University Metadata and Retrieval: The Death of the Library Catalog?

Judith C. Russell
Dean of University Libraries

Copenhagen
October 14, 2016

DCMI: Dublin Core Metadata Initiative
MISSION: The University of Florida is a large, comprehensive research university. We are a diverse community dedicated to excellence in education and research and shaping a better future for Florida, the nation and the world. Our mission is to enable our students to lead and influence the next generation and beyond for economic, cultural and societal benefit.

With an enrollment of over 50,000 students annually, UF is home to 16 colleges and more than 150 research centers and institutes. The colleges are:

- Agricultural & Life Sciences
- Dentistry
- Engineering
- Law
- Nursing
- The Arts
- Design Construction & Planning
- Health & Human Performance
- Liberal Arts & Sciences
- Pharmacy
- Veterinary Medicine
- Business
- Education
- Journalism
- Medicine
- Public Health

Over 31% of UF students are in graduate and professional degree programs. UF awards over 800 Ph.D., 1,198 Professional, and 4,000 Master’s degrees annually.
The annual budget of the University is over **US $3.89 Billion**, with over **US $700 Million (18%)** from externally funded research.

In 2013, among U.S. universities, UF ranked:
- **6th** in the number of start-up companies created with 16;
- **7th** in technology licenses with 140; and
- **11th** in the number of U.S. patent applications with 107.

**Gatorade®,** the world’s most popular sports drink, is just one of hundreds of commercial products resulting from UF research.

UF operates two major research hospitals, one with embedded cancer and children’s hospitals, a psychiatric hospital, a rehabilitation hospital, and an animal (veterinary) hospital.

UF faculty publish over 7,500 highly cited scholarly articles per year in a wide range of disciplines and a diverse set of journals.
LIBRARY MISSION: The Smathers Libraries partner with UF faculty, students and staff, as well as the University’s collaborators and constituents, to facilitate knowledge creation that contributes to UF’s standing as a preeminent public research university. The Libraries encourage creativity and inquiry necessary to support the University’s global ambitions and play an important role in attracting and retaining top students, faculty and staff.

LIBRARY VISION: The Libraries ignite curiosity, serve as the locus of knowledge management, and promote intellectual exchange within our diverse global learning community.

To accomplish its mission and vision, the Smathers Libraries will:
• Offer key services at the point of need to meet the requirements of the University enterprise;
• Initiate and participate in collaboration and community building; and
• Assure effective, efficient and equitable access to pertinent information resources for all library users.

The Strategic Directions for the Smathers Libraries are available at: http://ufdc.ufl.edu/IR00004144/00002.
The UF Libraries consist of seven libraries on the Gainesville campus and three other off-campus facilities.

The UF Libraries hold or provide access to large print and microfilm collections, including over:
• 5.45 million print volumes and 1.25 million e-books;
• 1.35 million maps and images and 1.26 million documents; and
• 8.26 million microfilms.
  Most of these items are located through the 4.8 million local MARC records in the Online Public Access Catalog (OPAC).

The UF Libraries also hold or provide access to large digital collections, including over:
• 152,000 current full-text print or electronic journals and newspapers;
• 1,100 electronic databases; and
• 11 million digital files in the UF Digital Collections (UFDC).
  Most are located through full text search of UFDC, vendor platforms and/or a commercial discovery tool (Summon). The discovery tool also incorporates the MARC records from the OPAC.
Challenge of Discovery: MARC records provide minimal descriptive and subject access and yet we rely on them heavily, especially for our print content.

- 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 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).

Even with “copy cataloging” and acquisition of MARC records with shelf-ready print books and some licensed electronic content, Cataloging remains our 3rd largest department, utilizing more than 10% of our salary budget, so we continue to make an enormous investment in a traditional method of identification which requires title by title effort by trained catalogers.

Historically, when we digitized materials from our print collections, the MARC record was the initial, and often the only, descriptive metadata and the subject headings from the MARC record were often the only “controlled vocabulary” assigned to the digital item.
Digital Content is increasingly important to allow for access anytime and anywhere. It serves all students, faculty and staff, including those with disabilities and those relying on remote access.

