Increasing Discoverability to ETDs through Machine Assisted Indexing: A UF Case Study

By Christy Shorey and Chelsea Dinsmore

September 13, 2018
USETDA 2018, Denver, CO

The University of Florida Libraries began working with Access Innovations in 2016 to build a Machine Assisted Indexing (MAI) tool for use on UF’s digital collections. Beginning with full-text scans, the system uses word frequency to suggest keywords for titles. A rules set prioritizes words and redirects to preferred terms, with human review as the final step. MAI helps minimize the time needed to identify relevant keywords, and enables us to use more natural language terms than are typical in traditional cataloging. We chose our ETD collection as our pilot project for this process.

Why ETDs?
- UFETD collection good size for test (25,000 titles)
- Both digitized and born digital files
- Good variation in quality and formatting over the years
- 1934-2016 (at time of pilot)

How did we start?
- Training with Access Innovations, Oct. 12-18, 2017:
  - Theory of Knowledge
  - Taxonomy Fundamentals
  - How does Search Work
  - What can I do with a Taxonomy
  - Introduction to Data Harmony (the MAI tool)

Our Thesauri Work
- Began with a thesaurus developed for Social Studies and Humanities content
- Enhanced an existing Geographical thesaurus to cover more Florida terms
- Developed a Florida Name Authority set to better identify relevant subject terms
- Will continue as an ongoing, iterative process

Inconsistent Metadata
- Full Metadata Fields
- Standards Vary Over Years
- Original Scan
- OCR Quality
- Improved OCR
- Cross System Communication
- Evaluation

OCR Quality
- Poor scans can result in poor OCR
- Poor OCR means the full text search for terms is ineffective

Cross System Communication
- Need:
  - Crosswalk from UFDC fields to Access Innovations fields
  - Mechanism to send XML, txt and PDF files to AI, and receive enhanced XML to ingest
- Solution:
  - Marshaling Applications Website (MAW)
  - In-house tool connecting UFDC and Access Innovations

Our pilot project for this process.
- Option 1
  - Create METS XML from existing metadata (MARC or XML)
  - Scan or upload item
  - Send through MAI
  - Enhance Metadata with MAI subject terms, then as needed

- Option 2
  - Create METS XML from existing metadata (MARC or XML) + MAI subject terms
  - Scan or upload item
  - Send through MAI
  - Enhance Metadata with MAI subject terms as needed

Success in this project is the process, not the final product. The pilot exposed areas of our system that need more attention:
- Quality changes of OCR over time
- Flaws in system, such as bad field mapping
- Areas that need to be updated to reflect new practices
- Metadata inconsistencies
- MAI is just that, machine assisted indexing, requiring humans to:
  - Validate selected terms
  - Create rules relevant to our collection

As we incorporate MAI into our daily routine, we must decide where it will best fit in the workflow.

Future Directions
- Apply MAI to more UFDC collections
- Use MAW to transfer content between systems
- Finalize workflow for enhancing existing digital records
- Develop workflow for newly digitized items

What is it? Can we measure it?
- We can measure how many records we enhanced by:
  - Improving subject terms based on full text
  - Adding “human readable” subject headings
- How do we measure impact on users?
  - Number of views and downloads?
  - Increased citations?
  - Other?

Lessons Learned
- Increased citations?
- Number of views and downloads?
- Improved OCR
- Inconsistent Metadata
- OCR Quality
- Cross System Communication
- Evaluation

Poster online at: http://ufdc.ufl.edu/IR00010552/00001