User Search Terms and Controlled Subject Vocabularies in an Institutional Repository

Scott Hanrath, University of Kansas
Erik Radio, University of Arizona

Accepted manuscript of article published in Library Hi Tech, *Vol. 35 Issue: 3, pp.360-367,* https://doi.org/10.1108/LHT-11-2016-0133

Introduction

The application of controlled subject vocabularies is a means of improving resource discovery. By consistently applying subject terms across resources, similar resources can be more effectively collocated when searching or browsing by subject. Controlled subject vocabularies may appeal to managers of institutional repositories (IRs) because such vocabularies have the potential to increase the visibility and interoperability of repository content by employing standard vocabularies used in similar repositories or within similar disciplines.

However, applying controlled subject vocabularies to IR records can incur significant costs. This is especially true in cases where the controlled vocabulary is to be applied retroactively to repository content that has been submitted by a variety of users. Submitters may include, for example, authors who are unfamiliar with principles of information organization and cataloging. IR workflows may include little quality control of the submitted metadata values. Compounding the problem, IRs often include a wide range of content, from articles and gray literature to institutional records, in a wide range of disciplines. Such scenarios can result in a great diversity of subject and keyword terms applied unevenly across content and over a significant period of time. After-the-fact metadata remediation and enhancement thus potentially requires a great deal of effort.

Repository managers are therefore faced with value proposition: given the potential costs of applying a controlled subject vocabulary, how likely is it to positively impact the user's success in discovering repository content?

As part of a larger pilot project to gauge the impact of applying a controlled subject vocabulary to its IR, KU ScholarWorks, the University of Kansas Libraries sought methods to understand the search behavior of users in regard to subjects and subject vocabularies. If the discovery and use of IR resources is taken as a goal—not unreasonable for repositories that seek to increase the reach of scholarly resources through open access—then measuring how well previously successful searches for repository content align with a given controlled subject vocabulary is one way to assess the potential value of applying that vocabulary to repository records.

This article describes an exploratory study that examined the extent to which user search queries were aligned with OCLC's Faceted Application of Subject Terminology (FAST) vocabulary (OCLC, n.d.). FAST is derived from the Library of Congress Subject Headings. It has a broad range of terms, and was designed to be easy to apply and use, particularly in search interfaces that use facets to provide subject access. Both of these aspects make it an appealing choice for the repository environment. User queries were compared to both the FAST controlled subject vocabulary and the uncontrolled legacy subject terms in repository records. Values from other metadata fields applied to the downloaded content were also considered. The results provide IR managers with potential criteria for determining whether or not to invest in applying a subject vocabulary to repository records.

Literature Review

Beginning in the early 2000s, IRs have become a common service of many academic libraries (Lynch, 2013). Shreeves and Cragin define an IR as "a set of services and technologies that provide the means to collect, manage, provide access to, disseminate, and preserve digital materials produced at an institution"(2008). Dissemination of an institution's research output by making repository content available for indexing by commercial search engines or searching within a federated search system has been seen as beneficial to IRs, particularly those with a mission of providing open access to research(Arlitsch and O'Brien, 2012). Markland notes that given user preferences for conducting what they perceive as "quick" and "easy" searches using commercial search engines, particularly Google, how an IR's content is represented in Google is an important consideration (2006).

Reliance on full-text indexing is one reason that many institutions may not invest effort into enhancing metadata that is often ingested in a "mixed metadata environment" that may include uncontrolled, author-supplied metadata or metadata ingested from disparate existing systems that employ different descriptive practices (Chapman et al., 2009). Yang examined Google keyword searches that resulted in visits to a digital repository and compared the keywords with the values of metadata fields for the visited content, finding that search keywords most often matched values for Dublin Core Title, Description, and Subject fields, with other fields values matching less frequently (2016). Yang concludes repository managers may wish to pay particular attention to these fields as a means of increasing traffic from commercial search engines(2016).

