Remodeling Archival Metadata Descriptions for Linked Archives

Brian Dobreski
School of Information Sciences
University of Tennessee
Knoxville, TN, USA
bdobresk@utk.edu

Jaihyun Park
School of Information Sciences
University of Illinois
Champaign, IL, USA
jaihyun2@illinois.edu

Alicia Leathers
School of Information Studies
Syracuse University
Syracuse, NY, USA
aleather@syr.edu

Jian Qin
School of Information Studies
Syracuse University
Syracuse, NY, USA
jqin@syr.edu

Abstract

Though archival resources may be valued for their uniqueness, they do not exist in isolation from each other, and stand to benefit from linked data treatments capable of exposing them to a wider network of resources and potential users. To leverage these benefits, existing, item-level metadata depicting physical materials and their digitized surrogates must be remodeled as linked data. A number of solutions exist, but many current models in this domain are complex and may not capture all relevant aspects of larger, heterogeneous collections of media materials. This paper presents the development of the Linked Archives model, a linked data approach to making item-level metadata available for archival collections of media materials, including photographs, sound recordings, and video recordings. Developed and refined through an examination of existing collection and item metadata alongside comparisons to established domain ontologies and vocabularies, this model takes a modular approach to remodeling archival data as linked data. Current efforts focused on a simplified, user discovery focused module intended to improve access to these materials and the incorporation of their metadata into the wider web of data. This project contributes to work exploring the representation of the range of archival and special collections and how these materials may be addressed via linked data models.

Keywords: Archives metadata, linked data, ontological modeling, item-level metadata, Linked Archives

Introduction

Archives are tasked with organizing, describing, and providing access to collections of historical materials. Despite being valued for their uniqueness, archival resources do not exist in isolation from other sources of cultural heritage information. In effectively representing these materials, it is critical to capture the context surrounding them by illuminating their connections to people, places, organizations, and other materials (Society of American Archivists, 2013). Using metadata to capture the relationships among archival materials and other entities can be accomplished in a variety of ways, though linked data approaches may be particularly useful in this regard and have been of significant interest within the cultural heritage domain in recent years (Gracy, 2015; Niu, 2016). Despite the nascent stages of archival linked data projects, a number of domain ontologies and metadata models are already available in this space, including CIDOC-CRM (http://www.cidoc-crm.org/), BIBFRAME (https://www.loc.gov/bibframe/), and EDM (Isaac, 2011).

While these models may facilitate the description of a variety of cultural heritage resources, they are complex and may not fully capture all relevant aspects of archival materials (Matienzo, Roke, & Carlson, 2017). At the same time, most existing linked data approaches to modeling and representing archival materials tend to occur at the finding aid or collection level, rather than the item level (Park, 2015). Item-level description of archival materials can reveal new connections and increase access and discovery of these resources, but poses particular challenges as well. Individually representing the variety of materials present in an archival collection with a single model becomes more difficult, especially for collections of

non-textual materials. Depicting heterogeneous materials, along with their potential digital surrogates, in an archival environment requires careful consideration of both resource characteristics as well as the functional requirements of metadata in these settings: organization, discovery, curation, provenance, and preservation.

Starting in fall 2018, we initiated a "Linked Archives" project after meeting with the archivist and metadata librarians at the Syracuse University Library Special Collections and Research Center. As a pilot study, this project was established to analyze the archival metadata descriptions at the item level, and based on this analysis, build an ontology model for transforming the existing metadata descriptions into linked data structures. In this paper, we report a linked data approach to modeling and representing archival media materials at the item-level, including both the physical originals as well as their digital surrogates. Using Syracuse University Library as a case, we analyzed data from three non-textual collections: still images, sound recordings, and video recordings. From this analysis, we developed and explored a new model based on linked data principles capable of capturing key properties and relationships for all of these materials. Anticipating a modular approach, the current model reflects metadata relevant to the functional requirements of end users, with potential future expansions addressing management and preservation requirements. We present our Linked Archives model below, along with challenges and further considerations associated with representing archival media materials. This work sets the foundation for continued development of a modular approach to heterogeneous item-level archival data while further revealing the challenges and potentials of linked data in the archival domain.

