IFLA WLIC 2018
KUALA LUMPUR

Submitted on: Jul 18, 18

Topic or Metadata Modeling for Cross-Disciplinary Scholarship: Challenges and Opportunities for Academic Libraries

Zheng Wang
University of Notre Dame, United States

Christina M. Leblang
University of Notre Dame, United States

Abstract

At the University of Notre Dame, we have been exploring automatic classification of texts via topic modeling and user-generated metadata to support cross-disciplinary scholarship. This effort originated in 2015 from a collaboration between the libraries and the Center for Civil and Human Rights to create an online comparative research tool to explore documents of Catholic social teaching and international human rights law. The library built the infrastructure for indexing, retrieving, and visualizing records while the researchers provided the controlled vocabulary and initial classification scheme.

From the onset, the project team realized there were limitations with current library classification standards and practices. To provide satisfactory discovery for cross-disciplinary content, the group "crowdsourced" the controlled vocabulary task to researchers and students of each respective discipline. Through the selection of controlled vocabulary, initial hand-tagging, and a more robust topic modeling, the researchers provided semantic linking of similar or different concepts at full-text and paragraph level. The modeling disambiguated terms (i.e., the use of child - biological vs. child of God) and bridged the gap between different disciplines description of equivalent concepts. For example, users can select the topic "solidarity/cooperation" and explore meaningful search results from the two fields about working together to improve human lives. The modeling enables a user from one discipline to

overcome the problem of nuanced vocabulary in the other domain and, hence, uncover relevant information that might otherwise remain hidden within the context of current classification schema.[1]

The project team is currently reconciling issues of transparency by providing detailed documentation on the application of the controlled vocabulary and in the process of implementing features for crowdsourcing data to enhance classification. The paper will present the updates of our topic modeling endeavor, and provide insights on considerations of the scalability and sustainability for academic libraries to support cross-disciplinary scholarship.

Keywords: Cross-Disciplinary Research, Topic Modeling, Classification, Controlled Vocabularies, Machine Learning

---

The Hesburgh Libraries at the University of Notre Dame, recognizing the increased need for the library profession to provide more sophisticated and innovative infrastructure, instruments, and tools to support trends in cross-disciplinary research, have been collaborating with the Center for Civil and Human Rights (CCHR) since 2013 to create an interdisciplinary research database that allows users to simultaneously search and compare documents of Catholic social teaching and international human rights law. The team completed phase one of the collaborative work in the spring of 2017; the online resource, *Convocate*, is now available for scholarly purposes as well as public usage. Throughout the project, the team realized there were limitations with current library classification standards and practices to support such a fusion endeavor. With an awareness of these limitations the library utilized both the domain subject expertise of research scholars and the automation of topic modeling to make *Convocate* a reality.

Given our experience with *Convocate*, the goal of the paper is to offer some insights for peers who are interested in similar projects as well as foster a global community to advance Library Science in the new area of machine learning and artificial intelligence. This paper starts with a review of the limitations of existing library knowledge systems as well as current library practices in facilitating cross-disciplinary research. In light of this review, one will note the need for libraries to gravitate towards new classification paradigms. We then present a case study of *Convocate* phase one work and ongoing tasks that details our own evolving approach for shifting classification paradigms.

Limitations of Existing Library Knowledge Systems

Over two decades ago, numerous visionary librarians identified multidisciplinary research as a significant issue of concern that would shape future library services (von Ungern-Sternberg, 1995; Searing, 1996; Palmer, 1996; Beghtol, 1998). Since then, Library Science professionals have continued to write about the significant challenges that multidisciplinary research poses for

[1] Solidarity is a pervasive concept in Catholic social teaching that recognizes the responsibility to work for the common good of others because we are all persons. Cooperation is a similar concept in legal documents that addresses the need for nations to work together to solve international problems of economic, social, cultural or humanitarian nature.

patrons and librarians in all disciplines (Denda, 2005; Ackerson, 2008; Romanowski, 2016). While there are differences between multidisciplinary and interdisciplinary research among scholars, authors tend to use the terms interchangeably when discussing cross-disciplinary approaches involving two or more academic disciplines.[2] From the perspective of Library Science and knowledge management, both multidisciplinary and interdisciplinary inquiry pose similar challenges in the description (cataloging) and discovery of resources. The shift to cross-disciplinary research in the academy has prompted some librarians to advocate for the optimization of conventional classification systems or the establishment of brand new methods and technological infrastructures to reconnect people to knowledge.

