Library data as linked open data

Jason Evans, National Wikimedian at the National Library of Wales

Libraries have changed. Around the turn of the last century new public libraries were being built in towns and cities all over Britain. The Scottish-American businessman and philanthropist Andrew Carnegie funded more than 2,500 new libraries around the world, and here in Wales the National Library was established in 1907. For many years the library catalogue was no more than ink on paper: card indexes in deep wooden drawers, painstakingly compiled by humans. Both the appetite for building libraries and the way in which they operate have changed somewhat in the intervening century but the core purpose of the library is very much the same – to give access to knowledge, literature and learning.

At the National Library of Wales we still use some of our old card indexes. Archive collections tend to be indexed by category and subcategory and many have been annotated by catalogers, noting related documents. In its simplest form, this is linked data – the ability to make connections across collections based on a number of different common factors such as author, subject or place.

Figure 1: An old card index at the National Library of Wales

So, if our aim is to give the best access possible to our users, both our MARC21 data for printed material and our Dublin Core archive and manuscript data have certain shortcomings. This is both down to the way the data is structured and the way we give access to it. At the National Library of Wales we have been exploring the benefits and challenges of converting to linked open data. To date this has been done by experimenting on a reasonably small scale using the established infrastructure of Wikidata.

Wikidata is part of the Wikimedia Foundation's family of websites – you may have heard of its flagship site Wikipedia. Much like Wikipedia, Wikidata is open, editable, and reusable by anyone. It is part of the Wikimedia Foundation's vision of a world in which everyone has access to the sum of all human knowledge, and, as the name suggests, Wikidata's contribution is in the form of data. Originally created about 6 years ago to help connect related Wikipedia articles in different languages, Wikidata has flourished into a massive open data set with data about just about everything, from the human genome to people, places, events, literature, and all sorts of other 'things'. At the time of writing, Wikidata describes nearly 85 million 'things'. What makes Wikidata so powerful, other than the fact that institutions and individuals can contribute or make use of the data freely, is that it is linked data, rather than text based data – it's 'things' not 'strings'. Behind the interface, Wikidata looks very much like RDF.[1] Each item has a unique identifier (a Q number) along with a label and description which can be added in over 300 languages, making this a multilingual dataset. Items are then described using an unlimited number of statements, also known as triples. The dataset already contains over a billion such statements.

