

University Metadata and Retrieval:

The Death of the Library Catalog?

Nineteenth International Conference on Grey Literature
Rome, Italy
October 23, 2017

Judith C. Russell
Dean of University Libraries
University of Florida

Marjorie M.K. Hlava
President
Access Innovations, Inc.

University of Florida

MISSION: *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 53,000 students, UF is home to 16 colleges and more than 150 research centers and institutes. UF operates two major research hospitals, a psychiatric hospital, a rehabilitation hospital, and an animal (veterinary) hospital.

Over 31% of UF students are in graduate and professional degree programs. UF awards over 800 Ph.D., 1,200 Professional, and 4,000 Master's degrees annually

University of Florida

The annual budget of the University is over US \$2.87 Billion, with over US \$724 Million (25%) from externally funded research.

In 2015, among U.S. universities, UF ranked:

- 10th in the number of start-up companies created with 15;
- 3rd in technology licenses with 261; and
- 10th in the number of U.S. patent applications issued with 118.

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



UF faculty publish approximately 8,000 highly cited scholarly articles per year in a wide range of disciplines and a diverse set of journals.

University of Florida Libraries

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.*

University of Florida Libraries

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>.

University of Florida Libraries

The Smathers Libraries consist of six 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 microforms.

Most of these items are located through the 4.8 million local MARC records in the Online Public Access Catalog (OPAC).

University of Florida Libraries

The UF Libraries also hold or provide access to large digital collections, including over:

- 175,000 current full-text print or electronic journals and newspapers;
- 1,100 electronic databases; and
- 13 million digital files in the UF Digital Collections (UFDC), many of which are not in the OPAC.

Most are located through full text search of UFDC, vendor platforms and/or a commercial discovery tool that provides access to the full text of these electronic resources as well as the MARC records from the OPAC.

Digital Content

Digital Content is increasingly important to allow for access anytime and anywhere.

- 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 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.

Challenge 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).

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 a controlled vocabulary (but not LSCH) is necessary to organize sub-collections and enhance the precision of retrieval even when full text is available.



UFDC
University of Florida
Digital Collections

Search Collection:



UFDC HOME
ADVANCED SEARCH
TEXT SEARCH
BROWSE PARTNERS

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.

Digital Content Through UFDC

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, Sobek^{CM}.

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.



Digital Content Through UFDC

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](#).

Access Innovations, Inc.

UF turned to Access Innovations, Inc. for assistance because of its unique expertise.

- It is a clever group of metadata and semantic analysts, based in Albuquerque, New Mexico, USA.
- The creator of the Data Harmony Software products for metadata capture, creation and enrichment.

Its Mission: *To add intelligence to data, increase the value of content, change search to found, and increase the accessibility and value of an organizations knowledge base by harmonizing the source data, as programmatically as possible and adding subject metadata, and allow repurposing of data.*



