Measuring the Impact and Effectiveness of Transitioning to a Linked Data Vocabulary

Erik Radio
Office of Digital Innovation and Stewardship, University of Arizona, Tucson, Arizona, USA
radio@email.arizona.edu

Scott Hanrath
Information Technology and Discovery Services, University of Kansas, Lawrence, Kansas, USA
shanrath@ku.edu

Abstract

Language used to describe resources in an institutional repository may benefit from the consistency offered by a controlled vocabulary as well as introduction into the larger linked data universe. Transitioning to a linked data vocabulary presents concerns regarding the effectiveness of mapping pre-existing terms and the potential for semantic loss. This paper describes such a transition to OCLC's FAST vocabulary in the University of Kansas' institutional repository. It analyzes the outcomes of this transition and its subsequent impact on resource usage when exposed as linked data.

Keywords: controlled vocabulary; institutional repositories; linked data; usage assessment; Faceted Application of Subject Terminology (FAST)

Though many of the structures required for its development are in place, the long-term vision of the Semantic Web for transforming the way information is organized and retrieved on the web is far from being achieved. An important component of the developing linked data (LD) ecosystem is the incorporation of machine-readable identifiers into resource metadata in the form of URIs. Much emphasis is placed on the potential benefits of LD-enhanced discovery. But LD records may also benefit from the consistency offered by use of a controlled vocabulary as necessitated by the use of unambiguous URI identifiers, particularly in contexts where such control had not previously been exercised. Because participation in LD activities requires adherence to the principles of a controlled vocabulary, it raises questions as to the appropriateness, effectiveness, and impact of adopting a particular vocabulary.

A principal consideration for identifying an appropriate LD vocabulary is assessing the descriptive needs of the corpus it will serve. An obvious benefit of uncontrolled terms is the ability to freely describe resources to any desired degree of specificity and granularity. While this is not without drawbacks, it does raise the question of how a LD vocabulary can be widely descriptive without being reductive. For a homogeneous corpus this is a more readily achieved criteria. A selection of similar resources will benefit from a more granular vocabulary. The opposite is potentially, though not necessarily, true as well: a more diverse corpus, both in type and content, will require a more broadly enumerated descriptive domain. A vocabulary that attempts to be universal in both scope and detail can become bloated and even run the risk of resembling an aggregated list of uncontrolled terms. Yet such a system of diverse resources is often an accurate description of an institutional repository (IR).

IR metadata has been frequently identified as inconsistent in quality, a term used hesitantly as there are various metrics for assessing this that are not all congruent. Incorrect use of terms and a lack of authority control are two commonly cited issues that hinder effective retrieval, particularly where metadata are created by authors, rather than catalogers or other information professionals. Adopting a LD vocabulary would appear to serve two needs: increase overall quality through data consistency, and greater resource exposure when records are serialized as LD. Among the questions faced by anyone seeking to apply a LD vocabulary to an IR are: how well does a new vocabulary reflect a more efficient and semantically meaningful version of the original user-created terms? If it doesn't, what does this mean for the role of older terms? Are there certain subject areas for which a LD vocabulary is more representative? Will there be any measurable impact on resource usage? The following article describes an exploratory transition to a controlled vocabulary in the University of Kansas' IR by measuring the accuracy of mapping preexisting terms to a new vocabulary and the consequent impact on resource discovery and usage.

Context

IRs are an increasingly common feature of academic institutions' digital collections. Lynch succinctly describes them as

> "a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its members." (Lynch, 2003).

As such they are most commonly used to disseminate scholarly articles, books, and similar resources, but have also grown to absorb a plethora of ancillary materials such as raw data, committee minutes meetings, conference presentations, and a host of other resource types that in some way reflect the intellectual and cultural nature of the institution (Duranti, 2010). As a kind of information system and cultural object, it is not difficult to see how the effective management of an IR is dependent on a logical structure and metadata that enables basic retrieval. Similarly, as a reflection of its institution it in many ways serves as a microcosm of the institution's academic and administrative landscape.

The scope of IR content is a known source of potential problems for metadata management. As Chapman et al. note,

> "Repositories often include metadata coming from a range of disciplines, each of which have different citation traditions and different emphases on the type of information they share...Metadata can be sparse or lack important contextual information particularly when that context is held at a collection level. The breadth and depth of disciplines across an academic institution means that use of controlled subject terms is possible at only the highest levels." (Chapman et al., 2009)

To compound the problem, the considerable number of heterogeneous items being ingested at any given time makes detailed item-level cataloging an expensive and usually untenable proposition for library personnel and resources. It was originally anticipated that an IR would rely on self-archiving by resource creators as a part of a regular ingest process, with the added benefit that their deep knowledge would allow for descriptively rich metadata. The opposite, as McDowell found, has tended to be the case and consequently many libraries have moved to a model where they assist in the description of resources (McDowell, 2007). Yet for a variety of reasons there is a high level of inconsistency across repository metadata and efforts to clean or enhance existing metadata are not widely undertaken (Chapman et al, 2009).