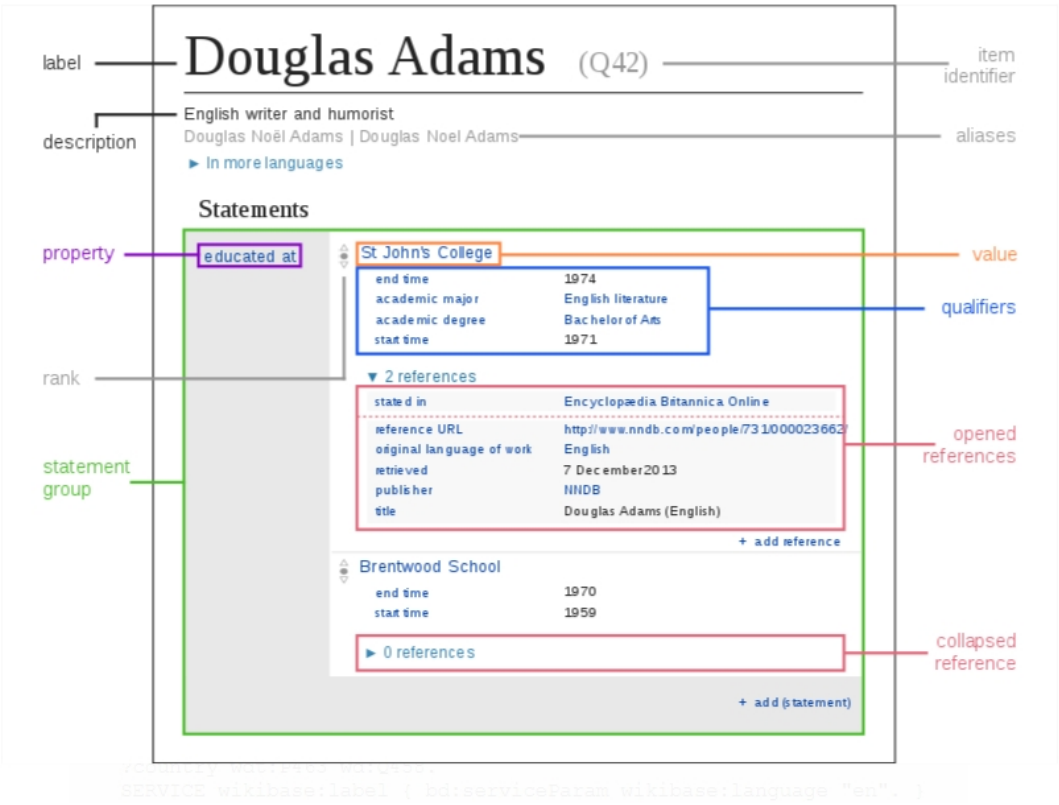

Figure 2: Annotated example of a Wikidata Item

[1] https://www.wikidata.org/wiki/Wikidata:Relation_between_properties_in_RDF_and_in_Wikidata

By converting catalogue data to linked open data on Wikidata we are essentially taking the metadata 'fields' and matching them to Wikidata 'properties'; in other words we are the converting text 'strings' that inhabit those metadata fields and matching them to Wikidata items (things). For example, if the value of a metadata field for place of publication is 'Aberystwyth' we will match this to the Wikidata item about Aberystwyth (Q213154). This means that our data no longer simply states 'Aberystwyth' as the place of publication in a human readable way, but creates a connection, in a machine-readable way, to all the data Wikidata holds about Aberystwyth, such as the type of town, its date of inception, its population, and its coordinate location. Wikidata also acts as a hub for external identifiers, and the item for Aberystwyth includes information such as the WorldCat ID, VIAF ID, and the Library of Congress authority ID for the town.[2]

As you might imagine, transforming standard metadata effectively into this format has its challenges. Mapping metadata field types to the relevant Wikidata properties is pretty straight-forward, but matching text strings to the correct Wikidata items takes more time. Wikidata has 13 items called 'Aberystwyth'. Most are paintings, but there is also an item for a hymn and one for a 19th century merchant ship. There are tools which allow you to filter out these irrelevant items, such as OpenRefine, but often an element of manual checking is required. This is particularly true when dealing with creators of works – untangling 50 different John Joneses and ensuring each has their own data item with at least some biographical data is no small task, trust me. Often this problem is compounded by a lack of consistency in the way data is presented in MARC fields.

Despite the challenges we have seen some interesting results. We took data for the Peniarth Collection – some 540 medieval manuscripts – and converted it to linked data. Some data fields mapped easily to the new format, but much of the useful descriptive data for the items formed a text description in a 'scope and context' field. This was teased apart using patterns in the text to identify scribes, authors, and works contained in the manuscripts. We found that many of the authors and scribes were already in Wikidata thanks in part to our previous work to share data for the Dictionary of Welsh Biography. Once all this data was matched up to Wikidata, we found that we had access to a much richer dataset than before. Using Wikidata's query service you can query every element of the data. So, for example, we could easily isolate works by an individual scribe, or discover which scribes worked on the same manuscripts. We could generate lists of manuscripts by genre and main subject, as well as access newly connected information such as biographical data about authors. So, for example, we could return a list of items in the collection authored by women born in the 15th century, or manuscripts that included works about the Roman Empire written in Welsh. And because Wikidata allows institutions to add their own identifiers to items on Wikidata we discovered that other institutions held different versions of some of the manuscripts in our collection, or other manuscripts with the same authors or scribe. The opportunity for research is massively increased both by the nature of the structured data and by the ease of access to the data through Wikidata's query service and suite of visualisation tools.[3]

---

[2] https://www.wikidata.org/wiki/Q213154

[3] https://blog.library.wales/treasured-manuscript-collection-gets-the-wikidata-treatment/

