Using Wikidata to Extract Cartographic Resources from Archival Collections

Summary: This paper discusses a project undertaken to make cartographic resources from the field notebooks of 20th-century cartographer Erwin Raisz discoverable on the Web, in relation to the works they influenced. Library catalog records of Raisz’s maps, books, and other materials in published or manuscript form, along with descriptions of his notebooks from an inventory spreadsheet, were imported into Wikidata, in order to show the benefits of a Linked Open Data environment for this kind of project. The case study that resulted shows how this work might be useful to a wide variety of professionals in the field of cartographic heritage, and how it can be replicated especially as applications improve. It also enumerates the benefits of doing so with Wikidata and other open source tools.

Introduction and Background

In the thirty years since the invention of the World Wide Web, librarians and cartographers alike have adapted their work to visual and hypertext-linked online spaces. Online public access catalogs (OPACs) and Geographic Information Systems (GIS) data are two of the innovations most familiar to everyday users, who have become accustomed to retrieving library materials over Internet connections far from a physical library building, and to seeing constantly updated maps in their news media, government reports, and educational materials. Those of us who work as map librarians use these technological tools in combination as a normal part of our daily work, such as when we catalog licensed GIS datasets or add georeferenced map scans to our holdings records. However, we have barely scratched the surface of what we could use the Web to accomplish. The World Wide Web Consortium (W3C) has been transforming itself into a more ambitious iteration over the past two decades, which its inventor, Sir Tim Berners-Lee, calls “the Semantic Web” because it parses meaning-based relationships between terms (Berners-Lee, Hendler, & Lassila, 2001). The Semantic Web is built on a system of subject-predicate-object statements outlined as the Resource Description Framework (RDF), which is interoperable across languages and resource types, making Web data internationally useable and easily discoverable. While once this idea of Internet use would have sounded utopian and fantastical, it is now a reality at least in part. Big data seems to be mentioned in every news cycle, governments create open data transparency protocols, and artificial intelligence (AI) applications use data across the Web to serve patrons far from a traditional computer screen. These changes have not gone unnoticed by information professionals, and librarians of all specialties and technology skill levels have been proposing creative responses. The International Federation of Library Associations and Institutions (IFLA) released an influential document, “Riding the Waves or Caught in the Tide? Insights from the IFLA Trend Report,” in 2013 that has been updated several times, most recently last year. The report and its updates cover everything from privacy concerns, to preservation of digital objects, to

* Harvard Map Collection, Harvard University Library, Cambridge, MA, USA [lena_denis@harvard.edu]
examples of how libraries can leverage big data within their extant Internet-based information exchange tools (Crepin-Leblond, Tuli, & Stielow, 2013). At the forefront of digital trends in libraries, information professionals have been seeking ways to expose all of this data – along with electronic library resources of all kinds – as Linked Data, the final ingredient in the Semantic Web. The W3C defines Linked Data as a “collection of interrelated datasets on the Web,” connected through RDF statements with their relationships clearly defined (W3C, n.d.). When set up successfully, Linked Data can deliver virtually any information across the Internet to a user who searches for it, because its being in RDF makes it machine-readable, translatable between human languages, and possible to query using Web standards (Bizer, Heath, & Berners-Lee, 2009). Because RDF is now a well-established standard for websites, the implications are enormous: if infinite Web and library resources can be made available as Linked Data, virtually any information on the Web could be queried as in a database. This enormous database is the vision of the Semantic Web, and it is already being built, in part by librarians.

From 2014 to 2016, Cornell University, Harvard University, and Stanford University collaborated through an Andrew W. Mellon Foundation grant, in a project called Linked Data for Libraries (LD4L), to develop library ontologies that would combine RDF and other Semantic Web standards with pre-existing authority records and classification schemes that libraries have already been using for decades (Krafft & Rayle, 2016). The project grew, gaining many more partner universities and changing its name and focus to include Linked Data production efforts. This expanded project is now in a second phase: Linked Data for Production: Pathway to Implementation (LD4P2). LD4P2 is working with the Library of Congress on creating tools to enhance traditional MARC (machine-readable cataloging) library records, as well as automating the production of new ones, as BIBFRAME (Bibliographic Framework Initiative) records. BIBFRAME is an ontology created by the Library of Congress, which expresses the statements and relationships typically found in library catalog records as Linked Data. Current efforts among the LD4P2 partners include a cloud-based cooperative cataloging environment that can leverage MARC to BIBFRAME records that have been converted elsewhere, but also store the records as RDF to allow for eventual exchange as Linked Data that can be exported to external resources. One external resource of interest for the project is Wikidata, the newest Wikimedia Foundation project, which stores the structured data that runs its sister projects, including the much more famous Wikipedia (Wikimedia Foundation, n.d.). Wikidata is interesting for LD4P2 and other library projects for at least two major reasons. First, though few non-Wikimedia editors have heard of it by name, Wikidata is already being used by massively popular AI applications and other Internet-based services, including those of Apple and Amazon (Simonite, 2019). Thus, its ability to reach a global audience is unquestionable. Second, not only is Wikidata Linked Data, it is Linked Open Data (LOD), defined by Sir Tim Berners-Lee as Linked Data published under an open license, and thus reusable for free (Berners-Lee, 2009). In the heavily-discussed context of Open Access and data transparency from academia to governance and everywhere in between, Linked Open Data is an invaluable framework.

