Submitted on: 22.06.2016

Classification & Indexing Satellite Conference
11 & 12 August 2016
State Library of Ohio, Columbus, Ohio, US

Aligning Author-Supplied Keywords for ETDs with Domain-Specific Controlled Vocabularies

Myung-Ja K. Han
Content Access Management, University of Illinois at Urbana-Champaign, Urbana, USA.
mhan3@illinois.edu

Patrick Harrington
Content Access Management, University of Illinois at Urbana-Champaign, Urbana, USA.
pharrin2@illinois.edu

Andrea Black
Content Access Management, University of Illinois at Urbana-Champaign, Urbana, USA.
atblack@illinois.edu

Deren Kudeki
Content Access Management, University of Illinois at Urbana-Champaign, Urbana, USA.
dkudeki@illinois.edu

Abstract:

Subject access can provide essential points of access for users to find, identify, select, and obtain various resources available in libraries. Subject access is not always available, however, due to the increasing amount of metadata created by non-catalogers (including author-supplied metadata), changes in libraries' discovery services, and a lack of best practices for aligning non-controlled vocabularies to authorized subject headings. This paper addresses the issue of author-supplied metadata, specifically how to align keywords submitted by authors of electronic theses and dissertations (ETDs) with Library of Congress Subject Headings (LCSH) and discipline-specific taxonomies by analyzing 32,696 keywords from 5,365 master's theses and doctoral dissertations submitted to the University of Illinois at Urbana-Champaign's institutional repository between 2010 and 2014. This paper shares findings from the data analysis, including that matching rates vary depending on college, with newer or rapidly-developing fields (such as the School of Molecular and Cellular Biology) having lower matching rates than traditional, well-established fields of study (such as the College of Agriculture, Consumer, and Environmental Sciences), and recommends that when keyword reconciliation is performed, it should be done with more than one authority in tandem for the *best results; when the LCSH and discipline-specific controlled vocabularies were combined, matching results were slightly or moderately increased.*

Keywords: electronic theses and dissertations, metadata reconciliation, subject access, author-supplied keywords, controlled vocabularies

Introduction

Subject access points have played an important role in providing a unique opportunity for users to find, identify, select, and obtain (IFLA 1998) resources that are described with the same or related subject terms and classifications. As identified by Hjørland & Kyllesbech Nielsen, subject access can be provided with different types of taxonomies, such as access points classified by provider or agent, i.e., author generated values, or access points classified by kind (2001, p. 260). For different types of works, libraries have employed controlled terms in closed systems, i.e., subject terms assigned by subject catalogers with taxonomies used by the library domain--mainly Library of Congress Subject Headings (LCSH)--in an online public access catalog (OPAC).

Subject headings improve discoverability of resources. According to Sapon-White et al. (1998) "[print] dissertations with subject headings ... are more likely to circulate (and circulate more often) than those without subject headings" (p. 291). However, most electronic theses and dissertations (ETDs) rely on author-supplied keywords, so libraries have to find a way to work with keywords for subject access. Many researchers argue that developing best practices which leverage author-supplied metadata to create controlled subject access could significantly reduce the time and expense of applying controlled vocabularies to resources without sacrificing the benefits of controlled terms in enhancing resource retrieval (Maurer et al. 2011, Lubas 2009, Richardson et al. 2008). As a way to identify best practices, Strader (2009) tried to see how these keywords matched with LCSH, and found that over half of author-supplied keywords (59.02%) did not have exact matches in LCSH. In replicating Strader (2009)'s results, Schwing et al. (2012) found that "keywords tend to represent more current, cutting-edge ideas, as well as terms that are more specific within the sciences" while "LCSH, in contrast, tends to be more stable and to connect to broader subjects" (p. 924).

More recently, the use of subject searching in the current discovery service environment has been a challenge to both libraries and users because of three outstanding changes: first, libraries have been moving from OPACs to web scale discovery services that enable access to both resources in the OPAC as well as articles and chapters available from major database subscriptions whose resources are described with more specific subject terms than are offered in LCSH (Larson 1991); second, libraries are now dealing with more and more metadata created by non-catalogers (e.g., author-supplied metadata) that often use subject terms not available in LCSH or other established controlled vocabularies; and third, libraries and vendors have not yet developed best practices to provide discipline-specific subject access services to users.

