From Authority Data, to Linked Open Data and Wikidata: The Case Study of a Hebrew Manuscript Catalogue

Gila Prebor[1] [0000-0001-6458-0831]
[1] Department of Information Science, Bar-Ilan University, Ramat-Gan, Israel 520009
Gila.Prebor@biu.ac.il

Abstract. Traditionally, library catalogues have served as a tool to manage library collections and as a bibliographic tool for information retrieval. Eventually this caused library catalogues to be data silos. In order to break down these metadata silos, the information must be accessible and free to use. The semantic web, and in particular, linked open data, are initiatives that can turn library catalogs into a real part of the Internet. Today libraries are an important player in the linked data arena. Converting catalogues to large linked data enables large-scale analysis of cultural heritage Big Data. By implementing linked data initiatives open library data is available for reuse in the information space. Libraries can share their open metadata with non-library communities. Wikidata is a collaboratively edited knowledge base hosted by the Wikimedia Foundation. It is one central database of human knowledge which contains structured and linked data. If more collections will be added to this huge linked open network it will enable researchers to investigate and find new discoveries thereby revitalizing our cultural heritage data, which has been persevered in closed silos for hun-dreds of years. The potential in Wikidata for the information world in general and for libraries in particular is illustrated by research done on the National Library of Israel Hebrew manuscripts catalogue.

Keywords: Linked Open Data, Authority Data, Wikidata, Manuscript Catalogue.

1 Introduction

Traditionally, library catalogues have served as a tool to manage library collections and as a bibliographic tool for information retrieval. In the 20th and 21st centuries libraries focus not only on bibliographic record but also on data. The content of the catalog record has been standardized according to international rules and standard protocols such as AACR, MARC, Z39.50 and RDA so it could be easily exchanged and duplicated. This enables the catalogues to be accessed from a distance both by human users and by machines. Library standards were intended to be used by librarians and the catalogues serve only the library community. Eventually this caused library catalogues to be data silos. In order to break down these metadata silos, the information must be accessible and free to use [1].

2 The Semantic Web, Libraries and Linked Data

The semantic web, and in particular, linked data, are initiatives that can turn library catalogs into a real part of the Internet. Today libraries are an important player in the linked data arena. Converting catalogues to large linked data enables large-scale analysis of cultural heritage Big Data. By implementing linked data initiatives open library data is available for reuse in the information space. Libraries can share their open metadata with non- library communities. Tim Berners-Lee, the inventor of the web, has proposed a five-star scheme for linked open data and it begins with open licensing (http://5stardata.info/en/). This orientation to openness and giving access aligns well with the ethos of libraries and archives [1, 2]. RDA - Resource Description and Access was published in the Metadata Open registry (http://metadataregistry.org/) as an element set in the RDF standard model.

Several libraries have already taken the initiative to convert their catalogs to RDF-based triples and to linked data [3]. For example, the Swedish Union Catalog, LIBRIS (libris.kb.se), was one of the first catalogues that began sharing linked data in 2008. Other libraries using linked data are the British National Bibliography (http://www.bl.uk/bibliographic/datafree.html) and the Library of Congress (https://id.loc.gov/). The Getty vocabularies are now available as Linked Open Data (http://www.getty.edu/research/tools/vocabularies/lod/index.html) [4].

3 Authority Control

One of the most important tasks in cataloging and building metadata records is authority control. This work has been done in libraries for decades in the era of catalog cards and a long time before the invention of the computer. Authority control has three major concerns: Consistency – to ensure consistency in the different forms used to represent entities; Relationship – showing the relationships among entities; Uniqueness - maintaining the uniquely of the entities. The process of authority work involves more than determining the authorized name or title. With the recent move to FRAD – Functional Requirements for Authority Data and RDA – Resource Description and Access, the process also regularly includes recording significant descriptive information that assists in identifying the entity. For example: In addition to the usual attributes: dates associated with the person, title of person, other designation associated with the person, FRAD adds many attributes like: gender, place of birth/death, country, place of residence, affiliation, address and more [5]. Figure 1 shows an example of Authority record in the RDA era.

All this information will enable future complex queries such as: Who are the women authors of American origin who write in French or novels of women authors who lived in France in the twentieth century?