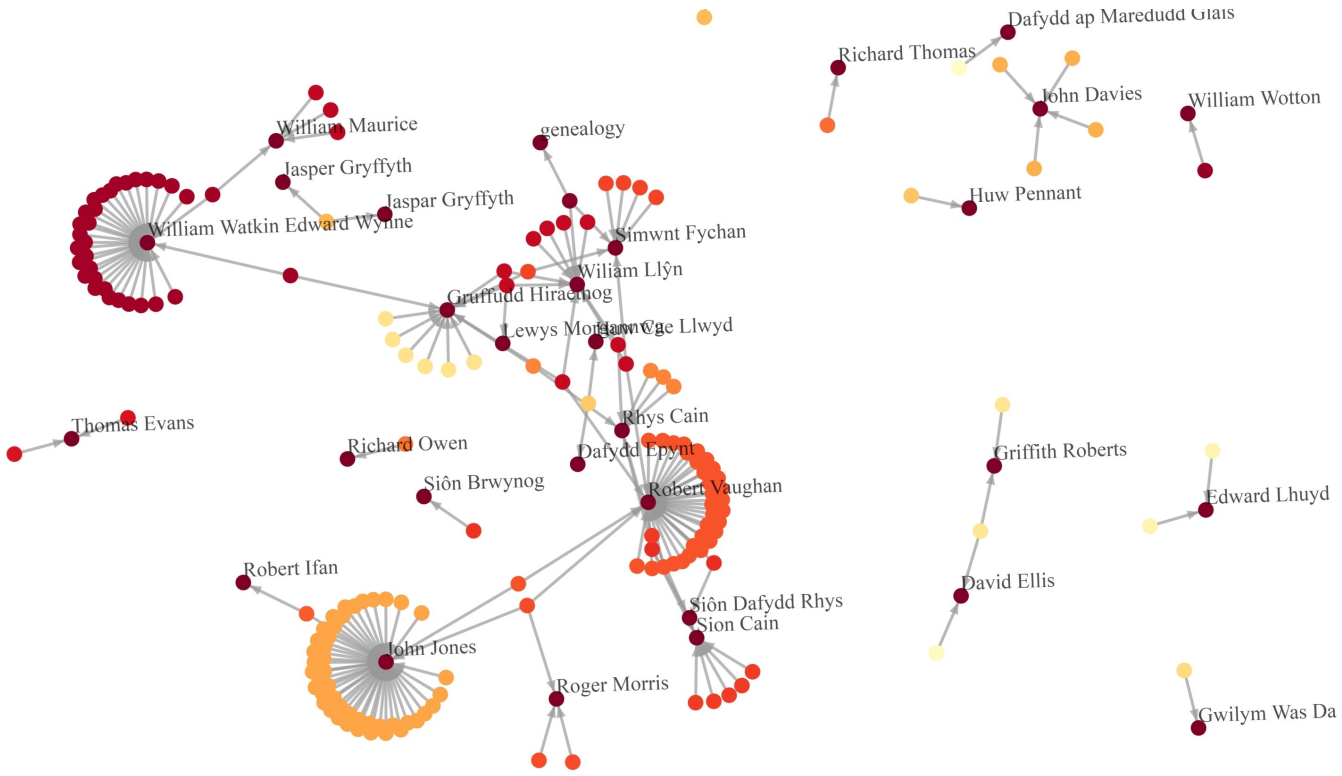

Figure 3: Scribes of Peniarth Manuscripts and the works with which they are associated

As well as sharing data through Wikimedia platforms we have also contributed to Wikipedia and Wikimedia Commons, where we have shared around 20,000 openly licenced images of photographs, artworks and other digitised content. Wikidata can then be used to describe this content using linked data. Again, the process of transforming our data led to an enrichment of our data record. Matching tags of places depicted in images to place items on Wikidata meant we gained access to coordinate data, and so we could explore the collection on a map for the first time. We also gained access to a wealth of information about the artists, the types of things depicted in images, and other institutions who also held either works depicting those things, or works by the same artists.

We have also seen other positive outcomes from making our data open. Wikidata applies a CC0 licence to all its data, and most of the images we have shared to Wikimedia Commons are in the public domain. This means that our content can be freely reused by anyone. We have seen our content used in thousands of Wikipedia articles, garnering millions of views every month. Authors, historians, curators, and the media have used our free open content to enrich their own resources, but increasingly we are seeing that by 'round tripping' our improved, structured open data we can develop and improve our own online services.