AMERICAN
PSYCHOLOGICAL
ASSOCIATION

ACR *American Association
for Cancer Research*

AMAX
AMERICAN
MEDICAL
ASSOCIATION



SPIE

ASCE
American Society of Civil Engineers

DU PONT



ASCO
American Society of Clinical Oncology

PLOS

AAAAS

ACS
Chemistry for Life®

Cargill™



Mc
Graw
Hill
Education

ASUG

ConsumersUnion®

IOP
Institute of Physics

ANSI
American National Standards Institute



AIP
American Institute
of Physics

PROJECT
HARVEY
FOUNDED IN 1958



P&G

Innovations Inc.®
simony taxonomy software

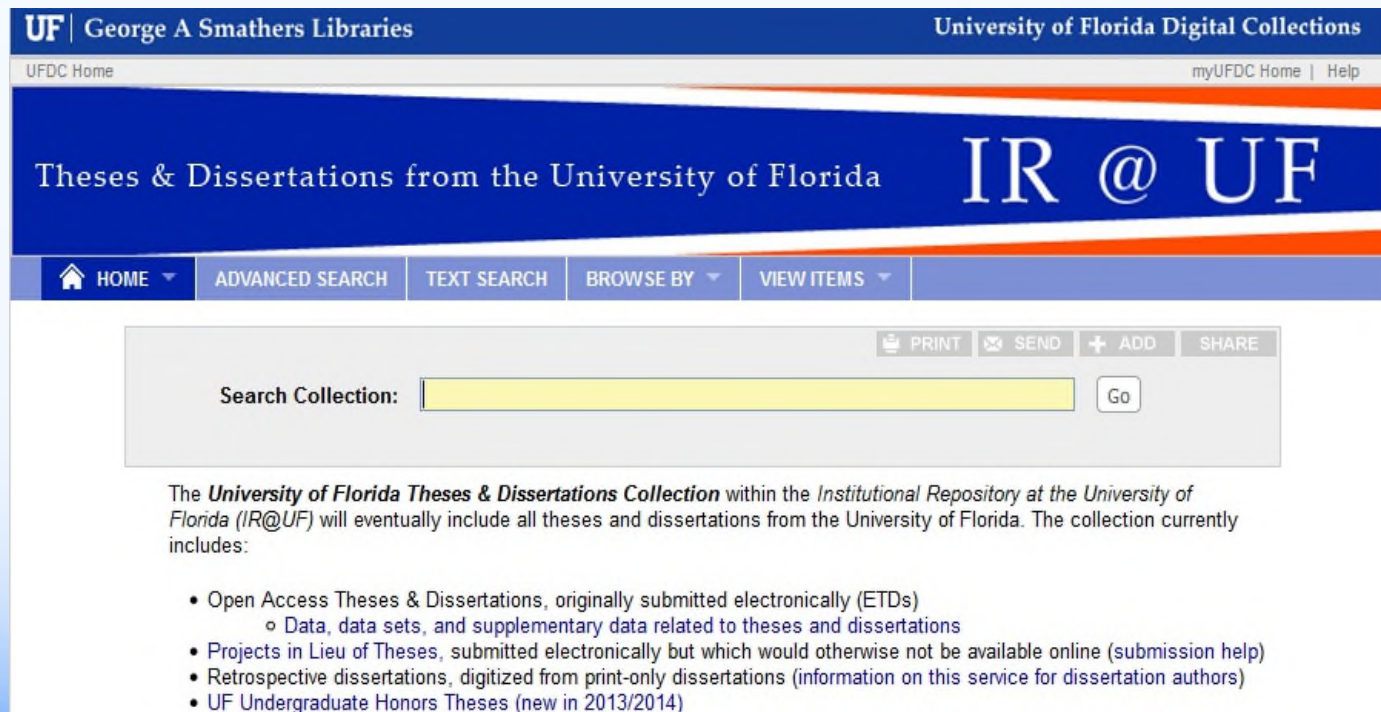
triumphlearning™

15

UF | UNIVERSITY of
FLORIDA

Florida Thesis 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](#).



The screenshot shows the University of Florida Digital Collections (UFDC) website. The header includes the University of Florida logo and the text "George A Smathers Libraries" and "University of Florida Digital Collections". Below the header, there is a navigation bar with links: "HOME", "ADVANCED SEARCH", "TEXT SEARCH", "BROWSE BY", and "VIEW ITEMS". The main content area features a search bar labeled "Search Collection:" with a yellow input field and a "Go" button. Above the search bar are buttons for "PRINT", "SEND", "ADD", and "SHARE". Below the search bar, there is a paragraph of text: "The **University of Florida Theses & Dissertations Collection** within the *Institutional Repository at the University of Florida (IR@UF)* will eventually include all theses and dissertations from the University of Florida. The collection currently includes:" followed by a bulleted list of items.

- Open Access Theses & Dissertations, originally submitted electronically (ETDs)
 - Data, data sets, and supplementary data related to theses and dissertations
- Projects in Lieu of Theses, submitted electronically but which would otherwise not be available online (submission help)
- Retrospective dissertations, digitized from print-only dissertations (information on this service for dissertation authors)
- UF Undergraduate Honors Theses (new in 2013/2014)

Florida Thesis Project

All digital and digitized UF theses and dissertations were selected as the content for the project. The objective was to identify the ones for which Florida is the subject matter and apply enhanced metadata derived using controlled vocabulary to each one.

