Measuring Collection Diversity via Exploratory Analysis of Collection Metadata

Jordan Pedersen

Presenter

ABSTRACT

As libraries commit to equity, diversity, and inclusion (EDI) initiatives, it is critical to measure the strengths and weaknesses of collections in response. Building off already established approaches of collection diversity audits, the author argues that it is possible to use MARC metadata to perform exploratory analyses at a large scale. This paper discusses the process of planning, preparing the data, and performing an analysis, following the trajectory of an analysis of the 2.8 million geographic subject headings present in the University of Toronto Libraries’ MARC records. The tools developed during this project are cited and available for reuse.

KEYWORDS

Collection diversity; metadata analysis; equity; diversity audit

As libraries, including the University of Toronto, increasingly make explicit their commitments to equity, diversity, and inclusion (EDI), it is critical to measure how well collections support these mandates, as well as identify areas for improvement. By approaching metadata as a data source that can be studied, rather than solely as descriptors about library materials, we can expand our capacity for self-reflection, and support our missions around EDI. While this session was intended to be generalizable and give participants the tools required to perform an analysis of their own metadata, it reported on a case study using the University of Toronto Libraries’ (UTL) metadata.

Collection diversity audits and metadata as scholarship

Collection diversity audits are tools that are commonly used in public and school libraries to determine the diversity of a collection, generally in relation to the diversity of the population the library serves. Karen Jensen developed a Guide for Conducting a Diversity Audit of Your Collections, which was shared with participants, including myself, of the Equity in Action: Building Diverse Collections workshop from Library Journal. Based off this framework, there are at least two types of collection audits: overall collection audits, which should be done on a regular basis on a sample of the collection; and book order audits, which are done on book orders prior to purchase. The processes of these audits are often labor-intensive, with auditors using a spreadsheet to mark off the measures of diversity that a resource satisfies. Strengths of these approaches are that they allow for depth of collection knowledge to be cultivated, and results may be stated with high degrees of certainty because the library materials have been studied closely. Because of the required labor, sampling is often used, with various sampling methods being chosen based on the best judgment of the auditors.

These approaches are valuable, but given the scale of the University of Toronto’s collections (approximately 12 million physical resources and 1.5 million electronic resources) and diversity of collections (there are resources in 341 languages, across forty-four libraries), working
with a sample seemed unsatisfactory. Therefore, all materials in the University of Toronto’s Integrated Library System (ILS) were included in this dataset. More detail will be given to the advantages and disadvantages that came with this approach in the planning an analysis section, but participants could also apply this exploratory analysis approach to a sample.

In addition, this analysis was developed to demonstrate the value of using library metadata as a data source for scholarship. As Rachel Sagner Buurma and Jon Shaw write, “The bibliographic records in libraries’ searchable online public access catalogs (OPAC) have recently taken on a new role as a source of bibliographic data that can be aggregated, shared, circulated, manipulated, transformed, studied, and interpreted.” They point to examples that use text mining and relationship analyses common in the digital humanities, but also to approaches which have been around for much longer, such as relying on techniques from book history and literary studies. Historically, much of this work has been done by faculty members and researchers, and there are fewer examples led by librarians.

Planning an analysis

Given the value of library metadata and collection analyses, seeing the detailed processes required to undertake this work can give other librarians the resources to start their own analysis. To begin an analysis, it is helpful to have guiding research questions, to understand the local context, and to confirm that data can be gathered to answer the questions being asked. In an exploratory analysis, these questions can change and often inform further work.

I began my analysis at UTL with the following questions: Can we measure the geographic diversity of our collections by using metadata? If so, are there areas of the world which are over- or under-represented? What impact does this representation have on the scholarship that can be done at our institution?

In asking these questions, something that I needed to consider was whether fewer subject headings would mean that we have fewer items about a region, or just poorer quality metadata. On one hand, these two problems would have two different solutions – one may require a review of acquisitions or collection practices, and the other, a review and augmentation of metadata. However, from a user’s perspective, they result in the same scenario: users cannot search and find these items, which in turn may impact what they choose to research. This is ultimately the most interesting part of the question to me, because it demonstrates the larger impact the library can have on the diversity of scholarship.

