Integration with Wikidata

An evaluation of Wikidata Properties and Wikibase as solutions for Curationist

Prepared by Sharon Mizota and Jessica Gengler
September 11, 2024
Updated October 15, 2024

Table of Contents

Introduction

In the spring of 2024, Curationist leaders expressed interest in exploring the use of Wikidata Properties as a metadata schema. Since its inception, Curationist has used Wikidata as the sole source of controlled descriptive terminology for the metadata it adds to records contributed by its partners. This report's authors, who create and manage this metadata, have long been interested in integrating the Curationist backend and frontend with Wikidata in a way that would make the search, selection, and management of this terminology more seamless and to take advantage of the rich connections between concepts, works, and people that Wikidata's linked open data provides. This report is the result of our investigation and evaluation of:

- Wikidata Properties as a possible new schema for Curationist metadata, and
- Wikibase as a possible repository for our controlled vocabularies.

What is Wikidata?

Wikidata is a database representing things, people, concepts, and their relationships to one another. It is an example of "linked open data," or data structured to facilitate making connections. This may mean establishing that one artist was taught or influenced by another, that a given sculpture depicts the same person as a particular painting, or that two works were created in the same time period or are made of the same materials. It performs a function known in library science as "authority control," in which each record stands in for a discreet and unique person, place, thing, or concept. This means it may contain records or Items for:

- jar, the general category of object described as a "rigid, approximately cylindrical container with a wide mouth or opening,"
- Jar, a particular example of a jar in the collection of The Metropolitan Museum of Art,
- Jar, "a tram and metro station in Oslo, Norway," and
- any number of other "jars," as long as they are unique.

Because of these overlapping meanings of the word "jar," authority control facilitates better, more precise online searching and discovery. If you are only interested in the Oslo tram station, you do not want to find information about containers. Wikidata helps computers understand the difference between these things by assigning a unique ID number, or "Q" number, to each Item. Each Item in Wikidata has a unique QID:

- jar, (**Q1207302**) the general category of object described as a "rigid, approximately cylindrical container with a wide mouth or opening"

- Jar, (**Q116244267**) a particular example of a jar in the collection of The Metropolitan Museum of Art
- Jar, (**Q6159423**) “a tram and metro station in Oslo, Norway”

Even if the name of the tram station were to change, we would still be able to keep track of it because its QID would remain the same. These persistent identifiers are a powerful way of building relationships between things online that are resistant to changes in naming and language.

This persistence is particularly important because, in addition to being freely available to all, Wikidata is open to being added to and edited by anyone. Anyone can change almost anything on Wikidata at any time. This means Wikidata is much more responsive to changes in culture and language than other sources of authority control, like the Library of Congress Linked Data Service, but it also means that it can be difficult to keep up with.

Curationist uses Wikidata because it is more up-to-date than vocabularies that are centrally managed and because it is editable. Since 2021, Curationist archivists have created over 50 new entries on Wikidata and have edited countless others. We often add more current, respectful, or accurate language to Wikidata for the terms we add to the records that come from our partner institutions.

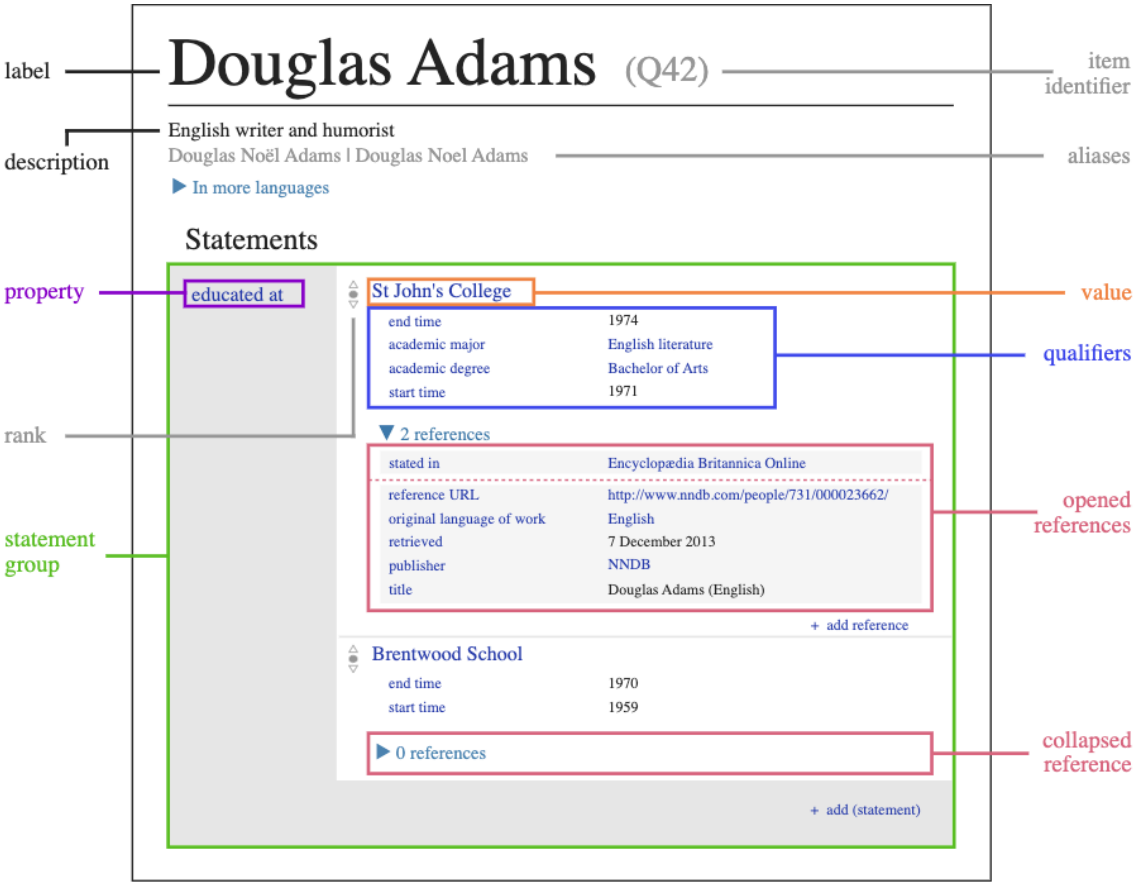
For the purposes of this report, it is also important to understand Wikidata Properties. Properties are the data elements available on Wikidata to describe an Item. The record for our first “jar” example above includes the following Properties:

- **subclass of:** the larger category of objects to which “jar” belongs
- **has use:** the main use of “jar”
- **image:** a photographic image of jars from Wikimedia Commons
- **pronunciation audio:** a recording of a voice saying “jar”
- **physically interacts with:** other substances jars come into contact with, like food
- **icon:** an illustration of a jar
- **described by source:** resources that define what a “jar” is
- …and many others

All of these Properties help computers understand what a “jar” is and how it might be used or related to other entities. Properties define what is knowable about an Item on Wikidata, and there are thousands of them.

Anatomy of a Wikidata Page

Here is a helpful guide to the various parts of a Wikidata page.



Douglas Adams (Q42)

English writer and humorist

Douglas Noël Adams | Douglas Noel Adams

▶ In more languages

Statements

educated at

St John's College

end time	1974
academic major	English literature
academic degree	Bachelor of Arts
start time	1971

▼ 2 references

stated in	Encyclopædia Britannica Online
reference URL	http://www.nndb.com/people/731/000023662/
original language of work	English
retrieved	7 December 2013
publisher	NNDB
title	Douglas Adams (English)

+ add reference

Brentwood School

end time	1970
start time	1959

▶ 0 references

+ add (statement)

By Charlie Kritschmar (WMDE) - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=49616867

To learn more about Wikidata, please see this Introduction to Wikidata.

What is Wikibase?

Wikibase is software that enables you to set up your own Wikidata-like database. It was developed by Wikimedia Deutschland. Its data structure is the same as Wikidata's but it is completely customizable in terms of the names and number of Properties included. As such, each installation of Wikibase can be totally independent of Wikidata and used for custom vocabulary management.

Learn more about Wikibase here.

Methodologies

Although this report has two main areas of inquiry, we approached both in the same way. Our research consisted of a literature review of existing documentation about Wikidata and Wikibase projects that are similar to the ones proposed above. This documentation included case studies, technical documentation, WikiProject descriptions, articles, blog posts, and recordings of presentations. Sharon also presented on our project at the LD4 Art and Design Affinity Group meeting in May and received helpful feedback and resources to add to our review. We also conducted some analysis of Wikidata and Wikimedia Commons data models and did some exploratory mapping. From there, we identified people associated with these projects, or through our existing networks, and conducted three unstructured consultations to present our ideas and get their responses and advice. We also consulted with Mitchell Parsons, Curationist Developer, to assess the feasibility of the technical solutions we uncovered along the way. In this process, we identified people to invite as panelists to the Metadata Learning & Unlearning Summit, which took place on September 19, 2024. We presented the findings from this report at this Summit and solicited feedback from the six panelists.

