Kansas Working Papers in Linguistics, Vol. 35 (2014), 1-8

Enhancing an Open-Access Linguistics Journal Archive with Library of Congress-like Metadata: A Case Study of the Effectiveness for Improving Discovery

Geoff Husic
University of Kansas Libraries

The University of Kansas has hosted the *Kansas Working Papers in Linguistic* (KWPL) in its digital repository, KU ScholarWorks, since 2007, but complete subject metadata, based on Library of Congress Subject Headings, was added only in 2013. The purpose of this study was to determine the efficacy of adding Library of Congress subject headings in enhancing discoverability and increasing downloads. Results have provided compelling evidence that the additional metadata has led to an overall increase of downloads from KWPL by 66% in the year following the addition of the enhanced metadata.

Keywords: metadata, Library of Congress Subject Headings (LCSH), federated searching, institutional repositories, open-access publishing, linguistics

1. Goal of the case study

The goal of this case study is to establish a possible correlation between the addition of Library of Congress Subject Headings (LCSH) as the primary subject metadata to records for a particular open-access online journal *Kansas Working Papers in Linguistics* (http://kuscholarworks.ku.edu/dspace/handle/1808/276) and any positive change in number of downloads from this collection after the addition of the enhanced metadata. By 'download' I mean an actual capture of the PDF file attached to the metadata. A download could be harvested by a web crawler, manually downloaded to a computer, or opened directly within a web browser. I felt that counting downloads rather than just record views would be more indicative of the impact of the papers on real-world users, as record views frequently occur as a result of machine indexing, without any human intervention.

This study is concerned only with the subject vocabulary employed rather than metadata scheme per se. It is hoped that the study will inform stewards of digital repositories of the benefits of supplying systematic subject metadata to collections with the view of enhancing discoverability.

2. Metadata and discoverability

As it is generally acknowledged that metadata, by the very nature of its intent and function, can enhance the discoverability of objects in a collection, whether digital or traditional, there has not been a great deal of discussion in the literature on the benefits of descriptive metadata in digital repositories per se. More work has recently been devoted to the topic of the challenges encountered in trying to add consistent, granular, and quality metadata to digital collections or how to assess these practices (e.g., Palavitsinis, 2014). However as far back as 1999, the ALA/

ALCTS/CCS/SAC/Subcommittee on Metadata and Subject Analysis released a report which made numerous recommendations about the use of controlled vocabulary as metadata in digital resources, and specifically recommended Library of Congress Subject Headings or the Sears Headings as appropriate options in many cases (ALA/ALCTS/SAC/Subcommittee on Metadata and Subject Analysis, 2013). The focus of more recent discussions has been rather centered on which alternative metadata schemes may be most appropriate for specific resources or collections.[1]

It is quite natural that a metadata scheme that is designed for a collection of botanical fossils will be more appropriate for describing the fossil collection than, let us say, a scheme designed for describing customs inventories. When dealing with the more familiar collections that libraries manage, e.g., books, journals, still and moving images, scores, etc.), there are a relatively smaller number of schemes that are commonly employed in the libraries of English-speaking countries (e.g., Library of Congress Subject Headings, or hereafter LCSH, the US National Institutes of Health Medical Subject Headings (MeSH), the NASA Thesaurus, or the Sears Subject Headings (Chan, 2010)).

A review of library-related discussion groups has encountered a number of discussions of supplying additional descriptive metadata to open-access digital repositories. Some libraries report using LCSH while others have relied on the Sears Subject Headings, which are constructed with somewhat simpler and less technical terminology than LCSH and are widely used in many journal-article databases. I have found no libraries in the literature that report using either these or other metadata schemes universally for all submissions. Most libraries, in fact, appear to have the submitters supply their own descriptive metadata, which are frequently called by users and labeled in submissions as 'keywords.' While somewhat useful for discovery, self-submission of keywords without consulting an authoritative thesaurus results in a body of metadata that will lack any degree of consistency, i.e., they are not a 'controlled vocabulary,' which in turn also naturally hinders efficient discoverability and collocation of works on similar topics.

Chapman et al. (2009) reviewed the metadata practices of DSpace institutional repositories at three American universities which seem to be fairly typical in their practices and institutional support. They found that enhanced, standardized descriptive metadata, while understood to be desirable, is often viewed to be either impractical due to workflow considerations, or not possible, due to lack of dedicated resources and expertise in the processing units. They also find a wide variation in the kind of subject metadata supplied for various collections, from some standardized vocabularies for some collections, keywords supplied by submitters, to none at all. Although five years have elapsed since the publication of this article, the generalizations discussed here still seem to hold true in many institutional repositories.