Today, cross-disciplinary study is a global norm rather than an anomaly as evidenced by the number of investments from universities, academies, corporations, and foundations, as well as governments. Nevertheless, researchers have a difficult time searching for information across disciplines because the terminology is unfamiliar (Ackerson, 2008). The causes of such suboptimal experience may be attributed to historically departmentalized knowledge classification systems and bibliographic controlled vocabularies.

Departmentalized Knowledge Classification Systems

Library bibliographic schema was originally designed assuming application for academic disciplines. This structural principle is no longer adequate. Multidisciplinary knowledge production and documents from communities of cooperation cannot be accommodated in a disciplinary structure (Beghtol, 1998). When the entire universe of knowledge is organized into disciplinary classes and subclasses and publications are written in languages catered towards specific fields, finding connections across classes and subclasses is complicated by the very inherited hierarchical nature of taxonomy systems. Von Ungern-Sternberg (1995) pointed out that it is crucial to reconsider the traditional reliance on discipline-based classification and it is necessary to try to solve the problems that orientation has created. This was one problem addressed in our work on *Convocate* that will be detailed in the case study section of the paper.

Bibliographic Controlled Vocabularies

A second shortcoming in library classification is the current standards for controlled vocabularies. Denda (2005) has indicated the challenges that libraries are facing to support multidisciplinary research in terms of relevance and usefulness of controlled vocabularies, such as LCSH. Controlled vocabularies are meaningful in setting authoritativeness, consistency, and standardization in terminology but with the creation of new knowledge, existing systems do not always explicitly represent newer territories and interdisciplinary associations that link them (Palmer, 1996). Often identifying new terms for an ever-increasing knowledge corpus does not occur on pace with the publication rate, which has also accelerated exponentially with the Digital Era. Many, such as Beghtol (1998) and Romanowski (2016), acknowledge the efforts from the library bibliographic control world to update controlled vocabulary on a regular basis, but cataloging rules and formats and library systems are out of synchronization and slow to evolve to meet these informational needs.

[2] Multidisciplinary research is understood to be research between two domains in which each domain remains in its own track. In contrast, Interdisciplinary research is a *fusion* of two fields.

In addition, the intricate relationships among concepts and ideas across areas are not adequately or intuitively expressed to facilitate productive research. Library vocabularies are not always in sync with the terms used by domain experts in their daily teaching and research. As a result, users who are interested in cross-disciplinary work have to be familiar with two more "taxonomy systems" beyond their native terminologies. Gross, Taylor, and Joudrey (2014) cited a study on author-assigned keywords and LCSH matches against abstracts of OSU ETDs. The results indicated that author-assigned keywords exactly matched words in the abstract 54.61 percent of the time, while cataloger-assigned LCSH correctly matched only 26.84 percent of abstract words; Keyword nonmatches occurred 10.59 percent of the time, and cataloger-assigned LCSH nonmatches occurred 31.08 percent of the time. Although they argued that the results indicate unique contributions of LCSH terms to discovery, one may also claim that it is time to augment LCSH terms with user-centered strategies.

Limitations of Current Library Systems

Over the past twenty years, library professionals have been experimenting with new techniques to resolve the aforementioned issues in an attempt to pave the path for cross-disciplinary activities. The objective of this work has been to identify relationships across disciplines in the hope of easing the difficulties for users in locating related resources. In theory or practice, librarians may have chosen to cross-assign notational access points under the present domain-based classification systems, such as Dewey or LLC. However, the success of a particular classification system at expressing a particular multidisciplinary topic through multiple notations will depend on the topic, the system, and how they interrelate (Beghtol, 1998). Given the current library climate, the number of shelf-ready collections and the massive quantity of electronic resources, this approach is neither sustainable nor scalable, Even more so it does not leverage the state-of-the-art technology to solve the puzzle.

Some suggested letting go of controlled vocabulary and solely relying on keyword searches. Although OCLC (2009, 2011) and librarians, for example, Gross (2014) and Romanowski (2016), indicated that keyword via search engines favored user search tactics, keywords alone cannot close the terminology gap across fields and may, therefore, leave specific resources invisible to users. A simple keyword search cannot identify the relationship between two words. Therefore, under the current technological context and profound disciplinary "silos," controlled vocabularies continue to be relevant in offering expert-guided discoveries. Even so, because of the limitations previously discussed, new relational instruments to connect concepts are critical to adapt to the change in research and scholarship. Some have created thesauri, systematically defining concepts and relationships, but Qin and Paling (2005) argue that they are less expressive and flexible as compared to machine-readable ontologies.

