Title
Exploring Bibliographic Records as Research Data

Authors
Hutchinson, Joshua
Wallbank, Sarah
Dickerson, Madelynn
et al.

Publication Date
2019-12-01

Peer reviewed

Exploring Bibliographic Records as Research Data

Sarah Wallbank Electronic Resources & Serials Cataloging Librarian, University of California, Irvine (**wallbank@uci.edu); Danielle A Kane** Digital Scholarship Services Emerging Technologies Librarian, University of California, Irvine (**kaned@uci.edu); Madelynn Dickerson** Research Librarian for Digital Humanities and History, University of California, Irvine **(mrosed@uci.edu); Joshua Hutchinson** Cataloging and Metadata Librarian, University of California, Irvine **(jchutchi@uci.edu)**

Exploratory Origins

This article describes the way in which a group of four librarians at the University of California, Irvine is exploring potential uses for bibliographic data from the library catalog for digital humanities (DH) research. The project started when the Research Librarian for Digital Humanities and History (Madelynn Dickerson), who has a background in collections and technical services, reached out informally to colleagues in the Cataloging and Metadata Services department. In a December 2018 email she wrote, “There are so many interesting intersections between DH and ‘technical services’ work. I would be really interested to work with you on something, and I’m particularly interested in ideas for demonstrating the value of library data for scholarly research if that’s something you’d ever be interested in working on too.” The Cataloging and Metadata Librarian (Joshua Hutchinson) heeded the call, and after a few exploratory conversations, the group ultimately included the Electronic Resources & Serials Cataloging Librarian (Sarah Wallbank) and the Digital Scholarship Services Emerging Technologies Librarian (Danielle Kane).

From these casual beginnings, our group developed the dual purposes of (1) showing campus researchers the possibilities of using the catalog metadata for their work, and (2) providing a hands-on educational project for the librarians involved. This project gave us an opportunity to practice managing and cleaning large datasets (using tools like C# MARC Editor and OpenRefine), and to practice using digital humanities research methods such as text analysis (using tools like Voyant Tools).

Exploratory Procedures

Scoping

Starting in December 2018, our group began regularly scheduled monthly meetings where we discussed our scope and planned procedures. We decided to focus on an analysis of women authors of print monographs in History (all monographs with Library of Congress call numbers within the C-F range). With this scope, we hoped to get a better understanding of how bibliographic data from UC Irvine’s library catalog could be used by researchers, what skills would be needed in order to successfully complete a similar project, and complete a use-case as a model for future work.

Our initial research questions included:

- Of all the monographs in our catalog with the call numbers C-F, how many were written by women?
- Are women historians likely to write about a particular topic within the discipline?
- Is it possible to accurately and ethically identify an author as a woman based on their name alone? How would one go about doing this for the purpose of scholarly analysis?

We chose these initial research questions because we found them interesting and relevant to our work, and because they helped us define a reasonable scope for the project while simultaneously forcing us to engage with bigger critical questions beyond solely number crunching. In addition, as all members of the team were somewhat familiar with the metadata included in bibliographic records, the scoping exercise for this project involved thinking about how this bibliographic data could be used to achieve interesting research results. For instance, because all records include publication information (generally place of publication, publisher and date) additional research questions related tothe diversity of the collection in terms of place of publication and publisher— does the UC Irvine print collection focus primarily on the Anglo-American world? Is history primarily published/collected from university publishers in the US and the UK? Has the library been collecting broadly across the decades, or are there interesting patterns that might be gleaned from studying the date of publication?

Downloading the Data

Once we had decided on a general scope, our next step was to download the bibliographic data from our library management system, Ex Libris' Alma. Our first data download from the Alma Analytics module took place on January 11, 2019 as a Binary MARC file that included records with LC Classification that began with C, D, E, F and had a location of Langson Library (the building that houses our humanities monographs, including history). The intention was to capture all history monographs with a physical copy. We excluded serials by accounting for the location within the library, and the bibliographic leader byte 7. We expected that some level of serials and electronic books would come into our data, but for the purposes of this exercise, a small amount of imperfection was deemed acceptable. Exporting from Alma Analytics had to be completed in four exports (one per letter) because it appeared that Alma Analytics was unable to export more than 65,000 records at a time. We then ran the MARC records through the C# MARC Editor program (https://csharpMARC.net/) in order to create a .csv file. This data file was 306 MB and had 184,105 rows (this number was later reduced as the data were refined and additional serials were removed based on the Leader byte 07) and 220 columns, with each row representing a single physical book and each column indicating an individual MARC field.

Playing with the Data

Once we downloaded the data, the first thing we did was simply play with it as a means of becoming familiar with it and exploring the possibilities of what we could learn. This dedicated "play time" informed later decisions about the direction of our

work and produced some preliminary statistics that, while not rigorous enough to draw scholarly conclusions from, gave us a sense of the real potential of the project.

