DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications

Finding Florida—Implementing Machine Aided Indexing in an Academic Library

Xiaoli Ma
University of Florida, US
xiaolima@ufl.edu

Chelsea Dinsmore
University of Florida, US
chedins@uflib.ufl.edu

Dave Van Kleeck
University of Florida, US
dvankleeck@ufl.edu

Laura Perry
University of Florida. US
lauraperry@ufl.edu

Abstract

From 2018 to 2021, a team of library professionals at the George A. Smathers Libraries worked to implement Machine Aided Indexing (MAI) in order to locate content about Florida places and spaces among the 16 million pages of hosted content at the University of Florida Digital Collections. This semi-automated process uses a combination of commercial software and locally developed applications. MAI consistently assigns subject terms from controlled vocabularies, aka, taxonomies, to thousands of items in a couple of hours. This method selects terms considering the frequency of the terms appearing in the texts and as well as the preset rules that define the concurrence of terms and other contextual restrictions. After three years' effort, the team identified 23% items, out of the 76,316 text-based single-volume content in English, are about named places in Florida and tagged all of these items with place names, adding 34,000 access points for these items. Most of these access points were not available prior to this process. On top of that, the team also compiled a Florida specific taxonomy--*Thesaurus of Florida Place Names.* This paper outlines the key components of this MAI process and details the challenges and lessons learned.

Keywords: metadata creation automation; machine aided indexing; digital libraries, taxonomy.

Introduction

The University of Florida (UF) Digital Collections (https://ufdc.ufl.edu/) aggregates digital content from all over Florida and around the world. Currently, UF Digital Collections host over 16 million pages, covering over 78,000 subjects in all disciplines. In preparing an online portal dedicated to gathering Florida-specific content, a team of library professionals at the George A. Smathers Libraries (the Libraries) wanted to know which items are about named places in Florida. Without knowing the quantity of the target group and having no consistent descriptive information about named places, finding Florida in UF Digital Collections sounds as daunting as looking for a needle in a haystack.

To tackle this task, from 2018 to 2021, the team experimented with the practicalities of implementing Machine Aided Indexing (MAI), using a combination of commercial software and locally developed applications. MAI consistently assigns terms from controlled vocabularies, aka, taxonomies, to thousands of items in a couple of hours. The process selects subject terms based on the frequency of the terms appearing in the text of scanned documents, and prioritizing and filtering based on preset rules that define the concurrence of terms and other contextual restrictions.

In order to inform the MAI process of the named places in Florida, the team compiled a new taxonomy solely for this project and named it *Thesaurus of Florida Place Names (*FLGEO*)*. It holds over 12,000 names across 23 categories. Over 3 years, working with MAI, the team identified 17,107 items that are about named places in Florida, which is approximately 23% of the total of 76,316 text-based single-volume items in English. These items include journal articles, research papers, technical reports, archival documents and many more. All of these 23% items are now tagged with the place names in Florida and can be used for building the Florida

https://doi.org/10.23106/dcmi.953132608

portal site. Beyond this, these newly assigned place names provide over 34,000 access points for these items. Most of these access points were not available prior to this process.

This accomplishment demonstrates that MAI is an efficient way of identifying and grouping text-based content at scale. It quickly improves the discoverability of a large amount of digital content. We also learned that to sustain MAI as a service and to continuously improve its results requires long-term commitment of resources.

UF Digital Collections and the Implementation Team

UF Digital Collections began in 2006. It is the home of large digitalized collections like Florida Digital Newspaper Library (https://newspapers.uflib.ufl.edu/), Baldwin Library of Historical Children's Literature (https://ufdc.ufl.edu/collections/juv) and the Samuel Proctor Oral History Collections (https://ufdc.ufl.edu/collections/oral). These digitalized collections were an institutional priority, and combined with the collection of born digital content, the UF Digital Collections grew at an extraordinary speed. As a result, the libraries face the very real challenge of describing incoming content in a quick and accurate manner. This was particularly important as library staff anticipated system upgrades which would require content reorganizations to optimize the search and browse experience.

