INVERTING THE LIBRARY CATALOGING PROCESS TO STREAMLINE TECHNICAL SERVICES AND SIGNIFICANTLY INCREASE DISCOVERABILITY AND SEARCH FOR SPECIAL COLLECTIONS

Marjorie M. K. Hlava
Access Innovations, Inc.
mhlava@accessinn.com

Judith C. Russell
George A. Smathers Libraries
University of Florida
jcrussell@ufl.edu

David "Win" Hansen
Access Innovations, Inc.
david_hansen@accessinn.com

Access Innovations, Inc.
Creator of the Data Harmony Software Suite

INTRODUCTION

Digitization of special collections saves space and increases access. But does digitization create a new kind of inaccessible media? OCR of non-searchable materials to generate searchable text is helpful, but the current results are usually not adequate to ensure access to all appropriate content and offer nothing more than free-text search. Using the Library of Congress Subject Headings (LCSH) and historical cataloging methodologies supports storage but limits access; traditional cataloging methods that involve human curation are expensive and slow.

This paper outlines a different approach. Maintaining MARC[1] records but significantly enhancing them with the addition of terms from the broadly based JSTOR thesaurus[2] and adding geographical location information for every record provides richer, more descriptive metadata. Using metadata and deep indexing instead of traditional cataloging records speeds the process, allows for automation, and increases searchability and discovery. Creating a metadata record and then exporting MARC records from the metadata record is an inversion of the traditional cataloging process (which entails creating MARC records first) and enhances discovery, speeds the entire process and, once established, reduces costs.

BACKGROUND

Digitization of special collections is a still new frontier for many organizations, particularly those with large stores of archival materials competing for space. How useful is capturing a digital image of a page or artifact? Is it just the new microfilm of the 21st century: painful to capture, hard to retrieve, and

[1] *The MARC formats are standards created by the Library of Congress for the representation and communication of bibliographic and related information in machine-readable form.*

[2] *The JSTOR thesaurus was used with special permission.*

even harder to view? How does one search for records? And how does one retrieve them from the digital abyss so that information can contribute to research?

In April 2016, Judy Russell, Dean of Libraries at the University of Florida, invited Marjorie Hlava, President of Access Innovations, Inc., to spend a few days analyzing the processes and methodologies used in the digital production and cataloging systems at the University of Florida. Of particular interest were processes for metadata capture and retrieval to support discovery and research for the university, especially within the special collections. The goal was to be able to surface the material in University collections and present a singular Portal of Florida History using the many special collections as well as the main library catalog. The special collections are currently siloed[3] in various systems and databases with inconsistent use of metadata fields and terminology, which does not facilitate distributed access to the library's treasures.

Each special collection of digital content provides different challenges, so the University of Florida investigated automated tools and defined processes that could be applied across the full spectrum. This was necessary because:

- The collections have been digitized over time for different purposes;
- Inconsistent scope, metadata standards, and vocabularies exist for each individual collection; and
- Multiple partners, both within the university and from external collaborators, have contributed to these inconsistent metadata standards and vocabularies.

[3] *An information silo is an insular management system in which one information system or subsystem is incapable of reciprocal operation with others that are, or should be, related. Thus, information is not adequately shared but rather remains sequestered within each system or subsystem. Such data silos are proving to be an obstacle for organizations wishing to use data mining to make productive use of their data.*

Florida History is one of the preeminent collections at the University of Florida, in both print and digital form, but the content is drawn from many different collections, including rare books, manuscripts, political papers, newspapers, maps, government documents, university archives, and more. The challenge is to identify all of the digital content, aggregate it, and present it as a coherent body of information: The Portal of Florida History.

Recent large-scale initiatives, such as the Portal of Florida History, focused attention on the need for significantly expanded and enhanced metadata for digital collections. Improved and consistent metadata practices needed to be defined and rigorously employed prescriptively as metadata for existing content is raised to the new standards. This required new tools as well as changing roles and responsibilities for cataloging and metadata staff.

The primary value of MARC records is as an inventory of print holdings and a means of identifying the availability and location of known items (a book by this author or with this title); MARC records provide minimal descriptive metadata but are heavily relied upon, especially for print collections. Primary subject access is via Library of Congress Subject Headings (LCSH), and Medical Subject Headings (MESH) are added for materials acquired for the Health Science Centre Libraries. Some MARC records are supplemented by licensed book jackets or tables of contents to improve retrieval precision.

OCLC and the Library of Congress understand the shortcomings of LCSH and have made considerable inroads with alternative methods, similar to those used by database publishers. With the rise of interest in special collections generated by both university-based researchers as well as outside parties such as educators, students, genealogists, historians and other non-academic researchers, perhaps it is time to reconsider traditional cataloging methods and seek a good foundation for a new approach.

To reframe the problem to be solved, the University of Florida can be thought of as a massive database publisher trying to use a traditional library approach to cataloging, metadata, and archival materials. A shift in thinking from *cataloger* to *database producer* could engender a massive change in thought process and workflow considerations.

