ARTICLE

Text Analysis of Archival Finding Aids

Collection Scoping and Beyond

Anne Bahde and Cara Key

ABSTRACT

Archival repositories must be strategic and selective in deciding what collections they will acquire and steward. Careful collection stewards balance many factors, including ongoing resource needs and future research use. They ensure new acquisitions build upon existing topical strengths in the repository's holdings and reassess these existing strengths regularly through multiple lenses. In this study, we examine the suitability of text analysis as a method for analyzing collection scope strengths across a repository's physical archival holdings. We apply a tool for text analysis called Leximancer to analyze a corpus of archival finding aids to explore topical coverage. Leximancer results were highly aligned with the baseline subject heading analysis that we performed, but the concepts, themes, and co-occurring topic pairs surfaced by Leximancer suggest areas of collection strength and potential focus for new acquisitions. We discuss the potential applications of text analysis for internal library use including collection development, as well as potential implications for wider description, discovery, and access. Text analysis can accurately surface topical strengths and directly lead to insights that can inform future acquisition decisions and archival collection development policies.

INTRODUCTION

Archival repositories must be strategic and selective in deciding what collections they will acquire and steward. Careful collection stewards balance many factors, including ongoing resource needs and future research use. They ensure new acquisitions build upon existing topical strengths in the repository's holdings and reassess these existing strengths regularly through multiple lenses. Collection analysis and assessment are critical tools to ensure ethical stewardship, to better discern collection scopes and new directions, and to better recognize whose voices are missing from the materials.[1]

Text analysis is a machine learning (ML) technique used to identify patterns and trends across large sets of unstructured data. Natural language processing (NLP), which supports text analysis techniques, enables ML for human language. Automated text analysis facilitates deeper comprehension of natural language in large corpora, interprets qualitative texts at a macro scale, and limits the inherent subjectivity in both composing and reading texts. As digital representations of physical materials that may be otherwise digitally inaccessible, archival finding aids hold great potential as a corpus for text analysis. The structured data found in Encoded Archival Description (EAD) finding aids contain natural language narrative elements. These include the archivist's description of a collection's history, scope, and contents; folder titles assigned by the creator or archivist generalizing the aggregated content within the folder; and item-level narrative details about visual or other materials. Taken together, these narrative fields

About the Authors

Anne Bahde (corresponding author) (anne.bahde@oregonstate.edu) is Special Collections Librarian for Research and Learning, Oregon State University. **Cara Key** (cara.key@oregonstate.edu) is Digital Repository Librarian, Oregon State University. © 2024.

Submitted: 29 February 2024. Accepted for publication: 11 July 2024. Published: 16 December 2024.

https://doi.org/10.5860/ital.v43i4.17065

present a rich textual corpus, related to but distinct from controlled fields such as subject headings.

In this study, we examine the suitability of text analysis as a method for analyzing collection scope strengths across a repository's physical archival holdings. We apply a tool for text analysis called Leximancer to analyze a corpus of archival finding aids to explore topical coverage. We then discuss the potential applications of text analysis for internal library use including collection development, as well as potential implications for wider description, discovery, and access.

BACKGROUND

The Special Collections and Archives Research Center (SCARC) at Oregon State University Libraries and Press was established in 2011 after a merge of the former Special Collections and University Archives departments. SCARC's early collecting scope was influenced by the previous emphases of these two units. The initial "signature areas" of collecting strength were determined to be university history, natural resources, history of science, and Oregon multicultural communities. Collecting attention was focused in those areas, but some were not fully defined or scoped at the time. In 2013, a further emphasis was added with the founding of the Oregon Hops and Brewing Archives; in 2014, the Oregon State Queer Archives was added.[2] In the ensuing decade, a robust influx of new accessions was acquired in these signature areas and others.

Recently, SCARC has begun work to rewrite its repository-level collection development policy. To date, policy drafting work has included assessing collection pursuits and new offers in a team setting to fully draw on the vast and rich collective knowledge carried by SCARC staff members. Additionally, data regarding new accessions since the 2011 merge has been gathered, discussed, and informally analyzed. The present text analysis project can contribute to these ongoing conversations as one method for synthesizing data about collection content.

LITERATURE REVIEW

Machine Learning and Text Analysis in Libraries and Archives

Practitioners in library and information science have long utilized text analysis and other ML methods for a wide variety of studies involving natural language data. Text analysis results can reveal hidden patterns and relationships among large and potentially overwhelming amounts of text and are often accompanied by visualizations that allow fuller interpretation.

Results from text analysis and related methods can inform library programming, collecting, and many other operations. Using a corpus of titles, abstracts, and subject headings extracted from more than 400 children's books, Joo, Ingram, and Cahill used a variety of analysis approaches including term frequency, bi-gram analysis, topic modeling, and sentiment analysis to conclude that library storytimes should be centered around a specific set of topics of interest.[3] Harden used topic modeling on a corpus of nearly 2,000 written student responses to gauge first-year student comprehension of concepts in the Framework for Information Literacy, revealing a deeper student connection to the material than had been supposed.[4] Sharma, Barrett, and Stapelfeldt used text analysis on library chat reference transcripts, finding a wealth of insights about common questions and using this information to inform virtual reference and staffing needs.[5]

Jane Greenberg argued in 1998 that though NLP showed promise for the electronic archival environment, it was not suited to archival operations such as collection description or development because it ignores critical archival context. Greenberg argued that NLP "provides limited support of the archival accountability and memory objectives; [and] completely fails to support the evidential objective."[6] However, she did acknowledge the powerful indexing and accessing potential of NLP and encouraged inquisitive exploration of its possible applications for archival work. Curiosity about the potential of NLP is a recurring theme in the subsequent literature as archival practitioners tested these methods in applied cultural heritage settings over decades, with numerous authors performing these studies citing exploration and discovery of the methods as a primary goal of analysis.[7]

These studies have largely focused on digital collections to demonstrate how ML can be applied across archival operations. Using an OCR'd digital collection, Gregory, Geiger, and Salisbury found that Voyant's text analysis visualizations and features "consistently identify the main themes of a collection and draw connections between general themes and specific words and phrases to produce cleaner, more precise data."[8] They argue that this approach can aid in automating metadata production and extracting more useful and accurate metadata buried within narrative text. When ML methods are built into regular descriptive work, they assert, we can use these "[tools to] create familiarity, knowledge, and understanding rather than to merely confirm that which already exists."[9]

Similarly, Glowacka-Musial also explored topic modeling using the R programming language on a digitized corpus of historical press releases, and like Gregory, Geiger, and Salisbury, found that the process of identifying dominant topics streamlines the production of descriptive metadata.[10] Cain used topic modeling and text mining with MALLET and the Topic Modeling Tool (TMT) on a collection of historical government documents digitized by the library, generating a list of topics and visualizations to show how this approach could both enhance and ease access and content description.[11]

Experimenters at the Bancroft Library at UC-Berkeley developed ArchExtract in 2015 as a proof-of-concept platform relating NLP to archival processing through named entity recognition, topic modeling, and keyword extraction. Though no longer under development, the program previewed some potential ways that these methods could be regularly integrated into archival work.[12] Recent surveys have shown the promise of how ML methods might be applied to archival materials and archival practice as regular products of work, but they have focused on digitized archival and born-digital materials with little attention to the possibilities as applied to finding aids for physical materials.[13]

Collection Assessment in Libraries and Archives

Many methods have been used by librarians to assess collections over the decades. In recent years, automated analytics tools such as OCLC's GreenGlass and built-in analytics within discovery systems such as Alma/Primo have made this type of analysis much easier for cataloged materials. Data visualization is a standard component of these analytics.[14]

These tools have allowed robust, meaningful assessment to inform collection development strategy for library materials. ML methods have also been employed for the same collection analytics purposes. Librarians at Texas Tech University created a predictive analytics tool for interlibrary loan data to predict future circulation and are using results to influence collection development decisions.[15] Kiri Wagstaff and Geoffrey Liu trained a ML classifier to model librarians' weeding priorities. Their results "suggest that machine learning classifiers can improve the efficiency of weeding projects by pruning or prioritizing the list of weeding candidates prior to their review by a librarian."[16] The authors stressed the use of results as a time-saving way to allow "librarians to focus their time and attention" on likely sets. They and others in the literature emphasize that results from these models can inform analysis, but all ML endeavors need to be done in concert with human interpretation and other analysis tools.

