Pragmatic Principles for Archival Linked Data

Elizabeth Russey Roke and Ruth Kitchin Tillman

ABSTRACT

Where does linked data fit in archival description? How do we get from promise to implementation? This article evaluates the benefits and limitations of current approaches to linked data in archival work. It proposes four pragmatic principles for the archival community to follow when determining how to pursue linked data. This approach engages with communities (both inside and outside cultural heritage institutions) already publishing linked data, accounts for institutional resource limitations, and recognizes the need for technological, educational, and social support for institutions and workers. Through an examination of the work of the Archives and Linked Data Interest Group with Schema.org and Wikidata, the article provides case studies that explore how these pragmatic principles for archival linked data create inclusive, rather than exclusive, communities.

KEY WORDS

Metadata, Linked data, Finding aids

Linked data, an emerging metadata framework that prioritizes connections between resources over local record-based description, has been pitched as the future of library and archival metadata for the last decade. As an approach that requires major resource commitments, its implementation is not to be undertaken lightly. It necessitates a fundamental shift in how libraries and archives create, maintain, and publish metadata. Since 2014, the many iterations of the Linked Data for Libraries (LD4) grant[1] have created tools and recommendations aimed at making shared BIBFRAME a reality.[2] In 2018, the Swedish National Library fully transitioned from MARC to BIBFRAME.[3] For all the work to date, this is the only example of its adoption at scale. Despite this, momentum is building through systems and practices that would integrate it into existing workflows. FOLIO, an open-source library services platform currently under development but beginning to be adopted in larger libraries,[4] models its data structures on BIBFRAME and Dublin Core, rather than on MARC.[5] And, at time of this writing, the Program for Cooperative Cataloging (PCC) is sponsoring a year-long project in which catalogers at seventy-one institutions develop the skills to perform authority work in Wikidata.[6]

Linked data in archives raises some important questions. Some are practical: Should finding aids be completely transformed into linked data structures? Should we create a specific archival standard to fully replicate the elements used in archival description? These are certainly possibilities. Other questions of linked data cut to the core of the archival profession itself. As the archival focus broadens from truth and evidence to include memory and community,[7] how is it influenced and supported by the structural standards and descriptive systems we employ? Are there advantages to publishing linked data, particularly in how we establish, document, and maintain our commitments to donors, researchers, and staff? As Kathy Wisser challenged in an interview in 2014, "If we're actually going to invest some meaning into [linked data] and add value, let's really explore what it means to link things and which things really need to be linked in order to be useful."[8]

In this article, we engage with these questions of linked data and archival description and propose a pragmatic, principle-based approach. We ground this work in the descriptive standards widely adopted within the archival community and in recommendations for the ethical creation of linked data. We suggest that the archival community should focus linked data initiatives on archival description efforts that engage with communities already well underway in the creation, publication, and reuse of linked data. Using or adapting existing platforms and ontologies not only empowers us to build links beyond the archival community, it brings our work into the much broader ecosystem of knowledge and relationships promised by linked data.

Linked Data's Potential for Archival Description

> *Meaning in archival records is revealed through their contexts as much as through their contents. Archivists expose contextual significance by describing records, agents, activities, and the relationships between them.*
>
> —from DACS Principle 4[9]

The context of archival records is fundamental to their description. Biographical notes provide information about the functional relationships between content creators and the records they create. Arrangement schemes preserve semantic connections between records such as an author's notes and drafts within a literary collection. Processing notes and custodial histories record the provenance of a collection including the agents, locations, and dates of transfer, appraisal, and processing. Scope notes document the content of the collection and often include additional context on the events leading to its creation. In other words, good archival description goes beyond a list of items in a collection and attempts to contextualize them, describing "what the records are, what they mean, and the historical process by which they were created and maintained."[10]

The management and publication of archival metadata takes many forms reflecting the many ways it can be used. Human-readable finding aids, such as paper or HTML or PDF, are primarily intended for end users and serve researchers working with individual collections. EAD(XML) documents were originally designed as a means to encode paper finding aids for discovery in an online environment. MARC records ensure that archival collections are discoverable alongside bibliographic materials in library catalogs. Database-backed systems (such as ArchivesSpace) store metadata alongside collection management data and allow one to export finding aids in any of these formats. With so many possible approaches to archival description, what is the place for linked data?

Defining Linked Data in an Archival Context

The term "linked data" is often used informally in libraries and archives to describe a range of data encodings,[11] not all of which are formal linked data as envisioned by its originator, Tim Berners-Lee.[12] For example, because linked data uses URIs (Uniform Resource Identifiers) to represent things and express relationships between them, relational descriptions that incorporate URIs are sometimes referred to as "linked data." Such "linked data" generally meets some but not all of the four characteristics Berners-Lee uses to define linked data.[13]

In this article, we use "linked data" to mean both Berners-Lee's vision and a middle ground of transition toward that vision. That middle ground carries

great potential for the pragmatic principles we promote, allowing more than just the best-resourced institutions to create and even use linked data. This approach has precedent. One of the best-known linked data projects, Wikidata, requires a transformation layer[14] to become true linked data.

In an archival context, linked data offers a way to traverse relationships between descriptions published in different spaces and to create a fuller understanding about the things being described. Such links provide fuller information about the things connected. These may be accompanied by additional statements that connect one identifier to something such as data-typed values (e.g., xsd:date) or uncontrolled textual values (e.g., "Berners-Lee").

Each Resource Description Framework (RDF) statement may be loosely said to resemble a single sentence.[15] As with a sentence, some statements rely on or are best understood in the context of each other, while others may stand alone or as small groups. Multiple RDF ontologies[16] are frequently used together to meet different descriptive needs. Because of RDF's granularity and the potential for multiple ontologies to complement each other, these individual statements or groups of statements may be extracted from one document, reused in another, or combined into a new set of statements.