Part One: Wikidata Properties

From its inception, Curationist has used Wikidata as a source for its controlled descriptive vocabularies. Since then, the use of Wikidata has increased across the digital cultural heritage sector, and more and more institutions are contributing their data to Wikidata. With this in mind, Curationist leaders became interested in whether Wikidata Properties—the data elements that Wikidata captures about each of the Items in its database—could be used as a new Curationist schema that would make it easier to map new contributor data. This interest in a new schema was also driven by the development of the Curationist Museum Services project, which plans to advise smaller cultural heritage organizations on digitization, metadata, and open access sharing. This report details our findings and recommendations as to whether a data migration from the Curationist schema to one based on Wikidata Properties is advantageous for ingesting new metadata to Curationist and facilitating sharing of that metadata.

Background

Curationist's current metadata schema is a modified version of VRA Core 4, a standard developed for visual materials that is currently maintained by the Library of Congress. Museum metadata is notoriously varied and detailed, and at the time the schema was developed, Curationist was interested in capturing as many of the available data elements as possible. VRA Core 4 was selected because it includes many of the elements our contributors capture in their metadata, and because it includes "attributes," which enable the recording of the source of each data element. This feature of VRA Core 4 facilitates the metadata layering that differentiates Curationist from other digital aggregators.

The schema was then customized to facilitate search engine optimization and the use of works in the site's editorial content. These customizations shifted the schema away from the standard VRA Core 4, although most data elements still correspond to the standard schema. Although VRA Core 4 is not widely used, we have found there is no one schema that predominates among our current contributors. Contributors use a variety of schemas, mostly custom ones.

As an example, here are the various element names for "date" metadata across current Curationist contributors:

Source Contributor	Elements that correspond to VRA Core 4 "date"
Art Institute of Chicago	date_display date_start date_end
Brooklyn Museum	object_date object_date_begin object_date_end period dynasty
Cleveland Museum of Art	creation_date creation_date_earliest creation_date_latest
Metropolitan Museum of Art	objectDate objectBeginDate objectEndDate
National Gallery of Art	displaydate beginyear endyear visualbrowsertimespan
Rijksmuseum	earliestDate latestDate periodName
Smithsonian	date
Statens Museum for Kunst	production_date/start production_date/end production_date/period
Walters Museum of Art	DateBeginYear DateEndYear DateText Period Reign Dynasty

As you can see, there is no consensus among our contributors as to what even the most common elements, such as "date," are called. Currently, onboarding new contributors requires mapping these element names to their equivalents in the Curationist schema on a case-by-case basis. This report looks at whether using Wikidata Properties as a schema would make this process easier.

Findings

Our research into using Wikidata Properties as a metadata schema took four directions:

- **Review of Existing Wiki Data Models.** We reviewed and compared data models for cultural heritage projects on Wikidata and Wikimedia Commons to gauge their similarities to our schema.
- **Mapping of Wikidata Properties to the Curationist Schema.** We mapped Wikidata Properties to our current schema to assess whether all of our elements could be adequately represented.
- **Documentation Review.** Examination of documentation of data transfer and data roundtripping projects by museums to assess the difficulty of sharing data to and from Wikidata.
- **Interviews with Wiki Experts.** We conducted three interviews with associates active in the Wikidata or Wikimedia Commons space to get their reactions to our proposals.

Review of Existing Data Models

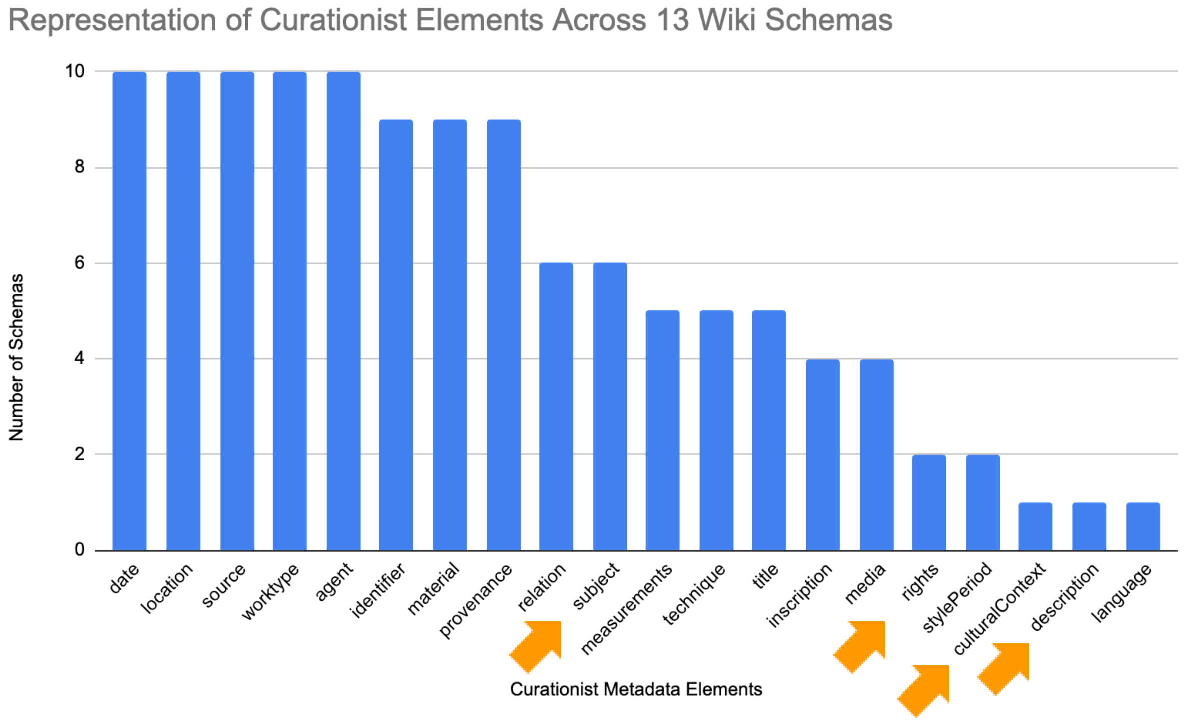
There are a number of cultural heritage and visual art WikiProjects that use a preselected group of Wikidata Properties or Wikitext elements as ad hoc schemas for creating/uploading data in Wikidata and Wikimedia Commons. These data models are used to guide Wikimedians on which data elements to focus on when they participate in a WikiProject.

We looked at twelve data models for Wikidata cultural heritage Items and one from Wikimedia Commons and compared them with our Curationist schema. While there was no one model that covered all of the elements of the Curationist schema, the closest was WikiProject Visual Arts, which has 47 elements that map to elements and subelements in the Curationist schema. However, this mapping is not always one-to-one. For example, there are multiple Wikidata elements that correspond to the single Curationist elements “location” and “provenance.” See the section “Data Models Reviewed” below for the complete list of schemas we considered.

Below are the main Curationist elements and the number of Wikidata/Wikimedia data models that have an equivalent element or elements, ranked from most to least common:

Curationist Main Element	**# of Models**	**% of Models**
date	10	77%
location	10	77%
source	10	77%
worktype	10	77%
agent	10	77%
identifier	9	69%
material	9	69%
provenance	9	69%
relation	6	46%
subject	6	46%
measurements	5	38%
technique	5	38%
title	5	38%
inscription	4	31%
media	4	31%
rights	2	15%
stylePeriod	2	15%
culturalContext	1	8%
description	1	8%
language	1	8%

Here is the same data as a bar graph:



Curationist's priorities and Wikidata's emphases for cultural heritage items are not exactly aligned. While “date,” “location,” “source,” “worktype,” and “agent” are all key pieces of data on both platforms, “subject,” “rights,” “culturalContext” and “description” are less represented in Wikidata/Wikimedia Commons data models. These latter (marked with ↗) are the elements where Curationist is most likely to add new metadata. The fact that they seem to be less important for Wikidata/Commons audiences reflects the different orientations of the data sources. Whereas Wikidata in particular is more interested in unambiguous data that can be used to create connections between Items, Curationist is more concerned with discovery, nuance, and interpretation of cultural heritage through a social justice lens.

