Disambiguating Descriptions: Mapping Digital Special Collections Metadata into Linked Open Data Formats

Jacob Jett, Timothy W. Cole
School of Information Sciences
University of Illinois at Urbana-Champaign
501 E. Daniel St., Champaign, IL 61820
jjett2@illinois.edu; t-cole3@illinois.edu

Myung-Ja Han
University Library
University of Illinois at Urbana-Champaign
1408 W. Gregory Dr., Urbana, IL 61801
mhan3@illinois.edu

ABSTRACT

In this poster we describe the Linked Open Data (LOD) for Digital Special Collections project at the University of Illinois at Urbana-Champaign and describe some of the particular challenges that legacy metadata poses for representation in LOD formats. LOD formats are primarily based on the World Wide Web Consortium's Resource Description Framework standard which demands both that entities be named by opaque universal identifiers whenever possible but also that metadata descriptions for entities be as unambiguous as possible. The challenges for disambiguating those descriptions are illustrated through examples drawn from digital special collections based at four different digital libraries.

Keywords

Linked Open Data, special collections, digital special collections, metadata.

INTRODUCTION

The LOD for Digital Special Collections project at the University of Illinois at Urbana-Champaign is a 20 month exploratory project funded by the Andrew W. Mellon Foundation. Its primary task is considering answers to the question: "After digitization, what more needs to be done to maximize the usefulness of these digitized resources?" (Cole et al. 2015, 1). In order to find answers to this question the project is investigating ways to:

1. Enrich the existing digital special collections' metadata with links to outside resources,
2. Map that metadata into linked data friendly vocabularies, and
3. Make that linked data visible to linked data consumers.

Our focus in this poster is to report on some of the challenges we have observed as we carried out the second of the measures described above, mapping the metadata that describes objects in a digital special collection to a linked data friendly vocabulary. Our exemplar case showcases an item from the Motley Collection of Theatre & Costume Design[1] (Motley Collection) and the challenges that occurred when we mapped it into the schema.org[2] vocabulary that is used as an encoding standard by major web search engines. After a discussion of the mapping challenges we present objects from several other digital special collections that are based at other institutions and conduct a comparative analysis in order to speculate on whether or not the particular mapping challenges we faced are general ones or are specific to the legacy metadata contained within the Motley Collection.

LINKED OPEN DATA

LOD is highly structured data that is linked with other highly structured data. It lies at a confluence of the Open Data (Auer et al. 2007) and Linked Data (Berners-Lee 2006) movements and has its own World Wide Web Consortium (W3C)[3] community group[4] dedicated to it. It sits at the top of a five-star tiered deployment scheme proposed by Berners-Lee that describes how accessible data is. It relies on the W3C's Resource Description Framework (RDF) standard[5] for its essential structure. RDF provides the basis for the fourth tier of Berner-Lee's 5-star deployment scheme. In contrast most library digital special collections metadata records situate themselves within the third tier of the 5-star deployment scheme. Their metadata is accessible via the web as structured data in a non-proprietary format (usually through the HTML[6] or XML[7] serialization formats). In order to achieve full 5-star LOD status, digital special collections metadata must be mapped into an RDF-compliant vocabulary and it must be linked to

ASIST 2016, October 14-18, 2016, Copenhagen, Denmark.

[1] http://imagesearchnew.library.illinois.edu/cdm/landingpage/collection/motley

[2] https://schema.org/

[3] http://www.w3c.org

[4] https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[5] https://www.w3.org/RDF/

[6] https://www.w3.org/TR/html5/

[7] https://www.w3.org/TR/2006/REC-xml11-20060816/

pertinent entities that directly relate to the entity being described.

METADATA FOR DIGITAL SPECIAL COLLECTIONS

Metadata for digital special collections are different from traditional library metadata in a Machine Readable Cataloging format. Since digital special collections are housed in a separate asset management systems, they frequently employ a simpler metadata schema. For example, digital special collections that are indexed and presented using CONTENTdm or Omeka start with Dublin Core metadata. The Dublin Core metadata schema brings with it a number of issues, among them is the 1-to-1 principle that requires metadata to only describe a single representation of a resource.[8] Literature has found that metadata for digital special collections almost always includes other information beyond the digitized resource. This additional information includes information about the physical resource and other related resources, such as local file name and contextual information (Jackson et al. 2008; Han et al. 2009).