This paper examines the possibility of aligning author-supplied subject terms (keywords) in the metadata for ETDs with LCSH as previously tested by Strader (2009) and Schwing et al. (2012), and investigates further whether those keywords could be aligned better with already-established discipline-specific controlled vocabularies. Based on the lessons learned from the research, this paper suggests ways to improve the ETD metadata creation process and discovery services by exploiting available information technologies, including linked data services.

Data and Methods

For this paper, researchers analyzed 32,696 keywords assigned by authors to metadata records for 5,365 ETDs (3,270 doctoral dissertations and 2,095 master's theses from 72 departments in 18 colleges) submitted to the University of Illinois at Urbana-Champaign (UIUC)'s institutional repository, IDEALS, from 2010 to 2014 to see how authors supplied keywords that best describe their own work. The number of ETDs submitted over that time period is shown in Figure 1.

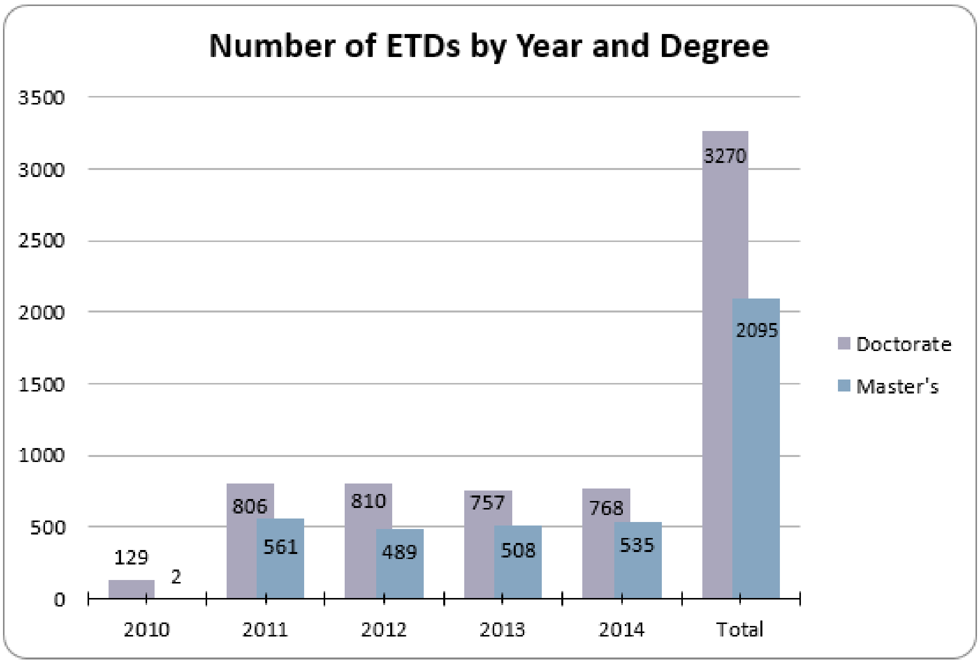

Figure 1: The sample data consists of 5,365 ETDs (3,270 doctoral dissertations and 2,095 master's theses) submitted to the University of Illinois at Urbana-Champaign.

The University of Illinois at Urbana-Champaign made electronic deposit of a thesis a requirement for all graduates in 2011 after a test trial in 2010, which explains the sudden jump in the number of ETDs from 2010 to 2011. When depositing a thesis to its ETD submission and management system, Vireo, a student is asked to provide keyword(s) that are then added into the IDEALS metadata in a Dublin Core Subject element and in data field 653 (Index Term – Uncontrolled) when transferred to a MARC record for an OPAC. Although adding a keyword is optional, all students but one provided keywords. In terms of utilizing these keywords, IDEALS currently provides a subject browse and quick search option because these are stored and labeled as subject(s) in its system. However, because these are not from controlled vocabularies, their performance is not ideal. For example, most of the keywords have only one associated thesis, so identifying related ETDs with the same, broader, or narrower subject term is not well-supported in IDEALS.

