

Testing Assumptions—Does Enhancing Subject Terms Increase Use?

Todd Digby and Chelsea Dinsmore
University of Florida, USA

In modern library systems, access to the digital content is heavily dependent on effective metadata. The University of Florida (UF) Digital Collections (UFDC) are an actively growing, open access, digital library comprising over 500,000 records. As with any large-scale digital library project, a well-known challenge is the varying quality and quantity of legacy metadata available for each title. Inconsistent metadata makes digitized materials harder to find. If users cannot find the content they are looking for, a great deal of human effort has been wasted and the investment in digital collections is not being realized. Subject terms can be one of the most efficient methods for accessing desired materials, and subject terms created from controlled vocabularies deliver the most consistent results. To date, applying and editing subject metadata has been a record-by-record, labor-intensive process, making the prospect of retrospective projects cost-prohibitive. The UF team is investigating the capacity of research library staff to implement a Machine Assisted Indexing (MAI) system to automate the process of selecting and applying subject terms, based on the use of a rule set combined with controlled vocabularies, to the metadata of a body of already digitized content. To execute the project, the Smathers Libraries team at UF is collaborating with Access Innovations (AI) consultants to implement a machine-assisted indexing system to mitigate the challenges discussed above.

Two collections in the UFDC were selected to test the MAI process on and then assessments were developed to determine if the process was functional and if it met the stated need to improve access. The first pilot focused on enhancing subject metadata across the Electronic Thesis and Dissertations (ETDs) collection. A second pilot assessment effort focused on a long run of a journal with strong historical ties to agriculture in Florida. Random issues of the title were selected for machine assisted indexing and the use of those issues will be measures against the use of the other issues in the series.

This paper addresses our methods and outcomes of these two pilot projects. Next steps and more in-depth assessment methodologies will also be discussed. Through this assessment, we look to improve and streamline our workflows and determine if our enhancements have increased access and discovery of these pilot digital collections.

Machine Aided Indexing—Overview

In a world that is now dominated by non-library based web search engines, with hidden search algorithms and full-text searching, many researchers rely on only the first page of results to find what they are looking for.¹

This approach has also been adopted in the world of searching through library resources where a single discovery layer will search across the multitude of catalogs, digital library platforms, journal databases, and other subject-specific indexes. As our access to full-text resources grows, the ability to hone in on specific and relevant information becomes increasingly more important. The increased volume of information that is now accessible has caused many to recognize that “current search engines yield good results for specific search tasks but are unsuited to the conceptual or subject-based searches requiring high precision and recall, common in academic research or serious public inquiry.”²

Indexing has been a part of the library world since before the electronic age and is defined, “according to the British indexing standard (BS3700:1988), [as] a systematic arrangement of entries designed to enable users to locate information in a document.”³ This process of manually assigning indexing terms has been taking place with limited changes as libraries moved from print indexing systems to electronic indexing systems.

Reason to Use Machine Aided Indexing

Within the context of the library catalogs/OPACs and library digital collections, the cataloging and indexing of these collections has been a manual process completed by catalogers, or in the case of theses or

dissertations, this may have been completed by the researcher's submission to the institutional repository. Library indexes are developed using both the cataloging record of these items and may possibly include the full-text of items, allowing for a wide discrepancy of the level and precision of the indexing available for our discovery systems to aid researchers finding relevant materials. Although cataloging and indexing within libraries has historically been a manual process, there has been a limited history of using an automated or computer-aided indexing method. NASA, for instance, has been using machine-aided indexing for a number of decades to index scientific and technical reports. This work was largely done to speed up the indexing and provide catalogers with a set of terms to review.⁴ Other efforts have also focused on extracting subject indexing through keyword or key phrase analysis.⁵ These efforts, however, have been limited and have not found their way into mainstream library-based cataloging and indexing practices.

The impetus to find more effective ways to generate and maintain current subject metadata at the University of Florida came from a proposal to build a digital collection around materials about Florida. This unexpectedly represented a significant challenge, since a term like "Florida" is both a location and found in the name of our institution, the University of Florida. Additionally, the terms "University" and "Florida" are found in the names of at least ten more institutions within the State University System of Florida. Given these challenges, a more precise method of updating geographic and other more general subject metadata was needed.

These metadata enhancement efforts were supported and championed by the library dean, who stated, "Recent large scale initiatives have focused on the need for significantly expanded and enhanced metadata for our digital collections, both retrospective and prospective."⁶ In looking for possible solutions to our needs, we engaged Access Innovations, a company that provides thesaurus construction and database management tools to publishers and other entities. Using their Data Harmony software, the University of Florida undertook two pilot projects to enhance our digital library metadata.

Two Pilot Projects Overview

Electronic Thesis and Dissertations (ETDs)

The initial pilot focused on an effort to apply MAI to enhance subject metadata across the Electronic Thesis and Dissertations (ETDs) collection. This collection has been populated by researchers at the University of Florida; broad subject terms (often supplied by the authors) have not provided precision findings. The objective of this pilot was to apply enhanced subject metadata generated—using a controlled vocabulary provided by JSTOR—to each of the 29,000 publications in the collection and test for improved findability. Using the Access Innovations software MAIstro™, the enhanced subject terms were extracted from the full text of the UF theses and dissertations before being added to the metadata records of the ETDs from the UF digital collections.

