Wikidata Search & Discovery project for Oxford GLAMs: final report

Martin Poulter
September 2019

Background 2
Project overview 3
Linked data in Google results 6
Reach metrics 9
  Google visibility problem 9
  Views on the proofs of concept 10
  Referrals to other sites 11
  Sharing by other institutions 11
Sustainability and crowdsourcing 12
  Volunteer engagement on Wikimedia platforms 12
Shaping working practices 14
  Training 14
  Documentation 14
  Answering queries 15
  Supporting research 15
Suggested future work 16
Summary and conclusions 17
Background

Employing one person half-time for two years, the Wikidata project explored the benefits of sharing Oxford GLAM collection records as Linked Open Data on the Wikidata platform. These data described paintings and sculptures from the Ashmolean, manuscripts from the Bodleian, drawings from the Pitt Rivers Museum and botanical artworks from the Sherardian Library.

On Wikidata, these records join up with many other data sources, including artist biographies and authority files such as VIAF or ULAN. Sharing these data made them visible through a variety of platforms and APIs. This project created three Wikidata-driven interactive sites, each giving multiple new visualisations of those collections. Independently of those proofs-of-concept, the data feed other sites and tools, including Google search results.

Part of the Wikidata representation of the Shahnamah of Ibrahim Sultan, a manuscript in the Bodleian collection. This view is created by Wikidata’s own visualisation tools.
Project overview

The data sets shared included:\footnote{The project wiki page has more details of the data sets, events and writings. \url{https://www.wikidata.org/wiki/Wikidata:GLAM/Oxford}}

- 3,409 items of Eastern art from the Ashmolean’s Jameel Centre database
- 2,106 works by the illustrator Ferdinand Bauer that are on Digital Bodleian
- 2,003 manuscripts and paintings from the Bodleian Libraries
- 319 astrolabes from the History of Science Museum and other museums
- 43 drawings in the Pitt Rivers Museum

Wikidata is a machine-readable knowledge base currently describing 65 million items, including around a million from cultural collections. These sites are open for editing by anyone and so serve as platforms for public engagement (described below). Joining up with other data sets in Wikidata gave us access to:

- Biographical links about artists and creators, including the Benezit Dictionary of Artists, Britannica and the Getty Union List of Artist Names.
- Alternate forms of names and names in alternative languages
- Authority-file identifiers such as VIAF (the Virtual International Authority File) and EMLO Location (Early Modern Letters Online)
- Co-ordinate locations of places
- Freely-reusable images of artists
- For plant species, Wikidata gives taxonomic information (their connections in the tree-of-life) as well as the common names of the species in multiple languages, links to the species’ entry in botanical databases and links to present-day photographs of the plants. A plan to link to the Oxford herbaria database was not achieved because that database did not allow incoming links to specific records.

The three proof-of-concept sites are:

- \textit{Collection Explorer}: combining items from the Bodleian, Ashmolean and Pitt Rivers Museum, it show galleries of items associated with a place, creator or depicted entity.
- **Astrolabe Explorer**: showing astrolabes, their owners, and places of origin with a map and timeline.

- **The Sibthorp-Bauer Expedition**: a combined map and timeline for this 18th century botanical expedition, showing the paintings made at each location and giving additional information about each species.

Collection Explorer was not explicitly publicized because of its prototype status. Its purpose was to be indexed by Google and to refer traffic to the target catalogues. Astrolabe Explorer was promoted within the scholarly community by email.

The Sibthorp-Bauer site was promoted by the Bodleian Libraries’ Twitter account and gained 15 replies, all positive, among a total of 675 engagements.

The data can be accessed through other tools, in an ecosystem which is continually developing:

- Wikimedia Commons and some versions of Wikipedia use Wikidata to generate ‘infoboxes’ with key facts about art works and manuscripts.

- [Reasonator](#) is an online application that summarises what Wikidata knows about an entity. The profile for an artist in Reasonator ([example](#)) shows links to related entities, connecting artists to their works and connecting people and places to works that depict them.