As an institution's resources and collecting emphases shift, it can be difficult to assess where collection strengths lie if they have not been developed programmatically over time. Maintaining descriptive consistency and topical cohesion can also be challenging, resulting in unreliable term usage and discovery difficulties. Martha O'Hara Conway and Merrilee Proffitt urged archivists into a new paradigm of archival collections assessment in 2011 when they showed how the critical assessment process can expose hidden collections, establish processing priorities, and enhance collection management.[17] Patricia Rettig demonstrated the high value of collection analysis for archival repositories. She devised controlled categories and assigned them to collections in a special-subject archive, collating these to show the utility of this approach for correcting assumptions about collection areas, aiding reference work, and providing transparency to researchers.[18]

Qiana Johnson notes that any collection assessment should be directly related to a library's strategic planning efforts: "Assessing a collection based on the library's collection development policy gives the library an idea of where it needs to focus its energy for the foreseeable future. Then, as the collection is continually assessed, new areas of focus will emerge."[19] This kind of continual assessment can bring a fuller, more holistic understanding of a collection. Such an understanding is a necessary baseline for anyone responsible for adding material to a special collections or archives. As stewards who hold historical collections for the benefit of the public, accurate and responsible assessment of collections is a critical component of our work to both maintain strengths and responsibly build new areas of focus.[20]

The authors identified a potential area of development in the literature for studies involving ML tools applied to natural language elements in archival finding aids and applying ML tools to archival collection development strategies. We conducted text analysis on a corpus of finding aids to better understand how the method can surface topical collecting strengths in a repository's holdings and suggest scoping parameters.

PROCESS AND METHODS

Choosing Tools

After exploring several tools, including Voyant and spaCy, the authors selected Leximancer as a user-friendly and comprehensive tool appropriate for the purposes and scope of the project, and purchased a one-year academic subscription with library award funds. Leximancer has been utilized by scholars in a variety of fields for over a decade, offering an entry point to machine-learning techniques with relatively low technical knowledge required. Leximancer automatically performs an unsupervised semantic and relational analysis on large volumes of unstructured data to identify high-level themes within the text. Through a co-occurrence frequency statistical analysis, it "quantif[ies] the relationships between concepts" and presents these via two products, the Concept Map and the Topic Guide.[21]

The Concept Map provides a birds-eye view of the "original high dimensional co-occurrence matrix," showing a graphical representation of concepts and how they emerge, relate, and cluster spatially into topical themes.[22] The ranked-concept lists and thesaurus accompanying the Concept Map also allow the researcher to see specific concepts in the context of the original texts.[23] The Topic Guide makes review of the corpus "more efficient and more effective" by automatically indexing subjects to create the most frequently co-occurring concept pairs within the text.[24]

These visualizations allow users to explore the data in unique ways. Haynes et al. comment on Leximancer's ability to enable both "zooming out" to see the higher-level concepts from a corpus and "zooming in" to see the nuances of the relationships between topics within the text: "Leximancer facilitates a highly inductive, data-driven process, providing an analytical 'fresh lens' and the potential for identifying novel linkages and groupings of specific terminology that might not be identified by manual" means.[25] As a tool for both analyzing and presenting the data, it offered a comprehensive and time-tested option for this project.

Selecting and Preparing the Data

Like many repositories, SCARC provides public access to archival collections that are in different stages of the arrangement and description process. Of the 1,371 publicly available collections included in the analysis, 878 collections are fully processed and comply with modern descriptive standards, including full description, subject headings, and container lists for larger collections.[26] 265 collections are preliminary collection-level descriptions including basic descriptive matter and minimal subject terms for materials that have not been fully described, arranged, or physically processed. Of these 265, 188 also have container lists made at the time of accessioning in varying levels of detail. These container lists, originally in PDF form, were OCR'd using ABBYY and converted to TXT files. They were included in the dataset in addition to the main CSV file as rich additional examples of natural language description from both collection creators and archivists. 228 collections have "stub" records with very minimal preliminary description.

At the time of this project, SCARC used Archon v3.21 for archival collection management. Compiling the dataset involved exporting 1,420 finding aid documents from Archon in XML format. The exported files were batch edited using the Oxygen XML Editor application to resolve validation errors. A loss of Archon functionality was deemed acceptable, so some components of the XML documents that were not relevant to the goals of the project were removed during this data cleanup, such as HTML markup and external reference links.[27] The resulting XML documents validated against the EAD 2002 W3C Schema.[28] Nearly 50 collections were closed to research for various reasons; these were excluded from subsequent processing, resulting in the final set of 1,371 collections.[29]

From the EAD element set, the authors identified descriptive elements of potential value for text analysis. As this study was focused on exploring collection content, the narrative fields were limited to the unit titles and scope and content notes which describe series, subseries, and folders in natural language.[30] Various identifying metadata, including Description Level and Collection Type, were also included to aid in tagging and tracking the results of the analysis. A custom XSLT 2.0 stylesheet was used to transform the collection of EAD XML documents to a single CSV output, where one row of data represents an object at any level of description (i.e., collection, subgroup, series, subseries, file, or item) according to a field mapping.[31]

Baseline Subject Analysis

We established a baseline understanding of the topical coverage in the corpus by performing an analysis on the Library of Congress Subject Headings (LCSH) assigned at the collection level by archivists. Subject headings and other controlled vocabulary terms were not included in the source text for Leximancer modeling. However, this fundamental aspect of traditional bibliographic description serves a similar function of abstracting collections to their general scope of coverage. The baseline subject analysis used the LCSH taxonomy to group the assigned headings into related topics, giving an aggregate picture of the collections.

To achieve this, 2,163 unique original LCSH were identified and extracted from SCARC EAD finding aids. Removing subdivisions resulted in an initial set of 1,282 unique subjects. Using an XSLT 2.0 pipeline, a topic model was constructed by recursively retrieving and merging consecutively broader terms for these subject headings from LCSH authority records until no novel broader terms remained.[32] Processing with this pipeline reduced the original subject headings to a set of 418 top-level terms, according to the knowledge organization available from the Library of Congress. Of these top-level terms, many have little or no hierarchical depth or represent a small number of subject heading occurrences; 257 of the top-level terms represent a single unique original heading each; 38 of those have only one instance of the original subject heading. Of the remaining top-level topics, several of the largest were so generic as to be nearly devoid of meaning; for instance, 42% of the unique original subject headings assigned to SCARC finding aids can be traced to the top-level term "Science." To address these issues, the resulting topic graph was filtered for "right-sized" concepts to meaningfully group subject headings into concept areas. Throughout all levels of the broader term tree, 485 LCSH were identified that each have between 5 and 50 unique originally occurring (nonexclusive) subject headings as descendants.[33] This set of right-sized terms served as a baseline against which to compare the Leximancer text analysis results.[34]

Large concentrations in topics related to science, people, and agriculture are prevalent, and SCARC's original signature areas can be discerned relatively easily. The sciences represented are wide-ranging, but terms related to biology, chemistry, botany, and zoology are frequent. People and society are strongly represented through broad terms such as *Public institutions*, *Creative ability*, and *Women*. Agriculture is also wide-ranging, including various elements such as *Agricultural industries*, *Food*, and *Livestock.* Education terms are prevalent, indicating the university concentration. The natural environment is strongly represented in "right-sized" terms such as *Land use*, *Environmental protection*, and *Nature conservation*. An emphasis on industries is also revealed, with "right-sized" terms such as *Primary commodities*, *Manufactures,* and *Industrial management* in the analysis.

Leximancer Modeling

Leximancer automatically generates concept seeds, which are those terms that appear most frequently in the corpus and that have other affiliated terms co-occurring in text segments. The lists of most frequently appearing words and names give an initial suggestion of concepts that organically emerge from the corpus. The user can tighten and enhance these by combining, adding, or removing words in the Concept Seed Builder interface, which also allows the user to add positive or negative evidence terms to refine the concept.[35]

In this project, for multiple iterations of the same dataset, this initial concept seed list remained the same or similar, even with changes to settings such as concept generality, classification threshold, and segment size. Leximancer's user guide advises that some concepts can be "bleached of semantic meaning" and encourages removal of terms that are meaningless or ambivalent without context.[36] In this project's initial list, a number of categories of such words were identified and manually removed, including adjectives such as *large* or *several*; time-based concepts such as *circa* or *during*; event-based words such as *received*; and other broad terms that did not indicate topical coverage, such as *views*, *study*, *efforts*, or *involvement*. General words related to archives, such as *collection*, *folder*, and *series*, were also taken out. No new terms were added.[37]

For this study, which was concentrated on topical coverage, all material type terms such as *photographs* and *correspondence* were also removed, along with terms related to material types, such as *glass* or *lantern* (slides). After such terms were removed, the initial concept seeds list was largely composed of frequently occurring nouns. Because the analysis is based on pattern recognition in the text, the level of specificity in this list was varied, including terms as broad as *military* and as specific as *cadet*.

Leximancer attempts to identify proper names by examining the corpus for words beginning with capital letters through a setting called "Identify Name-Like Concepts." Use of this setting resulted in many false positives such as abbreviations, ambivalent terms, collection titles, terms related to the repository, or singular names such as William or Smith. Instead, the most frequently occurring names surfaced in a word-only analysis, where they were reclassified from words to names.