Example of ETDs' Enhanced Subject Metadata

Subjects

Subjects / Keywords: dissolved, gastropods, grazing
Fisheries and Aquatic Sciences -- Dissertations, Academic -- UF

Genre: Electronic Theses
bibliography (n) theses (marg) Fisheries and A

Enhanced Subject Terms

Original Subject Terms

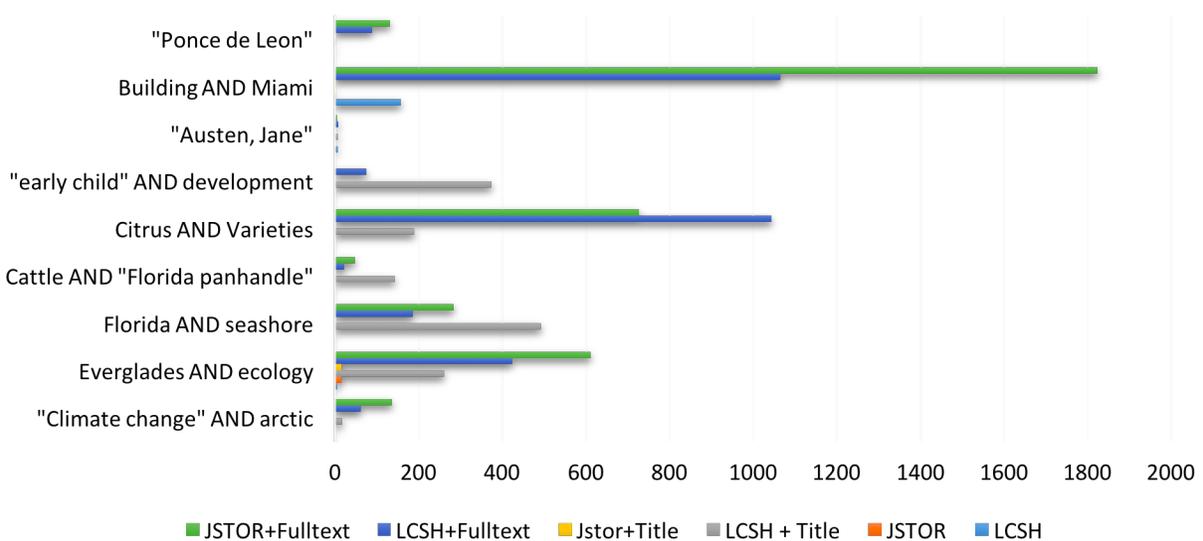
Subjects

Subjects / Keywords: dissolved, gastropods, grazing
Fisheries and Aquatic Sciences -- Dissertations, Academic -- UF
Snails (JSTOR)
Macrophytes (JSTOR)
Hypoxia (JSTOR)
River water (JSTOR)
Aquifers (JSTOR)
Karsts (JSTOR)
Florida -- Ichetucknee Springs State Park
Florida Springs (JSTOR)
Florida -- Weekie Wachee
Florida -- Tallahassee
Florida -- Homosassa
Florida -- Hernando County

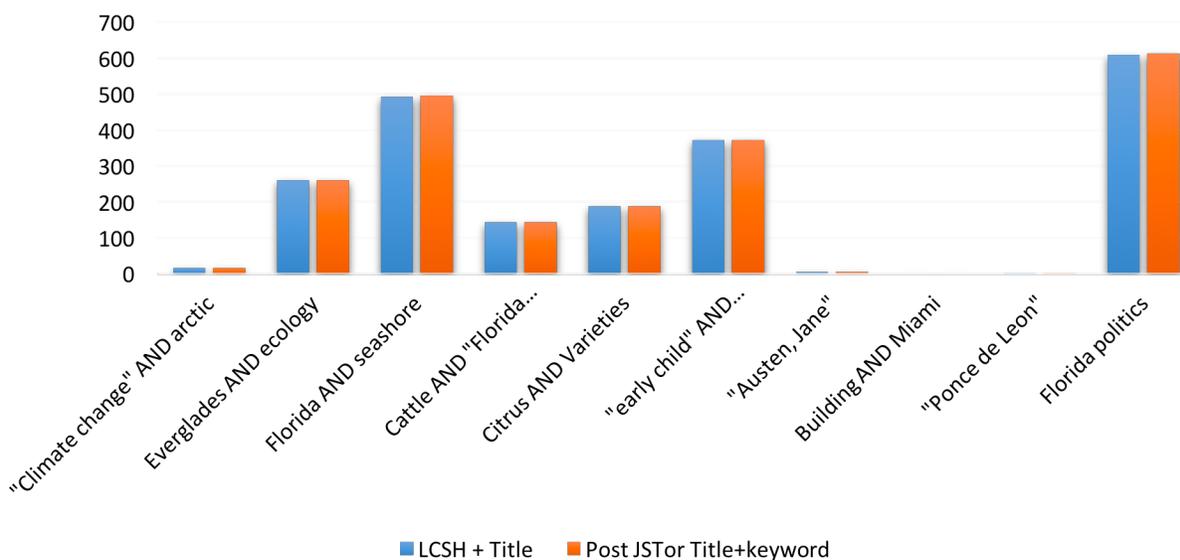
From: *Oxygen Mediated Grazing Impacts in Florida Springs*, Kristin Dormsjo (2008), <http://ufdc.ufl.edu/UFE0021801>.

