Archives and Special Collections Linked Data: Navigating between Notes and Nodes

OCLC Research Archives and Special Collections Linked Data Review Group

OCLC

MEMBERS OF THE OCLC RESEARCH ARCHIVES AND SPECIAL COLLECTIONS LINKED DATA REVIEW GROUP

- Erin Blake, Folger Shakespeare Library
- Itza Carbajal, University of Texas Austin
- Regine Heberlein, Princeton University
- Sarah Horowitz, Haverford College
- Jason Kovari, Cornell University
- Vanessa Lacey, University of Cambridge
- Cory Lampert, University of Nevada, Las Vegas
- Darnelle Melvin, University of Nevada, Las Vegas
- Holly Mengel, University of Pennsylvania
- Cory Nimer, Brigham Young University
- Maria Oldal, Morgan Library and Museum
- Merrilee Proffitt, OCLC
- Nathan Putnam, OCLC
- Arielle Rambo, Library Company of Philadelphia
- Elizabeth Roke, Emory University
- Eric de Ruijter, International Institute of Social History
- Dan Santamaria, Tufts University
- Karen Smith-Yoshimura, OCLC
- Weatherly Stephan, New York University
- Bruce Washburn, OCLC
- Chela Scott Weber, OCLC

July 2020

OCLC Research
Dublin, Ohio 43017 USA
www.oclc.org

ISBN: 978-1-55653-155-2
DOI: 10.25333/4gtz-zd88
OCLC Control Number: 1165363462

ORCID iDs
Erin C. Blake, Folger Shakespeare Library https://orcid.org/0000-0002-9275-738X
Itza Carbajal, University of Texas Austin https://orcid.org/0000-0001-5316-658X
Regine Heberlein, Princeton University https://orcid.org/0000-0002-0803-8025
Sarah Horowitz, Haverford College https://orcid.org/0000-0001-6230-9984
Jason Kovari, Cornell University https://orcid.org/0000-0002-7090-9071
Vanessa Lacey, University of Cambridge https://orcid.org/0000-0002-1905-9489
Cory Lampert, University of Nevada, Las Vegas https://orcid.org/0000-0002-9467-5214
Darnelle Melvin, University of Nevada, Las Vegas https://orcid.org/0000-0002-4614-3504
Holly Mengel, University of Pennsylvania https://orcid.org/0000-0001-5258-4640
Cory Nimer, Brigham Young University https://orcid.org/0000-0001-6572-2942
Maria Oldal, Morgan Library and Museum https://orcid.org/0000-0001-5258-4640
Merrilee Proffitt, OCLC https://orcid.org/0000-0002-2322-8337
Nathan Putnam, OCLC https://orcid.org/0000-0002-3984-3035
Arielle Rambo, Library Company of Philadelphia https://orcid.org/0000-0001-5655-8976
Elizabeth Roke, Emory University https://orcid.org/0000-0001-8495-7501
Eric de Ruijter, International Institute of Social History https://orcid.org/0000-0003-0244-9437
Dan Santamaria, Tufts University https://orcid.org/0000-0003-4097-5964
Karen Smith-Yoshimura, OCLC https://orcid.org/0000-0002-8757-2962
Weatherly Stephan, New York University https://orcid.org/0000-0002-6381-2036
Bruce Washburn, OCLC https://orcid.org/0000-0003-4396-7345
Chela Scott Weber, OCLC https://orcid.org/0000-0002-6358-5128

Please direct correspondence to:
OCLC Research
oclcresearch@oclc.org

Suggested citation:
OCLC Research Archives and Special Collections Linked Data Review Group. 2020. *Archives and Special Collections Linked Data: Navigating between Notes and Nodes*. Dublin, OH: OCLC Research. https://doi.org/10.25333/4gtz-zd88.

CONTENTS

INTRODUCTION

As libraries and other cultural heritage organizations plot a course to move toward a linked data future, it is important to ensure that all collections are represented in that future. While published works represented in monographs may be relatively easy to represent in linked data, this is not the case for special collections materials (rare books, manuscripts, photographs, institutional archives, etc.) that may be unique or may have special physical characteristics that are of particular interest for study.

To help fill this knowledge gap, the OCLC Research Library Partnership (RLP) convened an archives and special collections linked data review group. The RLP group was supported by OCLC Research and OCLC Global Product Management Staff. Because OCLC Research has a long and strong history of working both on behalf of the archives and special collections community and on behalf of the larger library community in linked data, taking on the issue of special collections and linked data under the auspices of the RLP was a natural fit.