Leximancer's developers suggest examining thesaurus results to remove or merge outlying, irrelevant, or similar terms. The process of generating the Concept Map must be repeated and adjusted multiple times, and relative locations should be compared for each run to check for model stability.[38] When results are consistently repeating, "the cluster map is likely to be representative."[39] Runs were stopped once similar results had been consistently achieved.

RESULTS

The final Leximancer model resulted in a tightly cohesive Concept Map and Topic Guide showing themes and topic pairs emerging from the corpus. Leximancer returned 140 concept terms and nearly 1,500 resulting Topic Guide pairs. Results were highly aligned with the baseline subject heading analysis, indicating accurate and thorough controlled vocabulary terms assigned to the fully processed collections by archivists. The concepts, themes, and co-occurring topic pairs surfaced by Leximancer suggest areas of particular collection strength and potential focus for new acquisitions.

In the map, theme color reflects relevance (with warm red and orange denoting the most relevant concepts, and cooler colors denoting less relevant concepts). Figures 1–3 present different views of the same Concept Map. Figure 1 shows that higher level themes can be removed to show underlying concepts. Figure 2 and Figure 3 show how circle size denotes the degree of connectivity with other concepts, and how the percentage of themes displayed can be adjusted to show more or fewer themes. Figure 4 shows a Topic Guide excerpt.

Figure 1. Concept Map concepts

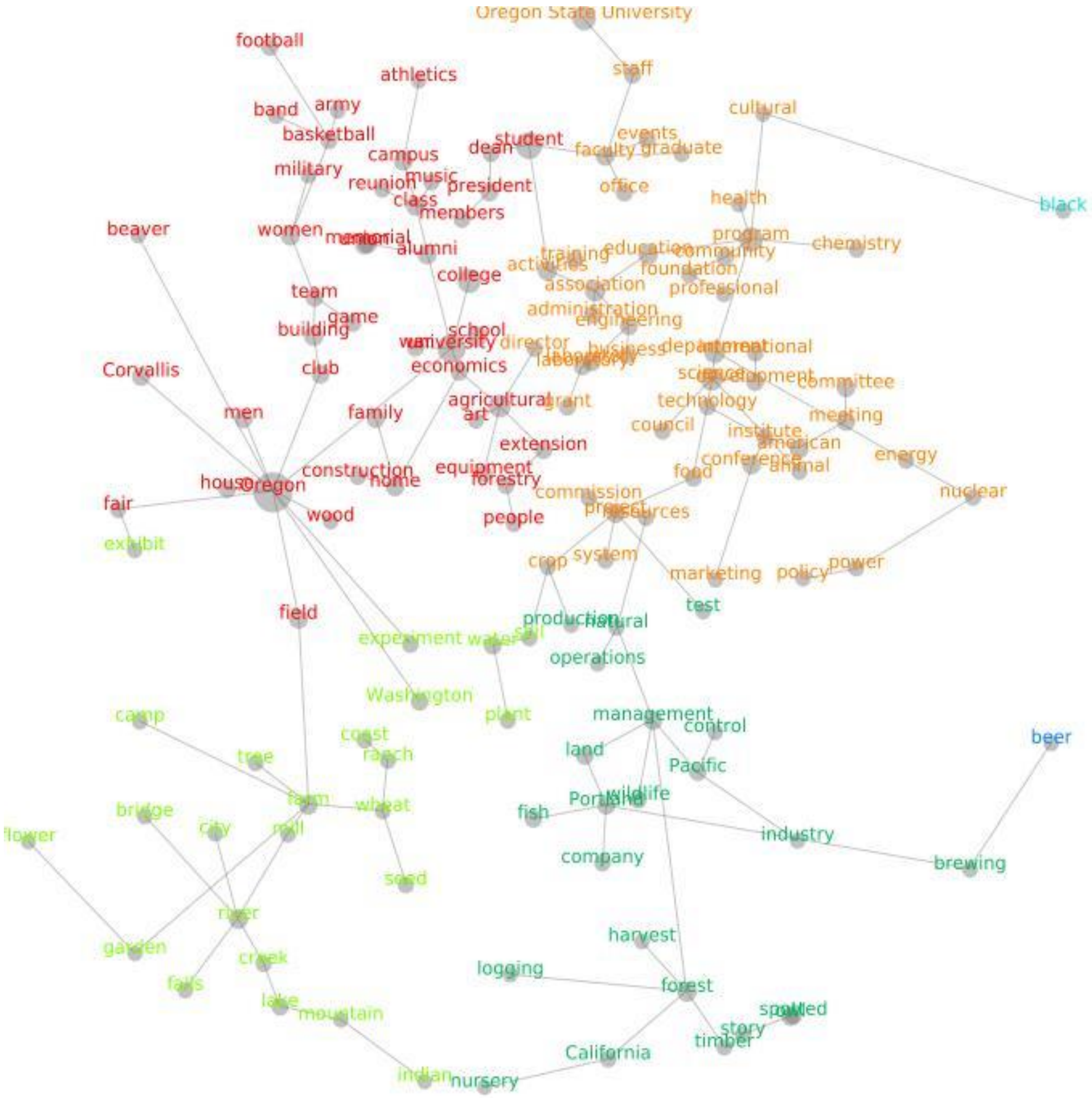


Figure 2. Concept Map themes (35%)

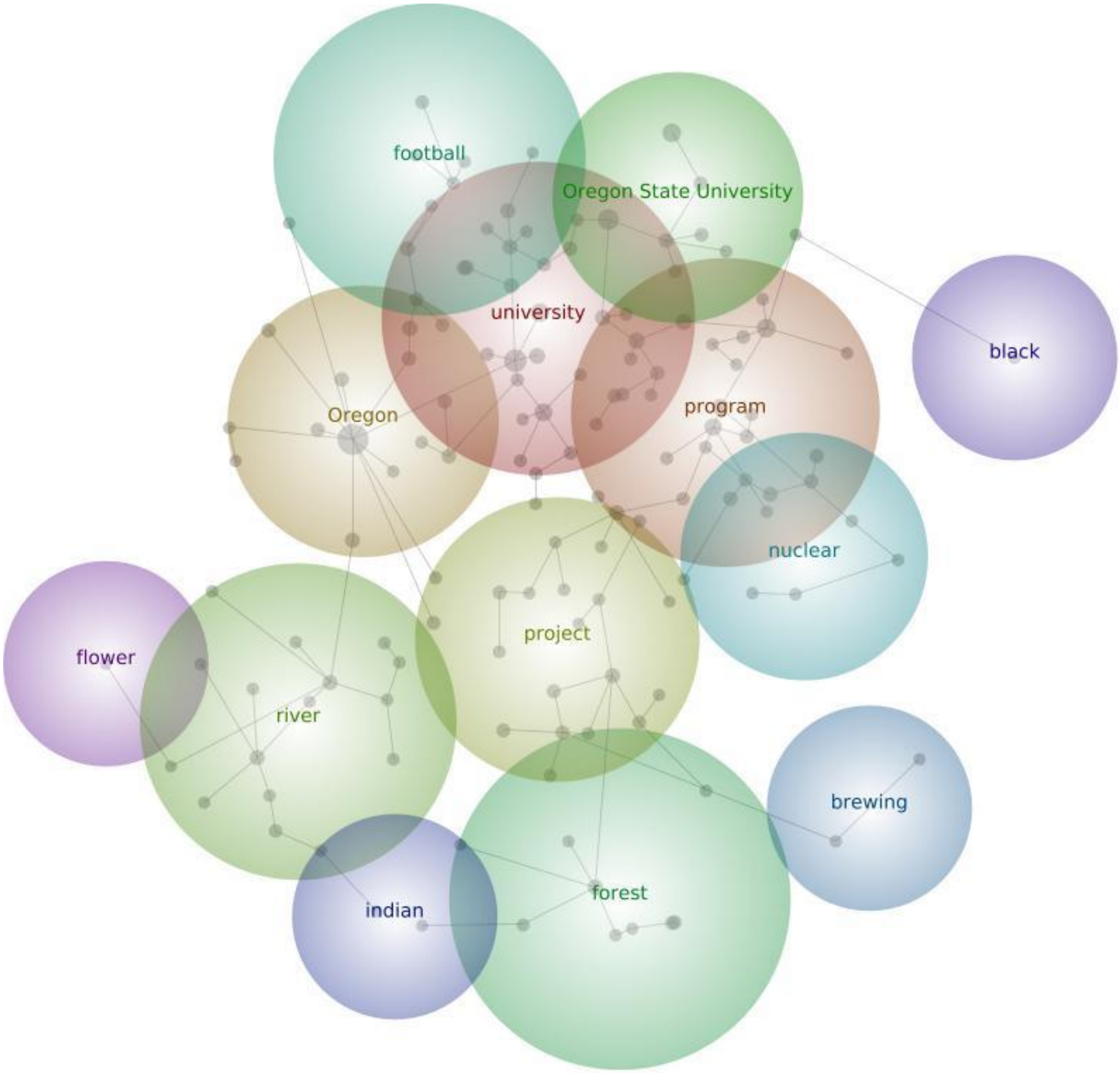


Figure 3. Concept Map themes (17%)