Literature Review

With the expansion of digital collections in the 1990s, archives and other cultural heritage institutions looked toward new standards for metadata creation and publication. *Metadata Encoding and Transmission Standard* (METS) became a particularly popular sharing and encoding standard for archival metadata at the item level (Ellings & Waibel, 2007). METS has guided the creation of much metadata for digital, archival objects, though as Semantic Web technologies have emerged and developed, cultural heritage institutions have become increasingly interested in pursuing these approaches. In 2011, Linked Open Data in Libraries, Archives, and Museums (LODLAM) was formed as a community for linked data enthusiasts in cultural heritage spaces (LODLAM, 2019). LODLAM has served as a resource sharing hub and demonstrates the growing relevance of linked data approaches in cultural heritage. In 2013, Mitchell profiled major linked data projects in the cultural heritage domain, including Europeana, Digital Public Library of America, and BIBFRAME, noting that each relied on RDF and a "big-umbrella" approach as opposed to specific content rules. As institutions and practitioners have become more knowledgeable in linked data approaches, more small-scale and institutional level linked data projects have emerged, particularly among archives. A challenge of such projects has been striking a balance between utilizing well-established, general standards and capturing specific, local data needs. For example, Park (2015) presented work combining DC, SKOS, and ISAD(G) into a linked data model capable of representing content from the National Archives of Korea, while Matienzo, Roke, and Carlson (2017) explored the challenges of using models such as Schema, BIBFRAME, EDM, and RICCM to represent archival descriptions.

Currently, a number of linked data compatible standards and approaches are available for archives wishing to publish their data. Reviewing projects and approaches in the archival domain, Niu (2016) noted that ontologies were a popular tool, dividing them into two types: those for ontology building (e.g., SKOS, OWL), and domain ontologies including FOAF, CIDOC-CRM, and LOCAH. Though the use of these and other ontologies can enable new kinds of access to archival data, Niu (2016) also indicated that many of these models are based on legacy archival description practices rather than current and emerging user needs. Understanding and addressing user needs and understanding in relation to archival data has been an ongoing issue, pre-dating the rise of semantic approaches (Yakel, 2004). In response, some archival linked data projects have looked beyond cultural heritage domain ontologies and models and toward other perspectives, such as the web indexer and developed supported Schema.org model

(Matienzo et al., 2017), or independent specialized models such as LODE (Linking Open Descriptions of Events) (Shaw, Troncy, & Hardman, 2009). Even though vocabularies such as DBpedia, Schema.org, and LODE promise potential benefits for archival description and information discovery, challenges and barriers remain in utilizing these vocabularies to transform existing archival metadata descriptions into linked data. In a study that compared elements in *Encoded Archival Description* (EAD) and *MAchine-Readable Cataloging* (MARC) to the classes and properties in the abovementioned vocabularies, Gracy (2015) discovered that metadata encoding standards such as EAD and MARC tend to closely align with linked data vocabularies at the class level. Since archival descriptions have more limited granularity than the linked data vocabularies do, this makes it difficult for archival descriptions to "take the advantage of the richness and depth of potentially relevant information through these sources" (Gracy, 2015, p. 277). Overall, a growing array of data models, ontologies, and other tools exist for archival linked data projects, though choosing appropriate standards and creating metadata to meet user and institutional needs remains a challenge.

While archives contain a great deal of textual materials such as letters, manuscripts, and books, non-textual media such as photographs or sound recordings are also common and pose particular challenges. Currently, many institutions with archival media collections turn to ready-made platforms to publish their data and digital surrogates, such as contentDM or Omeka (Andro, Asselin, & Maisonneuve, 2012). Some existing platforms have begun allowing institutions to enhance their current metadata through the addition of linked data from external sources (see for example, Waitelonis & Sack, 2009). Although such platforms are beginning to incorporate more semantic functionality, archival linked data projects have also explored the use of more specialized data models and tools, particularly in representing media. For example, Daquino et al. (2017) described efforts to represent the Zeri Photo Archive as linked data, finding the general archival model CIDCO-CRM to not adequately fit their data; rather, the project developed its own ontologies specifically for photographic material, based on the specialized standards F Entry and OA Entry. This conflict between generalized models and specialized materials is echoed in other works exploring linked data and media in general. Gracy et al. (2013), for instance, explored and compared traditional and linked data models for sound recordings and other music materials, finding issue with aligning general models of description to more specialized ones. Although specialized ontologies exist for particular classes of media materials, discrete and disparate approaches to these materials present their own challenges for archives in providing consistent, cross-collection organization and discovery. Remodeling existing, specialized data for heterogeneous collections as linked data thus continues to pose difficulties for archival institutions.

Methods

The Special Collections and Research Center (SCRC) at Syracuse University (SU) Library houses a large repository of archival collections. In addition to EAD formatted finding aids, each collection has item-level metadata descriptions generated from archival processing. SCRC distinguishes physical original items and their digital surrogates, with separate item-level descriptions created for each. While the findings aids describe archival materials at the collection level using EAD, item-level metadata descriptions utilize an in-house schema that was developed based on METS and Metadata Object Description Schema (MODS). Between collection and physical item there is a one-to-many relation, i.e., one collection contains many physical items. Similarly, a physical item may be associated with many digital items, e.g., a TV program video cassette may have two digital items: a video file and a transcript of the audio. Similarly, a physical photograph may have multiple digital items: the digital images of the photo and verso of the photo with an inscription. Physical items and their associated digital items each receive separate descriptions.

