SUBJECT KNOWLEDGE CARDS: ENHANCING ETD DISCOVERY THROUGH LINKED OPEN DATA

Lucas Mak[1], Devin Higgins[2]
[1]Michigan State University Libraries, makw@msu.edu
[2]Michigan State University Libraries, higgi135@msu.edu

Abstract:
Michigan State University Libraries creates subject knowledge cards to enrich discovery experience of ETD users through harvesting linked open data (LOD) from OCLC FAST (Faceted Application of Subject Terminology) Subject Headings, Wikidata, and DBpedia. By querying the above LOD datasets and tracing related URIs among them, we generate a knowledge card, on the fly, to show users broader/narrower/related concepts for each subject heading assigned to an ETD. This enables users to browse the subject hierarchy both vertically and laterally and discover other semantically related ETDs available in the repository. The knowledge card also captures selected data points from Wikidata and DBpedia to provide some context to the subject term.

Keywords: Linked Open Data, Knowledge Card, FAST, Wikidata, DBpedia

BACKGROUND

Michigan State University Libraries (MSUL) started its ETD program back in 2011 when the Graduate School began mandating electronic submission of theses and dissertations. Besides electronic submission of around 700 newly published theses and dissertations per year, MSUL's ETD collection has recently grown to over 16,000 with the ingestion of digitized retrospective titles. The ETD collection is hosted in an Islandora digital repository, which also is the home of other MSUL text, image, and audio collections. With the rapid growth of MSUL's digital repository, subject headings have emerged as an important way to collocate items both within and outside of the collection. Though Islandora provides faceted subject browsing, it is merely a listing of subject headings occurring in search results, which may or may not have relationships (i.e. broader, narrower, or related) among themselves. Moreover, subject headings assigned to ETD titles, especially STEM (Science, Technology, Engineering, Medicine) related ones, may be incomprehensible for students outside those disciplines. In the current interdisciplinary research environment, overly technical subject headings, though accurate, may do a disservice to users of the repository, who, with proper contextual information, may be able to see multidisciplinary connections in scholarship otherwise quite foreign to them.

PROJECT IDEA

Since 2012, Google has been providing general overview of the search subject on its result page through a knowledge graph (or knowledge card), which is powered by linked data from various discrete data sources. The concept of the Semantic Web, a network of data that can be

read and understood by machines, was first proposed by Tim Berners-Lee (2001). Library communities around the world, though slowly, jumped on the linked data bandwagon and published their authority and bibliographic data as linked data, including OCLC (Online Computer Library Center) which published its FAST authority file as linked data. In 2016, OCLC further experimented reconciling its FAST dataset with Wikipedia entries (Bennett et al. 2016). Given that the MSUL digital repository uses FAST as its default subject vocabulary and stores FAST URIs (Uniform Resource Identifier) in MODS (Metadata Object Description Schema) records, the repository development team started to investigate how the FAST dataset and its linkages to other datasets could be leveraged for discovery enhancement, especially in providing subject hierarchy browsing and contextual data about each subject.

IMPLEMENTATION OF SUBJECT KNOWLEDGE CARDS

Broader and related terms are recorded in RDFXML of each FAST authority record and can be captured by retrieving the RDFXML file using the FAST URI stored in the MODS record. To capture narrower terms of a particular subject requires sending an API query to identify terms that use the search string as their broader term. The idea of providing subject hierarchy browsing is to allow users to retrieve titles that are semantically related to the subject in context. Each retrieved broader, narrower, and related term has its URI searched against the repository SOLR index. Only those terms that have results in the repository will be shown to users, to avoid wasted clicks on zero-result searches.