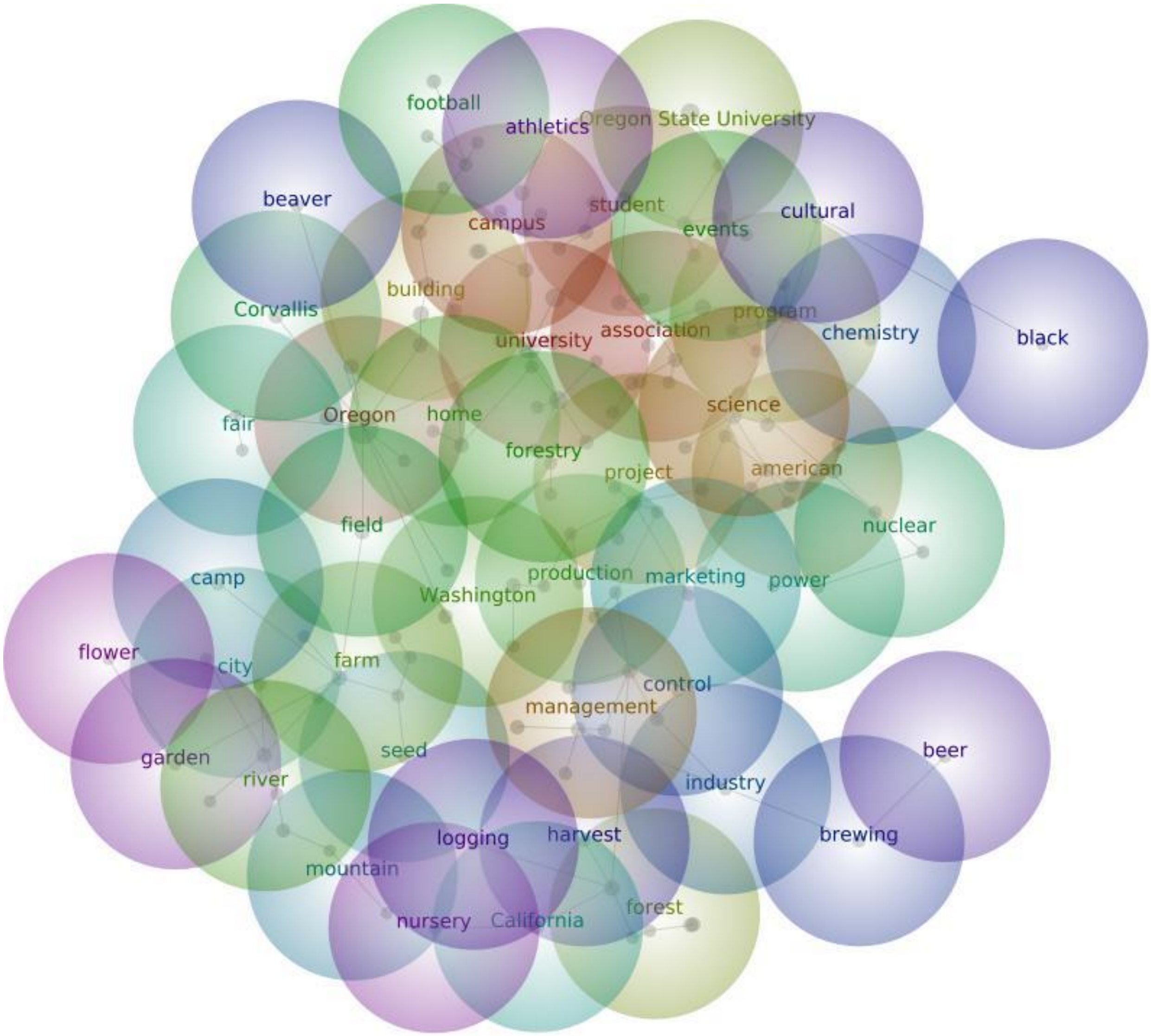
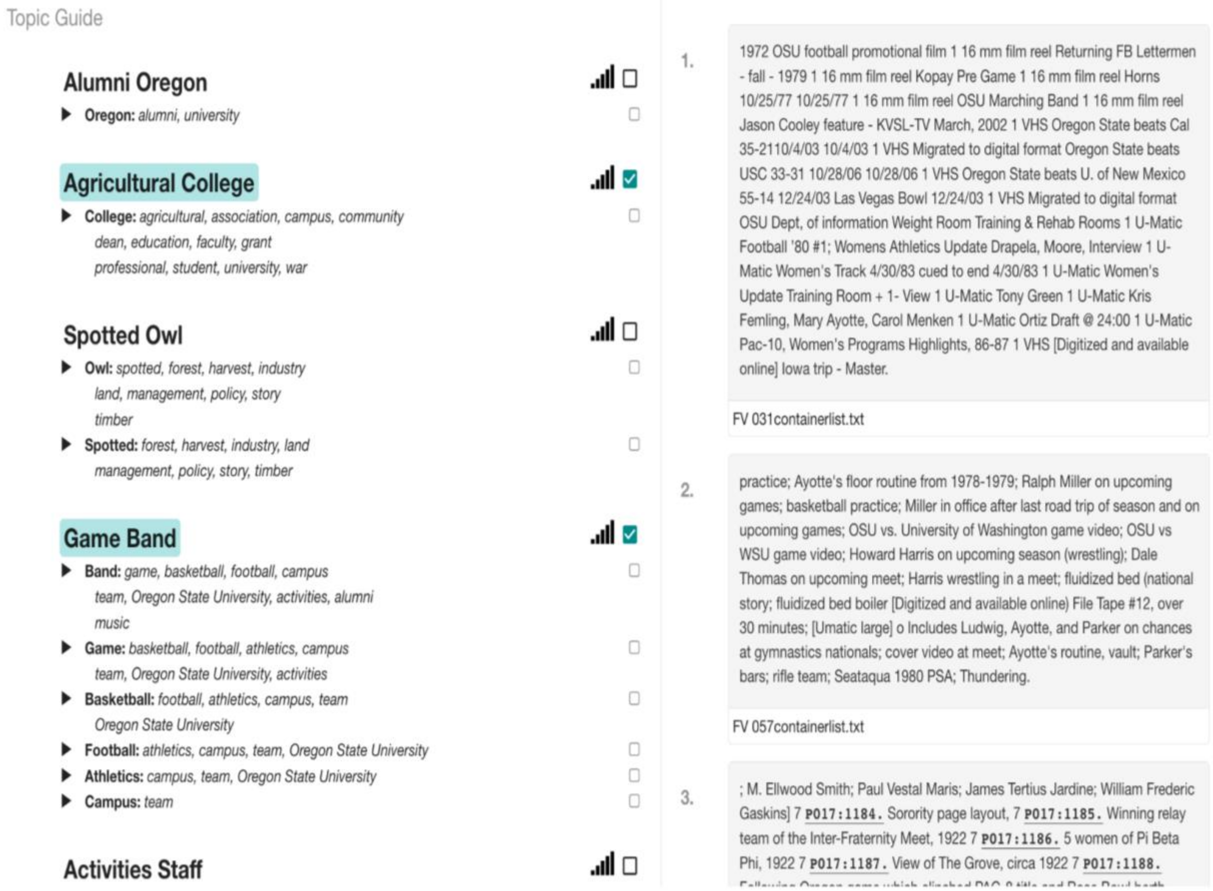


Figure 4. Topic Guide excerpt



Topic Guide

Alumni Oregon ☐
- **Oregon:** *alumni, university* ☐

Agricultural College ☑
- **College:** *agricultural, association, campus, community dean, education, faculty, grant professional, student, university, war* ☐

Spotted Owl ☐
- **Owl:** *spotted, forest, harvest, industry land, management, policy, story timber* ☐
- **Spotted:** *forest, harvest, industry, land management, policy, story, timber* ☐

Game Band ☑
- **Band:** *game, basketball, football, campus team, Oregon State University, activities, alumni music* ☐
- **Game:** *basketball, football, athletics, campus team, Oregon State University, activities* ☐
- **Basketball:** *football, athletics, campus, team Oregon State University* ☐
- **Football:** *athletics, campus, team, Oregon State University* ☐
- **Athletics:** *campus, team, Oregon State University* ☐
- **Campus:** *team* ☐

Activities Staff ☐

1. 1972 OSU football promotional film 1 16 mm film reel Returning FB Lettermen - fall - 1979 1 16 mm film reel Kopay Pre Game 1 16 mm film reel Horns 10/25/77 10/25/77 1 16 mm film reel OSU Marching Band 1 16 mm film reel Jason Cooley feature - KVSL-TV March, 2002 1 VHS Oregon State beats Cal 35-2110/4/03 10/4/03 1 VHS Migrated to digital format Oregon State beats USC 33-31 10/28/06 10/28/06 1 VHS Oregon State beats U. of New Mexico 55-14 12/24/03 Las Vegas Bowl 12/24/03 1 VHS Migrated to digital format OSU Dept, of information Weight Room Training & Rehab Rooms 1 U-Matic Football '80 #1; Womens Athletics Update Drapela, Moore, Interview 1 U-Matic Women's Track 4/30/83 cued to end 4/30/83 1 U-Matic Women's Update Training Room + 1- View 1 U-Matic Tony Green 1 U-Matic Kris Femling, Mary Ayotte, Carol Menken 1 U-Matic Ortiz Draft @ 24:00 1 U-Matic Pac-10, Women's Programs Highlights, 86-87 1 VHS [Digitized and available online] Iowa trip - Master.

