Computer-assisted Indexing Complements Manual Selection of Subject Terms for Metadata in Specialized Collections

Suzanne Cady Stapleton, Chelsea S. Dinsmore, David Van Kleeck, and Xiaoli Ma

Discovery of digital items by scholars and the public is highly dependent upon effective metadata to ensure inclusion and prioritization in search engines. Subject descriptions based on controlled vocabulary, such as Library of Congress subject headings (LCSH), are particularly useful to enhance discovery, but they may be expensive to provide. In this case study we compared the results of computer-assisted indexing with manual selection of subject terms as part of an effort to enhance findability of journal information at the issue level. Our results suggest that incorporating both computer and human processes provide the highest impact for discovery of distinctive digital collections.

Introduction

Digitization provides libraries a way to offer equitable access to their most distinctive holdings. Extensive resources have been expended on the process of selecting, scanning, and adding content to digital library collections around the world. Today, the focus of digitization tends to be on curated collections featuring unique items. However, if the metadata describing these items is insufficient for successful precision and recall from search queries, all of the work to make the digital surrogates available may be wasted. If high value research materials cannot be found by users online, then those materials might as well still be sitting in closed stacks.

It is well established that metadata (in other words, bibliographic description) as the descriptive representation of an item is critical to the success of patrons' discovery and usage of digital items.[1] Even in full-text search environments, Zhang and Dimitroff found that metadata influences discovery through prioritization of results.[2] Interoperable metadata can be shared efficiently between online systems, allowing for an increase in discovery and potential use.[3] Enhanced metadata is particularly valuable when collections feature a GLAM (galleries, libraries, archives, and museums) institution's distinctive holdings. Expanded metadata may include additional subject headings, abstracts, chapter titles, and even transcriptions of manuscripts or

*Suzanne Cady Stapleton is Agricultural Sciences & Digital Scholarship Librarian, email: suzanne@ufl.edu; Chelsea S. Dinsmore is Chair, Digital Support Services, email: chedins@uflib.ufl.edu; David Van Kleeck is Chair, Cataloging and Discovery Services, email: dvankleeck@ufl.edu; Xiaoli Ma is Metadata Librarian, Digital Support Services, email xiaolima@ufl.edu. All authors are faculty with the University of Florida George A. Smathers Libraries.

audio resources. However, creating robust metadata is labor-intensive. As Kuny and Cleveland state, "One important consequence of the information revolution is that the costs of organizing information are beginning to match the costs of producing the information."[4] Assessment of an automation option for metadata development are the subject of this research project. For this study we asked the question: to what extent can the assignment of subject terms be automated for digital collections of specialized materials?

To explore this question, a team at the University of Florida George A. Smathers Libraries (Libraries) tested computer-aided subject indexing software. This software was selected because it is widely used in the publishing industry but had not yet been used in an academic library setting. The team identified an appropriate controlled vocabulary emphasizing natural language over hierarchical arrangement. Natural language tools were preferred to better match user expectations and user experiences, such as with the popular Google search engine.

The research team compared the computer-aided selection of subject terms based on full-text analysis with the results from humans using a weighted rubric to select subject terms for the same publication. The challenges and benefits of each method in this case study were analyzed. As a result, we recommend a workflow that integrates computer-aided indexing with manual indexing for best efficiency in increasing discoverability of digital library collections. Results of this study contribute to the discussion described by Golub et al. on incorporating tools for computer-assisted index term suggestions into human indexing workflow.[5]

Literature Review

Yang and Perrin maintain that "one of the most important duties for digital librarians is to evaluate indexing effects of search engines on their local digital resources. By doing so, digital librarians are able to adjust metadata strategies to improve the discoverability of digital content on the Internet."[6] This evaluation can suggest metadata standards that are streamlined for maximum efficiency of item description and for interoperability with various search platforms (such as use of Dublin Core, MARC, and XML MODS metadata element fields). Standard minimal metadata (like title, author/creator, publisher, year of publication) may be inadequate for maximum item discovery. As Fagan and Willey state, librarians can improve search engine optimization (SEO), particularly when working with undergraduates.[7] The subject (for instance, keyword phrases, category, and subject description) in an item's metadata are significant. Subject indexing of digital objects enhances their discoverability and usage.[8] In a year-long analysis of user searches in Texas Technology University's institutional repository, Yang determined that title, description, and subject were the most significant Dublin Core metadata elements for item discovery through internet searches.[9] In 73,341 visits from "organic" search engine traffic, user search terms (keywords) appeared in item metadata as title (74%), description (55%), and subject (20%), alone or in combination.[10] Gross and Taylor defend the ongoing need for subject headings in metadata.[11] They found that keyword search terms appeared exclusively in subject metadata fields in 35.9 percent of search results; thus, these results "would be lost in the absence of subject headings."[12] In 2015, Gross and Taylor replicated this research and found that, even with the inclusion of tables of contents and summaries/abstracts, "an average of 27% of hits would be lost if the subject headings were not present in the records. While the proportion of hits that would be lost in the absence of subject headings is reduced with the addition of contents and summary data, it still represents a significant proportion of total hits (more than one fourth)."[13] They also "found that when limited to English, the loss is 24.8%,

demonstrating that subject headings in English are, indeed, helpful in locating materials in other languages."[14] Moreover, Hider et al. found that reindexing the records in the Australian government's Office for Learning and Teaching Resource Library database with additional subject terms significantly increased both precision and recall in search results.[15]

Refining search results via the use of facets is a common search strategy in those catalogs and discovery platforms that feature faceted search results. Most libraries using these platforms configure them to include subject terms as one of the prime facets. Woods, Gillespie, and McManamon, while conducting a usability survey of the resource discovery platform for the University of Liverpool's Library Service, found that "the use of refining and limiting facets was also endemic, with 96% [of the survey respondents] advising that they refined their searches to some degree."[16] Even in next-generation library catalogs that feature streamlined centralized search boxes to appeal to users, enhanced subject metadata can support a "browsable, high-level topical overview of the information space" as recommended by Cuna and Angeli.[17] This points to the value of subject terms in catalog and digital collection platform records.

However, subject indexing is time-consuming. "Subject terms play a crucial role in resource discovery but require substantial effort to produce."[18] "Subject indexing has been long considered one of the most challenging tasks in information organization due to the difficulties in accurately, consistently, and comprehensively describing what a document is about according to the anticipated use of the document."[19] Thus, there is interest in exploring automation of subject indexing for digital collections.

Research and development of innovative computer-aided subject indexing is an active area of research. The debate over the role of automation by computers versus human involvement in subject indexing is also ongoing. Wu and Li, for instance, recommend automatic key phrase selection to enhance metadata for more efficient discovery.[20] Machine assistance for selection of appropriate subject terms offers potential increase in quality (precision and recall) and potential time and cost savings in the creation step.

