Using Automation to Generate More and Better Metadata for Special Collections and Archives:

Association of Southeastern Research Libraries
Professional Development Webinar Series
February 21, 2018

Judith C. Russell
Dean of University Libraries
University of Florida

Chelsea Dinsmore
Director of Digital Production Services
University of Florida
Challenges of Discovery

MARC records provide minimal descriptive and subject access and yet we rely on them heavily, especially for our print collections.

• The primary subject access is with the Library of Congress Subject Headings (LCSH), although Medical Subject Headings (MESH) are added for materials acquired for the Health Science Center Libraries.

• Some MARC records are supplemented by licensed book jackets or tables of contents to improve the precision of retrieval.

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).
The University of Florida Digital Collections (UFDC) hosts more than 300 outstanding digital collections, containing over 13 million pages, covering over 78 thousand subjects in rare books, manuscripts, antique maps, children’s literature, newspapers, theses and dissertations, data sets, photographs, oral histories, and more for permanent access and preservation. Through UFDC, users have free and Open Access to full unique and rare materials held by the University of Florida and partner institutions.

The UF Libraries encourage and support faculty collaboration on digital collections and digital scholarship.

UFDC is constantly growing with new resources, new scholarship, and system enhancements to the Open Source SobekCM Software. The search box above searches across all the digital resources in all the collections. By clicking on the icons below, you can view and search individual collections.

UFDC is available at: http://ufdc.ufl.edu/.
Challenges of Discovery

Recent large scale initiatives focused attention on the need for significantly expanded and enhanced metadata for our digital collections, both retrospective and prospective.

• Natural language full text searching provides better results than searching of MARC records, but UFDC includes many maps, photographs, architectural drawings, movie posters, etc., with limited text for searching.

• Application of a controlled vocabulary (but not LSCH) is necessary to organize sub-collections and enhance the precision of retrieval even when full text is available.
Unique Collections provide unique challenges, so UF sought to acquire automated tools and define processes that can be applied across the full spectrum of our collections. This is necessary because:

- The information has been digitized over time for different purposes;
- Individual curators have defined the scope of each collection and chosen metadata standards and vocabularies that supported the specific needs of each project; and
- Multiple partners both within the university and from external collaborations have also resulted in inconsistent metadata standards and vocabularies.

The size of these digital collections makes it impossible to revise and enhance these records without sophisticated automated tools or to aggregate content for important subcollections, like the Portal of Florida History.
Florida Theses & Dissertation Pilot Project
In 2016, UF began a pilot project with Access Innovations to **acquire automated tools** and **define processes** that can be used to identify and organize digital content for the **Portal of Florida History**, including the development and application of **enhanced metadata** using **controlled vocabulary**.
Florida Theses & Dissertation Pilot Project

Scope: Digital and digitized UF theses and dissertations – 29,000 records.

Objective: Use Machine Assisted Indexing (MAI) to apply enhanced metadata derived using controlled vocabulary to each record across the T&D collection and identify titles for which Florida is a subject.

Process: Access Innovations developed an extended Dublin Core metadata schema for the project using its XIS® (XML Intranet System).

Once the schema was tested and approved, Access Innovations launched an XIS® project to bring the data into. Then the MAI was run across the OCR generated text files and inserted the new terms into the METS files which were brought back into UFDC.
Florida Theses & Dissertation Pilot Project

The Access Innovations XIS® project included the following steps:

• Information was extracted from UFDC
• Three thesauri (NewsIndexer, NICEM and JSTOR) were tested for indexing purposes.
• JSTOR was chosen.
• Access Innovations extracted an additional set of “Florida-specific geographic terms” to be used to identify theses and dissertations about Florida locations for possible inclusion in the Portal of Florida History. This new taxonomy was used for the theses and dissertations and will continue to be used to identify and tag records with Florida-related content.
Florida Theses & Dissertation Pilot Project

