Meeting the Challenge of Collocating Metadata: Explore Chicago Collections

Mary M. Case

One of the most exciting and fulfilling projects I have been involved with in my career is the founding of the Chicago Collections Consortium (CCC) (www.chicagocollections.org), an organization of libraries, archives, museums, and societies whose mission is to keep Chicago's history and culture alive. Early conversations with colleagues in the city began in late 2006, led by leadership at the Newberry Library and the University of Illinois at Chicago (UIC). Discussions ultimately led to a planning grant from the Andrew W. Mellon Foundation in 2011 and establishment of a newly incorporated organization in 2012. By 2014, the CCC achieved 501(c)(3) status and submitted a second grant to Mellon via UIC to build and implement its first major project, a web-based portal launched in late 2015.

From its 12 founding members, the CCC has grown to 27 members today from small historical and zoological societies to academic and public libraries to major cultural institutions. The CCC hosted its inaugural physical exhibit in 2015, has launched a lecture series, and provides programming for K-12 and community partners, and workshops for members. It is in the process of developing a major digital exhibit for late 2018 and envisions future grant-funded projects for digitization and preservation.

There are many aspects of the creation and development of this new organization that would be interesting to explore, but I thought the readers of *Technicalities* might be most interested in how

the organization met the challenge of developing a system that would be able to ingest metadata from multiple discovery systems (26 and counting) and present it to the public in a clear, easy-to-search interface under constraints that, I must admit, I thought would make the goal unachievable. My explanation below is at a fairly high level, but I have included citations to detailed documents that I hope those interested will pursue.[i]

Designing the Portal – Explore Chicago Collections

As mentioned, the first major project of the CCC was to build a portal through which researchers, students, and the interested public could discover more easily what materials were in our members' Special Collections related to the history and culture of Chicago. We knew that not all finding aids were online and not all finding aids were consistent within an institution, let alone across institutions; many collections were unprocessed; researchers had to go from institution to institution just to discover which held relevant collections; and even special collections librarians in the city did not know what other institutions held. A quick and dirty spreadsheet developed early on of collections from several members indicated that collections had actually been split among several institutions by donors in the past without current staff realizing it. This discovery alone led UIC to de-acquisition a small collection and send it on to another of our members who had the bulk of the related materials. The spreadsheet also quickly served as a new reference tool for Special Collections librarians who could now help patrons find needed materials elsewhere in the city.

During the planning phase, a number of decisions were made as a result of recommendations from teams of staff from member institutions in collaboration with consultants hired for the project. These decisions included defining the scope and format of materials of interest and determining the functional specifications and technology to be used. In terms of materials, we decided that manuscript collections and digital images would both benefit from public exposure and drive use. The collections would need to be *about* Chicago with Chicago being defined as the entire metropolitan area, including nearby counties in both Illinois and Indiana. Surveys, interviews and research helped identify the functionality required, after which an exhaustive analysis of potential technology was undertaken.

Among other functional requirements, it was determined early on that deposit of finding aids and images in a central database would be the most effective way to standardize and deliver content. In addition, several criteria were identified that created constraints on possible solutions. With institutions having hundreds of existing finding aids and thousands of image collections, no solution could require institutions to change their local standards or practice. No institution would be willing to maintain a separate version of metadata for the CCC, requiring cataloging of collections twice. Moreover, the consortium would not have a full-time programmer available to adjust portal code with every new format or member added. With these as fundamental requirements, the technology team ultimately selected XTF (eXtensible Text Framework) as the platform that would work best for the CCC.

XTF had been developed by the California Digital Library in 2002. It accommodates a wide range of formats in a multi-institutional setting and is tolerant of variations in metadata standards.

It had the advantage of not requiring high level programming skills to install, configure, and maintain. It could easily ingest and create indexes based on the full text of finding aids and image metadata records, while supporting separate "companion" metadata files that could include a controlled vocabulary, for example. But XTF did not have an administrative interface that would make it easy for archivists to deposit their content. The primary goal of the implementation grant from the Andrew W. Mellon Foundation was to build this interface, providing for the standardization of information from finding aids and mapping subject terms to a controlled vocabulary to be developed by the archivists.

Completing the project required the development of the public interface through which users would search the index and retrieve records. Decisions were made about what kinds of searches would be supported and how presented, how the search results would be displayed in standardized formats for finding aids and images, and what information about the holding locations would be helpful to users, among many other details. Additional funding was acquired to hire a branding and web development firm to design and finalize the details of the public interface for what was ultimately named Explore Chicago Collections (explore.chicagocollections.org).

Creating Browsable Metadata

While the intent of the portal was to provide keyword searching against an index of the complete finding aids and image metadata records, the identification of audiences and development of personas made it clear that there needed to be browsing options for easier access. Scholars were

definitely a primary audience identified early on, but so too were students, teachers, genealogists, lifelong learners, and the general public. To excite interest in Chicago history and increase use of our collections, it was imperative to make them easy to find and use.