The history of the IR at the University of Kansas, KU Scholarworks (KUSW), very much follows a trajectory common to other repositories. Initiated in 2003 with the shared perception that self-archiving would provide a steady stream of content for the repository, workflows have since shifted towards assisted deposits, as the perceived number of faculty able to regularly ingest their scholarly output was much smaller than anticipated. While some faculty continue to self-deposit, librarians, professional staff, and student workers now complete the majority of the work to recruit and deposit content. However, the mixed workflow and scope of content has led to the same problem encountered in other repositories: a high level of inconsistent metadata element usage across records.

While there are many places where adherence to an authority source, like a controlled vocabulary or name authority file, would improve data consistency in KUSW, subject terms have been of immediate focus. Over a decade's worth of mostly uncontrolled terms has lead to the creation of a substantial user-driven vocabulary. Unfortunately, the majority of these terms are not controlled vocabulary subjects and many are keywords (e.g. names of particular proteins). Determining the value of both is an important question and has been examined in depth by Gross et al. who maintain that a significant portion of recalled documents will be lost with the absence of subject terms in favor of reliance on keywords and other terms present in a record (Gross et al, 2014). While the concern over less effective retrieval is warranted, the positive side of transitioning to a controlled vocabulary is the opportunity to optimize current data and provide greater consistency across the repository. While KUSW administrators have done significant remediation work to standardize term variations, introduction of a controlled vocabulary would hopefully help to ensure that this kind of remediation work need not be repeated. A secondary goal is the possibility of enhanced discovery through exposure of records as LD.

FAST as Vocabulary

Ideally a controlled vocabulary adequately represents the subject content of the corpus it is describing by being comprehensive without becoming needlessly granular. The Library of Congress Subject Headings (LCSH) represent a rich selection of terms, but its complexity limits its usability by untrained catalogers (Dean, 2010). This is a central concern for KUSW administrators given that most cataloging of KUSW resources are currently undertaken by student workers; incorrect or improperly formatted headings would in many cases be less preferable than no heading at all. KUSW is not the only digital collection with a legacy of uncontrolled metadata terms at the University of Kansas, so finding a vocabulary that could be well-suited to other collections would be an additional benefit.

OCLC's Faceted Application of Subject Terminology (FAST) ("FAST", 2013) is a vocabulary adapted from LCSH with a particular emphasis on usability and functional appropriateness for various web platforms. Because an explicit objective of FAST is to be “faceted-navigation-friendly”, it lends itself well to the analogous faceted feature of the DSpace architecture KUSW runs on (“FAST”, 2013). From a content perspective, its origins in LCSH made it likely to be appropriately representative of the diverse resources in the repository. Indeed, a similar mapping of a diverse vocabulary, Ulrich’s subject headings, to FAST proved to be largely successful with only a few and easily resolvable constraints (Mitchell and Hsieh-Yee, 2007). Finally, FAST subject terms are supplied with a URI allowing for a serialization of resource metadata into LD.

Mapping to a controlled vocabulary raises questions about the historicity of terms and potential for metadata loss as the result of a large-scale mapping effort. Mapping from an already controlled vocabulary like Ulrich’s suggests that a successful mapping of one term would mean a more substantive system-wide transformation. Conversely, a successfully mapped free-text term would not necessarily create a large transformation across the corpus. In the case of KUSW many of the uncontrolled subject terms are only used once. Interestingly, a previous experiment in mapping user-generated subjects to several different controlled vocabularies demonstrated that LCSH provided the highest number of matches for topical terms given its broad scope (White, 2013). A similar outcome for FAST would be a logical conclusion.

Terms that do not have any clear match in a controlled vocabulary are problematic in a controlled environment. Regardless of their semantic value, the terms are potentially valuable as a kind of artifact of legacy metadata. An argument could be made that terms supplied by the creator of a resource are of more interest than those provided by library personnel in an assisted deposit. Whether unmapped terms should be left as keywords, reconciled against another vocabulary, or discarded if not actually subjects is addressed below.