OCLC's work in linked data stretches back over a decade, starting with publishing linked data (first FAST and VIAF followed by all of WorldCat.org). OCLC Research then investigated end user discovery applications across descriptions of people, organizations, places, events, concepts, and works, and then the processes needed to create original linked data. The two linked data pilots—Project Passage and CONTENTdm—determined how all these components work together to create, edit, use, and publish linked data in a single system. This collective work and expertise culminated in a Mellon Foundation grant in 2020 to design, build, and publish a Shared Entity Management Infrastructure to support linked data by libraries around the world.[1] This project takes advantage of all OCLC's linked data research, prototyping, and lessons learned to build a production-ready infrastructure.

There are several points where OCLC's work with special collections and linked data intersect, including:

- The OCLC Project Passage initiative (2017-2018), in which 16 libraries experimented with creating and editing linked data to describe resources.[2] This work revealed that more community discussion was needed among archivists and special collection librarians to redraw the line between structured data and the narrative textual data embedded in their current practices. Could linked data include the *context* required to discover and interpret a resource?
- OCLC's CONTENTdm Linked Data Pilot (2019-2020), which focused on transforming descriptions of digitized materials into descriptions of their associated entities (creative works, people, places, etc.).[3] Five institutions that use CONTENTdm participated in the pilot, representing public libraries, academic libraries, and private research institutions, and the digital collections they selected for study covered a wide range of cultural materials and formats, though primarily based on photographic or cartographic original sources. One of the areas of research concentration for the pilot was the processes and workflows for reconciling and identifying entities that were not already described in authority control systems and how the resulting entities could be managed and shared.

The Archives and Special Collections Linked Data Review Group extended this work on linked data,[4] and, in response to identified community needs, OCLC recruited 15 professionals from the Research Library Partnership to explore the key areas of concern in transitioning to a linked data environment. This group met between October 2019 and April 2020 during monthly virtual meetings. Members of the group presented on a variety of projects to help showcase promising areas for linked data for special collections, as well as to explore areas of concern. This OCLC Research publication is a summary of findings from those discussions.

Linked Data Features and Concerns

Review group members explored linked data as a series of opportunities for special collection description, as well as areas that will need additional work, thought, or attention.

FEATURES AND BENEFITS

Multilingual support

- In some linked data systems (e.g., Wikibase), multilingual support is foundational.

Entity relationships

- Structured data, with the exception of narrative-rich description, that defines relationships between entities is more expressive and efficient than record-oriented description.

Linking disparate data models

- The ability to define domain-specific data models and crosswalk to other shared models may be a way to balance prescriptive and permissive modeling.

Lower barrier to participation

- Wikidata is seen as a more accessible alternative to NACO participation/contribution.

BARRIERS AND CONCERNS

Context trapped in narrative description

- Descriptions of special collections and archives make extensive use of note fields. Archival collections and other collections (such as medieval manuscripts) make heavy use of semi-structured narrations that go on for paragraphs or pages. When descriptions for these items or collections are represented in systems that were designed primarily for bibliographic description, much valuable data is recorded in unstructured notes. In either case, it is labor intensive to extract entities and relationships from these unstructured notes. This data may be well served by linked data structures, but the number of unique cases for archival and other traditions of descriptive practice may make it difficult to move forward as a community.

Complex models

- Problems have arisen due to trying to model everything in the description of an item or collection when not all data maps well to linked data, such as notes.
- There are many local data models, and those institutions may resist adopting a shared model.

- Descriptions that adhere to complex models for very specific material types, or locally developed models make crosswalking into flatter linked data structures more difficult.

High-barrier infrastructure

- Software, data, and workflows take a lot of work to establish and often lack institutional buy-in and support.
- Lack of support from identifiers in the data, such as issues with minting identifiers and persistency.

Scalability

- Can linked data standards, practices, and workflows scale? Institutions that collect and steward these materials include both large, relatively well-funded institutions and small and less well-resourced repositories. There are few successful exemplars of linked data implementation in special collections; stories that exemplify success will be necessary to build support with resource allocators and to demonstrate utility to dispersed user communities who will directly benefit from linked data efforts.

Issues in Depth

Some areas of focus, whether opportunities or challenges, were discussed in depth and came up multiple times throughout the group's discussions.

