“An Entirely Too Brief History of Library Metadata and a Peek at the Future, Too”

Even if you feel unfamiliar with metadata, you may have heard the definition, “metadata is data about data.” This is technically true but not very illuminating. The title of a book is metadata. The length of a feature film is metadata. The date of a treaty is metadata. Metadata is information about other data, and in the case of libraries the “other data” is usually an information object - like a book, film, or government document. So, you may *feel* unfamiliar with metadata, but the truth is that we all rely on metadata all the time in our daily lives. If you work in a library (and you probably do), then you use metadata every time you search for a book or article. In fact, metadata probably plays a big role in a lot of what you do.

A conversation about metadata can get very theoretical very quickly. So, in honor of the History Librarians’ discussion group, I will present an entirely too brief history of library metadata as a way to contextualize our current metadata challenges. In libraries, many of these challenges originate in library data models developed in the nineteenth century and early twentieth centuries. Our choice of data models and standards has proliferated since then, and I hope this brief history makes the sea of library metadata a little less overwhelming.

Catalogs and Metadata

The card catalog is a good place to start. It is a very real and tangible example of metadata in action, and most of us remember using one. The precursor to the card catalog was the inventory - a simple list of all the holdings in a given collection.[1] Inventories are basically lists of metadata. Inventories are great for keeping track of all the things you have, but it is extremely tedious to read through an entire list of items to find that *one book* you need.

Enter the card catalog. The card catalog was pre-electronic data processing.[2] Card catalog entries are not dissimilar to what you might find in an inventory, but the portability and interchangeability of the cards allowed librarians to organize all the holdings in a collection by different indexes - an author index, title index, or subject index to name a few possibilities. Now, you can search for all the holdings on a given subject or by a given author without going through the entire list of holdings.

Librarians established rules that governed the language of the catalog. These rules were derived from specific functions the catalog was meant to fulfill. For example, in order to functionally collocate all the works by the author known as Mark Twain - we need to figure out a way to deal with Samuel Clemens. This is no less true in a digital environment than it is in the paper-machine that is the card catalog - but we are getting to that.

---

[1] Umberto Eco’s love letter to lists, *The Infinity of Lists: an Illustrated Essay* explores our human fascination with list making of all kinds as a means of expressing infinity.

[2] Markus Krajewski’s 2011 book, *Paper Machines: about card catalogs, 1548-1929* is an excellent source for an in-depth look at this technology.

In 1967, the rules of the catalog were codified and published as the *Anglo-American Catalog Rules* (AACR). A second edition was published in 1978, inventively titled AACR2. These rules outlined *what* information should be in a record, *how* to phrase that information, and how to deal with situations like Mark Twain and Samuel Clemens. This standardization was really powerful, because standardized parts are interchangeable within a system (think Ford's Model-T or Ikea). Libraries could *order* cards with all the metadata about a particular book pre-made! Users could search every library collection using the same technique and feel reasonably assured they would get the same results!

In the early 1960s the Library of Congress invested in the development of a data standard that would make catalog records computer-readable. This resulted in the creation the MARC (Machine-Readable Cataloging) standard. If AACR2 gave us the language to describe items in our collection, MARC was a way of formatting that language into a structure a computer could process. AACR2 and MARC are not the same thing, and this is an important distinction. You can populate a MARC record with metadata that does *not* follow the rules of AACR2, and you can follow the rules of AACR2 in a record that is *not* MARC. In the parlance of catalogers, AACR2 is a content standard, and MARC is a transmission standard. They are both metadata standards.

The card catalog is a data model. It is a navigable representation of the library that allows us to understand our collections in a way we cannot when we are standing in front of a row of shelves. A data model shows how the elements of data are organized and structured, *and* how they relate to each other and to properties in the real world. The card catalog showed you where a book *by* Mark Twain was located in the library, where that author's *other* books were located, and where books *about* Mark Twain were located. When the card catalog moved to a computerized environment, it needed to retain this same functionality.

An Integrated Library System (ILS) is our current-day stand in for the card catalog. Most (if not all) ILS are built to receive, exchange, and edit MARC records. Because MARC was specially designed by libraries and for libraries, it is not easy to exchange or read MARC data over the web or with other, non-library computer applications. Our inability to share the metadata that is already encoded in MARC is a huge barrier to the visibility of library collections on the web.

This brings us to the twenty-first century. The internet and rapid adoption of digital documents has been an enormously disruptive shift for virtually every industry on the planet. In an attempt to address this, the Library of Congress initiated the development of a new transmission standard for library metadata. BIBFRAME will (hopefully) replace MARC as the carrier of library metadata.[3] The "rules" have also gotten an update. Resource Description and Access (RDA) was written to replace AACR2 as the content standard for library metadata.[4] Both of these changes are designed to make library metadata more malleable, and easier to share on the

[3] See the Library of Congress' BIBFRAME site for more information: https://www.loc.gov/bibframe/
[4] See the Library of Congress' site on RDA for more information: https://www.loc.gov/aba/rda/

web. Library records in BIBFRAME will be easier to break down into their component parts, and will meet Resource Description Framework (RDF) specifications.