During the planning phase of the project, the Portal Committee conducted an in-depth analysis of Chicago-focused finding aids and digital images from five institutions to help determine the variation in metadata across collections. Almost 1,300 EAD files and 1,900 MARC records for archival collections were gathered, along with metadata records for over 20,000 images from 15 different image collections. The metadata files yielded over 13,000 unique subject headings with almost half of the terms used only once by any institution. Only 1,335 led to collections at more than one library. The greatest number of libraries at which the same term was used was six and this happened only seven times. It was quickly clear that existing subject terms were not going to be effective in collocating access to our members' Chicago collections. A shared vocabulary would need to be created to enhance discovery, and its implementation could not mean updating existing records.

With this analysis by the Portal Committee, a Controlled Vocabulary Task Force was established to work on a solution. They reviewed existing vocabularies such as FAST (Faceted Application of Subject Terminology) and Library of Congress Subject Headings (LCSH) and found that a very large proportion of the original metadata matched these headings with some varying degree of accuracy. The Task Force concluded that with the high number of matches, they could use the FAST and LCSH headings to create trigger terms for matching to a smaller vocabulary they then set about to create. After several months of research, discussion, and experimentation, the Task

Force decided on a two-tier structure with six top-level facets (Events & Movements, Government & Leadership, Daily Life & Identity, Creativity & Thought, Natural & Built Environments, and Work) and 88 topic terms. All of the topics led to more than one item and 84 led to more than one library. Fifty-seven of the terms led to materials at 6 or more libraries with the term with the largest match leading to materials at 15 member institutions.

We did not want users to browse by topic alone, however. Chicago is famous for its neighborhoods, some with official designations, many without. The Task Force combined files from the Chicago Data Portal with information contained in the *Encyclopedia of Chicago* to create geographic boundaries and identifying terms for over 180 neighborhoods. For cities and towns, the team took data from the 2010 Census and matched names to the Library of Congress Name Authority File, Geonames, and the Getty Thesaurus of Geographic Names, creating additional documentation for those places that did not match an existing authority record. Over 340 cities and towns are now available to browse in the portal. For personal and organizational names, a large list of authorized headings was created using the original metadata. If a record is entered with a name that cannot be matched, the new name is then added to the authority list.

All of these steps succeeded in providing a much simplified, easy-to-apply set of controlled vocabulary to facilitate browsing the collections and digital images in Explore Chicago Collections.

Metadata Hopper

The administrative module that was developed to mediate between archivists, XTF, and the user interface, was named "Metadata Hopper." The Metadata Hopper was intended to allow archivists to create a set of rules that would tell the system how the metadata in the finding aid or MARC record or image file of a particular collection or collections would map to the agreed upon standard headings/sections in the portal public display. No one needed to change how they constructed their current finding aids.

In terms of the controlled vocabulary, when the project was first conceived, the notion was that archivists would need to apply the various facets manually. Each facet might be presented as a drop-down menu with the controlled terms listed for selection. Automatic tagging, while clearly desirable, seemed too ambitious for the initial development work. But as the Metadata Hopper continued to be built out and tested with the archivists, it became increasingly clear that the automatic tagging needed to be developed up front. The additional time spent to program this functionality in the beginning would be more than made up for in the time saved by archivists from numerous institutions who would not have to manually tag content later.

To facilitate this automatic tagging, the Task Force created a table of trigger terms based on their earlier mapping of the CCC's controlled vocabulary back to FAST and LCSH terms. The FAST and LCSH headings not only provide context for archivists, but also alternate headings that were then incorporated as additional trigger terms. When finding aids are deposited, subject terms are searched against this file of trigger terms and controlled headings are automatically suggested. Archivists have the opportunity to review the terms and can change or add tags prior to

'publishing' the record (i.e., making it publicly accessible). The Task Force has reviewed the controlled vocabulary twice since it was created and continues to refine the automatic tagging.

As I write, ExploreChicagoCollections provides access to over 5,100 archival collections and almost 107,000 digital images from 22 of our members. While practices at member institutions may not have changed, it does seem that many used the opportunity of the portal to enhance their metadata before depositing to include such things as neighborhood or geographic identifiers or to add the personal or organizational collection name to the subject field. For smaller members without large legacy collections, the standards used for display and the controlled vocabularies helped these institutions define local EAD and metadata practice. Since its launch in October 2015, the portal has been searched over 300,000 times by users from 172 countries and has just received a glowing review posted on *The American Archivist* Reviews site https://reviews.americanarchivist.org/2017/09/11/explore-chicago-collections/.

The CCC will be looking to expand the formats of materials included in the portal over the next few years. The foundation of collaboration and creative problem solving will serve us well as we continue to build a robust gateway to Chicago resources.

My sincere thanks to Kate Flynn (kef@uic.edu), Portal Manager for the CCC, and Tracy Seneca (tjseneca@uic.edu), Head of Digital Programs & Services at UIC, for their patient explanations and review of this manuscript. Any remaining errors or confusions are mine.

[i] The Explore Chicago Collections site includes a brief description of the portal's Technical Background <http://explore.chicagocollections.org/tech_background/>. The code for the Metadata Hopper and its documentation can be found on BitBucket https://bitbucket.org/uiclibrary/portal-admin/overview