Metadata Challenges for RDF-based Vocabularies

Unlike Dublin Core and traditional descriptive metadata standard, like MARC, the RDF-compliant schema.org vocabulary is a mixed descriptive/narrative vocabulary that tries to represent both discreet entities and events.[9] One of its advantages is that it has had a large amount of uptake by web browser developers, search engines, and OCLC.[10] It provides a wealth of mapping options for most kinds of entities; however, the distinct lack of a certain entities in schema.org that are important in special collections' contexts, such as "stage production", creates some barriers for using schema.org in these contexts because RDF-based vocabularies are only at their most useful when they have entities to link to and among. One of the implications of this need for linking entities is that sparse metadata descriptions that lack links to or descriptions of contextualizing entities at all can be as problematic as extremely rich descriptions that conflate multiple entities together.

THE MOTLEY COLLECTION CONTEXT

Comprised of Margaret Harris, Sophia Harris, and Elizabeth Montgomery, the Motley Theatre Design Group created sketches for costumes and scenery for dozens of plays across a 44 year span. Found in 1932, their designs were integral parts of stage productions in venues such as the Royal Shakespeare Theatre, Broadway, and the Metropolitan Opera in New York City. The topically complex nature and rich descriptions provided through the Motley Collection's metadata afforded us ample opportunities to showcase both schema.org's utility for representing digital special collections' metadata and its gaps. In the particular case of the Motley Collection's metadata, we found that while very rich descriptions had been provided, they often violated the Dublin Core 1-to-1 principle (i.e., the metadata describes not only the digitized image, but also the work—the stage production—it was originally produced for as well as the play that the stage production was an example of).

Metadata Challenges for the Motley Collection

Figure 1 below showcases one of Motley Collection's 4,788 items as a user would encounter it through the library's CONTENTdm interface. This costume sketch, entitled "1914: Sergeant and Grocer" and which showcase the designs for costumes for a sergeant and a grocer character, were made for Peter Ustinov's 1967 play, *The Unknown Soldier and His Wife*. What is remarkable is that a sizeable amount of the metadata attached to the sketch actually describes the stage production that it is part of.

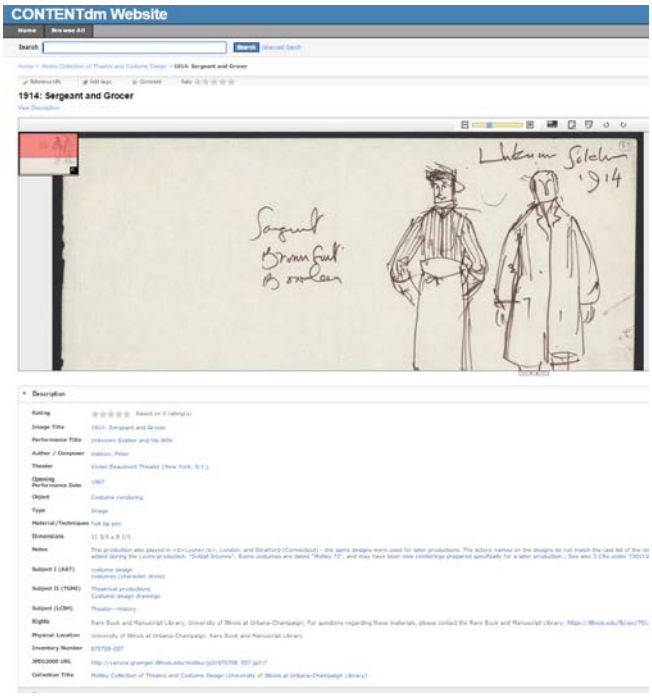


Figure 1. Costume Sketch for Sergeant and Grocer.

