Indexing of Special Collections & Archives for Increased Accessibility:
Transforming Access to Library Collections with Automated Metadata

Coalition for Networked Information (CNI)
Washington, DC
December 11, 2017

Judith C. Russell
Dean of University Libraries
University of Florida

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

Access Innovations Inc.® maker of Data Harmony taxonomy software

UF UNIVERSITY of FLORIDA

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.**


UF | George A Smathers Libraries
University of Florida Digital Collections
myUFDC Home | Help
PRINT
SEND
SHARE
UFDC
University of Florida
Digital Collections
Search Collection:
Go
1
2
3
4
5
6
7
8
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

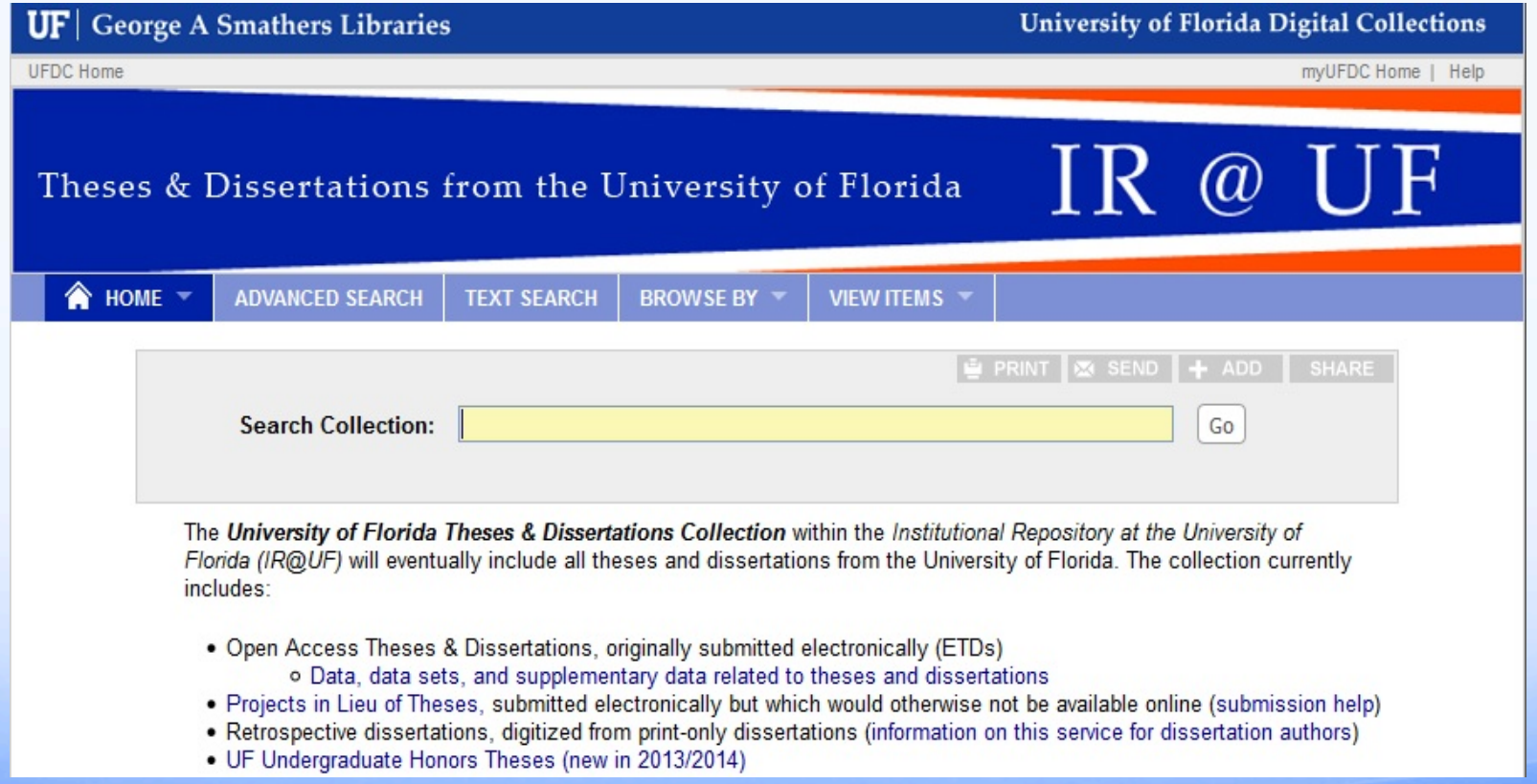
Unique Collections provide unique challenges, so UF sought to acquire automated tools and define processes that can be applied across the full spectrum of our collections. This is necessary because:

- 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 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.



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 record from the pilot project clearly demonstrates the enhanced metadata:

EFFECTS OF HABITAT TYPE AND STRUCTURE ON DETECTION PROBABILITIES OF AMERICAN ALLIGATORS (*Alligator mississippiensis*) DURING NIGHT-LIGHT COUNTS By CAMERON BLAIR CARTER (2010)

Original Record
Limited Author-Assigned Keywords

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.