   FV 031containerlist.txt

2. practice; Ayotte's floor routine from 1978-1979; Ralph Miller on upcoming games; basketball practice; Miller in office after last road trip of season and on upcoming games; OSU vs. University of Washington game video; OSU vs WSU game video; Howard Harris on upcoming season (wrestling); Dale Thomas on upcoming meet; Harris wrestling in a meet; fluidized bed (national story; fluidized bed boiler [Digitized and available online) File Tape #12, over 30 minutes; [Umatic large] o Includes Ludwig, Ayotte, and Parker on chances at gymnastics nationals; cover video at meet; Ayotte's routine, vault; Parker's bars; rifle team; Seataqua 1980 PSA; Thundering.

   FV 057containerlist.txt

3. ; M. Ellwood Smith; Paul Vestal Maris; James Tertius Jardine; William Frederic Gaskins] 7 **P017:1184.** Sorority page layout, 7 **P017:1185.** Winning relay team of the Inter-Fraternity Meet, 1922 7 **P017:1186.** 5 women of Pi Beta Phi, 1922 7 **P017:1187.** View of The Grove, circa 1922 7 **P017:1188.**

The Concept Map shows two broad clusters of university-related themes covering the upper portion of the map (see Figure 1). Terms such as *student*, *faculty*, *staff,* and *administration* are networked around academic areas such as *agricultural*, *science*, *engineering,* and *forestry.* Topic Guide pairs such as *Student Activities*, *Cultural Student*, and *Student Association* suggest an emphasis on student experiences. Sports themes emerge from terms such as *team* and *game* and pairs such as *Athletics Women* and *Football Team.* University research areas are indicated by a cluster of terms including *chemistry*, *science*, *business*, *nuclear*, *technology*, and *energy*, which are surrounded by wider university contributions and impacts in these areas such as *council*, *institute*, *conference*, *policy*, and *commission*.

Very near the center of the map are the terms *agricultural* and *extension*, indicating the land grant mission of the university and its reach throughout the state. Pairs such as *Extension Program* and *Extension Activities* echo this reach. The central focus on agricultural content emerges from terms and themes such as *farm*, *field*, and *food* and more than 200 agriculturally focused Topic Guide pairs such as *Crop Production*, *Food Agricultural, Garden Harvest*, *Nursery Seed*, and *Soil Resources*.

A large *Oregon* theme occupies the central left area of the map, bridging concepts in university and natural environment clusters. This area also includes overlapping social themes such as *home*, *family*, *club*, and *people*. The social element is echoed in the many pairs involving *people*, including *Business People*, *Tree People*, and *Ranch People*.

Natural environment terms are distributed throughout the lower portion of the map. Terms such as *water*, *plant*, *land*, *fish*, and *natural* border landscape features such as *coast*, *lake*, *river*, and *mountain*. Co-occurring pairs including *Natural Coast*, *Wildlife Fish*, *Land Water*, and *Natural Land* also indicate this emphasis.

The lower right of the Concept Map indicates an emphasis on industries with terms such as *production*, *operations*, *management*, *control*, and *company*. This emphasis is echoed in pairs such as *Industry Development*, *Operations Activities*, and *System Engineering*. This section also includes terms related to the lumber industry, including *logging*, *timber*, *forest*, and *owl*. Topic Guide pairs such as *Spotted Owl*, *Logging Operations*, and *Mill Timber* emphasize this topic area.

DISCUSSION

Though results resemble the institution's original "signature areas," there are intriguing nuances indicating new relationships, connections, and juxtapositions that could be considered for collection scoping and future acquisition decisions. Text analysis can accurately surface topical strengths and directly lead to insights that can inform future acquisition decisions and collection development policies.

Near the core of the Concept Map (see Figure 1) are the terms *crop* and *production*, indicating the centrality of this topic across the collections. Crop-related terms are scattered throughout the lower area of the map and include *seed*, *wheat*, *fish*, and *tree*. *Food* also clusters near the center and co-occurs most often with terms such as *agricultural*, *technology*, and *science*. *Agricultural*, *crop*, and *food* are broadly connected to other areas by co-occurring terms. This suggests enough of a central focus in the collections on agricultural production to call this out as a particular area of strength, along with surrounding sciences, operations, and industries.

Terms most related to the known brewing and hops concentration include *company*, *management*, and *industry*, but also include *field*, *farm*, *agricultural*, *chemistry*, and many others. This variety of co-occurring terms to *beer* and *brewing* indicates broad, solid interrelations to other known topical strengths in the corpus, positioning this area as a particularly significant example of cohesive power across collections. The collecting activities and relationships built to yield these collections over the past decade could be a model for developing stronger concentrations on other agricultural products and communities indicated within existing collections.

The natural environment (as indicated by terms *forest*, *river*, *wildlife*, and *water*) and the built environment (*bridge*, *city*, *farm*, and *ranch*) are often spatially close on the map. Industry terms are situated near the natural environments where they closely co-occur (*forest*, *coast*, *agricultural*, and *land*). Together, these suggest a tension between the natural and human-made worlds. Records from individuals, organizations, and companies working in areas such as environmental sustainability, natural engineering, and urban ecosystem conservation could thus be complementary acquisitions to this strength. Natural resource terms such as *wildlife*, *land*, *water*, and *forest* cluster near terms suggesting the physical and intellectual infrastructure needed for their management (*policy*, *resources*, *system*, and *commission*). This suggests the broad interrelation between natural resources and their management in the collections, and the potential for acquisitions from organizations and agencies bridging these.

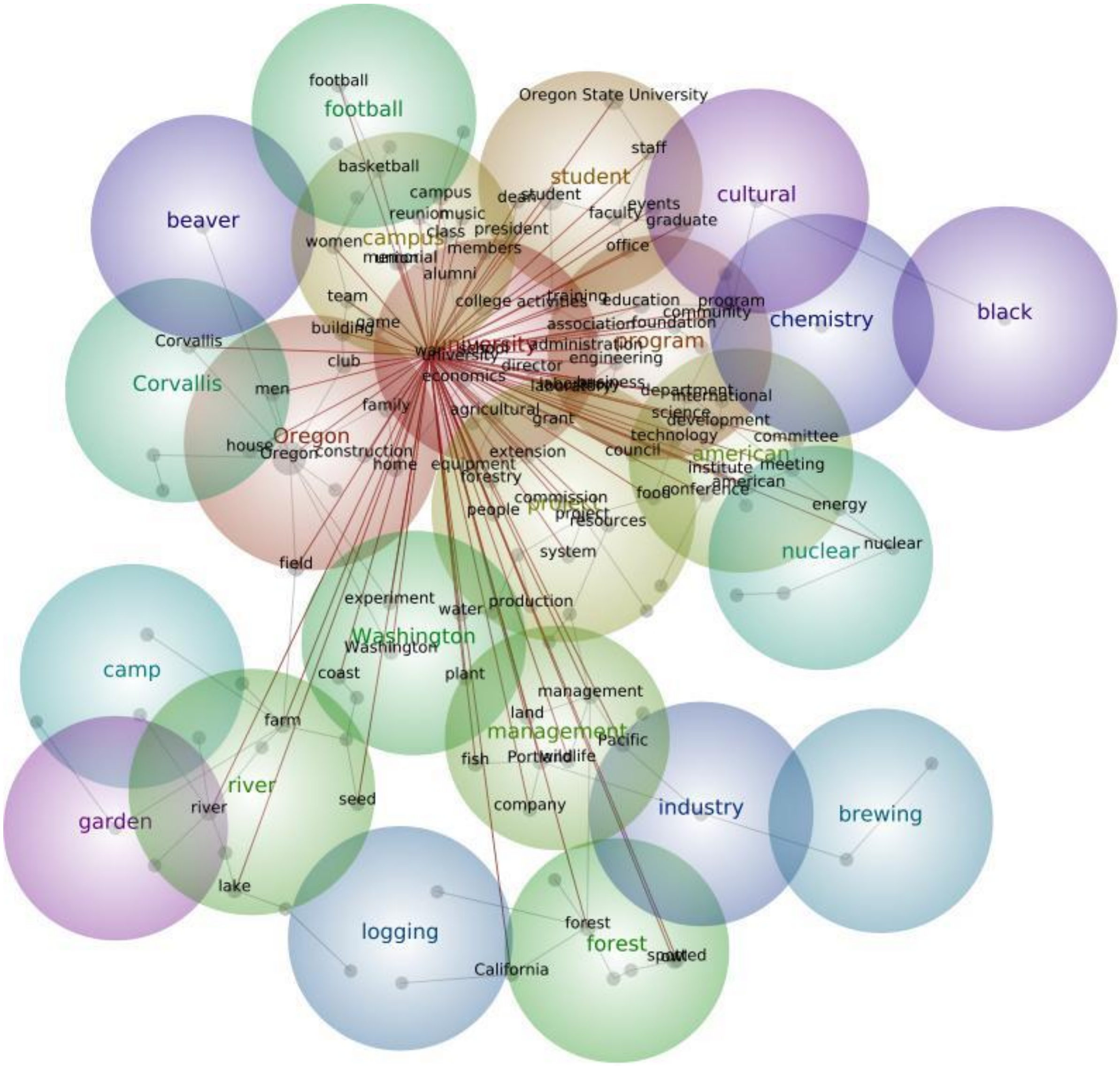
The subject heading analysis indicated geology, chemistry, entomology, engineering, and botany as top science-related terms, but only two of these appear in Leximancer results. The theme of *science* stretches broadly into other areas, as indicated by pairs such as *Chemistry Agriculture,* *Health Science*, *Engineering Science*, *Natural Science,* and *Energy Nuclear*. Science is also indicated by terms such as *laboratory*, *experiment*, *grant*, *study*, and *test*, and Leximancer's ranked concepts list shows *science* is most frequently located with terms such as *natural*, *food*, and *agricultural*. This information suggests the broad "signature area" of history of science could be sharpened to emphasize agricultural, environmental, and/or engineering in archival acquisitions. The emphasis on nuclear energy could be supplemented with historical materials related to energy production, natural resources, and effects on the environment and human health.