These Authority files require ongoing maintenance, adding years of death, changing place of residence, changing work place etc. Before the advent of digital online public

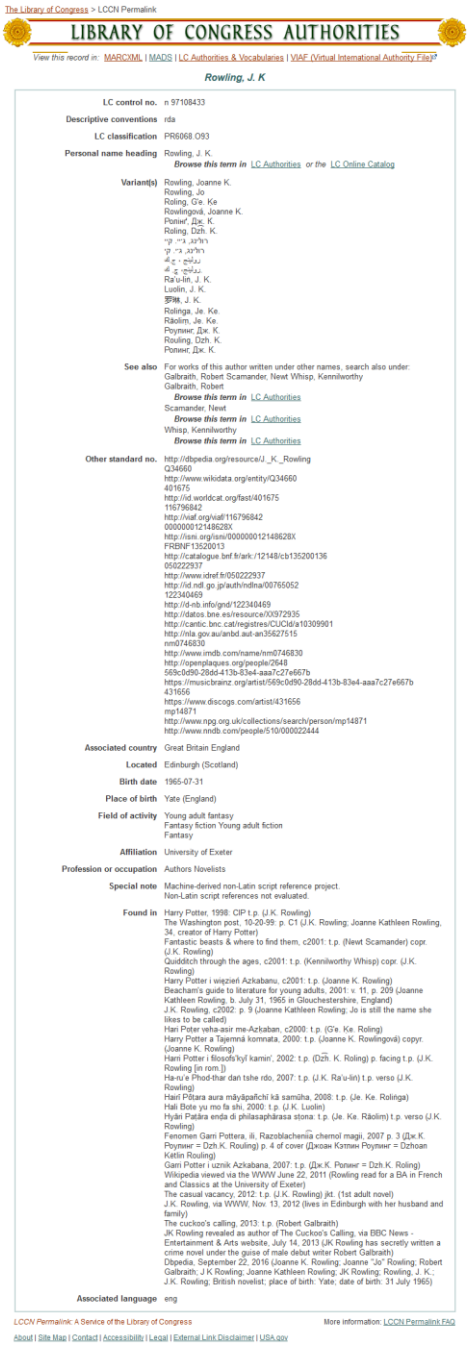
access catalogs and the Internet, the creation and maintenance of library authority files was usually carried out by individual catalog departments in each library. Naturally, then, there was a significant difference in the authority files of the various libraries, but it didn't matter. Today there are different approaches in the world to create uniformity and save work. Such an international effort is the Virtual International Authority (VIAF - http://viaf.org/) file, which is a collaborative attempt to provide a single title for a particular topic. Another example is the Integrated Authority File (GND www.dnb.de/gnd) held and used by many libraries in German-speaking countries and the Library of Congress of the United States. The idea is to create one global virtual authority file.

4 The Next Stage - WIKIDATA

Linked Open Data have the same or similar purpose in the Internet as authority files in the library world. Over the years, various databases of Linked Open Data have been created, and probably the largest database of all is WIKIDATA, which was launched in October 2012 and now includes millions of entries (https://www.wikidata.org/wiki/Wikidata:Main_Page).

Wikidata is a collaboratively edited knowledge base hosted by the Wikimedia Foundation. It is one central database of human knowledge which contains structured and linked data. Wikidata offers a lot of advantages like: anyone can edit it, new items can be added to Wikidata by every user if something is lacking, it can be read by people and machines, it is multilingual, it is on the wiki platform, Wikidata items contain other data and are linked to Wikipedia articles and it is entirely in a free license (CC).

The next stage is the integration of authority data from library catalogues to Wikidata. There are already several initiatives of this type in the world [1, 6, 7].



The Library of Congress > LCCN Permalink

LIBRARY OF CONGRESS AUTHORITIES

View this record in: MARCXML | MADS | LC Authorities & Vocabularies | VIAF (Virtual International Authority File)

