Inverting the Library Cataloguing Process to Streamline Technical Services and Significantly Increase Discoverability and Search for Special Collections

Marjorie M. K. Hlava  
Access Innovations, Inc., Albuquerque, New Mexico, USA.  
mhlava@accessinn.com

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

David “Win” Hansen  
Access Innovations, Inc. Albuquerque, New Mexico, USA.  
david_hansen@accessinn.com

Copyright © 2018 by Marjorie M.K. Hlava, Judith C. Russell, and David Hansen. This work is made available under the terms of the Creative Commons Attribution 4.0 International License: http://creativecommons.org/licenses/by/4.0

Abstract:  
Digitization of special collections saves space and increases access. But are we really just creating a new kind of inaccessible media instead? OCR is helpful and promising but the current results are usually not adequate to ensure access to all appropriate content. Using the Library of Congress Subject Headings (LCSH) and historical cataloguing methodologies supports storage but limits access. Traditional cataloguing methods are expensive and slow.

This paper outlines a different approach. Maintaining the MARC records but significantly enhancing them with the addition of terms from the broadly based JSTOR thesaurus and adding geographical location information for every record provides more touch points. Using metadata and deep indexing instead of traditional cataloguing records speeds the process, allows for automation, and increases searchability and discovery. Creating a metadata record and then exporting MARC records from that record is an inversion to the traditional cataloguing process and enhances discovery, speeds the entire process and, once established, reduces costs.

The following is a case study using three types of documents in the Special Collections of the George A. Smathers Libraries at the University of Florida (UF) as a proof of concept and then a pilot project of the inverted cataloguing process.

Keywords: Automatic Indexing, Taxonomy, Library Catalogues, Digital Archiving, Digital Preservation
Introduction
By acknowledging the decreasing reliance on the Online Public Access Catalogue (OPAC) for discovery, the libraries at the University of Florida are transforming access to library collections by re-envisioning the way they “catalogue” their collections. Digitization of archival materials is a still new frontier in many organizations, especially those with large archives competing for space. How useful is capturing a digital image of a page or artefact? Is it just the new microfilm of our century: painfully captured, hard to retrieve and harder to view? How do you search those records? How do you 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 digging into the processes and methodologies of the digital production and cataloguing systems at the University of Florida. Of particular interest were metadata capture and eventual 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 for Florida History from the many sources available within special collections and the main library catalogue. Those collections are currently siloed with inconsistent use of fields and terminology. This situation does not facilitate distributed access to the library’s treasures.

In addition to the OPAC, the University of Florida Digital Collections (UFDC) host more than 300 digital collections consisting of 545,000 items and over 13 million pages from the UF libraries and partner institutions.

Each collection of digital content provides differing challenges, so UF investigated automated tools and define processes that could be applied across the full spectrum. This was necessary because:
• The collections have been digitized over time for different purposes;
• Individual curators have defined the scope of each individual collection and chosen metadata standards and vocabularies that supported the specific needs of each project but are inconsistent among collections; and
• Multiple partners, both within the university and from external collaborators have resulted in inconsistent metadata standards and vocabularies.

Several major initiatives are driving these changes. Florida History is one of the preeminent collections at UF, both in 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, etc. 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, 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 perhaps not LCSH) is necessary to organize sub-collections and enhance the precision of retrieval even when full text is available.

Improved and consistent metadata practices needed to be defined and then 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 cataloguing and metadata staff.

MARC records provide minimal descriptive and subject access and yet are relied on heavily, 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.

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 Libraries vision statement calls for providing service at the point of need and their digital collections are essential to that. As a Land Grant University, they have agricultural extension offices in every county in Florida. They also have numerous research centres and other facilities located outside of Gainesville. Most on campus students take one or more online courses per semester and remote online degree programs are growing, so service to users who are not physically in the Libraries is very important – even though student use of the physical libraries is high and growing\(^1\). (1)

---

\(^1\) Internal study at the University of Florida Gainesville on Library usage.
OCLC and the Library of Congress understand the shortcomings of LCSH and have made considerable inroads in alternative methods, similar to what database publishers (a.k.a. Abstracting and Indexing (A&I) services) are using. With the rise of importance in the Special Collections and interest in them by outside parties such as educators, students, genealogists, historians and other non-academic researchers, perhaps it is time to consider the traditional cataloguing methods a good foundation for a new approach. In truth, the University of Florida is a massive database publisher trying to use a traditional library approach. A shift in thinking from cataloguer to database producer would cause a massive change in thought process and workflow considerations.