White describes a study in which subject terms applied to scientific datasets by both information professionals and scientists were then mapped to four different controlled vocabularies and coded on a 1 to 5 scale from no match to perfect match(2013). The study found that the more general Library of Congress Subject Headings (LCSH) vocabulary most frequently mapped to topical subject terms, though with largely partial matches, while the Integrated Taxonomic Information

System vocabulary most frequently mapped to scientific terms used for biological species described in the datasets, with stronger, more complete matches (White, 2013). White concludes that controlled vocabularies provide value to repositories and recommends that repository managers "look closely at the terms applied by their users and choose vocabularies that appropriately match user needs"(2013).

Vállez et al. compared the correspondence between user search queries and subject descriptors from a subject-specific thesaurus applied to an academic journal by two different methods, human indexers and a semi-automated process, finding an overall low overlap of use search queries and the controlled subject vocabularies, with a higher overlap between user queries and the descriptors supplied via semi-automatic indexing (2015). They propose a method for employing logged user queries to improve controlled vocabularies, suggesting that the "adaptation of controlled vocabularies is crucial to optimizing the indexing process"(Vállez et al., 2015).

Chan, et al. summarize the impetus for the development of the FAST vocabulary, describing a goal of drawing on the rich semantics of LCSH, while providing a simplified, post-coordination-friendly syntax better suited to the Dublin Core metadata schema and to application by persons without specialized training in information organization (Chan et al., 2001; Dublin Core Metadata Initiative, n.d.). Qiang describes the results of an evaluation of FAST subject headings conducted by the Association for Library Collections and Technical Services' (ALCTS) Subject Analysis Committee(SAC)(2008). The ALCTS SAC reviewed a sample of bibliographic records with both LCSH and FAST headings, finding that the FAST headings were sufficient to cover the "aboutness" of the items, but that some context was lost when pre-coordinated strings from LCSH were broken up into multiple FAST headings(Qiang, 2008). For a detailed description of how LCSH terms are faceted into FAST, see Can and O'Neill (2010).

Methods

Data collected through Google Analytics were used to provide instances of successful searches for repository content. A successful search was defined as a case when a user reached an item page in the repository (i.e., a page that contains item metadata and link to a file containing the full item content, such as a PDF of an article) from a keyword search in an external search engine and the user then downloaded content from the item page. A custom "BitStream Click" event was created and tracked in Google Analytics. The event was recorded when a user followed a link to download a file, or "bitstream" in terminology of DSpace, the application used for the repository, from an item page.

Cases were limited to those meeting three criteria. First, the Landing Page tracked in Google Analytics, "the page through which the user entered the site," was an item record page (Google, n.d.). Second, the user's next recorded interaction was a BitStream Click event. Third, a Keyword from a search source was recorded. (Google Analytics does not provide Keywords in some cases, including cases when the search occurred over HTTPS)(Google, n.d.). These data provided a set of items from the repository where users arrived from a search and subsequently downloaded content. The search queries used by the user to retrieve the items were also available.

This study examined these cases from a four-month period from January through April 2015. The time period was selected to provide data from multiple months during an academic semester. During that period 2,208 unique combinations of item and search queries were recorded in our Google Analytics data. The data examined for this study do not include cases where a user reached an item page but did not download a file. The downloading of an item's file was taken as a confirmation that the item reached by the search was relevant to the user's query (though this excludes cases where a user may simply have been interested in retrieving or confirming metadata values, e.g., for including a citation in a bibliography). Cases where a user downloaded a file directly from an external search were also not included. In such cases the item's metadata record is bypassed as the metadata values are not always included in the file itself.