We uploaded a spreadsheet with just the title data from our dataset, uncleaned, into Voyant Tools (https://voyant-tools.org), an open source, browser-based platform for text analysis. Voyant provides a default display featuring a visual word cloud of frequently occurring terms, and a corpus summary that includes statistics such as the total number of words in a document. According to Voyant, our (uncleaned, imperfect) corpus of book titles had 1,613,151 total words and 91,961 unique word forms. The most frequently occurring words in the corpus were *history* (12,903 occurrences), *war* (10,134 occurrences), *american* (9,582 occurrences), *la* (9,011 occurrences), and *world* (5,506 occurrences) (see Figure 1).

While unscientific, these exploratory results provided general insight into the makeup of UC Irvine's collection of history monographs. It is no real surprise that among history book titles, "history" is the most frequently occurring term. It is, however, a bit sad that "war" is the second most frequent. A term like "la" is likely a definite article appearing in multiple romance languages, and could potentially also be a reference to Los Angeles. The word "world" is interesting, especially in that via Voyant's "phrases" tool view, we see that the most frequent two-word phrase by far is "world war," which occurs 1,743 times (see Figure 2). Voyant's "phrases" view, as well as its separate "Contexts" widget are helpful in identifying the particular meaning of frequently occurring words by enabling researchers to see the other terms that appear in proximity.

We also wanted to look at author data. We had done some preliminary reading on similar projects, such as Peng et al's 2014 article, "Author Gender Metadata Augmentation of Hathitrust Digital Library," which explained the way that team had determined author gender using metadata available in the HathiTrust Digital Library. Peng et al used a range of name matching techniques, including Virtual International Authority File (VIAF) lookup, and matching data to baby name websites. We were not ready for that. What we did want to experiment with, however, were ways to simply tell which names appeared more frequently than others. As with the title data, we therefore uploaded a spreadsheet including only (uncleaned, imperfect) author name data to Voyant Tools. We're not prepared to make definitive declarations of author gender at this time, however Voyant Tools did display the most frequently occurring terms in our author data and they were *john* (5,312), *robert* (3,546 occurrences), *david* (3,423 occurrences), *william* (3,228 occurrences), and *james* (2,886 occurrences) (see Figure 3). Our author name dataset included both first names and last names, so it is likely that some of these frequently occurring names appeared as last names.

Text analysis is fairly unforgiving work, and uploading our uncleaned data highlighted many errors and discrepancies in our dataset that need to be cleaned. Playing around with the title and author data in this way helped us to see these errors so that we could target them in our next step, which was data cleaning. It also helped us to feel comfortable working with the dataset and to learn about the various features of Voyant Tools.

Cleaning the Data

Our first attempts at data cleaning were to discuss each of the columns contained in our dataset and to determine what data needed to be kept going forward. Our main focus was to start working with the publisher, location, and publication date fields. First, we created a file naming convention for versioning while cleaning the data. After reviewing the dataset, we determined that some columns of data were unnecessary for the current project and we removed these in order to reduce the size of the dataset before uploading the .csv into OpenRefine (http://openrefine.org/). OpenRefine runs in a browser window and the larger the dataset, the more memory OpenRefine will need to be able to work with it effectively. We would have run into memory errors without reducing the size of our dataset. As it was, we experienced lag on a number of clustering and merge operations.

OpenRefine was used to split the publication information (from the MARC 260 and 264 fields) into separate columns for place of publication, publisher, and date of publication by splitting multi-valued cells using the $ sign as a separator (see Figure 4). The results were reviewed and we made adjustments as necessary using Excel. It was important to make sure that the same information appeared in the same column and, due to differences in the amount of information contained in the 260 field, columns needed to be adjusted. Information that was originally contained in 1 column was split into and organized by hand using Excel into a total of 33 columns. We then re-uploaded the file, now with the publication information from the MARC subfields in separate columns, into OpenRefine. By using the facet tool, we first turned columns into a text facet. Then, using the cluster feature, we were able to remove extraneous information surrounding the data we were hoping to clean. We removed the subfield delimiters “a,” “b,”and “c”, and also extra punctuation such as periods, commas, colons, semicolons, and “less than” symbols. While we focused on standardizing the publication year from the MARC subfield ‘c’, we ultimately plan to standardize place of publication and publisher (MARC subfields ‘a’ and ‘b’).

Using the publication year, we considered splitting the single large data file into multiple smaller files by decade in order to make data cleaning and processing easier with limited computing power, and as a way to split the work between members. We tested sorting and splitting the data up by decade, but determined that this might introduce too many discrepancies between files so decided to keep all data together throughout the cleaning process. The fact that we had a file naming convention, and saved our files at each stage of the cleaning process, made it easy to take a step back in our process and proceed in a new direction.

Further, we made an effort to connect the MARC language and country codes with the corresponding English language term (e.g., *eng* would become English and *enk* would become England). We did this by querying the Library of Congress linked open data system, parsing the HTML for the page title, and then extracting the name of the language and country from that HTML. While this procedure worked in small samples, it is as yet unsuccessful for the full dataset, presumably due to the large size of the dataset and the computing capacity required.