To best interpret results, it is also critical to have a foundational understanding of the local context and collections strengths within the institution. For example, at the University of Toronto, there is the East Asian (Cheng Yu Tung) Library, the Richard Charles Lee Canada-Hong Kong Library, and the Petro Jacyk Central & East European Resource Centre, all of which likely increased the number of materials that are about the countries or regions of the world that the respective libraries are focused on. And so, when one is considering a study of their own metadata, it is valuable to also consult with colleagues for their perspectives on the context of the institution and the answers that might be found.

After understanding the context, it is time to determine which parts of the MARC record are valuable for answering the questions being asked. When using records that adhere to the MARC standard, the amount of structure and standardization of the data can impact which fields are best to use. When determining the best fields for an analysis, they can be broken into the following categories:

(1) Control field (i.e. LDR, 008)
(2) A field that uses a thesaurus (i.e. 6XX)
(3) Free text (i.e. 2XX, 5XX)
Control fields are the most structured type of field, and will likely require the least amount of cleaning, a time-consuming process which will be explored in more depth in the next section. Data in fields that use thesauri should adhere to a standard, but it is possible for more deviation to occur that is difficult to track than in a control field. For example, when working with 650 fields, it is possible that the second indicator is incorrectly encoded and a record contains local subject fields that have been coded as Library of Congress. These types of errors may be difficult to notice by looking at the data, but still may impact results.

Finally, the free text fields may pose the most challenges because they can contain much variation which can be difficult to measure. For example, a 500 general notes field may include information about whether an index is included, the source of a dataset, or content information that is unformatted, and it is hard to reconcile all these things in a way that gives an answer to a specific question. However, if the questions being asked lend themselves to natural language processing (NLP), these would be the fields to use because the more standardized fields do not contain enough context (in other words, natural language) to create meaningful semantic relationships. Some fun examples of analyses that use free-text fields include using named entity recognition to pull proper nouns out of notes fields if looking for local people or places, or using NLP toolkits to determine the similarity between different records across libraries.

In some cases, one may be able to get data to answer the guiding questions from multiple places within a MARC record and will need to choose which field or subfield is best for the research question at hand. For example, when I was looking at geographic diversity, I could have chosen to use place of publication (260/264), topical term (650 $a$ or $z$), or the geographic names (651 $a$). I decided that it was most interesting to me to look at materials that were about a specific region, which narrowed it down to the 650 and 651 fields, both of which contain a medium amount of structure. I then decided to begin with the 651 field, and to expand to 650 if time permitted.

It is worth acknowledging before going further that a limitation of this study is that the thesaurus that UTL uses (primarily Library of Congress), uses language that is not neutral when talking about places. These subject authorities are created to serve the needs of a government built on colonialism, and so place names also reflect colonial names and biases, for example giving more granular terms to places deemed important to the users, and using outdated and incorrect terms for places that are considered “other.” It becomes likely that the data will include both insensitive subject headings and biased ones. As Sanford Berman wrote:

But in the realm of headings that deal with people and cultures – in short, with humanity – the LC list can only “satisfy” parochial, jingoistic Europeans and North Americans, white-hued, at least nominally Christian (and preferably Protestant) in faith, comfortably situated in the middle- and higher-income brackets, largely domiciled in suburbia, fundamentally loyal to the Established Order, and heavily imbued with the transcendent, incomparable glory of Western civilization.4

While Berman’s criticisms extend much wider than to that of geographic subject headings, similar criticisms apply. Because the thesauri are developed from a particular, colonial perspective, it lacks nuance and representation from other groups. For example, placenames known by Indigenous Peoples are overwritten by the Canadian and American state terms, and in other cases outdated, offensive blanket terms are used, such as “Orient.”