Access Innovations 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.

Florida Thesis Project

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.
- 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.

Florida Thesis Project

Sample records from the pilot project clearly demonstrate the enhanced metadata:

Subjects

Subjects / Keywords: alligator, detectability, estimates, habitat, population, sightability, survey
Interdisciplinary Ecology -- Dissertations, Academic -- UF

Genre: Electronic Thesis or Dissertation
bibliography (marcgt)
theses (marcgt)
Interdisciplinary Ecology thesis, M.S.

Original Record

Enhanced Record

Subjects

Subjects / Keywords: alligator, detectability, estimates, habitat, population, sightability, survey
Interdisciplinary Ecology -- Dissertations, Academic -- UF
Alligators (JSTOR)
Population estimates (JSTOR)
Aquatic habitats (JSTOR)
Wet prairies (JSTOR)
Vegetation (JSTOR)
Night lights (JSTOR)
Water depth (JSTOR)
Area Surveys (JSTOR)
Florida -- Everglades
Florida -- Kissimmee
Florida -- Franklin County
Florida -- Taylor County

Genre: Electronic Thesis or Dissertation
bibliography (marcgt)
theses (marcgt)
Interdisciplinary Ecology thesis, M.S.

Florida Thesis Project

The enhanced metadata can be searched and edited in the XIS® Staff Review Panel

The screenshot shows a web browser window with the URL www.accessinn.com:8081/FloridaSearch/index.jsp. The page features the University of Florida logo and a search bar. The search bar is labeled "Search Florida Thesis database" and has a dropdown menu with the following options: "All fields", "Title", "Authors", "Keywords Floridathes", and "Keywords Geothses". The "Keywords Floridathes" option is selected, and a type-ahead search is in progress. The search term "brid" is entered, and a list of suggestions is displayed: "Arch bridges", "Bridge bearings", "Bridge decks", "Bridge engineering", "Bridge piers", "Bridge railings", "Bridle paths", "Concrete bridges", "Curved bridges", "Girder bridges", "Highway bridges", "Land bridges", "Liquid bridges", and "Railroad bridges". A blue arrow points from the text "Type ahead based keyword search on taxonomy terms" to the search bar.

R7
R8

Type ahead based keyword
search on taxonomy terms

Dia 20

R7

Inserted in teh UFDC Index?

Russell, Judith, 12-Oct-17

R8

This isn't showing the enhanced metadata, it is showing the search functionality in the staff interface.