> …subject indexing involves determining subject terms or classes under which a document should be found and what it could be used for; this goes beyond simply capturing what the document is about and is generally done better by people than by computer programs. However, automatic indexing algorithms that use rules learned from a good training set might find such terms, but human indexers who are not well trained might miss them.[21]

Whether computers or humans are employed in subject indexing, proper training is critical. Equally important is evaluation of digitization workflows, particularly those aspects influencing metadata.

This study compares the use of computer-aided indexing to human indexing methods for selection of subject terms to enhance metadata of a newly digitized specialized collection. The project serves as a test case of a new tool with the intention of using results to contribute to the development of best practices in metadata management for digital library collections.

Project Background

Land grant institutions in the United States continue the work begun in the early 1990s to digitize and preserve historic agricultural literature as a means to "capture the national char-

acter of Americans."[22] Caminita, Cook, and Paster describe the national effort and rationale to preserve and digitize agriculture and rural literature.[23] These efforts identified "core" historical literature through bibliometric analysis and consultations with subject matter specialists. The University of Florida participated as one of the nine original libraries in the National Preservation Program for Agricultural Literature and continues providing leadership in these efforts today.[24] In 2018, for example, the Libraries modeled cooperative digitization of significant historical agricultural literature with external publishers and material under copyright. With funding from Project Ceres, the Libraries completed digitization and print preservation of 54 years of the *Florida Cattleman & Livestock Journal*, a foundational publication to agriculture in the state and a nationally recognized publication. Published by the Florida Cattlemen's Association, the publication has strong ties to the University of Florida; every issue includes articles by University of Florida Institute of Food and Agricultural Sciences faculty sharing results of their research. The publication continues to be an important communication tool within the agricultural industry in Florida today. The *Florida Cattleman and Livestock Journal* is an example of specialized content in agricultural sciences, a field of importance to land-grant institutions. The digital content was added to the Florida Historical Agriculture and Rural Life collection (https://ufdc.ufl.edu/flag) of the University of Florida Digital Collection (UFDC).

The UFDC currently holds more than 14.8 million pages from more than 655,000 digital items, comprising nearly 170,000 different titles, many from highly specialized collections.[25] Metadata of published content is typically populated directly from MARC-formatted OCLC records, which are converted to METS/MODS format in the UFDC system.

This study explored one of the first uses of Access Innovation's Data Harmony Machine-Aided Indexer (M.A.I.) software in an academic library. Access Innovations is a software and metadata management company, which has provided semiautomated indexing in the publishing industry for more than 30 years. Machine-aided indexing systems use three methods to extract and recommend words or phrases: linguistic (criteria based on parts of speech), statistical (frequency of use), or rules-based (algorithms using if-then statements).[26] Data Harmony's M.A.I. uses a combination of programmed rules on top of criteria that extract terms based on part of speech or position within a sentence (a linguistic system). This system was selected for this study for its ability to return reliable results based on rules-based criteria.

Subject description terms identified using M.A.I. were compared to manual selection of subject terms, using the serial, *Florida Cattleman and Livestock Journal*. Once selected, additional subject description terms for each issue of this title are added to improve findability at the issue level. Future assessment to evaluate the impact of digital records with and without enhanced metadata on discovery and usage is warranted.

Materials and Methods

The Project Ceres award from the United States Agricultural Information Network (USAIN), Agriculture Network Information Collaborative (AgNIC), and the Center for Research Libraries (CRL) funded digitization of historic issues of the *Florida Cattleman and Livestock Journal*. Every issue in this monthly serial except one, from its inception in 1934 through 1988 was digitized and shared online for the first time. The magazine changed titles three times during this period. Issue sizes ranged from four pages in the early years to more than 100 pages, typical for issues published after the mid-1950s. This publication offers broad coverage of production topics such as cattle breeding, pasture improvement, identification and treatment of pests and

diseases, as well as marketing, policy, and food science. From the 616 historic issues digitized with funding by Project Ceres, a random sample of 35 issues were selected for this research project. Issues for the sample were identified with a random number generator from R.[27]

The research team evaluated options for an appropriate controlled vocabulary for this research using both a "top-down and a bottom-up approach."[28] For the top-down approach, search results for sample terms related to the content of this special collection were compared between existing vocabularies: a broad social studies and humanities thesaurus developed for JSTOR (https://www.jstor.org/) and the Agricultural Thesaurus of the National Agricultural Library (NAL Thesaurus, https://agclass.nal.usda.gov/agt.shtml). The NAL Thesaurus provides the controlled vocabulary for search engines such as AGRICOLA and PubAg. Of 54 sample production terms searched, JSTOR included 54 percent and NAL included 68 percent (excluding names of animal breeds, families, farms, or organizations). JSTOR was selected for the controlled vocabulary because of its natural language structure and the breadth of topics covered. The selection of JSTOR was also influenced by the intent to use the customized controlled vocabulary to enhanced metadata of electronic theses and dissertations, which cover a broad range of topics. The Libraries obtained permission to customize their own version of the JSTOR vocabulary for internal use. In the bottom-up approach to creating a relevant controlled vocabulary, terms of significance to the specialized collection were selected from the collection itself and through guidance with an advisory team.

A small advisory team composed of an animal scientist and an agricultural historian convened to work alongside the subject matter specialist librarian and student research assistant. Together, the team and researchers contributed expert knowledge in a variety of disciplines related to the topical coverage of the serial content: plant sciences and agricultural production, veterinary studies, anthropology/folklore/oral history, and animal science/meat science. Basic user profiles were developed to anticipate future use by UFDC patrons. The primary uses anticipated for the Cattleman's digital collection were family (or farm) genealogical research and animal studies/agricultural science research. The team studied topics covered in the historic issues and reviewed industry trends during this time period. Significant advances and challenges to the industry during this period were identified as topics of interest to agricultural scientists. For example, the passage of the fence laws in 1949 contributed to the successful eradication of the cattle fever tick *Rhipicephalus (Boophilus) annulatus* and *R. (B.) microplus* from Florida in 1962.[29] The fence laws changed the Florida cattle industry from open range management reliant on cattle branding to rotational grazing of improved pasture. A list of important terms related to identified topics was developed. Terms such as "fence law," "cattle fever tick," "open range," "range management," and "improved pasture" were identified as important.

Machine-aided Indexing Method

Digital object files were prepared for batch loading into Access Innovation's Data Harmony software, following a standardized naming convention for each .txt file. Reports were set up to generate the top 10 terms for each issue of the serial. M.A.I. reports display the preferred terms and their frequency in the full text, as well as any nonpreferred terms and their frequency. Four rounds of rule building were undertaken to develop the taxonomy for this project; in each round, specific terms were identified to be excluded, added, or in need of rule modification.

