country), is not always complete in Wikidata. The name of a historical place could be the same and its boundary is similar for different dynasties, but only one or a few of the dynasties are provided as the value of P17 (country). For example, 钱塘县 (Qiantang County, Q11650262)[65] is an ancient county in the Tang, Ming, and Qing dynasties, but this Wikidata item only has the Qing dynasty as the value in P17 (country). Is it a concern that other dynasty names are not listed? Can Qiantang County in the Qing dynasty be reconciled to this Wikidata entity? The Group is still discussing those questions.

Semantic discussion

Semantic discussion on describing Chinese people

The Chinese language is different from the Western languages, so are the Chinese cultural traditions. For example, for the writing system, Chinese has traditional characters and simplified characters. The traditional characters are typically more complicated with more strokes, and they are more in line with the original ancient characters. In Wikidata, the various names of the prose masters and poets in both traditional and simplified Chinese are included at the "In more languages" section, and can be added to various name properties. The properties P1559 (name in native language) and P1477 (birth name) allow names in Chinese characters to be added with transliterations. These properties are more of generic nature and can be applied to Wikidata entries that are not related to Chinese culture.

There are some properties that are unique to Chinese culture and the East Asian culture sphere, such as P1782 (courtesy name) –in Chinese "字", a name bestowed upon one at adulthood, and P1787 (art name) –in Chinese "号", a professional name used by Chinese artists, poets and writers, and P1786 (posthumous name) –in Chinese "谥号", an honorary name given to royalty, nobles, and others in China after the person's death. Traditionally, Chinese literati, officials and higher classes in ancient times, had courtesy names, art-names as well as posthumous names. For example, Su Shi (苏轼)'s courtesy name is Zizhan (子瞻), his art-name is Dongpo (东坡) and his posthumous name is Wenzhong (文忠). Thus, the Group decided to use traditional Chinese characters to represent the values of these properties. Hanyu Pinyin (Pinyin) is further added to indicate the pronunciation of the Chinese characters in Mandarin. In inputting Chinese characters for the name, such as Courtesy name, Wikidata gives one the option to add Pinyin with tones with the qualifier pinyin transliteration. It also allows one to specify the writing system that these characters belong to. Besides Pinyin, the older Wade-Giles normalization system developed by Thomas Wade in

the mid-19th century is another type of transliteration. Wade-Giles as well as the vernacular spellings of the names cannot be found under specific name properties at present. These spellings can be put under the more generic transliteration though.

In adding information for the time period, era name, and year, it was found that the Chinese express them in different ways. For example, how can "December 19, the 3rd year of Jingyou (景祐) during the reign of Emperor Song Renzong" be expressed in Wikidata? One option is to add both the equivalent date in the Gregorian calendar and the time period property (P2348) with the value of the Northern Song dynasty (Q319460). As for the Era name–Jingyou Era (Q11090134), it looks like it is constrained to be only linked to the Emperor Renzong of Song, but not to anyone who lived in that era.

It was also debated on whether to describe Chinese cultural heritage topics by expressing its cultural uniqueness, or by using a more universal approach. Courtesy name (P1782), art-name (P1787) and posthumous name (P1786) seem to be Asian or Chinese specific. Should the more widely used pseudonym property (P742) also be added? In describing the Chinese poets, for country of citizenship (P27), it looks like the dynasty names rather than China (Q29520) are recorded. From a more contemporary view, the value probably should be China (Q29520); however, the more refined value, dynasty names have been added to this property widely. Also, for the time period (P2348), shall the year range, the dynasty name, or the emperor's era name be recorded? Currently, dynasty names have been used.

Other types of information can be unique to Chinese and thus it is difficult to translate or record, for example, positions. The Group discussed whether to translate the position names more literally or more in line with the universal way. For example, for "大理评事, 签书凤翔府判官," which literally means "Dali (Supreme Court) Judge, Clerk (of) Fengxiang (Prefecture) Magistrate," it might not be ideal to simply translate it as "clerk." The various positions or titles can be difficult to translate. Shall the ancient unique names or the contemporary and universal labels be used? Sometimes there is no modern equivalent to the ancient names. For example, Jinshi (进士) is equivalent to Ph.D in some way, but not exactly; a Jinshi is an imperial scholar in the national civil service examination. Other types of information, such as field of work (P101) and employer (P108), pose similar challenges. Both Hanlin Academy scholar (Q107382552) (翰林学士) and Minister of Rites (Q47175595) (礼部尚书) are very unique Chinese titles. Shall the equivalent universal names be added for these property values? Another interesting phenomenon is that in ancient times, a literatus could be a pantologist who served as a poet, writer, politician, calligrapher, painter, historian, musician, essayist and other roles. These

roles are normally listed under occupation (P106), which can differ from its modern meaning.