- [Crotos](#) is an online application for browsing art works from hundreds of collections by many criteria. Items appear when there is an image of them in Wikimedia Commons as well as a representation in Wikidata.

- [Sum of All Paintings](#) gives an overview of paintings that are known to Wikidata. This list currently includes 1079 paintings in the Ashmolean, just under 1000 of which were provided by this project.

- Wikidata Query Service takes custom queries in the SPARQL language, such as *[Benezit Dictionary of Artists](#)* links for Ashmolean artists, or IIIF links for manuscripts in Latin.

- Various voice and natural-language interfaces for Wikidata have been built. [QAnswer](#) at the Université Jean Monnet accepts typed questions in ten languages. Asked for ‘paintings in the Ashmolean Museum’ or ‘astrolabes made of copper’, it provides an image gallery. [Platypus](#), hosted by École normale supérieure de Lyon, accepts typed
and voice questions in English and French. Asked for ‘depictions of Shiva’ it provides a long list including items from the Ashmolean and the Bodleian.

QAnswer showing images and links for History of Science Museum objects in response to a natural language request

Other aspects of the project involved embedding Wikidata and Wikimedia expertise in the university via training events and documentation, explained in later sections of this report.
Linked data in Google results

Wikidata’s tens of millions of items and billions of properties constitute what is called a knowledge graph. This knowledge graph is available without restriction to other software, most notably Google and Wikipedia. Google’s knowledge graph shapes how search results are presented and refined. Although this is harder to verify, some of the knowledge graph used by Google seems to be drawn from this project’s Wikidata contributions.

That Wikidata is of interest to Google is shown by the fact that, as of September 2019, Google is employing its own Wikimedian In Residence with a focus on Wikidata.²

Consider a search for the Chinese painter Ren Xiong who is represented in the Ashmolean’s collection. An Ashmolean page appears 30th in these Google results, with the artist’s Collection Explorer profile appearing 23rd. So users will only see Ashmolean-specific information on the third page of results. However, on the front page alongside the top search results is a Knowledge Graph box.

![Knowledge Graph Box](image)

This suggests some artworks. Clicking on ‘Orange flowers, loquats, and garlic’ takes us to refined results where an Ashmolean link is the top hit.

---

Alongside these search results is a Knowledge Graph box for the artwork itself, showing that Google recognises this painting as an entity in its own right. Of the three specific art works on offer, the third is ‘Bird sitting on a branch with pink flowers’. This link again gives us refined results, where the Ashmolean catalogue link is third, and a Pinterest pin of the catalogue page is second.

Similarly, a search for Wu Hufan has the Ashmolean catalogue at 20th place, but the Knowledge Graph box mentions some works, and clicking on ‘Magnificent view of landscape’ brings up refined results in which the relevant Ashmolean catalogue link is fourth.
A search for ‘Lou Baian’ brings up this suggestion:

The Wikidata entry for this artist, with his description and birth date, was created as part of the process of adding the Jameel data. Clicking the link gives slightly different results in which the Collection Explorer profile of Lou Bai’an is the fourth result.

By no means all the artists mentioned in Collection Explorer have these links visible in Google searches, and we have seen that Knowledge Graph results differ greatly in their content.
Reach metrics

One way to measure the success of the artist profiles on Collection Explorer is to compare their Google ranking to that of the catalogues to which they link (with Google’s personalisation switched off). Doing this for the hundreds of creators would be a large project in itself. From a sample, the results were extremely mixed, for example:

<table>
<thead>
<tr>
<th>Artist</th>
<th>Google rank of Collection Explorer</th>
<th>Google Rank of Ashmolean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang Liuqiao</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>Mori Yusen</td>
<td>8</td>
<td>Not in first 100 results</td>
</tr>
<tr>
<td>Zhou Jingxin</td>
<td>14</td>
<td>34</td>
</tr>
<tr>
<td>Feng Kanghou</td>
<td>16</td>
<td>27</td>
</tr>
<tr>
<td>Xiong Hai</td>
<td>Not in first 100 results</td>
<td>8</td>
</tr>
</tbody>
</table>

There are a lot of results for which neither site appears in the early results, often because the name of the artist is shared by other people.