**Background**
The University of Florida digital collections cross the boundaries between online databases represented by the digital collection, and traditional library systems (OPAC), represented by the Ex Libris Aleph system. Improved retrieval, especially for the digital collections, requires significant metadata enhancement and consistent vocabulary applications for enhanced retrieval and discovery both on the SobekCM (2) and the Aleph systems. This paper concentrates on SobekCM support for the digital library collections. The existing metadata is clean and well formed by library cataloguing standards; however, it is insufficient to support modern search needs.

Current discovery indicated a few concrete steps that could be taken to enhance the metadata. University of Florida runs a dual facility. One is a traditional cataloguing unit suitable for a traditional and large University Library (and all the appropriate accoutrements that go with it). The second is much more akin to a secondary publishing or aggregation system. The second unit has the needs of an A&I service, but only partially, as it still has a base in historic cataloguing. This makes it difficult to determine a suitable storage, search, and distribution platform: one which is not quite an archive service and not quite an A&I service, but with the budding requirements for such.

The mission of the Digital Processing Facility (DPS) is generally to acquire, digitize, archive, and disseminate the Florida Historic record through direct capture and linking to other authoritative resources. The tools in use for the DPS are different than those of the cataloguing unit. The outputs are different as well: the DPS outputs to XML and the cataloguing unit outputs to MARC. The metadata collected by the two systems is also different with each system prioritizing different information.
After examining all the fields used in the SobekCM system, the DPS, fields filled locally, and those imported from OCLC, a map was developed. That map provided a crosswalk for eventual contribution to the metadata for each collection and supporting information for metadata enrichment, especially in the SobekCM hosted data.

The second step for the metadata was to find what data was missing (yet to be enriched) to the point that enough information would be included to support a satisfactory search. The fields that contribute to better search and discovery [classification, abstract, keywords, titles, names, topical, ID, ETD, date, and geospatial] were assessed by collection.

With the assessments in hand, each data set or collection would have a plan developed for enrichment. This was especially true for semantic items (subject, topical taxonomy, and concepts), places, personal names, and organization names. Most of these were sparse in the existing metadata although they are the most likely items for search queries.

To adequately cover Florida history and heritage for search and retrieval, metadata tool sets were needed, including:

- a Florida Heritage Taxonomy: a full-fledged thesaurus with broader terms, narrower terms, related terms, and synonyms, plus editorial notes scope notes, and definitions by term;
- an authority file of Florida place names with accompanying date of use and geospatial coordinates;
- an authority file of people names with dates, location, and role;
- an authority file of organization names, with links to the people file, including place, dates, role or type; and
- an authority file for events with name, synonyms, dates, and places.

Many of these lists already existed, and subject metadata resources were also available. Four were tested:

- JSTOR thesaurus
- NICEM thesaurus
- NewsIndexer thesaurus

---

SobekCM is the internal content management system at the University of Florida, named for the Egyptian god with the crocodile head since the mascot of the University of Florida is the “Gator”. SobekCM is the software engine which powers both the University of Florida Digital Collections (UFDC) and the Digital Library of the Caribbean (dLOC) digital repositories. SobekCM allows users to discover online resources via semantic and full-text searches, as well as a variety of different browse mechanisms. For each digital resource in the repository there are a plethora of display options, which may be selected by an appropriately authenticated use. This repository includes online metadata editing and online submissions in support of institutional repositories. This software was developed at the George A. Smathers Library at the University of Florida by a team lead 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. SobekCM is released as open source software under the GNU GPL license and can be downloaded from the SobekCM Software Download Site.
Library of Congress Subject Headings

These were tested on 25,000 theses from the UF Collection by programmatically indexing the records and analysing the results based on accuracy statistics. It quickly became clear that the JSTOR Thesaurus, with local Florida enhancements, would be the best fit.

Design of the Proof of Concept
There were three major components to the pilot project:

- Create a metadata map or schema for the full information in the digital library digital records;
- Identify the current collection metadata coverage: assess all digital items at the individual accession or item number level and determine what metadata is currently applied and what metadata is missing. This was facilitated by creating a metadata map between the MARC record cataloging from OCLC plus original cataloging done at UF and comparing it to the metadata records for the digital library held on the SobekCM system; and
- Gather supplemental information depending on the format type and source of the original records for individual collections.