The spatial distance between the terms *forestry* and *forest* on the map may indicate a similar disconnect in the content of the collections worthy of closer attention. The educational focus in the university's College of Forestry collections may not be paralleled in the collections closer to timber industry concerns such as *wildlife* and *policy*. If this disconnect were addressed strategically, collecting relationships could focus on individuals and organizations who bridge forestry training into applied environments.

Local history is strongly suggested as an area of current concentration, but it is unacknowledged in the original "signature areas." Terms and pairs such as *Development Community*, *City Corvallis*, and *City Council* suggest the presence of material specific to the local community outside the university. Further development could focus on unity between local history and other strengths, focusing on community individuals and organizations whose efforts engage with environmental, scientific, or cultural concerns.

Another unacknowledged common theme is *war*, which is close to the university cluster but also has connections throughout, to *science*, *agriculture*, *forestry*, and landscape terms (see Figure 5). Connecting these materials to the known emphasis on world peace in the Ava Helen and Linus Pauling papers could strengthen this collection cohesion and form a foundation for future collecting in this area.[40]

Figure 5. Concept Map showing connections for term *war*



Underrepresented communities were widely present in the baseline subject heading analysis in terms such as *Native Americans*, *Hispanic Americans*, and *African Americans*. Though these are controlled vocabulary terms, they may not reflect the expressions of these communities in natural language.[41] In the Concept Map, the term *American* is broadly related across areas, but co-occurs most frequently with the terms *association*, *institute*, *Indian*, and *cultural*. *Association* and *institute* are broadly connected in context to a variety of social and scientific organizations. *Indian* in most context snippets refers to antiquated usage of the term American Indian. *Cultural* in context snippets often refers to underrepresented communities. The term *black* co-occurs broadly with other terms including *cultural*, *union*, and *American*, indicating Black presence in the collections (along with the pair *Cultural Black*). In context, however, the term is sometimes being used as an adjective descriptor, or within names of people, businesses, or places.

These representation issues point to inherent problems in ML and language. Leximancer is concerned with the frequency of terms; less-frequent terms are not included in results. Underrepresented communities may not easily surface in a text analysis if specific terms associated with these communities are infrequent in the corpus. Thus, the results must be as closely examined for what is *not* present as much as for what is represented. As Erin Wolfe and others have noted, computational text analysis is challenged to surface these voices, and possesses the biases of creators.[42] This underscores the urgency of reparative description efforts and suggests that deliberately assigning identity-based subject headings to relevant materials may be more effective at surfacing hidden voices than relying solely on natural language description.[43] Adding specific concept seeds (see below) could adjust results to focus more attention on these terms.

Leximancer's results suggest distinct areas of topical cohesion across collections and subtle ways to refine or restate collection parameters. Based on these results, current collecting strengths might be better defined as:

- the intellectual, physical, administrative, and personal histories of the university and its students, faculty, and staff
- agriculture and food production in Oregon across a range of operations and industries
- forest, water, wildlife, and land ecosystems and resources, their management, and the balance between the natural and built environments
- sciences related to agriculture, nature, energy, and the built environment
- communities and peoples of Oregon
- war and peace in the 20th century, especially its connections to science and agriculture

These results show that text analyses generated by Leximancer (or similar utilities) can effectively reveal areas of collection density and connection. The visualizations and analyses trace how terms and concepts relate to each other throughout the corpus and suggest potential collecting actions to narrow, broaden, or unify collection strengths for institutional cohesion. Though this model indicated strong collection interrelatedness within SCARC collections, models displaying wide gaps between subject areas would suggest a lack of collection relatedness. While some institutions may have good reasons for maintaining disconnect between acquisition areas, others may seek to use such results to bring topical coherence and integrity across the repository in support of other strategic goals.

LIMITATIONS

The study's limitations included Leximancer's unexpected learning curve. Initially there was difficulty pinpointing how configuration changes affected results, though this was largely resolved after intensive engagement with the program. Leximancer does not favor two users working simultaneously, which also slowed progress. The authors found that Leximancer models were unpredictable; for example, running the same model multiple times with minor adjustments caused top terms to drop from the frequency count for unknown reasons. Variations in Leximancer settings might yield different results.

Other limitations are related to the data involved in this study. Finding aids are abstractions of physical content described in the aggregate by archivists, who may allow unconscious bias to highlight or diminish specific aspects of a collection. Though finding aids are permeated with natural language, they are also idiosyncratic in that they are structured data with an inherent vernacular. This imposed structure can potentially overshadow the unstructured narrative elements. As word frequency is the single most important factor in this method of analysis, unevenness in the length and depth of descriptive matter found across the finding aid corpus led to the more described collections being more represented simply because their descriptions contained a higher number of words. This is in contrast to a subject heading-based analysis, in which a similar number of headings are assigned to most collections regardless of their size or other nature. Future iterations could selectively reduce or remove the most detailed finding aids described at the item level from the set to mitigate this influence.

The resulting model did not depart significantly from the original broad signature areas and did not surface more latent content. The scope of this project was to analyze broad topical strengths using natural language descriptive elements across a corpus of archival finding aids. To get the most organic picture of scope, there was no intervention in Leximancer's initial set of concept seeds. More sophisticated results might be expected by adding concept seeds of known areas of secondary strength using terms pulled from the baseline subject analysis and/or vocabularies suggested by collection archivists. These seeds could be added and refined in the thesaurus to force attention toward these areas for further definition.