Google visibility problem

The Google visibility of the prototypes was harmed by an error that was spotted late in the day. Since Wikidata stores labels (names) of items in hundreds of human languages, a feature was implemented early on where the applications detected the user’s language and retrieved names and labels in that language. Thus a Japanese visitor would get Japanese versions of the artist name and short description, as well as Japanese names for types such as ‘painting’ or ‘statue’.
It turns out that Google penalises sites if the same page gives different content to different users. I learned this in August 2019 via a Google employee. Thus the reach of the sites was inadvertently harmed by a feature intended to ease their use by a wider audience of people. This was a search-engine optimization (SEO) mistake rather than a problem with the open data platform.

The sites could be coded to fix how content is served in different languages, but there was not enough time in the closing weeks of this project to implement that change across all the sites. So I 1) created multiple pages for different languages’ common names of species in the botanical explorer; 2) fixed the language of other sites to English. This does not change that other language data are still retrievable directly from Wikidata, or via other Wikidata-driven tools.

Views on the proofs of concept
Over the August 2018 to July 2019 academic year, the proofs of concept received 11,308 page views from 1,907 users, according to Google Analytics. The pages per session averaged 4.31, meaning that users visited multiple pages of the site as densely interconnected structure encouraged them to do. Of these views, 39% were for the botanical explorer, 31% for Collection Explorer and 30% for Astrolabe Explorer. The users’ browser languages were 62% English, 38% other languages. Twenty-two percent of the users were based in the UK.
Referrals to other sites

In the 2018/19 academic year, the proofs-of-concept were responsible for 0.43% of users referred to the Ashmolean’s Jameel Centre web site. In terms of time spent on the Jameel site, users arriving by this route spent on average 4:31, compared to 2:44 for users in general. On Digital Bodleian, where a much smaller proportion of the content is described in Wikidata, 0.14% of the users came from the proofs of concept. As explained in the previous section, Wikidata’s main effect may be on Google, which is the source of 48% of the Jameel Centre users and 15% of Digital Bodleian’s users.

Sharing by other institutions

Another metric of success is the sharing by other institutions that was inspired by the activity in Oxford. The History of Science Museum sharing images and data about 165 astrolabes prompted the Science Museum and the Adler Planetarium to share data and images about their own collections. These 122 additional astrolabes brought the total described on Wikidata up to 700.
Sustainability and crowdsourcing

Wikimedia platforms are editable by the general public, so although this was not promoted as a crowd-sourcing project, an element of crowd-sourcing was inevitable. A ‘recent changes’ tool summarises the edits made to items within Wikidata. An open question was whether the records we shared would be vandalised or would improve over time. Although there were a few detrimental edits (such as merging together the records for two different astrolabes), these seemed to be genuine mistakes and were reverted by other users. Below are examples of the positive improvements made.

For the Jameel Centre artworks:
- Adding labels and descriptions in other languages
- Identifying the war between Rama and Ravana as the subject of the artwork ‘Rama and Ravana doing battle’
- Adding Shakuntala, lotus leaf and writing implement as depicted items in the artwork ‘Shakuntala writing a love letter on a lotus leaf’

For the astrolabe data:
- Adding measurements to an astrolabe, based on the online catalogue record
- Changing a ‘material’ property from ‘Juglans’ (the species of tree) to ‘walnut wood’ (the material)
- Adding foreign-language labels
- Creating a record for an astrolabe in the Mathematisch-Physikalischer Salon, Dresden, using a photograph that had been uploaded to Wikimedia Commons by another user
- Adding location of assembly/creator/material properties to summarise the Wikipedia article about an astrolabe, in turn based on the museum web site
- Adding ‘described by source -> Heilbrunn Timeline of Art History’ to one astrolabe