Here "describing" refers to a wide range of practices that include cataloging, tagging and indexing the digital content. With the advancement of Optical Character Recognition (OCR) technology, machines can convert scanned images of typed, handwritten or printed words into machine-encoded text files. This advancement in technology has led to great opportunities to expand what machines can do with the texts. Coming out of these opportunities is Machine Aided Indexing (MAI). MAI uses machines to locate target terms selected from taxonomies and then count their appearance in the texts. Based on the counts and preset logic that disambiguates similar terms or defines required context, MAI assigns selected terms to items as subjects. The process takes a much shorter time than conventional cataloging practices done by humans where catalogers read a good chunk of the documents before selecting subjects. Often, it takes a cataloger several minutes to select one subject while MAI assigns multiple subjects to thousands of documents in one hour.

Under the leadership of the Libraries' Associate Dean for Discovery, Digital Services & Shared Collections, the Libraries formed an implementation group of eight members from three departments: Resources Description Services, Digital Services, and Library Technology Services Based on their functional areas, members respectively worked as project advisors (2), project lead (1), digital asset advisor (1), taxonomists (3), and developer (1). The project advisors provided the team the project background, set up the visions and supplied the resources. The project lead articulated the advisors' visions into specific tasks, managed the team's work, communicated the programming requirements with the developer, and worked on all steps of the MAI process. The digital asset advisor instructed the team the best methods to retrieve media files from the storage, assisted with the analysis of OCR related issues, and supported the ingestion of updated XML files into UF Digital Collections. Taxonomists researched and compiled terms to form FLGEO, built term-level MAI rules, ran MAI and then reviewed results. To address term and rule-related issues, taxonomists adjusted term selections and MAI rules. They also discussed with the project lead other issues so those issues could be handled by members of the team with the relevant expertise. The developer worked with the project lead to create the local applications to scale up the MAI process.

The Process

The team worked with Access Innovations (https://www.accessinn.com/), an experienced subject index service and product provider to implement the local solution. The team used Access

https://doi.org/10.23106/dcmi.953132608

Innovations' MAIstro software as the environment to build the taxonomy and to test MAI results with single samples. MAIstro allows users to set the fields needed for taxonomy terms and provides different views to display the relationship among terms. MAIstro also supplies features to assist the rule building. As well, it lets users upload or paste text-based single files to test how MAI results respond to the built rules and the changes in the rules. In short, the fields, the views, the rule building features and the testing capacity compose the environment for building the taxonomy. The taxonomy was then stored and hosted on the Access Innovation' Server where Get Suggest Term API pulls out terms to index batches of thousand items. The API locates and counts target terms from huge amounts of texts retrieved from UF digital collections storage. When terms appear three or more times in the texts, and rank within one of the five terms used most often in one document, the API suggests them as subjects. To integrate the API to local workflow, the team developed a series of applications. These local pieces gather the MAI results from the API and display them so taxonomists can review the results. During the review, taxonomists identified and fix term and rule related issues in MAIstro. They also reported other issues to the team. After that, taxonomists run the local applications to write the approved terms to XML files. Then the project lead and digital asset advisor ingested updated XML files to UF Digital Collections.

In order to identify content about named places in Florida, the team compiled a taxonomy of 12,000 names across 23 categories that cover cities, towns, counties, beaches, historical sites, rivers, lakes and many more from *Thesaurus of Geographic Names* (TGN), the *Geographic Names Information System* (GNIS), *Wikipedia* and a few Florida-focused websites. The plan was to take in as many as names of geographic locations in Florida as could be identified.

The diagram below outlines the process and lists key components.

