Challenges of Mapping Digital Collections Metadata to Schema.org: Working with CONTENTdm

Patricia Lampron[1(✉)], Jeff Mixter[2], and Myung-Ja K. Han[1]

[1] University Library, University of Illinois at Urbana-Champaign, Urbana, IL, USA
{lampron2,mhan3}@illinois.edu

[2] OCLC Membership & Research at OCLC, Dublin, OH, USA
mixterj@oclc.org

Abstract. As digitized materials stored in content management systems become more prominent as a mode of resource access, the library community is experimenting with linked data to make these collections available in new ways. Applying Schema.org semantics to curated digital collections allows for enhanced search engine discovery, as well as the dissemination of metadata in ways that can connect resources across the internet. This paper shares the challenges encountered when mapping unique digital collections metadata to Schema.org semantics, and lessons learned from experimentation on CONTENTdm collections metadata at both the University of Illinois at Champaign-Urbana Library and OCLC.

Keywords: Linked data · Digital collections · Schema.org · Metadata · Contentdm

1 Introduction

Current trends in library metadata lean toward discovery and accessibility, not just within Online Public Access Catalogs and content management systems, but as shared information that relates across systems and is searchable through highly used search engines. According to a 2015 Library edition of the Horizon Report, “Popular search engines can only touch about 10 % of the Internet; the remaining 90 % are websites that are not indexed currently because most of this data is located in library catalogs in formats that cannot be searched or is guarded in secure areas that cannot be accessed by bots,” [1]. This lack of visibility in popular search engines has led to efforts by libraries to make their metadata available by incorporating Linked Open Data (LOD) technologies into their metadata management, e.g., creation, sharing and dissemination. This new direction can be seen clearly in recent efforts to overhaul current cataloging standards and practices, and the development of BIBFRAME [2], a vocabulary designed for bibliographic data that can be expressed using Resource Description Framework (RDF), led by the Library of Congress; as well, in the use of Schema.org and the Bib Extend Community Group’s recommendations for describing library materials as linked data [3], in order to enable users to discover these resources on the web. However, most of

E. Garoufallou et al. (Eds.): MTSR 2016, CCIS 672, pp. 181–186, 2016.
DOI: 10.1007/978-3-319-49157-8_15

the efforts are focused on the library's traditional collections whose metadata is in MARC format, and less attention has been given to the carefully curated special collections being digitized and housed in separate digital asset management systems.

The University of Illinois at Urbana-Champaign (UIUC) Library has been exploring ways to publish its bibliographic data and associated holdings and item data as linked data using Schema.org semantics [4, 5]. As a next step, the library is developing a metadata application profile consisting of a common set of properties to describe the wide variety of items across the 21 digital collections housed in their content management system, CONTENTdm [6], in attempt to make its rich and unique digital collections more discoverable on the web. OCLC has also been exploring how to map CONTENTdm data into RDF and simultaneously reconcile string values against the Virtual International Authority File (VIAF) [7] and Faceted Application of Subject Terminology (FAST) [8], and ultimately providing an N-Triple data dump of Schema.org data. These investigations both provided invaluable lessons in applying linked data principles and Schema.org semantics to digital collections described with customized, non-traditional metadata.

2 Exploiting Linked Data to Promote Digital Collections

As changes in user needs turn toward accessing information through sources outside the library (i.e. search engines), the library must find ways to make connections between the outside world and their resources. Linked data can help make these connections, but in order for libraries to benefit from the Web they must take into account W3C specifications and recommendations [9].

Schema.org is a linked data vocabulary designed and published by the major search engines and promoted as a structured vocabulary that they can all consume and understand [10]. When applied correctly, linked data using Schema.org semantics can provide search engines with well-structured data that can be harvested and that links to other resources on the Web. Notable search engines already support Schema.org semantics structured as microdata, RDFa, or more recently JSON-LD, and embedded within HTML pages, and are using this markup for indexing and display purposes, as well as building connections between information and resources, for example in Google Knowledge Graphs. These are the types of connections and exposure that libraries endeavor to create, and so embedding Schema.org enhanced metadata within CONTENTdm HTML pages could be a step toward better discoverability of digital collections.

3 CONTENTdm Collections and Their Metadata

CONTENTdm is a popular content management system used in libraries and archives for storage and access of curated digital collections. In CONTENTdm, each collection contains its own set of metadata fields, which are referred to as a "metadata profile". Each field in the collection has its own label, and can be mapped to a Simple or Qualified Dublin Core element, or the collection manager can choose to leave it unmapped. The system is organized much like a file cabinet, with the cabinet being the entire

CONTENTdm system, and the folders inside being collections that contain each metadata profile. This model provides detailed descriptive metadata specific to the individual collection. CONTENTdm also allows for the use of both local and established controlled vocabularies within a particular field or shared with multiple fields. Using controlled vocabularies ensures consistent metadata both within a collection as well as across multiple collections when the vocabulary is shared.