You can see the crosswalk between the Curationist schema and all thirteen data models in this Google Sheet.

Data Models Reviewed

This analysis is based on a review of the data models for the following projects/WikiProjects:

- WikiProject Collection data from Wikidata to Finna

Description of roundtrip data exchange where a selection of smaller museums upload their data to Wikidata and then connect it to their Omeka-S installations as linked open data for eventual sharing with **Finna.fi**. See “Review of Projects” below for more detail.

- **IDEA, International (Digital) Dura-Europos Archive (4 schemas)**
  Documentation of Syrian archaeological heritage taking place in Wikidata as a tool for international cooperation.
  About: https://www.wikidata.org/wiki/Wikidata:WikiProject_IDEA
  Includes Wikidata templates:
  - Papyri
  - Other Artifacts
  - Archival Docs & Photos
  - Inscriptions

- **Sum of all paintings**
  Item structure is very minimal.
  Sub-project: Sum of all paintings/Provenance
  - “The main properties we use for this are collection (P195), location (P276) & owned by (P127).”

- **WikiCommons Artwork Template**
  The only non-Wikidata model we looked at, it is mapped to Wikidata Properties but not all of its elements have a corresponding Wikidata Property. Sometimes one WikiCommons element maps to several different Wikidata Properties. This is not a schema proper in that it is just a template for a Wikitext page. Commons is now using Wikidata Properties as structured data (SDC) alongside this template, but this template is actually closer to the Curationist schema.

- **WikiProject Media Art**

- **WikiProject Provenance**
  An effort to get every item about an artwork described by ownership and/or location since its inception on Wikidata.

- **WikiProject Public Art**

- **WikiProject Textile Arts**

- **WikiProject Visual Arts**

This was the schema closest to Curationist’s that we found.

Wikidata Properties Make Sense… In Wikidata

Another observation that emerged from this review of data models was that the names of Wikidata Properties seem to be specific to Wikidata—they were developed not as a descriptive “system,” but were crowdsourced as people needed them—and do not necessarily correspond to existing and widely used schema like Dublin Core or Schema.org. This situation means that sharing with Wikidata is likely to always require mapping, reformatting, and generating new metadata.

It also means that Wikidata contains Property names that may be awkward, too precise, or not precise enough for cultural heritage collections. While there was a high degree of agreement among the WikiProjects listed above about which Properties to use, there are several whose names are not intuitive. Examples of odd-sounding Properties include “inception” (P571), which is widely used for “start date” or “creation date” (unless you are describing a person, and then you might use “date of birth” (P569) instead), and “instance of” (P31), which is widely used for “worktype,” “object type,” or “classification,” etc.

There are also Properties that have the same or overlapping meanings, such as:

- “creator” (P170) and “author” (P50)
- “based on” (P144) and “inspired by” (P941)
- “depicts” (P180)” and “main subject” (P921)

These overlaps can make thorough and complete searching of Wikidata a fool’s errand, as one can never be sure to have covered all of the relevant Properties in any one query.

Takeaways:

- **Different Ideas of “Essential” Data.** Wikidata/Commons data models and the Curationist schema each emphasize different aspects of cultural heritage objects. Wikidata in particular tends to favor unambiguous data that fosters connections between Items, whereas Curationist emphasizes providing multiple avenues of discovery and more nuanced interpretation.
- **Idiosyncratic Property Names.** Wikidata Properties are often named in ways that are not intuitive and may not make sense to a cultural heritage organization. They are also not internally consistent even within Wikidata because they are crowdsourced.

Mapping Wikidata Properties to the Curationist Schema

Starting with the data model closest to the Curationist schema—WikiProject Visual Arts—we attempted to map Wikidata Properties to the Curationist schema.

Of the 97 elements, subelements, and attributes that make up the Curationist schema, we were able to find equivalents in Wikidata Properties for 44. By this somewhat crude measure, only 45% of the Curationist schema is able to be represented in Wikidata. However, the actual percentage is lower, due to several issues we encountered in the mapping process.

You can see the full mapping attempt on the first tab of this Google Sheet.

Mapping Issues

There are several areas of incompatibility when mapping between Curationist and Wikidata. Sharing data from Wikidata to Curationist would not be a problem, but potential sharing from Curationist to Wikidata brings up several issues:

- **Wikidata Items:** Every Wikidata Property has a default “data type.” This means that each Property is designed to expect a certain type and format of data and cannot accommodate all forms of data. Most Wikidata Properties relevant to the Curationist schema have the data type “Item,” which means they expect a link to a Wikidata Item and do not accept strings or text values. By contrast, most of the data values on Curationist are strings/text. Only a small percentage of data contributed by partners and the data added by Curationist archivists corresponds to Wikidata Items. *Basically, we could migrate our current data to a schema based on Wikidata Properties, but the vast majority of our data would still not be shareable with Wikidata.*
    - **“type” subelements:** In the Curationist schema, these subelements are used to indicate the “type” of an element, such as the type of “location,” which could be “creation,” “discovery,” etc. on Curationist. This information is handled by various Qualifiers in Wikidata. However, Wikidata expects the values for these Qualifiers to be Wikidata Items, and although the Curationist values are controlled, they are not all Wikidata Items. Migrating this data to Wikidata would require converting the Curationist values to Wikidata Items, and this would only apply to data created by Curationist archivists. The vast majority of “type” data we receive from contributors would also not be easily mappable to Wikidata.
- **Tracking users:** Both Wikidata and Curationist keep track of contributions made by individual users, but their user accounts do not overlap and they use different methods to track them. Migrating to a schema based on Wikidata Properties

would not solve the problem of incompatible user accounts between the two systems, and sharing data from Curationist to Wikidata would require additional mapping between Curationist usernames and Wikidata usernames or, more likely, would require dropping Curationist usernames entirely in favor of an automated generic “Curationist” Wikidata user, likely a bot.

- **Language:** Wikidata keeps track of languages differently than Curationist. In Wikidata, languages are tracked based on the display language of the user, whereas Curationist keeps track of the language of each individual data element. This creates a mismatch in the mapping because Curationist stores language information per element.
- **Description element**: Wikidata uses this field mainly to disambiguate Items that share the same Label or name, and descriptions are usually short phrases or a single sentence. By contrast, descriptions on Curationist don’t always serve this purpose and are sometimes much longer and more discursive.
- **Display preference:** Wikidata does not have a convention for indicating display preference, as Curationist does. It does have “Ranks,” which indicate the quality and relevance of multiple values for a statement, but this is not the same as the display preference Curationist uses to display the most important or most prominent data point first.
- **Measurements**: In Wikidata, measurements are expressed by individual Properties for each dimension, whereas Curationist values vary widely and often express measurements as overall dimensions: “6 in. x 5 in. x 2 in.” or “170 cm wide.” Migrating this data would require parsing the Curationist values to determine which ones to migrate to multiple separate Wikidata Properties. This likely would be quite time-consuming with a high probability of errors.
- **Copyright**: Again, the vast majority of Curationist values are not expressed as Wikidata Items, which is what Wikidata expects for copyright data. Although all items on the Curationist site are public domain, CC0, or CC-BY, the actual expressions of these values are quite diverse and sometimes include longer blocks of text.
- **Provenance**: Wikidata expresses provenance across three different Properties that expect Wikidata Items as values; the longer, discursive, or list-like content common in Curationist values would not neatly or easily map.
- **Image**: Images that appear on Wikidata must be in Wikimedia Commons. Although it might not be a high priority to share images to Wikidata, it would require making sure that all images were uploaded to Wikimedia Commons. This would not only be time-consuming and resource-intensive but would require a complex deduping process for images that have already been uploaded by their source institutions.

There are also a number of smaller issues not listed here. Due to the large number of incompatibilities between Wikidata Properties and Curationist metadata elements, mapping between the two, with the eventual goal of sharing data to Wikidata, would be quite complex and incomplete.

Takeaways:

- **Incomplete Mapping:** Only about 45% of the Curationist schema can be represented in Wikidata, and this percentage is probably lower considering incompatibilities in “data type” between the two databases.
- **Many Mapping Issues:** The biggest barrier to sharing Curationist metadata with Wikidata is the fundamental incompatibility between the text-based data in the Curationist dataset and Wikidata’s expectation of links to other Wikidata Items for the majority of its Properties. Secondary to this are incompatibilities between data elements that are represented by one element in the Curationist schema and many in the Wikidata data model, as from the single element “measurements” to three or more: “width,” “height,” and “length,” etc.