Semantic Discussion on describing Chinese historical places

To describe and translate Chinese historical places to English, one has to provide the context of the time range. Each dynasty in China may use its own divisions of administrative units. The same Chinese character of an administrative unit may represent two or more different administrative levels in various time periods. For example, 州 (zhou) in Han dynasty is the first administrative level, whereas in Ming dynasty, 州 (zhou) is the tertiary administrative level. Even in the same time period, one Chinese character of an administrative unit can represent two or more administrative levels. For example, 县 (xian) can be either third or fourth administrative level in different areas in Ming dynasty. Therefore, it is unlikely to find a one-to-one translation for those administrative units from Chinese to English. There needs to be tremendous efforts in consulting or learning Chinese history to provide accurate administrative level information for places.

Concerns about Wikidata ontology

While women in ancient China had their personal names, they were not always recorded. A married woman was referred to by either both her surname and her husband's surname, or just her husband's surname. Shi (氏) is often added at the end of the surname(s), meaning surname or family name. For example, the wife of Qisou Wang (王齐叟), whose maiden name was Shu (舒), was referred to by Wang Shi (王氏) or Wang Shu Shi (王舒氏). When the Group tried to add family names to the selected Chinese women poets, a flag was noticed on their Wikidata pages because of a Wikidata constraint: if a family name (P734) is added for a name entity, a given name (P735) has to be added as well. However, this constraint cannot be applied to those women in ancient China since their given names were not often recorded. In addition, current ontology dictates that the values of P735 (given name) have to be established items in Wikidata and cannot be text strings. This does not work for Chinese given names in general, because unlike most Western given names, Chinese given names are formed by a combination of any one or more Chinese characters, which makes it nearly impossible to put them into an established category.

Other miscellaneous concerns

There are also other concerns that will affect the Wikidata item description, including social, political, cultural, and Diversity, Equity, and Inclusion

(DEI) considerations. For example, how to define Chinese poets? Poets who write in Chinese from all over the world? Poets who write in Chinese and are/were active in Chinese speaking countries? It was noticed that some Wikidata properties do address DEI considerations. For example, for sex or gender (P21), although the Group chose its value for the Chinese Women Poets project to be "female", P21 does allow other terms to be added; the value for P172 (ethnic group) can be Han Chinese people (Q42740) or other terms; and the value for religion or worldview (P140) can be Confucianism (Q9581) or other terms. In addition, some properties require established items as values unless strings are allowed; this has posed new challenges since the Group needed to create new Wikidata items from scratch in order to get the Wikidata identifiers.

Copyright and ownership issues are also involved in this study. Though the metadata from both MQWW and CBDB databases can be downloaded freely, it is unclear if they are out of copyright, under the public domain, or CC0-License. The Group did its best to confirm the copyright and asked for permission to reuse the data. From the reuser's perspective, the presence of an explicit copyright statement and a Creative Commons license may encourage wider reuse of this rich data.

Lessons learned

Establishing data models is a critical step for any WikiProjects so that Wikidata items can be consistently created and enhanced. It is important to define the property level, such as Core (the bare minimum), Constant (the bare minimum and with a constant value), and Extended (the optional). For example, Wikidata item 海宁州 (Q60992243)[66] did not meet the minimum requirements of the data model for Chinese historical places since it had no English label, no description or any statement about this historic place in China.

Another critical step is to provide references or sources to the origin of a Wikidata statement so as to ensure the high quality of Wikidata. Wikidata supports multiple perspectives. "References make it possible to record and represent multiple pieces of data on a subject, even if they contradict one another. As long as there is a reliable source for a statement it can be added to Wikidata."[67] For example, the birth date of Chia-ying Yeh (Q9334904) had two values on July 17, 2021: January 1, 1924 and July 1, 1924. One person should have only one birth date. Since Wikidata is a collaborative editing system, it allows multiple perspectives; however, statements should be ranked based on which value has higher confidence according to validated references.