After reviewing an initial M.A.I. report on the terms that appeared most frequently in the research sample, rules were modified. This was the first custom rule-building in an iterative

process to reduce false-positive search results. The rules we built in the first round sought to exclude generic, overly broad terms and to define abbreviations of significant and frequent terms. We determined that very broad terms would not be useful as enhanced metadata since they would repeat concepts in the publication title and would be unlikely to identify topics specific to a particular issue of the serial. Very broad terms are unlikely to aid patrons in finding relevant issues of the serial. Therefore, rules were written to exclude very broad terms (such as Florida, cattle, livestock, journal, herd, steer, cow; agriculture, dairy, ranches, counties) from computer-aided indexing reports. One rule was developed in response to poor optical character recognition (OCR), whereby "tiles" appeared as a highly used term. Researchers determined that "tales," "the," and column breaks were misread as "tiles" by the OCR software. Initially, eight rules were modified based on an initial review of the machine-aided indexing reports of the 35 sample issues. Historic publications, in particular, may contain numerous challenges to OCR (see challenges discussed later).

The advisory team reviewed the JSTOR vocabulary and found it lacked many terms of specific importance to Florida's cattle industry. These terms were added to create a local, customized version of the thesaurus. Examples of production terms significant to Florida's cattle industry that were not found in the JSTOR thesaurus include "Bang's disease," "cattle fever tick," "cattle rustling," "forage legume," "improved pasture," "quarantine," and "selenium" (a mineral often deficient in Florida soils). Added terms came from the National Agricultural Library (NAL) thesaurus whenever possible. Select local terms of significance that do not appear in either JSTOR or NAL thesauri were identified and added, such as "Cracker cattle," "fat steer," "fence laws," "ticky deer," and "beef checkoff." A total of 219 terms were added to the customized JSTOR thesaurus for this project, resulting in a total of 57,113 preferred terms. Separate thesauri are under development at the Libraries for prominent names and geographical locations important to the state. More than 200 names identified in this project were made available for reuse in these other metadata enhancement projects.

In the second round of review of M.A.I. reports, we worked with the advisory team to identify preferred and related terms of significance. Separate rules were written to define these terms into the rules-based thesaurus. For example, "cattle fever tick" was set as the preferred term for "ticky deer" and "fever tick"; "brucellosis" was set as the preferred term for "Bang's disease" (*Brucella aborta*). "Clover" was identified as a related term for "forage legume." We built other rules to address the high incidence of erroneous terms in the M.A.I. reports (for example, an exclusion rule was written for "box" to prevent results related to post office addresses, prevalent in the advertisements and classified sections of this serial).

In the third round of rule building, custom rules were modified as we became more familiar with the impact of the rules. For example, exclusion of individual terms does not exclude phrases containing those terms in JSTOR (for instance, excluding "cattle" did not exclude "cattle ranching"; excluding "dairy" did not exclude "dairy industry"). Rules to exclude generic individual terms needed to be revised to include additional phrases with broad terms such as "cattle ranching" and "dairy industry." As Miller attests, "knowledge of the principles of controlled vocabularies and semantic relationships can help local implementers to create more robust and reusable vocabularies."[30] A total of 29 terms were excluded due to their generic nature.

In the final iteration of thesaurus rule development, the subject specialist worked with the metadata librarian to build more nuanced and informed rules. For example, although beef

was excluded as too broad a term to be useful, "beef checkoff" was added as an important concept representing a national beef promotional campaign launched during the period covered by this serial. The term "grasses" was removed from "pasture" because grasses represent a separate subset of pasture plants; not all pasture plants are grasses. Improving pasture by incorporating legume plants (such as clover) was a significant subject covered by this specialized collection with the goal of improving cattle nutrition, beef quality, and sales of Florida beef. We added the terms "improved pasture" and its wild predecessors "open range" and "native range." Definitions of locally important terms were expanded to include, for example, "fat stock" for finishing cattle.

After each round of rule building, a new report of the top 10 terms generated by the machine-aided indexing was reviewed. In the end, four rounds of custom rule-building were deployed in this study. In sum, a total of 218 terms were added, 63 terms excluded, and 37 rules modified in the thesaurus. While these changes were important for this specialized collection, they involved only a fraction of terms in the original JSTOR thesaurus of 56,895 terms. From the M.A.I. report, the top five terms for each issue in the random sample were reviewed before being added to the issue metadata as additional subject indexing terms. Best practices informed our decision to limit additional subject terms to between five and seven; Library of Congress recommends a maximum of six subject headings for most cases.[31]

FIGURE 1
Excerpt of Manual Rubric Score Sheet for Two Sample Serial Issues, Vol 10(9) and Vol 23(3)

Year	Month	Volume	Issue	Subject term or phrase	Cover (+3)	TOC (+2)	FL Importance (+4)	Issue Theme (+2)	Repetitive (-2)	Preferred subject term or phrase	Sum
1946	Jun	10	9	Everglades	1		1				7
1946	Jun	10	9	Fat Cattle	1						3
1946	Jun	10	9	Hoof, Mouth	1		1	1		Hoof and Mouth disease	9
1946	Jun	10	9	Pasture	1		1	1		Improved pasture	9
1946	Jun	10	9	Escambia	1						3
1946	Jun	10	9	Moultrie	1						3
1946	Jun	10	9	Flood Control	1		1	1			9
1946	Jun	10	9	Marks and Brands			1	1			6
1946	Jun	10	9	Tannery			1				4
1946	Jun	10	9	Quarter Horse Association			1			Florida Quarter Horse Associatio	4
1946	Jun	10	9	Durrance Ranch			1				4
1958	Dec	23	3	Equipment	1	1		1		farm automation	7
1958	Dec	23	3	Florida Dairyman	1		1	1		Florida Dairy Farmers Federation	9
1958	Dec	23	3	Taxes	1			1		land value	5
1958	Dec	23	3	Bang's	1	1	1			Brucellosis	9
1958	Dec	23	3	Adams, Jr., Alto	1		1			Adams, Jr., Alto	7
1958	Dec	23	3	Hereford sale	1						3
1958	Dec	23	3	Cammack, Bill	1						3
1958	Dec	23	3	Boyd and Hall Dairy	1		1				7
1958	Dec	23	3	Swine Anemia		1					2
1958	Dec	23	3	IDFA		1				Independent Dairy Farmer's Asso	2
1958	Dec	23	3	Koger, J. French		1					2
1958	Dec	23	3	Watershed Project		1	1				6
1958	Dec	23	3	Nielsen, Alf R.		1					2
1958	Dec	23	3	Melear, V.B. "Boots"		1					2
1958	Dec	23	3	Callahan		1					2
1958	Dec	23	3	Quarter horse		1					2