While the schema.org vocabulary doesn't have a sense of what a stage production is, it does have a sense of particular theater performances which its community have defined as a kind of event—"schema:TheaterEvent".[11] Our first intuition is that this would be the most appropriate mapping. But upon closer examination it becomes apparent that the metadata contained in the record does not actually describe any particular performance of the play. Similarly, the costume sketch in Figure 1 was not produced for any

[8] http://wiki.dublincore.org/index.php/Glossary/One-to-One_Principle

[9] As opposed to strictly narrative vocabularies such as CIDOC-CRM (CIDOC 2010) and FRBRoo (IWG-FCH 2015) or strictly descriptivist accounts such as Bibframe (https://www.loc.gov/bibframe/docs/index.html) and the SPAR family of ontologies (http://www.sparontologies.net/).

[10] https://www.oclc.org/developer/develop/linked-data/worldcat-vocabulary.en.html

[11] https://schema.org/TheaterEvent

particular performance, it was produced as part of a more overarching entity, which might be interpreted as an event but could also be interpreted as a form of creative work.

Field Name	Mapping to schema.org – schema:VisualArtwork
Image Title	schema:name (Text)
Object	schema:genre (Text)
Type	schema:artform (Text or URL)
Material/Techniques	schema:artMedium (Text or URL)
Dimensions	schema:height & schema:width (schema:Distance or schema:QuantitativeValue)
Subject I (AAT)	schema:about (schema:Thing)
Subject II (TGMI)	schema:about (schema:Thing)
Subject III (LCSH)	schema:about (schema:Thing)
Rights	schema:copyrightHolder (schema:Organization or schema:Person)
Physical Location	schema:provider (schema:Organization or schema:Person)
Inventory Number	spc:standardNumber (Text or URL)
JPEG 2000 URL	schema:associatedMedia (schema:CreativeWork)
Collection Title	schema:isPartOf (schema:Collection)
[Design by]	schema:creator (schema:Organization) [always Motley in this case]
[is part of Stage Production]	schema:isPartOf (schema:CreativeWork, spc:StageWork)
Field Name	**Mapping to schema.org – schema:CreativeWork**
Performance Title	schema:name
Theatre	schema:locationCreated (schema:Place)
Opening Performance Date	schema:dateCreated (Date)
Notes	schema:description
[additional type]	schema:additionalType (URL) [spc:StageWork]
[production of]	schema:exampleOfWork (schema:Book, fabio:Play)
Field Name	**Mapping to schema.org – schema:Book**
Author/Composer	schema:author (schema:Person)
[additional type]	schema:additionalType (URL) [http://purl.org/spar/fabio/Play]
[Published Work]	schema:name
[publication date]	schema:datePublished (Date)
[part of]	schema:isPartOf (schema:CreativeWorkSeries) [when true]
[adaptation of]	schema:exampleOfWork (schema:Book or schema:CreativeWork) [when true]

Table 1. Motley Collection Mapping[12]

In particular, metadata fields illustrated in Figure 1, such as "Performance Title", "Author/Composer", and "Theatre" don't describe the costume sketch at all but refer to the stage production that it is a part of, or the play that stage production is an example of. Human beings can easily disambiguate the combination of descriptive and contextual information given to them through the metadata in Figure 1. Machines do not have such facility with ambiguity. In order for machines to draw the correct conclusions machine-readable data must be as free from ambiguity as possible.

In the case of the Motley Collection's items, that necessitated decomposition of the legacy metadata records into three separate accounts, mapping its metadata to the appropriate entity: the digitized resource (typically an image), stage production, and play as shown in Table 1 (opposite).[13] These mappings were further complicated by gaps in the chosen vocabulary—schema.org.

The schema.org vocabulary also has a sense of a very similar kind of entity to the creative work that needs to be described—"schema:Episode".[14] An "episode" is a kind of creative work, just like our stage production. It has contributors and creators with roles like actor, author, and director, among others. It has associated works, such as music, created for it. Not only does our interpretation of stage production as a kind of creative work (in the schema.org sense of work) seem correct, the schema.org community has already provided a strong foundation through entities like "schema:Episode", "schema:Radio[-]Episode", and "schema:TVEpisode".

SIMILAR MAPPING CHALLENGES

In order to see that whether the mapping challenges of the Motley Collection commonly occurred within other digital special collections, we set out to examine examples from three other digital special collections that were not involved with our LOD project: the Archie Givens, Sr. Collection African American Literature[15] at the University of Minnesota, the American Art Posters 1890-1920 collection[16] at the Boston Public Library, and the George Arents Collection's Cigarette Cards[17] sub-collection at the New York Public Library. We examined the metadata describing an exemplar from each collection to assess how similar or dissimilar it was from the metadata descriptions of items in the Motley Collection and to determine what particular challenges it posed for mapping into LOD.