The third critical step is to learn how to effectively use various Wikidata tools and also understand their limitations. For example, Wikidata SPARQL

query service is a powerful tool to extract Wikidata with a query composed of logical combinations of triples. The key of using this Service to pull data is to use general conditions/properties or common properties that all targeted Wikidata items have. In the initial trial to retrieve Eight Great Prose Masters of the Tang and Song dynasties, only seven entries were retrieved because the Wikidata item for one poet did not have the essayist (Q11774202) listed as occupation (P106). After the Group revised the query by removing essayist (Q11774202) from the query, Eight Great Prose Masters of the Tang and Song dynasties were retrieved. Another option is to add essayist (Q11774202) to the Wikidata item for that one poet so that all of the Eight Great Prose Masters of the Tang and Song dynasties have essayist (Q11774202) listed in occupation (P106). Another example is Wikidata Item Quality Evaluator.[68] Although the Evaluator gives quality scores and calculates their average value, it does not indicate improper entries of Wikidata items. Thus, it will be more helpful if tools are available in the future for both built-in field-level validation and built-in field-level constraint for the value of a property in a statement. In addition, there will be a learning curve to using the existing tools for batch-editing and workflow integration, such as OpenRefine, QuickStatements,[69] and Wikibot.

Opportunities

"Wikidata is actualizing the vision of a semantic web touted by Sir Tim Berners-Lee, creator of the world wide web."[70] Wikidata has provided both the library and the CH section with great opportunities, including:

- Wikidata can lower the barriers for libraries and the CH sector to adopt LOD.
- Wikidata enables creation of URIs and entity management on the Semantic Web. It enables humans and machines to work collaboratively to enrich Wikidata items.
- Wikidata powers a lot of other systems. Wikidata items are reused widely, for example, Wikidata statements can be referenced to both Wikipedia and other sources. The trending of the Wikidata being referenced more than 70 million times by Wikipedia, and more than 800 million times by other sources on a daily basis is available from Wikidata stats. Specifically, on April 12, 2022, Wikidata statements were referenced 73,178,984 times by Wikipedia. On the same day, Wikidata statements were referenced 847,353,272 times by other sources.[71]
- Wikidata enables machine processing of authority data, including reuse, inference, and distribution. It enables machines to make

inference based on the relationship of RDF properties and classes. The Wikidata community including libraries and the CH sector can use Wikidata as a distributor for authority data.
- Wikidata supports multilingual input and display of Wikidata items. For example, it allows the Group to input Chinese characters for Wikidata items that they created and enhanced.
- Wikidata enables creation of rich data that may be excluded or unavailable in traditional library data, for example, Wikidata often contains more extensive biographical data and images.
- Wikidata enables global representation and linking. Wikidata in the LOD cloud[72] is a typical example. Bibliographic data can be enriched in RDA/RDF and BIBFRAME with Wikidata and other external sources.
- Discovery and representation in knowledge graphs can be enhanced, for example, Wikidata has been brought into an info box as a knowledge panel in discovery interface or search engine interface. Wikidata also enables search support including typeahead feature during a user's search.

Conclusion

"Wikidata: WikiProject Chinese Culture and Heritage Group" has provided four main contributions to the Wikidata community and the CH sector:

- Created and enriched Wikidata items for a subset of Chinese women poets and the Chinese historical places; thus, increasing the discoverability of that data and contributing to the management of those entities;
- Contributed to the diversity of data in Wikidata by adding Chinese cultural heritage information;
- Demonstrated the benefits of Wikidata in the description and discovery of cultural heritage information; and
- The data models and semi-automated workflows that the Group has built might be of interest to other people doing similar work or Wikidata research.

In the future, the Group plans to provide more extended properties for Wikidata items of Chinese women poets, including art-name (P1787, Hao, 号) and courtesy name (P1782, Zi, 字), keep working on Chinese historical places, establish best practices for describing Chinese historical places, propose to the Wikidata community to consider adding a new property number for the MQWW database and revise the constraint surname, and

start creating and enhancing Wikidata for the notable works of Chinese women poets. The group also plans to expand the scope of the project to reconcile Wikidata items that may appear to be partial matches as well as to incorporate other identifiers, such as those from the Library of Congress Name Authority File (LCNAF), Virtual International Authority File (VIAF), Hong Kong Chinese Authority Name (HKCAN), OCLC WorldCat Identifies, etc.