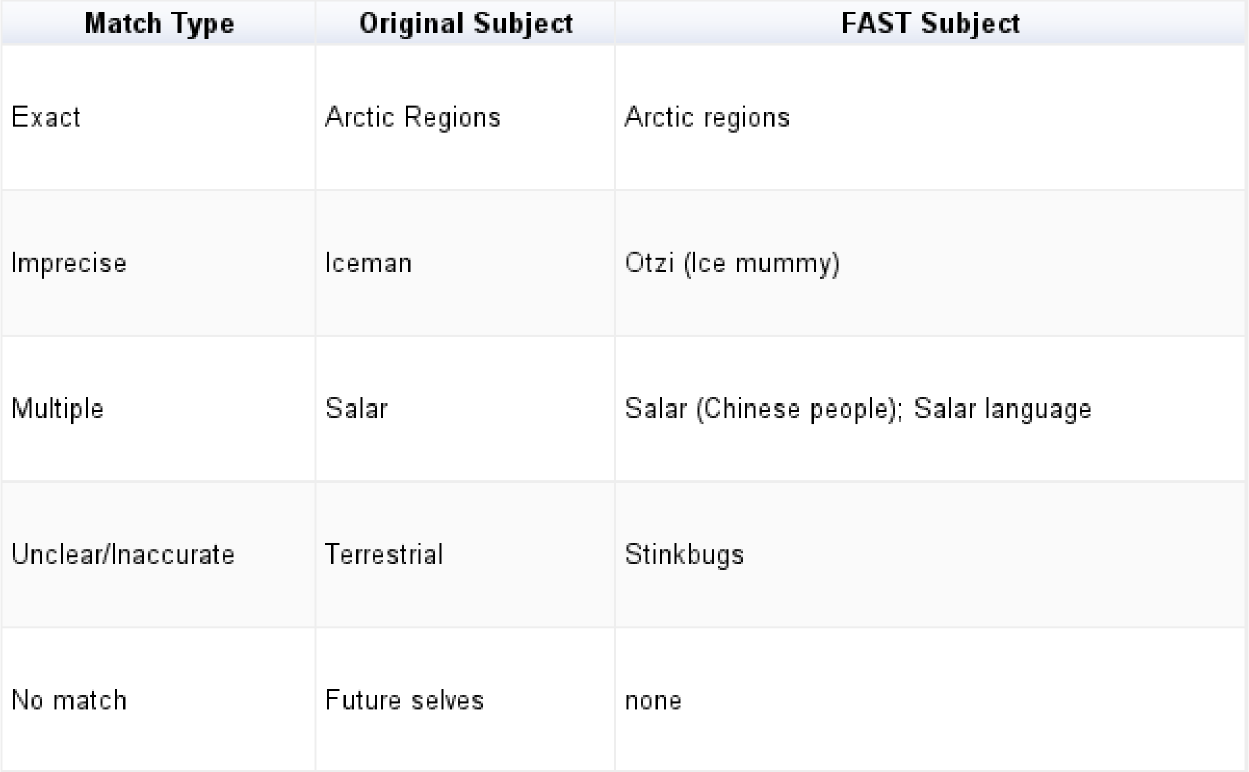
Method

Implementing a controlled vocabulary presents considerable concerns for workflows. Even as an automated process, oversight is still required for ensuring quality and accuracy in updating existing repository records. Finding a systematic way to ensure future terms adhere to the vocabulary is also a challenge. For these reasons it was determined that before a repository wide transition was considered, a smaller, representative sample of records would serve as a test case to identify the effectiveness of the mapping and hopefully draw attention to any issues that might arise in the process.

DSpace communities are organizing entities that contain collections of resources (items, in DSpace’s terminology) within a repository. In KUSW, communities have been used to define various departments and disciplines. The Anthropology, School of Business, and Geology communities in KUSW were identified as being distinctly different in subject matter to allow us to determine if and how well FAST would meet the descriptive needs of these particular subject areas. From both Anthropology and Business we selected 35 records containing subject terms. Geology had slightly fewer records with subjects, which only allowed for 33 to be selected. The chosen records represent less than 1% of the total items in the repository. Diachronicity, or how terms relate over time, was not considered to be an important factor in record selection as a repository scale transformation would involve mapping all terms to the present state of the FAST vocabulary without regard to how terms have evolved.

A member of the cataloging department mapped the records. Given the manageably small sample size for this experiment we deemed it feasible and desirable to have human judgement guide the mappings so we might more clearly see potential ambiguity in term relations that might be missed by automated means. The cataloger took each of the provided subject terms and entered them in OCLC's online searchFast tool, which uses an autosuggest feature to match strings against the existing vocabulary and return the closest possible match ("searchFast", 2015). The cataloger noted when there was an exact match, imprecise match, multiple matches, an unclear or inaccurate match, and no clear match.