In order to explore working with library records as both BIBFRAME and external resources like Wikidata, the LD4P2 partners had to decide on several experiment datasets with which they could begin testing. One of the datasets chosen was a historically significant group of materials from the Harvard University Library’s Map Collection, a 200-year-old, 500,000-map and 12,000-atlas collection of cartographic resources ranging from 1493 to the present day. It is this resource group,
combined with the goal of expressing library data as externally sourced Linked Open Data, that inspired the project discussed in this paper.

The Erwin Raisz Cartographic Notebooks

The subject of this project is a collection of cartographic field notebooks that were the work of 20th-century cartographer Erwin Raisz, who taught and worked at Harvard from 1931 to 1947. He was extraordinarily influential in encouraging the understanding of cartography as a historic discipline, distinct from geography and cosmotheography. Raisz wrote the first English-language textbook on the subject in 1938, General Cartography, invented the Armadillo Projection, and published many maps and atlases as well as additional textbooks on cartographic techniques and principles (Robinson, 1970).3

Figure 1. Scholia page, generated from Wikidata, listing major publications of Erwin Raisz.

While at Harvard, he assisted the American government – as did many academic cartographers – during World War II, and then spent post-war years mapping many regions of the world with sociologically-minded texts to accompany his maps. Cartographic historians often agree that he did so most significantly in collaboration with Gerardo Canet, a Cuban geographer with whom he collaborated on Canet’s Atlas de Cuba of 1949 (Reyes Nuñez, 2016).

3 Raisz’s influence on cartographic study has been discussed at length elsewhere and will not be here.
To the great fortune of the Harvard Map Collection, where Raisz spent much of his time while he was working at the University, he left there an archival collection of 66 field notebooks, containing notes and drawings stretching from 1910 (during his childhood) to the year of his death, 1968. They range in style from art books to mathematical studies, and contain inestimably valuable drafts of the work for his maps and books. They are in many languages, and cover places Raisz knew very well as well as those he knew barely at all.

This archive’s value to historians and geographers is self-evident, far beyond those who are interested in Raisz as an individual. The Map Collection, as a special collections unit within the
Harvard Library system, uses traditional library cataloging to disseminate information through MARC records on the Harvard Library OPAC. MARC records expose a great deal of useful and discoverable information about individual sheet maps, map series, atlases, and reference books. Map Collection staff are well-acquainted with these catalog formats and use them everyday. However, an individual MARC record is extremely limited for describing an archival collection, because it cannot expand to item-level within a collection-level description. (See Figure 4.)

Figure 4. Collection-level MARC record for Raisz notebooks.

There are solutions to this problem in other library cataloging environments, such as finding aids (typically used in archives) or XML schema-based collection descriptions (typically used for digital objects). However, these formats are not suitable for a MARC-based environment in which materials are rarely acquired – the Raisz notebooks being an exception based on special circumstances – that need other formats to meet library discoverability standards. The software

---

4 Due to multiple large-scale acquisitions, location changes, and other events beyond the scope of this paper, the Map Collection’s historic holdings are still in the process of being cataloged, along with the cataloging of new, ongoing acquisitions. As of this writing, about 50% of Collection materials are discoverable in the OPAC.
changes and staff training that would be required to implement a new cataloging system would occupy significant time and financial resources, unjustifiable for the sake of one collection. Therefore, the field notebooks had nothing but a MARC record to describe them, with no enumeration of their individual components or their influence on Raisz’s published works.

I was the staff member who created the record for the notebooks, and as I received more reference requests for Raisz’s materials in recent years from researchers who did not realize what the notebooks contained, I wanted to do more. Additionally, I learned about the LD4P2 project and its earlier phases and wondered if I might present a testable use case with the Raisz notebooks. I could use Linked Data to create item-level information under the umbrella of a collection-level entity, with relationships between the various items and other works by the same creator made visible. In December 2018, I approached an LD4P2 partner member about this possibility, and they helped me set up and implement a production plan.