The researchers worked with author-supplied keywords in two ways: first, matching author-supplied keywords with LCSH terms and second, matching them with domain-specific controlled vocabularies. For the first matching process, LC's linked data service (id.loc.gov) was used for all keywords. For the second matching process, all keywords were divided into 18 colleges and then by degrees as used in IDEALS "communities."[1] The researchers then selected four sample colleges for the test: College of Agriculture, Consumer, and Environmental Sciences (ACES); College of Applied Health Sciences; College of Education; and College of Fine and Applied Arts, and identified domain-specific controlled vocabularies for each. These colleges were chosen to represent different disciplines of study and because all have strong domain-specific controlled vocabularies that support web search services.

[1] https://www.ideals.illinois.edu/community-list

Findings

Data Sets

Initial data analysis showed that authors provided an average 5-6 keywords per ETD (see Table 1) with one thesis having 177 keywords. The average number of keywords is similar in all 18 colleges. Between degrees, doctoral dissertations have slightly more keywords than master's theses. In terms of length of the keyword, the majority of keywords are made up of one word (10,945 keywords or 33.48%) or two words (14,637 keywords or 44.77%). There are 39 keywords that contain more than ten words, such as Sloan Digital Sky Survey Data Release 7 (SDSS DR7) Galaxy Angular Power Spectrum,' and two keywords that are 19 words long, such as 'electroactive polymers EAPs robotics flexible hyper-redundant robotic arm design partial differential equation boundary control PDE boundary control experimental validation.' Also noticed was the frequent use of acronyms, such as FPGA, ADV, and GPGPU. These acronyms were entered with or without full phrases giving their expanded meaning.

		2010	**2011**	**2012**	**2013**	**2014**
Doctorate	Average	1.07	6.37	6.53	6.62	6.86
	Min no. of keyword	1	1	0	1	1
	Max no. of keyword	10	37	70	54	34
Master	Average	4.5	5.43	6.03	5.86	5.33
	Min no. of keyword	3	1	1	1	1
	Max no. of keyword	6	21	70	177	33

Table 1: Authors add an average of 5-6 keywords per ETD.

Matches with LCSH

As the first step, the researchers tried to find matches with LCSH as Strader (2009) and Schwing et al. (2012) did in their earlier research. The matching results of this research were exactly the same as theirs. At this stage, a match simply meant that the web search service of the controlled vocabulary in question would return one or more results when a keyword was input, regardless of the accuracy of the results. When using the matching algorithm available in LCSH,[2] 15,552 among 32,696 keywords (47.6%) had matches with LCSH terms. To see whether there were any differences between colleges, these keywords and results were examined by college. Since the numbers of keywords were different, this comparison did not adequately represent how well LCSH describes one specific discipline or another. However, it shows that while traditional disciplines such as the School of Art and Design and the College of ACES have high matching results, new and emerging disciplines like the School of Molecular and Cellular Biology have low matching results. This finding is similar to findings by Schwing et al. (2012) as mentioned above. Matching results for different colleges are shown in Table 2.

College	**Keyword**	**Matches**	**Percent Matches**
School of Art and Design	61	37	**61.0**
Institute of Aviation	27	16	**59.0**

[2] http://id.loc.gov/authorities/subjects.html

School of Social Work	147	86	**59.0**
College of Applied Health Sciences	766	424	**55.4**
College of Veterinary Medicine	269	148	**55.0**
Neuroscience Program	236	129	**54.7**
Division of Nutritional Sciences	312	170	**54.5**
College of ACES	3,427	1,848	**53.9**
School of Earth, Society, and Environment	510	272	**53.3**
College of Fine and Applied Arts	1,501	783	**52.2**
Graduate School of Lib. & Information Science	494	258	**52.2**
College of Education	1,464	755	**51.6**
College of Liberal Arts and Sciences	9,063	4,675	**51.6**
College of Law	36	18	**50.0**
College of Media	68	32	**47.0**
College of Business	331	140	**42.3**
College of Engineering	13,548	5,586	**41.2**
School of Molecular and Cellular Biology	356	141	**39.6**

Table 2: Keywords divided by colleges and by percent matches with LCSH.