Wikidata has a huge potential to meet the needs of entity management for describing and discovery of cultural heritage information. The Wikidata platform is open to anyone who pledges to abide by the Wikidata community practices and guidelines. As a result, Wikidata items can be created and enhanced not only by library professionals but also subject experts outside of the library community. Wikidata supports multilingual input and display of its items, which can be critical for discerning cultural heritage resources because of the potential loss of meaning when they have not been described in their own languages. Wikidata is flexible enough to accommodate some unique cultural characteristics, for example courtesy names, which provide a rich context describing people from non-English speaking countries and regions. However, Wikidata might not be able, at least not at its current state to replace traditional tools used by libraries for entity management because of the concerns of its inconsistent quality (since everyone can edit anything in Wikidata) and lack of best practices or community agreement on how to describe and manage various entities. As more and more people from the library community and the CH sector investigate and explore the potential of Wikidata, like what this Group has tried to do, Wikidata may become one of the viable and suitable platforms that can support libraries' and the CH sector's entity management endeavors in the near future.

Disclaimer

Amanda Xu contributed to this article in her personal capacity. The views and opinions expressed within are those of the author and do not necessarily represent the views of the Agricultural Research Service, USDA or the United States Government.

Acknowledgement

This article is a summary of the projects conducted by the "Wikidata:WikiProject Chinese Culture and Heritage Group" which was formed in October 2020 with Chinese American librarians from several institutions. The goals of this Group are to promote Chinese

cultural heritage information in Wikidata, expand the members' horizons in Linked Open Data for libraries, and seek collaboration opportunities. The Group has undertaken multiple projects and presented their initial findings about characteristics of Wikidata related to Chinese cultural heritage, challenges in creating and enhancing Wikidata items, data models for Chinese women poets and historical places, and different visualizations using Wikidata SPARQL Query Service. The Group coordinator is Xiaoli Li (University of California, Davis). The Group members are: Sai Deng (University of Central Florida), Sophie Dong (Syracuse University), Greta Heng (San Diego State University), Jie Huang (Penn State University), Jing Jiang (California Digital Library), Clara Liao (Library of Congress), Hsianghui Liu-Spencer (Carleton College), Cindy Tian (University of Notre Dame), Amanda Xu (National Agricultural Library), Yan Yu (University of Notre Dame), and Lihong Zhu (Washington State University). Special thanks to China Biographical Database Project (CBDB), Ming Qing Women's Writings Project (MQWW), and China Historical Geographical Information System Project (CHGIS).

ORCID

Lihong Zhu http://orcid.org/0000-0001-8485-6938
Amanda Xu http://orcid.org/0000-0002-2091-8244
Sai Deng http://orcid.org/0000-0002-3681-4888
Greta Heng http://orcid.org/0000-0002-3606-6357
Xiaoli Li http://orcid.org/0000-0001-5362-2151

Notes

1. Jennifer Liss, "The Semantic Web, Authority Control, and You," presented at the MCLS Linked Data Users Group Virtual Meeting, May 1, 2018, https://www.mcls.org/files/3715/2543/9356/LDUG_Authority_Pres_05022018.pdf
2. Chelsea Dalgord, "Shared Entity Management Infrastructure Project Update," June 24, 2020, https://www.loc.gov/bibframe/news/source/bibframe-from-home-oclc-update.pptx
3. "Getty Thesaurus of Geographic Names Online," https://www.getty.edu/research/tools/vocabularies/tgn/ (accessed December 10, 2022).
4. "Europeana Data Model," https://pro.europeana.eu/page/edm-documentation (accessed December 10, 2022).
5. Eero Hyvonen, "Cultural Heritage Linked Data on the Semantic Web: Three Case Studies Using the Sampo Model," invited talk to be published in the proceedings of: VIII Encounter of Documentation Centres of Contemporary Art: Open Linked Data and Integral Management of Information in Cultural Centres Artium, Vitoria-Gasteiz, Spain, October 19–20, 2016, https://seco.cs.aalto.fi/publications/2017/hyvonen-vitoria-2017.pdf
6. "Wikidata:Relation between Properties in RDF and in Wikidata," last updated July 18, 2022, https://www.wikidata.org/wiki/Wikidata:Relation_between_properties_in_RDF_and_in_Wikidata
7. "Wikidata:Notability," last updated December 15, 2022, https://www.wikidata.org/wiki/Wikidata:Notability#:~:text=It%20refers%20to%20an%20instance,in%20other%20items%20more%20useful
8. "Help:Statements," last updated November 18, 2022, https://www.wikidata.org/wiki/Help:Statements