Russell, Judith, 12-Oct-17

Florida Thesis Project

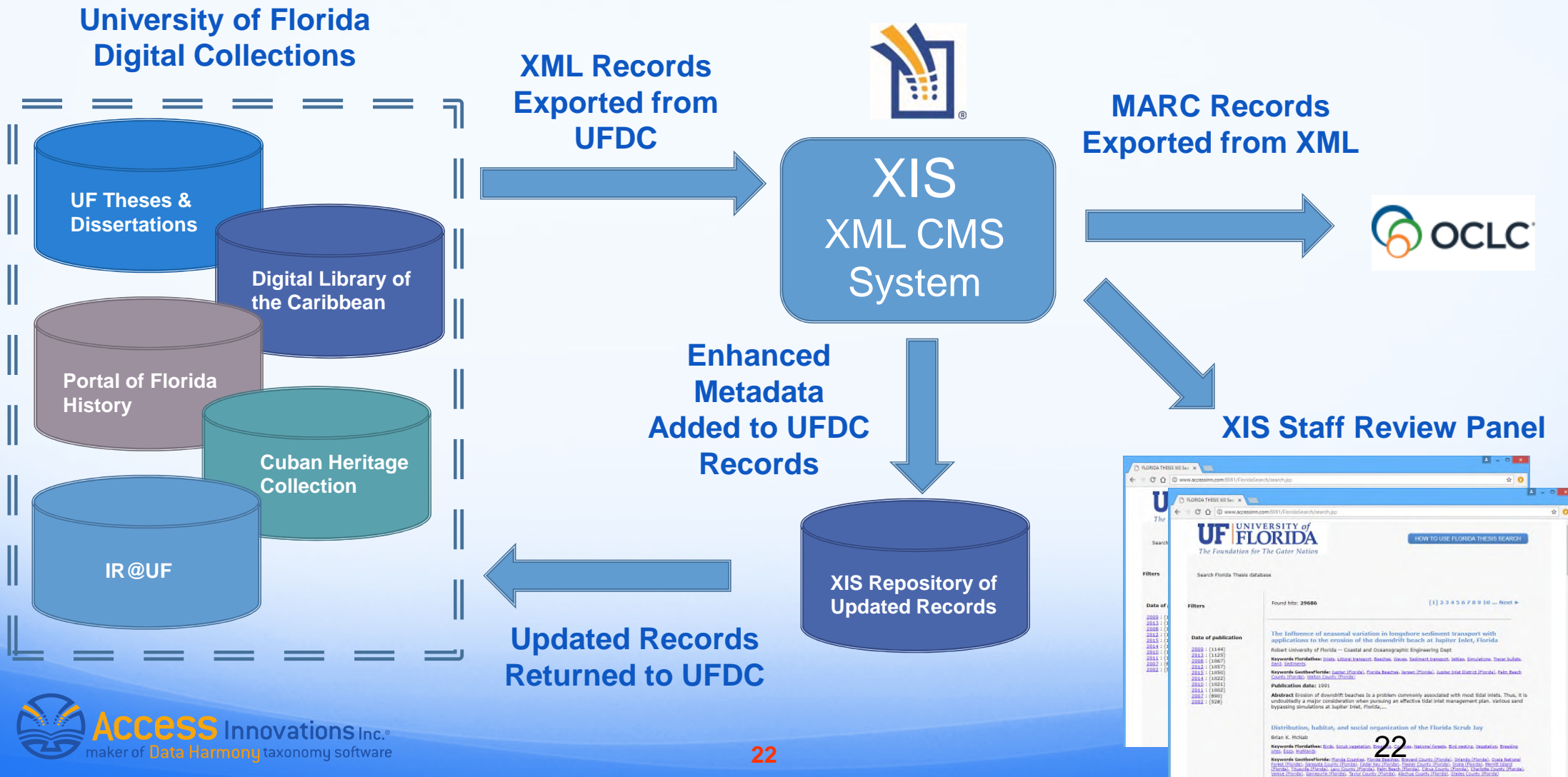
XIS® Staff Review Panel

Side filters for additional search refinement

Linked keywords for easy topic browsing

The screenshot shows a web browser window with the URL www.accessinn.com:8081/FloridaSearch/search.jsp. The page features the University of Florida logo and a search bar. On the left, there are side filters for 'Date of publication' with a list of years and their corresponding hit counts: 2013 (1818), 2015 (1691), 2012 (1685), 2009 (1649), 2010 (1562), 2014 (1515), 2008 (1456), 2011 (1448), 2007 (1230), and 2003 (914). The main content area displays search results for 'Found hits: 49696'. The first result is 'A Toolkit for managing XML data with a relational database management system' by Herman Lam. It includes keywords for Florida (XML, Warehouses, Databases, SQL, Engines, Oracles, Document titles, Relational databases, Data models, Data management) and Geothesis (Levy County (Florida)). The publication date is 2001. The abstract states: 'This thesis presents the underlying research, design and implementation of our XML Data Management Toolkit (XML toolkit), which provides the core functionality for storing, querying, and managing XML data using a relational database management system...'. Below this, another result is partially visible: 'Source specific query rewriting and query plan generation for merging XML-based semistructured data in mediation systems' by Sanguthevar Rajasekaran, with keywords for Florida (Document titles, Bibs, XML, International standard book numbers, Domain ontologies, Engines, Bibliographies, Data models, Databases, Warehouses) and Geothesis (Gainesville (Florida), Levy County (Florida)). The publication date is 2001. The abstract for this result is partially visible: 'ABSTRACT: This thesis describes the underlying research, design, implementation and testing of'.