Matches with Domain-Specific Controlled Vocabularies

To assess the possibility of utilizing domain-specific controlled vocabularies for subject access services, the researchers tried to match author-supplied keywords with domain-specific controlled vocabularies identified for four colleges listed in Table 3. These were selected based on its web services and usage in the discipline. The keywords for each college were then searched against the domain-specific controlled vocabularies identified. Surprisingly, none of the domain-specific controlled vocabulary matches were better than the LCSH matching results. As shown in Table 4, only the National Agricultural Library's Agricultural Thesaurus (NAL-AT) had almost the same matching percentage as LCSH. Matches with other domain-specific controlled vocabularies are significantly lower than with LCSH.

College	Controlled Vocabulary
College of ACES	National Agricultural Library's Agricultural Thesaurus (http://agclass.nal.usda.gov/dne/search.shtml)
College of Applied Health Sciences	Health and Ageing Thesaurus Search (http://www9.health.gov.au/thesaurus/ThesaurusServlet?layout=initial)
College of Education	Education Resources Information Center (ERIC) Thesaurus (http://eric.ed.gov/)
College of Fine and Applied Arts	Getty's Art & Architecture Thesaurus® Online (http://www.getty.edu/research/tools/vocabularies/aat/index.html) LC Thesaurus for Graphic Materials (LCTGM) (id.loc.gov)

Table 3: Domain-specific controlled vocabularies used for four colleges.

College	LCSH Match (%)	Domain-Specific CV Match (%)
College of ACES	53.9	53.2

College of Applied Health Sciences	55.4	29.6
College of Education	51.6	34.0
College of Fine and Applied Arts	52.2	31.0 (Getty) 16.0 (LCTGM)

Table 4: Matches from LCSH and domain-specific controlled vocabularies.

These results showed that LCSH has more matches with author-supplied keywords than domain-specific controlled vocabularies, but not how many unique terms are actually matched. As such, the researchers looked further into the match results to see how many unique terms and matches were available from LCSH and domain-specific controlled vocabularies. The analysis revealed the same results; although there are several unique terms that are only available in domain-specific controlled vocabularies, the majority of the matches were found in LCSH. For example, among 766 keywords from the College of Applied Health, 203 keywords (26.5%) were uniquely found in LCSH, and only 6 keywords (0.8%) were uniquely matched in the Health and Aging Thesaurus (Figure 2). Only the domain-specific controlled vocabulary for the College of ACES showed similar matching results with LCSH. Among 3,427 keywords, 1,505 (43.9%) had matches in both LCSH and NAL-AT, while 343 (10.0%) had matches only in LCSH and 317 (9.3%) had matches only in the NAL-AT.