DESCRIPTIVE DATA MODELS DO NOT CURRENTLY SERVE SPECIAL COLLECTIONS

The dominant data encoding standards are not well adapted for special collections in a linked data environment. The special collections community has tried to adapt to those standards. The result has been the "note-ification" of descriptive information, with a lot of information found in note fields. On the one hand, we have an opportunity in this moment of change to develop more agency and to shift the conversation to the benefit of special collections description and discovery. On the other hand, current archival standards allow for a high amount of variability, and enforcing a single way of doing things could be a hard sell for those accustomed to variability (or for those who rely on tools that promote or prescribe different ways of doing the same thing). Discussions about linked data can be leveraged to develop community consensus, especially as adoption rates increase for tools, systems, and software.

DISCURSIVE DESCRIPTION PRESENTS A CHALLENGE TO ENTIFICATION

Linked data forces you to structure your data. But there are clear limits to being able to clearly express things that are in paragraph form. Archival description makes especially extensive use of note fields. So much of our data is perceived to be rich and not easily adapted to identifier-based systems. Migrating this data to entities is time consuming and requires as-yet underdeveloped data models; it therefore should be approached as an iterative process. This process can be aided by machine processes (API lookups, Open Refine processing, etc.). Several of the projects looked at discursive text that explores the limits of entification, revealing in many cases "a hard remainder" that some data cannot be entified (such as interpretive passages) or will be difficult to convert to entities (such as common names that are hard to differentiate).

Perhaps entification should be viewed as an iterative process, one that can be carried out in parallel with other descriptive practices. Entification provides better linking to the information passages, but it will not replace relevant discursive texts, such as exhibition labels, long catalog entries, or other interpretive passages. Ignoring these would result in data loss during conversion, but there is also a danger in creating linked data models in name only by permitting textual data to remain in the data structure. There must be a balance. Our data models and descriptive practice will demand entified data alongside longer form descriptions, such as discursive notes that are not entifyable.

This shift from notes to entities should influence descriptive practice and take the profession toward data that wants to be structured (or from a discovery standpoint, data that benefits the end user by being structured) and away from materials that require a narrative. This is not to say that all data can be structured, but descriptive standards must place a premium on controlled value lists over unstructured narrative. This will be a major shift in mindset and practice that should be supported by tools, but it may not be as difficult as it seems. Within discursive description, institutions often ascribe to utilizing standard phrases that appear to be free text but are in fact carefully controlled; these could be pulled out by machine as structured data. Newer repositories, or collections that have not been cataloged, may be advantaged in not having legacy practices or descriptions.

POTENTIAL FOR BETTER DISCOVERY

Many discovery-related tasks are challenging or impossible in the current library discovery environment. A linked data infrastructure would likely make these tasks possible. For example:

- Inclusion of identity markers of people represented in the collections.
    - e.g., gender: "I'm looking for women in printing in the 16th / 17th century."
- The need for information on correspondents' networks, within and across collections, including geographically or organization-based circles.
- The ability to study social mobility by connecting information on people with their locations to see how they moved around, spread knowledge, and met other people.
- Inclusion and increased visibility of item-specific information as well as variances and the surfacing of resource uniqueness in a shared bibliographic environment where manifestation descriptions are standard.
    - e.g., "I'd like to find all the bindings/endpapers designed and produced by X."
- Even though linked data structures will bring significant advantages to discovery, end users will need to adapt to new ways of interacting with linked data structures.

PRESCRIPTIVE VS. PERMISSIVE DATA MODELING

Linked data approaches allow for more permissive modeling than we have been accustomed to. Current data models, where they exist, are shaped by disciplinary practice (e.g., descriptions of coins are shaped by needs of numismatists and archeologists). Some elements of description that are shared or appear to be shared between disciplinary practices may mesh well together (bindings on rare books and artist's books, for example), whereas other elements may not mesh together so easily. Because the data may not be harmonized, there is a need for "statements of equivalence." This is a problem of siloed data and the data models for different disciplinary practices, such as those for describing art, versus archival materials, books, coins, etc. These items are not the same, but crosswalking materials together can be difficult assuming mappings or crosswalks exist at all.

ETHICAL ISSUES AND COMMUNITY ENGAGEMENT

Linked data in the library community is often assumed to be "open," and that openness is broadly considered to be an inherent good. However, there are many good reasons for us to have "closed" data that is not shared, or data that is only partially shared (regardless of whether it is linked). Publishing data can put people and organizations in harm's way, such as information about human rights organizations, opposition political groups, etc. Making this data available in linked open data structures can be particularly risky because it is easy for this information to be exploited and propagated into other systems. This said, data should not be viewed as open/closed in a binary or definitive state. As situations change and evolve, data that is closed or restricted can be made more open later. This is not dissimilar to archival practice around embargoing access to collection content for a predetermined amount of time.