Access to archival materials is typically mediated, with end users limited mostly to collection level descriptions (Miguez, 2018). While collection-level metadata have an established standard (i.e., EAD) for description and front-end discovery and presentation, the item-level metadata at SU have largely been hidden behind the scenes and less utilized due to policy and operational reasons. How to bring this hidden

item-level metadata to the front for developing intuitive, effective discovery applications was thus the motivating goal for this project. Our research of the current landscape of linked data for archives shows that much of the current effort has been focused on the collection level; linked data for archives at the item-level is less developed in existing efforts, which is especially true for media-oriented archival collections. As a pilot test, this project aimed to develop a linked data model for item-level archives metadata, and to generalize workflows and rules for automatic mapping of existing metadata descriptions to this linked data structure.

Data

The SCRC at SU Library provided sample metadata records for three media types: music recordings, video recordings, and photographs. An in-house metadata schema was developed by SCRC based on METS, which contains elements for physical item-level and digital item-level respectively. There were 36 metadata elements for objects, 49 for items, and 22 for names. Table 1 shows the distribution of metadata element categories.

Table 1. Distribution of metadata element categories

Cate-gory	Elements common to both	Physical items	Digital items
Administrative	collection_id, internal_id, object_id, object_type_id, time_stamp_created, time_stamp_export, time_stamp_updated	alt_repo, bibid, series_id date_issued, donor, draft, index, location, notes, object_deleted, object_draft, open_closed, related_items, rights, summon_content_type	item_id, Item_draft, item_deleted, digitized_by, linked_objects, item_html, item_download, checksum_display, checksum_archive, notes, orig_gen, orig_format, orig_notes, dig_notes
Descriptive	title	coverage, date_issued_display, date_orig_display, description, geo_code, language, media_type, series, subjects, subject_local, title_alt, type	date_digital, duration, color_bw, dimensions, dimensions_digital, physical_description, sound,
Technical (digital items only)	file_display, file_archive, general_technical_information, file_compression_archive, file_ppi_archive, file_quality_archive, file_scan_hw_archive, file_scan_sw_archive, file_type_archive, file_size_display, file_size_archive, internal_file_path, file_bit_depth, sampling_rate_audio, sampling_rate_video, sampling_ratio, codec, file_format, tech_info_file, tech_info_preservation		

It is clear from Table 1 that not all elements used for physical item descriptions are in digital item descriptions, and vice versa. Physical item descriptions focus on the representation of the object as a raw resource, while digital item descriptions focus more on the technical and administrative metadata generated from digitization. As such, the physical item description contains more information useful for discovery purposes while the digital item description is used for managing and presenting digitized items. Both levels contain similar amounts of administrative elements but obviously vary in semantics and purposes.

The item-level metadata we obtained belong to three collections, the finding aids for which are available on SU Library's website:

- Belfer Cylinders Collection: music and spoken word recordings dating from 1890 to 1929. There are 1729 physical item records, 1729 digital item records, and 3000 name records for individuals, groups, and other entities in various roles.
- Ronald G. Becker Collection of Charles Eisenmann Photographs: photographs of 19th century sideshows, circuses, and performers, most taken by Charles Eisenmann or his successor Frank Wendt, dating from 1836 to 1960. There are 1,414 physical item records, 1,416 digital item records, and 1,504 name records describing various roles.
- Ted Koppel Collection: videos of ABC News television programming with Ted Koppel, including approximately 6,600 episodes of Nightline (March 1980-November 2005). There are 7416 physical item records, 13,610 digital item records, and 72,988 name records of individuals appearing in various roles.

We reviewed this metadata, focusing particularly on what is of use for discovery and selection activities. We analyzed the structure and relations to identify key concepts and the relationships among them, as well as which metadata elements would be absolutely necessary for supporting the tasks of interest.

Modeling Approaches

Our literature review and the data available for this pilot project helped us make two important decisions. First, existing ontology models for general purposes are not suitable for the Linked Archives model, although some classes and properties can be adopted in our model. Second, the model will neither create new metadata, nor take all elements (or properties) in SCRC's current schema into the new model. The rationale for these two decisions was that the goal of the current Linked Archives model would be mainly for discovery purposes. The current relational data structure in place supports administrative tasks well, so there was no need to transform most administrative and technical metadata into the Linked Archives model at this time. The discovery function covers two kinds: discovery by human users (e.g., scholars, students, general public) and discovery by machines. The former involves using Linked Archives data to develop discovery tools for users, with the latter entailing publishing the datasets in linked data format suitable for machine consumption.