Florida Thesis Project



Application of XIS® to All UFDC Content

The Florida Thesis Project used Lucene Solr for search since both XIS® and Sobek^{CM} use that software:

- Initial testing and evaluation of the search results was done in XIS®.
- Enhanced metadata and connections to the Lucene index in UFDC were added through the regular Sobek^{CM} load process and reassessed.
- And the Pilot Project was concluded.

We are very encouraged by the quality and quantity of metadata created using these automated tools.

Application of XIS® to All UFDC Content

With the Pilot Project concluded, the Data Harmony (DH) software from Access Innovations will be linked to UFDC via an API.

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 records will be created in the XIS® Data Input Panel and then loaded into UFDC, and submitted to the UF Libraries Discovery Service and OPAC as well as OCLC.
- When appropriate, these records will be identified in UFDC as part of the Portal of Florida History.
- 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.

Unique Collections in UFDC

Identifying materials for the **Portal of Florida History** remains a high priority, but other unique collections also require attention and provide unique challenges.

Large scale collaborations create special challenges, and UF has two particularly important, large scale, collaborative digital initiatives:

- Over 10 years ago, Florida International University and the University of Florida received a federal grant to establish the Digital Library of the Caribbean (dLOC).
- In June 2016, the Smathers Libraries signed an agreement with the Biblioteca Nacional de Cuba José Martí (BNJM) to create broad and deep open access to the Cuban Heritage Collections.

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.

- 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.

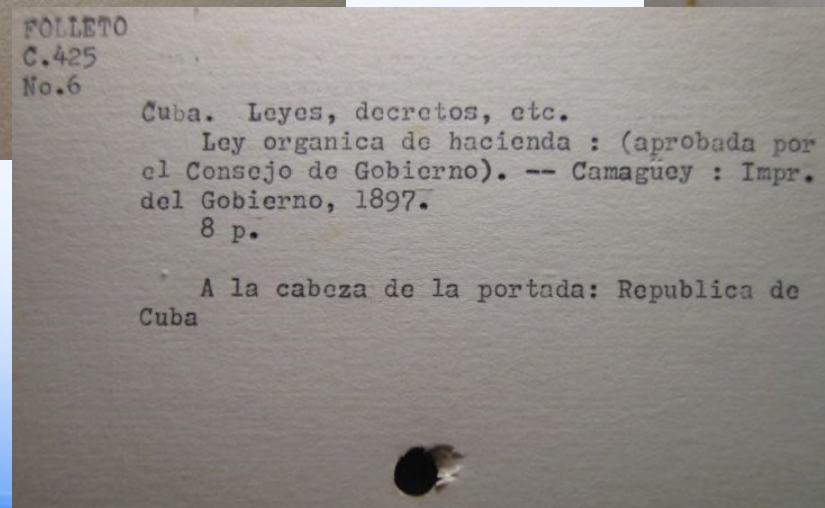
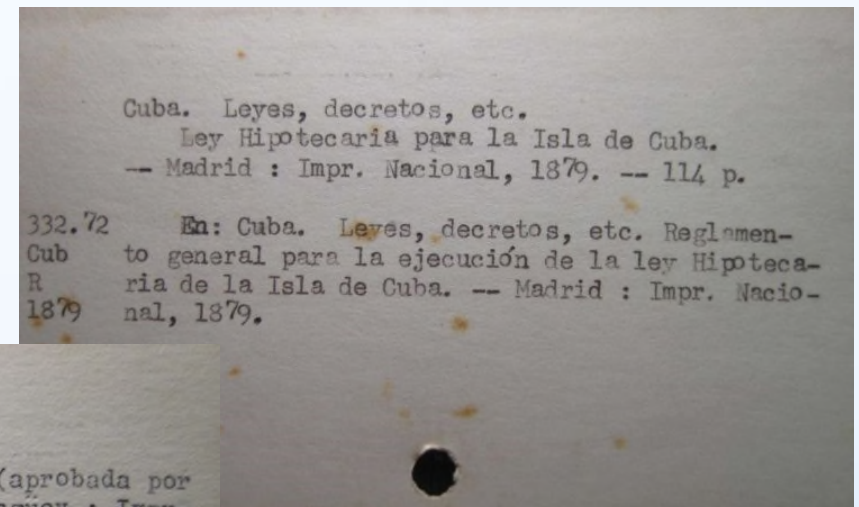
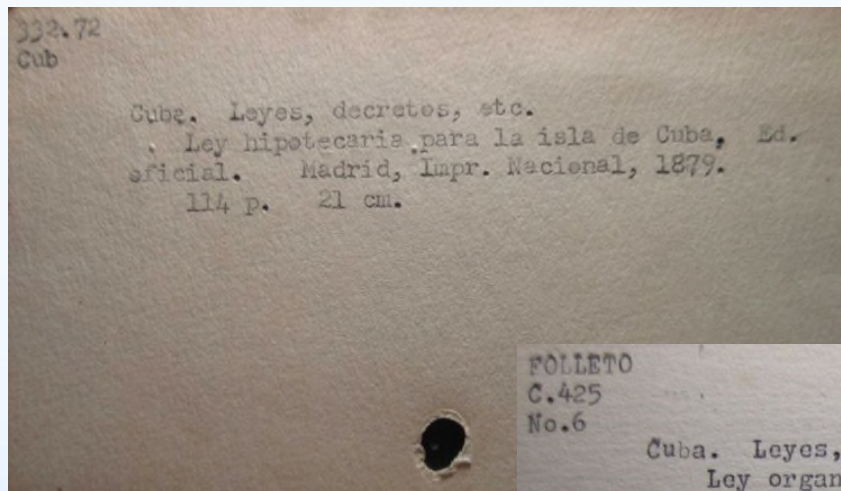
Cuban Heritage Collections