Review of Projects Sharing Data with Wikidata

We found several documented examples of projects where institutions were sharing their metadata with Wikidata or roundtripping it: uploading and then downloading records enhanced by crowdsourcing back to their own repository, or using Wikidata as a waystation to share their data with other repositories or aggregators. We studied three of them that seemed particularly relevant to our own situation to better understand the possibilities, process, and requirements.

Swedish National Heritage Board Study of Wikidata Data Roundtripping

This report describes three pilot projects where user-created Wikidata/Commons data was ingested to enhance museum records. In each case, data was mapped and ingested from/to selected fields only, so it does not deal with Wikidata Properties as a schema. The report illuminates the complications encountered in passing data back and forth, in particular trust and accuracy concerns around crowdsourced data.

The Swedish Performing Arts Agency developed a tool for crowdsourcing translations of documents uploaded to Wikimedia Commons but the project *did not result in ingests of the translations because of their poor and inconsistent quality* (which was sometimes due to use of the Wikitext markup on Commons). The project team was able to ingest the Commons links to show which images in their collection had been uploaded.

The Nationalmuseum pilot involved ingesting crowdsourced IDs from authority sources that had been added to Wikidata for the artists represented in their collection. They had already uploaded the artist data to Wikidata and used a SPARQL query to export a CSV with the data. Notably, they then had to *work with a 3rd party software developer* to convert the CSV into SQL queries in order to make the update in their system, and had to add a field to their database to accommodate Wikidata URIs.

The Nordic Museum pilot uploaded images to Commons with a Pywikibot-based script. They then developed a tagging tool for users to use the Europeana Fashion Vocabulary (EFV) to describe images in Structured Data on Commons (SDC). However, their CMS didn't support ingestion of vocabularies external to the vendor's, and they had to remap their vocabulary to EFV and then *hire the vendor to do the imports.* They also experienced an issue with the one-to-many relationship between their CMS and Commons. Their CMS had a record for one item, but Commons could have multiple images of the same item, resulting in much merging of records. The users engaged in the tagging project also found EFV limiting and wanted to use other Wikidata terms. There was also some duplication between crowdsourced and preexisting descriptions resulting in the museum deciding to review contributions on a per user basis as a way to assess their quality.

Takeaways:

- **Enrichment of Metadata in WikiProjects is Limited.** "Traditionally Wikimedia GLAM partnerships have focused on importing content and data from institutions to Wikimedia projects. Ideally, the data will get enriched, corrected and returned to the originating institution by volunteers in Wikimedia Commons. However, enriching is often limited to categorizing the content. Correcting metadata happens far less often." (Final Report, p. 16)
- **Wikidata vs. Wikimedia Commons:** "The export of information from Wikidata was exponentially easier than exporting data from Wikimedia Commons." (Final Report, pp. 16-17)
- **What is Described in Each System May Be Different:** "Wikidata stores items of works, whereas Structured Data on Commons handles representations (images) of the works, but also photographs and digital items, which may be works in their own right. In addition to this, as seen in the case of the Nordic Museum, the divide is handled yet differently in collection management. This complexity must be taken into account when designing processes for the environments." (Final Report, p. 18)
- **Upload Debt:** Data is always changing across multiple sources and can become difficult to keep up to date. For example, when an institution has previously uploaded smaller files with stricter rights but opens up their access later, there are then duplicates online with conflicting copyright information.
- **Difficulty of Roundtripping:** "While Structured Data on Commons will offer new ways to express data that has previously been squeezed into the descriptions,

the rich descriptive texts in Wikimedia Commons are unlikely to disappear. Translating descriptions and returning these texts back to the institutions is however problematic. Rich text is generally not consumable by the collection management systems." (Final Report, p. 18)

- **Biggest Barriers are Technical Staffing, Money, and Incompatible Tech:** "The lack of technical resources and human or financial resources were cited as the largest barriers for adopting third party contributions in the pre-pilot survey. These pilots have however shown that a major contributor to these challenges is the underlying collection management system. All the participating museums had different collection management systems and the import process differed between all of them." (Final Report, pp. 18-19)

Collection Data From Wikidata to Finna

This project involved three museums in Finland—Pori Art Museum, Lönnström Art Museum, and Lapua City Museums, together with a Wikidata user called Zache as a chief technician and University of Jyväskylä Open Science Centre—who designed a roundtrip data exchange where they uploaded their collection data to Wikidata and connected it to an Omeka S installation as linked open data, which then facilitates their sharing their data with the aggregator, Finna.fi. "This way small museums will also have the opportunity to publish their collections online using existing open source software. The project will pilot the use of Wikidata as a metadata repository, with Omeka S as an image bank." (WikiProject page).

This project involves upload and download of data to/from Wikidata. However, they are doing mapping on both ends between Wikidata Properties and Dublin Core, so it is not an example of using Wikidata as a schema. The project also monitors changes made to the Wikidata entries associated with their collections and then updates the ones in Finna. They have developed a REST API tool to do this.

To get the data into Wikidata, they used OpenRefine to do a bulk upload. New items are added to Wikidata using Cradle, but the process seems to be one at a time. Then, the metadata in Wikidata is harvested to Finna.fi using OAI-PMH. More details are available on the WikiProject page linked above.

Takeaways:

- **Mapping, Not Migrating:** While this project seemed relatively issue free, it is worth noting that the museums worked with a relatively limited set of Wikidata Properties (eleven main Properties and nine "other" ones) and mapped their existing data to these Properties specifically for sharing. It requires an extra step for them to share their data to Wikidata, but once the mapping is done it does not need to be repeated every time they share new records.

- **Use of Existing Tools:** It seems they were able to make this project happen with existing tools like Omeka, OpenRefine, and Cradle, except for the edit tracking tool that was developed by the University of Jyväskylä Open Science Centre.
- **Omeka as a Repository:** The project used Omeka instead of Wikimedia Commons as an image repository, facilitating greater control and avoiding the potentially messy interface with Commons. Omeka is open source software used mainly for smaller digital collections and online exhibitions. Although it is free, it does require a server and technical support to install and administer it.

The Metropolitan Museum of Art: Wikidata/Commons contribution strategies for GLAM organizations

The Met has been contributing a selection of its collections data to Wikidata. They hand-selected a thousand items to upload as “collection highlights” as well as eight thousand works from the Heilbrunn Timeline of Art History. This effort was led by Andrew Lih, who presented about it at Wikidatacon in 2019.

Here is a high level description of the process from Data Roundtripping: a new frontier for GLAM-Wiki collaborations by Sandra Fauconnier:

- On the museum’s side, data from The Met’s API and CSV is exported and converted to a Python (pandas) dataframe;
- On the Wikidata side, a SPARQL data export of The Met’s objects from Wikidata in JSON format is converted to a Python dataframe as well;
- Next, in a third, bi-directional comparison (diff), both dataframes are compared;
- Necessary updates to Wikidata are then exported to, and performed by, the QuickStatements batch editing tool, or with a bot (Pywikibot).
- Mappings between The Met’s metadata and Wikidata (for instance, vocabulary terms and names of creators) take place in a Google spreadsheet, which is updated and maintained by curators from The Met and by knowledgeable Wikidata volunteers.

The Met uses these Wikidata Properties (notably, they do not share provenance data with Wikidata):

Base set:

- Label
- Description
- Aliases

- Instance of
- Image
- Inception
- Creator
- Collection
- Inventory number
- The Met object ID

Nice to have

- Location
- Genre
- Material used
- Depicts
- Height
- Width
- Copyright status

According to Lih, these elements were selected to privilege findability over completeness, but getting the Met's data prepared for Wikidata still requires complex mapping, deduping against Wikidata, and transformations of existing collections data.

Here are slides from his presentation.

This Wikidata Linked Open Data Workflow also gives a sense of the steps and the options available to organizations that want to contribute their data to Wikidata.

Takeaways:

- **A Multi-step, Highly Technical Process:** The process of accurately and non-duplicatively sharing data with Wikidata is hardly seamless and requires a lot of technical intervention.
- **Mappings Are Still Manual:** Despite all the technological knowhow involved in this project, the mapping between the Met's data and Wikidata is maintained in a spreadsheet.

Interviews

We asked our interviewees what they thought about using Wikidata Properties as a metadata schema and the possibility of using Wikidata in concert with Wikimedia Commons as an online repository for smaller organizations looking to share their collections online and/or with Curationist. We expect to gather more perspectives from a wider audience of Wiki experts at our Metadata Learning & Unlearning Summit.

Evelin Heidel, Program Director, Wikimedistas de Uruguay and Curationist Board Member
Evelin has been working for over a decade in the Open GLAM (Galleries, Libraries, Archives, and Museums) space, including over four years at Creative Commons.