The modeling process proceeded through a number of steps: 1) create a list of top classes based on major entities common across collection and item levels; 2) design object and data properties in the model; and 3) test the model with individual instances and modify the model as needed. In the modeling process we used multiple tools to communicate and document the model and its iterations, e.g., Google Team Drive for data and file sharing, spreadsheets for listing metadata elements and mapping them with the data properties in the model, and the Protégé software to create the model in Web Ontology Language (OWL). Our approach was a combination of top-down and bottom-up, which means that we identified the classes based on what exists in the current metadata descriptions and cross-validate them by comparing them with those top classes in CIDOC-CRM, BIBFRAME, and Schema.org. Relations between classes were analyzed from logical, semantic, linguistic, and inferential perspectives, which formed the basis for decisions on what should be treated as object properties or data properties.

We used the ontology software Protégé to remodel the metadata elements into a discovery module in the envisioned Linked Archives ontology. Protégé is an open source ontology tool developed by Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine. It is W3C standards compliant and can generate multiple ontology formats. The modeling process was iterative and constantly tested with individual instances. The classes and relations as well as value domains and ranges were carefully mapped to the ontology. Individual data entry was used as a way to validate the logic and correctness of classes and object properties in the ontology. When a problem was identified, we adjusted the class arrangement and property definition and then use the object and data properties to test any changes for logic and correctness.

RESULTS

Classes in the Linked Archives Model

The metadata for the three archival collections included in this pilot study is structured by collection, physical item, and digital item levels. Each collection contains multiple physical items, while physical items may be associated with one or more digital items. The physical item metadata describes the physical object itself, while the digital item metadata describes technical details of the associated digital files. Physical items in the Koppel collection tend to have two digital items associated: a video file and a transcript file. It was also common in these collections for multiple items to bear an identical title, because the same individual was photographed with the same theme but in different poses or the same piece of music work was recorded with different arrangements. For example, three physical items in the Belfer Collection have the identical title *Forget Me Not*, but each was performed with a different arrangement: songs with orchestra, popular instrumental music, and popular music. Similarly, four different portraits in the Becker-Eisenmann Photograph Collection were taken of the same person, but all contain the same title, "Nora Hildebrandt, Tattooed Lady" (SCRC, 2019) (Figure 1).

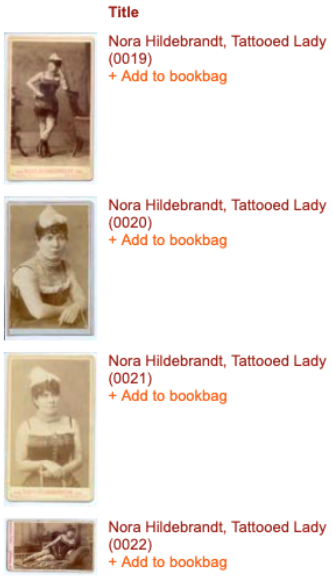

Figure 1. Photograph objects with an identical title

Across collection and item levels, a number of entities appeared frequently and served different roles in representing the content. Given the purpose of this project, we intended the model to be lightweight, that is, focus on elements to aid discovery of knowledge hidden at the item level and leave the administrative and technical functions to the current technology and infrastructures. Figure 2 is the initial model that illustrates the entities and relations between them as well as the three levels of archival artifacts. We originally used *Agent* as a class and listed *Person* and *Organization* as its subclasses, but when testing the model with individuals, the *Agent* class appeared to be cumbersome and added an unnecessary layer that could complicate the data structure. It would serve only the role of grouping like classes together rather than holding individuals in the ontology. After weighing the benefits and drawbacks of several options in modeling *Person* and *Organization*, we decided to create separate classes for person and organization. The entities on the left side in Figure 2 were singled out because they were the key access points as well as the common content structure for archival representations.

Figure 2. The initial Linked Archives model

It should be pointed out that *Series* is a special case for the classes defined in this model. An item could be part of a bibliographic series, an archival record series, or both. A bibliographic series originates with the publisher or creator of a work, while a record series is created at the time an archive receives a collection. A collection could contain both types, with an item belonging to both a collection-level record series and an external bibliographic series. For example, the Belfer Cylinders Collection contains a series assigned during production ("Edison Blue Amberol"); the Becker-Eisenmann Collection contains several record series in its finding aid ("Charles Eisenmann Cabinet Cards"); and the Ted Koppel Collection

includes episodes of several different television series, such as "Nightline," but is also divided into record series based on these pre-existing bibliographic series ("Nightline 1980-2005"). For this project, we focused on the bibliographic definition of series whenever possible as more applicable to discovery through linked data.