A plan for gathering the metadata by subset was needed and options for acquiring missing metadata included:

- Checking the metadata records on the OCLC or the original UF cataloging for additional information;
- Adding information from the individual finding aids for the special collections;
- Data mining for people, places, and things. This was done either against authority files from places such as the state library, if available, and from geographic location lists from maps and gazetteers of Florida and other sources;
- Translation, reformatting, and addition of dates for searching based on periods (antebellum, Civil War, etc.) and supplemental information from the several places in the MARC record where such data is recorded, and moving it to a single consistently formatted date field in the digital library record;
- Gathering all date information available from the MARC records and translating it to a standard yyyyymmdd format to significantly improve date range searching; and
- Further creation of a synonym list for date ranges and period descriptions to aid in discoverability. That is, 1860, 1861,1862,1863,1864, and 1865 could each be a synonym as dates for the US civil war.

It was determined that a team of at least two people would take on the task of processing the data. One would do the assessment and add the metadata as appropriate. The second person needed some programming skills so that they could process in batches, adding metadata and making the records ready for quality control (QC). Then that data was added to the records (if they were already present) in the collection on the SobekCM host. It was expected that each of the smaller collections would take a week to process and the larger collections would take longer. Newspapers were a special case as they have both broad and diverse coverage, whereas
the collections are mostly narrowly defined. Both collections abound with names of people and places.

Doc ID
Call number
Author (including <roleterms> in the XML)
Author birth and death dates
Author location
Title
Publisher
Place of publication
Date of publication
Size
Subject heading
  Topic
  Geographic
Document type
Holdings statement
JSTOR Term
LSCH Term
News indexer Term
Entity
SuDocs numbers
Catalog number in Ex Libris Aleph
Catalog number in OCLC
Rights

Diagram 1 - Field list for Florida Thesis and History Portal

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
<mods:name type="personal">
<mods:namePart>Rackham, Arthur</mods:namePart>
<mods:namePart type="given">Arthur</mods:namePart>
<mods:namePart type="family">Rackham</mods:namePart>
<mods:namePart type="date">1867-1939</mods:namePart>
<mods:role>
<mods:roleTerm type="code" authority="marcrelator">ill</mods:roleTerm>
<mods:roleTerm type="text">Illustrator</mods:roleTerm>

Diagram 2 - Sample Author record

Florida Thesis Project
The digital preservation group within UF provided an excellent test bed for the pilot project. Digital UF theses and dissertations were selected as the content for the project. These records provided a significant volume of information across a wide spectrum of disciplines available in full text so they represented the broad coverage area UF 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 application. Once the schema was tested and approved, Access Innovations launched an XIS® project to accommodate the data.

The Access Innovations XIS® project included the following steps:

- Information was extracted from UFDC, including the full text and the existing metadata.
- Three thesauri (NewsIndexer, NICEM and JSTOR) were selected and tested for indexing purposes, as well as the LCSH as the existing subject metadata set.
- Indexing tests were run to determine which thesaurus would be preferred, and JSTOR was chosen.
- 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 cataloguing 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. This enabled moving from a potential list of over 650 fields to a list of searchable fields and supporting metadata of 23 fields.
- The next step was deciding on a set of resources to use in tagging the records with topical, location and named entities. After some experimentation with several options, it was decided to use the JSTOR Thesaurus and an average of 7 topical terms (instead of the LC MARC usual 3).
- People, Places and Things would be best covered by a place names thesaurus which was then built, called Florida Geographic Terms, using the Access Innovations GeoThes as a basis. It is now registered as a recognized Library of Congress authority file for place names.
• Addition of a name authority list, now called Great Floridians, built with names gathered from the current catalogue 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 the Data Harmony XIS® (XML Intranet System) was created to store and retrieve the records. Because the records were now in a consistent and deep record, it was possible to export them as MARC records and ingest them into OCLC World Cat as well as to the Library Catalogue in SobekCM.

Figure 3: Comparative sample records for the Pilot project providing additional metadata from the JSTOR and GeoThes Thesauri. These terms are automatically generated based on the full text OCR of the original records.

The first two and last four lines remain the same in both records, but the automated process added 12 additional controlled vocabulary terms, several of which clearly identified this as a thesis about specific geographic areas in Florida and therefore appropriate for inclusion in the Portal of Florida History. This provides the user several more points for retrieval to the data in the repository.
UF 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 as the default search option in UFDC provides 85% accuracy in retrieval. Simultaneous use of the taxonomy and full text search increases the precision of the search results over that by 6-7% because of the extraneous items that are retrieved.