More recently Bundza (2014) has described a three-tiered taxonomy tagging model that is employed in the Digital Commons platform, the lowest level of which contains the equivalent to subject keywords (the two higher tiers correspond to academic discipline and academic department respectively). This system is flexible in that it allows non-specialists in metadata, such as faculty and researchers to self-submit and tag their submissions with subject terms. However, with only 1000 terms available in the pull-down menu system to cover all possible disciplines, this seems rather inadequate to tag collections for which a much greater degree of granularity is desirable. Her conclusion is that attempting to rely on the paper-submitters themselves to supply quality metadata has not proven to be successful.

[1] One such recent discussion can be viewed at CODE4LIB Archives (accessed 03/21/2014).

3. Background on KWPL

Kansas Working Papers in Linguistics (hereafter KWPL) is an annual publication of the University of Kansas Linguistics Graduate Student Association. Its extent is currently 351 articles in 34 volumes. KWPL is intended as a forum for the presentation of the latest original research by the students, faculty, and emeriti of KU's Department of Linguistics and other related departments, such as Anthropology, Slavic Languages and Literatures, etc. Submissions are carefully reviewed by the faculty and grad student editors to ensure that they meet the quality standards and accepted research methodologies for the field. Prior to 2005, KWPL had been published by the KU Linguistics Department in print format only, and was shared with a limited audience beyond KU, primarily with other universities with prominent Linguistics Departments, in exchange for their own working papers. Often these publications languished on meeting room bookshelves of the various linguistics departments, as they were not routinely indexed in the major citation indexes for the field such as *Linguistics and Language Behavior Abstracts* or the *MLA Bibliography*. Around this time, several universities had in fact announced that they would be suspending publication, as the printing and distribution costs had become too expensive.

When I was alerted about KU's Linguistics Department's dilemma, I contacted the KWPL faculty advisor and graduate editor with a bold proposal: that we move KWPL to a 'born-digital' model. At first there was a modest amount of skepticism, based primarily on their concern that KWPL would be broken into two separate entities, the older paper serial, and a new digital manifestation starting with volume 28. They also held a concern, common at that time, that an online publication wouldn't have the same appearance of legitimacy as a traditionally published paper journal. Nevertheless, they were receptive to the idea.

During this period KU librarians were being encouraged by our administration to think ahead to the possibilities of the new digital environment and to envision projects that could be done locally to leverage the newly acquired technological tools, such as KU's Spencer Research Library's improved scanning equipment and DSpace (the software which runs KU ScholarWorks, which hosts KWPL). I approached the then Associate Dean of Scholarly Communication with a request for funding to cover the digital scanning of the entire previous 24 paper volume back file of KWPL to deposit in KU ScholarWorks as a pilot project. He immediately lent his enthusiastic support. During the next several months, I worked with a librarian in our Special Collections Department and her student assistant, who scanned each article in each issue as a separate PDF file. This procedure went fairly quickly since the Linguistics Department had extra copies that we were able to guillotine so that they could be more easily feed into a scanner, as in 2005 there were not as many options for quicker bulk scanning available to us. We also were not able to perform OCR on most of the older issues due to the quality of the print and lack of resources to carefully edit the OCR output. In these cases, only the author, title, and abstract, if present, were searchable in KU ScholarWorks.

While waiting for the scanning to be completed, I had begun to edit each PDF in Adobe Acrobat to add consistent headers with the KWPL volume information, and adjusting PDF numbering to match the original pagination to avoid possible confusion when these works were cited. Once the scanning was complete, we created a separate container in KU ScholarWorks for each issue and I uploaded each article into KU ScholarWorks, copying the abstract, if available, into the abstract field. When this portion of the project was completed, all 24 former-paper

issues, and the first born-digital issue were available and the Linguistics Department was delighted with the results.

Beginning with issue 28 (2006) KWPL was issued exclusively in electronic format. With that issue, I continued to deposit each article in each volume with an abstract, if available, and I created one based on my own summary if it was lacking. In addition, from that point on I supplied Library of Congress-like (LC-like) subject descriptors (more on LC-like descriptors below) in order to enhance discovery. I tried whenever possible to specifically add descriptors that were succinct, neither too narrow nor too broad, and that might also contain terms that did not already appear in either the title or the abstract (also more on my approach to metadata below).