Identifying and Gathering Datasets

First, we made the decision to express the information in Wikidata, for several reasons.

1. Production capacity compatible Linked Data description environments are not yet readily available, making a library-based Linked Data tool impractical.
2. Wikidata was easy for me to learn, because I already edit Wiki pages, for citations and accuracy, in my capacity as a librarian.\(^5\)
3. Wikidata has several well-tested tools for editing and exporting records in large batches, making the project scalable for importing many records in a short period of time.
4. Wikidata is open data, meaning my input would be enhanced as Linked Open Data.
5. Excitingly for AI enthusiasts, Wikidata items can be created through automated processes, so many pages are created or updated by bots. In fact, I discovered that a Wikidata item and unique identifier – formatted as a number following the letter Q for Wikidata entities – already existed for Erwin Raisz, created several years ago by a bot that scrapes Wikipedia pages to create new items.

---

\(^5\) See the 1Lib1Ref project for context on the collaboration between Wikipedia and librarians.
Over the following weeks, I manually edited some Wikidata pages, adding information as I do in Wikipedia, to learn the vocabulary and the structure of the RDF triples. I also learned basic SPARQL, the query language for RDF formats, and made visualizations such as Figure 6 to practice displaying relationships between both hierarchically and non-hierarchically related items.
Figure 6. Results of a SPARQL Query I made to show Wikidata entities for types of casseroles and my favorite casserole ingredient.

Then, I put together two lists of items to batch import as datasets. The first list was all the MARC records I could find in the Harvard OPAC that listed Erwin Raisz as a creator, and the Map Collection as their location. These specifications gave me 97 records, adjusted to 70 when I accounted for multiple editions of the same work. While certainly too large a number of records to edit individually in a reasonable timeframe, this group – showcasing a variety of formats ranging from educational pamphlets, to wall maps, to second edition textbooks – was a manageable number for experimentation with batch editing processes. My colleagues from the LD4P2 project and I devised a simple workflow to prepare this dataset for import:

1. Export the record list from our OPAC as binary MARC files.
2. Use a (freely available) tool called MARCEdit to consolidate and convert the binary files as a JSON file.
3. Open the JSON file in an open-source editor called OpenRefine, which can clean data and reconcile contained entities to external databases, including Wikidata.
4. Export the data from OpenRefine, once all clean-up and reconciliation was complete, into QuickStatements, another open-source tool that can turn OpenRefine data into a set of commands that Wikidata recognizes.
5. Use the QuickStatements commands to create 70 new Wikidata items from the dataset.

An LD4P2 colleague showed me how to begin the file conversion process and use OpenRefine and QuickStatements. I learned, started, and finished this five-step process within a month, and that was
while completing my other work tasks, because much of this process went quickly once set in place and could run in the background while I did my normal job duties.

Figure 7. Screen capture of OpenRefine during dataset reconciliation process.

![OpenRefine Screenshot](image.png)

Figure 8. Example of QuickStatements item creation commands.

```plaintext
Now that the first dataset was complete, I turned to my second one. The second list of items was an inventory of the Raisz notebooks that I had compiled in a spreadsheet some years before – mainly for my own benefit, but with the hope it would one day be useful in another format. The inventory was not encoded and was thus unusable in an OPAC, but it served as a shelflist for finding notebooks by date and location depicted.
```
Results: Archival Collections + Wikidata = Cartographic Resources

As of this writing, I have used OpenRefine and QuickStatements to generate notebook records from this shelflist. Not all of them are full Wikidata items yet, for technical reasons beyond the scope of this paper. However, some results I had hoped for are already visible.

Figure 9. Query showing notebook relationships.

This visualization is made from a SPARQL query I created to show the works of Erwin Raisz with his field notebooks as contributing factors. The information here about Raisz himself is the same bot-generated information seen in Figure 5. The Wikidata item I made for the notebook collection as a whole, based on the MARC record for them, displays an author link to Raisz. That item expands further, to show two of the individual field notebooks that have their own Wikidata items already. These two notebook items link in turn to the works that they influenced: Raisz’s maps and illustrations in the Atlas de Cuba (seen in Figure 2) and the text and drawings for a 1962 textbook he wrote, Principles of Cartography. Even with the current dataset still in progress, Wikidata has made it possible to extract cartographic resources from an archival collection in ways that were nearly impossible – certainly extremely time-consuming and difficult – with traditional cataloging. This process will become more robust when I finish my own work on this project, but what is most encouraging about this work is that because this is Linked Open Data, anyone can add more information. Whether researchers who see the Raisz notebooks and catch connections I have missed, librarians from other institutions with Raisz materials, or some other unforeseen stakeholder, anybody – anywhere in the world and writing in any language – can enhance my contributions to make knowledge about Raisz’s work more complete.