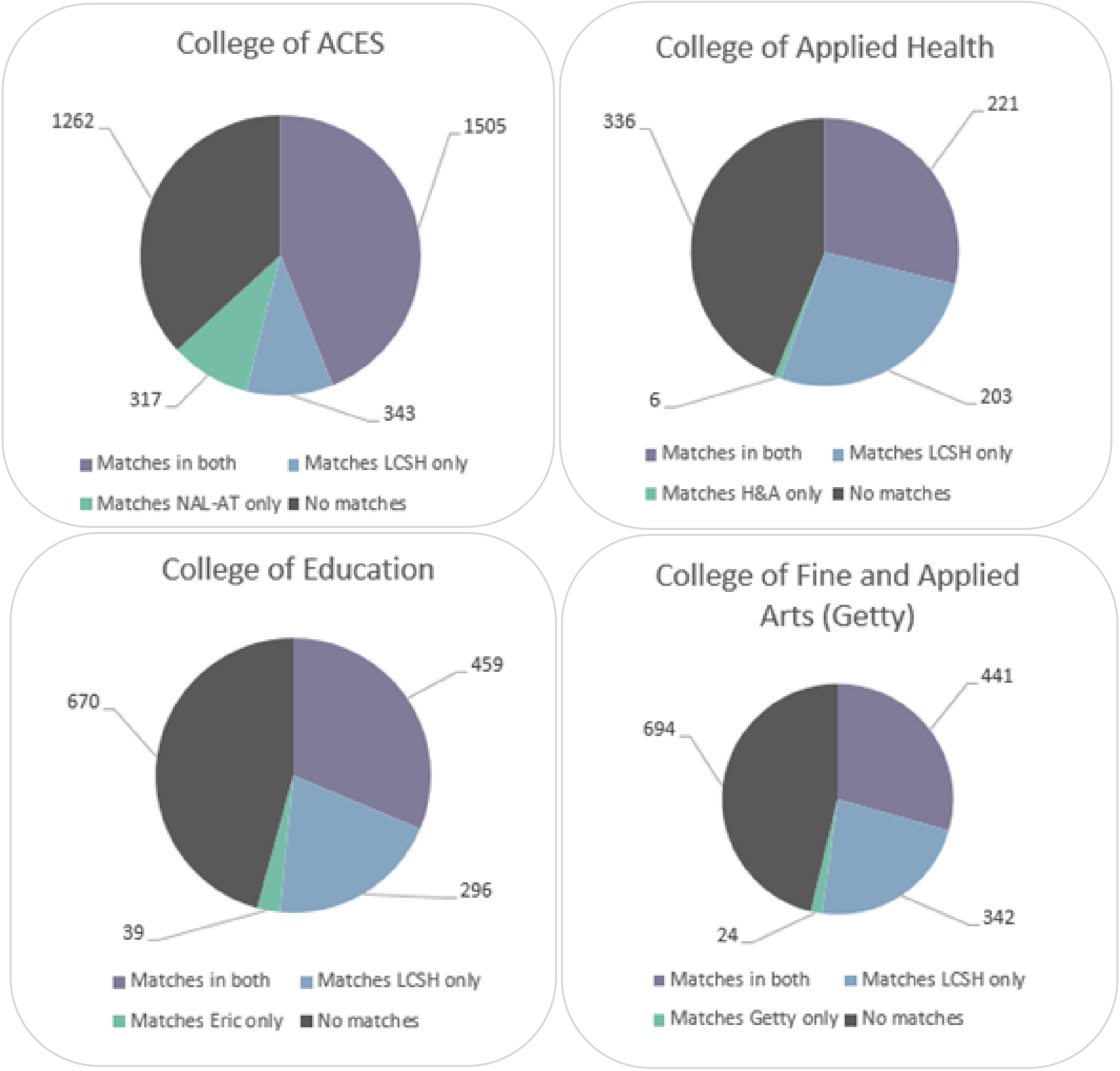

Figure 2: Matching results of author-supplied keywords with LCSH and domain-specific controlled vocabularies for four colleges.

The researchers speculate that this is because while domain specific vocabularies include specific subject headings, they do not include general subject headings. For example, while the LCSH-GM and Getty have matches for keywords like 'stained glass' and 'sketches' that are specific to the College of Fine and Applied Arts, they do not have keywords like 'project management' or 'convergence,' that are included in LCSH. This shows that LCSH works better for general subject headings while domain specific controlled vocabularies work better for specific subject headings.

The matching results of author-supplied keywords with LCSH and domain-specific controlled vocabularies also revealed that if combined, matching results are improved compared to results with just LCSH (Table 5). This is concordant with the previous finding that domain-specific controlled vocabularies include terms that are not available in the LCSH. For example, combined matching results for keywords in the College of ACES increased to 63.2% from 53.9%. Although the increase is not as big when compared with College of ACES, the other colleges' matching results also increased: 55.4% to 56.1% (College of Applied Health Science), 51.6% to 54.2% (College of Education), and 52.2% (LCTGM) to 53.8% (Getty) (College of Fine and Applied Arts).