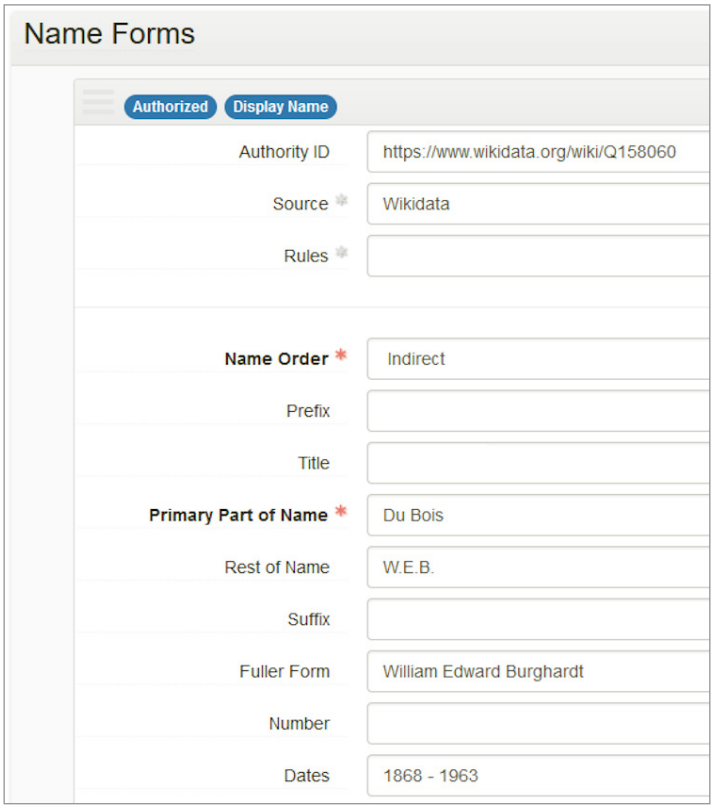

FIGURE 1. Screenshot of the ArchivesSpace Agent record

The examples that follow are small groups of RDF statements that demonstrate how links might be created between a hypothetical ArchivesSpace record for a Du Bois collection and Du Bois's Wikidata record. These groups were chosen to demonstrate how such links might reveal new relationships.

When the Agent record (see Figure 1) for W. E. B. Du Bois is assigned to a record in a creator role, the public user interface will generate the following embedded data using Schema.org:

```
{
"@context": "http://schema.org/",
"@id": "https://archives.example.edu/repositories/2/resources/35",
"@type": [
"Collection",
"ArchiveComponent"
],
"name": "W.E.B. Du Bois Papers",
"creator": [
{
   "@id": "https://archives.example.edu/agents/people/35246",
   "@type": "Person",
   "name": "Du Bois, W. E. B. (William Edward Burghardt), 1868-1963",
   "sameAs": "https://www.wikidata.org/wiki/Q158060"
}
], ... {
```

Example: Wikidata

```
@prefix wd: <http://www.wikidata.org/entity/>
@prefix wdt: <http://www.wikidata.org/prop/direct/>
wd:Q158060 wdt:P69 wd:Q49087, wd:Q13371, wd:Ql152087, wd:Q151510.
```

In the first example, the object of the schema:creator statement is W. E. B. Du Bois's Wikidata record. The statement could be read in English as:

> "this resource's creator is the entity represented by Wikidata entity Q158060 (Du Bois)."

The second record contains a single statement set from that Wikidata record indicating that Du Bois (wd:Q158060) was educated at (wdt:P69) Fisk University (wd:Q49087), Harvard University (wd:Q13371), Humboldt University of Berlin (wd:Q152087), and Heidelberg University (wd:Q15151). Fisk University's record contains a statement that it is an instance of "historically black colleges and universities" (HBCUs).

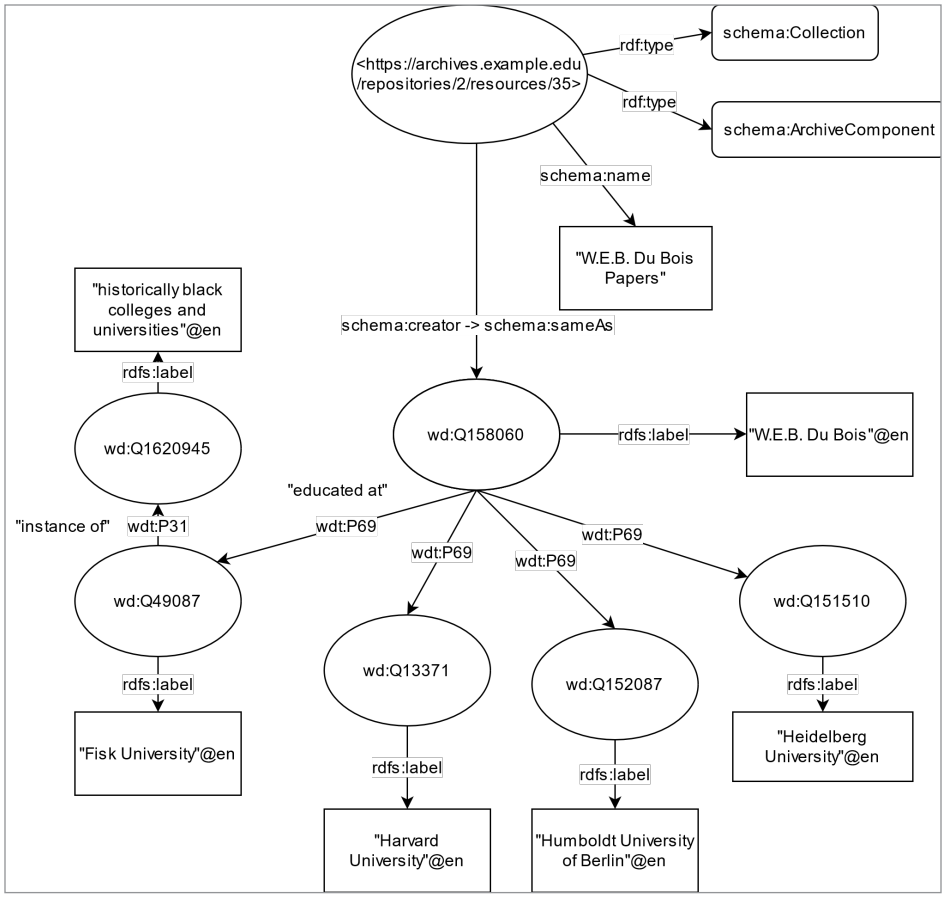

FIGURE 2. Graph demonstrating implied connections between ArchivesSpace and Wikidata

Although the first example does not record Du Bois's ISAAR-CPF 5.3 ("Relationships with other corporate bodies, persons, or families") in any machine-actionable way, its connection to the second example allows for the querying of such information and more from the linked Wikidata entity (see Figure 2). One might use this information to enhance either record, to form a new record, or to facilitate a complicated query that draws upon information beyond local description, such as "Which archives hold the papers of HBCU graduates?"[17] or "Can we identify HBCU attendees in our archives?"