Shuffle Along – Archie Givens, Sr. Collection

As can be seen in Figure 2 below, there are certain similarities between the playbill's description and that of the costume sketches in Figure 1. The playbill itself

[12] Metadata for Motley Collection can be divided into three types when mapping to schema.org semantics, VisualArtWork, CreativeWork, and Book. Rows in great represent information that is added into metadata to make the fuller Linked Data representation.

[13] http://publish.illinois.edu/linkedspcollections/files/2016/06/Motley-Mapping.pdf

[14] https://schema.org/Episode

[15] https://www.lib.umn.edu/givens

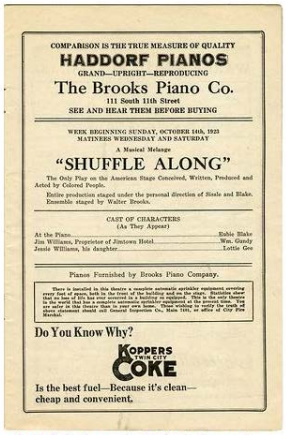
[16] https://www.digitalcommonwealth.org/collections/commonwealth:w0892d586

[17] http://digitalcollections.nypl.org/collections/cigarette-cards#/?tab=navigation

describes a stage production rather than any specific performance of the play. In this case, the metadata conforms to the traditional practices of archival description. It doesn't describe anything beyond the playbill itself and is more focused on the object's provenance rather than giving an account of its particulars.

This would seem to reinforce the view that the Motley Collection's idiosyncratic metadata and the difficulties mapping it into a LOD format may be unique to the Motley Collection. In fact though, if we try to link this object to related entities it becomes difficult to reconcile several of the assertions it makes. One of the claims is that the playbill is topically about Noble Sissle, the play's lyricist.

view full size image | download reference image



view full size image | download reference image

Title	Shuffle Along
Date	1923
Description	Playbill of the play "Shuffle Along," written, produced, and performed by African A
Physical Form	Playbill
Type of Resource	Text
Subject	African American Musical Ragtime Jazz Operetta Noble Sissle

Figure 2. Shuffle Along.[18]

In the case of topical headings such as this, it is clear that the metadata is conflating the digital object with work that it is describing. Before its metadata could be mapped into a RDF-compliant vocabulary like schema.org, its factual inconsistencies would need to be repaired. An entity representing the Metropolitan's production of the musical would need to crafted from scratch as beyond the evidence provided by the playbill itself, no record of this musical being performed at the Metropolitan in Minneapolis exist. It is mainly known from its Broadway production, which was active during the preceding two years (1921-1922). The productions would need to be linked together and a link to the play itself as form of written work would need to be made.

About Paris – American Art Posters 1890-1920

Figure 3 showcases a poster used to advertise a novel. Its metadata is very precise with regards to exactly what is being described. Once again, the precise focus on describing the poster, avoiding any apparent 1-to-1 descriptive practice conflicts, would seem to indicate that the idiosyncrasies of the Motley Collection's metadata is particular to it.

Figure 3. About Paris by Richard Harding Davis.[19]

This view is reinforced by considering how easily it is linked to the novel it advertises using schema.org's "schema:about" property within its own description. The primary problem is that this doesn't adequately articulate the role that the poster plays in regards to Davis' novel. To properly articulate that, a brief description of the novel would also need to be provided for the linked data consumer. If this was done then the "schema:image" property could be employed in novel's description to provide an explanation of precisely what role the poster plays in relation to it.

Miss Maie Ash – George Arents Cigarette Cards

The image in Figure 4 depicts a cigarette card, a kind of turn-of-the-twentieth century trading card sold with packs of cigarettes, and its description follows archival practice norms.