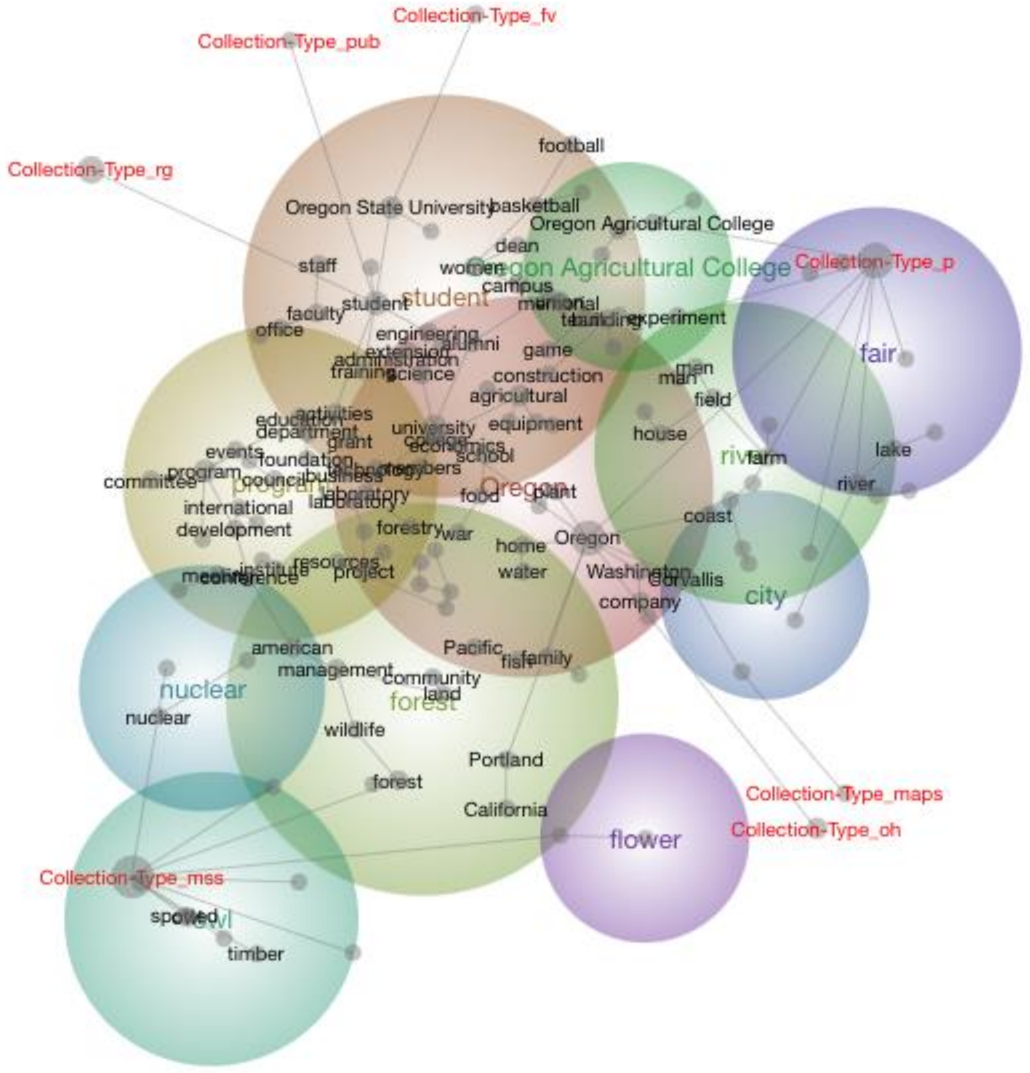
LOOKING AHEAD

Leximancer allows the user to tag different categories within the data to compare how these will cluster in the concept map. While outside the scope of the current study, retaining the material type terms in concept seeds and/or adding Description Level or Collection Type tags would bring an additional layer of understanding that would affect decisions about acquisitions or processing priorities. Figure 6 shows how topic coverage distribution is affected by the added element of collection type.

In addition to collection scoping and development, text analysis on a corpus of finding aids has numerous potential applications to archival description and processing. As Gregory, Geiger, and Salisbury have shown, this type of concept mapping and topical co-occurrence guide could drive description tasks such as selecting controlled vocabulary terms, writing the related materials note, and describing content.[44] Results could also drive other department operations such as reference or instruction, which represent research strengths to the public and connect researchers with relevant materials.

Comparing these results with those from similar processes using other established text analysis tools such as spaCy and the Natural Language Toolkit, as well as with emerging ML models, would yield deeper understanding. Further comparative elements are expected, including unprocessed accessions, digital collections metadata, oral history transcripts, and/or OCR'd text of digitized materials. To study the effect of archival descriptive intervention, the dataset could be limited to creator-generated metadata. At a larger scale, with a dataset including multiple repositories, a text analysis approach could identify overlaps, connections, and areas of distinction within a consortium or coalition of institutions, along with potentially split or missing collection areas across institutions.

Figure 6. Concept Map with Collection Type tags



Leximancer is also capable of semantic analysis and extracting social networks, offering tempting potential for other cultural heritage discovery applications. Leximancer's User Guide lists "document base navigation" as a potential application of the tool, specifically for legal e-discovery; but that use might just as easily be applied to archival search environments.[45] Though Leximancer's interface can be nonintuitive, individual functions within the interface have interesting implications for cultural heritage discovery. The Topic Guide allows for selection of terms to be searched together for co-occurrence, and the powerful ranked-concept searching allows users to see how often and in what context a term is used with other terms. This feature shows great promise and begins to address Greenberg's early caution that NLP analysis must "permit any retrieved record to be viewed with the context of the recordkeeping system from which it emerged."[46]

Adding this layer of searching nuance to the cultural heritage research process would yield a much richer and deeper understanding of any collection and would offer timesaving ways to increase research efficiency. The Topic Guide features context searching for keyword pairs, which yields search results that contrast sharply with the results for the same singular keywords in the current discovery interface. This type of topical pairing presentation hints at more efficient research methods by better showing what specific avenues of research may be most fruitful among the collections at hand. Further, concept frequency and co-occurrence could suggest keyword norming or other ways to unite search results in less-than-ideal search systems. A network visualization showing how terms connect to each other (see Figure 5) can also aid discovery and surface hidden connections.[47] As a tool, Leximancer fulfills all the design principles and workflow considerations suggested by Hutchinson for ML tools to be functionally adopted in library and archival workflows. It is usable, interoperable, flexible, iterative, and configurable, which signals that a platform such as this has immense potential to affect library and archival discovery.[48]

CONCLUSION

Text analysis can yield clear insights when applied to assessment of archival scope and collection strengths. This project found that the results of text analysis using natural language archival description surfaced distinct topical strengths from the corpus and suggested ways to strategically cohere and refine the repository's collecting scope. Repositories seeking to align collecting scopes for more strategic decisions may find utility in this approach, though all results will require further interpretation and should not "replace the work of judgment, inference, and interpretation."[49]

ML and artificial intelligence (AI) techniques are rapidly becoming crucial tools within the cultural heritage and archival communities. ChatGPT is already being tested for application to extracting and matching entities, improving discovery, and enhancing texts, and other tools are developing quickly.[50] As the community adjusts to the presence of these tools and begins to see new applications of this technology, there will be great potential for further investment. Ryan Cordell suggests library ML projects can "serve as *diagnostics* to the biases and gaps in existing digitized or born-digital collections, as *rebuttals* to claims—whether from scholars or Silicon Valley—about ML objectivity, or as a *synecdoche* that helps patrons, scholars or students better understand the historical stakes of library collection and archives."[51] The present project extends that diagnostic reach to nondigitized, non-born-digital collections, showing that data from archival finding aids can be used for the same purpose, for the same insights, for the same audiences.

To be responsible collection stewards moving forward, we must leverage the full power of ML and AI. Internal, diagnostic applications can improve decision-making and assessment for acquisitions and resource expenditures; external discovery environments can make the collections we steward more thoroughly accessible to the public. By examining collecting data with intention and inquisitiveness, cultural heritage professionals can use these emerging tools to build more cohesive, relevant, and inclusive collections for the future.

ACKNOWLEDGMENT

The authors would like to acknowledge the support of Kiri Wagstaff, Ryan Wick, Kevin Jones, and the Robert W. Lundeen Award Fund of the Oregon State University Libraries and Press.

ENDNOTES

[1] Mary Kidd et al., *Total Cost of Stewardship: Tool Suite* (OCLC Research, 2021), https://doi.org/10.25333/4bqc-5k43; Martha O'Hara Conway and Merrilee Proffitt, *Taking Stock and Making Hay: Archival Collections Assessment* (OCLC Research, 2011), https://doi.org/10.25333/C33S6M.

[2] Tiah Edmunson-Morton, "Oregon Hops and Brewing Guide," https://guides.library.oregonstate.edu/brewingarchives; Natalia Fernández, "OSU Queer Archives: OSQA," https://guides.library.oregonstate.edu/osqa.

[3] Joo Soohyung, Erin Ingram, and Maria Cahill, "Exploring Topics and Genres in Storytime Books: A Text Mining Approach," *Evidence Based Library and Information Practice* 16, no. 4 (2021): 41–62, https://doi.org/10.18438/eblip29963.

[4] Melissa Harden, "First-Year Students and the Framework: Using Topic Modeling to Analyze Student Understanding of the Framework for Information Literacy for Higher Education," *Evidence Based Library & Information Practice* 14, no. 2 (June 2019): 51–69, https://doi.org/10.18438/eblip29514.

[5] A. Sharma, K. Barrett, and K. Stapelfeldt, "Natural Language Processing for Virtual Reference Analysis," *Evidence Based Library and Information Practice* 17, no. 1 (2022): 78–93, https://doi.org/10.18438/eblip30014.

[6] Jane Greenberg, "The Applicability of Natural Language Processing (NLP) to Archival Properties and Objectives," *The American Archivist* 61, no. 2 (1998): 421, https://doi.org/10.17723/aarc.61.2.j3p8200745pj34v6.