Properties in the Linked Archives Model

The relations between *Collection* and *Physical Item* (inside the rectangle in Figure 2) and content representation classes (classes outside of the rectangle) laid the ground for building properties for the classes. According to the Web Ontology Language (OWL) specifications, classes are sets of individuals and have two types of properties: object properties that connect pairs of individuals and data properties that connect individuals with literals (Motik & Parsia, 2012). Defining data properties was straightforward as it entailed mapping the metadata elements in the sample records to metadata standards. Out of 36 physical item elements, 6 contained no information, and it was determined that 11 out of the remaining 30 would aid in discovery (title, media_type, type, title_alt, date_issued, description, language, subjects, coverage, date_orig_display, date_issued_display). These elements were then mapped to Dublin Core and Schema.org. Of the 22 elements describing persons and organizations, nine were considered useful for disambiguation (role_name, name_last, name_middle, name_first, prefix, years, name_alternate, lcnaf_id, name_firm). These nine elements were then mapped to Schema.org and Library of Congress Name Authority File, with the Virtual International Authority File (VIAF) as a possible alternative. In the process, we also reviewed a number of other standards so that we could reuse as many existing vocabularies as possible, including Library of Congress Linked Data Service, Schema.org (particularly CreativeWork, MediaObject subsets), and PREMIS 3. In general, metadata for collections, objects, and items could be readily mapped to existing standards, while metadata for other classes required additional effort. An example is the data properties for the *Subject* class, which was best mapped with Simple Knowledge Organization System (SKOS, https://www.w3.org/2004/02/skos/) vocabulary.

The physical item properties were used to connect entities to address questions about who, what, when, and where concerning the object-level representations. Any entities in Figure 2 may serve as the subject for collections and/or objects. The *Event* class is selected to counter the linear organization of physical archival materials, because archival materials related to an event may not be in the same group or series. *Event* is a special case of subject in that it is characterized by the presence of a theme, time, and place. For example, "Iran Hostage Crisis, 1979-1981" is a subject term in LCSH that satisfies the three necessary conditions for an event. The objects related to this event that were scattered across multiple items in the Ted Kopple Collection from different times can be brought together by this event term. When the Linked Archives datasets are published, it will create an opportunity for connecting other archival materials related to the same event together. Besides the usual function of bringing all items on this topic together as a subject term, it also provides specific event-based (a type of topic-based) searching and browsing for humanities scholars, which has been demonstrated by prior research to allow for new research questions to emerge (Anderson & Blanke, 2015).

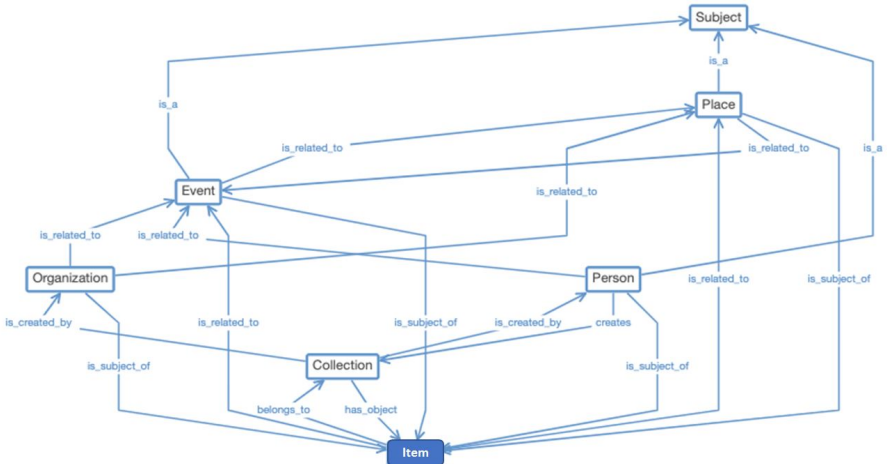

Figure 3. Illustration of relations between classes in the Linked Archives Model (portion, generated by Protégé, web version).

After a few iterative

revising rounds, we settled on a small number of physical item properties that would be potentially effective in connecting individuals of two classes and formulating rules for metadata transformation to the linked data format. Figure 3 shows a portion of the relations between classes. For each physical item property in the model, we defined the label, International Resource Identifier (IRI), domain, and range. Inverse relations were also defined where appropriate. For example, an item (range) belongs to a collection (domain) and in reverse, a collection has multiple items. Similarly, an item may be related to an event while the event is an individual of subject. As of this writing, the model is still evolving because there could be several ways to represent the same relation between individuals of two classes. The final decisions on relation defining will continue to be guided by the objectives mentioned at the beginning of this paragraph.