Figure 1. Examples of match types for original and FAST terms



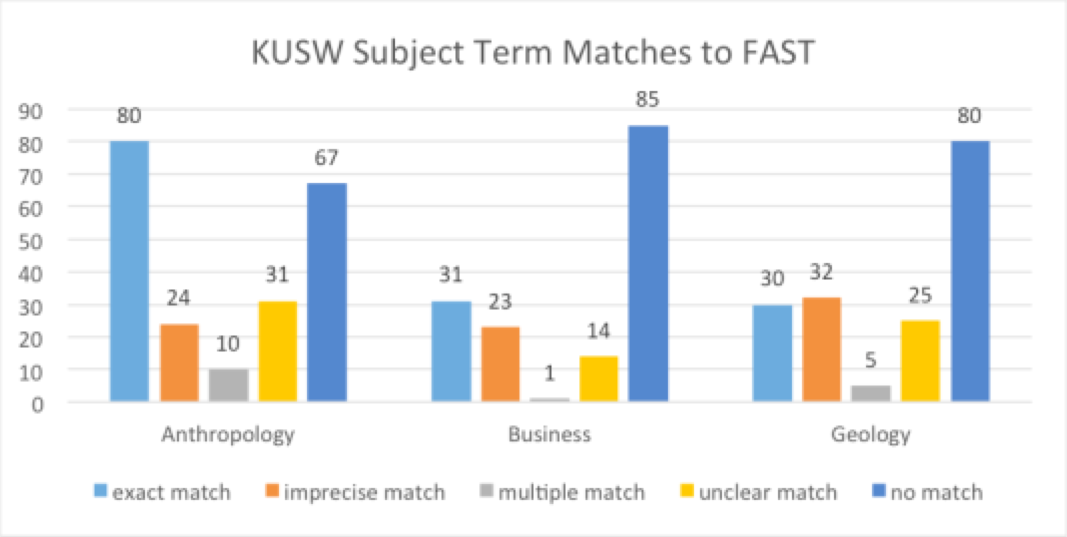
Match Type	Original Subject	FAST Subject
Exact	Arctic Regions	Arctic regions
Imprecise	Iceman	Otzi (Ice mummy)
Multiple	Salar	Salar (Chinese people); Salar language
Unclear/Inaccurate	Terrestrial	Stinkbugs
No match	Future selves	none

Multiple matches can be considered a kind of imprecise match as both types provide to varying degrees a new FAST term that is of debatable equivalency. For example, /Stratigraphy/ leads FAST to suggest /Beds (stratigraphy)/ and /Sequence stratigraphy/ which, while related, may be inaccurate to the resource content given their granularity. Unclear or inaccurate matches are similar situations but ones in which a terms or terms are clearly recognizable as being too specific or clearly outside the semantic field. Being able to map terms without having to reexamine the content of a resource makes having exact, and to an extent imprecise and multiple matches, a desirable outcome.

Results and analysis

The results shown in Figure 2 indicate that for each discipline a high number of terms had no matching, equivalent, or related term as provided by FAST's suggest feature. That there would be a number of uncontrolled terms outside of the FAST vocabulary was a predictable outcome given that uncontrolled subject terms can suffer from over-specificity and sometimes reflect something closer to genre terms. But that these would make up by and far the majority of the terms in Business and Geology and remain a substantial portion in Anthropology is a surprising result. A look at those terms in Business and Geology that had no FAST correlate reveals vocabularies containing terms like /seasonality/, /shallow/, and other instances that are adjectival and ambiguous when isolated.

Figure 2. KUSW Subject Term Matches to FAST



Yet there were several terms that would be good candidates for an equivalent FAST term. For example /market efficiency/ when entered into FAST only returns /Market Reform Act of 1990 (United States)/ which while possibly relevant could not without a reexamination of the resource be comfortably mapped. However, while entering the string through FAST's autosuggest function the incomplete /market eff/ did recall /Efficient market theory/ which semantically seems a more viable candidate. Since we were only looking at full strings, this possibility was one that would likely have remained hidden in the context of an automated mapping. However, should autosuggest be incorporated into the ingest process this aspect may prove a useful feature.

There remains the possibility that the terms with no clear match may be considered inaccurate and perhaps even imprecise, lowering their overall count. But doing so would require a second evaluation of the terms and more likely than not a return to the resource content for final interpretation. Even then this would only serve to even out the wide gap between unclear and disparate terms which would still comprise a significant portion of the terms overall. Since a review of the final mapping will involve human judgement there may be instances where a clear dichotomy between a term and its mapping will lead to an ignoring of the term and instead relegate it to a keyword field. Indeed, given the results and the high number of unique non-

FAST terms, a keyword field for terms of semantic value seems like an appropriate compromise so as to capture important aspects of the resource that are out of FAST's scope.