• When available, electronic resources are preferred over print for most acquisitions, including monographs.
• 87% of the materials budget is allocated to electronic resources, resulting in over 5 million downloads of licensed content by library patrons last year.
• Over 50% of our materials budget is used for collaborative acquisitions to minimize unnecessary duplication and reduce costs.
• Most monographs, whether print or electronic, are selected through Patron Driven Acquisition programs, resulting in less purchasing for future researchers and more focus on the needs of current users.
• The Libraries are active participants in HathiTrust, Internet Archive and other digital initiatives to expand access to high value electronic content for UF students, faculty, researchers and clinicians.
The Smathers Libraries invest heavily in our own digital and digitized content and development of a robust and expansive digital collections infrastructure, including the open source platform, SobekCM.

The University of Florida Digital Collections (UFDC) host more than 300 digital collections, consisting of 500,000 items and over 11 million pages. The digital collections provide permanent access and preservation for over 78 thousand subjects in rare books, manuscripts, antique maps, children's literature, newspapers, theses and dissertations, government documents, data sets, photographs, oral histories, and more.

- Over 1.1 million pages are added each year.
- Most UFDC content is available for public access.

The Institutional Repository (IR@UF), which is part of UFDC, offers a central location for the collection, preservation, and dissemination of scholarly, research, and creative production from UF authors and UF colleges. University theses and dissertations are preserved and made accessible in the IR@UF through digital deposit and retrospective digitization.
UFDC provides preservation and access to over 11 million pages of digital content from the UF Libraries and partner institutions. These unique materials are viewed over 85 million times per year, providing access to over 7.78 million users world-wide. UFDC is available at: http://ufdc.ufl.edu/.
Challenge 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 controlled vocabulary (but not LSCH) is necessary to organize sub-collections and enhance the precision of retrieval even when full text is available.
- UF digital collections are so large, and growing so quickly, that automated processes are essential.
- Improved and consistent metadata practices must be defined and then rigorously employed prospectively and metadata for existing content must be brought up to the new standards.

This requires new tools and changing roles and responsibilities for cataloging/metadata staff.
Unique Collections provide unique challenges, but UF is seeking to acquire automated tools and define processes that can be applied across the full spectrum of our collections – and facilitate our response to new uses our content and service to additional/different users.

• Genealogy – 18 months ago we began a partnership with Family Search and Internet Archive to identify and digitize genealogical content in our collections. This requires presenting the content in a manner that supports genealogical research, a new lens with which to examine content acquired for different purposes.

• Portal of Florida History – 24 months ago we began a project to identify, aggregate and organize all of our digital content that supported research on Florida history, broadly defined. We believe that a substantial portion of the content in UFDC, perhaps as much as 20-25%, could be within the scope of the Portal of Florida History, but the available metadata does not allow for easy identification, aggregation and presentation of that content or support effective search.
  • 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.
A Pilot Project with Access Innovations is helping us acquire automated tools and define processes that can be used to identify and organize content for the Portal of Florida History, including the development and application of enhanced metadata using controlled vocabulary.
• All digital and digitized UF theses and dissertations were selected as the content for the project. The objective is to identify the ones for which Florida is the subject matter and apply enhanced metadata derived using controlled vocabulary to each one.
• Access Innovations has developed a metadata schema for the project using its XIS® (XML Intranet System). It is an extended Dublin Core application.
• Once the schema was tested and approved, Access Innovations launched an XIS® project to accommodate the data.

NOTE: SobekCM supports 11 subject metadata fields, which have NOT been used consistently over time. Standardizing the use of the fields is essential to the development and application of enhanced metadata and to support advanced search focused on one or more fields.
• All topical subject terms will be assigned to a single field.
• Geographic terms, place names, personal names, and corporate names will each be assigned to a single field.
• Existing terms will be mapped to the appropriate fields and/or replaced with controlled vocabulary terms in the appropriate fields.
• Three thesauri (NewsIndexer, NICEM and JSTOR) were selected.
• Access Innovations extracted information from the existing metadata for the UF theses and dissertations to establish a rule base for the XIS® project.
• Using the thesauri, the rule base and the full text from the UF theses and dissertations, Access Innovations ran a batch process to add the selected subject metadata to the appropriate fields in the record for each item.
• UF is currently reviewing the subject metadata to confirm the selection of the thesauri.
• Once the thesauri are confirmed, Access Innovations will extract a set of “Florida-specific terms” to be used to identify candidate theses and dissertations for inclusion in the Portal of Florida History. This taxonomy will include Florida place names, notable people and other terms indicative of Floridian content. It will be used to identify and tag records for the Florida History collection.