**Proving the Model**

To further test that the model works properly it was applied it to a very different collection: the Cuban Heritage collection of the Catálogo de la Biblioteca Nacional "José Martí" de Cuba (BNJM), also known as the Cuban National Library. It is a member of the Digital Library of the Caribbean (dLOC) which is composed of 225 organizations. More than 50 of those institutions digitize materials from their own collections and upload them to dLOC on a common platform, hosted by UF:

- 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.

To keep up with the inflow of this data, it was necessary to apply the automated tools and the techniques to the newly 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.
BNJM estimates that 58% of its Cuban Heritage materials are uniquely held in Cuba. Digitization of those materials will be their responsibility and challenge. UF and its partners outside of Cuba will seek to identify sources for and digitize the other 42%. Bibliographic control is essential to avoid duplication of effort and make the collection as comprehensive as possible. BNJM has shared is bibliographic records with UF and they have agreed to establish an OCLC symbol for BNJM, managed by UF, and make sure all of their Cuban Heritage records are available in WorldCat. They will also use the metadata in the project management database; however, ~16,000 of the BNJM records are not yet in MARC format. BNJM shared ~16,000 scanned, physical catalogue cards for conversion into MARC records for OCLC and into metadata for the project management database.

The Cuban OCR was excellently captured and accurate; however, it was not fielded and some of the data, captured line-by-line in the OCR, was mashed together. In order to take the raw OCR files and make them into metadata records for inclusion in the UF repositories, it needed to be separated into XML fields to load to the database. Although many library catalogues would add various date data into a single field, for search in the new repositories it was necessary to place the publisher name, location, and date information into separate fields in order to use them in search. Once separated into component parts, XML records are deposited into XIS® for accurate and detailed search by field and exported to OCLC as MARC records. Without this step, the data is a full text block of undifferentiated material and does not provide reliable search results.
16,000 of the Cuban catalogue records were deemed to be fully unique and not in WorldCat. These records were programatically separated into fields and then indexed using the JSTOR thesaurus. An automatically translated Spanish version of JSTOR was also used to ensure broader and deeper indexing. The records were then loaded into XIS® repository and available for search on the platform.

Results
The pilot project on the Florida Thesis data and the Cuban test case expanded the model for processing digital records and ingestion, successfully testing the indexing and the metadata model schema. Work is now underway to fully process the Cattleman’s Journal, including testing how to gather articles spread on multiple pages to a single record. The University of Florida’s Bryant Collection was used as another test case for the ingestion of scanned personal papers of a Great Floridian. The 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. UF is eager to apply these tools and processes to the rest of their digital collections.

Next steps
Even as UF is working with XIS® to improve the metadata records for its digital collections, it is preparing to transition from doing primary cataloguing in OCLC to XIS® for all UF cataloguing and metadata creation then exporting the appropriately formatted records in MARC, XML, and other ingestion formats as needed to support all UF collections – print and electronic:

- Records will be created in the XIS® Data Input Panel, which will prompt for the correct placement and use of thesaurus terms.
- Records will be exported from XIS® to OCLC, the UF Libraries Discovery Service and OPAC, and, when appropriate, to UFDC and dLOC.
- Direct submission by users of The Institutional Repository at the University of Florida (IR@UF), the digital archive for the intellectual output of the University of Florida community, and includes research, news, outreach, and educational materials.
IR@UF items will be processed through XIS® to provide consistent metadata, including use of thesaurus terms, ensuring that the records in UFDC and the OPAC/Discovery Service are consistent and result in submission of more complete records to OCLC.

Figure 7: Planned record creation for all UF Records – create in XIS® as XML then export to all the appropriate consuming applications.

Figure 7 shows the revised workflow for title-by-title cataloguing. Cataloguers use the XIS® Data Input Panel, which will prompt for the appropriate thesaurus terms. Once the XML files are in the repository, records are output in the format needed by the recipient services the consuming applications. This streamlined process supports as many exports as desired – there are four at present – to consuming applications. It covers the cataloguing process from any source whether digital or print.

Conclusion
Increasing emphasis on digital collections and reducing the reliance on cataloguers to create MARC records, while simultaneously increasing the investment in automated metadata creation (which can generate MARC records as long as needed for the catalogue), is inverting the traditional cataloguing 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” cataloguing, as a title-by-title effort of applying metadata by trained cataloguers, may require as little as 3% of the budget and only 4 or 5% of the library employees.