Gross (2014) cited a 2010 study by Nowick, Travnicek, Eskridge, and Stein. Their study investigated the use of controlled vocabulary versus keywords identified by automated text analysis or word clustering techniques for documents in an online environment. They explored similarities among search terms from users, the text of the documents themselves, and controlled vocabularies. Their findings showed that "the controlled vocabulary terms were better matched to both users' search terms and document terms than documents to users. Correlations between users and controlled vocabularies were 2–3 times higher [than] between users and documents.

This suggests that, through controlled vocabularies, libraries do provide a bridge between users and relevant documents. These results would indicate that human catalogers are the ideal way to organize documents into a library. However, given the limitations of humans to undertake a complete catalog of the internet, there may be ways to refine cluster-based organizing algorithms for digital libraries."

Ontological methodologies and semantic linking are two hot topics for facilitating enhanced search and discovery. There is a bright future for this category of strategies to aid in cross-disciplinary research as well. Once concepts and sub-concepts have been mapped and their relationships identified, computers can help point users to the precise information they are seeking. However, one must not underestimate the significant amount of work involved in mapping these concepts and sub-concepts. This mapping requires not only librarians but also domain specific experts and scholars. In addition, this type of practice will need to engage with computers and algorithms to automate and boost productivity.

Authors firmly believe librarians, as experts in classification systems, taxonomy, ontologies, and semantic/linked data, can have a tremendous contribution to influence the direction for solving the problems discussed. However as a profession, librarians must accept the equally important participation by domain experts of various disciplines, as well as the essential partnership with machines for scalability, extensibility, and sustainability.

Libraries ability in innovation in knowledge management determines the profession's relevance to patrons. The technological advancement is an overarching switch, which has altered users information seeking behaviors, and the changing nature of research has been calling for bibliographic classification systems' accurate and immediate response to all possibilities (Beghtol, 1998).

Convocate: A Case Study

The Problem as Understood from a Research Scholar's Point of View

The very idea for *Convocate* arose from research scholars' recognition of the gap in current resources for bridging the fields of Catholic social teaching and international law. Previously, for someone to successfully navigate cross-disciplinary research of these two fields, a research scholar would have to individually access the databases of not only Catholic social teaching documents and international law documents but also each individual organizations' database of documents. In other words, for example, to write coherently about child labor practices in Latin America in light of a strong Catholic identity and formation in social teaching, a research scholar would potentially search the databases of the United Nations, the International Labour Organization, and the Organization of American States for international law documents alongside the databases of the Vatican, and the Episcopal Conference of Latin America for Catholic documents. To streamline this process, research scholars wanted to create a tool that would bring documents from both fields together on one platform for simultaneous searching and side-by-side textual comparison. Beyond this general idea for such a tool, research scholars were unsure of the best methods to analyze documents and store document data, how to manage document permissions, and how to build the backend or user interface for such a tool. Given all of these limitations, research scholars turned to the Center for Digital Scholarship housed in the

libraries for consultation and eventually expanded this relationship to partner more broadly with the Hesburgh Libraries.

Creation of a Unique Controlled Vocabulary

The Hesburgh librarians encouraged the research scholars to use LOC subject headings for document classification. Given the set of foundational Catholic documents of interest, research scholars realized that most if not all documents would be returned for almost all of the possible searches related to the intersection of social teaching and human rights. Thus the decision was made to narrow in on each paragraph of text as a unique search result. Given the detailed level of searching desired, research scholars found LOC subject headings to be lacking in the specificity with which they wanted to apply to texts. Instead of utilizing LOC subject headings for a set of controlled vocabulary, research scholars created a unique set of controlled vocabulary. Initially, they gathered a list of potential terms from common search topics on significant websites and databases, such as the United Nations and the United States Conference of Catholic Bishops. The terms from both disciplines were then merged into one coherent hierarchical list of controlled vocabulary. This list was sent to senior scholars whose primary area of research lies in the intersection of international human rights and Catholic social teaching for review and critique. Finally, the controlled vocabulary was tested against the dataset by applying the controlled vocabulary for document classification. Through this last step, holes in the controlled vocabulary were identified and addressed. Besides, some terms were eliminated as they were found to be too specific and not widespread throughout the dataset.

Application of Controlled Vocabulary

Once the controlled vocabulary was fully identified, research scholars read through a set of documents from both fields applying the controlled vocabulary to each paragraph. No limits were set on the number of controlled vocabulary that could be applied to each individual paragraph. In addition, as a default, paragraphs in question were tagged with controlled vocabulary terms. The rationale for such an approach was that other research scholars would be given the opportunity to see these paragraphs on the periphery of a particular concept and make their own decisions as to whether or not to include these texts in their own personal work. Eventually, given the number of paragraphs (over 11,000) and controlled vocabulary terms (over 250), it became evident that the task of reviewing each paragraph for each term was too time-consuming. With the help of both librarians and computer scientists, research scholars turned to topic modeling to semi-automate the tagging process. The controlled vocabulary terms were mapped to various topics from the topic modeling algorithms in order to tag the remaining paragraphs. Due to the complexity of the chosen hierarchical controlled vocabulary, the terms were not necessarily mapped one-to-one with the topics from the computer algorithms. Currently the team is working to create a more automated approach of tagging that will utilize the current dataset as a training module for classification of new documents as they are added to the database.