Mapping Individuals

The item-level metadata schema used at the SU Library is a combination of METS and MODS. The metadata records are natively stored in XML format but can also be exported in CSV format. The schema structure is simple and straightforward and the metadata records for items, persons, and subject headings can be transformed to match the data properties in the model without much difficulty. The challenging part is the mapping of physical item properties as they are essentially non-existent in current metadata records and require some additional work. Using the event "Iran Hostage Crisis, 1979-1981" as an example, we entered two physical items and five persons as individual instances for the *Physical Item*, *Event*, and *Subject* classes as shown in Figure 4 below. The three persons played different roles in the TV program episode (physical item) and the individuals in the three classes were connected by the relations defined in the model. That is, the model made it possible to reveal the item-level content in a more refined way that a collection-level representation would not have been able to provide. Such content-focused representation addresses the needs of historical research on archives described by Anderson and Blanke: "We [archivists] think of the [archive] as different collections; they [researchers] think of the [archive] as different subjects (Interview with archivist)" (Anderson & Blanke, 2015, p. 1188).

```
<!-- http://linkedarchive.syr.edu/collection/object/12983 -->
<owl:NamedIndividual rdf:about="http://linkedarchive.syr.edu/collection/object/12983">
    <rdf:type rdf:resource="http://linkedarchive.syr.edu/collection/object/"/>
    <rdfs:label>Nightline: Iran Hostage Crisis: Day 142</rdfs:label>
</owl:NamedIndividual>

<!-- http://linkedarchive.syr.edu/collection/object/56770 -->
<owl:NamedIndividual rdf:about="http://linkedarchive.syr.edu/collection/object/56770">
    <rdf:type rdf:resource="http://linkedarchive.syr.edu/collection/object/"/>
    <property:is_related_to rdf:resource="http://id.loc.gov/authorities/subjects/sh85067917"/>
    <rdfs:label>Nightline: Iran: Day 149</rdfs:label>
</owl:NamedIndividual>

<!-- http://linkedarchive.syr.edu/person/12153 -->
<owl:NamedIndividual rdf:about="http://linkedarchive.syr.edu/person/12153">
    <rdf:type rdf:resource="http://linkedarchive.syr.edu/person/"/>
    <property:is_related_to rdf:resource="http://id.loc.gov/authorities/subjects/sh85067917"/>
    <property:is_related_to rdf:resource="http://linkedarchive.syr.edu/collection/object/12983"/>
    <property:role>TV host</property:role>
    <rdfs:label>Koppel, Ted</rdfs:label>
</owl:NamedIndividual>

<!-- http://linkedarchive.syr.edu/person/12530 -->
<owl:NamedIndividual rdf:about="http://linkedarchive.syr.edu/person/12530">
    <rdf:type rdf:resource="http://linkedarchive.syr.edu/person/"/>
    <property:is_related_to rdf:resource="http://id.loc.gov/authorities/subjects/sh85067917"/>
    <property:is_related_to rdf:resource="http://linkedarchive.syr.edu/collection/object/12983"/>
    <property:role>Reporter</property:role>
    <rdfs:label>Kashiwahara, Ken</rdfs:label>
</owl:NamedIndividual>

<!-- http://linkedarchive.syr.edu/person/38696 -->
<owl:NamedIndividual rdf:about="http://linkedarchive.syr.edu/person/38696">
    <rdf:type rdf:resource="http://linkedarchive.syr.edu/person/"/>
    <property:is_related_to rdf:resource="http://id.loc.gov/authorities/subjects/sh85067917"/>
    <property:is_related_to rdf:resource="http://linkedarchive.syr.edu/collection/object/12983"/>
    <property:role>Interviewee</property:role>
    <rdfs:label>Morefield, Dorothea</rdfs:label>
</owl:NamedIndividual>
```

Figure 4. Sample individuals in the Linked Archives Model

Through the modeling process, we learned to start with a few, "fail fast," and modify quickly. Although we had a basic idea about what the model would look like, not everything worked the way we had hoped for. An example is the *Agent* class. While it is a widely adopted entity in many ontologies, it did serve an immediate purpose in our current model. As soon as we realized the drawbacks of *Agent* class after testing it with individuals for the class, we quickly made the decision to remove this class and used *Person* and *Organization* directly instead. This reduced an extra layer in representation and can save future troubles in creating transformation code.