Rowling, J. K

LC control no.	n 97108433
Descriptive conventions	rda
LC classification	PR6068.O93
Personal name heading	Rowling, J. K. ***Browse this term in*** LC Authorities *or the* LC Online Catalog
Variant(s)	Rowling, Joanne K. Rowling, Jo Roling, G'e. Ke Rowlingová, Joanne K. Ролинг, Дж. К. Roling, Dzh. K. רולינג, ג׳יי. קיי רולינג, ג׳. קי رولينج ، ج. ك رولينج، ج. ك Ra'u-lin, J. K. Luolin, J. K. 罗琳, J. K. Rolinga, Je. Ke. Rāolim, Je. Ke. Роулинг, Дж. К. Rouling, Dzh. K. Ролинг, Дж. К.
See also	For works of this author written under other names, search also under: Galbraith, Robert Scamander, Newt Whisp, Kennilworthy Galbraith, Robert ***Browse this term in*** LC Authorities Scamander, Newt ***Browse this term in*** LC Authorities Whisp, Kennilworthy ***Browse this term in*** LC Authorities
Other standard no.	http://dbpedia.org/resource/J._K._Rowling Q34660 http://www.wikidata.org/entity/Q34660 401675 http://id.worldcat.org/fast/401675 116796842 http://viaf.org/viaf/116796842 000000012148628X http://isni.org/isni/000000012148628X FRBNF13520013 http://catalogue.bnf.fr/ark:/12148/cb135200136 050222937 http://www.idref.fr/050222937 http://id.ndl.go.jp/auth/ndlna/00765052 122340469 http://d-nb.info/gnd/122340469 http://datos.bne.es/resource/XX972935 http://cantic.bnc.cat/registres/CUCId/a10309901 http://nla.gov.au/anbd.aut-an35627515 nm0746830 http://www.imdb.com/name/nm0746830 http://openplaques.org/people/2648 569c0d90-29dd-413b-83e4-aaa7c27e667b https://musicbrainz.org/artist/569c0d90-29dd-413b-83e4-aaa7c27e667b 431656 https://www.discogs.com/artist/431656 mp14871 http://www.npg.org.uk/collections/search/person/mp14871 http://www.nndb.com/people/510/000022444
Associated country	Great Britain England
Located	Edinburgh (Scotland)
Birth date	1965-07-31
Place of birth	Yate (England)
Field of activity	Young adult fantasy Fantasy fiction Young adult fiction Fantasy
Affiliation	University of Exeter
Profession or occupation	Authors Novelists
Special note	Machine-derived non-Latin script reference project. Non-Latin script references not evaluated.
Found in	Harry Potter, 1998: CIP t.p. (J.K. Rowling) The Washington post, 10-20-99: p. C1 (J.K. Rowling; Joanne Kathleen Rowling, 34, creator of Harry Potter) Fantastic beasts & where to find them, c2001: t.p. (Newt Scamander) copr. (J.K. Rowling) Quidditch through the ages, c2001: t.p. (Kennilworthy Whisp) copr. (J.K. Rowling) Harry Potter i więzień Azkabanu, c2001: t.p. (Joanne K. Rowling) Beacham's guide to literature for young adults, 2001: v. 11, p. 209 (Joanne Kathleen Rowling, b. July 31, 1965 in Glouchestershire, England) J.K. Rowling, c2002: p. 9 (Joanne Kathleen Rowling; Jo is still the name she likes to be called) Hari Poter veha-asir me-Azkaban, c2000: t.p. (G'e. Ke. Roling) Harry Potter a Tajemná komnata, 2000: t.p. (Joanne K. Rowlingová) copyr. (Joanne K. Rowling) Harri Potter i filosofs'kyĭ kamin', 2002: t.p. (Dzh. K. Roling) p. facing t.p. (J.K. Rowling [in rom.]) Ha-ʻu'e Phod-thar dań tshe rdo, 2007: t.p. (J.K. Ra'u-lin) t.p. verso (J.K. Rowling) Hairī Pōtara aura māyāpañchī kā samūha, 2008: t.p. (Je. Ke. Rolinga) Hali Bote yu mo fa shi, 2000: t.p. (J.K. Luolin) Hyārī Paṭāra enda di phīlasaphārasa stona: t.p. (Je. Ke. Rāolim) t.p. verso (J.K. Rowling) Fenomen Garri Pottera, ili, Razsblachenii͡a chernoĭ magii, 2007 p. 3 (Дж.К. Роулинг = Dzh.K. Rouling) p. 4 of cover (Джоан Кэтлин Роулинг = Dzhoan Kėtlin Rouling) Garri Potter i uznik Azkabana, 2007: t.p. (Дж.К. Ролинг = Dzh.K. Roling) Wikipedia viewed via the WWW June 22, 2011 (Rowling read for a BA in French and Classics at the University of Exeter) The casual vacancy, 2012: t.p. (J.K. Rowling) jkt. (1st adult novel) J.K. Rowling, via WWW, Nov. 13, 2012 (lives in Edinburgh with her husband and family) The cuckoo's calling, 2013: t.p. (Robert Galbraith) JK Rowling revealed as author of The Cuckoo's Calling, via BBC News - Entertainment & Arts website, July 14, 2013 (JK Rowling has secretly written a crime novel under the guise of male debut writer Robert Galbraith) Dbpedia, September 22, 2016 (Joanne K. Rowling; Joanne "Jo" Rowling; Robert Galbraith; J K Rowling; Joanne Kathleen Rowling; JK Rowling; Rowling, J. K.; J.K. Rowling; British novelist; place of birth: Yate; date of birth: 31 July 1965)
Associated language	eng