[7] Jonathan O. Cain, "Using Topic Modeling to Enhance Access to Library Digital Collections," *Journal of Web Librarianship* 10, no. 3 (2016): 210–25, https://doi.org/10.1080/19322909.2016.1193455; Kate Gregory, Lauren Geiger, and Preston Salisbury, "Voyant Tools and Descriptive Metadata: A Case Study in How Automation Can Compliment Expertise Knowledge," *Journal of Library Metadata* 22 no. 1/2 (January 2022): 1–16, https://doi.org/10.1080/19386389.2022.2030635.

[8] Gregory, Geiger, and Salisbury, "Voyant Tools and Descriptive Metadata," 14.

[9] Gregory, Geiger, and Salisbury, "Voyant Tools and Descriptive Metadata," 15.

[10] Monika Glowacka-Musial, "Applying Topic Modeling for Automated Creation of Descriptive Metadata for Digital Collections," *Information Technology & Libraries* 41, no. 2 (2022), https://doi.org/10.6017/ital.v41i2.13799.

[11] Cain, "Using Topic Modeling to Enhance Access to Library Digital Collections," 24.

[12] Discussed in Tim Hutchinson, "Natural Language Processing and Machine Learning as Practical Toolsets for Archival Processing," *Records Management Journal* 30, no. 2 (2020): 12, https://doi.org/10.1108/RMJ-09-2019-0055.

[13] Ryan Cordell, "Closing the Loop: Bridging Machine Learning (ML) Research and Library Systems," *Library Trends* 71, no. 1 (2022): 132–43, https://doi.org/10.1353/lib.2023.0008;

Christopher A. Lee, "Computer-Assisted Appraisal and Selection of Archival Materials," in *2018 IEEE International Conference on Big Data (Big Data)* (Seattle, WA: IEEE, 2018), 2721–24, https://doi.org/10.1109/BigData.2018.8622267; Thomas Padilla et al., "Final Report—Always Already Computational: Collections as Data," Zenodo, May 22, 2019, https://doi.org/10.5281/zenodo.3152935.

[14] See also the work at SPARC Data Analysis Working Group, "Data Visualization for Collection Development," last updated August 16, 2021, https://sparcopen.org/our-work/negotiation-resources/data-analysis/data-visualization-for-collection-development/.

[15] R. Litsey and W. Mauldin, "Knowing What the Patron Wants: Using Predictive Analytics to Transform Library Decision Making," *Journal of Academic Librarianship* 44, no. 1 (2018): 140–45, https://doi.org/10.1016/j.acalib.2017.09.004.

[16] Kiri L. Wagstaff and Geoffrey Z. Liu, "Automated Classification to Improve the Efficiency of Weeding Library Collections," *The Journal of Academic Librarianship* 44, no. 2 (2018): 238–47, https://doi.org/10.1016/j.acalib.2018.02.001.

[17] Conway and Proffitt, *Taking Stock and Making Hay*, 2011.

[18] Patricia J. Rettig, "Collecting Water: An Analysis of a Multidisciplinary Special-Subject Archives," *The American Archivist* 80, no. 1 (2017): 82–102, https://doi.org/10.17723/0360-9081.80.1.82.

[19] Qiana Johnson, "Moving from Analysis to Assessment: Strategic Assessment of Library Collections," *Journal of Library Administration* 56, no. 4 (2016): 496, https://doi.org/10.1080/01930826.2016.1157425.

[20] Chela Weber et al., *Total Cost of Stewardship: Responsible Collection Building in Archives and Special Collections* (Dublin, OH: OCLC Research, 2021), 8, https://doi.org/10.25333/ZBH0-A044.

[21] Leximancer, *Leximancer User Guide, Release 4.5*, Leximancer Pty Ltd, March 10, 2021, p. 104, https://static1.squarespace.com/static/5e26633cfcf7d67bbd350a7f/t/60682893c386f915f4b05e43/1617438916753/Leximancer+User+Guide+4.5.pdf.

[22] Leximancer, *Leximancer User Guide*, 104.

[23] Exports of Leximancer outcomes and all other supporting materials for this project can be found in the OSF project repository Text Analysis of Archival Finding Aids: Supporting Materials: https://osf.io/auxf4/.

[24] Leximancer, *Leximancer User Guide*, 11.

[25] Emily Haynes et al., "Semiautomated Text Analytics for Qualitative Data Synthesis," *Research Synthesis Methods* 10, no. 3 (September 2019): 459, https://doi.org/10.1002/jrsm.1361.

[26] *Describing Archives: A Content Standard (DACS)* (Chicago: Society of American Archivists, 2015).

[27] A description of the individual edits is available in the OSF project repository Text Analysis of Archival Finding Aids: Supporting Materials: https://osf.io/auxf4/.

[28] EAD 2002 W3C Schema, http://www.loc.gov/ead/ead.xsd.

[29] The detailed finding aid for the Ava Helen and Linus Pauling Papers (which fills six printed volumes) was not included; the truncated version was included: https://scarc.library.oregonstate.edu/findingaids/?p=collections/findingaid&id=286.

[30] The Biographical/Historical note was excluded from the set. Though it too is a narrative note, the typical Biographical/Historical note describes the background of the individuals and/or organizations which created the materials, and often does not directly relate to the content of the collection. Relational analysis of the Biographical/Historical note can yield incredible insights about the relationships between people, organizations, and families, as the Social Networks and Archival Context (SNAC) Project has shown: https://portal.snaccooperative.org/about.

[31] The code and generic documentation are available in GitHub: EAD2CSV, https://github.com/osulp/ead2csv. Documentation specific to this project is included in the OSF project repository.

[32] The code and documentation are available in GitHub: LCSH Broader Term Subject Analysis, https://github.com/osulp/LCSH-BT_subject_analysis.

[33] Components of the baseline subject analysis including the complete LCSH topic model for SCARC finding aids, the original set of subject headings, and overviews of both top-level and right-sized topics are available in the OSF project repository.

[34] Some percentage of broader-term chains follow highly questionable pathways from an original subject heading to a terminal term, and the reliability of this method of analysis warrants further examination.

[35] Settings documentation is available in the OSF project repository.

[36] Leximancer, *Leximancer User Guide*, 105.

[37] Full lists are available in the OSF repository.

[38] Leximancer, *Leximancer User Guide*, 105; Leximancer Monthly Webinar April 2022, https://www.youtube.com/watch?v=XQNnEcs7D2I.

[39] Leximancer, *Leximancer User Guide*, 105.

[40] See the online exhibit "Linus Pauling and the International Peace Movement," https://scarc.library.oregonstate.edu/coll/pauling/peace/narrative/page1.html. The peace emphasis is more developed in the detailed finding aid and less so in the truncated version.

[41] Alexis A. Antracoli et al., "Archives for Black Lives in Philadelphia: Anti-Racist Description Resources," Archives for Black Lives in Philadelphia's Anti-Racist Description Working Group,

last updated September 2020, p. 5, https://archivesforblacklives.files.wordpress.com/2020/11/ardr_202010.pdf.

[42] Erin Wolfe, “Natural Language Processing in the Humanities: A Case Study in Automated Metadata Enhancement,” *The Code4lib Journal* 46 (2019), https://journal.code4lib.org/articles/14834; Ryan Cordell, “Machine Learning and Libraries: A Report on the State of the Field,” LC Labs, Library of Congress, July 14, 2020, pp. 12, 32–33, https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf?loclr=blogsig.

[43] SCARC’s Anti-Racist Actions LibGuide details the department’s collective actions toward this effort since 2020: https://guides.library.oregonstate.edu/scarc-anti-racist-actions.

[44] Gregory, Geiger, and Salisbury, “Voyant Tools and Descriptive Metadata,” 15.

[45] Leximancer, *Leximancer User Guide*, 7.

[46] Greenberg, “The Applicability of Natural Language Processing,” 421.

[47] Yewno Discover, for example, creates knowledge graphs of related materials and was integrated into Ex Libris’ Primo discovery in 2020. Scott Scheutze, “Enhance Resource Exploration through A New Yewno App,” ExLibris, October 19, 2020, https://exlibrisgroup.com/blog/enhance-resource-exploration-in-primo-through-a-new-yewno-app/; Anne Bahde, “Conceptual Data Visualization in Archival Finding Aids: Preliminary User Responses,” *portal: Libraries and the Academy* 17, no. 3 (2017): 485–506, https://doi.org/10.1353/pla.2017.0031.

[48] Hutchinson, “Natural Language Processing,” 14–15.

[49] Haynes et al., “Semiautomated Text Analytics for Qualitative Data Synthesis,” 460.

[50] Ben and Sarah Brumfield, “Ten Ways AI will Change Archives,” email message, January 26, 2024. See for example Erin Wolfe, “ChronoNLP: Exploration and Analysis of Chronological Textual Corpora,” *Code4Lib Journal* 57 (2023), https://journal.code4lib.org/articles/17502.

[51] Cordell, “Closing the Loop,” 140, referencing Rediet Abebe et al., “Roles for Computing in Social Change,” in *FAT* ’20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM, 2020): 252–60, https://doi.org/10.1145/3351095.3372871.