The Library of Congress opened its catalogs to the world. Here's why it matters

The Conversation.com

July 20, 2017

Imagine you wanted to find books or journal articles on a particular subject. Or find manuscripts by a particular author. Or locate serials, music or maps. You would use a library catalog that includes facts – like title, author, publication date, subject headings and genre.

That information and more is stored in the treasure trove of library catalogs.

It is hard to overstate how important this library catalog information is, particularly as the amount of information expands every day. With this information, scholars and librarians are able to find things in a predictable way. That's because of the descriptive facts presented in a systematic way in catalog records.

But what if you could also experiment with the data in those records to explore other kinds of research questions – like trends in subject matter, semantics in titles or patterns in the geographic source of works on a given topic?

Now it is possible. The Library of Congress has made 25 million digital catalog records available for anyone to use at no charge. The free data set includes records from 1968 to 2014.

This is the largest release of digital catalog records in history. These records are part of a data ecosystem that crosses decades and parallels the evolution of information technology.

In my research about copyright and library collections, I rely on these kinds of records for information that can help determine the copyright status of works. The data in these records already are embodied in library catalogs. What's new is the free accessibility of this organized data set for new kinds of inquiry.

The decision reflects a fresh attitude toward shared data by the Library of Congress. It is a symbolic and practical manifestation of the library's leadership aligned with its mission of public service.

Some history

To understand the implications of this news, it helps to know a bit about the history of library catalog records.

Today, search engines let us easily find books we want to borrow from libraries or purchase from any number of sources. Not long ago, this would have seemed magical.

Search engines use data about books – like the title, author, publisher, publication date and subject matter – to identify particular books. That descriptive information was gathered over the years in library catalog records by librarians.

The library's action sheds light on this unseen but critical network. This infrastructure is invisible to most of us as we use libraries, buy books or use search engines.

For many, the idea of a library catalog conjures up the image of card catalogs. The descriptions contained in catalog records are "metadata" – information about information. Early catalog records date back to 1791, just after the French Revolution. The revolutionary government used playing cards to document property seized from the church. The idea was to make a national bibliography of library holdings confiscated during the Revolution.

For many years, library collections were organized individually. As the number of books and libraries grew, the increased complexity demanded a more consistent approach. For example, when the Library of Congress purchased Thomas Jefferson's personal library in 1815, it arranged its collections around Jefferson's personal system organized around the themes of memory, reason and imagination. (Jefferson based this on Francis Bacon's own model.) The library sought to arrange its collections on that model into the 19th century.

As the number of books and libraries grew, a more systematic approach was needed. The Dewey Decimal System appeared in 1876 to tackle this challenge. It combined consistent numbers ("classes") with particular topics. Each class can be further divided for more specific descriptions.

In the 1890s, the library developed the Library of Congress Classification System. It is still used today to predictably manage millions of items in libraries worldwide.

Catalogs, cards and computers

By the 1960s, systematic descriptions made the transition from analog cards to online catalog systems a natural step. Machine-Readable-Cataloging (or MARC) records were developed to electronically read and interpret the data in bibliographic cataloging records. The structured categorization coincided naturally with the use of computers.

Now, MARC records too are on the way out, making room for more modern and flexible standards.

The Library of Congress remains a primary – but not the only – source for catalog records. Individual libraries produce catalog records that are compiled and circulated through organizations like OCLC. OCLC connects libraries around the globe and offers an online catalog. WorldCat coordinates catalog records from many libraries into a cohesive online resource. Groups like these charge libraries through membership fees for access to the compiled data. Libraries, though, typically do not charge for the catalog

records they produce, instead working cooperatively through organizations like OCLC. This may evolve as more shared effort and crowdsourced resources can be combined with the library's data in ways that improve search and inquiry. Examples include SHARE and Wikipedia.

One month later

In the short time since the Library of Congress' data release, we see inklings of what may come. At a Hack-to-Learn event in May, researchers showed off early experiments with the data, including a zoomable list of nine million unique titles and a natural language interface with the data.

For my part, I am considering how to use the library's data to learn more about the history of publishing. For example, it might be possible to see if there are trends in dates of publication, locations of publishers and patterns in subject matter. It would be fruitful to correlate copyright information data retained by the U.S. Copyright Office to see if one could associate particular works with their copyright information like registration, renewal and ownership changes. However, those records remain in formats that remain difficult to search or manipulate. The records prior to 1978 are not yet available online at all from the U.S. Copyright Office.

Colleagues at the University of Michigan Library are studying the recently released records as a way to practice map-making and explore geographic patterns with visualizations based on the data. They are thinking about gleaning locations from subject metadata and then mapping how those locations shift through time.

There's a growing expectation that this kind of data should be freely available. This is evidenced by the expanding number of open data initiatives, from institutional repositories such as Deep Blue Data here at the University of Michigan Library to the U.S. government's data.gov. The U.K.'s Open Research Data Task Force just released a reportdiscussing technical, infrastructure, policy and cultural matters to be addressed to support open data.

The Library of Congress' action demonstrates an overarching shift in use of technology to meet historical research missions and advance beyond. Because the data are freely available, anyone can experiment with them.

Source: http://theconversation.com/the-library-of-congress-opened-its-catalogs-to-the-world-heres-why-it-matters-78570