https://doi.org/10.23106/dcmi.953132608

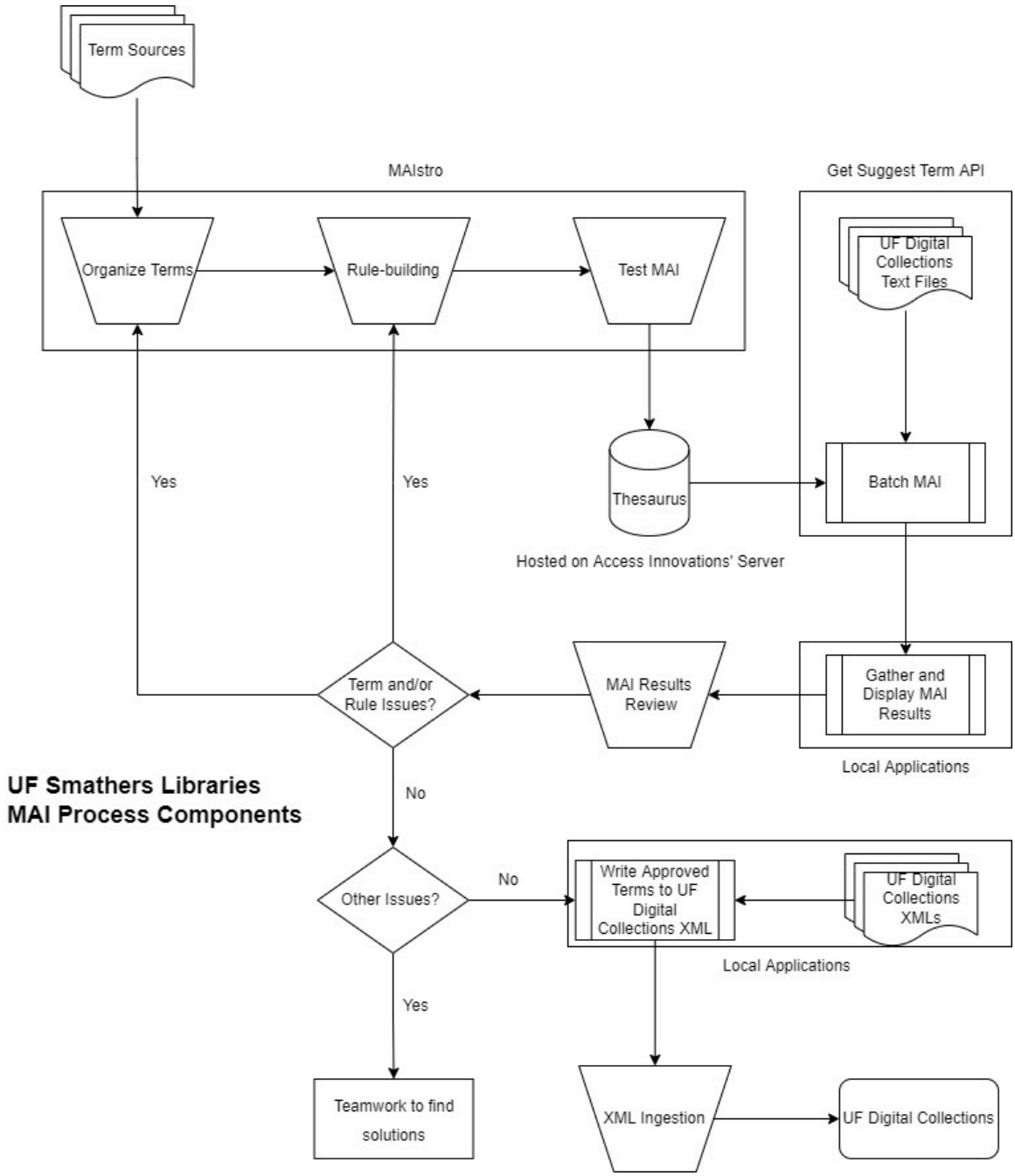

FIG. 1. MAI Components

Frequency, the number of times terms appear in the documents, is the fundamental method to identify the terms that should be used to indicate the key content or feature of the documents. However, frequency counts alone cannot always produce good results. In order to assign appropriate terms to mark the content, machines need sufficient knowledge about the context in which the terms appear, as well as knowledge of the relationship between related and similar terms. For instance, Florida, California and Virginia all have a county named "Orange County". As a Florida specific project, the team set an umbrella rule to every term so that the MAI process

https://doi.org/10.23106/dcmi.953132608

counts the appearance of the terms only when "Florida" or other names of Florida appear in the context.