Review of Approaches to Archival Linked Data

Early experiments in linked data for archives[18] focused on simply converting EAD XML to RDF. The Linked Open Copac and Archives Hub (LOCAH) project (2010–2011) and its continuation, Linking Lives, (2011–2012), one of the

first large-scale projects to work with archival data, mapped elements in EAD to existing linked data properties and vocabularies such as Friend of a Friend (FOAF), SKOS, RDFS, and Dublin Core.[19] While the prototype mapped many EAD elements successfully to RDF, project participant Jane Stevenson reports in her evaluation of the project that it was difficult to model inherited data from the EAD hierarchical structure, resulting in the homogenization of complex concepts and missing data.[20] Other projects using this conversion approach focus on enhancing existing metadata with linked data-based vocabularies such as VIAF, DBpedia, and GeoNames, providing a path for archives to engage in linked data, even while remaining in XML-based data structures and relying on existing systems for access and discovery.[21] While the conversion approach to linked data support and encourage the development of linked data skill sets among archivists, the resulting data are only supplemental, providing an alternative approach to description that fulfills some use cases, but never fully replacing the need to maintain traditional EAD-based metadata.

A second category of linked data work focuses on investigating the potential of computational methodologies for researcher-driven research and for improved discovery interfaces. These experiments, while based on linked data structures, focus mostly on the outcomes of linked data and the research it makes possible rather than the data modeling on which it relies. The archives at Emory University partnered with the Emory Center for Digital Scholarship in 2015 to explore how network graphs derived from data in finding aids could reveal new insights about a cluster of collections documenting the Belfast Group, an Irish writing workshop.[22] The Linked Jazz Project at the Pratt Institute[23] used oral history transcripts of jazz musicians from several archives as the basis for building an interface that not only exposes social networks but also crowdsources to discover previously unknown relationships between entities. The Carnegie Hall Data Lab has created a "learning space" for experiments in linked data, leveraging data sets published in linked data not only as a resource for end users but also as a way for staff to expand their understanding of linked open data, semantic technologies, and data-driven research.[24] While these types of projects can be less labor-intensive than a complete system redesign, they depend upon institutional support (including commitment to their maintenance), space to experiment, and staff expertise, and they rarely result in institutional shifts to linked data structures.

A final category of linked data work in cultural heritage institutions focuses on completely reconceptualizing cultural heritage metadata using linked data. This approach, which includes the Europeana Data Model (EDM), Records in Context (RiC-O), BIBFRAME, and Linked.Art, aims to develop new ontologies specifically for library, archives, and museum communities. While these ontologies may map to earlier encoding standards, their primary goal is to develop new

semantic models based on existing content standards. The EDM, for instance, was created to support the description of digital and digitized cultural heritage objects aggregated from a wide variety of sources with different descriptive practices. It reuses fields from several major vocabularies and mints new properties and classes only when needed to complete the model. Linked.Art, which is still under development, follows a similar approach, combining the CIDOC Conceptual Reference Model and Getty vocabularies.[25]

RiC-O is an effort led by the International Council on Archives Expert Group on Archival Description that brings together concepts from the ICA content standards into a unified ontology. It takes an opposite approach to the EDM and Linked.art, defining a complete set of classes and properties. While developing entirely new linked data frameworks avoids mapping problems and provides an opportunity to rethink the structure of our data in a new environment, developing new ontologies is only the first step in a difficult and time-intensive process. Ontology projects must also support implementation and develop a user community that derives tangible benefits from new data models. BIBFRAME, a multiyear project begun in 2012 to model bibliographic metadata and replace MARC, surfaces some of these challenges. Although the second version of BIBFRAME was released in 2016, there are still only a handful of adopters to date because of the need to develop new editing and discovery technology, workflows, and policies in addition to the ontology itself.

The path to linked data in the wider cultural heritage community has necessitated the creation of specific projects to support linked data creators, publishers, and consumers. The LD4 series of grants aims to fill some of the aforementioned gaps by building technical, functional, and social infrastructure to support adoption of the BIBFRAME standard. From late 2017 to September 2018, the OCLC-sponsored Project Passage sought to address metadata librarians' reports that linked data creation and editing tools are "difficult to use" and that their use requires prior knowledge of the technical side of linked data.[26] OCLC's research and engineering staff adapted Wikibase and MediaWiki, the open-source software underlying Wikidata, by adding visualization and querying layers that immediately display the connections inferred from participants' edits, providing participants with a ready-made linked data editing interface.[27] Participants tackled examples based on their experience describing materials, rather than their experience with linked data. OCLC's work to provide a technical platform for linked data work gave participants an accessible entry point into linked data that built on their expertise as catalogers without requiring them to understand the more technical aspects of linked data probably best left to developers.

The recently launched PCC Wikidata Pilot focusing on name authority work extends these shared infrastructure experiments into the wider world of

linked data. Its participants work directly in Wikidata, rather than in a local Wikibase, benefiting from and extending linked data created by others. Nearly one-third (31 percent) of those involved come from non-PCC institutions,[28] a promising sign when PCC membership is often a reflection of institutional privilege and capacity. Experiments that provide a shared cataloging environment adapted from existing systems are an encouraging development. While requiring more trust in a centralized entity for development and maintenance, such projects open the door to those who do not have the technical knowledge or institutional support to build from the ground up.

Functions and Benefits of Linked Data for Archives

Archival systems comprise many functions, from inventory control to discovery. Historically, these functions have often been grouped together. In an era before relational databases and personal computers were available, the most efficient way to ensure that an archival worker could find all the information about a collection was to keep as much as possible in a paper finding aid.[29] As American archivists moved their descriptions online, they "initiated an almost straight transfer of the . . . analog form of the finding aid to the digital realm" and simply encoded it in EAD to support browser search and display.[30] To continue to support the practice of one document for all use cases, EAD included an "audience" attribute, which allowed encoders to hide data intended only for internal use. Wideman and others argue that the focus on these legacy practices undermined the potential offered by machine-actionable description.[31] An oft-cited benefit of linked data is granular, parsable data, but the move to database-driven archival management systems such as ArchivesSpace and Access to Memory (AtoM) achieves a similar result and has enabled the storage, search, and expression of descriptive and administrative data beyond what EAD-based systems made possible. We must evaluate, then, what other functions linked data supports that are not already replicated in XML or existing systems.