Discussion

Modeling metadata for any collection requires deep consideration of the nature of the objects and their potential uses. During this project, we focused on archival collections with the goal of bringing item-level metadata to the forefront to support end user discovery activities through linked data. Recognizing that existing linked data models for archives are complex and may not capture the idiosyncratic nature of these materials (Matienzo et al., 2017), we took a combination of top-down and bottom-up approaches in examining existing data and relationships among the materials. Analysis of the three media collections in our sample revealed a number of entities of interest, though the preponderance of discovery metadata at the physical item level signaled the need to focus on this class in particular. The *Digital Item* class, representing the digitized surrogate materials, was largely associated with internal-use metadata of a technical and administrative nature, and much of this metadata was thus excluded from current modeling efforts. As a result, we implemented a modular strategy to our overall approach, focusing first on a discovery module and the supporting descriptive metadata, while leaving technical, administrative, and preservation functions to future modules to be developed.

The current modeling process thus required us to focus on elements of relevance to end users in discovery environments. Even within this tight scope, however, we confronted challenges related to the general nature of archival materials. Though we were able to incorporate many properties from existing, widely used vocabularies into our model, including Dublin Core, Schema.org, and various Library of Congress standards, some aspects of our collections required us to create new properties. For example, absent from these standards were some of the object properties that connect individuals of two classes. This demonstrates the idiosyncratic ways in which many archival collections have been organized and represented at their holding institutions. In addition, though some persons and organizations in the collections were well-known and had universal identifiers available through sources such as VIAF, most entities in the collection bore only local identifiers, requiring us to mint and rely on our own identifiers during individual instance modeling. Further work will be needed to record equivalencies between our identifiers and currently existing ones in other systems. Other challenges associated with working with this archival data included quality issues (inconsistent or missing elements), and the complexity of relationships among entities. Prominent examples of the latter included variations in the series relationship (publisher series, archival series, television series), and physical items with relationships to multiple digital items (e.g., video tape with both a transcript and a digital video file). Clearly modeling and labeling these relationships required careful consideration, particularly as archival terminology and relationships are often opaque to end users (Yakel, 2004).

Though modeling metadata for media materials presents its own considerations, no single media type posed significant challenges during this study's modeling process. Modeling for individual media characteristics in the current discovery module was straightforward; however, greater challenges were posed by the nature of uniting heterogeneous collections under one linked data model. For instance, in the existing metadata, the dimensions element covered all measurements for physical and digital items (e.g., height and width for photographs, pixel measurements for files). The nature of the measurements given in this field varied according to media type, making it infeasible to map to more granular existing elements from sources such as Schema.org. Keeping metadata applicable to all media types can enhance consistent user discovery across collections, but runs the risk of ignoring the unique properties of certain media

types that could also support discovery. Given the presence of specialized, media focused ontologies (for instance, Daquino et al., 2017), it is worth considering if a more modular approach to different material types could be employed, united under one top-level, general archival model. This approach is worth further consideration. Another notable challenge is how to maximize the benefits of publishing item-level archival data as linked data. Though no archival collection exists in isolation, in the present study, no immediate crossover was found between the three collections chosen for this project. No names, works, events, or subjects appearing in one collection also appeared in another; linked data publication of these datasets thus offers little new in terms of cross collections discovery. Still, all three collections are concerned with entertainment in 20th century America, and it is possible that digital humanities scholars may yet find and create connections among these materials through their work. Full publication of our data online also affords the opportunity to link with collections and entities from other institutions and data sets. Making these connections will require additional effort in the future, but hold potentials for increased user discovery of cultural heritage materials across the web of data.

Conclusion

Though archival resources may be valued for their uniqueness, they do not exist in isolation from each other, and stand to benefit from linked data treatments capable of exposing them to a wider network of resources and potential users. To leverage these benefits, existing, item-level metadata in archives must be remodeled as linked data, a particularly challenging process for heterogeneous collections. This paper presents the development of a linked data solution capable of making discovery-focused metadata available for collections of archival media materials, including photographs, sound recordings, and video recordings. The Linked Archives model was developed from an examination of existing metadata, alongside comparison to and incorporation of elements from pre-existing ontologies and vocabularies such as Schema.org and Dublin Core. In its current form, the model offers an initial discovery module capable of presenting archival media resources as linked data. Several limitations of note affected the development process, including the small sample size (three collections), and the restriction of data to one archival institution. Further development of the discovery module will rely on subsequent testing with additional collections and the incorporation of data from other institutions. Still, the current model presents a simplified solution to supporting further linking of unique but interrelated data in archives. Its development contributes to work exploring the range of archival and special collections and how these materials can be addressed under encompassing linked data models.