LCCN Permalink: A Service of the Library of Congress — More information: LCCN Permalink FAQ

About | Site Map | Contact | Accessibility | Legal | External Link Disclaimer | USA.gov

Fig. 1. An Authority record of Rowling, J. K. in Library of Congress Authorities Record (https://lccn.loc.gov/n97108433)

5 The Case Study of a Hebrew Manuscript Catalogue

The potential and importance of authority files and the use of wikidata can be demonstrated through research conducted at Bar-Ilan University [8]. In this research the authors tried to develop a new semi-automatic methodology for the construction of event-based ontology from the NNL library catalogue of Hebrew manuscripts. Based on the constructed ontology, a new framework was developed and implemented for catalogue data enrichment, correction and its systematic quantitative analysis.

One of the major problems encountered in the research is the lack of an authority control file in the NNL manuscript catalog. The research showed that 44,338 unique persons involved in manuscript lifecycles, but only 10,867 of them have an authority record and a record in VIAF. This means that three quarters of the people do not have an authority record, and a lot of information is missing, mainly years of life and geographic information. In order to complete the missing information for various manuscript events, an automatic inference procedure based on the ontological entities linked to these events was devised (as illustrated in Fig. 2) [9].

In general, the missing dates and locations of the Manuscript_Biographic_Event can be inferred from the life period dates of its Agent (a person involved in the event). For example, for manuscripts missing the composition date and place for the composition events, this information was calculated from authors' life periods and activity sites, respectively. To complete the missing creation dates and places for the creation event, we used the corresponding scribes' entities, and to estimate censorship events' dates and places, censors' activity periods were utilized.

When people's life periods and locations were missing both in the catalogue and in the external resources, an automatic inference algorithm used to complete this data from the corresponding events related to these people in the ontology. For example, creation events' dates and locations were used to complete information about scribes. After having completed people's data, the inference procedure for events was applied again, until all possible completions were performed in the ontology.

The problem that arose was that some of the inferences were not correct because in the absence of authority control many of the people appearing in the manuscripts, the Inferences created were wrong. For example, the name *Shlomo ben Shlomo* appears seven times, in different roles (Fig. 3).