To determine how well the keyword search queries matched the FAST vocabulary, the set of successful search queries was imported into Open Refine. Open Refine (formerly known as Google Refine) is a software tool for cleaning, transforming, and extending data, particularly "messy" data("Open Refine", n.d). Open Refine provides convenient methods to match, or reconcile, a string of text or an identifier with an external vocabulary. The search queries were run again an Open Refine reconciliation service that uses OCLC's assignFAST autosuggest API to return likely FAST matches for a supplied query(OCLC, n.d.). Human judgment (the authors') was used to examine the matches returned by the reconciliation service and confirm cases where the FAST terms did in fact match the keyword search queries. For example, the query "baba yaga" was reconciled to the FAST entry for "Baba Yaga (Legendary character)".

To further explore the alignment of user queries to FAST terms, a small sample of 300 search query-item combinations was selected from the total population of 2,208 unique query-item combinations and manually analyzed by splitting queries into discrete concepts. The sample was obtained by assigning a random number between 0 and 1 for each query-item combination; query-item combinations were them sorted by the random number ascending order and the first 300 were selected. The sample size of 300 provides a confidence level of 90%, with a confidence interval of +/- 4%. Human judgment (the authors') was used to determine cases were search queries contained multiple discrete concepts. Those discrete concepts were then split into separate queries and used for a second round of reconciliation against FAST. For example, the query "cognitive disability and internet" was split into two queries, “cognitive disability” and “internet”, which were reconciled against FAST individually.

The study examined the 300 sample queries for three additional factors: the prevalence of known item searches, the prevalence of Google Scholar related article searches, and the degree of correspondence between search queries and item metadata for title and subject.

Human judgment was used to classify queries as known item searches, for example queries containing citations, article identifiers, or combinations of fragments of author names and titles. Such known item searches are unlikely to match a controlled subject vocabulary and so the relative prominence of such searches is a useful in determining the potential value of applying a controlled subject vocabulary.

For similar reasons, queries were classified as Google Scholar Related Article searches if they matched the form queries used in Google Scholar to retrieve articles related to a citation, i.e., “related:[identifier]:scholar.google.com”. The presence of such searches in the sample was not anticipated, but represents another case where applying a controlled subject vocabulary might not be expected to have a direct benefit for the discovery of content.

Finally, the degree of correspondence between search queries and item metadata for title and subject was examined by employing approximate string matching techniques to gauge the similarity between the queries and the metadata values. Levenshtein Distance (the number of single character replacements necessary to transform one term into another) was used to determine the similarity between the queries and the metadata values for title and subject. (For an overview of Levenshtein Distance and other methods for approximate string matching techniques, see Navarro, 2001). The similarity scores were normalized by string length, so that scores ranged from 0 to 100, where 100 indicated an exact match between the two strings, calculated as

$$score = 100\left(1 - \frac{\text{Levenshtien Distance }(Term1, Term2)}{\text{Max}\big(\text{Length}(Term1), \text{Length}(Term2)\big)}\right)$$

Values for the title and subject fields were compared with the original query term. In cases where multiple values for subject were present for an item, the best of score of all the comparisons was selected for the item. For example, the query “mt. fuji tourism” lead to an item with the subject includes values “Mt. Fuji”, “Tourism”, “Japan”, “Environmental Protection”, “Fujisan”, and “Unesco World Heritage Sites”. In this case, when the query was compared against each subject term, the highest individual score was 46.

Values for the subject field were also compared with manually spilt query terms. In cases were queries were manually split, each split value was compared against each possible subject value, with the best score selected for the item. For example, "mt. fuji tourism" was split into "mt. fuji" and "tourism". When each of these terms were compared with the each of the subject field values above, the highest individual score was 100.

Results

From January through April 2015, 2,208 unique item-search query combinations resulting from successful searches were collected. They represent approximately 15% of all BitStream click events from that period. The other 85% of BitStream click events occurred in cases other than a user landing directly on an item page from a search, or where Google Analytics did not record a search Keyword. Of those 2,208 item-search query combinations, 46 (2%) were successfully reconciled against the FAST vocabulary (see Table 1).