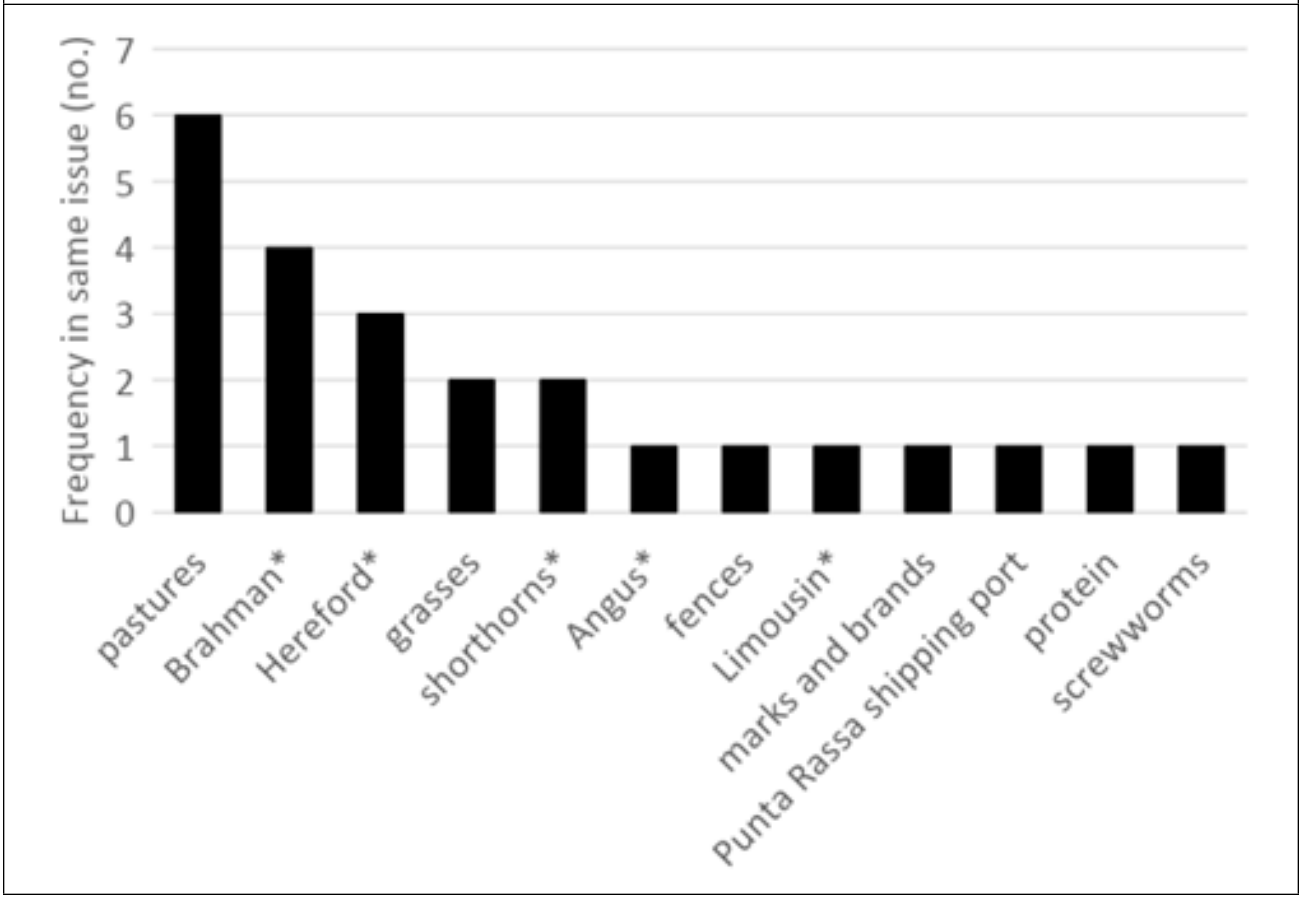
Note: Each subject term or phrase is identified as a name (blue highlight) or production term (green highlight). The provenance and importance of each term or phrase is identified by issue front cover (cover), table of contents (TOC), importance in Florida (FL importance), theme of serial issue (Issue theme), whether the term is repetitive (repetitive). Points are assigned to each term based on its provenance and importance. Preferred terms are noted (preferred subject term or phrase) and the sum of scores for each term is listed.

Manual [Human] Rubric Method

An alternative method to machine-assisted selection of subject terms to use as enhanced metadata for issues of the *Florida Cattleman & Livestock Journal* relied upon manual use of a rubric. We developed a rubric to standardize the manual selection of subject terms. The rubric is a scoring system whereby terms are weighted by their provenance (see sample excerpt in figure 1). Each time a term appeared on an issue cover, for example, it was assigned three points and two points for each time it appeared in the table of contents. Beginning in 1953, issues of the *Florida Cattleman & Livestock Journal* featured a theme. For example, the theme for each January issue was Brangus breed of cattle, a cross between Brahman and Angus developed to better handle Florida's subtropical climate.

Two researchers applied the rubric to the 35 sample issues. Terms were selected and color coded as either Names or Production terms. Names included people, places, farms/ranches, animals, and organizations. During the period covered, many industry support organizations launched, for instance, the state-sponsored livestock markets for cattle sales, the Florida Quarter Horse Association, and new research facilities. Production terms included concepts related to all aspects of cattle production from calving, feeding and nutrition, pasture management, disease management, to marketing, sales, and beef consumption, as well as related industries including horse, swine, dairy and topics celebrating rural life such as rodeos and festivals.

FIGURE 2
Twelve Terms Appeared in the Top Five Ranking in Both Machine-aided Indexing and Manual Rubric Reports across 35 Random Sample Issues of the *Florida Cattleman & Livestock Journal*