6 An inspiration for this visualization came from this blog post by Martin Poulter, Wikimedian in Residence for the Bodleian Libraries of Oxford University.
Conclusion: What Happens Next

This small case study, using Linked Data to demonstrate relationships between a library’s cartographic resources previously hidden in an undiscoverable format, is a neatly self-contained example of how this technology might benefit those interested in cartographic heritage. However, I believe it is scalable, and should be adapted for wider use. It was quick to learn and easy to implement, and has already created usable items that can be expanded, since it was implemented as open data with open tools. Because such excellent tools already exist for creating and converting information to the RDF format required for Linked Data, including ones that use AI, I barely had to create any new information within the records I had. I simply had to convert them into RDF to reveal the information they already contained, with minimal human and manual intervention required for entity reconciliation. As more of the Web becomes Semantic Web, this will get even easier and more widespread, without users even needing to understand how it works – just as many users of the Web in its current state are unable to explain how it works, since no understanding of its backend processing is necessary to take advantage of it. As libraries adopt more Linked Data practices in the style of the LD4P2 project, the sheer numbers of Linked Data(sets) will create a critical mass, so that the case for its utility will not have to be proven. As tools emerge that are currently in production for front-end data input and data querying, information professionals will have user interfaces that make the export of Linked Data possible from workflows that they already follow in their daily work. Rather than devising workflows that take weeks of data clean-up and manual intervention as mine did, one day (soon) it will be possible for anybody who catalogs library data to add the information they produce to the Semantic Web, with minimal changes to their practices.

Practitioners interested in Linked Data for cartographic materials also anticipate some new features in Wikidata itself, which is always evolving and expanding its own processes. A strength traditional cataloging has that Wikidata does not yet have is a way of describing bibliographic type data. The term “map” in Wikidata is very broad, and has several sub-types. The one used for Raisz’s published maps is “map edition,” but this is not yet standardized, and could itself change and required a future batch edit.  

Wikimedia is aware of this issue, and projects such as WikiCite are already underway to try and provide better sources for bibliographic citations.

Another important point to stress about using Wikidata is that it has a global community of people who can collaborate and assist each other. Though I relied on colleagues within my institution who could help me with basic starting steps, I also posited questions to Facebook groups and Wiki talk pages when I had small technical issues or wanted to know a best practice, and volunteer Wiki contributors from around the world helped me.  

Wikidata has a powerful infrastructure that can be used in tandem with other tools – open or not – to great effect for discovering cartographic heritage online and using it to show relationships that link seemingly disparate formats. Doing so has utility for many kinds of research and digital humanities projects that harvest cartographic data, especially from formats like archives that can be difficult to parse. It could also be used effectively to shape one of Erwin Raisz’s favorite topics: cartographic practices of the future.

---

7 A SPARQL query for map subclasses shows only 6812 maps and a query for map editions only shows 50.
8 As with other Wikimedia Foundation projects, a very large number of the data contributors and editors pioneering innovative work with Wikidata are doing so as unpaid volunteers, which LD4P2 and other similar projects make great effort to respect. The Wikimedia community’s stress on openness of information reflects the openness of its own structure and the global and professional diversity of its stakeholders.
Acknowledgements

I am inexpressibly grateful to Marc McGee, who first made me aware of the LD4P2 project, taught me everything I know about Linked Data and Wikidata, and patiently guided me through the subsequent steps of this project. I also thank the LD4P2 Partners more generally, especially Christine Fernsebner Eslao, for helping me conceptualize a working plan when I was still unfamiliar with the tools I could use. Lastly, I thank Bonnie Burns and Marty Schreiner in the Maps and Media Services Division of Harvard Library, for allowing me to pursue this project and making it logistically feasible to do so.

Bibliography


Simonite, T. (2019, February 18). Inside the Alexa-Friendly World of Wikidata | WIRED. Retrieved March 22, 2019, from https://www.wired.com/story/inside-the-alexa-friendly-world-of-wikidata/?fbclid=IwAR0NUROwIWTk9gyBB0GseHOaGgGO5xUcEXpq3e_ZxJFBYbHNL3k6BdEunPo