9. Ibid.
10. "Wikidata," last updated December 7, 2022, https://en.wikipedia.org/wiki/Wikidata#cite_note-20
11. Stephan Bartholmei, Rachel Franks, James Heilman, Mylee Joseph, Vicki McDonald, Anna Raunik, Mia Ridge, and Mark Robertson, "Opportunities for Academic and Research Libraries and Wikipedia," endorsed by the IFLA Governing Board, December 2016, https://www.ifla.org/wp-content/uploads/2019/05/assets/hq/topics/info-society/iflawikipediaopportunitiesforacademicandresearchlibraries.pdf
12. Karen Smith-Yoshimura, "Analysis of 2018 International Linked Data Survey for Implementers," *Code4Lib* 42 (2018), https://journal.code4lib.org/articles/13867
13. "Wikidata:WikiProject PCC Wikidata Pilot," last updated July 22, 2022, https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot
14. "Wikidata:WikiProject LD4 Wikidata Affinity Group Calls," last updated December 13, 2022, https://www.wikidata.org/wiki/Wikidata:WikiProject_LD4_Wikidata_Affinity_Group/Affinity_Group_Calls
15. "Wikipedia:WikiProject Women in Red," last updated December 15, 2022, https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red
16. Tim Berners-Lee, "Linked Data," last updated June 18, 2009, https://www.w3.org/DesignIssues/LinkedData.html
17. "CulturaSampo," http://www.kulttuurisampo.fi/ (accessed December 10, 2022).
18. "DigiCULT," https://www.digicult.info/pages/index.php (accessed December 10, 2022).
19. "CASPAR," http://www.casparpreserves.eu/ (accessed December 10, 2022).
20. Donald Sturgeon, "Constructing a Crowdsourced Linked Open Knowledge Base of Chinese History (PNC)," presented at the 2021 Pacific Neighborhood Consortium Annual Conference and Joint Meetings, https://doi.org/10.23919/PNC53575.2021.9672294
21. "Chinese Text Project," last updated October 10, 2016, https://ctext.org
22. Konstantinos N. Vavliakis, Georgios Th. Karagiannis, and Pericles A. Mitkas, "Semantic Web in Cultural Heritage after 2020," presented at the 11th International Semantic Web Conference 2012 (ISWC 2012), https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.492.4352&rep=rep1&type=pdf
23. Eetu Makela, Eero Hyvonen and Tuukka Ruotsalo, "How to Deal with Massively Heterogeneous Cultural Heritage Data - Lessons Learned in CultureSampo," *Semantic Web* 3, no. 1 (2012): 85–109, https://doi.org/10.3233/SW-2012-0049
24. Johan Oomen and Lora Aroyo, "Crowdsourcing in the Cultural Heritage Domain: Opportunities and Challenges," *Proceedings of the 5th International Conference on Communities and Technologies. New York, June 2011* (New York, NY: Association for Computing Machinery, 2011), 138-49, https://doi.org/10.1145/2103354.2103373.
25. Ibid.
26. Thomas Baker, Karen Coyle and Sean Petiya, "Multi-Entity Models of Resource Description in the Semantic Web: A Comparison of FRBR, RDA and BIBFRAME," *Library Hi Tech* 32, no. 4 (2014): 562–582, https://doi.org/10.1108/LHT-08-2014-0081
27. Carol Jean Godby and Karen Smith-Yoshimura, "From Records to Things: Managing the Transition from Legacy Library Metadata to Linked Data," *Bulletin of the Association for Information Science and Technology* 43, no. 2 (2017): 18–23, https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/bul2.2017.1720430209
28. "OCLC Awarded Mellon Foundation Grant to Develop Infrastructure to Support Linked Data Management Initiatives," last updated January 9, 2020, https://www.oclc.org/en/news/releases/2020/20200109-oclc-awarded-mellon-grant-linked-data-management-infrastructure.html
29. Marçal Mora-Cantallops, Salvador Sánchez-Alonso, and Elena García-Barriocanal, "A