For Bodleian manuscripts:
- Adding Russian, Georgian and Turkish names and descriptions to Georgian manuscripts
- Adding Hebrew labels and descriptions to the Hebrew Avicenna Canon

Volunteer engagement on Wikimedia platforms

Since 8,000 Bodleian images were shared on Wikimedia Commons in 2015, 340 volunteers have made edits to the files, including changing category tags and translating labels. The three top volunteers have made more than a thousand edits each. These numbers do not count volunteers who made use of the images, for instance by adding them to a Wikipedia article with a caption. In the current project, we have also seen volunteers editing the Wikidata representations of Oxford’s collections, in the process creating new pathways by which items can be discovered. A few active users were profiled in the article ‘The hyper-engaged Wikidata volunteers helping Oxford GLAM’ but these are just a sample of a much larger group.
These highly-engaged and productive volunteers maintain and improve the platforms through which most people find out about art, and do work for which staff lack capacity. GLAM institutions ought to actively track these volunteers (using the tools that Wikimedia makes available) and include them in engagement metrics. Siobham Leachman — a New Zealand-based Wikidata volunteer who shares knowledge about biodiversity and museum collections — has won recognition including a travel scholarship and a medal from the institutions she benefits.\(^3\) Thanking and celebrating volunteers this way should be a lot more common, and they should be considered as stakeholders in digital activities. In the terminology of the GLAM Audience Framework, they are facilitators and engaged explorers.

During this project, Wikimedia Commons introduced a feature that allows users to rapidly tag what is depicted in an image, using Wikidata identifiers. I created a software module, ‘Depict-o-tron’, showing that Wikidata could be used to auto-suggest and record depictions for images that are available through an IIIF server, storing the depiction data on Wikidata. New Wikidata-driven applications that let users tag depictions in artworks (Wikidata Art Depiction Explorer and Wikidata Image Positions) have since come along and represent a more fully fleshed-out version of this idea.

Two kinds of depiction data could be crowd-sourced. Simple judgements of whether a painting is a portrait or not, or depicts a tree or not, can be done by almost anyone and can be assisted by machine learning. This is the approach taken by a Met/ MIT/ Microsoft project in which an online community use a Wikidata-driven mobile game to verify depiction statements that are generated by feeding the Met’s digitised paintings to Microsoft’s object recognition service.\(^4\)

There are more specific judgements, such as identifying statues that depict the Maitreya Buddha or still life paintings that show a particular type of porcelain vase. The expertise for these statements is held by a small part of the population, but the Wikidata crowd-sourcing platforms may be the most efficient way to engage them. The data created by these crowdsourcing tools can go far beyond the level of detail in the catalogue, and can reflect a different set of interests. The data are available to many of the Wikidata-driven applications mentioned in this report, such as Reasonator, Crotos, and custom queries of Wikidata. Hence these depiction data make the artworks more visible.

The aspects of a work that interest someone may not be those highlighted in the catalogue record. A painting may be catalogued as a portrait of a particular samurai, but there may be people interested in the weapon he is carrying, the clothes he wears, or in the flora and fauna depicted, rather than the person.

Shaping working practices

By giving training events, blogging and answering queries, I have helped researchers and librarians get data from Wikidata and shown how the different Wikimedia projects work with each other to extend the reach of images, text and data.

Training

Among the training delivered:

- Lectures in the Digital Humanities at Oxford Summer School (in 2017, 2018 and 2019) on using Wikidata in Digital Humanities, the most recent in an optional session chosen by 80 attendees
- A training workshop for GLAM staff on SPARQL basics, using Wikidata queries for items in Bodleian and Ashmolean collections
- A training session at a Digital Humanities hackathon hosted at TORCH
- An overview session on Wikidata for the Oxford e-Research Centre
- A lecture and training session for GLAM staff on Wikidata as a platform for museum catalogue data
- Two seminars on the overlaps between open data, open culture, open science and open education.
- A workshop in the Berlin/Oxford Open Science Summer School about Wikipedia and Wikidata

I also presented at the Museums+Tech conference in London, at the GLAM Digital Showcase and in various meetings, including the GLAM Collections Committee and groups within IT services dealing with GLAM catalogues.