Of the random sample of 300 queries that were further analyzed, 97 queries (32%) were manually split into discrete topics and the queries were again reconciled against the FAST vocabulary. Of this set of queries, 84 of 300 queries (28%) were successfully reconciled against FAST. That is, in 28% of cases either the entire original query or at least one of the manually split query terms matched a FAST value (Table 1). That represents a large increase from the 2% of queries successfully reconciled from the entire set of queries without any attempt to split the query terms into discrete topics.

Fifty-two (17%) of the random sample of 300 queries were judged to be known item searches. A further 10 queries (3%) where found to be Google Scholar Related Item searches. Together, the known item and Google Scholar Related Item queries make of 20% of the sample, representing a fairly large number of queries that might not be expected to match terms from a subject vocabulary. Removing these queries from the sample, 34% of non-known item, non-Google Scholar Related Item searches (84 of 238) were successfully reconciled against the FAST vocabulary using Open Refine, a six percent increase over the sample as a whole (Table 1).

The results indicate that a low percentage of user search queries were close, one-to-one matches with terms from the FAST vocabulary. However, more than a quarter of user search queries were close matches to terms from the FAST vocabulary when user search queries were split into discrete topics. When obvious known item searches are excluded, just over a third of queries closely match FAST terms. The primary value of this finding is an indication that a substantial portion of user search queries to an IR appear to be topical in nature. Attention to subject description in IR records may then provide an opportunity to improve the search visibility of the content.

Of the 300 items in the sample of unique item-search query combinations, 298 items were found to have at least one value for Title in the metadata record. When the Levenshtien Distance score for the original query terms and the Title values was calculated, 8% of queries had a score of 75 or greater against the Title values, judged as a "good" match (see Table 2). For example, the query "lyric memorization techniques" was found to be a good match with a score of just over 79 for an item with the existing subject term "Memorization techniques." Just over 5% of queries had a score of 95 or greater, judged as a "very good" match. For example, the query "urban farming" was found to be a very good match with a score of 100 for an item with the existing subject term "urban farming" (indicating that the terms are identical).

One-hundred and ninety-seven items in the sample had at least one value for the Subject field (meaning that 103 items, or 34%, lacked a value for Subject). For Subject field values compared against the original query terms, 4% had a similarity score of 75 or greater (see Table 2). Almost 3% had a similarity score of 95 or greater. When the manually split query terms were compared with the values for Subject, 12% had a similarity score of 75 or greater, and 9% had a score of 95 or greater. Though not as large as the increase in FAST reconciliation, the manually split set of queries showed a nearly 7% increase in similarity scores compared to the full original queries.

The correspondence between user search terms and existing Subject field values was relatively low at under 5% of user queries closely matching Subject field values. As when comparing search terms with FAST terms, when manually split user queries are compared to Subject field values the closely matched percentage increases to nearly 10%. The match rate for Subject is higher than that for Title, but neither field seems to match a particularly substantial percentage of the examined user search queries. The comparatively high rate at which user search queries matched values from FAST compared to matches with existing subject terms again suggests that attention to subject description may improve the visibility of IR content.

Conclusion

A limitation of this study is that it does not address how metadata values contribute to the search visibility of IR content compared to the contributions of the full-text indexing of the content. While the results do indicate how closely queries match Subject field terms, there is no clear indication how the Subject field values influence the indexing of the IR items or their placement in search results. Arlitsch and O'Brien conducted a detailed study which found that providing metadata structured according to Google Scholar guidelines (e.g., transforming Dublin Core values to High Wire Press tags and including them as meta tags in the HTML document) can greatly improve the indexing ratio of IR content, though how individual fields or values contribute is less clear (2012). Moreover, Arlitsch and O'Brien note that a variety of technical factors (e.g., server performance, sending inappropriate HTTP response codes) may also impact the indexing of IR content (2012).

The mechanics of actually applying a controlled subject vocabulary to IR content can pose challenges, even cases where IR managers decide that expected benefits are worth the effort. As in this study, the use of approximate string matching techniques such as Levenshtien Distance and tools such as Open Refine may prove useful. Machine learning techniques for document classification may also be valuable, such as the approach Liu describes to improve SHARE metadata using data from the Public Library of Science (2016).