Figure 4. Miss Maie Ash.[20]

[18] Image retrieved from: http://umedia.lib.umn.edu/node/780429

[19] Image retrieved from: https://www.digitalcommonwealth.org/search/commonwealth:w0892j157

Once again, to get the most out of mapping the object's metadata into LOD it will be necessary to create metadata about intermediary entities that were closely linked to the cigarette card's provenance and role. Similar to the Motley Collection's need for a stage production entity, it might even be necessary to extend the schema.org vocabulary with an additional property, "includes" or "includedWith" in order to provide sufficient information for a machine to build the correct linkages between the cigarette card and the cigarettes it was sold with.

CONCLUSIONS

We identified three challenges when mapping digital special collections metadata into linked open data: conflating distinct entities together, missing interlinking entities, and gaps in RDF-based vocabularies. As revealed in our metadata analysis of the Motley Collection, in some instances, digital special collections' metadata includes descriptions about not only the digitized resource, but also other related physical and contextual resources. In these cases, metadata mapping requires additional steps to separate metadata into appropriate types of descriptions. Even more frequently, there seems to be an absence from the metadata descriptions of the kinds of physical and contextual resources that make linking to outside resources possible at all. Metadata enrichment work with linked data sources must also be considered and well planned if digital special collections are to realize the benefits afforded by LOD.

In the case of the Motley Collection, we are actively engaged in trying to identify established and reliable linked data sources for names (personal, theater, and play) that appear in metadata and are adding the matching URIs into metadata when there are matches. However, sometimes the string values added in metadata are not authorized forms and, data cleanup and enrichment are important parts of the reconciliation process. Additionally, it does not seem to be unusual for existing RDF-compliant vocabularies to have gaps which need to be filled through extensions. Our "spc:StageProduction" extension, a sub-class of schema.org's "schema:CreativeWork" is one such example of this process.

NEXT STEPS

As the LOD for Digital Special Collections project moves forward, we will be extending our mapping to an additional digital special collection based at the University Library—Portraits of Actors, 1720-1920.[21] In addition to mapping metadata, we will also be exploring how to enrich TEI-encoded[22] documents with LOD. The challenges encountered and solutions identified will be shared through future papers or posters like this one.

ACKNOWLEDGMENTS

We gratefully acknowledge the Andrew W. Mellon Foundation for generously funding the LOD for Digital Special Collections project.

REFERENCES

Auer, S. R., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. *The Semantic Web. Lecture Notes in Computer Science 4825*, p 722. doi:10.1007/978-3-540-76298-0_52.

Berners-Lee, T. (2006). Linked Data. *Design Issues*. [revised 2009]. Retrieved from: https://www.w3.org/DesignIssues/LinkedData.html

CICDOC CRM Special Interest Group (CIDOC). 2010. Definition of the CIDOC Conceptual Reference Model, version 5.0.2. Retrieved from: http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.2.pdf

Cole, T. W., Han, M.-J., and Szylowicz, C. (2015). Exploring the benefits for users of Linked Open Data for digitized special collections. Grant Proposal Narrative. Retrieved from: http://publish.illinois.edu/linkedspcollections/files/2015/11/Cole-UIUC-LODProposal-4-WebSite.pdf

Han, Myung-Ja, Christine Cho, Timothy W. Cole and Amy S. Jackson. (2009). "Metadata for Special Collections in CONTENTdm: How to Improve Interoperability of Unique Fields through OAI-PMH." Journal of Library Metadata 9(3): pp.213-238.

International Working Group on FRBR and CIDOC-CRM Harmonization (IWG-FCH). *FRBR object-oriented definition and mapping from FRBR$_{ER}$, FRAD and FRSAD (version 2.4)*. 2015. CIDOC-CRM. Retrieved from: http://www.cidoc-crm.org/docs/frbr_oo/frbr_docs/FRBRoo_V2.4.pdf

Jackson, Amy S., Myung-Ja Han, Kurt Groetsch, Megan Mustafoff, and Timothy W. Cole. (2008). "Dublin Core Metadata Harvested through OAI-PMH." Journal of Library Metadata 8(1): pp.5-21.

[20] Image retrieved from: http://digitalcollections.nypl.org/items/510d47da-7956-a3d9-e040-e00a18064a99#

[21] http://imagesearchnew.library.illinois.edu/cdm/landingpage/collection/actors

[22] http://www.tei-c.org/index.xml