College	**LCSH Match (%)**	**Combined with Domain-Specific CVs (%)**
College of ACES	53.9	63.2
College of Applied Health Sciences	55.4	56.1
College of Education	51.6	54.2
College of Fine and Applied Arts	52.2	With Getty: 53.8 With LCTGM: 52.3

Table 5: Matching results were improved when LCSH and domain-specific controlled vocabularies are combined.

Matching Results Analysis

As a next step in the data analysis, the researchers examined the keywords and the quality of retrieved controlled terms from LCSH and domain-specific controlled vocabularies. The matching results were reliant on each site's search-and-retrieval services and underlying database designs. Whether the underlying database supports hierarchies and relationships with other terms made an impact on matching results. The researchers wanted to see how many results were an exact match, false match, or a partial match by examining all keywords with retrieved terms one by one.

College	**LCSH**		**Domain-Specific CV**	
	Match	**Exact Match**	**Match**	**Exact Match**
College of ACES	1,848	1,056 (57.1%)	1,822	1,009 (55.4%)
College of Applied Health Sciences	424	157 (37.0%)	227	111 (48.9%)
College of Education	755	359 (47.6%)	498	391 (78.51%)
College of Fine and Applied Arts	783	240 (30.7%)	Getty: 465 LCTGM: 240	128 (27.5%) 83 (34.6%)

Table 6: Among the all matching keywords, domain-specific controlled vocabularies often have more exact matches than LCSH.

Two researchers hand-coded each match for LCSH and the five domain-specific controlled vocabularies. When reviewing terms, the researchers accepted different capitalizations, acronyms and full spelling (e.g., Branched chain amino acids and BCAA), term in singular and plural (e.g., African American and African Americans), words connected with a dash (e.g., After school programs and After-school programs) as exact matches. Otherwise, all other matches were treated as partial matches if the phrase includes the same term(s) or has the same meaning. The analysis revealed that although LCSH has a larger number of matching keywords, percentages of exact matches are lower than those of domain-specific controlled vocabularies (except in the cases of Getty and NAL-AT) as shown in Table 6. For the College of Education, the ERIC thesaurus shows 78.5% exact matching (391 of 498 matching keywords) compared with 47.6% exact matching in LCSH. Figure 3 shows detailed information about matching results.