Enhanced Record
Additional Computer-Assigned Controlled Vocabulary Topical & Geographic Terms

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

University of Florida Digital Collections

UF Theses & Dissertations

Digital Library of the Caribbean

Portal of Florida History

Cuban Heritage Collection

IR@UF

XML Records & Full Text Exported from UFDC for Analysis

XIS XML CMS System

MARC Records Exported from XML

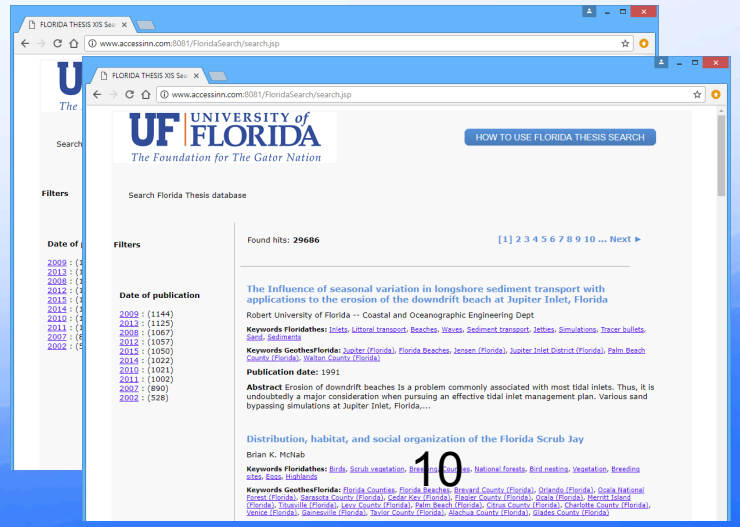
OCLC

Enhanced Metadata Added to UFDC Records

XIS Repository of Updated Records

Updated Records Returned to UFDC

XIS Staff Review Panel



Access Innovations Inc.®
maker of Data Harmony taxonomy software

10

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.

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.

Application of XIS® to All UF Cataloging

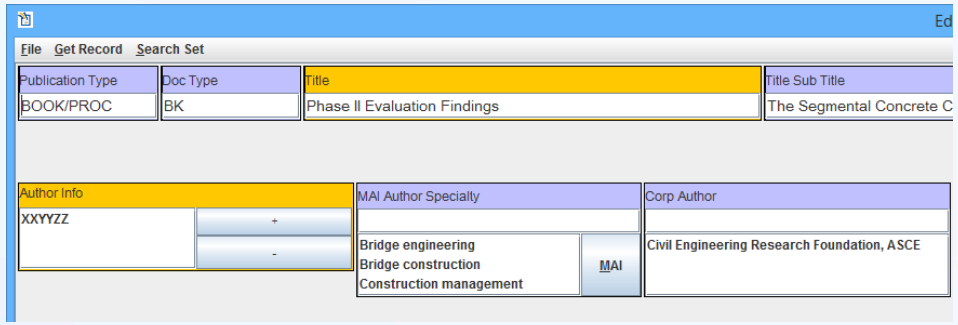
Planning has begun for a transition to XIS® for all cataloging/metadata creation and subject indexing, not just for UFDC, but for cataloging and metadata creation for all UF collections.

- 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.

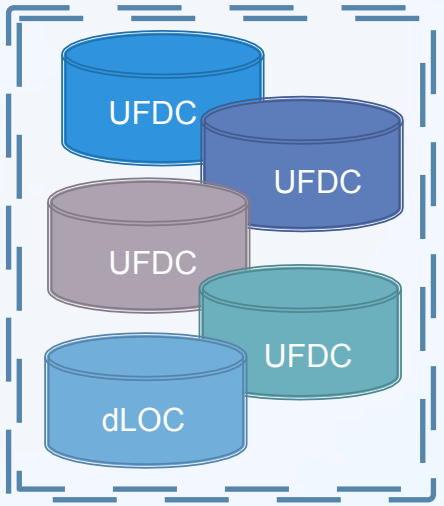
Planned Florida Record Creation

All Records Created in XIS

XIS Data Input Panel



XIS
XML CMS System



XML Records Exported from XIS to UFDC/dLOC

XIS Repository of UF Records

UF Libraries OPAC/ Discovery Service

MARC Records Exported from XIS to the UF Libraries OPAC/Discovery Service

MARC Records Exported from XIS to OCLC

OCLC

Transforming Access to Library Collections with Automated Metadata

Does this transition imply the death of the Library Catalog? Not yet, but we are placing increasing emphasis on digital collections and reducing our reliance on catalogers to create individual MARC records while increasing our investment in automated metadata creation (which can generate MARC records as long as we continue to 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, will require substantially less of our budget and a much smaller number of employees. Traditional cataloging will be used primarily for special collections and materials for which there is no digital surrogate.

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

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