University of Florida Digital Collections

XML Records & Full Text Exported from UFDC for Analysis

XIS XML CMS System

MARC Records Exported from XML

XIS Staff Review Panel

UF Theses & Dissertations

Enhanced Metadata Added to UFDC Records

XIS Repository of Updated Records

Updated Records Returned to UFDC
Example of Enhancements

Original Subject Terms

Example of Enhancements

Original Subject Terms:
- Gothic, horror, video
- Dissertations, Academic -- English -- UF
- English thesis, Ph. D
- bibliography (marcgt)
- theses (marcgt)
- non-fiction (marcgt)

Enhanced Subject Terms:
- Gothic, horror, video
- Dissertations, Academic -- English -- UF
- English thesis, Ph. D
- Video games (JSTOR)
- Horror fiction (JSTOR)
- Game theory (JSTOR)
- Family structure (JSTOR)
- Fairy tales (JSTOR)
- Visual materials (JSTOR)
- Femininity (JSTOR)
- Social structures (JSTOR)
- Computer technologies (JSTOR)

Example of Enhancements

Phase 2: Papers of Governor C. Ferris Bryant
Phase 2: Papers of Governor C. Ferris Bryant

Subjects

Bryant, Ferris, 1914- (LCSH)
United States. Office of Emergency Planning. (LCSH)
Florida. Board of Control (LCSH)
Florida Turnpike Authority (LCSH)
Florida. State Road Dept. (LCSH)
Marjorie Harris Carr Cross Florida Greenway (Fla.) (LCSH)
Politics and government -- 1951- -- Florida (LCSH)
Bryant, Ferris, 1914- -- Correspondence (LCSH)
United States. Congress. Senate -- Elections, 1970 (LCSH)
Segregation -- Florida -- St. Augustine (LCSH)
Political campaigns -- Florida (LCSH)
Elections -- Florida (LCSH)
Governors -- Florida -- 20th century (LCSH)
Example of Enhancements

Original Terms applied across folder/series

Enhanced Terms, also applied across folder/series
Lessons Learned

• Optical Character Recognition results are variable

• All collections are special – One size rarely fits all

• You still need experts
Looking Ahead
Digital Library of the Caribbean (dLOC)

More than 50 institutions digitize materials from their own collections and upload them to dLOC on a common platform, hosted by UF.

• With 202,068 items and 3,196,789 pages, dLOC is about 20% of the UFDC
• Multiple partners contribute digitized content with their own metadata schema and vocabularies.
• Content and metadata are available in multiple languages, including English, Spanish, French, Dutch, Creole, Papiamentu, and Hebrew.
• Reprocessing of the metadata with consistent use of fields and controlled vocabulary will greatly improve discovery and use of this material.
• Need to apply the automated tools and the techniques to the existing collections in dLOC and to apply those tools and techniques to new content as it is submitted for dLOC, including the Cuban Heritage Collections.
Planned Workflow for Digital Content

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

Changes pushed into copied SOBEK records

XIS Staff Review Panel

SOBEK backup copies maintained in parallel

SOBEK records with changes pushed back into previous databases
Application of XIS® to all UFDC Content

XIS® will become the metadata creation and subject indexing module for the entire UFDC content to identify and provide enhanced metadata for all UFDC content.

• Existing records will be extracted from UFDC to be “cleaned” and to perform the metadata enhancement and then reloaded into UFDC.

• New metadata records will be input with the XIS® Data Input Panel and then loaded into UFDC, and submitted to the UF Libraries Discovery Service and OPAC as well as OCLC.

• Records will be identified in UFDC as part of the Portal of Florida History via applied Florida Geographic terms.

• XIS® has the ability to batch correct large amounts of data in a single process. This is essential for retrospective record processing and intake of large new data sets.
Planned Florida Record Creation

All Records Created in XIS

- XIS Data Input Panel
- XMRC Records Exported from XIS to OCLC
- XML Records Exported from XIS to UFDC/dLOC
- UF Libraries OPAC/Discovery Service
- MARC Records Exported from XIS to the UF Libraries OPAC/Discovery Service
Thank you!

Judy Russell
Dean of University Libraries
George A. Smathers Libraries
jcrussell@ufl.edu

Chelsea Dinsmore
Director of Digital Production Services
George A. Smathers Libraries
chelseaz@ufl.edu

Marjorie Hlava
President
Access Innovations, Inc.
mhlava@accessinn.com