There are a few possible explanations for the higher representation of anthropology terms in FAST. While approximately 30 articles were taken from each discipline, the number of total subjects in each differed considerably. Anthropology had 214 terms, Geology 189 terms, and Business 156 terms. While Anthropology has more terms than the others, it is not so much higher to create the kind of disparity presented since overall percentages are roughly similar. Another possibility is that Anthropology terms are simply better represented in FAST though this statement cannot be responsibly be made without an analysis that is beyond the scope of the current discussion. Finally, an examination into the personnel completing the submission process and providing the terms is another possible source of answers through determining whether they were employing a controlled vocabulary or taking terms directly from a record created elsewhere. Certain disciplines and their scholars may be better oriented towards employing subject terms or may be more familiar with controlled vocabularies. A thorough examination into the source of uncontrolled terms would similarly be a worthwhile endeavor.

Hesitation to employ FAST as a representative repository vocabulary is an understandable reaction to the results of these test mappings. That LCSH is not without flaws concerning the currency and bias of subject terms is well-known (Fischer, 2005) and that these factors would consequently be at play in FAST as well is predictable. Yet this has not detracted others from identifying the value of FAST and explains its continued use despite ongoing debate concerning its relevance. No subject vocabulary can ever really be considered complete; at best it can hope to be comprehensive. FAST, despite its relatively poor performance for two of these three disciplines is functionally still an attractive vocabulary for reasons listed above regarding its scope and depth. The alternative would be to identify a subject specific vocabulary for particular disciplines, though doing so would potentially begin to bias the repository against those subjects for which there currently is no vocabulary.

The possibility of introducing disciplinary bias into a collection of bibliographic data takes on additional considerations as LD. The potential benefits of serializing metadata as LD for increased exposure means that not doing so will possibly relegate resources to less immediately discoverable channels. Disciplines that are more comprehensively represented in a LD vocabulary stand to gain more visibility than others. But there is also a risk of resource description claiming less immediately accurate subject terms, perhaps even only tangentially accurate, in order to gain greater exposure not unlike how current social media tagging trends often try to grab the widest possible audience by ascribing multiple terms of varying relevance. This semantic polarization, or the instance of an accurate term being accompanied by related but less accurate term, is a shift away from terms as emerging from resource content and more as an alignment with LD vocabulary terms. Again, this is not a new phenomenon but one with more serious implications in the LD environment. That this possibility exists is all the more reason for continued focus on LD vocabulary developments with a critical aspiration towards equal subject coverage.

Converting to RDFa

The second part of our experimental transition to a LD vocabulary involved serializing the data into a LD format. Formats available for exposing structured LD on the web include HTML Microdata, RDFa, and JSON-LD (World Wide Web Consortium, 2013, 2014, 2015). Each has unique characteristics that make them more amenable to specific contexts. Serializing our test corpus as RDFa was the most attractive option as it could be accomplished by a relatively simple reconfiguration of the DSpace architecture.

While Dublin Core (DC) terms, the default schema for resources in KUSW, can be expressed in RDFa, translating the values to the Schema.org (http://schema.org/) vocabulary was a logical direction. As the shared vocabulary that the major search engines have provided for making web data explicitly structured and consequently more easily indexed, it provided the best possible schema for enabling enhanced discovery. How search engines index the data and incorporate it into search results is difficult to predict. As Ronallo notes, Google, for example, provides limited support for only some Schema.org types in their Rich Snippets features and whether or not such snippets appear is dependent on a search query (Ronallo, 2012). Still, using the desired vocabulary at least fulfills the extent of what data providers can satisfy for successful LD exposure. Our records used terms from the Schema.org *ScholarlyArticle* type (http://schema.org/ScholarlyArticle). In an alternative approach, Mixter, O'Brien and Arlitsch describe an example of creating vocabulary extensions to Schema.org for institutional repository content (Mixter, O'Brien & Arlitsch, 2014). The resulting RDFa markup for this experiment was validated using Google's Structured Data Testing Tools (https://developers.google.com/structured-data/testing-tool/).

Perhaps counterintuitively we did not markup and map every DC term in our records to Schema.org terms. The reason for this was primarily methodological. A previous study undertaken at the University of Kansas found that by adding controlled foreign language terms to a record for an article in the same language, or that was the focus of that language, increased overall usage by 66% in the following year (Husic, 2014). The current experiment took this as one of its starting points by inquiring into how controlled subject terms and exposure as LD would affect usage. Other identifiers such as ORCiD identifiers for authors would consequently obfuscate the results of the following analysis. A full mapping to Schema.org of the appropriate record terms would be the next step in a total serialization into RDFa.

Limitations

The most prominent outside factor affecting the experiment concerns search engine behavior. While one can structure resource metadata to conform to Schema.org's specifications and in so doing optimize the data for search engines, any knowledge beyond that of how the data is being indexed and used to populate search rankings cannot be explicitly known. How conforming to these specifications impacts usage is of course the aspect this experiment sought to assess, but the specifics of how the data is used when exposed as LD remains a primary source of uncertainty and which makes it difficult to more finely tailor the data for further optimization.

