The Future is Wiki: National Libraries’ Wikidata Projects and the Culture of Connection

INM304

Leah Perry

3346 words
Introduction

Libraries and information services have long since been champions of equitable access to information. As the popularity of open access grows, the GLAM (Galleries, Libraries, Archives and Museums) sector investigates how to further develop their service and enrich their collections through knowledge collaboration. With the increased production of information, diverse range of publishable formats, and the world’s perpetual digitisation, information services aim to utilise the varied information available through semantic linking, centralising institutions’ disparate, but complementary, information. This is being implemented in national libraries across the world with the adoption of Wikidata. This online database can act as an integrated information cache, connecting institutions and creating relationships between their collections. This essay will begin by examining what Wikidata is and what benefits it can reap for information services. It will detail specific applications of Wikidata including authority control, enriching metadata, interface accessibility and user experience, through projects at Library of Congress, The National Library of Wales, and The National Library of The Netherlands. There will be overlap between concepts and practices, as they share the same focus and outcome of sharing previously siloed information to an open-access, collaborative, centralised space to enrich collections and foster ease of interoperability between institutions. The essay will conclude with reviewing how Wikidata and a subsequent culture of connection has and will continue to shape the future of the library.

Wikimedia

The Wikimedia Foundation is a non-profit organisation which aims to diminish barriers to the world’s knowledge. It hosts free, open-access and collaborative information platforms under a creative commons license, meaning all information, data, and media on their sites is freely
available for use without permission, even for commercial purposes. At the heart of their cause is the idea that, “Knowledge belongs to all of us.” (The Wikimedia Foundation, 2021). There is a clear accord between The Wikimedia Foundation and the library as an institution, as IFLA (2019) asserts, “the belief that people, communities and organisations need universal and equitable access to information, ideas and works of imagination for their social, educational, cultural, democratic and economic well-being,” and the quality of this information provision should not be compromised by its accessibility or scope. Alongside Wikidata, The Wikimedia Foundation hosts other projects such as Wikipedia, Wikimedia Commons, Wikisource, and many more. In the last five years, GLAM organisations have increasingly recognised the importance of Wikimedia projects for librarians and users of an information service alike. An understanding of Wikimedia is becoming essential for an institution’s online presence (Leva and Chemello, 2018), as the foundation’s projects grow in size, popularity, and effectiveness. The rise of the Wikimedian in Residence role demonstrates the value that Wikimedia has for collections, and its continuing success is increasingly calling for librarians and information professionals to be trained in Wikimedia projects: creating, editing, coding, enriching, and innovating use of records, alongside cultivating a deep awareness for how Wikimedia contributes to the ecology of contemporary information and information provision (Bartholemei et al, 2016). Wikidata classes for librarians prove invaluable for improving cataloguing processes (Monk, 2021), and libraries hosting Wiki projects such as hackathons, edit-a-thons, and workshops testify to the advantageous products of collaborative information work.

Unlike its sister sites Wikipedia, Wikimedia Commons and Wikisource, Wikidata supports structured data, making its content both human and machine-readable, acting as a central repository for external sources (Vrandečić, 2013). Intelligent devices such as Alexa and Siri often mine Wikidata to answer queries, Google uses it to enhance its search results, and YouTube utilises it to provide further information about video topics. Wikidata is multilingual, interoperable, and open, and so it fosters not only connections between institutions, but community collaboration (Allison-Cassin and Scott, 2018). It is the core of other Wikimedia projects (Kanke, 2020) and provides live updates (Abián, Guerra, Martínez-Romanos and Trillo-Lado, 2018), meaning data need only be edited in one place. This happens outside of the Wikimedia bubble, too, as Wikidata informs and is informed by a diverse population of external datasets, from local metadata and specialised repositories to international authority files and open-access databases (Erxleben et al, 2014).