In addition to assigning rules that applied to all terms, the team also noticed data patterns that required rules to be built. City names are good examples. The official city names mostly follow the pattern "City of" and a noun, but when mentioning particular cities, people use different names. The most common way to refer to cities ignores the "City of" and only uses the noun in city names. For instance, the City of Miami is most commonly called "Miami." Therefore, for all cities, the team put in the rule that when seeing the names without "City" or "City of", count them as an appearance of the official names. However, "Miami " could also appear frequently as a part of newspaper names like *Miami Herald* or football team names, for instance "Miami Hurricanes football". For these cases, the team also used rules to prevent from counting the appearance of "Miami" towards "The City of Miami". Another data pattern that the team identified is triggered by punctuations like "," "()",":", "." included in the place names. Using the example "St. Augustine" MAI cannot process the dot correctly. It won't take "St. Augustine" as one word. That is why custom rules are needed to determine when to suggest "City of Saint Augustine".

Other situations that need rules to guide the subject selection include but not limited to, when terms of broad connotation appear with similar terms of a more specific scope, when several related terms are used interchangeably in the same context, and when the same terms mean different concepts in various contexts. Leaving these situations untreated leads to inaccurate or wrong suggestions. For example, "Alachua" is the name of a city as well as a county in Florida. It is also the name of a large-scale outdoor sculpture by internationally known artist John Henry. The artist named the sculpture after Alachua, the county where the sculpture resides. This sculpture is a landmark on the UF campus. In this case, needless to say, software needs to know that "Alachua" can mean different concepts depending on the context. To disambiguate, the team provided in the rules the crucial words, like "city", "county" and the artist's name, so that the MAI process should consider to determine the meaning of "Alachua".

MAI Results Review

The team did the first round of rule-building immediately after compiling the terms for the taxonomy, and then devoted a review step in the MAI workflow to continuously and routinely identify marginal or incorrect MAI term suggestions. The team would then address these issues via deleting or expanding selected terms and / or conducting further rule building .

The reviewing process started with sample checking. The team sample-checked roughly 1% of every MAI batch, for instance, for a batch of 2459 items would have 25 items reviewed. The taxonomists gained an understanding of the content first by reading the Title and Abstract of the document, then if needed, by reading a part of the main text. Based on this understanding, the taxonomists evaluated whether the software had done a decent job in suggesting the terms. This evaluation depends on the reviewers' past cataloging training and experience that inform them the best practices of tagging the key content with subject terms from authoritative sources. The taxonomists followed a decision tree (see below) to evaluate the top three terms, and logged any issues along the way for further analysis. The taxonomists and the project lead met to discuss their observations on a biweekly basis.