Systematic Literature Review on Wikidata," *Data Technologies and Application* 53, no. 3 (2019): 250–68, https://doi.org/10.1108/DTA-12-2018-0110
30. Theo Van Veen, "Wikidata: From 'an' Identifier to 'the' Identifier," *Information Technology and Libraries* 38, no. 2 (2019): 72–81, https://doi.org/10.6017/ital.v38i2.10886
31. Paromita Biswas, "Exploring Possibilities in Wikidata." *Technicalitie* 40, no. 6 (2020): 11–15, http://www.casparpreserves.eu/
32. Gabriel Amaral, Alessandro Piscopo, Lucie-Aimée Kaffee, Odinaldo Rodrigues, and Elena Simperl, "Assessing the Quality of Sources in Wikidata Across Languages: a Hybrid Approach," *ACM Journal of Data and Information Quality* 13, no. 4 (2021): 1–35, https://doi.org/10.1145/3484828
33. Lucie-Aimée Kaffee, Alessandro Piscopo, Pavlos Vougiouklis, Elena Simperl, Leslie Carr, and Lydia Pintscher, "A Glimpse into Babel: an Analysis of Multilinguality in Wikidata," *OpenSym '17: Proceedings of the 13th International Symposium on Open Collaboratio,. Galway, Ireland: OpenSym, August 2017*: 1–5, https://doi.org/10.1145/3125433.3125465
34. Merrilee Proffitt, "Wikidata, Wikibase and the Library Linked Data Ecosystem: An OCLC Research Library Partnership Discussion," last updated September 17, 2020, https://hangingtogether.org/wikidata-wikibase-and-the-library-linked-data-ecosystem-an-oclc-research-library-partnership-discussion/
35. Alex Stinson, Sandra Fauconnier, Liam Wyatt, Susanna Ånäs, and Jane Darnell, "Why You Should be Paying Attention to Wikidata and GLAM," last updated August 23, 2016, https://diff.wikimedia.org/2016/08/23/wikidata-glam/
36. "Wikidata:WikiProject LD4 Wikidata Affinity Group," last updated October 14, 2022, https://www.wikidata.org/wiki/Wikidata:WikiProject_LD4_Wikidata_Affinity_Group
37. "Wikidata:WikiProject PCC Wikidata Pilot," last updated July 22, 2022, https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot
38. Keala Richard, "Smithsonian Libraries and Archives & Wikidata: Chinese Ancestor Portrait Project," last updated March 30, 2022, https://blog.library.si.edu/blog/2022/03/30/smithsonian-libraries-and-archives-wikidata-chinese-ancestor-portrait-project/#.Ytw7jHbMLrd
39. "Wikidata:WikiProject PCC Wikidata Pilot/Smithsonian Libraries/Projects/Chinese Portraits," last updated April 19, 2022, https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot/Smithsonian_Libraries/Projects/Chinese_Portraits
40. "Wikidata: WikiProject Historical Place," last updated December 15, 2022, https://www.wikidata.org/wiki/Wikidata:WikiProject_Historical_Place
41. "Wikidata:WikiProject CJKV Character," last updated November 1, 2021, https://www.wikidata.org/wiki/Wikidata:WikiProject_CJKV_character
42. "Wikimedia," https://www.wikimedia.org/ (accessed December 15, 2022).
43. "DPLA + Wikimedia," https://pro.dp.la/projects/dpla-wikimedia (accessed September 6, 2022).
44. "Wikidata:Notability," last updated December 15, 2022, https://www.wikidata.org/wiki/Wikidata:Notability#:~:text=It%20refers%20to%20an%20instance,in%20other%20items%20more%20useful
45. "China Biographical Database Project," https://projects.iq.harvard.edu/cbdb (accessed July 22, 2022).
46. "Ming Qing Women's Writings," https://digital.library.mcgill.ca/mingqing/english/index.php (accessed December 10, 2022).
47. "China Historical Geographical Information System," https://gis.harvard.edu/ (accessed December 10, 2022).