Linked data, which creates semantically precise connections between things, meets the functional need in archives for expressing specific, defined relationships between materials. Each RDF statement can carry semantic and domain-specific precision, and linked data allows for the judicious[32] combination of vocabularies. Using existing vocabularies allows us to benefit from work already done by expert communities. For example, one could use PREMIS, a preservation metadata standard, to describe the preservation actions taken on materials while also expressing its exhibition history using an entirely different vocabulary. One could record a geographic relationship to Bellefonte, Pennsylvania, using its Geonames representation https://www.geonames.org/5179799, which provides its name, geographic coordinates, and administrative information. Or

one could use https://www.wikidata.org/wiki/Q1133274, which carries a great deal of additional information about Bellefonte's history, civic structure, and acts as a hub linking to its representation in other linked data systems from Geonames to the Library of Congress.

Similarly, existing linked data models and ontologies can be used or adapted to model the complex relationships that make up a custodial history or describe a complex entity such as a family or corporate body. For example, corporate records may reflect multiple iterations of original order resulting from mergers of distinct companies into a single entity or holdings split across multiple institutions. As demonstrated in the Du Bois example, linked data can bring disparate data sources together into a more comprehensive whole that expands our understanding of the thing being described. This semantic precision in describing the types of relationships within and beyond a set of records encourages clearer description of the often-hidden influences on a collection.[33]

How might the use of linked data benefit our researchers? Linked data structures already support discovery and relational context in systems outside our own, from search engines to street maps to Wikimedia spaces. Relational context is particularly powerful when applied to a multi-institutional data set. Within archives, one may look at the Social Networks and Archival Context (SNAC) project, a database that expresses relationships using EAC-CPF, to see the potential for archival discovery systems when relationships between discrete pieces of data are linked together. SNAC aggregates and interlinks descriptions of people, collections, and other resources from a variety of sources, including MARC records from WorldCat, authority records from the Library of Congress, and finding aids. The SNAC entity for Flannery O'Connor,[34] for instance, contains biographical information from finding aids, name forms from her Library of Congress Name Authority File heading, and relationships to fifty-five persons/corporate bodies and sixty collections. As no single collection of her correspondence exists, a researcher can use SNAC as a hub to identify the many institutions where her letters and manuscripts can be found as stand-alone collections or within others' papers. Encoding relationships instead as linked data lays the groundwork for a similar outcome without the massive amount of remediation and database management required by SNAC. Effort could then be put into developing systems that dynamically process these relationships and display resulting networks of people, places, events, and records.

Beyond improving discovery in our own systems, linked data has the potential to assist users in finding our collections through the most common starting place: search engines. The largest search engines use data structured in the Schema.org linked data vocabulary to improve their indexing and enhance their results. By publishing our records in more ways than EAD/HTML/PDF, we have the potential to make them more visible to those who may not have even

thought to look for archival collections created by or related to the entity they're researching.

In short, linked data supports archival principles that encourage user-centered, flexible, iterative, and distributed approaches to archival description and access such as those outlined in the SAA Core Values.[35] By making our data more accessible on the web using the semantic structures and language of linked data, we meet our users where they are, rather than expecting them to come to us. Description of traditional collections could be improved by the integration of linked data created by communities to describe themselves. This could be as simple as using vocabularies such as Homosaurus,[36] a linked data vocabulary of LGBTQ-specific terminology, or as relying upon community-developed data repositories such as Wikidata for biographical histories. Linked data provides a technical framework for capturing the provenance and custodial history of archival collections, which has become even more important as institutional lines blur in postcustodial collecting approaches. Linked data would enable the creation of finding aids that draw upon multiple sources of information, both within and outside archives. This expansion of descriptive authority that archival workers have traditionally reserved for themselves would result in archival description that is more flexible, accurate, and responsive to communities and social change.[37]

All of these functions relate to how we publish and share data, rather than how we internally manage our collections or store descriptive data. Linked data provides no advantage for many archival functions, such as inventory management, and only equivalent support for the creation of document-type finding aids and narrative descriptions. As Julie Hardesty observes, simply converting metadata from XML to an RDF-based data model "does not make that data more shareable or more easily understood unless there are end user applications for using that data."[38] We see no advantage to rebuilding entire systems as linked data unless linked data offers concrete improvements for functions performed by the system.

Pragmatic Principles for Creating Archival Linked Data

Implementation of linked data frameworks in cultural heritage institutions will ultimately require enormous resources and much effort. This should be unsurprising if we look at the history of new standards in the library and archives community. The adoption of MARC and EAD required coordinated community efforts that, in some cases, stretched for decades. In the case of EAD, reaching the current implementation level[39] meant the development of long-standing workshops, a "cookbook" for data modeling, the creation of an SAA roundtable dedicated to EAD, and even archivist-created websites, such

as EADiva, that offered additional support to new users. The BIBFRAME linked data initiative echoes this history. As seen in this project, a fundamental transformation of our systems requires not just technical resources for building new platforms, systems, vocabularies, and ontologies but also significant staff time for skill-building and experimentation. When institutions struggle to fund core functions such as processing, only the most well resourced can take on the additional time, money, and training costs of experimenting in linked data or shifting metadata creation to a linked data ecosystem once one exists.[40]

This high cost of implementation has resulted in a concentration of linked data projects at only a few institutions. In *Perplexed Librarian*, Carlson et al. remind us that this concentration of linked data expertise and development in only a few organizations results in a vicious cycle that "stagnates the technical infrastructure that would welcome others into Linked Data Library Land" and encourages the perception that linked data is merely "vaporware, a much hyped project that will never come to fruition."[41] If the cultural heritage community fails to develop pathways for all to participate, not just those who have the resources and expertise to venture out on their own, the promise of interlinked, contextual discovery across multiple institutions and domains will ultimately be an empty one.