Relatedly, while the 90 day time period of this study would seem to be long enough to allow for search engines to crawl the new structured data, it cannot be definitively said that all of them did during this period. Monitoring usage throughout the course of the coming year will provide a clearer indication of any effect the serialization had. Of course this also allows for other uncontrolled factors to be introduced. For example, a particular resource being assigned as a reading in a course would likely lead to a sharp rise in usage unrelated to the experiment. Similar scenarios are not difficult to imagine.

Results and Analysis

To test the impact of the transition to FAST we gathered usage statistics for 90 days prior to the introduction of the controlled vocabulary and the subsequent 90 days, as collected by DSpace’s Solr statistics feature, which filters out requests from known web crawlers. Usage statistics examined were item views (web requests for an item page in DSpace, a surrogate page for the resource that includes item metadata and links to files) and downloads (web requests for files associated with an item). View and download statistics were examined for the repository as a whole, and then for resources with FAST terms (FR) and those without (NFR). Figures 3-5 shows the sum (total downloads) for each 90 day period for the different groupings.

Figure 3. Non FAST Resource Usage

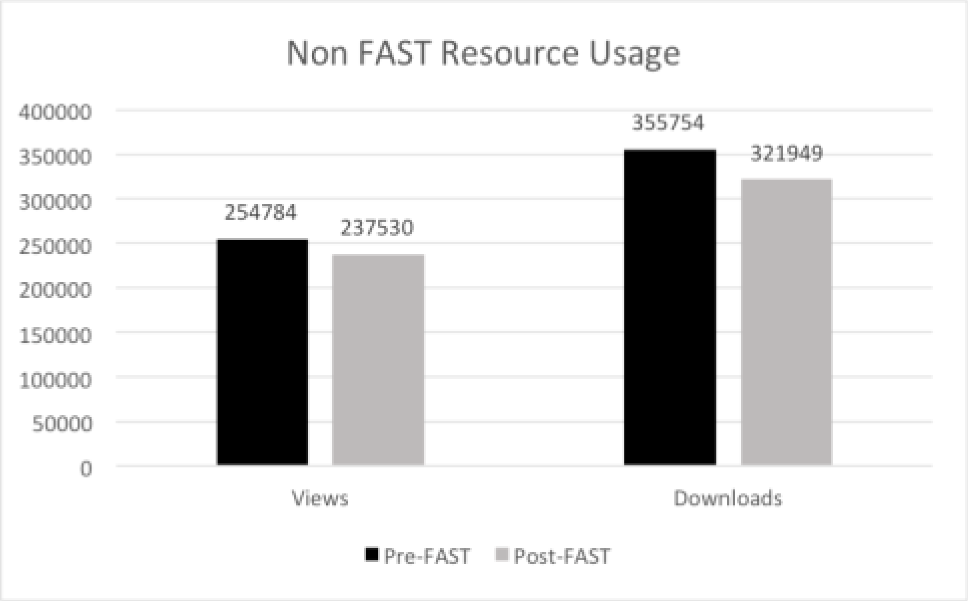


Figure 4. FAST Resource Usage

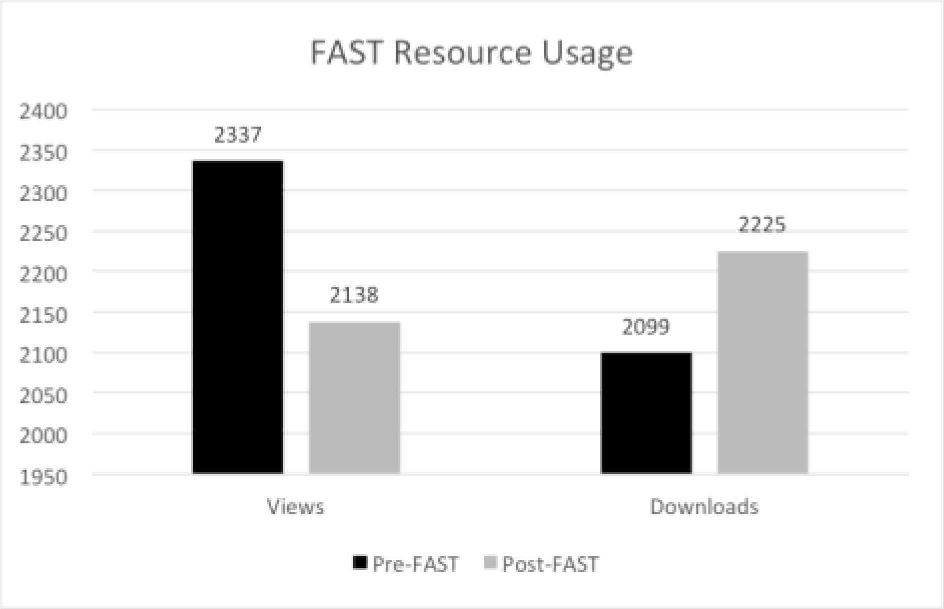
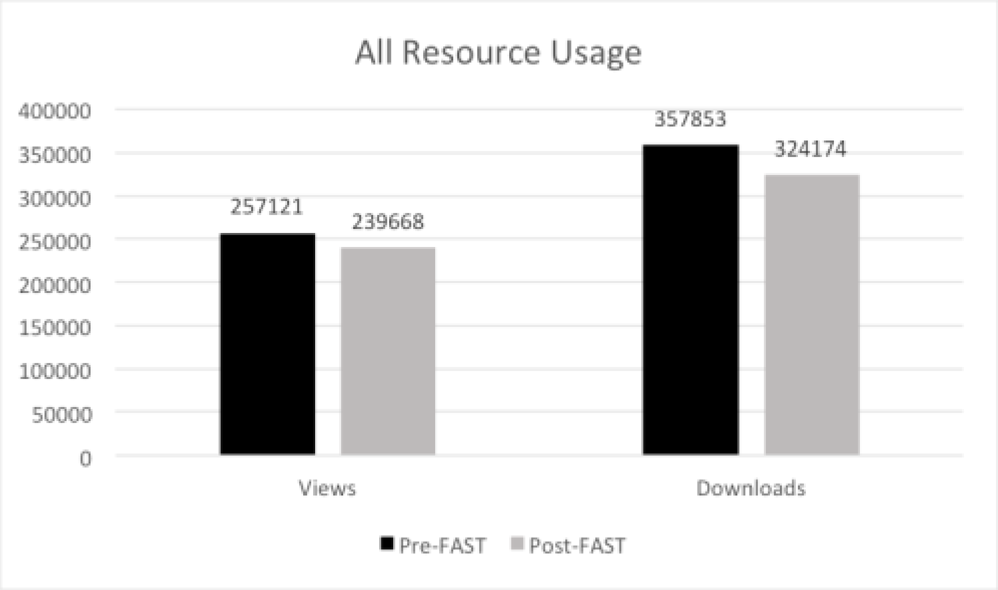


Figure 5. All Resource Usage



It is important to note that that repository usage experienced an overall decrease in usage during this total time period. Given the relatively small size of FR to the entire repository, it is understandable to see how NFR usage closely matches that of the overall repository both in item views and downloads. Item views followed a similar trajectory for FR though the decrease is even more pronounced. Given the overall decline in usage it is surprising then to observe that FR saw an increase of 6% in the number of downloads. While this represents only a modest increase, it does pose interesting questions particularly when viewed within the context of overall usage.

