Visualizing Library Metadata for Discovery

Poster

Myung-Ja K. Han
University of Illinois at Urbana-Champaign, USA
mhan3@illinois.edu

Stephanie R. Baker
University of Illinois at Urbana-Champaign, USA
srbaker@illinois.edu

Peiyuan Zhao
University of Illinois at Urbana-Champaign, USA
pzhao12@illinois.edu

Jiawei Li
University of Illinois at Urbana-Champaign, USA
jiaweil3@illinois.edu

Keywords: metadata; discovery; visualization; catalog records; data structure; MARC 21

Abstract

The benefits of visualization have been discussed widely and it is already implemented into library services. However, use cases for visualization have been mostly focused on collection analysis to improve collection development policies and budget management, not for discovery services that take full advantage of the rich information contained in library catalog records.

One of the challenges of working with library catalog records for visualization is the sheer volume of elements (such as control field, data field, subfield, and indicators) and information included in the MAchine-Readable Cataloging (MARC) format records. As is well-known, there are more than 1,900 fields in the MARC 21, which is just too many to use for effective visualizations (Moen and Benardino, 2003). In addition, some fields are used for recording the same information, for example, the control field 008 positions 7 to 14 and the subfield $c of the data field 264 are used for the production related date information. Instead of showing a clear relationship between resources, the large number of elements and duplicated information included in the catalog record may muddle those relationships in any visualization. The question then is which information added in which fields of the MARC 21 format catalog records should be considered essential information to be included in library catalog data visualizations for discovery.

According to Mischo, Schlembach, and Norman's research (2009) on users' search query terms analysis, users tend to use more than three words as search terms (i.e., known item search) rather than simple keyword searches. Many users also use full citations as search terms, thus showing that library users very often already know what they want when they come to the library gateway. Consequently, for the purpose of supporting Functional Requirement for Bibliographic Record (FRBR) User Tasks (IFLA, 2017), such as finding, identifying, selecting, and acquiring (along with browsing), library discovery service systems *do not* need to index all of the elements included in MARC 21 format catalog records. What is needed instead is only the key information that affects the discovery services, such as access point and authorized access point that connect FRBR Group 1 entities (e.g., work, expression, manifestation, and items) defined by the Resource Description and Access standards (Library of Congress, 2017).

TABLE 1. Data used for the prototype discovery service that employed visualization tool.

Data	MARC data fields	Data	MARC data fields
Name (Agent)	100, 110, 111, 700, 710, 711	Subject	050, 082
Title (Work)	130, 245, 246	Date	260 $c or 264 $c
Bibliographic record identifier	001 (Local bibliographic record ID)	Holdings record identifier	004 (Local holdings record ID)

Since visualizations work best for showing relationships between resources, the researchers at the University of Illinois at Urbana-Champaign Library developed entity relationships between 'work (title),' 'name' and 'subject.' Those relationships are displayed through visualizations that provide opportunities for users to understand, identify, and find related and similar resources in a more effective and organized manner. The information used for these relationships was extracted from a sample of 300,000 randomly selected library catalog records (from 7.4 million total catalog records) as shown in table 1 above. A prototype discovery service that employs the visualization tool D3.js (https://d3js.org/) was created.

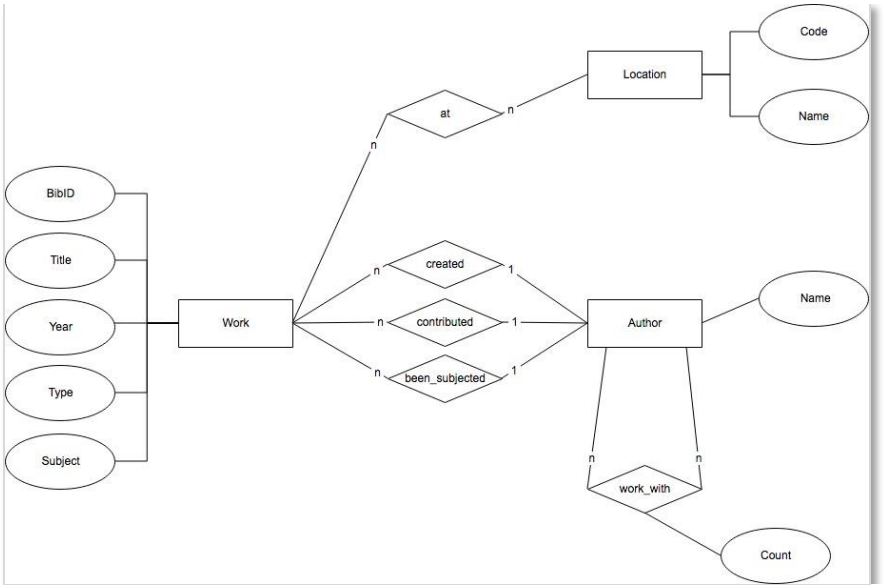

FIG. 1. Entity relationships diagram for the work search.