Wikidata can be queried by SPARQL, an RDF query language, structurally similar to
Wikidata, both sharing efforts to contribute to the Semantic Web by cultivating an interconnected network of information (Malyshev et al, 2018). To do this, Wikidata creates statements that work in triples: item, property, value. Items and values are labelled by human-readable text and a machine-readable unique identifier with the format Qx, where x is any positive number. Properties, with the format Px, denote the link between items and values. For example¹, Zadie Smith (Q140052) received the award (P166) James Tait Black Memorial Prize (Q392586). Items, properties, and values all have their own independent webpages with a URI, allowing the data to be linked in statements. There are more detailed concepts that offer additional information to or contextualise the item, property, value relationship, such as qualifiers and references. It is important to note that Wikidata does not claim to or aspire to provide truthful information, but rather, operates as a storage space that gathers information provided elsewhere, thus references and supporting sources are crucial. An item’s webpage may store contradictory information, for example, Zadie Smith’s page has two records for her date of birth: 25th October 1975 and 27th October 1975. Of course, Smith does not have two birthdays, but there are instances of both dates being used by external sources, references of which are provided. Wikidata is not concerned with truthfulness, but verifiability (Poulter and Sheppard, 2020). This is parallel to libraries, as different institutions may label the same document with conflicting metadata, which is not to say that either are exclusively correct. This necessary variety can be remedied by Wikidata.

Ultimately, Wikidata can diminish barriers to enriched collections, redistributing power to local institutions and individuals, but first, institutions must input their metadata and adopt Wikidata into their working strategy. Wikidata can be a “hub of hubs” (Stinson, Fauconnier and Wyatt, 2018, p. 26), creating fluid connections between separate organisations and their works. The more data that is stored in Wikidata, the more effective it becomes, and national libraries have been paving the way for the information profession, integrating the tool into their services using original and exciting methods (Byrne and Wyatt, 2021). Tharani (2021, p. 4) examined the increasing body of literature on Wikidata projects to find its top applications in information services, identifying the most popular uses to be: “integration and interoperability of authority data”, “improve metadata quality and processes”, and “enriching local catalogue user interface and experience.” This essay will now explore how three national libraries are using Wikidata to fulfil these applications.

Authority Control and Library of Congress

GLAM institutions retain their own authority files for consistency in cataloguing and retrieval. These can differ between institutions, as language – proper nouns and otherwise – are fuzzy, and can be expressed in multiple ways. This means that two libraries holding the same document may not use the same format for an authority name, fluctuating retrieval results and leading users to wrongly surmise that libraries do not have the desired document in their holdings. A universal standardisation for authority files is merely an idealist solution, as individualised collections and varying institutional approaches necessitate different metadata and cataloguing methods (Candela, 2019). This is especially apparent regarding grammatical structures within and across languages, and the differing descriptive approaches called for by textual documents compared to that called for by historical artefacts. Libraries are inherently concerned with interoperability, and linked data offers a solution that allows institutions to be connected while operating conflicting authority data. Library of Congress (LC) houses linked data authority files, collating various institutions’ authority data, and in 2019 began linking to Wikidata. Wikidata and LC collaborate and share information, becoming centralised spaces in which to navigate a wealth of resources from many institutions. Later, this essay will explore the proposal from van Veen that en masse, institutions could adopt the Wikidata identifier, therefore not requiring linked authority data, and creating stronger connections between collections. Wikidata provides more detailed bibliographic information, and so LC directs the user from this information towards the relevant resources, whether this is the authority file or a document on an institution’s website. This can be implemented for any digital library or repository, including any GLAM institution but also government services (Miller, 2019A). Linked data is a mutually rewarding practice, benefiting all parties.

Of course, to be most effective, as many institutions as possible must input as much of their data into Wikidata as possible. Library of Congress call numbers (LCCN) have been added to Wikipedia entries for a long time but the job is a large one, and it is ever-ongoing. In the last five years, LC has heartily adopted this practice towards Wikidata, inputting LCCNs – mostly Library of Congress Name Authority Files (LCNAF) but also Library of Congress Subject Headings (LCSH) – to improve institutional interconnectivity, enable further access to the LC collection, and encourage other institutions to follow suit (Miller, 2019B). 11% of
LC authority terms are linked to Wikidata, and on average, every two days 220 LCNAF links are added (Miller, 2020). LC cataloguing standards do not permit a large number of multiple labels, most of them defined by just one, but most Wikidata records have many labels provided by different institutions, and this aggregation of name variations offers a route to enriching one’s own collection metadata. Another benefit LC is reaping by inputting their data into Wikidata is furthering an awareness of the institution and promoting their collection. Local, small, or specialised libraries may have more granular, or accurate, authority data than national libraries or collections with wide breadth and scope, and providing this information in Wikidata allows other institutions to both refine and expand their own authority data (Myntti, Lewis, McCormack and Rockwell, 2020).