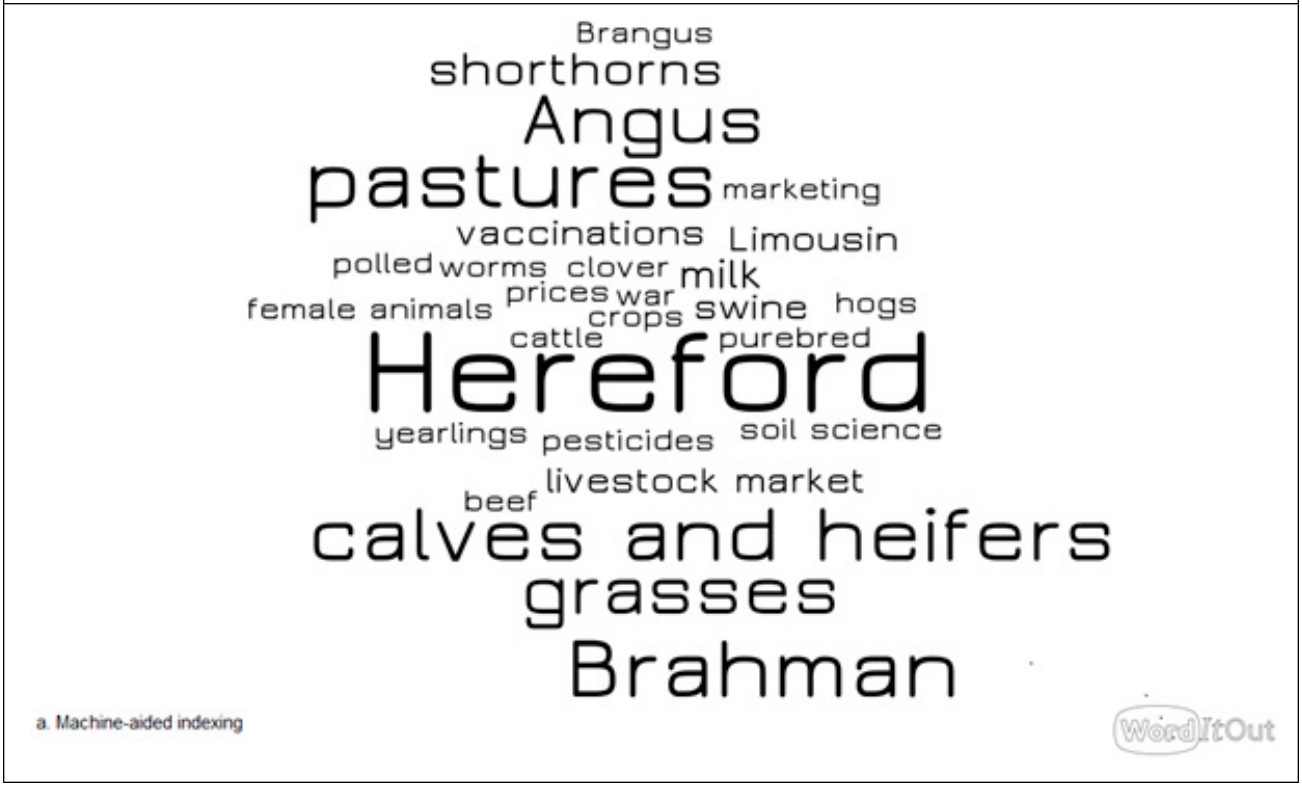
From the manual rubric, the top five scoring production terms for each issue in the random sample were reviewed before being added to the metadata for each of the 35 sample issues. All terms with tied scores were included. As this was a proof-of-concept pilot project, the team chose not to address the amount of time and effort required in the evaluative process. Limitations in the ability to evaluate results within the UFDC platform did not impact the researchers' ability to compare results returned by computer-aided selection relative to results from human review. Having uncovered some shortcomings within the platform, including returning issue-level results, the research team now has a plan of action to investigate solutions that will allow for further evaluation, particularly in cost and effectiveness.

Results: Comparison of Machine-aided Indexing and Manual Rubric Methods

There were 24 instances where a subject term ranked in the top five for the same serial issue in both M.A.I. and manual rubric reports. These 24 instances included 12 subject terms across the 35 sample issues (see figure 2). Five of these terms refer to specific cattle breeds (Brahman, Hereford, Shorthorn, Angus, and Limousin). For five serial issues, the top-ranked term was common for both subject selection methods: Brahman, Hereford, and pastures.

From this sample of 35 issues, M.A.I. reports generated 36 unique terms ranked in the top five (see figure 3a). The manual rubric generated 125 unique terms ranked in the top five from the same sample (see figure 3b). M.A.I. reports are produced by computer-aided review of term frequency in full text in accordance with controlled vocabulary and its associated rules for excluding or substituting a preferred or related term. The frequency of the top five terms

FIGURE 3a
Word Clouds of Unique Subject Terms Generated from Random Sample of Thirty-Five Issues of the *Florida Cattleman & Livestock Journal* from (a) Machine-Aided Indexing and (b) Manual Rubric Method



a. Machine-aided indexing

FIGURE 3b
Word Clouds of Unique Subject Terms Generated from Random Sample of Thirty-Five Issues of the *Florida Cattleman & Livestock Journal* from (a) Machine-Aided Indexing and (b) Manual Rubric Method

in the M.A.I. reports varies, but there is a significant focus on a limited number of terms (see figure 3a). Comparison of figure 3a and figure 3b reveal the differences in the two methods of subject term selection. The manual rubric generates much greater variety of terms than the M.A.I. reports. The M.A.I. method efficiently identifies the most frequently used terms, while a manual rubric highlights the most significant (important) subject terms.

Discussion

> Descriptive metadata is important because it provides *intellectual access* to the contents of a digital collection… [it serves] the two overarching functions of *identification* and *retrieval*…. It allows users to identify the content, context, and meaning of digital resources, both individually and in relation to one another, and it allows users to retrieve individual resources and sets of related resources based on any number of shared characteristics. Metadata supports both the searching and browsing methods of information retrieval… without metadata, the [digital] collection would be virtually useless.[32]

Rich subject indexing is important to the discovery and use of digital special collections. This project compared the use of computer-aided indexing against a manual rubric to select five to seven additional subject terms for metadata enhancement of digital objects. The objective was to evaluate and compare the results of these two methods to identify additional subject

terms to aid discovery of digital objects in an academic library digital specialized collection. The project also served as one of the first test cases of Access Innovation's Data Harmony software suite in an academic library setting. A distinctive collection of digital content was selected for this project, as the extra commitment to enhance metadata may be most likely to be employed in these themed, unique collections.

Training was necessary to become familiar with Access Innovations Data Harmony software suite, a new tool for librarians at the Libraries. The original research project with objective measurement of impact was modified by necessity in response to technological challenges encountered. The research team also pursued one iteration of the research method, to compare only the front cover plus table of contents of each issue in both machine-aided indexing and manual rubric, only to abandon this approach later. To measure the true power and gain maximum benefit of computer-aided indexing, the full text of each issue was included in the machine-aided indexing process.

M.A.I. can report the frequency of terms in a weighted word count, prioritizing terms that appear in the title or abstract, for example. Weighting of terms avoids problems, as Zhang, Smith, Twidale, and Gao attest, "as [digital] collections grow and more federated searching is carried out, the absence of weights for subject terms can cause problems in search and navigation."[33] No abstracts are published in the *Florida Cattleman & Livestock Journal*, so term preference by M.A.I. in this study occurred only as a result of thesaurus rules applied to the full text. With the manual rubric, terms were weighted in an attempt to efficiently capture the important topics in the issue based mainly on the provenance of terms, with priority assigned to terms on the front cover and table of contents. However, provenance of terms was not the only criteria in this method. Terms were also weighted heavier if they represented the issue theme or an important topic for the industry as identified by the Advisory Team. It was not uncommon for colloquialisms or term abbreviations to be used on the cover or table of contents, thus requiring consultation with the corresponding article(s).