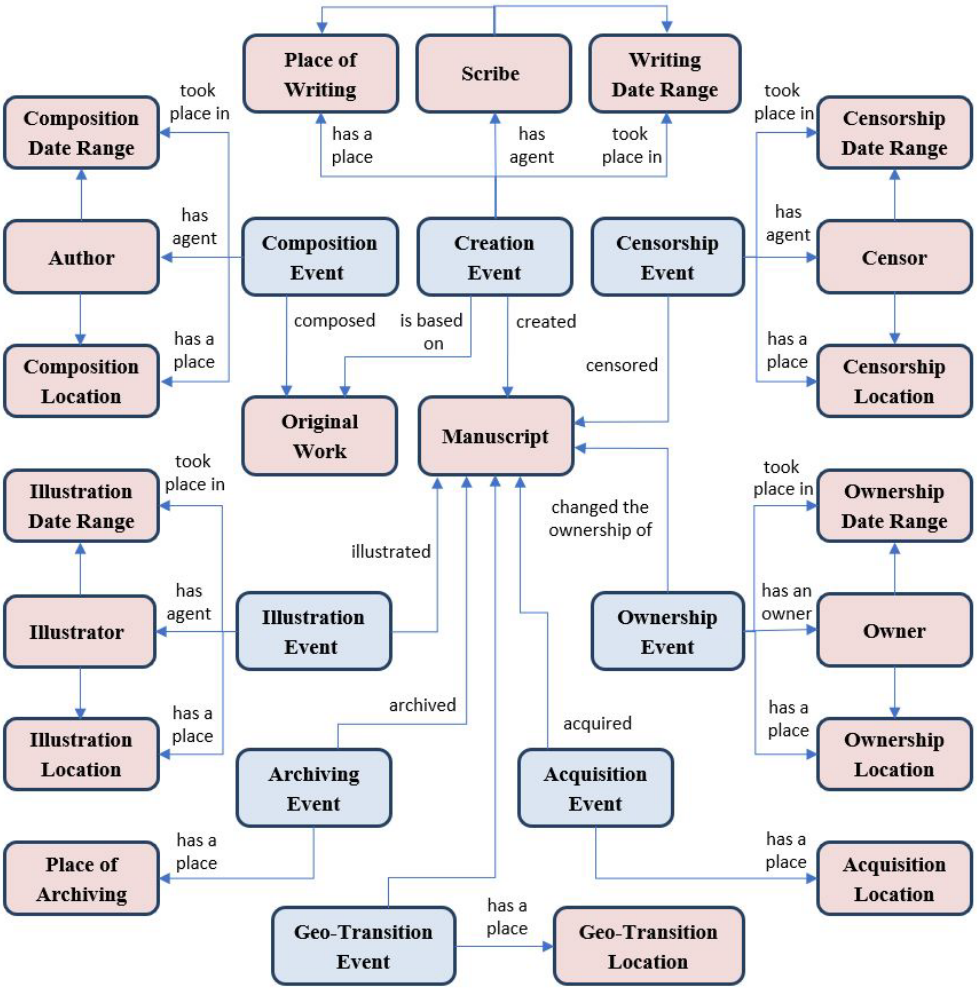


Fig. 2. The event-driven ontology scheme. From [9]).

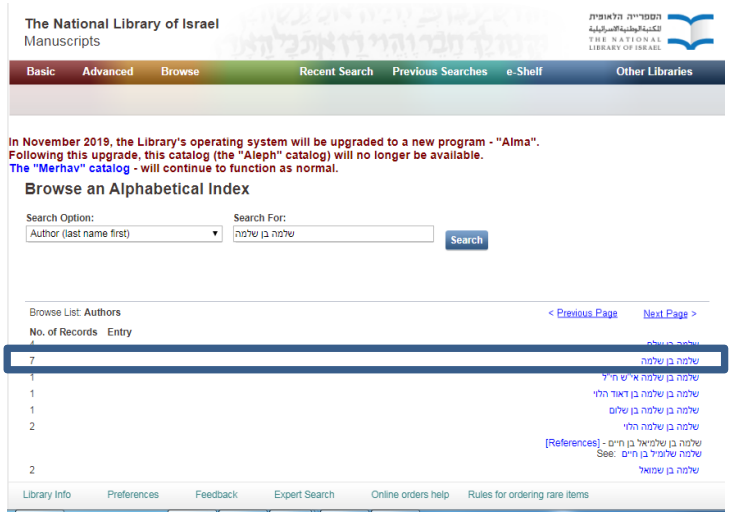


Fig. 3. The name Shlomo Ben Shlomo in NNL Catalogue

One of the 7 manuscripts, manuscript 41455, was copied by Shlomo Ben Shlomo in 1597 in Yemen (this data appears explicitly in the catalog). Based on this information it was determined that the scribe Shlomo Ben Shlomo worked in Yemen. The problem is that there were a number of people in different places and at different times with this name. On the other hand, a lot of information has been completed correctly. In figure 4 it can be seen that all the places highlighted in green were missing in the catalog and were correctly inferred.

The way to solve some of the problems is to convert the NNL library catalogue of Hebrew manuscripts collection metadata to linked open data and to create a Wikidata item for each of the manuscripts. This will lead to the enrichment of data and easy access to tools for querying and visualizing the collections. It will improve the inference from the catalog data and a lot more, like linking the data to additional information on the web such as books where people are mentioned, bibliographies, etc. An example of such a project can be seen in the project done in the National Library of Wales [10].