Our profession sorely needs an approach to linked data creation and consumption that lowers implementation barriers and extends the culture and benefits of linked data. In this section, we set out principles for a pragmatic approach to the creation and publication of archival description as linked data. By defining an approach, rather than endorsing an ontology or proposing a particular project, we recognize the variety of functions performed by archival description, the different capacities of archival workers and institutions, and the range of community-supported ontologies that exists outside our field. These pragmatic principles for linked data echo and build on frameworks for archives, such as Greene and Meissner's "More Product, Less Process," Theimer's Archives 2.0, and Santamaria's Extensible Processing, that prioritize users, encourage flexibility, and normalize iterative approaches to archival description and access.[42] The pragmatic principles we propose are designed to guide linked data exploration and adoption in tandem with enhancing existing systems and participating in projects outside our own discipline.[43]

Principle 1: Pragmatic Linked Data Strives for "Good Enough," Not Perfection

We can represent archival concepts, resources, and contexts as linked data without developing a unified archival ontology or entirely replacing the traditional finding aid. A "good enough" approach to linked data is an iterative and realistic approach to the generation and publication of linked data to meet specific needs and use cases. It recognizes that not all data needs to be published as linked data and that publishing linked data does not require storing archival description as linked data. It supports the reuse of ontologies created outside our field when their semantic definition suits our purpose.

We should focus our energies on areas where linked data excels and explore areas where it shows potential. In doing this, we are supported by the statement-based nature of linked data, in contrast to the document-based nature of XML. Linked data frameworks need not replicate the traditional finding aid nor fully replace databases such as ArchivesSpace to be worth pursuing. The move to linked data in archives must be an iterative process, with room made for varying levels of implementation.[44] Examples of this principle include projects that move the community down the road to linked data, such as adding identifiers to names and places or linking archival collections to linked data systems like Wikidata.

Principle 2: Pragmatic Linked Data Builds from Existing Standards, Ethical Principles, Platforms, and Processes

We can create useful linked data without building entirely new infrastructure. If a platform or process already exists and meets a need, how might we use or expand it? Such reimagination must be grounded in archival standards and ethics[45] but should extend beyond our current communities of practice. This might mean enhancing the platforms and processes already used in archival work or connecting them to outside systems. It also means engaging with those working outside the cultural heritage fields and bringing our knowledge and our collections into new communities. If the goal of linked data is to create a shared web of information, cultural heritage communities should participate in and collaborate with communities beyond their own.

While linked data allows us to build connections with disciplines outside our own, this must begin by engaging core archival standards. We identify core archival standards as the international and national content standards that apply to one's context and the ethical principles that shape one's practice. American archivists' content standards are ISAD(G), ISAAR-CPF, ISDIAH, ISDF, and DACS.[46] Working toward a linked data expression of archival description

begins with identifying which content standards and elements within those standards support the desired function.

Engaging with communities outside our own also requires a recommitment to our domain's ethical principles and norms. As the LD4 Ethics in Linked Data Affinity Group states, "ethics should be a forethought, not an afterthought, of any initiative."[47] As archivists and librarians, we have an uneven track record in the kinds of description we create and how we treat the people represented in our collections.[48] Over the past seven years, we have seen a reconsideration of ethical practices in which we should ground our work. We are asked to consider honest description,[49] incorporating radical empathy and a feminist ethic of care[50] acquiring consent,[51] supporting human rights,[52] decolonizing description,[53] respecting traditional knowledge,[54] and considering the privacy of the individuals represented in our materials.[55] The 2019 DACS Statement of Principles[56] now prioritizes our ethical responsibility to our users and to the people represented in our collections. Grounding ourselves in ethical principles is as important to undertaking linked data work as grounding ourselves in content standards.

Principle 3: Pragmatic Linked Data Encourages Innovation and Experimentation while Recognizing Varied Personal and Institutional Realities

Our personal or institutional capacity to implement complex technologies does not reflect the value of our collections or expertise. Supporting linked data innovation and experimentation at all levels of capacity leads to description that represents more priorities and perspectives than just those of the best-funded institutions. Project Passage, as described in the review of linked data approaches, is an excellent example of how a well-resourced entity can provide the technological support and assistance that allows others to focus on bringing their descriptive expertise to the project. By collaborating toward the creation of "good enough" linked data, we expand representation of the many workers, repositories, communities, and collections across the field. The goal must be finding a place for everyone to be able to participate and to see a role for themselves in the linked data ecosystem, whether it be by adding entities to linked data systems to document people or places represented in local collections, by developing and maintaining technology platforms that use linked data, or by teaching researchers to use new querying methods for search and discovery.

Pragmatic linked data recognizes that linked data implementations must be sustainable, not just within the archival community at large but also at the level of the individual institution, whether it be commitments to new workflows, systems, or infrastructure. These implementations should rely on

incremental and scalable approaches that respond to educational, financial, and technological barriers.[57] The "right" approach to linked data is locally sustainable, ensuring that archives engage in responsible stewardship of our resources whether they be physical, digital, or intellectual (i.e., data based).[58]

Principle 4: Pragmatic Linked Data Focuses on Use Cases that Contribute to Both the Communities We Serve and the Communities with which We Partner

We should create linked data when it is useful to create linked data. The generation of linked data and the ontologies, vocabularies, and systems needed to support it is an expensive proposition and cannot merely be an end unto itself.[59] Our prioritization of use cases for linked data should respect the communities we serve[60] and our own ability to complete the work, and not simply come from a desire to engage with a new technology or distant promises of metadata workflow efficiencies.

We discover potential impact by engaging in user-centered design processes that are grounded in archival standards and responsive to user needs—assessing our data, listening to those we serve, and engaging with those doing similar work across domains. Such use cases may include, but should not be limited to, discovering archival collections outside our silos; fostering shared descriptive ecosystems; supporting communities that wish to share their knowledge; expanding multilingualism in our interfaces; and surfacing suppressed perspectives.

Principles in Practice

What do pragmatic principles look like in practice? How does one decide which linked data projects or ontologies might prove useful in archival description work and what to do with them? Even a "good enough" approach requires serious consideration and time spent on evaluation. Before using a linked data project or ontology, one must assess such aspects as its mission, stability, and governance, as well as its functions. If the project proceeds, one must then determine where those functions support archival use cases and how to productively use and engage with them. While doing so will not entirely prevent missteps, it allows one to scope the project for achievable outcomes.

In this section, we describe two projects in which we engaged through the Archives and Linked Data Interest Group:[61] 1) extending Schema.org to support description of archival repositories and collections; and 2) developing guidance for archival description in Wikidata. We walk through a high-level assessment of each vocabulary, describe our engagement, and identify ways our approach

followed what would become the pragmatic principles and how this led to positive outcomes. These two projects demonstrate different ways to approach linked data work based on the governance and function of each.