In comparison, getting contextual information from Wikidata and DBpedia is less straightforward. In the first iteration, the development team tried tracing existing linkages, created by OCLC's reconciliation experiment, between the FAST dataset and Wikidata, and subsequently to DBpedia, to capture selected data points for building a knowledge card showing contextual information of the subject in question. This can be done in a few API calls per subject heading. However, only 50% of the 15,000 unique FAST subject headings available in the repository have at least one Wikidata link. As a result, a significant number of subject headings would not have their own knowledge card. At the same time, there are slightly more than 2,100 FAST headings with multiple links, with one heading having as many as eleven links. Hence, the tracing mechanism would have to be smart enough to pick the appropriate Wikidata entry for these two thousand headings. Our experiment has shown that selecting the appropriate WIkidata entry on-the-fly is difficult. For example, the FAST heading "Signs and signboards" had Wikipedia's abstract for "Ampersand" in its knowledge card in our experiment even though "Signage", probably the most appropriate corresponding entry in Wikidata, was among the seven linkages recorded in the FAST RDFXML. Besides reconciling against Wikidata, OCLC also includes linkages to other datasets from library communities, including VIAF (Virtual International Authority File). Since the VIAF dataset has been reconciled against Wikidata, it is possible to trace from FAST to Wikidata through VIAF. This can serve as an additional tracing, especially for FAST headings that don't have direct links to Wikidata. However, tracing through VIAF also poses its own issues. Since

LCSH (Library of Congress Subject Heading) is outside the scope of VIAF, FAST headings that are derived from LCSH wouldn't benefit from this method. Also, reconciliation between VIAF and FAST is not totally reliable, resulting in mismatches. For example, the FAST heading for the country "Turkey" is reconciled to a VIAF cluster for a town called Turkey in Texas, which causes the resulting knowledge card to show the Texas town instead of the country. Similar to FAST, Wikidata has been linked to multiple vocabularies, including VIAF, GeoNames, and various controlled vocabularies from Library of Congress (LC). Since most FAST headings are derived from LC controlled vocabularies and the one-to-one relationship is recorded in the FAST RDFXML, it is possible to query Wikidata's SPARQL endpoint using the LC term's control number or URI. This would be more efficient than sending multiple calls to different APIs to trace URIs between datasets. Moreover, since Wikidata is a crowdsourced dataset, one can easily add or correct linkages to other datasets when those links are missing or wrong. In comparison, there is no good way to correct any incorrect linkages or to suggest new linkages in the FAST dataset, besides emailing OCLC and hoping they will take steps to make an update or correction. In the end, the development team decided to capture Wikidata and DBpedia data points by querying the Wikidata SPARQL endpoint using the control numbers of LC terms, the VIAF ID, and the GeoNames ID as recorded in FAST RDFXML file, and using the existing reconciled linkages between FAST and Wikidata as a fallback.

LIMITATIONS

Though adding and correcting linkages in Wikidata helps the overall linking reliability and success rate, differences in data modeling between FAST and Wikidata prohibit perfect reconciliation. For example, "asparagus officinalis" (species) is a variant term for "asparagus" (genus) in FAST, which means FAST collapses the genus with the species. On the contrary, Wikidata has a separate entry for each of them (and also one for "asparagus" the vegetable). A similar example is the handling of name changes. Michigan State University (MSU) was known as Michigan State College (MSC) before 1955. In FAST, these are treated as separate entities. Each of them has its own authority record, and hence URI. However, Wikidata treats them as the same entity, with MSC just an alias of MSU.

FUTURE WORK

Given the digital repository only contains a very small portion of the resources MSUL holds, the development team plans to include selected results from the library catalog and scholarly databases to further enrich the discovery experience. Recently, names and subject access points in the MSUL library catalog have been enriched with LC URIs. Given the one-to-one correspondence between FAST and LC terms, it is easily doable to convert a FAST URI stored in a MODS record into its corresponding LC URI. We can use the latter to search the library catalog and retrieve selected results for display on the knowledge card. On the other end, a FAST term can be translated into subject terms used in some popular article databases or disciplinary repositories, through Wikidata. API calls can then be made to these databases

and repositories to retrieve lists of scholarly articles on the same subject matter for display to users.

REFERENCES

Berners-Lee, Tim. James Hendler and Ora Lassila. "The Semantic Web." *Scientific American 284*, no. 5 (2001): 34-43

Bennett, Rick, Eric Childress, Kerre Kammerer, Diane Vizine-Goetz. "Linking FAST and Wikipedia." Paper presented at IFLA WLIC 2016, August 2016. http://library.ifla.org/1980/1/S12-2016-bennett-en.pdf.

"FAST (Faceted Application of Subject Terminology) Released as Linked Data." OCLC. Published December 14, 2011. https://www.oclc.org/research/news/2011/12-14.html.