Choosing Metadata for Discovery and Description in a Mixed-Content Single-Subject Library: A Case Study

Genevieve Podleski
Federal Reserve Bank of St Louis
One Federal Reserve Bank Plaza
St. Louis, MO  63166-0442
genevieve.m.podleski@stls.frb.org

ABSTRACT
This poster illustrates the decision-making and implementation testing processes for development of a consistent and coherent unified metadata plan for an established digital library that is both an institutional and a subject repository.

Keywords
Metadata, digital libraries, MODS, institutional repositories

INTRODUCTION
In 2004, the Federal Reserve Bank of St. Louis Research Library saw an opportunity to build a digital library focused on economic and banking history and created the Federal Reserve Archival System for Economic Research (FRASER). Beginning with historical economic statistical publications produced by both the federal government and the Federal Reserve Banks, FRASER’s digitization projects focused on granting researchers access to data and primary and secondary source documents illustrating historic economic conditions in the United States. As FRASER grew, we added monetary policy documents to provide further context, including congressional hearings, speeches and testimonies from Federal Reserve Chairman and Board members and developed deep collections on monetary policy history and early central banking in the United States. Through partnerships with other Federal Reserve Banks, we have digitized institutional publications not widely available to the public. In recent years, we have expanded beyond institutional documents and added significant archival collections relating to economic and banking history. Because of this collection development plan, FRASER has become both an institutional repository for the output of the Federal Reserve System and a subject repository for economic history. This mix of grey literature (United States government documents and the quasi-governmental output of the Federal Reserve), published books and periodicals, and archival material posed significant challenges to organization and discovery.

Beginning in early 2013, FRASER librarians began work on a coherent metadata plan that would reorganize the digital library under a single, comprehensive metadata and description schema. The reasons for this timing were threefold: a change in management of the digital projects team, a plan to expand awareness and use of FRASER in time for the Federal Reserve’s centennial, and a desire to integrate FRASER content into the Digital Public Library of America (DPLA), set to launch in April of 2013. Because there had been no consistent approach to metadata, a preliminary metadata audit such as that described by Westbrook et al. (2012) was not undertaken.

TECHNOLOGICAL LANDSCAPE
FRASER was entirely homegrown and initially used no metadata standards. Because each type of material had its own library or archival organizational conventions, the back-end of the FRASER website had been set up with discrete and sometimes conflicting metadata schemas. Special collections, publications, statistical releases (a special category of data publications put out by the Federal Reserve Board of Governors) and “historical” items each had their own separate table in the SQL server setup, despite similar presentation on the public FRASER site.

Figure 1. Existing database structure (reduced in size to show complexity rather than detail)
The volume of existing content and the limitations of IT support time made a complete overhaul of the underlying back-end of FRASER unworkable. At the end of 2012, FRASER had over 38,000 individual items with at least minimal metadata (title, date, author) and projects were already underway that would nearly double that count, to over 61,000 items by the end of 2013. It was necessary, therefore, to craft a metadata plan that could be grafted onto the existing database structure.

METHODOLOGIES

Following a brief, informal review of literature, we looked to digital libraries both in the U.S. and abroad for examples and case studies. Using the OpenDOAR index of open repositories (http://www.opendoar.org/), we identified a handful of established digital libraries with mixed content. The joint digital library of the Kansas State Historical Society (KSHS) and the State Library of Kansas (SLK), Kansas State Publications Archival Collection (KSPACe), was one such exemplar. (Note that http://kspace.org no longer seems to be available and may have been incorporated into the larger digital library Kansas Memory, www.kansasmemory.org.) Because KSPACe, like FRASER, included both historical and government documents, its use of modified Dublin Core seemed an ideal model. We also consulted the State Library of Massachusetts Archive (another Dublin Core repository) and the digital collections of the University of Michigan and the University of Utah. In considering the subject-based repository aspect of FRASER, we also examined the metadata standards for ArXiv, the e-print repository for physics, mathematics, and other quantitative research fields (http://arxiv.org/).

With an eye towards maintaining easy interoperability with standard library cataloging practices (FRASER items are cataloged locally and included in the Federal Reserve Bank of St. Louis’s listed holdings in WorldCat), we also consulted the Library of Congress Metadata for Digital Content standards (Library of Congress, 2012). In addition, we used the DPLA Metadata Application Profile (Digital Public Library of America, 2013) as a resource in order to better facilitate our database’s future compatibility with DPLA. Although the team was open to a variety of metadata standards, including the possibility of a homegrown standard, we decided, based on our preliminary research, to develop a draft metadata outline using modified Dublin Core. This decision was partially influenced by the standard’s current use in similar repositories, by DPLA, and the high number of repositories using stock or modified Dublin Core (Park & Tosaka, 2009: 105; Lopatin, 2010).