Schema.org: Extension and Infrastructure

The project to extend Schema.org's support for describing archival repositories and collections is an example of how such work may require time, formalized process, and infrastructure changes but can lead to an outcome that supports more than just the participants. Schema.org is a linked data standard intended to support structured data on the Internet in a way that can be parsed by search engines. This section walks through the group's identification of whether and how Schema.org might be useful to archives, our engagement with others doing similar work to improve the vocabulary, and the infrastructure updates that support implementation.

Evaluating Function, Governance, and Utility to Archives

Developed by Google, Yahoo, and Microsoft in 2011, Schema.org is a collaborative effort to improve the Web by encouraging the publication of structured data to describe things on it. Although Schema.org is a linked data vocabulary, it is designed to be embedded within web pages as JSON-LD or within the HTML markup.[62] The Schema.org vocabulary provides a set of core types and properties that cover a number of topics, including people, places, and creative works. There is also an extension designed specifically for bibliographic information (SchemaBibExtend).[63] Such extensions are traditionally done through WC3 community groups and then reviewed by the broader community on GitHub, while the day-to-day operations are overseen by a steering group that consists of representatives from the founding companies and several community partners.[64]

In 2017, the Archives and Linked Data Interest Group began exploring Schema.org as a low-barrier methodology for publishing linked data about archives. At the time, no specific ontology existed for archives, and we wanted to find a solution that did not require archival repositories to completely rework their data models and systems. We also hoped that using the language of search engines could help expose archival collections to a wider audience. Search engines prefer pages with structured data, so a clear discovery benefit exists for pages containing this type of information.[65]

After a landscape review of other models, including the Linking Lives project, Bibframe Lite, and Europeana, the group decided to focus on Schema.org. We chose this path for several reasons: its simplicity and suitability in providing a basic representation of entities identified in archival description; the desire to

contribute domain expertise to the W3C Schema Architypes Community Group, a separate group that had begun exploring a Schema.org extension for archives in 2015;[66] and the potential for creating Schema.org markup directly from archival management systems and embedding it in discovery layers. Ultimately, we hoped to demonstrate the "potential of Schema.org as a minimally viable mechanism for publishing linked data about archives."[67] As a lightweight solution to enhancing discovery of archival collections, the project offered an opportunity to examine how well archival data could fit into existing non-library-centric ontologies and whether a simple representation of an archival collection could adequately represent and facilitate access to its contents.[68]

Mapping Fields

Although the project was designed to identify a simple, low-barrier approach to creating linked data for search, it was grounded in archival standards. In addition to exploring how well archives could map to Schema.org, we also sought to determine whether the descriptive elements deemed critical by our national and international standards provided useful and sufficient structured data to support search and discovery of archival collections on the Web. The group began by undertaking a simple mapping exercise to evaluate how well archival metadata could "fit" into the Schema.org ontology: creating a spreadsheet with individual sheets for Agents, Collections, and Repositories in which elements from ISAAR-CPF/DACS, ISAD(G)/DACS, and ISDIAH/DACS respectively were paired with properties available in Schema.org. Additional columns contained information about the elements' recommended use and space for general notes. We also mapped each element to any existing Schema.org properties/classes, as well as database fields in the two most commonly used software programs for archival description: ArchivesSpace and AtoM.[69]

The process of mapping existing descriptive standards and application-specific data models to Schema.org was mostly straightforward. We found that much of the description control information for archival collections was not able to be mapped to existing Schema.org properties, but we felt it was of little direct use for the types of searching Schema.org supports, given the focus on publication information about the collections and materials rather than metadata about the descriptions themselves. We also were not able to satisfactorily map information about the level of description, although Schema does provide a mechanism for expressing hierarchical relationships using the hasPart and isPartOf properties. Finally, certain kinds of information expressed in textual notes, such as information about appraisal, accruals, arrangement, physical characteristics or technical requirements, and references to originals or copies were not mappable using existing types and properties in Schema.

org, but again, we felt that this type of information was beyond the use cases Schema.org supports, namely search and discovery on the Web.

Engagement and Improvement

In 2019, the W3C Schema Architypes Community Group and the Archives and Linked Data Interest Group collaborated on a final proposal to Schema.org, which was accepted and added to the Schema.org code.[70] The resulting model is simple, adding only two subclasses of creative work and six new properties. Once the proposals were made on GitHub, the process became a collaborative engagement between members of our group and others in the Schema.org community. Members of the community recommended minor changes to improve our proposal.

Although our mappings ultimately showed that Schema.org is not a replacement for archival description, they met the minimum needs to support description, which is useful for search engine indexing/display and also fulfills ICA/DACS minimal descriptive standards. This aligns with the pragmatic principle of striving for "good enough," as well as with Rob Sanderson's paraphrasing of a quote attributed to Einstein, that ontology design should be "as simple as possible and no simpler."[71] The new Schema.org properties supplement existing description, translating archival concepts into the language of the Web and providing a bridge between different data communities.[72]

Outcome

This project demonstrated that archives shouldn't have to rethink their infrastructure or their description to publish linked data. Additionally, the project has resulted in improvements to applications designed to support discovery of archival information suggesting a more sustainable and accessible approach to linked data work than localized projects. One of the motivations for our pragmatic approach was the difficulty of migrating to a new metadata standard or profile. Even for a comparatively small change such as this, an individualized approach to adoption of the new Schema.org properties would require each archives to configure its database, XSLT, or other mechanism for publishing finding aids. Focusing on existing systems and the institutional realities of the average archivist (aligning with Principles 2 and 3), team member Mark Custer of Yale updated data mappings in the widely used ArchivesSpace platform to add these new classes and properties. Once institutions updated to version 2.7, their public user interface (PUI) came with embedded Schema.org data for all their publicly described collections and, consequently, repositories using the PUI are publishing linked data about their collections without needing to do additional

work. Another team member, Bruce Washburn of OCLC, added a Schema.org-based descriptive profile to ArchiveGrid,[73] the OCLC research project aggregating over five million records describing archival materials. Integrations at such scale and across the many institutions using ArchivesSpace encourage search engines to make functional use of the data.