Documentation

Seventeen blog posts, totalling 19,000 words, demonstrated what can be done with the data sets and reported how GLAM institutions around the world are using Wikidata. One theme was how Wikidata interacts with the IIIF image format, by pointing to the IIIF representations of collection items, and identifying the region of an image that depicts a particular entity.

Japan Search, a Linked Open Data interface to Japanese museum collections, launched during this project. I added Japan Search identifiers to Wikidata and worked out how to syndicate between the two platforms. This means that in principle Japan Search could harvest relevant Oxford items, and the Ashmolean could link to other works by its Japanese artists.
Answering queries

This is a selection of the many queries about Wikimedia platforms from within the GLAM Division and other parts of the university:

- An Ashmolean cataloguer wanted an authority file of painters who died before 1800, with their nationalities and genders. A Wikidata query generated 12,000 such painters.
- I helped the Bodleian German subject librarian get data about authors and locations associated with a corpus of German literature.
- In response to a request from BDLSS, I shared links to entries in the Bodleian’s manuscript catalogues, so the Wikidata entry for a manuscript includes the manuscript catalogue link, the Digital Bodleian link and the IIIF manifest link.
- I advised on metadata considerations for a project looking at preservation of Digital Humanities projects.
- I advised the Cultures of Knowledge team on getting Wikidata links into their own database, getting more of their own identifiers into Wikidata, and using Wikidata as a hub to get other kinds of identifier.
- I advised the OxLOD (Oxford Linked Open Data) project about using Wikidata to find identifiers for items in data sets.

Supporting research

I am a DPhil candidate in the School of Geography and the Environment studying how people’s attitudes towards different species of plants and animals impact efforts to conserve wildlife and biodiversity. My research uses digital ‘big data’ to track people’s attitudes. As one of the biggest open-access websites in the world and a unique resource for research, Wikipedia has formed the basis of much of my work [...] I have used Wikipedia data as a way to quantify human attitudes towards nature across hundreds of languages, thousands of species, and hundreds of millions of internet users. Our results have received widespread news coverage.

At the outset of this research, I met with Martin to discuss the methods for using Wikipedia data and he was instrumental in providing insights in how to extract and interpret information from Wikipedia. He was the only person at Oxford University that I was able to find with specific expertise in Wikipedia and thus provided unique assistance in moving my research forward. I am currently continuing to collaborate with him on an upcoming publication describing methods for using Wikipedia data for research in the conservation context.

John C. Mittermeier
School of Geography and the Environment
Suggested future work

Collection Explorer itself was built with a restriction to only show items from Oxford GLAM collections. It would be more useful to research and education, and more used, if this restriction were lifted. Taking the artist Matsumura Goshun, Wikidata knows of works in the National Gallery of Australia, Cleveland Museum of Art, Yale University Art Gallery, Detroit Institute of Arts and Art Gallery of New South Wales in addition to the Ashmolean. Somebody looking for information on this artist is better served by a profile that is not artificially restricted to one institution's collections. The same considerations apply when we consider entities depicted in art: though there may be people interested exclusively in depictions of Parvati or of Jahangir that are housed in Oxford, a wider audience will be interested in artworks without that restriction.

Among the images shared in 2015 were 1,200 from the Curzon Collection of Political Prints. Community edits have linked the images to Wikidata entries for depicted individuals. This semi-structured data set could be harvested into Wikidata, joining up with political cartoons from the same (Napoleonic) era held by the Library of Congress and University of Washington which are also represented on Commons. Querying this dataset could bring up all the caricatures of a given figure by a given artist, or show relations of influence among cartoons.