Looking to the Future (Exploratorily)

This is only the beginning for the project. Our biggest questions remain unanswered and we look forward to taking the first steps towards our original goal of determining and evaluating authors' gender. Our next steps might include:

- Substituting MARC codes with text for human readability;
- Applying what we learned cleaning location and publication data fields to title and author data;
- Dividing the spreadsheet by decade for individual team members to do further data cleaning and temporal analysis;
- Performing a reconciliation of names in our dataset against databases that provide "name registries," such as Wikidata.

Conclusion

Playing with the data before actually cleaning it may seem out of order, but playing with it has helped us to know what we want to do, or need to do, for cleaning. Because one of the goals of this project is to use this as a demonstration of the potential of bibliographic metadata for research, we thought it was important to spend some time with our data, thinking about what trends and patterns we could glean from it. In addition, we were interested in challenging ourselves to learn new tools, and to make sure that all members of the team gained new skills. We are looking forward to tackling the next steps of the project in order to more definitively address our initial research questions.

References

"C# MARC Editor." Accessed November 8, 2019. https://csharpMARC.net/

Doboš, Ján, and Tibor Csóka. "CRISP-DM Data Evaluation Phase of Simple Bibliographic Data Evaluation Platform Design." *Innovative Methods in Education and Research*, n.d.

Lahti, Leo, Jani Marjanen, Hege Roivainen, and Mikko Tolonen. "Bibliographic Data Science and the History of the Book (c. 1500–1800)." *Cataloging & Classification Quarterly* 0, no. 0 (January 7, 2019): 1–19. https://doi.org/10.1080/01639374.2018.1543747.

Marciano, Richard J., Robert C. Allen, Chien-Yi Hou, and Pamella R. Lach. "'Big Historical Data' Feature Extraction." *Journal of Map & Geography Libraries* 9, no. 1–2 (January 1, 2013): 69–80. https://doi.org/10.1080/15420353.2012.732020.

Muñoz, Trevor. "Recovering a Humanist Librarianship through Digital Humanities." In *Laying the Foundation*, edited by John W. White and Heather Gilbert, 7:3–14. Digital Humanities in Academic Libraries. Purdue University Press, 2016. https://doi.org/10.2307/j.ctt163t7kq.4.

O'Dell, Allison Jai. "Book Artists Unbound: Providing Access to Creator Metadata with EAC-CPF." *Art Documentation: Journal of the Art Libraries Society of North America* 33, no. 2 (2014): 267–78. https://doi.org/10.1086/678527.

"Openrefine.Github.Com." Accessed November 8, 2019. http://openrefine.org/.

Pattuelli, M. Cristina, Karen Hwang, and Matthew Miller. "Accidental Discovery, Intentional Inquiry: Leveraging Linked Data to Uncover the Women of Jazz." *Digital Scholarship in the Humanities* 32, no. 4 (December 1, 2017): 918–24. https://doi.org/10.1093/llc/fqw047.

Peng, Zong, Miao Chen, Stacy Kowalczyk, and Beth Plale. "Author Gender Metadata Augmentation of Hathitrust Digital Library." *Proceedings of the American Society for Information Science and Technology* 51, no. 1 (2014): 1–4. https://doi.org/10.1002/meet.2014.14505101098.

Prescott, A. "Bibliographic Records as Humanities Big Data." In *2013 IEEE International Conference on Big Data*, 55–58, 2013. https://doi.org/10.1109/BigData.2013.6691670.

Royal, Peter. "Untangling Text: Voyant Tools' Knots for Text Analysis." *Medium* (blog), November 10, 2015. https://medium.com/dh-tools-for-beginners/voyant-tools-2-0-less-common-tools-for-text-analysis-a922cfcd85cb.

Sheffield, Philip, and Sam Saunders. "Using the British Education Index to Survey the Field of Educational Studies." *British Journal of Educational Studies* 50, no. 1 (2002): 165–83.

Suarez, Michael. "Towards a Bibliometric Analysis of the Surviving Record, 1701–1800." The Cambridge History of the Book in Britain, October 2009. https://doi.org/10.1017/CHOL9780521810173.003.

Helsinki Computational History Group. "Supplementary Material." Accessed January 14, 2019. https://comhis.github.io/2019_CCQ/.

Tools for Bibliographic Data Analysis. Contribute to COMHIS/Bibliographica Development by Creating an Account on GitHub. R. 2015. Reprint, COMHIS, 2018. https://github.com/COMHIS/bibliographica.

Tuppen, Sandra, Stephen Rose, and Loukia Drosopoulou. "Library Catalogue Records as a Research Resource: Introducing 'A Big Data History of Music.'" *Fontes Artis Musicae* 63, no. 2 (2016): 67–88. https://doi.org/10.1353/fam.2016.0011.

"Voyant Tools." Accessed November 8, 2019. https://voyant-tools.org/.