The other major outcome of the project was the development of group members' experience and knowledge in data modeling and ontology development (Pragmatic Principle 3). The group's membership included a range of skill sets, from those who had built other ontologies and systems to those just getting started in linked data. The collaborative nature of the mapping exercise, familiar to many archivists who work with digital repositories and metadata, provided a low-barrier, hands-on opportunity to learn to read and work with RDF-based metadata.

Wikidata: Lowering Community Barriers

We began our exploration of Wikidata's potential for archivally informed linked data with the same open exploration of its potential use as we did with Schema.org. Wikidata is a public domain-licensed knowledge base using structured, linked data to describe and connect things. It already contains many linked data records about entities related to the collections we hold. In this section, we describe our assessment of Wikidata and why we decided to create documentation and tools for describing archival collections, archival repositories, and the people and organizations related to them.

Evaluating Function, Governance, and Utility to Archives

Wikidata came online in 2012 and has been active and expanding ever since. It began as a place to describe the people, places, events, things, concepts, and more found in the many Wikimedia Foundation sites. It now extends to include data about local businesses, scholarly articles, archival collections, data sets, and much more (a total of over 92 million entities), making it an excellent aggregator of information.[74] Its guidance for notability includes the kinds of people, organizations, and things that might be found in our collections.[75] It also allows citation of primary sources, a critical concern for those creating description using archival materials. As with other wiki-type tools, it can be edited and queried by anyone without creating an account, installing special tools, or hosting software.[76] Wikidata's strength is its community of engaged users who create and update records and propose new fields to support description.

The Wikimedia Foundation hosts and maintains the Wikidata infrastructure. Its data have long been included in VIAF (the OCLC international linked

data authority project), and the Library of Congress recently began adding Wikidata URIs for people to the linked data versions of their authority records. The SNAC Cooperative does not yet appear to incorporate Wikidata but does include its URLs in its list of related descriptions. Its data are also reused in search engine results by such large entities as Google, and tools exist that allow anyone to query for reuse.[77] The support from, integration with, and reuse by such major entities led us to assess it as sufficiently stable to merit engagement.

Because Wikidata is a centralized database, we could assess both the data standard and the data simultaneously. We found significant work had been put into describing people and organizations, including the many relationships between them. Although such relationships have long been considered an important part of archival description, the time and technological burdens of creating and using structured EAC-CPF have led to low rates of the standard's adoption. Even when we create such records, they are rarely interoperable across archival institutions, let alone outside of them. Wikidata, on the other hand, provides a wealth of such data that anyone can adopt and reuse in whole or in part.[78] Additionally, Wikidata functions as a reconciliation service for entities, storing identifiers from national, international, and local systems into a single database.

Like other Wikimedia Foundation projects, Wikidata is community driven and encourages communities with expertise to collaborate on improving both its data and the descriptive options for creating that data. The Archival Description WikiProject, for example, focused on improving description of archival collections within Wikidata.[79] Members of this project had been responsible for the creation of such key properties as "archives at," which allows archivists to link a person or organization to the repository where their archives are held. Creating new properties and classes is fairly simple in Wikidata, providing a low barrier to entry and allowing those with domain expertise to participate even without an ontology/linked data skillset.[80] Wikidata's governance is a combination of user consensus and elected administrators who address major disputes and review possible bad-faith actions.[81] The Wikidata community also supports tool development. Its Tools page[82] includes those that support the transformation and batch ingest of data from external sources such as our own databases. Engagement with Wikidata, therefore, appeared to be an excellent opportunity to benefit from others' descriptive work and to augment existing descriptions with information found in our collections.

Mapping Fields

To determine how we might engage, we began mapping Wikidata properties to descriptive elements. We used a copy of the spreadsheet with content

elements described in the Schema.org section as a starting point and mapped agent, repository, and collection elements to Wikidata properties, identifying any gaps. Because much work we needed for archival description had been done as described, we reflected on what other challenges might exist and how we could support other archivists interested in using Wikidata.

Because Wikidata is community based and interactive, many members of our group experimented and connected during this time with other people and communities doing similar Wikidata work. Some joined the LD4 Wikidata Affinity Group,[83] where they learned about tools to support batch Wikidata description and practices related to describing persons and organizations in Wikidata. Gloria Gonzalez, who works for the linked data infrastructure provider Zepheira, used her many connections with the WikiCite/Wikidata community to invite guests who could address specific questions raised by the group. Several others engaged with and shared the work of the Archival Description WikiProject. The group invited Dominic Byrd-McDevitt, the National Archives and Records Administration (NARA) Wikpedian at the time and a member of the Archival Description WikiProject, to share his insights from creating Wikidata records for NARA collections. His expertise greatly assisted with the conceptual mapping we have described.

One of the larger challenges we encountered during our research was determining the specific meaning and use of certain properties and fields. While some were self-explanatory, we found others to be less than transparent, requiring significant investigation. For example, start- and end-time qualifiers on a statement apply to when the statement was true and should not be used for other time ranges, such as an institution's dates of creation/dissolution. Sometimes our exploration resulted in using more generic properties for archival concepts. Other times, we decided that certain concepts, such as the unique record identifier, were met by functional elements of Wikidata's software.

Determining Productive Engagement

We determined that the most productive work our group could contribute would be creating documentation and tools. By doing this, we aimed to lower barriers for others getting started with Wikidata contributions. We wrote four guidelines describing people, corporate bodies (organizations), repositories, and collections. We also created standards-based templates for the Wikidata tool "Cradle,"[84] an interface for quick, template-driven record creation.

Each guideline consists of two sections. The first section lists recommended or required Wikidata properties specific to archival description.[85] These are mapped to the appropriate DACS/ICA elements along with required or recommended Wikidata properties, such as "label" and "instance of" and several

properties created by the Archival Description WikiProject that do not directly map to a standard but are archivally specific and greatly support description. Each row includes information about the expected or required value[86] and instructions that provide further context and guidance for using the property. The second section of each guideline consists of examples of how each property is used in Wikidata to describe real persons, organizations, repositories, or collections. By using real examples, we intend to encourage readers' exploration of Wikidata entities similar to those they're describing.

Outcome

Many library- and archives-based projects use Wikidata to create and publish linked data. This includes the already-mentioned Carnegie Hall Data Lab; the Canadian Archive of Women in STEM, designed to bring attention to and promote the discovery of the archival records of women in science, technology, engineering, and math held by Canadian institutions; and York University's use of Wikidata to describe Indigenous communities.[87] The motivations and use cases for doing this work include authority control, multilingual support, or discovery of "hidden" resources in special collections. Even simple, uncomplicated implementations of Wikidata may yield unexpected results. In 2020, Emory University linked all of its archival collections to Wikidata by adding the "archives at" property to the appropriate records, resulting in a significant bump in referral traffic and improved collection discovery.