Michael Sullivan's books about Chinese artists were very useful for creating profiles for artists unknown to Wikidata and other online biographical sources. It would help the linked open data representation of the Ashmolean’s collections, and of Chinese art generally, if key facts from those biographies could be released as open data.

Other art collections, especially of figurative art, such as the Ashmolean’s ‘The Elements of Drawing’ could be described in Wikidata using the same process that was used for the Jameel Centre data set. This would make the collection visible in other tools and enable more detail such as crowdsourced tagging.

As the Ashmolean online catalogue is unified, the Jameel Centre links that have been added to Wikidata will be superseded. The Quickstatements tool could be used to replace these links in bulk, marking the old links as deprecated.

I investigated using Wikidata to scan for dead links in the many language versions of Wikipedia. With some help from Wikidata’s developers, I found that it could not do this directly, so I made a custom script. This uses Wikidata to get a list of items in a given collection and find Wikipedia articles about those items in any language. The script then scans each article for links to Oxford domains. When an article has an outdated domain (such as image.ox.ac.uk or ashmus.ox.ac.uk), the script gets the correct link(s) from the Wikidata record. This has been used to fix dozens of out-of-date links and could be used in an on-going process.
Summary and conclusions

Linked open data offers the opportunity to connect together hundreds of different resources about each person or concept, and to express relations between art works, people, places and objects. Representing cultural heritage in this way is a desirable goal for institutions and the public. There are various ways to reach this goal, whether by institutional, national or thematic platforms. This project has shared data about multiple Oxford collections on a multi-purpose, web-scale platform, integrating it with other kinds of data and using this to provide novel interfaces including contextual information about people, concepts and places associated with an item.

Building a linked open data platform for cultural heritage involves an enormous effort. With Wikidata, we benefit from the work already done by others: more than a billion edits by tens of thousands of regular contributors who use hundreds of languages. We do not need to describe Ferdowsi, Isfahan, the family of Timur, or Sir Gore Ouseley because Wikidata already represents those things and millions more.

Wikidata gives cheap and immediate ways to make collections visible as linked open data. The project also suggests two ways to prospectively reduce costs:

1. Reduce barriers to the use of existing catalogue data by making them explicitly open with a CC-Zero licence.
2. Build or commission cataloguing interfaces that allow for identifiers to be retrieved from an API. The identifiers and auto-suggestion offered by Wikidata can easily be built into any suitably customisable system. When working with Chinese, Japanese, or Arabic names, storing names as strings, rather than using identifiers, can create technical debt in the form of an ambiguity that is expensive to resolve.

Sharing on Wikidata benefits the data by introducing authority file identifiers, by making connections across different kinds of data and across different collections around the world. It benefits the source catalogue by giving new kinds of path through which people can find and interact with the collection, and by giving contextual information such as artist biographies. Astrolabes can be set out on a map or timeline, and species can be browsed by a taxonomic tree. As a paper in the MuseWeb19 conference⁵ argues, the sustainable way for museums to engage with Wikimedia will involve sharing selected catalogue data on Wikidata, harvesting back enriched data and using this to create innovative public interfaces. One of the examples given in the paper is the Collection Explorer built by this project.

The data sets will continue to be available to the public for the long term. We found that public edits improve these data over time. There was no vandalism or misinformation added. The ability of the public to edit the data should not be seen primarily as a risk, but as a multilingual volunteer workforce who will improve the usability of the data set, if given the freedom to do so.

Developers of Wikidata-driven sites, apps and search engines should likewise be regarded as potential partners whose work will enhance public access to our collections. The risks are on the upside: the worst thing that could happen to an open data set is that it is not used, but opening it up creates the possibility of new public interfaces, including interfaces yet to be thought of.

This document is licensed under a Creative Commons Attribution-ShareAlike (CC-BY-SA) 4.0 licence.