This method corroborated the team's premise that machine-aided indexing is valuable in speeding up the process for identifying relevant terms but does not replace the need for human review. The process increases efficiency in that it reduces time for the initial steps of the overall application and review of subject terms.

Specialized collections cover specific content with unique terms that may need to be added to a thesaurus. Miller notes that "it is not uncommon in designing and creating metadata for digital collections that there is no established controlled vocabulary that meets local needs," which leads to development of local controlled vocabularies.[34] Baxter, Trott, and Hale identified "distinguishing characteristics" during metadata creation of a historic agricultural serial and assigned subject terms for prominent topics to "encourage patrons outside of the field of agriculture to intentionally and serendipitously discover these materials."[35] This project contributes to current conversations on best practices for metadata creation and enhancement of digital objects. Adding terms based on keywords from the title to existing LCSH has been recommended.[36]

This study illustrates the differences resulting from machine-aided indexing and a manual rubric method of selecting relevant subject terms to describe digital material. The importance of subject terms in bibliographic description (metadata) for improved discovery is established.[37] Use of these terms in both the item metadata and the search interface is needed for best use.[38] There is an added cost to enhancing the metadata of a digital item; thus, efficiencies in these

endeavors are of interest to all digital collection managers. Highly specialized content, such as found in digital special collections, requires additional preparation of an appropriate thesaurus to be able to select subject terms of significance to users. Modifying an existing thesaurus requires more investment of humans before the benefits of automatic indexing can be reaped.

The large differences in terms identified in the two methods employed in this study illustrate the influence of subject term selection methodology. We argue that the two methods employed will provide best results when used together. The machine-aided indexing provides an efficient means to highlight terms that appear with the most frequency in full text. The manual rubric, weighting subject terms by provenance and incorporating subject knowledge expertise to favor terms of significance, resulted in a greater number of significant subject terms. Although the M.A.I. terms are weighted by rules in the customized thesaurus, we found that a final review of M.A.I. terms by subject specialists is warranted. The manual rubric provided richer, more nuanced content, which was better able to describe changes in the history of the Florida cattle industry over the 50-year time period covered by this digital collection. A combined approach where the efficiencies of computer software to analyze full text is harnessed and final selection of terms is conducted by human subject matter experts is advised.

The process of developing such a workflow is proposed and appropriate for any specialized collection. The first step in the process is to assemble an advisory team with subject matter expertise, then select the most appropriate thesaurus, or thesauri, for the goals of the project. Develop a manual rubric to weight terms of significance to the specialized collection in question. Separately, run an initial M.A.I. report on a random sample of the specialized collection. We recommend generating reports with the top 10 to 15 terms per digital item. Use the first run to identify any problems with OCR and then refine results, focusing on their utility to researchers. Results from the two methods can be compared to help further develop thesaurus rules, as needed. Use an iterative process to build rules into the thesaurus to elucidate appropriate terms. Finally, employ subject matter experts to review the top-ranked terms from the M.A.I. method. Adjust the most frequently used terms (M.A.I. method) if needed to include the most significant terms (manual rubric method). Add the top five to seven terms as subject description terms into the metadata for each digital item.

The recommended process, then, is machine-assisted indexing with human-augmentation. Discernment from humans is still critical to identification of relevant subject terms. Creating an advisory team of subject experts with a range of perspectives on the content will improve the process of selecting subject terms. Close communication among the team is important to bring together expertise of indexers and subject matter experts into the ontological development. Baxter, Trott, and Hale describe a team of subject specialist librarians working with cataloging and digital librarians to digitize and make accessible historic agricultural serials from the University of Tennessee.[39] Testing and training on the impacts of customized rules is important. Close consultation with an ontologist would be useful to develop rules appropriate for the controlled vocabulary of a local project. Review of metadata as an essential ongoing process, rather than an infrequent large-scale endeavor, is advised.[40] Diagle imagines robust metadata structured in contextual layers to emphasize improved user experience of digital special collections, providing additional information only as the user requests.[41]

Of the several challenges encountered during this project, three are worth mentioning. This project prompted the project team to reevaluate and replace existing tools for Optical Character Recognition (OCR) of the newly digitized text. In this specialized collection, content

design and layout along with binding impediments proved challenging for the OCR generating system (Prime). The *Florida Cattleman* serial, especially the historic issues, includes decorative font, diagonal text, columns, images with embedded captions, and colloquialisms. These aspects, which may not be uncommon for collections of this period, created challenges for machine-reading. As a result of this study, the Libraries' Digital Support Services team plans to undertake a study to compare the OCR creation software tools, Abby Fine and Tesseract, against Prime OCR recognition software. Testing a sample of text from the body of work to be studied in the designated OCR recognition software is recommended as an initial project step in digital initiatives of content that was not born digital.

A second significant challenge encountered in this study is the subjectivity of subject determination. Subjectivity of indexing is well documented in the literature as catalogers attempt to summarize the significance of an item for anticipated future use.[42] Additionally, indexer bias in assigning descriptive terms may also be a concern.[43] Development and application of a rubric helped provide specific parameters to standardize human assessment of appropriate subject indexing terms. Yet, we found that, even with the rubric, prior experience influenced each human's perspective and interpretation of significance. In this study, the veterinary studies research assistant was more likely to select terms related to pathology and treatment of animals while the plant scientist prioritized pasture improvement. Thus, it is recommended to assemble an advisory team with broad, representative backgrounds.

Finally, a third challenge researchers faced was determining when the customized thesaurus was ready to use. In fact, thesauri are never complete, yet use of them must move forward. New terms emerge and the use and meaning of terms change over time. In this study, for example, "fat stock" is now referred to as a "finished animal," so vocabulary rules must always be updated. The Library of Congress, for example, undergoes routine revisions of subject headings to keep library catalogs current.[44] For transparency and reproducibility of research, we highly recommend capturing the thesaurus employed at the time of a study for reference.

This study was unusual in its aim to focus on issue-level metadata for a specialized collection serial. Recommended practices from this study are more easily applied to special monograph collections, such as electronic theses and dissertations. Since this study, the Libraries have expanded their capacity in metadata and computer programming support. Efforts are underway to address technological constraints encountered that restricted the ability to provide issue-level results and thus impeded our ability to capture measurable impact in this study. Knowledge and skills gained from this project are aiding the Libraries in development of processes to incorporate and maintain the local controlled vocabulary for use across the digital library.