Lessons from Convocate

Librarians and research scholars gained valuable insights through collaborative work on the *Convocate* project. These insights have reinforced what many librarians have voiced as the limitations of current library practices and highlighted the importance of controlled vocabularies,

domain-subject expertise, and the use of topic modeling and other automated computer techniques for document classification.

1. Importance of Controlled Vocabulary

While both fields speak about the same concepts, these concepts are represented by different vocabulary or technical jargon. For example, in regards to labor issues, Catholic documents largely use the language of "wage," and legal documents largely use the language of "remuneration." Even so, both wage and remuneration can be found in both types of documents. To capture more fully the results that speak about compensation for work, a person would need to search both "wage" and "remuneration." Moreover, supposing that someone is more heavily immersed in the legal field, that person might only search remuneration. Search results would still be returned in both fields but the scholar would miss a significant number of results, especially from the Catholic documents. Because scholars from different disciplines speak largely different "languages" to describe similar concepts or ideas, a set of controlled vocabulary as compared to a keyword search is necessary for a successful cross-disciplinary search.

2. Importance of Domain-Subject Expertise

The emphasis that academic libraries place on professional cataloging competence leads them, naturally enough, to trust and consult trained and skilled cataloguers to produce reliable and useful records. However, cataloging expertise is not the same as subject expertise, and no cataloguers can be expected to have an equally passionate interest in every subject (Campbell & Fast. 2004). Interdisciplinarity requires key knowledge of the concepts in more than one field, as well as familiarity with theoretical methodologies from different disciplines. Nature above requires that libraries involve scholars of those fields to establish an ontology for exploring topics (Denda, 2005). For example, from the Convocate project, solidarity is a pervasive concept in Catholic social teaching that recognizes the responsibility to work for the common good of others because we are all persons. Cooperation is a similar concept in legal documents that addresses the need for nations to work together to solve international problems of economic, social, cultural or humanitarian nature. In creating the controlled vocabulary for Convocate, research scholars chose to combine these two concepts: "solidarity/cooperation." By choosing "solidarity/cooperation", users will be able to explore meaningful search results that reflect the common threads from the two fields about working together to improve human lives. The topic search enables a user from one discipline to overcome the problem of nuanced vocabulary in the other discipline and, hence, uncover relevant information that might otherwise remain hidden within the context of current classification schema. It begs to ask how algorithms might recognize the link between the terms cooperation and solidarity without human domain experts to provide the initial connections.

3. Importance of Machine Automation/Topic Modeling

The quality of information retrieval can be measured according to the degree of match between the user and resources. If tools, such as ontologies, mirror the paradigm of a particular research community, research should be more effective (Denda, 2005). However, during the first phase of Convocate's development, the team discovered one of the most significant limitations was the arduous and time-consuming task of human creation and association of metadata. While it always will be necessary to train topic modeling systems to develop classification and concept associations, the current level of human intervention cannot scale to millions of documents. It is

clear that automation will be required. Gleaning from existing scholarly research on automated metadata creation and topic modeling, we know that libraries and organizations like OCLC have made progress in this area (Golub, 2006). Caragea et al. (2014) describe an automatic classification method applied to documents harvested from the Web. Danilevsky et al. (2014) outline a framework for topical keyphrase generation and ranking, based on the output of a topic modeling of short documents. The approach of Bijalwan et al. (2014) is similar to what we are considering, namely, "first categorize the documents using KNN based machine learning approach and then return the most relevant documents."(2014) From our perspective, none of the literature posits methods for comparing & contrasting seemingly disparate corpora.

Conclusion

The future of how librarians and libraries contribute to research and discovery is rife with emerging opportunities. Given that libraries are expanding how they can support research and the depth of their expertise in knowledge management, opening to innovative approaches can transform how research tools can support end users. Therefore, as cross-disciplinary research continues to shape future library services and affects patrons and librarians in all disciplines, now is the time to engage domain experts, librarians, and computer scientists to explore new possibilities in advancing cross-disciplinary knowledge creation based on user-centric design principles.