Many CONTENTdm fields can be mapped to the same Dublin Core element, however, they still often hold distinct descriptive information that is specified by their local field name. For example, although a field will be mapped to the Simple Dublin Core <dc:description> element, a local field name for this field could include contextual information, such as <Transcription>, <Inscription>, or <Translation>. The ability to refine metadata through field names while mapping to Dublin Core works well in allowing detailed descriptive metadata for discovery and access, and display in CONTENTdm, while still providing interoperable metadata for service providers and harvesters through the OAI repository. Within the CONTENTdm website, searching across collections is facilitated through both local field names as well as their mapped Dublin Core elements. The customized fields are also indexed for advanced search and discovery within an individual collection.

4 Mapping to Schema.Org

Although there are many benefits to the customizable nature of CONTENTdm collection fields, converting the metadata to linked data also presents challenges. The UIUC Library analyzed their collections' metadata fields in preparation for mapping to Schema.org semantics in order to develop a linked data based metadata profile, and discovered that while the customized fields, across collections, might share commonalities in their Dublin Core mappings, many of the Dublin Core elements are so broad in scope that overlapping fields could have a wide range of meanings. The most prominent example of this is the mapping of <dc:description>, which was mapped 116 times across 21 collections, and while <dc:description> is an extreme case, the majority of the Dublin Core terms in the library's CONTENTdm collections have multiple mappings, some even within the same collection. Another such example can be seen in <dc:contributor>. There are 13 unique field names across the 27 fields mapped to <dc:contributor>, including "Printer", "Speaker", "Architect", "Composer", "Lyricist", "Artist", and so on, all used for defining specific roles.

This analysis illustrates that a straight system wide mapping of Dublin Core elements to Schema.org semantics is not the most effective way to disseminate curated metadata to the web. As noted in the <dc:contributor> example, many of these field names can be represented by using either Schema.org properties or by employing the structured nature of Schema.org in combination with <schema:Role> to define specific terms that are common in the CONTENTdm collection metadata profiles, but this work must be performed by staff who have an understanding of the collections and how the fields are being used. It should be noted, however, that while Schema.org types and properties are designed to describe a wide variety of "things", it was originally created with commercial interests in mind, and so there are still areas in which information can be potentially lost or not represented. A number of extensions to the schema have been proposed and are in use, such as the Schema Bib Extend which has been adopted by the Schema.org vocabulary as an official extension in the bib.schema.org namespace [11]. Extensions like this can help fill the voids, but it is unclear whether extensions to Schema.org will be recognized in the future by search engines. Nonetheless, these extensions provide a set of semantics for exposing collections that contain more specific metadata through RDF.

Another difficulty in mapping CONTENTdm metadata to RDF is pulling apart conflated descriptions. It is very common for CONTENTdm records to contain statements about both a physical thing and its digital surrogate. For example, a single CONTENTdm record might contain both a <dcterms:dateCreated> value of '1904' and a <dc:format> value of 'JPEG'. It is clear that what is being described in this record is actually two items, the first being the original photograph taken in 1904 and the second being the digitized JPEG. This is problematic when mapping because in RDF both the physical item and the digital surrogate would be separately described and connected together with a property that shows the relationship between the two. In Schema.org this is done by describing a <schema:CreativeWork> (for the physical item) and a <schema:MediaObject> (for the digital item) and then connecting the *CreativeWork* to the *MediaObject* with a <schema:encoding> property. To achieve this type of granularity in mapping CONTENTdm metadata to RDF, it will be necessary to build templates that have a contextual understanding of metadata fields and can route field values to the appropriate RDF entity. This is again where local metadata fields can be useful. If the metadata fields are already mapped to Dublin Core elements it would be very difficult to distinguish between date created and date digitized using the <dc:date> element, but if local field values are retained through customized field names, there is a chance that the individual field values could carry with them enough semantics to inform a conversion template (i.e. <dcterms:dateCreated> and <dcterms:dateDigitized>).