Conclusion

This study employed a new tool for academic libraries to enhance the metadata of a newly digitized collection. Based on this study, we recommend a collaborative indexing workflow between computer automation and humans to maximize the inherent powers of each. This combination maximizes the power of computer automation to assess full text with the discernment of subject matter experts to develop an appropriate thesaurus and prioritize the final selection of subject descriptions. The findings also reveal a number of challenges that can inform future similar projects. After subject term selection, automated processes can be used to add the identified subject terms into metadata. In this manner, digital collections of unique content can be described most effectively for greater discovery and usage. These findings con-

tribute to special collections management as well, where evidence supports "more product, less processing" to reveal hidden collections with minimally sufficient descriptions that meet user preferences to search for "aboutness" of digital content through search engines.[45] Our research may also contribute to current efforts to provide ethical subject headings. In 2016, for example, the Library of Congress decided to revise the subject heading "illegal alien" in recognition of its "pejorative tone in recent years."[46] The American Library Association's Subject Analysis Committee Working Group on Alternatives to LCSH "Illegal aliens" is developing recommendations for libraries to implement subject term changes on their own or within their consortial catalog. Our research may offer solutions to challenges encountered in library efforts to use inclusive and respectful subject headings such as selecting alternative vocabulary/terminology, identifying inconsistencies across records and addressing workload/staffing issues.[47] This research may also contribute to discussion of best approaches to develop semantically rich subject metadata in library catalog and discovery services. A similar discussion of how to incorporate both the high-level reasoning of human indexers with powerful algorithms to quantitively process information automatically in library catalogs is underway.[48] The ability to develop welcoming user interfaces situated within hierarchical information organizations to support complex searches may rely heavily on rich subject metadata.

This project explores elements of machine learning (ML), particularly in regard to automated and human text analysis workflows to improve information retrieval through relevant subject descriptors. Publishers are actively developing a variety of ML tools, some that may prove useful to libraries in time, such as search platforms powered by natural language processing and recommendations pointing users to related information items. Librarians, as experienced information professionals, can contribute invaluable experience to these developments. Challenges to these developments in library settings include the traditional silos of computer science expertise, publishing, and library research; adequate financial resources; and a library culture of strict user data privacy.

Results of our research support the recommended best practice to incorporate computer-aided indexing tools with human subject matter experts for selection of subject terms to most effectively enhance metadata of digital specialized collections. Conclusions of this research contribute to ongoing conversations on how to efficiently and effectively enhance metadata of digital collections for greater discovery. Adding subject metadata to improve digital item discovery is an ongoing need, particularly important as GLAM institutions shift their focus away from mass digitization to feature distinctive collection materials.

Acknowledgments

The authors wish to acknowledge technical support from Robert Phillips, Application Programmer II, George A. Smathers Libraries and Access Innovations; student research assistant Lauren Cooney; and guidance on controlled vocabulary offered by advisory team members Robert L. Stone, Outreach Coordinator, Florida Cattlemen's Foundation and Dwain Johnson, Professor Emeritus, Department of Animal Sciences, University of Florida Institute of Food and Agricultural Sciences.

Declaration of Interest Statement:

Authors declare they have no conflicting or competing interests with this research.

Notes

1. Frederick Wilfrid Lancaster, *Indexing and Abstracting in Theory and Practice*, 3rd ed. (Champaign, IL: University of Illinois, 2003).

2. Jin Zhang and Alexandra Dimitroff, "The Impact of Metadata Implementation on Webpage Visibility in Search Engine Results (Part II)," *Information Processing and Management* 41 (2005): 691–715.

3. Eduardo M. Corrado and Rachel Jaffe, "Transforming and Enhancing Metadata for Enduser Discovery: A Case Study," *Italian Journal of Library, Archives, and Information Sci*ence 5, no. 2 (2014): 33–48, https://doi.org/10.4403/jlis.it-10069; Stephanie C. Hass et al., "Darwin and MARC: A Voyage of Metadata Discovery," *Library Collections, Acquisitions, and Technical Services* 27, no. 3 (2003): 291–304, https://doi.org/10.1016/S1464-9055(03)00071-X.

4. Terry Kuny and Gary Cleveland, "The Digital Library: Myths and Challenges," *IFLA Journal* 24, no. 2 (1998):107–13, https://www.ifla.org/files/assets/hq/publications/ifla-journal/archive/jour2402.pdf.

5. Koraljka Golub et al., "A Framework for Evaluating Automatic Indexing or Classification in the Context of Retrieval," *Journal of the Association for Information Science and Technology* 67, no. 1 (2016): 3–16, https://doi.org/10.1002/asi.23600.

6. Le Yang and Joy M. Perrin, "Introduction to the Special Issue of Digital Collection Metadata & Internet Discovery," *Journal of Web Librarianship* 11, no. 3/4 (2017): 153, https://doi.org/10.1080/19322909.2017.1383136.

7. Jody Condit Fagan and Malia Willey, "The Discoverability of Award-winning Undergraduate Research in History: Implications for Academic Libraries," *College & Undergraduate Libraries*, 25, no. 2 (2018): 164–86, https://doi.org/10.1080/10691316.2018.1456994.

8. Jung-Ran Park, "Metadata Quality in Digital Repositories: A Survey of the Current State of the Art," *Cataloging & Classification Quarterly* 47, no. 3/4 (2009): 213–28, https://doi.org/10.1080/01639370902737240; Ngoc-My Guidarelli, "Subject Data in the Metadata Record," *Library Collections, Acquisitions, and Technical Services* 24, no. 4 (2000): 499–500, https://doi.org/10.1016/S1464-9055(00)00179-2; Dan Dorner, "Cataloging in the 21st Century—Part 2: Digitization and Information Standards," *Library Collections, Acquisitions, and Technical Services* 24, no. 1 (2000): 73–87, https://doi.org/10.1016/S1464-9055(99)00099-8; Kuang-Hwei (Janet) Lee-Smeltzer, "Finding the Needle: Controlled Vocabularies, Resource Discovery, and Dublin Core," *Library Collections, Acquisitions, & Technical Services* 24 (2000): 205–15; Jin Zhang and Alexandra Dimitroff, "Internet Search Engines' Response to Metadata Dublin Core Implementation," *Journal of Information Science* 30, no. 4 (2004): 310–20, https://doi.org/10.1177/0165551504045851.

9. Le Yang, "Metadata Effectiveness in Internet Discovery: An Analysis of Digital Collection Metadata Elements and Internet Search Engine Keywords," *College & Research Libraries* 77, no. 1 (2016b): 7–19, https://doi.org/10.5860/crl.77.1.7.

10. Yang, "Metadata Effectiveness in Internet Discovery," 15.

11. Tina Gross and Arlene G. Taylor, "What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results," *College & Research Libraries* 66, no. 3 (2005): 212–30, https://doi.org/10.5860/crl.66.3.212.

12. Gross and Taylor, "What Have We Got to Lose?" 219.

13. Tina Gross, Arlene G. Taylor, and Daniel N. Jourdrey, "Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching," *Cataloging & Classification Quarterly* 53, no. 1 (2015): 1–39, https://doi.org/10.1080/01639374.2014.917447.