LC Labs’ Prints and Photographs (PnP) tool is the realised product of linking authority files to Wikidata (LC Labs, 2019). Users are able to navigate the holdings using Wikidata properties as filters. This can be as basic as selecting gender and/or country, or, by querying aggregated authority files on Wikidata, it can be as advanced as filtering for works covering female American poets born in the 1890s who have received the John Simon Guggenheim Memorial Foundation Fellowship. This shows that linking authority data to Wikidata allows the collection to be explored in new and granular ways, and with its open-access collaborative values, it is more time, cost, and labour efficient than other avenues (Cooey, 2019).

### Metadata and The National Library of Wales

Similar to Library of Congress’ direction with authority files, other bibliographic information supported by Wikidata allows for metadata enhancement and development of services. Properties provide great profile details through statements – for example, Zadie Smith’s page (Q140052) includes the properties: social media followers, official website, audio recording of subject’s spoken voice, influenced by, genre, notable work, blood type, to name but a few – which are contained as structured data to be easily accessed and queried by humans and machines alike. The National Library of Wales (NLW) has undertaken many creative Wikidata projects, capitalising on the wealth of structured metadata available to query their collection in new and innovative ways, most effectively with visualisation tools and unique
approaches to exploring their visual art holdings. The classic objective of libraries – to provide equitable access to information – has shifted, through the rise of the digital library and indeed the inexorable digitisation of our everyday lives, to providing a richer information experience through knowledge collaboration. It is now important for libraries to partake in creating and facilitating, “new artefacts as ontologies, semantic content descriptions, data links and new forms of collaboration using social networks, specialized communities, wikis, collaborative games and mashups.” (Hallo, Luján-Mora, Maté, A. and Trujillo, 2016, p. 117) NLW have used Wikidata to produce comprehensive and interconnected information for their records using visualisation tools such as maps, graphs, charts, timelines and interactive querying of such visualisations (Evans, 2021). Collections can be examined through linked data as Evans, Wikimedian in Residence at NLW, explains with the Peniarth Manuscripts collection, “we can identify all manuscripts which contain correspondence, and then see who the main subject of those correspondence are, and because Wikidata is linked data we could then access biographical data about those people.” (Evans, 2019A)

The key to this success is Wikidata’s structured data, which allows the collection to be queried in ways not easily facilitated by traditional metadata text strings (IFLA, 2020). Patterns, visualisations, groupings, and connections can be discerned by traditional metadata methods, but not to the same exhaustive extent or fluency as Wikidata. Having a centralised space for metadata that is free, collaborative, and publicly accessible, decreases the expense and effort required of libraries to provide high quality services. The more institutions that input their metadata into Wikidata, the more effective and widespread the results will be, and the more relationships and deeper connections can be built between the world’s knowledge. Furthermore, Wikidata’s consolidation of open-collaborative efforts and reputable source datasets “guarantees a good balance between the authority and the quality” (Europeana, 2017), and battles institutional silos and language barriers.

Dwynwen is a modified version of Crotos, the open art repository, created specifically for NLW’s collection. A system informed by the structured data on Wikidata, users are able to search and query Dwynwen with Wikidata properties, and documents are linked to their own Wikidata pages for further bibliographic information and identifiers in other institutions’ holdings. Library catalogues are efficient at searching for specific monographs using keywords, titles, authors, etc. but cultural heritage collections often consist of a diverse range of documents, formats, and items, which necessitate different methods of searching, and Dwynwen provides a process where users can “browse or explore [the] substantial archive of digital content in a visual and engaging way.” (Evans, 2019B) It has been recognised that
traditional keyword search is an ineffective way to navigate a cultural heritage collection, where undirected browsing and creative exploration should be favoured (Damiano, 2019), essentially translating the experience of the physical space, which encourages fluid inquiry, into the digital sphere (Madden, Webber, Ford and Crowder, 2017). Dwynwen observes the usefulness of narrative searching by providing capacity to browse using Wikidata properties and linking to further information.

**Catalogue Interface, User Experience, and The National Library of The Netherlands**

GLAM organisations hold an abundance of information in their bibliographic records. While promotion of individual institutions is important, cultural heritage and libraries fundamentally support the sharing of knowledge. As explored earlier, having an open centralised hub for variant authority files and other metadata for institutions to link to improves the quality of collections and provides opportunity to view information in a new light, which in turn prompts further innovation and discovery. In this way, digitisation has changed the library, as it is now not only expected to be a repository of information in its own right, but to be able to connect to other sources of information (Warraich, 2020). In order to provide the best quality service, library catalogues should facilitate inter-institutional search and connection to other external sources through user-friendly features (Yadagiri and Ramesh, 2013). Wikidata is only useful if its content is growing and being widely implemented, equally, it is only useful if the users of a library service where it is implemented can understand and appropriately use the features.