We created a metadata model specific to FRASER, using as models the DPLA Metadata Application Profile and the documentation of the Mountain West Digital Library’s implementation of Dublin Core (Mountain West Digital Library, 2011) for their digital collections. We then created a crosswalk of the existing FRASER metadata models to the Dublin Core schema and selected sample documents and publications from the FRASER collection as test cases. Members of the team independently applied the draft schema to Federal Reserve publications, published monographs, government documents including legislation, archival collections and individual items, and statistical data publications (Fig. 2, “Test”). As a group, we then read and reviewed the accuracy and completeness of each other’s draft applications of the metadata (Fig. 2, “Evaluate”) and revised both our applications and the schema (including usage notes) as necessary. Items were then re-described using the revised schema and the process was repeated.

After three rounds of testing, evaluation, and revision, we determined that the complexity of our documents and their organization required a more flexible approach than was provided by Dublin Core. The hierarchical nature of many of the FRASER items and the interrelatedness of others – particularly Congressional hearings and reports related to pieces of United States federal legislation – required a more robust schema that would allow for greater description of entity relationships. Furthermore, the high percentage of items with alternate titles, multiple contributing authors in addition to one or more main authors, and the conflation between creator and publisher for institutional documents began to strain the ability of Dublin Core to describe the FRASER collection without significant modification or extensive application notes. Because consistency in application and use of metadata standards is what gives them value, our inability to accurately, consistently, and completely describe our collection (Diao & Hernandez, 2014: 133), caused us to scrap our initial plan and again examine other common metadata standards for suitability.

The next approved standard on the DPLA ingest list (Digital Public Library of America, 2014) was MODS.
Created by the Library of Congress and based on MARC standards, MODS, (Metadata Object Description Schema) (Library of Congress, 2009) offered a more flexible method of representing and describing our collection, and its MARC-compatible structure allowed for greater automation of metadata updates for items already cataloged in OCLC. Taking these factors into consideration, MODS was chosen as the next candidate for FRASER metadata. A fresh metadata outline was created using MODS and incorporated concerns from earlier phases of drafting and testing and the testing procedure was carried out on the new schema. The testing and drafting process, complete with detailed crosswalks from the existing metadata content in FRASER, demonstrated the ability of MODS to meet our needs without significant deviations from standard practice.

Benefits of using MODS for this collection include a single, repeatable field (mods:name) that covers creator, publisher (for corporate documents), second author, etc.; a genre field (mods:genre) compatible with a Library of Congress genre authority list and a resource type field (mods:typeOfResource) equal to standard bibliographic types in MARC; and a repeatable, flexible RelatedItem (mods:relatedItem) field that allows for complex hierarchical relationships between items both inside and outside of the FRASER collection to be simply described. In addition, MODS' depth of descriptive fields, particularly the variations on the note field, allow us to migrate extensive and lengthy item notes to more specific and relevant fields. For example, mods:accessCondition can be used for copyright information, mods:abstract for content descriptions, and mods:originInfo for publication history. Unlike Dublin Core, MODS retains the general note field for both public and private general notes, similar to the MARC 500 field.

The new MODS-based metadata schema and a controlled vocabulary, both of which include documentation accessible to both FRASER librarians and clerks who do the majority of the day-to-day work on FRASER, are scheduled to be fully implemented in 2014. Preexisting records in the FRASER database are in the process of being retrofitted using a combination of OCLC ingest of MARC records to MODS via the MarcEdit software (for items where usable, complete records are available) and manual edits of item records. Future record creation will be done in a new back-end data entry portal that will map directly to the new schema.

CONCLUSION

Recent discussion in metadata for digital libraries has highlighted quality control and course-correction to how more closely to a metadata plan already in place (Palavitsinis et al., 2014; Westbrook et al., 2012), but has offered few case studies of retrofitting metadata to an existing collection. Our experience with FRASER highlights another challenge that will, as digital library projects age, become more and more prevalent.

Although FRASER’s content is unusual, the methods used to analyze, draft, and test metadata standards for legacy digital collections can be used by any digital library seeking to adapt or implement a more comprehensive metadata standard.

REFERENCES