Libraries, archives, and museums are shifting to approaches that encourage ethical engagement of communities. Web-based, accessible tooling can help to support various forms of engagement; linked data structures afford one of many avenues to invite in communities that are closer to the material to describe as well as enrich and correct the information we have. This moment offers an opportunity to create pathways to disrupt our constructs of authority and to provide a space and a place for others with more expertise (researchers, community members) to contribute their knowledge. Institutions should be careful when balancing risk and opportunity for those who may be impacted.

CHALLENGES AROUND MULTILINGUALITY

The group saw an opportunity for linked data approaches to support multilinguality at every level. There are serious challenges in sourcing authority records in multiple languages in the current environment; in some cases, sources exist but are unavailable. In other cases, major contributors of standards do not place emphasis on reaching multilingual audiences with the default and only language being English. A major challenge is to make subject vocabularies flexible, with some level of specificity but in a way that they don't need to be continually updated. We have the opportunity to clearly identify the original language from translated terms.

SUSTAINABILITY

To date, the majority of linked data efforts have been grant-funded or special one-off projects. This has impacted the perception of value and utility, especially for library administrators, and has made sustainability of these projects problematic. Existing systems (such as Wikibase) are perceived as novel, and it is burdensome to set up and receive approval for them. There is an educational barrier that is difficult to overcome ("the trouble with triples"). We are ultimately looking at a lot of labor with little commitment to financing that labor.

On the positive side, open structures allow for institutions that are not NACO contributors to create records/entities at the point of need.

Everyone wants easy-to-use and implementable tools embedded in a sustainable infrastructure.

ARCHIVAL ISSUES

Archival description attempts to document both content and context and represent hierarchical and other relationships among and across records. The recent revision of "Describing Archives: A *Content Standards Statement of Principles*"[5] emphasizes the need to describe relationships when creating archival description. In particular, Principle 4, which states: "Records, agents, activities,

and the relationships between them are the four fundamental concepts that constitute archival description."[6] Current systems reflect hierarchical arrangements of records but often struggle to represent additional relationships in meaningful ways. Linked data offers additional possibilities.

The focus on backlogs and minimal processing since the publication of "More Product, Less Process" in 2005[7] has encouraged thinking iteratively about descriptive efforts with an emphasis on using professional judgement to find a golden minimum for most descriptive work and on describing archival material in aggregate. Much linked data work in archives has focused on boutique projects involving a subset of collections related to a specific subject or topic and on remediating and entifying existing description, which is time and labor intensive. This work remains out of reach for most small and medium size archives and special collections, though these repositories house rich and unique holdings. There will need to be a shift in mindset to think about how to structure archival description natively as linked data and develop the tools and standards necessary to create linked data for archives in scalable production environments. One first step might be the ability to include stable, authoritative identifiers in tools and systems already commonly used in creating and managing archival description.

THE NEED TO EXPRESS RELATIONSHIPS AND CHANGE OVER TIME

Relationships are critical to special collections: we have built and invested in collections purposefully because of their relationship to one another and to a topical focus or subject matter. Expression of relationship is often central to understanding the collection as a research object—who created it and under what circumstances, who owned or used it and in what ways, how it evolved over time and under whose influence. These are complex and often layered relationships between people, organizations, geographies, contexts, events, and records. These relationships are (or should be) documented in scope notes, provenance notes, etc. and in the implied inheritance of hierarchical structures in archival description. There are opportunities for linked data to express these relationships more efficiently and in a way where the data can be exploited more easily to reveal new ways of looking at collections and to facilitate discovery.

We need to express both that people and organizations have a relationship to records and to specify the nature of those relationships and how they change over time. Current systems don't support the clear expression of these relationships and only comfortably express a limited set of relationships between agents and resources. These relationships go beyond the boundaries of institutions, collections, and cross formats, but current systems do not surface those connections. The real opportunities are in surfacing these relationships where they can be utilized in a discovery environment, allowing for different constellations of how resources are related (such as by a common creator, a common source of acquisition, a common subject, a common format, or a common event).