Once digitized, the Cuban 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 uniquely held in Cuba will be contributed by the BNJM to the Digital Library of the Caribbean.
- Multiple partners will contribute digitized content from their own collections 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.

Cuban Heritage Collections

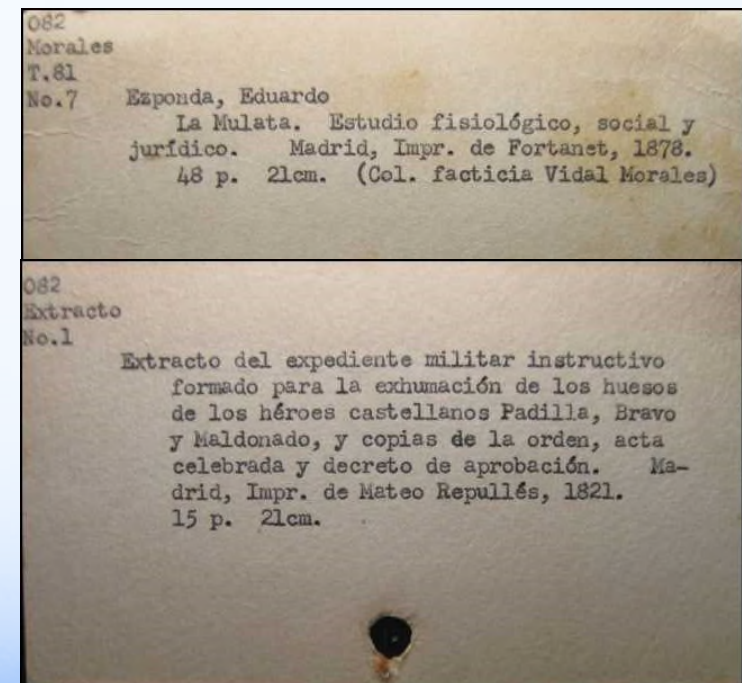
Many of the historic publications of Biblioteca Nacional de Cuba José Martí (BNJM) are identified only by catalog cards.



Cuban Heritage Collections

BNJM shared ~16,000 scanned catalog cards for conversion into MARC records for OCLC and into metadata for the project management database.