We worked with volunteers to develop a prototype for a linked data portal for exploring our open digital images, by adapting existing open source software. The result is the 'Dwynwen' website which allows users to search over 17,000 images in ways that are simply not possible with our standard catalogue data. You can even view content on a map based on places depicted in the images.

We will also shortly be releasing an interactive timeline for the Dictionary of Welsh Biography which is powered by Wikidata, open images from Wikimedia Commons, and text from Wikipedia.

As Wikidata has grown it has increasingly become a hub for describing publications. Scholarly articles in particular now make up 31% of Wikidata.[4] One of the driving forces behind the use of Wikidata in this way is to improve accuracy of content on Wikipedia. Creating a central, open, structured dataset of scholarly content makes it easier for Wikipedia editors to locate and cite reliable content for Wikipedia. Discussions as part of the WikiCite initiative suggest that in the long term all citations on Wikipedia could be standardized across all languages using linked data. It would mean Wikipedia users could access all those external identifiers used on Wikidata to see who holds copies of a book or paper, where it was available online, and how it was licenced for reuse. With this in mind the National Library of Wales recently shared over 30,000 catalogue entries from the Welsh bibliography on Wikidata.

This task presented a number of challenges. Because each piece of data has to link to another Wikidata item, text strings for publishers and authors either had to be matched to existing data items or the data had to be added. We found that one publisher might appear in dozens of different formats in our metadata. Some records gave publisher names in Welsh, others in English. This meant a lot of checks and processing using OpenRefine and Excel in order to clean the data. Hundreds of Welsh publishers and authors were added to Wikidata, and whilst finding, disambiguating, and locating information about all the publishers was a manageable task, the same cannot be said for authors. 13,000 works are now linked to items for over 2000 unique authors. However, we were not able to identify and match the remaining authors in the time we had.

Many of the authors were identified by matching our data with other datasets. For example, 4536 items were matched to OCLC bibliographic records,[5] 11386 have universal ISBN numbers,[6] and 6941 were matched to Open Library records.[7] This approach of connecting with different datasets is key to cleaning up and converting to linked data. Restructuring our data in this way at scale will depend on close collaboration across the sector.

Although the transition to linked data in libraries is still a fairly new idea, there are some notable examples of the use of not only linked data, but also open and collaborative data, being used to enrich user experiences. The Biblioteca Nacional de España (BNE) in Spain has created linked data for its entire catalogue, allowing them to create a powerful search engine which incorporates open text from Wikipedia.[8] At the National Library of the Netherlands, the OCR text of their online newspaper archive has been enriched with Wikidata entities, improving search and linking users to a wealth of additional information about people, places and events identified within the text. The National Library of Sweden has also begun sharing their national bibliography with Wikidata.

There seems to be a growing acceptance that linked data will form part of the next chapter in the ever changing landscape of library and information services. It offers more powerful search and discovery tools, more consistent and complete data, and better access for users. It will also allow us to begin connecting and aligning data between institutions, creating one huge global dataset. The open and collaborative elements of a system like Wikidata mean that we can work together with our partners and with our communities to develop, enrich, and research our data in new ways. This collaborative approach also helps mitigate some of the obvious and inevitable challenges the adoption of linked data presents, and allows us, as a sector, to pool resources in order to improve data, and ultimately to continue to serve as essential providers of knowledge.

---

[4] https://www.wikidata.org/wiki/Wikidata:Statistics
[5] Wikidata query for NLW works with OCLC ids
[6] Wikidata query for NLW works with ISBN numbers
[7] Wikidata query for NLW works with Open Library ids
[8] The BNE 'Datos' linked data portal
[9] van Veen T., Lonij J., Faber W.J. (2016) Linking Named Entities in Dutch Historical Newspapers. In: Garoufallou E., Subirats Coll I., Stellato A., Greenberg J. (eds) Metadata and Semantics Research. MTSR 2016. Communications in Computer and Information Science, vol 672. Springer, Cham