4. Discussion of the metadata supplied to enhance each KWPL article

As mentioned above, some articles already had abstracts, more so in the later issues than the earlier ones, as at some point abstracts became mandatory. In those cases with abstracts already present I attempted to add specifically those Library of Congress Subject Headings (LCSH) that would contribute metadata for terms not already found in the abstract or the title itself, thus adding additional elements for the purposes of discovery.

I should also note here that KWPL's DSpace platform does not provide any means to replicate the authority and cross-reference structures which form the backbone of most library catalogs, again, leaving skillful keyword searching as the only option for retrieving desired content.

I am not claiming that LCSH is necessarily the best choice for a metadata thesaurus to describe a linguistics collection, in fact LCSH in many ways is woefully inadequate and many of the available terms are out of date or constructed in a particularly counterintuitive fashion (cf. the tortuous heading ‘Grammar, Comparative and General—Syntax’ when something like ‘Syntax (Linguistics)’ would be much more functional. Nevertheless, it is the thesaurus that I am personally most familiar with. It also has the added advantage that in those libraries that employ federated search tools (KU uses Primo by Ex Libris), and where users have been instructed about the best LCSH's for their areas of research, these LCSH-like descriptors can help retrieve similar contents across databases. This is especially advantageous in searching the library catalog, which in most libraries will likely be using LCSH for books, journals, etc.

Above I have used the term ‘LCSH-like’ metadata. I would like to explain briefly what I mean by this term. As mentioned above, DSpace is designed to allow searching only by keyword. LCSHs, on the other hand, were designed to added great functionality in a catalog (card or electronic) due to its hierarchical structure. So, in essence, while these are in fact LCSHs, they currently lack in DSpace some of the functionality that is meant to be part and parcel of LCSHs. Nevertheless, they still provide valuable context for keyword searching.

In assigning headings to each article, I tried to adhere to the following principles: 1) I tried in all cases to make at least one LCSH for each article, even if that heading would not result in any additional unique descriptors. My rationalization for doing so was as follows: 1) In future search tools we may wish to rank these somewhat higher than a lengthy abstract when retrieving search results by relevancy, and 2) I attempted to adhere to the spirit of assigning LCSHs whereby one choses the most concise heading that is neither much narrower nor much broader than the concept being described. However, in the cases of lesser-known language I sometimes added a heading for the higher language family. As an example of this later case, for the article

"Theoretical Implication of the Great Menominee Vowel Shift" I included the following LCSHs: **Algonquian languages** and **Menominee language—Vowels**.

The following are a few examples of the kinds of additionally supplied metadata that I have added to enhance discoverability in the KWPL collection. One very frequent simple and effective tactic was adding the full periphrasis **X language** for the name of specific languages discussed in each article. It is nearly universal that the names of languages appear in abstracts and titles as the name only, e.g., "A Study of Quantifier Phrases in Thai." Merely adding the fuller descriptor, e.g., **Thai language**, can greatly assist discoverability in a database such as KU ScholarWorks, where searching is by keyword only and does not provide left-anchored browse searches as are available in most online library catalogs. The KU ScholarWorks DSpace software does however allow some Boolean searching, so that searching 'Thai language' will be a much more productive search strategy than searching 'Thai' alone.

Other common examples included:

- Cases where there are several possible terms for a particular linguistic concept. The most common is not necessarily always the one used in LCSHs. **Syllabication** (LCSH) versus **Syllabification**; **Nasals** versus **Nasality** (LCSH), or **Accents and accentuation** (LCSH) versus the more commonly written term **Stress**. In some cases the LCSH added will only apply to certain languages, e.g., **Indirect object** (LCSH) versus the sometimes related term **Dative** in those languages with this type of inflection. In these cases I have added the LCSH when it differs from the form found in the title or abstract, as it is useful for discovery.

- Cases where there are very different competing names for the same language: **Diegueño language** versus **Kumiai language** (LCSH). Similarly there are cases where the authorized form of the language name may be absent. In the article "Optimality Account of the Variability of the Third Tone Sandhi Domain in Mandarin" the term **Chinese language (**LCSH used for all Chinese dialects as well as for Han Chinese) in fact appears nowhere in the title or abstract.