It is tempting to conclude that serializing FAST data in RDFa was responsible for the increase in downloads during a period when NFR experienced the opposite trend. This possibility cannot be discounted, particularly as it would align nicely with linked data's broad tenet of increased discoverability. But unfortunately neither can one definitively be said to be the case. There are a host of other variables in this situation that were beyond the control of the experiment concerning search engine indexing and resource popularity. For example, the small increase in usage may be entirely due to interest from a small group of researchers in the subject area totally unrelated to any serialization provided by the experiment. Contrasted with Husic's findings which were over a longer time period and for a different type of vocabulary, it would not be an unreasonable to conclude that the increase in usage is likely unrelated to inclusion of FAST and its LD serialization (Husic, 2014). Any conclusions are speculative at this point; we can point to possibilities but further testing of a larger sample over a longer period would be required before any conclusion could be more responsibly made. At present we can only say that any impact the serialization had on usage was at best negligible.

Conclusion

Transitioning to a LD vocabulary is not a trivial endeavor and becomes especially complicated when coupled with the task of reconciling legacy metadata. As Woodley notes, element crosswalking has been well explored, but mapping data values is still in need of further exploration (Woodley, 2008). This paper has attempted to document the process of such a process by examining term mappings and resulting impact on usage. On a more specific level it was observed that FAST as a vocabulary did not suit itself well to two of the three test collections. Inadequate subject representation has always been a source of concern for controlled vocabularies but takes on a renewed importance when it involves a resource's ability to interact with the growing LD environment. This indicates that questions of authority and representation will require increased attention and participation to identify areas of under or misrepresentation in LD vocabularies. Finally, looking at the overall impact on resource usage after the transition suggests that one should not expect any sort of watershed moment in increased usage. There is no explicit promise that exposing resources as LD will result in such an event, but not participating in LD activities may prove to be a missed opportunity.