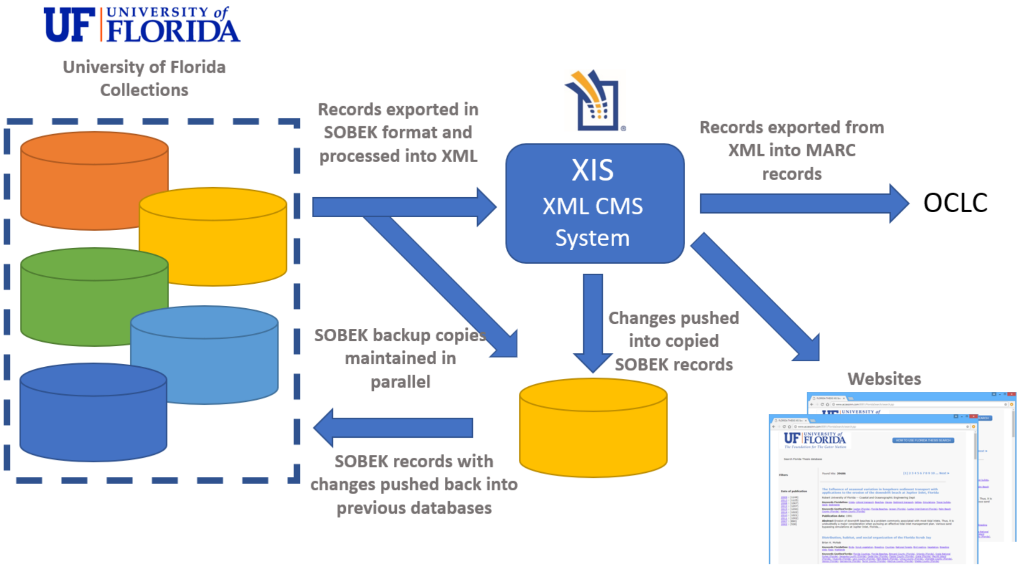
FLORIDA THESIS PROJECT

The digital preservation group within the University of Florida provided an excellent testbed for a pilot project, and digitized University of Florida theses and dissertations were selected as the content for the proof-of-concept. These records provided a significant volume of information across a wide spectrum of disciplines available in full text, so they represented the broad subject coverage the University of Florida desired to address. Over 25,000 items were available for the pilot to verify that size constraints were not an issue for the system.

Access Innovations developed a metadata schema for the project using its XIS® (XML Intranet System), an extended Dublin Core[4] application. Once the schema was tested and approved, Access Innovations launched a XIS® project to accommodate the data. Access Innovations extracted an additional set of "Florida-specific terms" to be used to identify candidate theses and dissertations for inclusion in the Portal of Florida History. This new taxonomy includes Florida place names, notable people, and other terms indicative of Floridian content. It was used for the theses and dissertations and will continue to be used to identify and tag records for the Florida history collection.

The existing Florida metadata was available in the MARC/METS XML format following standard library cataloging practice. That data was mined and the number of fields collapsed from the full MARC record approach to a Database Systems Approach using Dublin Core as a guideline.

[4] *The Dublin Core Schema is a small set of vocabulary terms that can be used to describe digital resources (video, images, web pages, etc.), as well as physical resources such as books or CDs, and objects like artworks.*


UF UNIVERSITY of FLORIDA
University of Florida Collections
Records exported in SOBEK format and processed into XML
XIS
XML CMS System
Records exported from XML into MARC records
OCLC
SOBEK backup copies maintained in parallel
Changes pushed into copied SOBEK records
Websites
SOBEK records with changes pushed back into previous databases

It was decided that people, places, and things would be best covered by a place-name thesaurus which was subsequently built, called Florida Geographic Terms, using the Access Innovations geography thesaurus (GeoThes) as a basis. Addition of a name authority list, now called Great Floridians, was constructed from names gathered from the current catalog and supplemented by entity extraction from the full text records themselves.

The JSTOR thesaurus, the Florida Geographic Terms, and the Great Floridian vocabularies were loaded as individual projects in the Data Harmony MAIstro® software for automatic tagging of the collection. The results from each list were put into a separate field in the database. A repository using XIS® was created to store and retrieve the records. Because the records were now consistent and deeply enriched with metadata, it was possible to export them as MARC records and ingest them into OCLC World Cat as well as to the Library Catalogue in Sobek^{CM}[5].

[5] *Sobek^{CM} is the internal content management system at the University of Florida. Sobek^{CM} is the software engine that powers both the University of Florida Digital Collections (UFDC) and the Digital Library of the Caribbean (dLOC) digital repositories. This software was developed at the George A. Smathers Library at the University of Florida by a team led by Mark Sullivan, with ongoing effort spanning several departments. Additional user testing, input, and resources have been contributed from other libraries, universities, and archives around Florida, the Caribbean, and beyond.*

The University of Florida now has a taxonomy of Florida-specific terms that it can maintain and expand and use to manage both print and digital collections. Use of this taxonomy alone as the default search option in the University of Florida Digital Collections (UFDC) provides 85% accuracy in retrieval. Simultaneous use of the taxonomy and full text search increases the precision of the search results to 91 – 92%.

RESULTS

The pilot project on the Florida Thesis data expanded the model for processing digital records and ingestion, successfully testing the indexing and the metadata model schema. User studies indicate significantly increased discovery and retrieval accuracy and all parties are encouraged by the quality and quantity of metadata that was created using these automated tools. The University of Florida is eager to apply these tools and processes to the rest of their digital collections.

CONCLUSION

Increasing emphasis on special digital collections and reducing the reliance on catalogers to create MARC records, while simultaneously increasing the investment in automated metadata creation (which can generate MARC records as long as needed for the catalog), is inverting the traditional cataloging process. At least at the University of Florida, MARC records will no longer be the original format used to generate most metadata. Instead, automated tools will be used to generate MARC records. They predict that within 10 years, or perhaps even sooner, “traditional” cataloging, as a title-by-title effort of applying metadata by trained catalogers, may require as little as 3% of the budget and only 4-5% of the library employees.