She identified a tension for GLAMs between the desire to share complete records and the limitations of aggregated metadata, like that available on Curationist, where, in her view, completeness isn't relevant. She sees the value of aggregation as making connections with other collections. She recommended aiming for simplicity, considering what the minimum relevant Properties might be, and leaving everything else aside. She advocated asking ourselves how Wikidata Properties would aid in the discovery process.

Evelin has seen that US GLAM institutions experience a bottleneck in getting their collections online: cultural heritage workers are overcommitted and there are not enough people to do the work of uploading things to Commons. She felt that using Commons and Wikidata as a bridge to Curationist was a good idea and that there is a good use case for it, but that doing so would make Curationist into a service provider. She encouraged doing a proof of concept project in the US and thinking about how it could scale.

Virginia Poundstone, Senior Product Manager, Wikimedia Foundation
Virginia was the former Platform Director at Curationist and is uniquely positioned to understand both Curationist's goals and challenges and the priorities and requirements of using and interfacing with Wikimedia products.

She recommended the combination of Wikimedia Commons and Wikidata as a free solution and a way for organizations to gain an online audience, with the caveat that it's not the best software for storing digital content in the long term. She said organizations should think of it as a "data lifeboat" or a "very public back up," rather than a primary archive. She said that there could be significant technical hurdles to sharing, however, as the Commons' Action API is old and cumbersome to understand, and updating it is not a priority for Wikimedia. She also emphasized that building anything interoperable with Wikimedia Commons, such as backend interfaces that connect to it, would require an engineer with Wikimedia experience. Although there aren't that many of these engineers, her number one recommendation was to work with a MediaWiki or Wikimedia Commons engineer. Still, she thought it would be a valuable service to smaller organizations and an opportunity to fill a gap that Wikimedia Foundation can't currently fill.

Takeaway:

- **Wikidata/Commons As a Bridge is a Good Idea, with Caveats.** Both interviewees found that using Wikidata/Commons as a low-cost way to get collections online was a good idea, with some caveats:
    - It would require sharing only a subset of data currently available
    - It is not a long-term solution for digital storage, and should not be an organization's only digital repository
    - It would involve significant technical hurdles, possibly turning Curationist into a service provider, and would require the skills of an experienced Wikimedia engineer

Preliminary Wikidata Properties Conclusions

Based on the research above, we have come up with the following conclusions.

- **Wikidata & Curationist Are Different Types of Repositories.** While Wikidata/Wikimedia Commons Properties and the Curationist schema have significant areas of overlap, they emphasize different things. Wikidata/Wikimedia Commons projects are more focused on data elements that are relatively concrete, often quantifiable, and rarely disputed, such as "date," "location," and "source" (source institution). By contrast, the data elements where Curationist makes the most impactful contributions are more subjective, such as "description," "cultural context," and "subject." In this sense, Wikidata and Curationist are slightly misaligned in the type of data they are designed to foreground. Wikidata was designed as a repository for authority records; Curationist was designed as a repository for artworks and artifacts. Although Wikidata includes records for many artworks and artifacts, it was not designed to hold more nuanced and discursive information about them, which is where Curationist excels.
- **Data Incompatibilities.** The vast majority of data currently in the Curationist database is not in a format that is shareable with Wikidata. We could map it to a schema based on Wikidata Properties, but would still not be able to upload it to Wikidata. Transforming the data in over 4 million records that are not consistent with each other would be a formidable task. From our perspective, there is no point in using a schema based on Wikidata Properties if we are not able to share our data with Wikidata.
- **Sharing with Wikidata is Complex.** Institutions that are already contributing their data to Wikidata have to perform multiple data cleansing, deduping, and other technical steps to prepare and upload it properly. Sharing data with Wikidata is not a task that should be undertaken lightly.

- **Wikidata/Commons Could Be a Waystation.** Wikidata/Commons could be used as a low-cost waystation for smaller organizations who want to share their data with Curationist, but providing the technical solutions and training to do so would be a very heavy lift. Curationist would also need to build the infrastructure to receive this data from Wikidata, which is not currently available. Finally, if Curationist were to recommend Wikidata/Commons as a waystation for an organization's metadata and images, it should emphasize that these repositories should not be the only places where these resources are stored, and that the organization should not use Wikidata Properties as their base metadata schema, but should use a schema that more closely suits their needs, both internally and externally.

Preliminary Wikidata Properties Recommendation

We do not recommend migrating Curationist's data to a new schema based on Wikidata Properties.

Converting the Curationist metadata schema to one based on Wikidata Properties will not necessarily facilitate the sharing of Curationist metadata with Wikidata. Also, Wikidata Properties are not internally consistent, are subject to change, and the language they use may not be appropriate or provide the proper framing for every collection. *Instead, we recommend creating a mapping between Wikidata and the Curationist schema that could be used to share Wikidata-ready metadata with Wikidata.*

Part Two: Wikibase for Controlled Vocabulary

Background

As noted above, Curationist has used Wikidata as a controlled vocabulary for all of the terms it adds to “Work” records on the site. (“Work” refers to any artwork, artifact, or document represented on the Curationist site.) The current process looks like this:

1. The Curationist archivist determines what terminology is needed or desired for a given Work record and searches Wikidata for that term
2. If the term or a synonym is available, the archivist determines whether the current Wikidata Label is respectful and accurate in regard to the entity it represents. Guidelines for terminology selection are documented in the Curationist Taxonomy Guidelines.
    a. If yes, the archivist adds the term to the Curationist Taxonomy spreadsheet and to the spreadsheet for the Work record.
    b. If no, the archivist confirms whether the desired term is already one of the Aliases on the Wikidata page.
        i. If an Alias will work, the archivist adds it to the Curationist Taxonomy and the Work record.
        ii. If there are no appropriate Aliases, the archivist adds the desired Alias to the Wikidata page and to the Curationist Taxonomy and Work record.
3. If the term is not found on Wikidata, the archivist adds the term to Wikidata, the Curationist Taxonomy spreadsheet, and the Work record.

This process involves shuttling between Wikidata, the Curationist Taxonomy spreadsheet, and the spreadsheet for the Work. We would like to streamline this process so that Wikidata is more seamlessly integrated with the CMS that will soon replace our system of spreadsheets. We envision this integration will not only make the process of term discovery and assignment easier, but also provide advantages for searching and browsing the Curationist site. Having a linked open data vocabulary on the Curationist backend will also have benefits on the frontend, enabling us to:

- Make all Wikidata terms assigned to Work records into links that bring up all other Works with the same term
- Link out to Wikidata so people can see the source of the term
- Facilitate better search by providing a readymade source of synonyms (Wikidata Aliases)

An example of what the first two items might look like can be seen on the Digital Culture of Metropolitan New York (DCMNY) site. The Wikidata terms on the right side of this record link to other items in the same collection with that term; the “W” icons link out to the Wikidata page for each term.

We investigated Wikibase as a possible solution to make these improvements through a review of project documentation, and also talked with the team at DCMNY to learn about how they implemented the connection to Wikidata.

Findings

Project Examples: Other Wikibases

We reviewed the documentation for twelve Wikibase projects that overlapped with our interest in Wikibase. Below are summaries of five we found to be the most useful in our evaluation of Wikibase as a solution for the Curationist vocabularies.

Rhizome ArtBase

Rhizome, established in 1999, utilizes Wikibase for managing its ArtBase, a comprehensive archive of born-digital art. Wikibase facilitates the storage of ArtBase data in the Resource Description Framework (RDF) format, which is structured around semantic “triples.” Each triple consists of a subject, predicate, and object, creating a graph of interconnected data that can be queried with flexibility and depth.

Technical implementation: Rhizome’s ArtBase employs a SPARQL endpoint with a graphical user interface for querying RDF data. This setup enables complex queries that allow for advanced data retrieval and interaction.

Takeaways:

ArtBase does not support faceted searching, so the reliance on SPARQL introduces a steeper learning curve for effective querying. Additionally, the current ArtBase interface lacks contextual relationships between the artworks, leading to only a single entry point for each work, reducing discoverability. Their implementation of Wikibase doesn’t fully align with our needs—we require a more complex setup—but the visual representation of SPARQL queries is intriguing.

Digital Scriptorium

The Digital Scriptorium is a consortium of North American institutions that aggregates their collection data of global premodern manuscripts. Wikibase is used to manage and enrich the metadata of these manuscripts by linking to existing authorities, enhancing their visibility and interoperability.