**Preparing the data for analysis**

Exploratory analyses are undertaken when a researcher is interested in becoming familiar with their data, sometimes prior to a statistical analysis. They may start with some loosely formulated research questions, but they are not committed to a specific test or thesis. Because of this structure, it is important to note that the findings of an exploratory analysis do not necessarily lead to conclusive results.
When working with MARC metadata for an exploratory analysis, the process will generally follow these steps:

1. Extract or obtain the data
2. Clean the data and store it in a useful way
3. Explore the data
4. Share results and formulate further plans for analysis or additional questions

In the analysis of the UTL MARC records, the data (specifically the record number and 651 fields) were extracted from a file that contained all the bibliographic records in UTL’s ILS, rather than working with a sample. To do this, I developed a Python program that takes a MARC file as an input, then retrieves the record number stored in the 001 field along with the 651 fields (one per line), and exports the results as a pipe-delimited text file. The program is shared and free for reuse on GitHub.5 This process was developed so that all the tools being used for the analysis were system agnostic and could be reused by the larger library community.

To clean and store the data, I knew that I would need to develop larger categories (i.e. country) in order to count occurrences in a meaningful way. Otherwise, a count of the specific subject headings, would not tell me much about the areas of the world best represented, unless I could find a way to map each heading to a literal map of the world that could interpret the free text. These considerations required me to keep the tabular format of my data, but to add two additional columns, one for country and another for region, giving categories to my data that could be counted.

Unfortunately, this process was not as straightforward as I had initially hoped. The first challenge was to pick a dataset to standardize the country names. I settled on the ISO-3166 country names and codes because it is an internationally recognized standard, and I was able to find a freely available comma-separated values (CSV) version that also included regions and sub-regions per code; however, other datasets are available that may return slightly different results, based on whether specific states or nations are recognized.6

Once the country dataset was chosen, both the MARC data and the ISO-3166 country dataset were imported into an SQLite database, with one table for each of them. Queries were developed to update the MARC table with the appropriate country, region, and subregion. The data are available on GitHub for reuse.7 The reason that an SQLite database was used was so that the queries could be saved and reused (thereby making the analysis reproducible), and because it is an efficient way to manage large amounts of tabular data (the UTL MARC dataset included more than 2.8 million rows) within a free tool.

As the country codes were being assigned to subject headings, there were decisions that had to be made, including how to handle diacritics, varying levels of granularity within thesauri such as Library of Congress, historical place names, multiple subfield a’s for the same country within a single MARC record, and multiple countries within a field. For places which were expressed sometimes with diacritics and sometimes without, I did not make a change, but instead kept this in mind as a limitation of my data that could influence the counting of distinct headings for a specific country. To tackle the problem of varying levels of granularity, I wanted to avoid introducing bias by relying on my personal knowledge of world geography, so I chose not to assign a country if the country was not in the subject heading itself. To handle the issue of varying levels of granularity and multiple subfield a’s for the same country within a record, it became apparent that I would both have to look at a count (of all the times a specific country occurred), but also a count of distinct records (to account for records that had multiple subject headings about the same country). For historical place names, I assigned the country “Historical_Place” with the intention of revisiting later to avoid arbitrary assignation based on modern borders; although it is possible that some historical place names escaped this filter. Similarly, where there were multiple countries within a field, they were given “Multiple_Countries” as their country name, with the intention of revisiting. Once cleaned, 1,993,165 subjects were assigned
a country (72%), 771,196 were not yet assigned a country (28%), 32,188 were assigned “Historical_Place” (1%) and 15,512 were assigned “Multiple_Countries (0.5%). This meant that I felt I had enough to work with to start exploring the data.

**Performing the exploratory analysis**

While it would have been possible to explore the data solely by using SQL Queries, knowing that I would like to share these findings with my colleagues, I opted to use other tools which would enable more visual representations. The majority of the analysis was carried out using Python libraries (NumPy, pandas, and matplotlib) within a Jupyter Notebook. The questions asked were:

- Which countries are best/worst represented in the UTL collection?
- Which countries are best/worst represented when measuring by number of titles?
- Which regions are best represented/occur most frequently?
- Which regions are best/worst represented when measuring by number of titles?
- Which countries had the most distinct 650 $a$ subfields?