- Cases where the name of the language has a variety of spelling conventions: **Kutenai** versus **Kootenai language** (LCSH), **Lakhota** versus **Lakota dialect** (LCSH), or **Uyghur** versus **Uighur language** (LCSH). These were an especially common addition.

- Cases in which the LCSH adds additional information that might not necessary be available in the title or abstract but may nonetheless draw out related and important topics, e.g., for the article entitled "Stress patterns in Hijazi Bedouin Arabic" I added the LCSH **Arabic language—Dialects—Saudi Arabia—Hejaz—Accents and accentuation**, which not only provides more useful context through the geographic aspect highlighted in the headings but also with the variant spelling 'Hejaz' as it appears in the LCSH.

In several instances I added some descriptors in order to help disambiguate similar concepts, e.g., **Assimilation (Phonetics)** versus **Assimilation (Sociology)** both of which are likely to appear in text as simply **Assimilation**. In a similar example, I added the descriptor **Politeness (Linguistics)**, which is the official LCSH. This could help those searching the entire KU

ScholarWorks database (or indeed Google Scholar which also indexes it) for topics on politeness but who are not searching for linguistic content, which are more appropriately covered by the LCSHs **Etiquette** and **Courtesy**. In this example the presence of that descriptor will probably be or more help to the end user who is scanning the list of results for useful items, or a more sophisticated searcher who is aware of how to use Boolean operators to exclude results not desired.

In two cases I deviated slightly from LCSH practice. In the first case I additionally supplied a subject term which is not the LCSH authorized form because I know it to be a term that is widely used in writing and speech by researchers of the topic, e.g. alongside **Persian language** (LCSH) I also supplied **Farsi language**. The second case relates to headings that I transcribed as, e.g. **English language—Dialects—United States—Minnesota**. In the authorized LCSH form of this heading the element **United States** would not appear (for all other countries except the United States in the case of a location within a particular country the name of the country appears before the smaller unit, c.f. **Chinese language—Dialects—China—Huian Xian** in LCSH practice. I felt supplying the United States in this case would be useful to users researching American English dialects.

5. Methodology

I chose to track statistics for one year from the point when I added the LCSHs to the collection (August 1, 2013) and then compare the results with the monthly averages for the preceding five years. Having monthly statistics for whole years is useful due to different usage patterns for scholarly resources generally seen in North American libraries at different times in the academic year. While I had access to statistics for the collection going back to 2007, I chose to begin my sample with August 2012 for two primary reasons: 1) That was the date that our DSpace repository KU ScholarWorks became indexed and accessible through our library's federated search tool (an Ex Libris Primo implementation), which appears to have immediately contributed to a larger number of views and downloads of KWPL papers; and 2) only 4 out of the 351 total papers were added to KWPL after the Primo implementation, a small number that was not likely to significantly alter the statistical results without being separately accounted for. For the purpose of this study I am therefore examining monthly statistics from August 1, 2012 to July 31, 2014.

6. Results

As mentioned above, my period of statistical observation begins with August 2012, in order to begin the sample after the implementation of our federated search tool, Primo, which had led to more hits of both record views and downloads, as expected.

For the year from August 1, 2012 to July 31, 2013, the average monthly downloads of papers from KWPL was 1681. This was somewhat higher than the average of 1163 monthly downloads in the five years before Primo, and this 44% increase is presumably due to the Primo indexing of KU ScholarWorks.

The LCSH metadata, which I had populated in a spreadsheet, was added to the KWPL collection on July 31, 2013 in a batch upload, so the new sample period was for August 1, 2013 through July 31, 2014.

From the period of August 1, 2013 to July 31, 2014, the average monthly downloads for the collection rose to 2788. This represents an approximately 66% increase in monthly downloads

from the collection in comparison to the period before the addition of the enhanced metadata and after the Primo implementation.

There were some rather dramatic increases in downloads for specific papers immediately after the addition of the LCSHs. Most noticeably, I noticed a major increase in downloads initiated in Portugal as based on IP. It made sense to suspect that a web crawler might have been crawling for materials specifically pertaining to Portugal or the Portuguese language. Indeed, the five papers in the collection to which I added 'Portuguese language' were among those that saw the most immediate dramatic increase in downloads. For example the paper "Oral Vowel Reduction in Brazilian Portuguese" (http://hdl.handle.net/1808/470) had been downloaded at an average monthly rate of 2.7 times between January and July of 2013, while in August the downloads rose to 14. After August, and for the remainder of the year the downloads fell back to a monthly average of 4.5, which is still higher than before the addition of the LCSHs, but consistent with the overall increase in the entire collections after those LCSHs were added. This increase is typical for the other four papers tagged with 'Portuguese language,' as well.