Technical implementation: Data from various consortium members is extracted and enriched using existing linked open data (LOD) authorities and vocabularies. Tools like OpenRefine are used to ensure that the metadata is interoperable and discoverable across institutions.

Takeaways:

While the Digital Scriptorium is a metadata aggregator, they do not correct or add to the source metadata. Each catalog entry is connected to a unique item within their Wikibase, allowing Digital Scriptorium to bypass Wikidata's notability requirement for item inclusion.

Smithsonian Institution

The Smithsonian Institution's PCC (Program for Cooperative Cataloging) Wikidata Pilot Project explores the potential of linked data for cataloging and authority control. The initiative sought to integrate the Smithsonian's extensive collections with global knowledge networks through the use of Wikidata, enhancing the discoverability and connectivity of Smithsonian resources. They explored using Wikibase as a central, open platform for managing names across its various units. This platform would help reconcile name data across the institution's databases, which currently operate with different standards and systems.

Technical implementation: As part of the pilot, the Smithsonian installed a local Wikibase instance to manage authority data, particularly for name authorities. The project focused on experimenting with linked data principles to create a more interconnected cataloging system. Similarly to Rhizome's ArtBase, the Smithsonian used RDF triples to allow for more interconnected data.

Takeaways:

The Wikibase installation is currently accessible only to Smithsonian staff, allowing for ongoing experimentation and refinement. The goal here was to foster interoperability between Smithsonian units without the need for complex manual intervention. Although they succeeded in enhancing the representation of Smithsonian organizations in Wikidata, they struggled to fully integrate publications due to the complexity of disambiguating author names and the labor-intensive process of linking publications to authors. Their work highlighted the challenges of working within a decentralized platform like Wikidata.

Luxembourg Ministry of Culture

The Luxembourg Ministry of Culture, in collaboration with eight GLAM institutions, uses Wikibase to share authority information across its member institutions. This project aims to create a unified and enriched data model that enhances the accessibility and impact of digitized cultural heritage collections.

Technical implementation: The project utilizes Wikibase for managing authority files related to persons, corporate bodies, and other entities. By mapping these authority files to the CIDOC-CRM ontology and validating them using Shape Expressions (ShEx), the project ensures that the data is both consistent and interoperable. This implementation also follows Wikidata's qualifier/reference structure, and is essentially an example of setting up a separate version of Wikidata.

Takeaways:

Most of the technical hurdles encountered in this project were related to mapping from CIDOC-CRM to the Wikibase data model, which is structured like Wikidata's. We would not have that issue because we would be sourcing our data from Wikidata and our current "authority file" is in spreadsheet form, which could more easily be reformatted for ingestion into a Wikibase.

OCLC National Libraries

Overview: The OCLC, in collaboration with sixteen member libraries, conducted a pilot project to explore the use of Wikibase for managing library metadata as linked data. The project, known as Project Passage, evaluated the framework's ability to reconcile, create, and manage bibliographic and authority data within a linked data environment.

Technical implementation: The Wikibase instance used in Project Passage included 1.2 million entities from various sources, such as Wikidata, VIAF, and WorldCat. The project demonstrated the platform's capabilities in supporting multilingual data, crowdsourcing, and linked data creation. Custom features like auto-suggest, SPARQL Query Service, and an Explorer UI were added to enhance usability. The pilot highlighted the need for interoperability between Wikidata and other Wikibase instances to ensure long-term sustainability and avoid isolation of data collections.

Karen Smith-Yoshimura summarized OCLC's experiments with Wikidata and Wikibase in this blog post ("Experimentations with Wikidata/Wikibase," Hanging Together, June 18, 2020). She offered this list of pros and cons:

Metadata managers noted that institutional buy-in is needed to support ongoing Wikidata/Wikibase work. Among the reasons for using Wikidata or Wikibase in the library environment:

- Expose institutions' resources to the larger web community
- Support institutional outreach to local communities
- Can create an entity description with a stable, persistent identifier immediately that can be reused by others
- Create labels in multiple languages and scripts and that are more respectful to marginalized communities
- Infrastructure supports collaboration across communities and countries
- Relatively low-barrier way to contribute to linked data and gain experience with "entifying"
- Tools are available such as the Reasonator, which displays Wikidata entries as well as related data and generates timelines that current library systems cannot

Among the barriers to using Wikidata or Wikibase in the library environment:

- Steep learning curve
- Uncontrolled metadata could result in inconsistent data quality
- Modeling and entities differ from library standards and practices
- The data you enter could be overwritten by someone else
- Duplicates or overlaps authority work
- Concern about scalability and long-term sustainability
- Installing a local Wikibase instance requires IT effort

Takeaways:

While our project shares many goals with OCLC's experiment, it is less focused on the need to be interoperable across multiple databases. Creating a "Curationist Vocabulary" that can be shared for others to use is not our main objective. Also, OCLC's concerns about the scalability and the IT expense required to implement and maintain a Wikibase are relevant to our project. If we are going to deal with the unpredictability and crowdsourced nature of Wikidata—whether we use Wikibase or not—it may make more sense to explore a direct connection to Wikidata rather than setting up a Wikibase in between Curationist data and Wikidata that then needs to be maintained and synced.

Use Cases for Institutional Wikibase Instances

This document was developed in May 2020, informally, by library staff at Columbia University, Harvard University, New York University, and the University of Pennsylvania.

It provides a number of use cases for Wikibase in order to help organizations decide whether Wikibase is appropriate for their needs.

The document describes the movement from string-based metadata (data expressed as strings of text) to a URI-based "identity management" approach (linked data), which is more efficient because it allows for centralized control of terminology. Because all terms are linked to a central authority, updates to terms can be enacted in one place that updates all records instead of across thousands of individual records.

According to this resource, **Wikibase would be appropriate** for what we want to do.

The following use cases cited in the document are relevant to our situation:

- **We have entities (people, mostly) that do not meet Wikidata's criteria for notability and identification.**
  We have experienced issues with other people deleting the terms we created on Wikidata because the terms were perceived to not meet Wikidata's criteria for notability (the degree of "known-ness" of a term that justifies its inclusion in Wikidata). Wikibase would allow us to create records for these terms and manage them locally.
    - Concern: "Custom displays or tools may need to be generated to enable staff users to effectively browse through non-unique access points, merge them, and update them. Tools for these purposes exist in the Wikidata ecosystem and could potentially be reused in a local Wikibase instance." ("Use Cases for Institutional Wikibase Instances," second use case)
- **We want to use locally preferred labels while pointing to the main Wikidata labels.**
  Sometimes (although these cases are increasingly rare), we want to use a different form of a term than the main Wikidata one. We usually add our preferred term to the Wikidata record as an Alias, but Wikibase would allow us to display our preferred term on our site while maintaining a link to the Wikidata entry.
    - Concern: Some custom properties and policies may need to be developed, but Wikibase supports this functionality out-of-the-box.
- **We want our terminology to appear as links in the Curationist public interface, both to other Works with the same terms and to the Wikidata page for each term.** In the "Use Cases" document, this is described as "Cross-platform discovery within a single institution."
    - Requirement: Integration of Wikibase with discovery layer.
    - Although Curationist does not have "multiple platforms," integration of the Wikibase with our Curationist public interface will enable us to link all

items that have a single term assigned to them, which is a useful tool for discovery.

- **Multilingual discovery**
  Although Curationist is not fully multilingual, Wikibase would support this functionality when it becomes available.
    - Requirement: Multilingual label support is built into Wikibase.
    - If we do start to add metadata in multiple languages, it would be supported by Wikibase.
- **Publishing of local thesauri and authorities for external use**
  Although this is a lower priority, Wikibase would allow us to publish our Curationist Vocabulary as linked open data for others to use, and would allow us to indicate on Wikidata which terms are part of the Curationist Vocabulary.
    - Providing deferenceable URIs and API/SPARQL access for local vocabulary terms so that others outside of the institution can use these in their own projects.
    - Requirements: Stable URIs for items, an API, and/or a SPARQL endpoint, all of which are built-in to Wikibase.
- **Pipeline to Wikidata and broader web discovery**
  Eventually, we could share our local terminology, and potentially its links to Works, to Wikidata, which would help improve the cultural heritage ecosystem.
    - Requirements: Adaptation of existing tools like QuickStatements, OpenRefine, PyWikibot, or development of new tools to port data from a local Wikibase into Wikidata. Development of federated Wikidata properties by Wikimedia is a step in the right direction here (see "Federated properties" below).

Takeaway:

According to this use case assessment document, Wikibase would be a good solution for Curationist, as it fulfills many of the requirements we have for our descriptive vocabularies.