![Figure 1. Count of unique titles per country.](image-url)
Unsurprisingly, given the University of Toronto’s context as a large Canadian institution with a substantial East Asian library and Canada-Hong Kong Library, the countries that appeared to be best represented across these questions are the United States, the United Kingdom, China, France, and Canada (see Figure 1). Countries that occur least often are often small Islands, such as Cocos (Keeling) Islands, Saint Barthélemy, Sint Maarten, and São Tomé and Príncipe. This is similar whether measuring by total number of subject headings or by number of titles.

When looking at the region level and the total number of subject headings, Europe was the best represented (43% of the subject headings assigned a country were in the European region), followed by the Americas (27%), Asia (24%), Africa (5%), and then Oceania (1%). The results are comparable when measuring by total number of titles (see Figure 2).

Finally, when looking at which countries had the most distinct subject headings, China, the United Kingdom, France, Spain, and Italy came out on top, with Canada and the United States falling into position seventy-six and seventy-seven respectively, out of a total of 249 countries (see Figure 3). This lower representation may be due to the way that Library of Congress and other North American bodies assign subject headings about states, provinces, regions, and cities within North American countries, where the country name is not included.

Figure 2. Count of total titles per region.
Conclusions and next steps in the UTL analysis

For the UTL analysis, my next steps will include assigning country codes to smaller entities (such as the provinces, states, and cities that were missed in this round of analysis), potentially deciding on a method for using both 651 $a$ and 650 $z$ in this analysis, and most importantly, determining what other metadata elements would be of value to ask more detailed questions about the University of Toronto’s collection.

For example, a more detailed question may be to search for where books were published in conjunction with the geographic subject headings, to see if we can determine if resources are being published about local places, or if they are being studied from abroad. This question is similar to the “Own Voices” measurement used when assessing the diversity of works of fiction, which began as a way of measuring representation in fiction collections by assessing whether books with characters from underrepresented groups are published by authors that are also members of the group. Another valuable inquiry may be to determine how old the records are in conjunction with the geographic subject headings, which may demonstrate whether the diversity of our collection is increasing over time. Finally, investigating which topical terms are used in conjunction with geographic subject headings and whether there is any evidence to suggest that these combinations of terms reinforce prejudices would be valuable, as all representation is not necessarily positive representation. This case study with data from University of Toronto Libraries has opened my eyes to the possibilities afforded by analyzing MARC metadata, but it feels like only the beginning of asking increasingly more insightful questions around collection diversity.

Notes

1. Karen Jensen, “Conducting a Diversity Audit of Your Collection” (presentation, Library Journal: Evaluating, Auditing and Diversifying Your Collections, online, October 20, 2020) https://www.libraryjournal.com/event/evaluating-auditing-and-diversifying-your-collections-april-2020.
2. Ibid.
3. Rachel Sagner Buurma and Jon Shaw, “Slow Metadata,” PMLA/Publications of the Modern Language Association of America 135, no. 1 (2020): 188–94, accessed May 18, 2021. doi:10.1632/pmla.2020.135.1.188.
4. Sanford Berman, *Prejudices and Antipathies: A Tract on the LC Subject Heads Concerning People* (Jefferson, NC: McFarland & Company, Inc.), 15.
5. Jordan Pedersen, "pymarc-extraction," 2021, https://github.com/JordanPedersen/pymarc-extraction (accessed May 18, 2021).
6. Peter Sorensen, “ISO-3166-Countries-with-Regional-Codes” dataset, 2020, https://www.kaggle.com/petersersen360/iso3166countrieswithregionalcodes (accessed May 18, 2021).
7. Jordan Pedersen, "nasig_2021," 2021, https://github.com/JordanPedersen/nasig_2021 (accessed May 20, 2021).
8. Jensen, “Conducting a Diversity Audit.”

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Notes on contributor**

*Jordan Pedersen* is a Metadata Librarian at the University of Toronto Libraries, Toronto, Ontario, Canada.