7. Discussion

Ideally, discussion concerning the appropriate metadata for any digital collection should occur at the very beginning of the planning stage. However the fact that there was a several-year lag time between the initial deposit of our KWPL papers (beginning 2007) and the addition of systematic LCSHs to each paper (July 31, 2013), allowed me to actually demonstrate the effectiveness of the metadata in improving discoverability in this particular collection.

It would have been instructive to examine more closely the origins of downloads from the collection to determine whether they were initiated by, say, academic institutions versus automatically harvested. However, due to privacy regulations, the software underlying KU ScholarWorks was designed to limit information that can be collected about the incoming IP to the country of origin where downloads occur.

I have not attempted to make any measurement of the cost-to-benefit ratio of processing this additional metadata by analyzing the cost of the indexer, in this case, myself, to see if the time and cost commitment is justifiable, from say, an administrative perspective, as this was beyond the scope of this modest study. But a good overview of such costs can be found in Miller (2011).[2]

It should also be mentioned that although in this case the additional metadata has proven to be effective, there are some drawbacks to the approach followed here. Ideally an institutional repository would have some kind of authority-controlled vocabulary system akin to what is usual in online library book catalogs. This is obviously much easier to achieve in a close-vocabulary ecosystem such as an online library catalog using LCSHs as the primary descriptive metadata. As mentioned above and in the literature, institutional repositories have been shown to employ a wide range of descriptive metadata schemes with wide ranges of consistency. In this case, the happy medium seemed to be to use the most widely used metadata scheme found in academia to supplement the words found in the papers' titles and abstracts, and thus this approach has achieved its goal in enhancing discoverability as evidenced by increased downloads.

Results for other kinds of collections using the methodology outlined here may yield quite different results. Collections, the majority of objects in which have detailed abstracts and full-

[2] Miller (2011:100) in particular presents a good, brief discussion of some of the costs associated with this kind of metadata creation.

text indexed, especially in disciplines such as many scientific fields where vocabulary is naturally highly standardize, may not gain as dramatic additional benefit by adding subject metadata using a very general thesaurus such as LCSHs. However, fields such as the humanities and the social sciences, where terminology is very labile, may see substantially improved discoverability. Perhaps it is the recognition of the benefits of improving metadata practices that is behind the recent dramatic growth in the number of advertisements for library metadata specialists in 2013 and 2014.

References

ALA ALCTS/SAC/Subcommittee on Metadata and Subject Analysis (Issued July, 1999). *Subject Data in the Metadata Record: Recommendations and Rationale*. http://www.ala.org/alcts/resources/org/cat/subjectdata_record (Accessed12/05/2013).

Bundza, M. (2014). The Choice Is Yours! Researchers Assign Subject Metadata to Their Own Materials in Institutional Repositories. *Cataloging & Classification Quarterly*, *52*(1), 110–118. doi:10.1080/01639374.2013.852439

Chan, L. M. (2010). *FAST: Faceted Application of Subject Terminology: principles and applications*. Santa Barbara, Calif: Libraries Unlimited, p. 6. "Library of Congress Subject Headings (LCSH) is the largest and most widely used of all extant subject headings lists."

Chapman, J. W., Reynolds, D., & Shreeves, S. A. (2009). Repository Metadata: Approaches and Challenges. *Cataloging & Classification Quarterly*, *47*(3-4), 309–325. doi:10.1080/01639370902735020

CODE4LIB Archives. https://listserv.nd.edu/cgi-bin/wa?A1=ind1308&L=code4lib#175 (accessed 03/21/201)

Miller, S. J. (2011). *Metadata for digital collections: a how-to-do-it manual*. London: Facet Publishing.

Palavitsinis, Nikos, Nikos Manouselis, and Salvador Sanchez-Alonso. "Metadata Quality in Digital Repositories: Empirical Results from the Cross-Domain Transfer of a Quality Assurance Process." *Journal of the Association for Information Science and Technology*, April 1, 2014 (online preprint). http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%292330-1643/earlyview

Author contact information:

Geoff Husic: husic@ku.edu
ORCID ID: http://orcid.org/0000-0002-0617-7160
University of Kansas Libraries
1425 Jayhawk Blvd, Lawrence, KS 66045