**NOTE:** Use of this taxonomy as the default search option will provide 85% accuracy in retrieval. Simultaneous use of the taxonomy and full text search reduces the precision of the search results to 6-7% because of the extraneous items that are retrieved. This is the value of controlled vocabulary indexing.
• Both XIS® and SobekCM use Lucene Solr for search.
• Initial testing and evaluation of the search results will be done in XIS®.
• Once the collection is satisfactory, the metadata and connections to the Lucene index in UFDC will be added through the regular SobekCM load process and reassessed.
• Adjustments will be made, if necessary, and the Pilot Project will be concluded.

Application to the remaining UFDC content:
• Once the Pilot Project is concluded, the Data Harmony (DH) software from Access Innovations will be linked to SobekCM via an API or web service.
• DH will be used with XIS® as a content creation and subject indexing module for the entire UFDC content to identify and provide enhanced metadata for all Florida History content.
• 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. Where needed, records will be extracted from UFDC to be “cleaned” and to perform the appropriate metadata enhancement and then reloaded to the Portal of Florida History in UFDC for search.
Unique Collections provide unique challenges, and large scale collaborations create special challenges. We are engaged in two extremely large and valuable collaborations that will benefit from expanded and enhanced metadata and the application of controlled vocabulary using the automated tools being developed for the Portal of Florida History.

- **Cuba** – In June the Smathers Libraries signed an agreement with the Biblioteca Nacional de Cuba José Martí (BNJM) to create broad and deep open access to digital collections of Cuban Heritage materials. We will coordinate partnerships with U.S. and Canadian institutions for digitization of maps, legal materials, and monographs, as well as journals and newspapers, in consultation and cooperation with the BNJM. The BNJM will serve as advisor to each project and contribute materials not held elsewhere. Once digitized, these sources will be available to all participating libraries and to users worldwide interested in studying and researching the history, literature and culture of Cuba. Digital materials will be contributed by the BNJM to the Digital Library of the Caribbean.
  - Multiple partners will contribute digitized content with their own metadata schema and vocabularies.
  - Content will be in Spanish and English and perhaps other languages.
  - Reprocessing of the metadata with consistent use of fields and controlled vocabulary will greatly improve discovery and use of this material.
• Digital Library of the Caribbean (dLOC) – Over 10 years ago, Florida International University and the University of Florida received a federal grant to establish dLOC. More than 43 institutions digitize materials from their own collections and create their own metadata and upload both to a common platform, hosted by UF. Each institution determines the content and metadata it will provide, although there is an initiative to encourage digitization of Caribbean newspapers.
  • 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.

We intend to apply the automated tools and the techniques learned from the Portal of Florida History project with Access Innovations 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 initiatives.
dLOC provides preservation and access to over 3.1 million pages of digital content from the 43 participating institutions. These unique materials were already viewed over 23.5 million times this year, providing access to over 6.1 million users world-wide. 
dLOC is available at: http://dloc.com/.
The Death of the Library Catalog?

• It remains the best tool we have to inventory our print holdings and identify the availability and location of known items and we don’t yet have something better to replace it.

• The service bureau for the 28 public colleges and 12 public universities in Florida is investing several million dollars and untold staff hours to install a “Next Generation ILS” that provides electronic resource management and other services, but is still primarily a Library Catalog utilizing MARC records.

• While there are automated tools for digital content, cataloging (especially for unique and relatively unique materials in our special collections) remains an item by item process.

• Large scale collections of digitized books, like HathiTrust and Internet Archive, provide the promise of full text that can be matched to our print collections and used to create enhanced metadata using automated processes.

The Death of the Library Catalog? Not yet, but we are placing increasing emphasis on digital collections and reducing our investment in the creation of MARC records while increasing our investment in enhanced metadata (which can generate MARC records as long as we need them for the Catalog). I predict that within 10 years (perhaps sooner) “cataloging” may be 5% of our budget for the Cataloging & Metadata Department.
Thank you!

The Smathers Libraries seek partners for collaboration, particularly in digital initiatives. We welcome visiting scholars who wish to do research in our collections.

Judy Russell  
Dean of University Libraries  
George A. Smathers Libraries

jcrussell@ufl.edu

UF Students, Faculty and Alumni are known as the Florida Gators!