- 082 Morales T.81 No.7
 - Ezponda, Eduardo
 - La Mulata. Estudio fisiológico, social y jurídico.
 - Madrid, Impr. de Fortanet, 1878.
 - 48 p. 21 cm. (Col. facticia Vidal Morales)
-
- 082 Extracto No. 1
 - Extracto del expediente militar instructivo formado para la exhumación de los huesos de los héroes castellanos Padilla, Bravo y Maldonado, y copias de la orden, acta celebrada y decreto de aprobación.
 - Madrid, Impr. de Mateo Repullés, 1821.
 - 15 p. 21 cm.

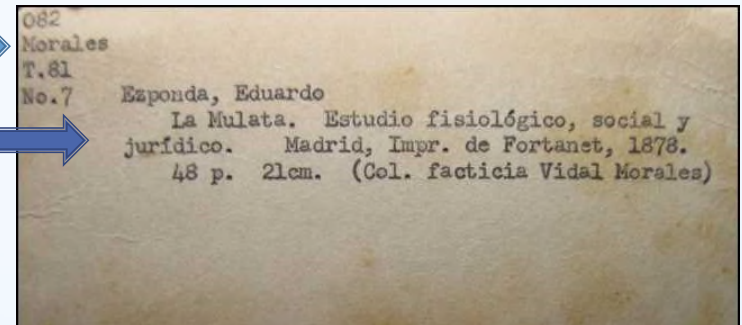
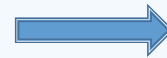


Cuban Heritage Collections

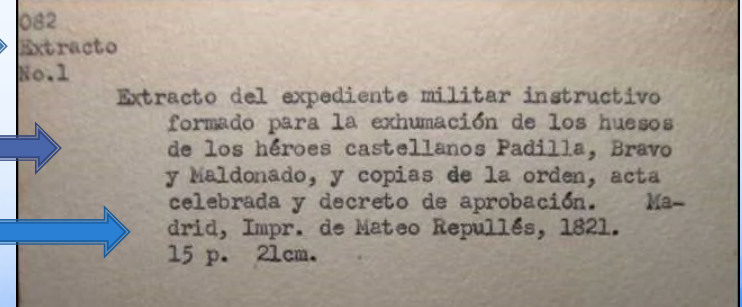
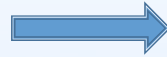
High quality OCR was used to digitize the catalog cards. However, the nonstandard formatting of the cards meant the resulting text files were difficult to parse into fields.

- Call numbers have a variety of lines and formats
- Formatting of the bibliographic information varies
- A variety of abbreviations are used
- Records are in Spanish

4 lines



3 lines



Cuban Heritage Collections

A combination of Bayesian inference and Boolean logic was developed to break the cards into fields.

- Folleto
 - C.35
 - No.7
 - Gordillo, Miguel [1843-1878]
 - Compendio de geografía física de la Isla de Cuba.
 - Habana, Impr. Viuda de Barcina, 1880.
 - 23 p.
- A string starting with “No.” or similar variant and the location in the upper left corner identify the call number. If that is not found, we search for a long string of letters indicating division between call number and bibliographic information.
- Initial capitalized words separated by commas usually indicate names. Must consider “del” and other non-capitalized name components.
- Long letter strings usually indicate titles
- Authority files flag publisher information lines by identifying locations.
- Number combinations with “p.” and “cm.” indicate pagination and dimensions