https://doi.org/10.23106/dcmi.953132608

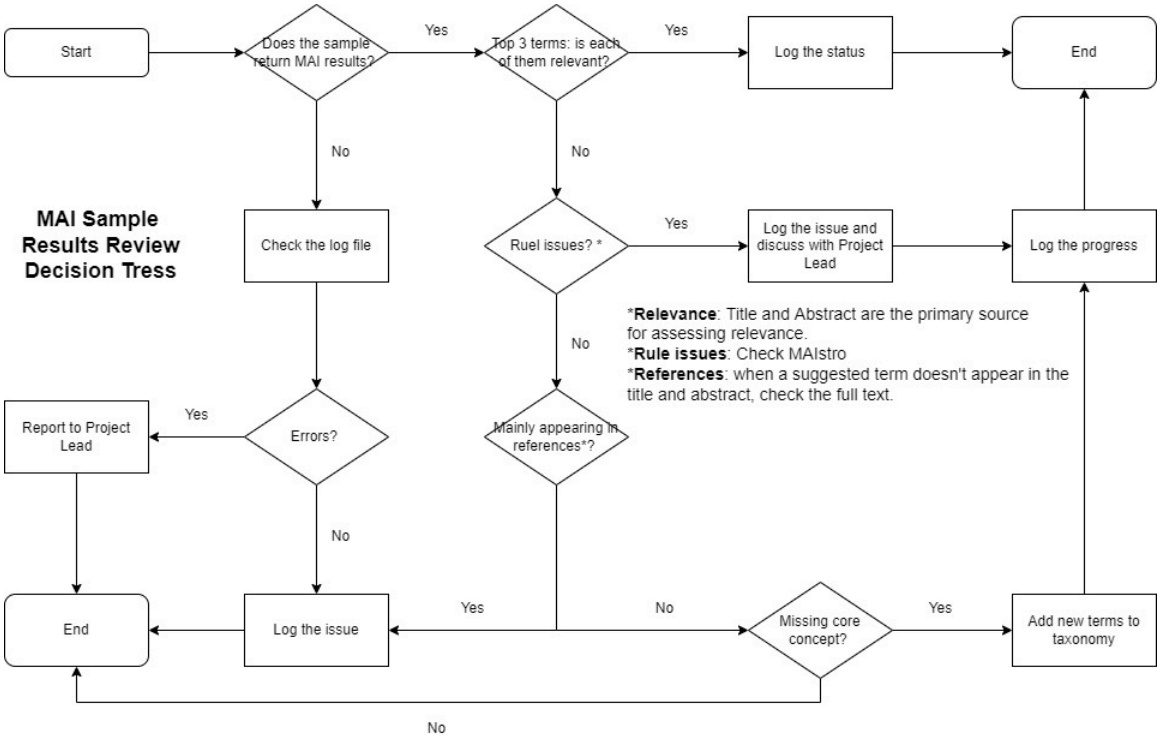

FIG. 2. MAI Sample Results Review Decision Tree

Reviewing is the most time-consuming step of the whole MAI process. However, it is essential and crucial as it ensures the quality of MAI results. It took taxonomists several minutes to evaluate the results of one item, while the MAI process indexed thousands of items in an hour. The issues identified during this step would then be addressed. New rounds of MAI would be followed by new rounds of review to check if the issues were resolved as expected. One resolved issue can improve the MAI results of all affected items that could hundreds or thousands.

Beyond common term and rule related issues, during the reviews, the team saw a few other issues. They include poor OCR quality, "noise" introduced by references, and a lack of important place names in the taxonomy.

Challenges

OCR, the technology that converts images into machine readable texts, doesn't always create sufficient quality text files for further text analysis tasks like MAI. If the item's OCR quality is so poor that the generated texts are not legible, MAI cannot produce any term suggestions. Many factors can lead to poor OCR quality, for instance, the low quality of the scans and the layouts of the original prints. If the scans didn't capture the content of the item well, the OCR process won't be able to yield good results. If the layouts of the original prints are so complicated that OCR programs cannot properly understand them, these programs can generate noise like extra words or special characters in the OCR results. Moreover, OCR technology progressed throughout the years. As a result, the OCR quality varies by the year according to the tool in use at the time OCR was done. Different causes of OCR issues demand different fixes, but in general, these issues call for a dedicated project to analyze, categorize and organize the work that improves the OCR quality to meet the most updated standards.

Not only confined by OCR quality, MAI capacities are also affected by issues around the content "noise", concepts that appear often but don't contribute to the theme. When reviewing

https://doi.org/10.23106/dcmi.953132608

MAI results, the team noticed that 9% of the sampled items suffered from an issue where the assigned place names only appeared in reference lists that were found at the end of the document. That is, the place names were mentioned as a part of bibliographical information, not as key geographic topics. For instance, a presentation about an exhibition in a museum located in Orlando that uses "Orlando" as a part of the museum name can quote the museum's name frequently and list "Orlando" many times in the bibliographic list, however, this presentation is not about Orlando. This situation confused the MAI process. Since "Orlando" appeared so often, machines suggested it as the dominant place name and assigned "Orlando" as a geographic subject. This is a case of incorrect indexing. To avoid this type of error, the team programmed a function to remove a certain percentage of the document from the MAI process. The biggest challenge with this was to decide the "magic" percentage to remove. Removing an insufficient amount cannot resolve the problem. Removing too much can take away the crucial content from the MAI process and therefore, produces poor machine suggestions. After six rounds of tests with a good variety of sample materials at different lengths and from multiple disciplines, the team concluded that removing 5% of the document significantly improved the results. It keeps the content that should be kept and removes the noise effectively. Since this function was put in place, we stopped seeing this issue.

Another major issue is the lack of important place names in the taxonomy. For instance, originally, the taxonomy only included "Central Florida", "South Florida" and "North Florida" to locate content that discusses regions in Florida. As a consequence, content that used other regional terms like "Northeast Florida", "Florida Panhandle" could not be surfaced via MAI. To bridge the gap, the team expanded the taxonomy to include ten terms to cover all major regions in Florida.

While this gap was obvious and the team addressed it, the team believes more gaps like this exist. Constant review of the MAI results is the most effective method to expose the gaps and then act upon them. Of course, the original design of the taxonomy also defines its limitations.