To pose the title of this article as a question, we might answer curtly that the effectiveness of the transition was passable, if flawed, and its impact underwhelming. But one should recall that impact is not measured solely by usage. By adopting a controlled vocabulary one can introduce order and decrease the amount of noise in an information system. If it is too much to expect immediate increased discoverability in the web at a large, one can at least clarify this process locally within the universe of the repository as a part of general curation activities. The original advantages of using a controlled vocabulary then retain those same aspects in a LD environment with the added benefit of the Semantic Web's future potentialities.

While we will continue to monitor usage for this experimental corpus, testing a more subject specific vocabulary on a collection would be an important way to determine if this would be a more worthwhile direction to ensure adequate subject representation. With a longer amount of time to monitor the current test as well we will be able to gather a more concrete picture of any actual impact of this serialization. Subject terms, however, represent just one of several potential metadata types suitable for exposure as LD. Future directions for increased experimentation would be to include ORCiD identifiers in records and serialize names with this identifier. As LD continues to evolve there will undoubtedly be new opportunities and methods to not only expose but draw connections across data sources which will necessarily engender a reflection on the semiosic aspects of this ecosystem; not only that data is linked but what these relations signify and the means by which it is accomplished.

Bibliography

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

Dean, R. J. (2010). FAST: Development of Simplified Headings for Metadata. *Cataloging & Classification Quarterly*, *39*(1-2), 331–352. http://doi.org/10.1300/J104v39n01

Duranti, L. (2010). The long-term preservation of the digital heritage : A case study of universities institutional repositories. *Italian Journal of Library and Information Science*, 1(10), 157-168. doi:10.4403/jlis.it-12

"FAST (Faceted Application of Subject Terminology)" (2013). Retrieved from http://www.oclc.org/research/themes/data-science/fast.html

Fischer, K. S. (2005). Critical Views of LCSH, 1990–2001: The Third Bibliographic Essay. *Cataloging & Classification Quarterly*, *41*(1), 63–109. http://doi.org/10.1300/J104v41n01_05

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

Husic, G. (2014). Enhancing an Open-Access Linguistics Journal Archive with Library of Congress-like Metadata: A Case Study of the Effectiveness for Improving Discovery. *Kansas Working Papers in Linguistics, 35.* http://dx.doi.org/10.17161/KWPL.1808.15945

Lampert, C. K., & Southwick, S. B. (2013). Leading to Linking: Introducing Linked Data to Academic Library Digital Collections. *Journal Of Library Metadata*, *13*(2-3), 230–253. http://doi.org/10.1080/19386389.2013.826095

Lynch, C. A. (2003). Institutional Repositories: Essential Infrastructure For Scholarship In The Digital Age. *Portal: Libraries and the Academy*, *3*(2), 327–336. http://doi.org/10.1353/pla.2003.0039

Mcdowell, C. S. (2007). Evaluating Institutional Repository Deployment in American Academe Since Early 2005. *D-Lib Magazine*, *13*(9/10). http://doi.org/10.1045/september2007-mcdowell

Mitchell, V., & Hsieh-Yee, I. (2007). Converting Ulrich's™ Subject Headings to FAST Headings: A Feasibility Study. Cataloging & Classification Quarterly, 45(1), 59-85. doi:10.1300/J104v45n01

Mixter, J., O'Brien, P., & Arlitsch, K. (2014). Describing Theses and Dissertations Using Schema.org. *International Conference On Dublin Core And Metadata Applications,* 138-146. Retrieved from http://dcpapers.dublincore.org/pubs/article/view/3715/1938 "searchFast" (2015). Retrieved from http://fast.oclc.org/searchfast/.

Ronallo, J. (2012). HTML5 Microdata and Schema.org. *The Code4Lib Journal*, *16*. Retrieved from http://journal.code4lib.org/articles/6400

White, H. (2013). Examining Scientific Vocabulary: Mapping Controlled Vocabularies with Free Text Keywords. *Cataloging & Classification Quarterly*, *51*(6), 655–674. http://doi.org/10.1080/01639374.2013.777004

Woodley, M. (2008). Crosswalks, Metadata Harvesting, Federated Searching, Metasearching: Using Metadata to Connect Users and Information. In Baca, M. (Ed.), *Introduction to metadata* (3rd ed.). Los Angeles, CA: Getty Research Institute.

World Wide Web Consortium. (2013). HTML Microdata. Retrieved from http://www.w3.org/TR/microdata/.

World Wide Web Consortium. (2014). JSON-LD 1.0. Retrieved from http://www.w3.org/TR/json-ld/.

World Wide Web Consortium. (2015). RDFa Core 1.1 - Third Edition. Retrieved from http://www.w3.org/TR/rdfa-core/.