Relevant Projects and Tools

Federated properties

This feature would allow us to use the same Properties as Wikidata in our Wikibase. It would ensure that our Wikibase has the same Properties (and their associated "P" numbers) as Wikidata. This alignment would facilitate syncing our data with Wikidata and future sharing of our data back to Wikidata. There are a few considerations:

- This feature is currently for testing only.
- It is turned off by default in Wikibase. Here are instructions to turn it on.

- This feature cannot be used with local properties. We would not be able to add properties of our own, but would be stuck with Wikidata Properties.

Wikibase Sync

The QA Company has developed an open source tool that syncs a local Wikibase with Wikidata:

- This tool could be very useful keeping our Wikibase in sync with Wikidata and seems quite powerful: We could clone records from Wikidata and add to them—our local terminology preferences, hierarchies, etc.—in our Wikibase.
- We reached out to QA Company for more information and emailed with Dennis Diefenbach <dennis.diefenbach@the-qa-company.com>:
  When you search a Wikibase with their tool enabled, it also searches Wikidata. If the item exists in Wikidata but not on the Wikibase, selecting it in the search dropdown will clone it to Wikibase. Changes that are local are prioritized, but if there is a change in Wikidata, the tool will try to reconcile it. If you search the Wikibase for a term that is a synonym in Wikidata, it will bring up the synonymous Item in the Wikibase, even if the synonym is not in the Wikibase.

Takeaways:

Wikibase appears to be supported by a robust community of people creating open source tools that enhance its interoperability and functionality, particularly as it interfaces with Wikidata. There is no indication that the popularity and widespread use of Wikidata is slowing, but the future of these tools is uncertain and depending on them involves a certain amount of risk.

Interviews

Allison Sherrick and Diego Pino, Digital Culture of Metropolitan New York (DCMNY)

Allison and Diego have built a feature into Archipelago, their open source digital repository software, that reconciles source terms with Wikidata by querying Wikidata live every time new metadata is imported into the system. Metadata is imported as a spreadsheet which is mapped to JSON. They have been able to reconcile up to 10,000 terms at a time without throttling, and in general have found that their reconciliation process is faster than the one in OpenRefine, which provides similar functionality. (Sharon has used OpenRefine reconciliation and she could only reconcile a few hundred terms at a time.) Their system finds and reconciles only unique terms from any one ingest, whereas OpenRefine tries to reconcile every term individually, even if there are duplicates. Like Curationist, they are able to store and display both the term from the source institution and the reconciled term from Wikidata (or the other vocabularies they reconcile against, including those of the Getty and the Library of Congress).

Building the reconciliation feature was not difficult once they had created the plugin interfaces to fetch data from the remote sources. Since Archipelago is open source, developers can review the code to see how they did it.

When asked about Wikibase as an intermediate terminology repository, they stressed that trying to keep a Wikibase synced with Wikidata would be “pain, pain, pain.” They noted that the German National Library started out using Wikibase and now just uses Wikidata.[1] In their system, they reconcile terminology once when the metadata is ingested, but do not continuously update the terms displayed when Wikidata changes. If there is an occasion to reingest the metadata, it is re-reconciled with Wikidata at that point. They are not adding or creating their own preferred terms as Curationist does, however.

Evelin Heidel, Program Director, Wikimedistas de Uruguay and Curationist Board Member

Evelin thought it made a lot of sense to have a Wikibase instance that maps to Wikidata and not to rely on Wikidata alone. She thinks Wikidata has some scaling and capacity issues, particularly around queries and in its interactions with Wikimedia Commons. (Some of this may have to do with the majority of the data on Commons being in Wikitext format, and thus not easy to query using SPARQL.) She thinks having a Wikibase would allow for more effective and useful querying of the data that perhaps can’t be done fruitfully in Wikidata because it is so big and varied.

Mitchell Parsons, Curationist Developer

We asked Mitch for his thoughts on how feasible it would be to create a Wikidata connection like the one in Archipelago, versus setting up a Wikibase to manage our terminology. He said that our main Curationist database is “a NO-SQL db (AWS DynamoDB) for our source of truth, and we have a read replica in another database called OpenSearch for just search queries.” He also said, “We can use any database we want for queries or creating views of the data,” and that he sees Wikibase as the easiest solution from an implementation perspective because it is readymade for the purpose of managing terminology. With its use of Docker (See “Wikibase Suite” section under “Technical Considerations” below.), Wikibase would be relatively straightforward to install. From his perspective, it makes the most sense to set up a Wikibase and keep it synced with Wikidata, perhaps using The QA Company’s WikiSync tool. There will be other maintenance and support needs, including setting up a schedule to sync the

[1] The German National Library was looking into using Wikibase for its authority data, or GND, but the project is currently listed under “Dormant & completed projects” (“Ruhende & abgeschlossene Projekte”) on the GND-Wiki. There is no mention of Wikibase, Wikidata, or a SPARQL endpoint on the current webpage for the GND.

Wikibase and keeping the Wikibase software up to date, but making this improvement involves additional work no matter what.

Curationist Archivist Perspectives

In addition to interviewing others, we realized it was important to do some self-reflection about our own experiences creating, editing, and managing Curationist-supplied metadata. We came away from this process with three considerations for our Wikibase decisionmaking:

- **Wikidata changes… eventually.** Although initially we couldn't find some of the preferred terminology we wanted to use on Wikidata, we've found that over time Wikidata often eventually changes to reflect our preferences. This suggests that the use case where Wikibase allows us to use our own version of a term may be less appealing, as Wikidata may eventually change to include that term.
- **The importance of hierarchy.** We've found that understanding where a term falls in a hierarchy—between terms with broader and narrower meanings—is helpful for selecting appropriate terms. Wikidata is not great at expressing hierarchy. Due to its crowdsourced nature, its hierarchies are inconsistent and sometimes circular. Having our own Wikibase would allow us to control the hierarchical relationships between terms, facilitating better and easier term selection than if we were using Wikidata alone.
- **Organizing terms by metadata element.** In our current Curationist Taxonomy, terms are grouped according to the metadata element where they are most likely to be used. Although terms can be used in any metadata element, this grouping helps Curationist archivists select appropriate terms more efficiently.

Takeaways:

Our interviews and self-reflection suggest that Wikibase may be both the easiest and most desirable solution for storing, managing, and sharing our terminology, although keeping it synced in a useful way with Wikidata will require ongoing maintenance and dedicated staff time. Both Evelin and Mitch highlighted the ease of querying and viewing the data as an advantage, although it is unclear how much we would be doing this aside from term selection. Certainly having a local Wikibase may have advantages for faceted search, although the number of terms we are currently managing is relatively small compared to the terms sourced from contributor institutions. However, despite the small number of terms currently in the Curationist Taxonomy, being able to assign and control the hierarchical organization of terms and associate them with Curationist metadata elements would be a helpful feature for Curationist archivists.

Technical Considerations

If we decide to go with Wikibase, there are three different options for installing/using it:

- Basic Wikibase
- Wikibase Suite
- Wikibase Cloud

The first two require a server with compatible hardware and software installed. The third, Wikibase Cloud, allows you to create a Wikibase hosted by Wikimedia.de in Germany.

Wikibase On Our Server

There are two ways to install and run Wikibase. The "Basic" installation is just the Wikibase software; "Wikibase Suite" is "containerized" and includes other services that augment the use of Wikibase, including QuickStatements and Elasticsearch.

"Basic" Version

Wikibase is an extension of MediaWiki; in order to install Wikibase, you must first install MediaWiki. Installation instructions for this "Basic" version are here. It uses Composer, a dependency manager for PHP libraries. Other requirements from the instructions include:

- A web server software: "MediaWiki is broadly compatible with all major web servers that can invoke a compatible version of PHP. Most installations use the Apache HTTPD web server. Nginx (configuration example) is a good choice as well."
- PHP: "For the latest stable version of MediaWiki, at least **PHP 7.4.3** is required. See the page on Compatibility for further information."
- Database server: "You will need *one* of the following database servers to run the latest version of MediaWiki:
    - MariaDB 10.3.0+ or MySQL 5.7.0+
    - PostgreSQL 10.0+
    - SQLite 3.8.0+

  Using MariaDB or MySQL is recommended as Wikimedia uses MariaDB."

Wikibase Suite

Wikibase Suite is much more well documented. From GitHub: "**Wikibase Suite** (WBS) Deploy is a containerized, production-ready Wikibase system that allows you to self-host a knowledge graph similar to Wikidata. In addition to Wikibase on MediaWiki, WBS Deploy includes the Wikidata Query Service (WDQS), QuickStatements, Elasticsearch, and a Traefik reverse proxy with SSL termination and ACME support. The service orchestration is implemented using Docker Compose." There is also a Wikibase Docker installation tutorial.