The plan for the taxonomy was to take in as many as names of geographic locations as possible, but at the same time, the team made the decision not to include any organization names, for example, names of schools and universities. These names complicate the "identifying" process. An article could extensively discuss a topic where UF was mentioned many times because some UF scholars worked on the research, but this article was not about Florida. While this decision has helped the team focus on the goal, it also limits what the MAI can produce. To summarize, gaps in the taxonomy in use can negatively affect the quality of MAI results, but no single taxonomy can include everything. The scope of the taxonomy should go closely with that of the project.

Sustainability

UF's MAI process is one of only a handful of projects actively assigning subjects to text-based documents using a semi-automated process in an academic library setting. JSTOR, a digital library of academic journals, books and primary sources, also uses Access Innovations' MAIstro software to grow and maintain its subject thesaurus. Its thesaurus team generously helped the UF team by sharing documentation and experience. UF's case served as a test to see if MAI technology combined with local infrastructure is ready to initiate as a new service in academic libraries. Because of its fast speed, this service has great potential to advance a library's response to the rapid ingest of digital documents, from reactive to proactive approaches.

The UF team's experience has demonstrated that the MAI tools packaged by Access Innovations can successfully meet the needs of an academic library. However, with any cataloging/metadata effort, this MAI method requires a long-term commitment in order to produce and maintain effective results.

To localize the workflow required extensive inter-departmental collaborations. Though the effort to localize MAI workflow took three years, the Libraries worked with Access Innovations

https://doi.org/10.23106/dcmi.953132608

for a total of five years. The first two years were spent testing the technology and sending data back and forth to Access Innovations in order to develop efficient workflows. In addition to the project team that developed the rules and thesaurus, information technology staff, unit heads from cataloging and metadata teams, as well as members of the special collections department attended trainings, meetings, and conferences to discuss the applicability of the tools and methods for integrating them into UF's workflows. Total billed costs for the software subscriptions, training time, and consultations on custom development needed to connect the Access Innovation products to the UF system exceeded $300,000 over 5 years. The authors have not been able to calculate the cost of hundreds of staff hours needed to bring this project to fruition.

Conclusion

Implementing MAI locally has allowed the Libraries to understand the factors that affect the quality of MAI results and has also provided the Libraries the opportunities to run MAI process in parallel with other daily tasks. This experience has prepared the Libraries to leverage more auto or semi-auto technology to speed up the work of classifying digital content.

One outcome from the team's three years' effort is the Florida specific taxonomy -- FLGEO. This taxonomy has been registered with the Library of Congress (LC) as a known authority and listed on the LC standards page (https://www.loc.gov/standards/sourcelist/subject.html). The team plans to share this taxonomy more widely. Already, one artificial intelligence project at UF has taken in a copy of FLEGO as a part of the vocabulary database.

Working with MAI, the team identified 17,107 items that are about named places in Florida, that is approximately 23% of the total of 76,316 text-based single-volume content in English. These items include journal articles, research papers, technical reports, archival documents, etc. All of these 23% items are now tagged with the Florida place names as geographic subjects. When needed, these items can be singled out from the 16 million pages hosted by UF Digital Collections to help form the Florida portal site. Beyond this, these newly assigned place names provide over 34,000 access points for these items. Most of these access points were not available prior to this process. Users can also use these subjects to browse and to group similar content together. In short, the team was able to find the "Florida" needle in the haystack of digital content and also improved the discoverability of Florida-related materials.

The process involved a great deal of staff time devoted to creating the FLGEO thesaurus, building rules within the software for their application, manually reviewing the results, and refining the rules via an iterative process of dialing in their accuracy with various sets of records. As the reader can imagine, this was a lengthy, time-consuming, and expensive process. Once this iterative process was completed though, the automated results were acceptably accurate for use in serving the needs of our digital collections users.

Acknowledgements

In Memory of Robert V. Phillips, III (1954-2020)

The authors would like to acknowledge the contributions of Robert V. Phillips, a close colleague who made several critical contributions to the programming side of the pilot project. Robert passed away during the later stages of the effort. Without his contribution, this project could not come to this stage.

https://doi.org/10.23106/dcmi.953132608