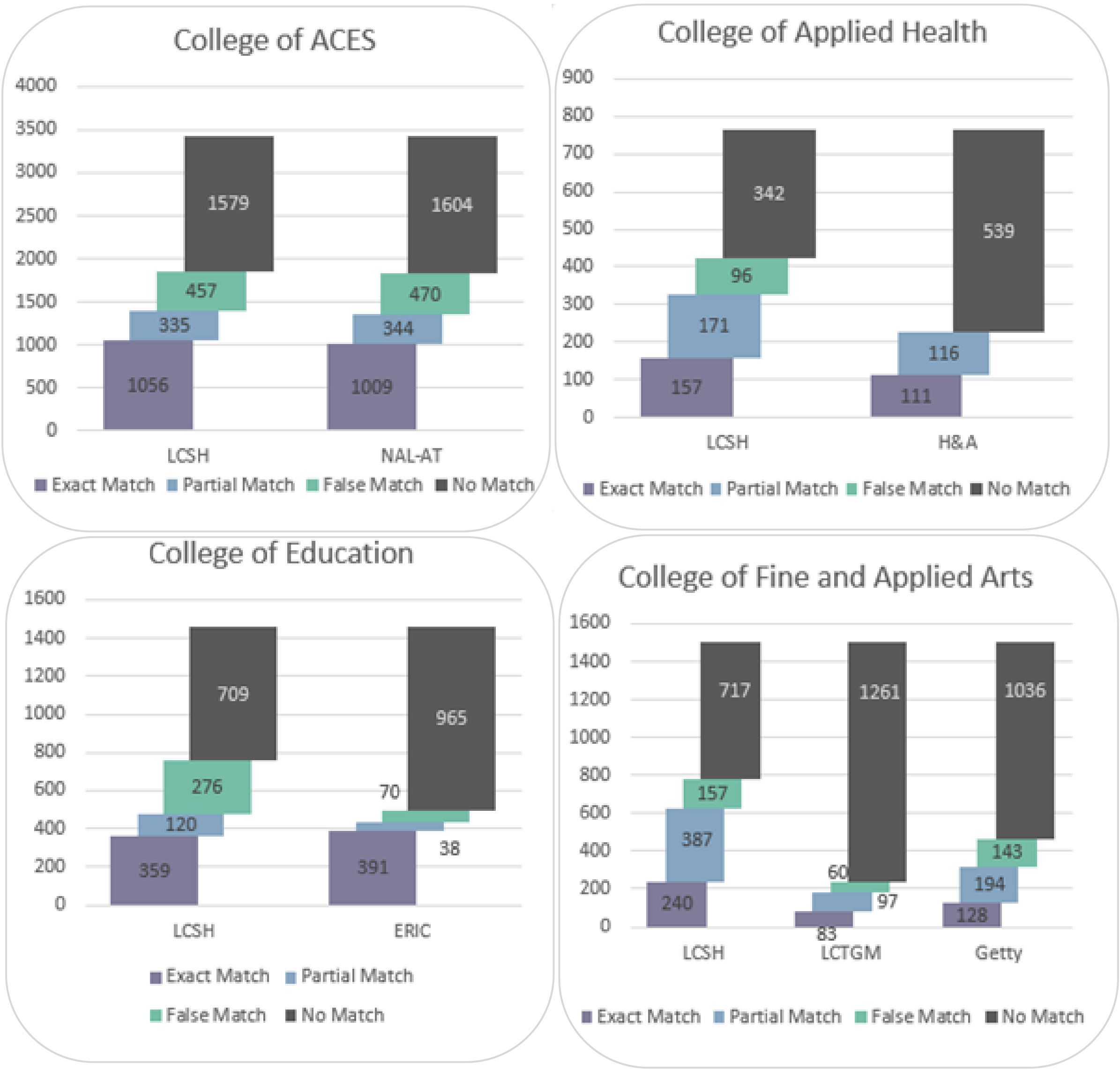

Figure 3: Types of keyword matches for LCSH and domain-specific vocabularies in each college examined.

Conclusion

Although most graduates who deposited their works through the ETD management system provided keywords, the system has not utilized them well enough in discovery or browsing services because they are not controlled terms. This study started with the hope and hypothesis that the discovery, use, and dissemination of ETDs could be increased by taking advantage of author-supplied keywords to increase subject access services by aligning them with already established domain-specific controlled vocabularies in addition to LCSH. Previous research that tried to find ways to use author-supplied keywords for providing subject access service by working with LCSH concluded that matching results are not ideal, and the metadata reconciliation work would be challenging.

This study tried to look further to see whether domain-specific controlled vocabularies match more often and more closely than LCSH with author-supplied keywords for ETDs. Based on the initial analysis of 32,696 keywords that describe 5,365 ETDs, 47.6% (15,552) of keywords match with LCSH, a result that is similar to previous findings. However, when matches were looked at more closely, the results were slightly better in traditional disciplines than new or rapidly-changing disciplines in schools such as the School of Molecular and Cellular Biology, the College of Engineering, and the College of Business, corroborating Schwing et al. (2012)'s finding that LCSH is not updated as quickly as new fields of study emerge in academia "due to the length of the review process for new LCSH terms" (p.905).

When matching the keywords from the ETDs of four colleges with five domain-specific controlled vocabularies, the matching results were lower than those with LCSH. However, the data analysis revealed that if LCSH and domain-specific controlled vocabularies are combined, the matching results are increased between 0.1% and 9.3%, depending on the controlled vocabulary used. In other words, domain-specific controlled vocabularies have unique terms that are not available in LCSH that could be useful in aligning additional keywords if remediation or reconciliation work is considered.