14. Gross, Taylor, and Jourdrey, "Still a Lot to Lose," 31.

15. Philip Hider et al., "Reindexing a Research Repository from the Ground Up: Adding and Evaluating Quality Metadata," *Australian Academic & Research Libraries* 47, no. 2 (2016): 61–75, https://doi.org/10.1080/00048623.2016.1204589.

16. Jeff Woods, Elizabeth Gillespie, and Catherine McManamon, "Discovering Discovery: Lessons Learnt from a Usability Study at the University of Liverpool," *Insights: The UKSG Journal* 29, no. 3 (2016): 258–65, https://doi.org/10.1629/uksg.320.

17. Andrea Cuna and Gabriele Angeli, "Improving the Effectiveness of Subject Facets in Library Catalogs and Beyond: A MARC-Based Semiautomated Approach," *Library Hi Tech*, ahead of print (2020), https://doi.org/10.1108/LHT-07-2019-0132.

18. Golub, "A Framework for Evaluating Automatic Indexing or Classification in the Context of Retrieval," 3.

19. Kun Lu, Jin Mao, and Gang Li, "Toward Effective Automated Weighted Subject Indexing: A Comparison of Different Approaches in Different Environments," *Journal of the Association for Information Science and Technology* 69, no. 1 (2018): 121–33, https://doi.org/10.1002/asi.23912.

20. Yi-fang Brook Wu and Li Quanzhi, "Document Keyphrases as Subject Metadata: Incorporating Document Key Concepts in Search Results," *Information Retrieval* 11 (2008): 229–49.

21. Golub, "A Framework for Evaluating Automatic Indexing or Classification in the Context of Retrieval," 6.

22. Nancy E. Gwinn, "A National Preservation Program for Agricultural Literature," *Journal of Agricultural & Food Information* 2, no. 2 (1994): 25–49.

23. Cristina Caminita, Michael Cook, and Amy Paster, "Thirty Years of Preserving, Discovering, and Accessing U.S. Agricultural Information: Past Progress and Current Challenges," *Library Trends* 65, no. 3 (2017): 293–315.

24. Caminita, Cook, and Paster, "Thirty Years of Preserving, Discovering, and Accessing U.S. Agricultural Information," 293–315.

25. University of Florida George A. Smathers Libraries, "University of Florida Digital Collections," available online at https://ufdc.ufl.edu/ [accessed 6 July 2020].

26. Larry Compton, "The Search for Machine-Aided Indexing: Why a Rule-Based System Is the Cost-Effective Choice," *Access Innovations*, available online at https://www.accessinn.com/the-search-for-machine-aided-indexing-why-a-rule-based-system-is-the-cost-effective-choice/ [accessed 20 July 2020].

27. The R Project for Statistical Computing, available online at https://www.r-project.org/.

28. Steven J. Miller, *Metadata for Digital Collections: A How-to-do-it Manual* (London, UK: Neal-Schuman Publishers, Inc., 2011).

29. *Florida Cattle Ranching: Five Centuries of Tradition*, eds. Robert L. Stone and Susannah S. Booth (Kissimmee, FL: Florida Cattlemens Foundation, Inc., 2013).

30. Miller, *Metadata for Digital Collections*, 147.

31. *Basic Subject Cataloging Using LCSH: Instructor's Manual*, ed. Lori Robare (Washington, DC: Association for Library Collections and Technical Services, 2011).

32. Miller, *Metadata for Digital Collections*, 9–10.

33. Hong Zhang et al., "Seeing the Wood for the Trees: Enhancing Metadata Subject Elements with Weights," *Information Technology and Libraries* 30, no. 2 (2011): 75–80.

34. Miller, *Metadata for Digital Collections*, 144.

35. Isabella J. Baxter, Louisa Trott, and Meredith Hale, "Coordinating Expertise to Preserve and Increase Discoverability of Key University of Tennessee Agricultural Serials," *Journal of Agricultural and Food Information* 21 (2019): 15–31.

36. *Basic Subject Cataloging Using LCSH*, ed. Robare.

37. Gross and Taylor, "What Have We Got to Lose?"; Yang, "Making Search Engines Notice: An Exploratory Study on Discoverability of Dspace Metadata and Pdf Files," *Journal of Web Librarianship* 10, no. 3 (2016a): 147–60.

38. Zhang, "Seeing the Wood for the Trees," 75–80.

39. Baxter, Trott, and Hale, "Coordinating Expertise to Preserve and Increase Discoverability of Key University of Tennessee Agricultural Serials," 5.

40. A.L. Carson and Carol Ou, "Metadata Revisited: Updating Metadata Profiles and Practices in a Vendor-Hosted Repository," *Library Resources & Technical Services* 63, no. 4 (2019): 204.

41. Bradley J. Diagle, "The Digital Transformations of Special Collections," *Journal of Library Administration* 52 (2012): 244–64.

42. Ronald E. Day, *Indexing It All: The (Subject) in the Age of Documentation, Information and Data* (Cambridge, MA: MIT Press, 2015); Frederick Wilfred Lancaster, *Indexing and Abstracting in Theory and Practice*, 3rd ed. (Champaign, IL: University of Illinois, 2003).

43. David Bodoff and Yaffa Richter-Levin, "Viewpoints in Indexing Term Assignment," *Journal of the Association for Information Science and Technology* 71, no. 4 (2020): 450–61.

44. Library of Congress, "Process for Adding and Revising Library of Congress Subject Headings," *The Library of Congress Cataloging and Acquisitions*, Library of Congress https://www.loc.gov/aba/cataloging/subject/lcsh-process.html [accessed 30 November 2020].

45. Jennifer Schaffner, "The Metadata *Is* the Interface: Better Description for Better Discovery of Archives and Special Collections, Synthesized from User Studies," *OCLC Research Report*, Online Computer Library Center, Inc. (2009), https://www.oclc.org/content/dam/research/publications/library/2009/2009-06.pdf.

46. Library of Congress, "Library of Congress to Cancel the Subject Heading 'Illegal Aliens'," Library of Congress report, Library of Congress (March 22, 2016), https://www.loc.gov/catdir/cpso/illegal-aliens-decision.pdf.

47. SAC Working Group on Alternatives to LCSH "Illegal aliens," "Report of the SAC Working Group on Alternatives to LCSH 'Illegal aliens'," American Library Association's Association for Library Collections and Technical Services Division (June 19, 2020), https://alair.ala.org/bitstream/handle/11213/14582/SAC20-AC_report_SAC-Working-Group-on-Alternatives-to-LCSH-Illegal-aliens.pdf?sequence=1&isAllowed=y.

48. Cuna and Angeli, "Improving the Effectiveness of Subject Facets in Library Catalogs and Beyond."