While the initial work of the Convocate project has demonstrated the value and need of cross-disciplinary approaches to discovery tools, it can have an even more meaningful impact by extending the methodology and framework of Convocate to other disciplines. Scaling the capabilities of the platform and broadening its scope of cross-disciplinary content to science, social sciences, and humanities could augment library science theory, knowledge, and practices with computational methodologies for supporting discovery and scholarship. Moreover, for libraries, there is an outstanding opportunity to build upon traditional library strength, such as classification, metadata creation, and collection stewardship, with new values and transition library employees to the knowledge workers of the future. Additionally, as attested in Convocate, this class of work positions libraries as collaborators in the overall scholarly endeavor. The authors hope this paper would call for a diverse community that engages in facilitating the creation and dissemination of cross-disciplinary research.

References

Ackerson, L. G. (2008) Challenges for Engineering Libraries, *Science & Technology Libraries*, 21:1-2, 43-52. DOI: 10.1300/J122v21n01_05

Beghtol, C. (1998). Knowledge Domains: Multidisciplinarity and Bibliographic Classification *Systems. Knowledge Organization*, 25(1), 1-12.

Bijalwan, V., Kumar, V., Kumari, P., & Pascual, J. (2014, 02). KNN based Machine Learning Approach for Text and Document Mining. *International Journal of Database Theory and Application, 7*(1), 61-70. DOI: 10.14257/ijdta.2014.7.1.06

Campbell, & Fast. (2004). Academic Libraries and the Semantic Web: What the Future May Hold for Research-Supporting Library Catalogues. *The Journal of Academic Librarianship*, 30(5), 382-390. DOI: 10.1016/j.acalib.2004.06.007

Caragea, C., Wu, J., Williams, K., Gollapalli, S. D., Khabsa, M., & Giles, C. L. (2014). Automatic Identification of Research Articles From Crawled Documents. Paper presented at the 2014 WSDM Workshop on Web-Scale Classification: Classifying Big Data. http://php.scripts.psu.edu/users/k/i/kiw5209/papers/2014/wscbd2014_caragea.pdf

Danilevsky, M., Wang, C., Desai, N., Ren, X., Guo, J., & Han, J. (2014). Automatic Construction and Ranking of Topical Keyphrases on Collections of Short Documents. *Proceedings of the 2014 SIAM International Conference on Data Mining*, 398-406. DOI: 10.1137/1.9781611973440.46

Denda, K. (2005). Beyond subject headings - A structured information retrieval tool for interdisciplinary fields. *Library Resources & Technical Services*, 49(4), 266-275, DOI: http://dx.doi.org.proxy.library.nd.edu/10.5860/lrts.49n4.266.

Golub, K. (2006) Automated subject classification of textual web documents. Journal of Documentation, Vol. 62 Issue: 3, pp.350-371, DOI: https://doi.org/10.1108/00220410610666501

Gross, T., Taylor, A., & Joudrey, D. (2014). Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching. *Cataloging & Classification Quarterly*, 1-39, DOI: 10.1080/01639374.2014.917447.

OCLC. (2009). *Online catalogs what users and librarians want : An OCLC report*. Dublin, Ohio: OCLC. Retrieved from https://www.oclc.org/content/dam/oclc/reports/onlinecatalogs/fullreport.pdf

OCLC. (2011) *Perceptions of libraries, 2010 context and community : A report to the OCLC membership*. Dublin, Ohio: OCLC. Retrieved from https://eric.ed.gov/?id=ED532601

Olson, H. (1996) *Between control and chaos: An ethical perspective on authority control*. Paper presented at the Authority Control in the 21st Century: An Invitational Conference. http://worldcat.org/arcviewer/1/OCC/2003/06/20/0000003520/viewer/file97.html.

Palmer, C. L. (1996). Navigating among the disciplines: The library and interdisciplinary inquiry - Introduction. *Library Trends*, 45(2), 129-133.

Romanowski, C. A. (2016) *A comparative analysis of the distinct evolution of cataloging and information technology towards the creation of the next generation library system*. Paper presented at the IFLA WLIC 2016 – Columbus, OH – Connections. Collaboration. Community in Session 93 - Cataloguing and Information Technology. http://library.ifla.org/1323/1/093-romanowski-en.pdf

Searing, S. E. (1996). Meeting the Information Needs of Interdisciplinary Scholars: Issues for Administrators of Large University Libraries. *Library Trends*, 45(2), 315–42.

Qin, J. & Paling, S. (2001). Converting a Controlled Vocabulary into an Ontology: The Case of GEM. *Information Research*, 6(2). Retrieved from http://informationr.net/ir/6-2/paper94.html