48. "China Biographical Database Project," https://projects.iq.harvard.edu/cbdb (accessed July 22, 2022).
49. "Ming Qing Women's Writings," https://digital.library.mcgill.ca/mingqing/english/index.php (accessed December 10, 2022).
50. "China Historical GIS," https://gis.harvard.edu/china-historical-gis (accessed July 22, 2022).
51. "Wiki-Bot," https://discord.bots.gg/bots/461189216198590464 (accessed December 10, 2022).
52. "jsjiang/ChineseWikiClub," https://github.com/jsjiang/ChineseWikiClub (accessed December 12, 2022).
53. "Pywikibot," https://github.com/wikimedia/pywikibot (accessed December 10, 2022).
54. "Ming Qing Women's Writings," https://digital.library.mcgill.ca/mingqing/english/index.php (accessed December 10, 2022).
55. "China Biographical Database Project," https://projects.iq.harvard.edu/cbdb (accessed July 22, 2022).
56. "Wikidata: Data Import Guide," https://www.wikidata.org/wiki/Wikidata:Data_Import_Guide (accessed December 10, 2022).
57. "OpenRefine User Manual," https://docs.openrefine.org/ (accessed December 10, 2022).
58. Tom Heath and Christian Bizer, "Querying Local Data with SPARQL," in *Linked Data: Evolving the Web into a Global Data Space* (Morgan & Claypool, 2011): 96.
59. "Wikidata SPARQL Query Service GUI," https://query.wikidata.org (accessed December 10, 2022).
60. "Wikidata:SPARQL Tutorial," https://www.wikidata.org/wiki/Wikidata:SPARQL_tutorial (accessed September 3, 2022).
61. "Wikidata SPARQL Query Service GUI," https://query.wikidata.org (accessed December 10, 2022).
62. "Pei Shu (Q45625354)," https://www.wikidata.org/wiki/Q45625354 (accessed June 28, 2022).
63. "裴淑," https://baike.baidu.hk/item/%E8%A3%B4%E6%B7%91/2445087 (accessed August 11, 2022).
64. "(Q60992243)," https://www.wikidata.org/wiki/Q60992243 (accessed June 28, 2022).
65. "Qiantang County (Q11650262)," https://www.wikidata.org/wiki/Q11650262 (accessed July 21, 2022).
66. "(Q60992243)," https://www.wikidata.org/wiki/Q60992243 (accessed June 28, 2022).
67. "Wikidata:Tours/References," last updated April 15, 2022, https://www.wikidata.org/wiki/Wikidata:Tours/References
68. "Wikidata Item Quality Evaluator," https://item-quality-evaluator.toolforge.org/ (accessed December 10, 2022).
69. "Help:QuickStatements," last updated October 30, 2022, https://www.wikidata.org/wiki/Help:QuickStatements
70. Monika Sengul-Jones, "The Promise of Wikidata: How Journalists Can Use the Crowdsourced Open Knowledge Base as a Data Source," last updated February 10, 2021, https://datajournalism.com/read/longreads/the-promise-of-wikidata
71. "Wikidata Stats," https://wikidata-todo.toolforge.org/stats.php (accessed September 25, 2022).
72. "LOD cloud," https://lod-cloud.net/ (accessed December 10, 2022).

Appendixes

Appendix 1. Data model for Chinese women poets (only core and constant listed).

Level	Property	Value	Usage note
Const	instance of (P31)	human	
Const	sex or gender (P21)	female	
Const	occupation (P106)	poet	
Const	field of work (P101)	poetry	
Core	name in native language (P1559)	name in native language	
Core	native language (P103)	native language	Recommended to pair with "name in native language".
Core	languages spoken, written or signed (P1412)	languages spoken, written or signed	
Core	family name (P734)	family name	
Core	date of birth (P569)	date of birth	
Core	date of death (P570)	date of death	

Appendix 2. Selected properties in CBDB and MQWW.

	CBDB	MQWW
Full Name	Y	Y
First Name	Y	Y
Last Name	Y	Y
Literary Name (Hao, 号)	N	Y
Courtesy Name (Zi, 字)	N	Y
Other Zi, Hao	N	Y
Gender	Y	Y
Ethnicity	Y	Y
Major Works	N	Y
Marital Status	N	Y
Year of Birth	Y	Y
Year of Death	Y	Y
Death Age	Y	Y
Dynasty	Y	Y

Appendix 3. Data model for historical places.

Data Model for Historical Places	Value	Note/Example
Basic description		
English label	English Description	For 县, English description should be: ancient county of China
Chinese label	Chinese description	中国历史地名, 今XXXX
Simplified Chinese label	Simplified Chinese description	
Traditional Chinese label	Classical Chinese description	
Core statements		
instance of (P31)	Wikidata item for type of place	ancient county of China (Q28739697)
country (P17)	China (Q29520)	
official name (P1448)	String value	In traditional Chinese
native label (P1705)	String value	In traditional Chinese
coordinate location (P625)	Coordinates	
located in time zone (P421)	UTC + 08:00 (Q6905)	
CHGIS ID (P4711)	String value	