One potential advantage to applying a controlled subject vocabulary to IR content, even given the uncertainties of search engine indexing, lies in Linked Data. Yoose and Perkins provide a useful overview of Linked Data principles and initiatives in the context of libraries, archives, and museums(2013). IR managers may wish to expose repositories as Linked Data a means of increasing the interoperability of IR metadata, or prepare repository content to be migrated to a Linked Data-based system in the future. A controlled subject vocabulary with defined URIs for values (as opposed to simple string values) would support these cases. For example, use of the FAST term "Environmental protection" can be expressed using the URI http://id.worldcat.org/fast/913324, allowing for machine-readable inferences about the object to which the term is applied. Southwick, et. Al describe considerations that surfaced in an exploratory project to create or update controlled vocabularies to be useful in a Linked Data context, noting that staffing, work place changes, and costs are to be expected as part of such a transition(2015).

On a much broader plane the effort required to apply a controlled subject vocabulary should be measured against efforts to apply similar authority lists such as names and places. Names and subjects, as language, are all subject to diachronic transformation. IR managers may wish to identify those types of values that are most resilient to change and that also provide the greatest advantages for enhanced retrieval. Once these are determined a greater consistency can be achieved in a repository which serves not only the immediate moment but in some amount of perpetuity as well. Future research might examine the correspondence between controlled subject terms and user search queries within specific disciplines or specific genres of IR content to better understand how these may impact the value of applying greater control to names and subjects in IR records. Such assessments may also benefit from larger sample size providing greater explanatory power than the one used in this exploratory study

The role of controlled vocabularies has been frequently re-examined for its continued utility. The University of California Libraries issued a report that while controlled vocabularies can be useful, their value for subjects is less clear, especially where automatically extracted information like tables of content might serve as a surrogate for subject classification(Bibliographic Services Task Force of the University of California Libraries, 2005). By contrast, Gross, Taylor, and Joudrey (2015) found that subject vocabularies still provide a valuable discovery function for research.

That user queries appeared to more closely align with terms from the FAST vocabulary than uncontrolled, pre-existing subject terms suggests that there is a role for controlled subject vocabularies in enabling more effective retrieval in the context of an IR. But retroactively applying a controlled vocabulary, as well as designing the mechanisms for its ongoing implementation, can consume considerable time and resources. The benefit of applying a controlled vocabulary to subject terms must be weighed against its costs, including the opportunity cost of time and resources that could be used to enhance other IR metadata.

Tables

Table 1. User search queries successfully reconciled against FAST

	N	**Count reconciled**	**Percent reconciled**
All search queries	2,208	46	2%
Sample, spilt into discrete topics	300	84	28%
Sample, split into discrete topics, known item and Google Scholar related item searches removed	238	84	34%

Table 2. Correspondence between sample queries and metadata values in fields in downloaded item

Field	**Number of items with value for field**	**Percent of queries in score range**		
		< 75	**75 - 95**	**95 to 100**
Title	298	91.6%	3%	5.4%
Subject	197	95%	1.5%	2.5%
Subject, split into discrete topics	197	87.8%	3%	9.1%

References

Arlitsch, K. and O'Brien, P.S. (2012), "Invisible institutional repositories Addressing the low indexing ratios of IRs in Google Scholar", *Library Hi Tech*, Vol. 30 No. 1, pp. 60–81.

Bibliographic Services Task Force of the University of California Libraries. (2005), "Rethinking How We Provide Bibliographic Services for the University of California: Final Report", December, available at: http://libraries.universityofcalifornia.edu/sopag/BSTF/Final.pdf (accessed 1 November 2016).

Chan, L.M., Childress, E., Dean, R., O'Neill, E.T. and Vizine-Goetz, D. (2001), "A Faceted Approach to Subject Data in the Dublin Core Metadata Record", *Journal of Internet Cataloging*, Vol. 4 No. 1-2, pp. 35–47.