To assess changes in findability, a series of simple and complex searches were run against the original UFDC set of materials (LCSH and author-submitted keywords). The same searches were run on just the MAI JSTOR supplied subject terms, held in the Access Innovations XML database (XIS). Searches were run looking at just subject terms, subject terms and titles, and finally, subjects and full-text in both systems. Once these numbers were obtained, the JSTOR terms were added to the UFDC records and the searches were run again, allowing the UF team to compare result rates of the ETDs collection before and after enhancing the subject metadata.

Results of Different Search Types



Number of Search Results before and after JSTOR added



Results were limited on this study. It was determined that the XIS system was not well-suited to performing complex searches. Additionally, the JSTOR terms were added to the UFDC metadata records, but there was a problem with the system when it came to reindexing such a large batch of updated records, so the combined subject results did not reflect full findability on the new terms. It was decided that, in addition to correcting the indexing barrier, a qualitative research study is needed to truly assess the value of the added terms.

Cattleman's Pilot

A second pilot assessment effort focused on a long run of a journal with strong historical ties to agriculture in Florida. Randomized issues of the title were selected for machine-assisted indexing and the rate of use of those issues were measured against the use of the other issues in the series. This pilot used the same MAI system and process as the ETD project.

Assessing the impact of this project was initially scoped out to examine access rates between the MAI-enhanced article serial records compared to those issues where the MAI was not performed. In the process of implementing this project, there were issues identified within our article level searching capabilities in our digital library system that were not going to allow for the enhanced MAI records to be searchable in the ways that were initially envisioned. We are currently examining additional ways to assess the impact of these changes to the metadata of article level items.

Conclusion—Next Steps

The initial goal of our overall project was to enhance the metadata to improve accessibility, findability, and, by extension, use of the impacted content. At the beginning of these projects, we believed that this assessment on our two pilot projects was something that was going to be relatively straightforward and give us results that could guide us in future decisions to extend the use of MAI and extend this process to additional digital collections housed by the library. Our assessment as originally conducted has resulted in findings that we did not anticipate. We found that current indexing and searching abilities within our collections had deficiencies which affected our study results. Although these searching deficiencies impacted our ability to gather and assess how our updated metadata can be searched, they have guided us in

planning for modifications and future system development that can be done to provide a more effective search system for accessing our digital collections. For example, this will include modifications to our SOLR indexing system. To get additional usage data, it will be necessary to rework our study and we may select other collections and material types—for instance not a serials collection—to get better usage and findability data. Finally, as we look to the future, we will implement the MAI process to more of our retrospective collections in addition to incorporating it into our regular digital collections workflows.

—Copyright 2019 Todd Digby and Chelsea Dinsmore

Endnotes

1. Griffiths, Jill and Brophy, “Student Searching Behavior and the Web.”
2. Golub, Koralijka, “Automatic subject indexing of text.”
3. American Society for Indexing, “Frequently Asked Questions.”
4. Silvester, Genuardi, and Klingbiel, “Machine-Aided Indexing at NASA.”
5. Medelyan and Witten, “Thesaurus based automatic keyphrase indexing.”
6. Hlava, Russell, and Hansen “Inverting the Library Cataloguing Process to Streamline Technical Services.”

Bibliography

American Society for Indexing. “Frequently Asked Questions.” <https://www.asindexing.org/about-indexing/frequently-asked-questions/>.

Griffiths, Jill and Peter Brophy. “Student Searching Behavior and the Web: Use of Academic Resources and Google.” *Library Trends* 53, no. 4 (2005): 539–554.

Golub, Koralijka. “Automatic subject indexing of text.” *Encyclopedia of Knowledge Organization. International Society for Knowledge Organization*. Last modified November 13, 2018, <http://www.isko.org/cyclo/automatic>.

Hlava, Marjorie MK, Judith C. Russell, and David Hansen. “Inverting the Library Cataloguing Process to Streamline Technical Services and Significantly Increase Discoverability and Search for Special Collections.” Paper presented at IFLA WLIC 2018 – Kuala Lumpur, Malaysia – Transform Libraries, Transform Societies, August 26, 2018. <http://library.ifla.org/2219/1/115-hlava-en.pdf>.

Jacquemin, Christian, Beatrice Daille, Jean Royauté, and Xiavier Polanco. “In vitro evaluation of a program for machine-aided indexing.” *Information processing & management* 38, no. 6 (2002): 765–792.

Medelyan, O. and I. H. Witten. “Thesaurus based automatic keyphrase indexing.” In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, June 2006: 296–297.

Silvester, June P., Michael T. Genuardi, and Paul H. Klingbiel. “Machine-Aided Indexing at NASA.” *Information Processing and Management* 30, no. 5 (1994): 631–645.