THE LONG TAIL OF AUTHORITIES/IDENTIFIERS IN SPECIAL COLLECTIONS

Because of the rare or unique—and often local or regional nature of special collections—they need to represent many people, families, and corporations that are not in authority files like the LC/NACO authority file. The barrier to being involved in NACO and similar is too high (cost, training, etc.) for many archives and special collections. Relatedly, these name authorities are likely to be used across types of collections—library, archive, museum. How do we create and manage authorities across systems and institutions in a lighter-weight way that still can take advantage of the network? It is not possible (or desirable) to create entities for everything; instead, we should develop approaches that help to balance local versus global IDs much like NACO and local authority files do now. Such an approach should allow domain experts to enrich and link to local identifiers.

CONCLUSION

Libraries are in a state of moving from experimentation to production with linked data, and it is vitally important that the needs for descriptions of special collections materials are not left behind at this critical moment. This document helps to outline some of the challenges and opportunities in our shared linked data future. Our thanks and deep appreciation to all who shared discussions and ideas on this journey.

ACKNOWLEDGMENTS

This paper benefitted from the many years of work that has been done by OCLC around linked data. This work started with experiments in publishing linked data sets beginning in 2009 (FAST, VIAF, etc.), and continued with the EntityJS Research Project (2013) and the Person Entity Lookup Pilot (2014). More recent experiments such as the ContentDM Metadata Refinery (2015-2016), Project Passage (2017-2018), and the ContentDM Linked Data Pilot (2019-2020) added to this knowledge base. We are grateful to OCLC staff and OCLC member libraries for this significant body of work.

Thanks to OCLC Software Architect Jeff Young for joining this group to share linked data concepts. Special thanks are also due to group member Cory Lampert from University of Nevada, Las Vegas, who inspired the "notes and nodes" theme for the publication title.

We are always grateful to our OCLC Research Library Partnership (RLP) collaborators who contribute to efforts such as this. The Archives and Special Collections Linked Data Review Group faced a special challenge as libraries globally shifted gears dramatically due to the impact of the COVID-19 pandemic in March 2020. This came during a critical time in this project, and these circumstances could have easily and understandably delayed the work of this group or the final stages of this publication. However, because of the diligence and dedication of group members who continued to devote effort toward the final stages of this project, we were able to bring it across the finish line under extraordinarily challenging circumstances.

This report couldn't have been published without the efforts of the OCLC Research publishing team, including Erica Melko and Jeanette McNicol. Thank you for always making sure our Is are dotted, our Ts are crossed, and our notes are properly formatted.

NOTES

1. OCLC WorldCat: OCLC and Linked Data. "Shared Entity Management Infrastructure." https://www.oclc.org/en/worldcat/oclc-and-linked-data/shared-entity-management-infrastructure.html.

2. Godby, Jean, Karen Smith-Yoshimura, Bruce Washburn, Kalan Davis, Karen Detling, Christine Fernsebner Eslao, Steven Folsom, Xiaoli Li, Marc McGee, Karen Miller, Honor Moody, Holly Tomren, and Craig Thomas. 2019. *Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage*. Dublin, OH: OCLC Research. https://doi.org/10.25333/faq3-ax08.

3. OCLC. 2020. "CONTENTdm Linked Data Pilot." https://www.oclc.org/research/areas/data-science/linkeddata/contentdm-linked-data-pilot.html.

4. OCLC. 2020. "The OCLC Research Library Partnership: Archives and Special Collections Linked Data Review Group." https://www.oclc.org/research/partnership/working-groups/archives-special-collections-linked-data-review.html.

5. Society of American Archivists (SAA). 2019. Technical Subcommittee on Describing Archives: A Content Standard (TS-DACS). Version 2019.0.3. GitHub. https://saa-ts-dacs.github.io/.

6. SAA. 2019. "Principles of Archival Description, 4." In TS-DACS: Statement of Principals. Version 2019.0.3. GitHub. https://saa-ts-dacs.github.io/dacs/04_statement_of_principles.html#4-records-agents-activities-and-the-relationships-between-them-are-the-four-fundamental-concepts-that-constitute-archival-description.

7. Greene, Mark, and Dennis Meissner 2005. "More Product, Less Process: Revamping Traditional Archival Processing." *The American Archivist* 68, no. 2 (Fall/Winter): 208-263. https://doi.org/10.17723/aarc.68.2.c741823776k65863.

For more information about our work please visit our website at:
oc.lc/research

6565 Kilgour Place
Dublin, Ohio 43017-3395

T: 1-800-848-5878
T: +1-614-764-6000
F: +1-614-764-6096
www.oclc.org/research

ISBN: 978-1-55653-155-2
DOI: 10.25333/4gtz-zd88
RM-PR-216774-WWAE 2007