Entity reconciliation and data inferencing during the conversion of non-RDF data into RDF is another challenge. When mapping a subject field to Schema.org, all of the various subject values become entities connected back to the item using a <schema:about> property. The idea of reconciling entities allows the subject strings to be mapped to existing linked data datasets. For example, one could take the subject string 'Ohio' and map it to the FAST URI <http://id.worldcat.org/fast/1205075>. Doing this helps connect the converted data to the wider web of linked data and alleviates the burden of having to create a new persistent URI for every subject value. Data inferencing is a result of mapping flat metadata to RDF. While it has been previously noted that the flat nature of CONTENTdm can be a benefit in the conversion process due to the simplicity involved in direct mapping, it does require that the mapper infer statements and sometimes entities that are not directly relatable to the original CONTENTdm record. For example, if a CONTENTdm record describes a recorded play there might be a customized <datePerformed> metadata field. When converting this record to Schema.org, the <datePerformed> field and value will have to spawn a new entity <schema:Event> which will be used to connect the value in the <datePerformed> field back to the play being described in the record.

5 Discussion

One of the frequent questions that comes up when discussing linked data is "Why?" There are two predominant perspectives on this question and both have valid arguments to support them. The first is an outward looking perspective that focuses on linked data syndication. The argument is that linked data can help improve the discovery of digital collections by improving the search engine optimization through metadata. As search engines put more emphasis on harvesting structured data, applying structured data to digital collections access systems using a vocabulary designed, published, and promoted by search engines seems like a logical and worthwhile effort. The second perspective is more inward looking and focuses on using linked data to help support and bolster internally maintained and curated data. As the name implies, linked data links resources on the wider web and it is believed that these connections can be leveraged to create a better end-user experience. Connecting to outside resources like FAST, VIAF, WikiData and GeoNames provides access to data such as maps, biographies, alternate names and foreign language data that can all be used to help provide a richer end-user experience.

While the work of implementing linked data in CONTENTdm collections is beneficial in many ways, it also presents its own unique challenges. Both the UIUC Library and OCLC experimentation on CONTENTdm collections metadata provided three invaluable lessons. First, it is nearly impossible to create one linked data profile that meets the needs of all special collections. Because each collection differs from the others in its descriptive metadata, the implementation of linked data transformation should be done at the collection level, rather than at the institution or system level, in order to ensure preserving and presenting the uniqueness of each collection to users. Second, unique collections require metadata reconciliation work, including the incorporation of links from various authority data, not just vocabularies that are standard to the library community, but also outside sources, for example the Internet Movie Database [12] or the Union List of Artist Names [13]. This work should be conducted with metadata creators and collection specialists to insure that the proper authority data is being chosen. Third, more communication among special collections curators, metadata specialists, and system administrators are required to make these unique digital collections available on the web. Because these collections are described through non-traditional library metadata standards, and are stored in and accessed through non-traditional library systems, sharing each other's needs and experiences would greatly benefit all stakeholders working with special collections, and ultimately users who are discovering these unique resources, both within the CONTENTdm environment and on the web.

References

1. Johnson, L., Adams Becker, S., Estrada, V., Freeman, A.: NMC Horizon Report: 2015 Library Edition. The New Media Consortium, Austin, Texas (2015). http://cdn.nmc.org/media/2015-nmc-horizon-report-library-EN.pdf
2. BIBFRAME. https://www.loc.gov/bibframe/docs/index.html
3. SCHEMA BIB EXTEND Community Group: Recipes and Guidelines. https://www.w3.org/community/schemabibex/wiki/Recipes_and_Guidelines

4. Cole, T.W., Han, M.-J., Weathers, W.F., Joyner, E.: Library MARC records into Linked open data: challenges and opportunities. J. Libr. Metadata **13**(2–3), 163–196 (2013)
5. Han, M-J., Cole, T.W., Lampron, P., Sarol, M.J.: Exposing library holdings metadata in RDF using Schema.org semantics. In: Proceedings of the International Conference on Dublin Core and Metadata Applications 2015, pp. 41–49 (2015)
6. CONTENTdm. http://www.oclc.org/en-US/contentdm.html
7. Virtual International Authority File. http://viaf.org/
8. Faceted Application of Subject Terminology. http://fast.oclc.org/searchfast/
9. Berners-Lee, T.: Linked Data (2009). https://www.w3.org/DesignIssues/LinkedData.html
10. Official Google Blog; Introducing schema.org: Search engines come together for a richer web. https://googleblog.blogspot.com/2011/06/introducing-schemaorg-search-engines.html
11. Schema.org Hosted Extension: bib. http://bib.schema.org/
12. Internet Movie Database. http://www.imdb.com/
13. Union List of Artist Names. http://www.getty.edu/research/tools/vocabularies/ulan/