Chan, L.M. and O'Neill, E.T. (2010), *FAST, Faceted Application of Shubject Terminology: Principles and Applications*, Libraries Unlimited, Santa Barbara, CA.

Chapman, J.W., Reynolds, D. and Shreeves, S.A. (2009), "Repository Metadata: Approaches and Challenges", *Cataloging & Classification Quarterly*, Vol. 47 No. 3-4, pp. 309–325.

Dublin Core Metadata Initiative. (n.d.). "Dublin Core Metadata Element Set, Version 1.1", available at: http://dublincore.org/documents/dces/ (accessed 14 March 2017).

Google. (n.d.). "Google Analytics", available at: https://www.google.com/analytics/ (accessed 15 October 2016a).

Google. (n.d.). "Campaigns and traffic sources", *Google Analytics Help*, available at: https://support.google.com/analytics/answer/6205762 (accessed 15 October 2016b).

Gross, T., Taylor, A.G. and Joudrey, D.N. (2015), "Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching", *Cataloging & Classification Quarterly*, Vol. 53 No. 1, pp. 1–39.

Liu, J. (2016), "Classifying Research Activity in SHARE with Natural Language Processing", *SHARE News*, available at: http://www.share-research.org/2016/05/classifying-research-activity-in-share-with-natural-language-processing/ (accessed 1 July 2016).

Lynch, C.A. (2013), "Institutional Repositories: Essential Infrastructure for Scholarship In the Digital Age", *Portal: Libraries and the Academy*, Vol. 3 No. 2, pp. 327–336.

Markland, M. (2006), "Institutional repositories inthe UK: What can the Google user find there?", *Journal of Librarianship and Information Science*, Vol. 38 No. 4, pp. 221–228.

Navarro, G. (2001), "A Guided Tour to Approximate String Matching", *ACM Comput. Surv.*, Vol. 33 No. 1, pp. 31–88.

OCLC. (n.d.). "FAST Linked Data", available at: http://fast.oclc.org/ (accessed 5 January 2016a).

OCLC. (n.d.). "assignFast API", available at: http://www.oclc.org/developer/develop/web-services/fast-api/assign-fast.en.html (accessed 7 March 2017b).

"Open Refine". (n.d), , available at: http://openrefine.org/ (accessed 21 February 2017).

Qiang, J. (2008), "Is FAST the Right Direction for a New System of Subject Cataloging and Metadata?", *Cataloging & Classification Quarterly*, Vol. 45 No. 3, pp. 91–110.

Shreeves, S.L. and Cragin, M.H. (2008), "Introduction: Institutional Repositories: Current State and Future", *Library Trends*, Vol. 57 No. 2, pp. 89–97.

Southwick, S.B., Lampert, C.K. and Southwick, R. (2015), "Preparing Controlled Vocabularies for Linked Data: Benefits and Challenges", *Journal of Library Metadata*, Vol. 15 No. 3-4, pp. 177–190.

Vállez, M., Pedraza-Jiménez, R., Codina, L., Blanco, S. and Rovira, C. (2015), "Updating controlled vocabularies by analysing query logs", *Online Information Review*, Vol. 39 No. 7, pp. 870–884.

White, H. (2013), "Examining Scientific Vocabulary: Mapping Controlled Vocabularies with Free Text Keywords", *Cataloging & Classification Quarterly*, Vol. 51 No. 6, pp. 655–674.

Yang, L. (2016), "Metadata Effectiveness in Internet Discovery: An Analysis of Digital Collection Metadata Elements and Internet Search Engine Keywords", *College & Research Libraries*, Vol. 77 No. 1, pp. 7–19.

Yoose, B. and Perkins, J. (2013), "The Linked Open Data Landscape in Libraries and Beyond", *Journal of Library Metadata*, Vol. 13 No. 2-3, pp. 197–211.