The new prototype discovery service supports only two simple search options, work and name, with the assumption that users will perform a 'known item search,' as mentioned earlier. The search results page displays related resources by visualizing relationships between entities as shown in figure 1. For example, if a user starts the search with a title, then the result page allows the user to browse related resources (works) by the same author, on the same topic, with the same publication date, or having the same holdings library(ies). If a user starts the search with a name, the search results page allows the user to browse by relationships associated with the name, such as works created by the name, close collaborators, and subject areas of all works associated with the name. Figure 2 shows an example of the resulting page from a name search.

This is different from the current faceted browsing services provided in many existing library discovery services. Instead of displaying numbers of items with the same information as a list, the prototype discovery service combines the results together and displays them as a visualization. When the user selects the item (work), then the prototype discovery service displays all indexed information including holdings libraries. On the same page, the prototype provides a link to the full

catalog record page in case the user wants to see all of the information included in the MARC 21 format catalog record that may help users to identify and select the resource.

This experimentation confirmed that visualizing library catalog data is not that easy even with small set of data from a sample records. The challenges include inconsistent terms used in the records, data quality, and granularity of data in certain MARC21 data fields. Although we decided to select the data associated with the access points for this experimentation, notions of what constitutes key information for discovery services is not clear yet. However, the prototype showed the benefits of using a selective set of data critical to discovery and visualization, as opposed to using all of the information included in an entire catalog record.

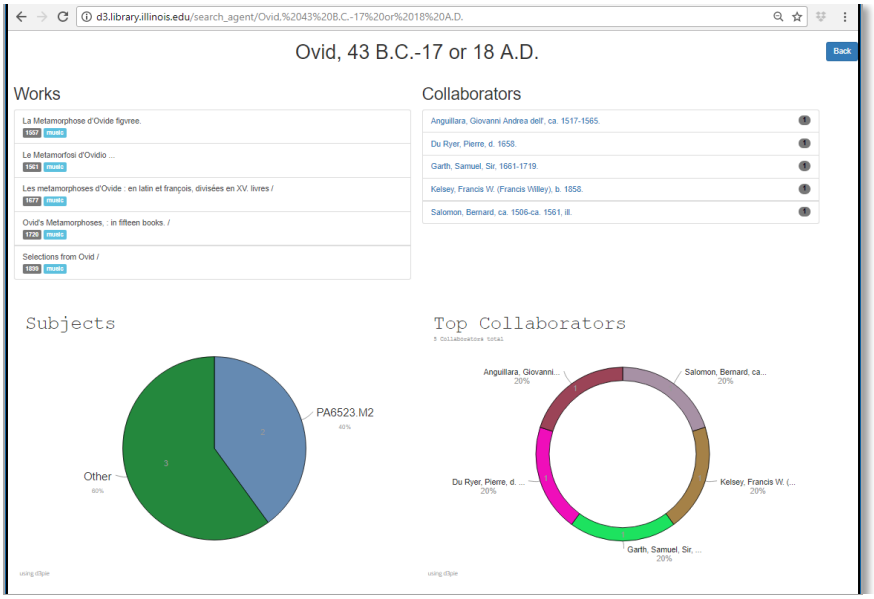

FIG. 2. A search result page that shows works, subject and collaborators related with the name.

For the next step, we will work with a complete set of library catalog records to test the full functionality of the service and the impact of the visualization. It will also include using URIs of linked data sources for entities that would significantly improve the library's visualization-based discovery service. With the maturing of the BIBFRAME ontology, our project will use the set of BIBFRAME vocabularies describing entities, such as work, instance, item, agents, subjects, and events (Library of Congress, 2016) for visualization by adapting workflows established from this experimentation. We also hope that a proper user testing should be conducted to identify an ideal set of bibliographic data used for the discovery services.

References

International Federation of Library Associations and Institutions (IFLA). (2017). Functional Requirements for Bibliographic Records (FRBR). Retrieved, May 5, 2018, from https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/8915.

Library of Congress. (2016). Overview of the BIBFRAME 2.0 Model. Retrieved, July 17, 2018, from https://www.loc.gov/bibframe/docs/bibframe2-model.html.

Library of Congress. (2017). Resource Description and Access (RDA). Retrieved, May 5, 2018, from https://www.loc.gov/aba/rda/.

Mischo, William. H., Mary Schlembach, Michael Norman. (2009). Modeling search assistance mechanisms within web-scale discovery systems. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, pp. 431.

Moen, William E., and Penelope Benardino. (2003). Assessing Metadata Utilization: An Analysis of MARC Content Designation Use. Proceedings of the 2003 Dublin Core International Conference. Retrieved, May 5, 2018, from http://dcpapers.dublincore.org/pubs/article/view/745/741.