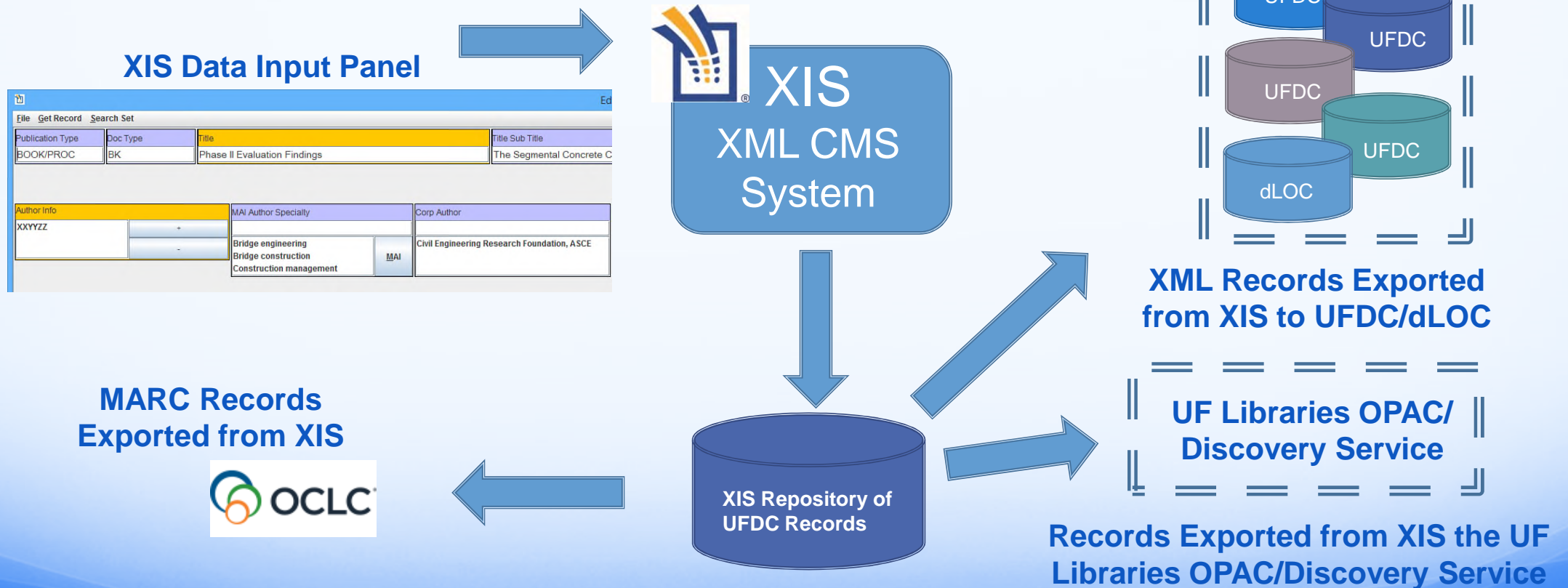
Application of XIS® to All UF Cataloging

Planning has begun for a transition to XIS® for all cataloging/metadata creation and subject indexing, not just for UFDC, but for cataloging and metadata creation for all UF collections.

- 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 IR@UF will be processed through XIS® to provide consistent metadata, including use of thesaurus terms.
- This will ensure that the records in UFDC and the OPAC/Discovery Service are consistent and result in submission of more complete records to OCLC.

Planned Florida Record Creation

All Records Created in XIS



Death of the Library Catalog?

The library catalog 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 Integrated Library System” (ILS) that provides electronic resource management and other services, but is still primarily a Library Catalog utilizing MARC records.

Our students, having grown up with Google, are much more likely to use the single search box from the Discovery Service than the OPAC – or just to use Google.

Death of the Library Catalog?

Not yet, but we are placing increasing emphasis on digital collections and reducing our reliance on catalogers to create MARC records while increasing our investment in automated metadata creation (which can generate MARC records as long as we need them for the Catalog).

We are inverting the traditional cataloging process. MARC records will no longer be the original format used to generate most metadata. Instead, automated tools will be used to generate MARC records.

I predict that within 10 years (perhaps sooner) “traditional” cataloging, applying a title by title effort by trained catalogers, may require as little as 3% of our budget and only 4 or 5% of our employees.

Death of the Library Catalog?

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.

But discovery of those records is less and less likely to occur through the OPAC and the thesaurus used for generation of automated metadata will supplement and eventually replace LCSH as the vocabulary used by those catalogers.

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.

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

Access Innovations helps clients with innovative search and data management solutions with its software and services.

Marjorie Hlava

President

Access Innovations, Inc.

mhlava@accessinn.com

Photograph by Catherine Russell



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