Beyond further developing the current discovery module, future work on the Linked Archives model encompasses other efforts as well. Further pilot conversion of individual instance data is underway, and expected to reveal additional refinements for the current model. A project to begin interlinking instance data with other data sets, including VIAF and Library of Congress Linked Data Service, is also in development. Given the importance of discovery user tasks to the current work, it will also be necessary to solicit the opinions and needs of archivists and users of archival data in assessing and strengthening the model; this could be accomplished through user testing, focus groups, or interviews. The project's investigators also look forward to the development of administration and preservation modules, and potentially others, to support the array of archival user tasks beyond resource discovery.

Acknowledgements: We thank Deirdre Joyce and Michele Combs of Syracuse University Library for providing the metadata records and informational assistance for this project.

References

Anderson, S. & Blanke, T. (2015). Infrastructure as intermeditation – from archives to research infrastructures. Journal of Documentation, 71(6): 1183-1202. https://doi.org/10.1108/JD-07-2014-0095

Andro, M., Asselin, E., & Maisonneuve, M. (2012). Digital libraries: Comparison of 10 software. Library Collections, Acquisitions, and Technical Services, 36(3-4), 79-83.

Daquino, M., Mambelli, F., Peroni, S., Tomasi, F., & Vitali, F. (2017). Enhancing semantic expressivity in the cultural heritage domain: Exposing the zeri photo archive as linked open data. Journal on Computing and Cultural Heritage (JOCCH), 10(4), 1-21. doi:10.1145/3051487

Elings, M. W., & Waibel, G. (2007). Metadata for all: Descriptive standards and metadata sharing across libraries, archives and museums. First Monday, 12(3).

Gracy, K. (2015). Archival description and linked data: a preliminary study of opportunities and implementation challenges. Archival Science, 15(3): 239–294.

Gracy, K., Zeng, M. L., & L. Skirvin. (2013). Exploring methods to improve access to Music resources by aligning library Data with Linked Data: A report of methodologies and preliminary findings. Journal of the American Society for Information Science & Technology, 64(10), 2078–2099. https://doi.org/10.1002/asi.22914

Isaac, A. (2011). Europeana data model primer. https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf

LODLAM: Linked open data in libraries archives and museums (2019). Retrieved from http://lodlam.net/

Matienzo, M. A., Roke, E. R., & Carlson, S. (2017). Creating a linked data-friendly metadata application profile for archival description. arXiv preprint arXiv:1710.09688.

Miguez, M. (2018). Linked Data for archivists: Graphs and Rhizomes. Society of Florida Archivists Journal, 1(1): 6-12.

Mitchell, E. T. (2013). Three case studies in linked open data. Library Technology Reports, 49(5), 26-43.

Morgan, E. L. & LiAM. (2014). Linked Archival Metadata: A Guidebook. http://sites.tufts.edu/liam/

Motik, B. & Parsia, B. (2012). OWL 2 Web Ontology Language: Structural Specification and Functional-Style Syntax (2nd ed.). https://www.w3.org/TR/2012/REC-owl2-syntax-20121211/

Niu, J. (2016). Linked Data for Archives. Archivaria 82, 83-110. Association of Canadian Archivists. Retrieved February 11, 2019, from Project MUSE database.

Park, O. N. (2015). Development of linked data for archives in Korea. D-Lib Magazine, 21(3), 6.

Pearce-Moses, R. (2005). A Glossary of Archival and Records Terminology. Chicago: The Society of American Archivists.

SCRC, Syracuse University Library. (2019). Special Collections Online. https://scrconline.syr.edu/xtf/search?brand=scrc;repository=scrc;f1-collection=Ronald%20G.%20Becker%20Collection%20of%20Charles%20Eisenmann%20Photographs;startDoc=21

Shaw, R., Troncy, R., & Hardman, L. (2009). Lode: Linking open descriptions of events. In Asian semantic web conference (pp. 153-167). Springer, Berlin, Heidelberg.

Society of American Archivists. (2013). Describing archives : A content standard (2nd ed.). Chicago : Society of American Archivists.

Stevenson, A. (2018). Is Linked Data an appropriate technology for implementing an archive's catalogue? Archives Hub Blog, https://blog.archiveshub.jisc.ac.uk/2018/05/09/is-linked-data-an-appropriate-technology-for-implementing-an-archives-catalogue/

Waitelonis, J., & Sack, H. (2009, September). Augmenting Video Search with Linked Open Data. In I-SEMANTICS (pp. 550-558).

Yakel, E. (2004). Encoded archival description: Are finding aids boundary spanners or barriers for users?. Journal of Archival Organization, 2(1-2), 63-77.