As a member of The National Library of The Netherlands’ (NLN) research team, van Veen is a keen advocate for fully integrating GLAM holdings and Wikidata, proposing that instead of collating disparate identifiers in Wikidata, it would be more effective if institutions globally adopted the Wikidata identifier (van Veen, 2019). As the infosphere becomes more advanced and complex skills are increasingly required of users, and metadata in Wikidata swells, one authoritative machine-and-human readable, language-independent identifier could simplify and improve discovery and retrieval. Van Veen (2017) argues that, “minimizing [the] number of variations for identical queries in different databases and minimizing the
required knowledge will lower the barrier to connect to external sources,” blurring the lines between collections across the world, creating a global shared and collaborative knowledge base.

NLN runs many Wikidata projects (‘GLAM/Koninklijke Bibliotheek Nederland’, 2021), including Album Amicorum of Jacob Heyblocq, a digitisation project of a ‘friends book’ from the mid-17th century. This item lends itself particularly well to the linked nature of Wikidata, as it is a diary of sorts, containing entries from many well-known figures Heyblocq met throughout his lifetime, offering many access points to the collection and outwards. The project is designed to examine how “visible, searchable, discoverable, sharable, reusable and linkable” the work is before being shared to the Wikimedia environment and after, with the hypothesis that it will exhibit new methods of interacting with and building upon the collection (‘WikiProject Alba amicorum National Library of the Netherlands/Jacob Heyblocq’, 2021). Access points are vital for hosting inter-institutional browsing and discovery.

Wikidata tools provided by libraries should accommodate ease of searching on their catalogue for their users, as SPARQL queries cannot be expected to be understood by all patrons. This is exemplified in another NLN project, Newspapers+, combining the efforts of named entity recognition and automatic linking to external sources in digitising Dutch newspapers. In this interface, the user is provided by a simple search box with separate terms divided into brackets. The system conducts a “best guess”, translating the user’s search into a SPARQL query to approach Wikidata (van Veen, 2019). Another important means for enhancing the user experience is for institutions to not just link Wikidata pages to their collection, but link their collection to Wikidata in order to support a cyclical interlinked information ecosystem, seamlessly transitioning between the shared knowledge.

Conclusion

To conclude, Wikidata’s informational interconnectivity and language-independent structured data can be utilised to enhance library and information services internationally. Institutions can input their authority data into Wikidata to connect collections and related works that otherwise may have been detached due to varying authority terms. Storing this authority
information in an accessible open platform such as Wikidata allows institutions to enrich their own collection, and individuals to contribute to the world’s knowledge. This is especially applicable for authority files in smaller, local services compared to national libraries with broad scope and collection volume, as local institutions may use more granular language. Equally, Wikidata entities including authority data in different languages reduces information borders between countries, allowing users to explore international collections, and which institutions can use to accommodate for various languages on their catalogue.

Applying other metadata hosted by Wikidata brings institutions and patrons to create original visualisations and deduce relationships between documents, in their own collection and others. This is particularly helpful for cultural heritage services that hold diverse formats, as traditional keyword search is not the most effective technique for browsing. Innovation prompts new methods to interact with the collection, breeding further discoveries and richer information creation. Wikidata acts as a centralised knowledge base for institutions to connect metadata to one another, and indeed to the greater information environment.

While these novelty techniques provide new insight and benefits for libraries and their users, it is important that to sustain and encourage this, institutions should make their services accessible and simple for their users. This can include ensuring the collection is linking both inward and outward to Wikidata and making this clear and comprehensible on their catalogue, thus effectively contributing to the linked information cycle and facilitating smooth and consistent navigation. It can also include implementing tools that perform SPARQL queries for the user, as complex computer language skills cannot be an obstacle to innovative information discovery. Perhaps global adoption of the Wikidata identifier would prove beneficial for users to navigate between institutions; this may be the next step for Wikidata projects.

Wikidata provides many exciting opportunities for libraries and information services to enrich their collection and contribute to the world’s shared knowledge. With time will come wider and deeper connections as more libraries and information services input their data onto Wikidata, and incorporate it into their own collection. The future of the library is connected, and Wikidata is the coordinator.
References


DOI: 10.1145/3314183.3323870.