Requirements for Wikibase Suite include:

- Hardware
  - Network connection with a public IP address
  - AMD64 architecture
  - 8 GB RAM
  - 4 GB free disk space
- Software
  - Docker 22.0 (or greater)
  - Docker Compose 2.10 (or greater)
  - git
- Domain names

  You need three DNS records that resolve to your machine's IP address, one for each user-facing service:
  - Wikibase, e.g., "wikibase.example.com"
  - QueryService, e.g., "query.example.com"
  - QuickStatements, e.g., "quickstatements.example.com"

Wikibase Cloud

Wikibase Cloud is a complete installation of Wikibase hosted by Wikimedia.de in Germany. The obvious advantage of using a cloud-based Wikibase is that no server maintenance or installation is required. Sharon was able to create a test Wikibase in Wikibase Cloud without any technical knowledge of server hardware or software.

However, we have some reservations about Wikibase Cloud. First, because it is hosted in Germany, we would need to create an imprint for the Wikibase, or "a legally mandated statement of ownership and authorship that must be displayed on websites and that contains information about the site's publisher." While Curationist could certainly provide this information, there may be some legal consultation required to ensure the imprint is done properly. Second, it is unclear whether third-party software such as that developed by The QA Company or others can be run on Wikibase Cloud. For example, to use The QA Company's WikidataSync tool, a "manifest" is required and Sharon was unable to figure out how to create one in Wikibase Cloud.

Migration

Once a Wikibase is set up, we would need to migrate our existing terminology data into it. Our Curationist Taxonomy is currently in a spreadsheet format and would need to be reformatted in order to populate a Wikibase. We could do this using QuickStatements but would have to learn its syntax and reformat the spreadsheet accordingly, which might involve some trial and error.

Takeaways:

Implementing Wikibase on a server that we manage would require technical expertise, not only for setup and configuration but for ongoing maintenance and support. Implementing Wikibase Cloud would require less technical support and investment but would likely be more limited in its functionality and interoperability with other tools and the Curationist database. In either case, IT resources would need to be dedicated to the ongoing support and management of a Wikibase installation to ensure that it remains useful and relevant. There would also be a learning curve for digital archivist staff in moving from the spreadsheet data to Wikibase, both in terms of setup and ongoing use and updates.

Preliminary Wikibase Conclusions

We are uncertain! While Wikibase seems like a great, out-of-the-box solution for managing our terminology, it is not without its drawbacks. We considered arguments on both sides—installing Wikibase, and creating a direct connection to Wikidata.

On the one hand, a direct connection to Wikidata, like the one created by the Archipelago team, seems most expeditious, obviating the need to set up a Wikibase and taking advantage of the crowdsourced and constantly changing updates to Wikidata. This approach would also eliminate the risk of a Wikibase becoming a “dead-end” repository that remains static and only serves Curationist. However, connecting directly to Wikidata would require significant development work to design, implement, and maintain, and we would be at the mercy of Wikidata updates and structure that might make finding and assigning the correct terminology more difficult than in our current, spreadsheet-based system.

On the other hand, setting up a Wikibase would allow us to use whatever terminology is most appropriate, and to control the hierarchy and metadata elements within which terms are relevant. There are tools to keep our Wikibase synced with Wikidata, although there is always the risk that, without regular maintenance, our Wikibase would fall out of sync with Wikidata and become a metadata backwater that is only relevant to Curationist. (We are experiencing that now with our spreadsheet.) Although Wikibase Suite comes with many included tools and features, it would still require installation and ongoing maintenance—both technical and in terms of metadata—to keep it up to date, functional, and useful.

For these reasons, we are undecided as to whether Wikibase is the right solution.

Learnings from the Metadata Summit

On September 19, 2024, we presented this report to a panel of Wiki experts:

- Anni Saisto, Pori Art Museum
- Ben Vershbow, Wikimedia Foundation
- Christos Varvantakis, Wikimedia Deutschland
- Evelin Heidel, Curationist board member
- Hanno Lans, CopyClear
- Sandra Fauconnier, GLAM-Wiki Specialist
- Siobhann Mccafferty, Australian Research Data Commons

We asked them for their initial impressions and responses without having read the report in advance. We also received a lot of great information and suggestions (and offers of help) in the chat. Here is a summary of the suggestions and recommendations.

Speakers did not engage with the question of using Wikidata Properties as a schema, since we have decided that is not a good option for Curationist.

They did suggest using Wikibase, although Christos warned it will not necessarily be easy to ensure it is properly mapped or cross-referenced to Wikidata. Sandra and Christos both suggested creating a test case with Wikibase Cloud or Wikibase to do a proof of concept.

Siobhann and Anni reminded us that there is no “right” choice, just the choice we make and then being nimble and open to making changes. They suggested that being in conversations with the Wiki ecosystem was a way to keep from making too many big missteps. Christos and Sandra agreed, suggesting we continue to have conversations with Wikimedians and third-party developers in the Wiki system.

Evelin suggested having conversations with our contributors to ascertain if they are having similar problems. She said that large institutions have a lot they can contribute and bridging the gap between them and smaller organizations would be an interesting role for Curationist. Christian Dawson cautioned against letting large organizations be gatekeepers while maintaining a degree of collaboration, in particular the donation of metadata enhancement back to our source institutions. Evelin noted in the chat that this is often difficult because the source institution can be overwhelmed with the updates/changes.

Anni suggested we could use the rich data in Wikidata on artists to make interesting connections between the Works on our site. We agreed, but thought this is perhaps a step or two down the road from where we are currently.

Sandra mentioned in the chat that she would be willing to help us work with art and culture collections that are not in institutional collections, as she does a lot of work in Wikidata and Wikimedia Commons on public art.

There were a number of suggestions and helpful links in the session chat. Here are a few of the ones we found most relevant:

- James Hare, who has “experience managing large data migrations to Wikidata and Wikibases so I would be happy to help,” suggested that Wikibase with a cross-reference to Wikidata seemed to be the way to go.
- Laurence Parry noted, “Federated properties can help there, using Wikidata's properties, although the v2 solution for that which allows mixing local and foreign properties is not quite finished.” But James responded: “Federated properties adds too much complexity; it is much simpler to maintain a soft coupling between local Wikibase version of a property and Wikidata version.”
- Laurence also provided some links for new Wikibase users:
    - Wikibase.cloud Telegram chat if you need some help/advice - https://t.me/joinchat/FgqAnxNQYOeAKmyZTIld9g -
    - and a mailing list as well at https://lists.wikimedia.org/postorius/lists/wikibase-cloud.lists.wikimedia.org/ .
    - For Wikibase software it's https://t.me/+WBsf9-C9KPuMZCDT and https://lists.wikimedia.org/hyperkitty/list/wikibaseug@lists.wikimedia.org/

The session was recorded and you can view the recording here.

Overall Recommendations

We learned a lot about Wikidata, Wikibase, and Wikimedia Commons in the course of researching this report, and no doubt, there is still a lot to learn. However, based on our research, we offer the following recommendations.

- **Wikidata Properties as a schema: No.** Curationist should not migrate its metadata to a new schema based on Wikidata Properties, and instead should create a mapping between its current schema and select Wikidata Properties and

Wikitext Templates. These mappings will facilitate the sharing of Wikidata-ready metadata currently on the Curationist site to Wikidata, and between potential Museum Services partners who may use Wikidata and Wikimedia Commons to share their metadata and images.

- **Wikibase for controlled vocabularies: Yes, with reservations.** Our ideal solution for managing descriptive terminology would be integrated into our CMS. It would allow archivists to search the Curationist Taxonomy *and* Wikidata to find appropriate terms that correspond to each metadata element, and then would allow the archivist to apply that term, along with its QID, to a record. Each term would then appear in the public interface as a link that would bring up all the other Curationist records to which that term has been applied. Ideally, there would also be a secondary link to the term's page on Wikidata or in the Curationist Wikibase. Having our own Wikibase would allow us to use the exact term we want, regardless of how the main term is expressed in Wikidata, and control the hierarchy and metadata element within which the term appears to facilitate searching by an archivist. Our main reservation about using Wikibase relates mainly to the IT and metadata resources required to keep it synced with Wikidata, which is changing all the time. The risk of establishing a Wikibase that is not regularly synced to Wikidata is that over time it will become an isolated resource that doesn't take advantage of linked open data connections and crowdsourced updates.