RDF is, in a nutshell, a metadata model that is optimized for the web.[5] Right now a library record is a static thing made up only of itself. The smallest unit of data is "the record" just like a card in a card catalog is the smallest unit of data. It is very difficult to break a single data point (like "the author") out of "the record." When records are formatted to meet RDF specifications, each piece of data is its own entity. Records can be produced on the fly from different data points by linking the data together. For example, the entity "Samuel Clemens" would have a relationship with "Mark Twain" and a relationship with "A Connecticut Yankee in King Arthur's Court" and relationships with many other individual entities. These entities are interconnected and fluid. So, just as librarians could *order* cards for the card catalog ready-made - a well-formed entity on the web is a ready-made data point that a librarian can link to instead of typing into a record. With RDF when a user performs a search, we are trying to pull in data from the web to form records on-demand instead of hosting a collection of indexed records in our online catalog.[6] This way of structuring data is also called Linked Data. It is a very hot term right now among library metadata-types and catalogers.

Metadata and Standards

Different technologies have had their own metadata standards grow up around them. Standards to structure data tended to develop in highly specific ways.[7] Library metadata grew up around the concept of the card, and new ways of formatting our data will allow us to break out of that mold. But, there are many standards. The Visual Resource Association Core schema (better known as VRA) is a metadata standard that can be used in MARC and is specifically suited to describe art objects.[8] Dublin Core (DC) is a very simple metadata standard that has been adapted for a wide variety of purposes. It is a set of fifteen elements that can be deployed using different vocabularies, so DC can be adapted to describe different kinds of things (books, art, data sets, etc.). DC is very popular in libraries. A number of institutional repository and digital collections software tools are designed to work with DC metadata, including DSpace, Omeka, and CONTENTdm.[9] There is truly a multitude of metadata standards.[10] Selecting the "right" standard for describing a collection of data really depends on the project, who will be using the data, how the data will be used, and if the data needs to be shared.

It is possible to transform metadata from one standard or structure to another, but it is not always easy. Often, there are no clear guidelines, and there is always a risk of data-loss with data transformation. The Frances G. Spencer Collection of American Popular Sheet Music at

---

[5] See the W3C's site on RDF for more information: https://www.w3.org/RDF/
[6] Google knowledge graph is a good example of what a record made "on the fly" can look like.
[7] Medical records are very structurally different from a library record or a report card.
[8] Learn about VRA Core here: http://www.loc.gov/standards/vracore/
[9] The Dublin Core Metadata Initiative (DCMI) site has everything you ever wanted to know about DC: http://dublincore.org/
[10] See xckd's take on standards: https://xkcd.com/927/

Baylor University is an example of data transformation at work in a library. This collection is comprised of over thirty-thousand sheet music titles from the late eighteenth to early twentieth centuries. The collection was cataloged on cards, using a locally-developed (at Baylor) vocabulary to describe the styles and subjects of the music. In an attempt to make this collection accessible to our users, we have been cataloging each title in a MARC record that we can load into our ILS, which users can search through our online catalog.[11] In the early 2000's this collection was selected for digitization and ingestion into our libraries' digital collections.[12] Our digital collections software is CONTENTdm, which uses Dublin Core metadata to represent items in the digital collection. The sheet music titles had already been cataloged twice - once on cards and again in MARC records. We could catalog each title *again* in DC, *or* we could transform the metadata we already had from MARC to DC. These two standards do not match each other one-to-one. MARC is much more granular than the simple fifteen elements in DC. The rules that govern how to go from one standard to another is called a *metadata crosswalk*.[13] Currently, we transform these records in batches using an open-source data transformation tool called OpenRefine.[14] The decisions we made about which field in MARC translates to which field in Dublin Core is our metadata crosswalk. Our librarians have used other methods, but we are constantly re-evaluating the process to find solutions that are less time-intensive and result in better data quality.

This is why working with metadata is confusing and frustrating and truly fascinating. The process we established to transform the Spencer sheet music titles from MARC to DC in 2001 is *not* the same process we use now. The software we use to transform the metadata now *did not exist* in 2001. The tools and the problems are constantly changing. Metadata is as important as ever for finding and retrieving items, but it also serves for managing large sets of things - books, online collections, data sets. It is as important to know what your users have access to as it is to know when an item was purchased, how much it cost, and when it was delivered (metadata we usually hide from our users). This kind of technical work involving metadata has come under scrutiny in recent years as libraries have had to "prove their worth" to administration and adapt to more streamlined operations. This has, of course, created additional challenges for many libraries. I hope this very brief history has given you some concrete examples of what metadata is, and some ways it works in libraries.

Kara Long
Catalog and Metadata Librarian
University Libraries
Baylor University

---

[11] In 2008 we contracted with Flourish Music Contract Cataloging to catalog these titles in MARC on our behalf.
[12] See the Frances G. Spencer Collection of American Popular Sheet Music here: http://digitalcollections.baylor.edu/cdm/landingpage/collection/fa-spnc
[13] There are a lot of guidance documents about how to do this, but there are not hard and fast rules. Many of the decisions about crosswalking metadata will have to do with the specific data set you are working with.
[14] Find out more about OpenRefine here: http://openrefine.org/