Another interesting dimension of the data analysis is the exact matching results. After examining keywords with matching terms from both LCSH and domain-specific controlled vocabularies, the researchers found that the domain-specific controlled vocabularies have better exact matching results than LCSH. The researchers speculate that this is because domain-specific vocabularies are more similar to the terms students use to describe their theses compared to LCSH and they have more flexibility when updating their terms than LCSH.

As libraries move toward linked open data and the semantic web, data cleanup will be required, preferably using controlled vocabularies from authorities with linked open data capabilities. Consequently, many libraries are contemplating and performing metadata reconciliation work. This study sheds light on two issues in the reconciliation of author-supplied keywords. First, no one authority alone is sufficient for reconciliation work; keywords for ETDs can contain very specific terms only used within a particular domain. As shown in the data analysis, there are terms that only appear in domain-specific vocabularies. When considering reconciliation work, libraries should use a combination of two or more authorities, not just LCSH or a domain-specific controlled vocabulary. Second, not all matches are exact matches, as noted in Europeana's *Report on Enrichment and Evaluation* (2015). All match results are based on the structure of the database and web search services provided on each site. Returning, for example, the first appearance of a keyword in any controlled subject heading instead of the most exact (e.g., returning "BAAV (Bovine adeno-associated virus)" when the keyword "bovine" is queried instead of returning the heading "bovine") further emphasized the need for cooperative development of best practices for both service providers and their users, e.g., libraries that use the service.

In addition, as many controlled vocabularies provide API services, perhaps it is time for libraries and system vendors to look at ways to integrate these services into the ETD submission process, allowing students to choose appropriate terms from domain-specific controlled vocabularies. This will ultimately improve subject browsing and searching services, as well as saving libraries from metadata reconciliation and remediation work.

Acknowledgments

Thanks to Seth Robbins for providing all ETDs metadata for this research.

References

Europeana. (2015). *Report on Enrichment and Evaluation*. Available at http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Enrichment_Evaluation//FinalReport_EnrichmentEvaluation_102015.pdf. (Accessed Feb 2, 2016)

Hjørland, B. & Kyllesbech Nielsen, L. (2001). Subject Access Points in Electronic Retrieval. *Annual Review of Information Science and Technology*, 35, 249-298.

IFLA. *Functional Requirements for Bibliographic Records, Final Report 1998*. Available at http://archive.ifla.org/VII/s13/frbr/frbr3.htm. (Accessed Feb 2, 2016)

Larson, R. R. (1991). The Decline of Subject Searching: Long-Term Trends and Patterns of Index Use in an Online Catalog. *Journal of the American Society for Information Science*, 42(3), 197-215.

Lubas, R. L. (2009). Defining Best Practices in Electronic Thesis and Dissertation Metadata. *Journal of Library Metadata*, 9(3-4), 252–263. http://doi.org/10.1080/19386380903405165

Maurer, M. B., McCutcheon, S., & Schwing, T. (2011). Who’s Doing What? Findability and Author-Supplied ETD Metadata in the Library Catalog. *Cataloging & Classification Quarterly*, 49(4), 277–310. http://doi.org/10.1080/01639374.2011.573440

Richardson, W., Srinivasan, V., & Fox, E. (2008). Knowledge Discovery in Digital Libraries of Electronic Theses and Dissertations: an NDLTD Case Study. *International Journal on Digital Libraries*, 9(2), 163–171. http://doi.org/10.1007/s00799-008-0046-9

Sapon-White, R. E., & Hansbrough, M. (1998). The Impact of Subject Heading Assignment on Circulation of Dissertations at Virginia Tech. *Library Resources & Technical Services*, 42(4), 282-91.

Schwing, T., McCutcheon, S., & Maurer, M. B. (2012). Uniqueness Matters: Author-Supplied Keywords and LCSH in the Library Catalog. *Cataloging & Classification Quarterly*, 50(8), 903–928. http://doi.org/10.1080/01639374.2012.703164

Strader, C. R. (2009). Author-Assigned Keywords versus Library of Congress Subject Headings. *Library Resources & Technical Services*, 53(4), 243-250.