Scribe name	MS ID	Year of creation	Place of creation in the Catouloge	script type	Inferered place of creation
Rachamim ben Solomon	41046	1850	missing	Persian	Iran
	41065	1850	Iran	Persian	
	43879	1920	Israel	Sefardic	
	42419	1848-1853	missing	Persian	Iran
	41077	1850	missing	Persian	Iran
	131391	19th century	missing	Persian	Iran
Sangvineti, Refa'el Yehi'el	99916	1781	Trino (Italy) & Reggio Emilia (Italy)	Italian	
	103997	17th century	Trino (Italy) & Reggio Emilia (Italy)	Italian	
	87374	1787	candiano (Italy) & Reggio Emilia (Ita	Italian	
	184917	1781	Trino (Italy) & Reggio Emilia (Italy)	Italian	
	174943	Centuries 17-18	missing	Italian	Trino (Italy) & Reggio Emilia (Italy)
	61500	1781	Trino (Italy) & Reggio Emilia (Italy)	Italian	
	86788	Centuries15-18	missing	Italian	Trino (Italy) & Reggio Emilia (Italy)
	99919	18th century	Trino (Italy) & Reggio Emilia (Italy)	Italian	
	69637	1780	missing	Italian	Trino (Italy) & Reggio Emilia (Italy)
	79770	1778	Trino (Italy) & Reggio Emilia (Italy)	Italian	
Franco Mendes, David	171820	Centuries 18-19	missing		Amsterdam
	171594	Centuries 18-19	missing	Sefardic	Amsterdam
	95311	1788	missing	Sefardic	
	95318	1774	Amsterdam	Sefardic	
	148972	1782	missing	Sefardic	Amsterdam
	148945	18th century	missing		Amsterdam
	149165	18th century	missing	Sefardic	
	95359	1740	Padua (Italy)	Sefardic	
	185535	Centuries 18-19	missing	Sefardic	Amsterdam
	185515	1792	Amsterdam	Sefardic	
	33786393	1735	Amsterdam		
	148893	1780	Amsterdam		
	2511318	18th century	missing		Amsterdam

Fig. 4. List of places that were missing from the catalog and correctly inferred (highlighted in green).

6 Conclusion

I see great potential in Wikidata for the information world in general and for libraries in particular, and I hope the world of librarianship will take part in these developments. Wikidata acts as a hub, joining collections together in a web of cultural heritage data. If more collections will be added to this huge linked open network it will enable researchers to investigate and find new discoveries thereby revitalizing our cultural heritage data, which has been persevered in closed silos for hundreds of years.

References

1. Bermès, E.: Enabling your catalogue for the Semantic Web. In Chambers, S.(Ed.) Catalogue 2.0 : the future of the library catalogue. pp. 117-142. Facet, London (2013).
2. Varnum, K. J., New Top Technologies Every Librarian Needs to Know: A LITA Guide. Chicago, American Library Association (2019).
3. Dunsire, G.: Linked data for manuscripts in the Semantic Web. Summer School in the Study of Historical Manuscripts, (2012). http://www.gordondunsire.com/pubs/docs/LinkedDataForManuscripts.pdf, last accessed 2019/09/10.
4. Hastings, R.: Linked data in libraries: status and future direction. Computers in Libraries, 35, 12-16 (2015).
5. Joudrey, N. D., Taylor, A. G.: The organization of information. 4th ed. Libraries unlimited, California
6. Allison-Cassin, S. Research Libraries and Wikimedia: A Shared Commitment to Diversity, Open Knowledge, and Community Participation. Wikimedia Blog (2017). https://blog.wikimedia.org/2017/10/04/libraries-wikipedia-york-university-project/ last accessed 2019/09/10.
7. Forziati, C., Lo Castro, V.: La connessione tra i dati delle biblioteche e il coinvolgimento della comunità: il progetto SHARE Catalogue-Wikidata. JLIS.it 9(3), 109-120. (2018)
8. Zhitomirsky Geffet, M., Prebor, G.: Towards an Ontopedia for historical Hebrew manuscripts. Frontiers in Digital Humanities, 3 (3) (2016).
9. Zhitomirsky Geffet, M., Prebor, G., miller, Y.: Ontology-based Analysis of the Large Collection of Historical Hebrew Manuscripts. Digital Scholarship in the Humanities. (2019).
10. Evans, J.: Treasured Manuscript collection gets the Wikidata Treatment, National Library of Wales Blog (2019) https://blog.library.wales/treasured-manuscript-collection-gets-the-wikidata-treatment/ last accessed 2019/09/10.