Both our own and other projects that support the use of Wikidata to create description are excellent examples of the kind of work that fulfills all four pragmatic principles. The recommendations we published built from existing archival data standards and could immediately be used on Wikidata (Principle 2). Low barriers to getting started on the site and its strong infrastructure allow anyone to participate, whether they have time to become deeply involved or never create an account and only occasionally update data (Principle 3). This encourages experimentation from more than the usual suspects. We benefit significantly from the number of corporations, persons, and (sometimes) families already described by the Wikidata community and can improve that data with our specialized knowledge or unique data sources (Principle 4).

And, finally, in fulfilling the first Principle, by accepting "good enough" description, Wikidata demonstrates one of the strengths of the pragmatic Principles. In creating or updating Wikidata records, we need not aspire to having a full conception of the person, group, or thing we are describing. This frees us up to add minor statements to flesh out existing records or to create "good enough" records that at least ensure representation in a shared

database. It allows us to benefit from others' specialized knowledge combined with our own.

Conclusion

Engaging in linked data does not have to require time- and resource-intensive projects or systems that don't yet exist. Our work should prioritize individual/institutional needs and users, whether by enhancing discovery, improving and reusing data in shared knowledge ecosystems, advocating for ethical data practices, or simply encouraging learning among archivists. By focusing less on transforming our entire descriptive framework and more on the simple, pragmatic approaches we have described, we create space for everyone to do linked data.

We've seen encouraging developments toward such approaches. The PCC Wikidata Pilot's choice to open its project to those outside the better-resourced PCC institutions and its ability to do so show how pragmatic approaches expand participation. Similarly, Project Passage's participants benefited from the work already done on the user-friendly Wikibase system and the import of Wikidata for reuse. The new LD4 community aspires to connect individuals, institutions, nonprofit organizations, government repositories, and commercial entities. "Together," its mission statement declares, "we explore, learn and collaborate to raise awareness and know-how, encourage adoption, and foster an ecosystem of interoperable standards, tools and services."[88]

We still face concerns over the lack of infrastructure to use linked data and the sustainability of project-based work, the tension between choosing the best project and just getting started,[89] and the current demographic makeup of contributors to linked data projects. We recognize that technological barriers may render the ideal of an equitable community of contributors to data structures unattainable. We intend our principles to be a series of guideposts and a starting place for our community to tackle these questions, not by creating a new standard but by working together and learning from each other.

Linked data is not a singular problem to be solved or an application to be built. Its strength lies in decentering ourselves and collaborating with others. Its ethos rejects strict adherence to monolithic standards in favor of flexible solutions. While such flexibility opens the door to inconsistent data or malicious falsehoods, it also offers the opportunity to connect the data we have with the data known and prioritized by other communities. Such communities might be in related fields, such as museum professionals working to improve data about an artist whose records we hold. They might be from a community project working to increase representation of BIPOC authors in Wikispaces. Without the mindset that we have something to offer and we can learn from others, linked

data becomes yet another encoding standard. If we let go of our need for control and embrace a more pragmatic approach, we will learn from and contribute to a world of data much bigger than our own.[90]

Notes

1 These include the LD4P and LD4L suite of Mellon grants, https://wiki.lyrasis.org/display/LD4P2.

2 For example, Sinopia, which was developed through Linked Data for Production: Pathway to Implementation (LD4P2), https://sinopia.io.

3 Niklas Lindström, "National Library Platform Based On BIBFRAME" (presentation at Third Annual BIBFRAME workshop in Europe, Stockholm, Sweden, September 17, 2019).

4 Including Michigan State, Cornell, Duke, and Texas A&M.

5 FOLIOProject, "Codex Metadata Model," https://wiki.folio.org/pages/viewpage.action?pageId=1415393, captured at https://perma.cc/6K6G-6TUW.

6 See the PCC Wikidata Pilot project page, https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot, captured at https://perma.cc/4ZHN-6KEN.

7 Terry Cook, "Evidence, Memory, Identity, and Community: Four Shifting Archival Paradigms," *Archival Science* 13 (2013): 95–120, https://doi.org/10.1007/s10502-012-9180-7.

8 Virginia L. Ferris, "Beyond 'Showing What We Have': Exploring Linked Data for Archival Description" (master's thesis, University of North Carolina at Chapel Hill, August 2014), 39.

9 Society of American Archivists' Technical Subcommittee, "Statement of Principles," in *Describing Archives: A Content Standard* (Chicago: Society of American Archivists, 2019), https://files.archivists.org/pubs/DACS_2019.0.3_Version.pdf, captured at https://perma.cc/6WHH-RQG7.

10 Society of American Archivists' Technical Subcommittee, "Statement of Principles."

11 For those seeking a clear overview of the concepts behind linked data and its potential, we recommend the first two chapters of Carlson et al.'s *Linked Data for the Perplexed Librarian* (Chicago: ALA Editions, 2020).

12 Tim Berners-Lee, "Linked Data Design Issues," last updated June 18, 2009, https://www.w3.org/DesignIssues/LinkedData.html, captured at https://perma.cc/585E-WXW3.

13 First, URIs represent things. Second, those URIs are HTTP based so that they can be looked up online. Third, when someone looks up such a URI, they can get useful, standards-based information about the thing it represents. Although not required, Berners-Lee strongly advises making that information available through a SPARQL service so that users could perform more complicated queries. Fourth, the information users retrieve contains URIs that provide more information.

14 Paul Wilton, "Wikidata - Q41483," April 3, 2018, https://datalanguage.com/news/wikidata-q41483, captured at https://perma.cc/QZT8-U53L.

15 RDF, the Relational Data Framework, is the standard data model for expressing linked data (much as XML is the way one writes EAD). Most RDF statements are "triples," consisting of a subject (the thing being described), a predicate (the descriptive property/relationship/"verb" being used to describe it), and the object (the value of the statement). The first two should be HTTP URIs. The third should be an HTTP URI when possible but may also be a value with a controlled datatype (such as an xsd:date) or, when nothing else is appropriate, free text (a "string").

16 A "linked data ontology" is a representation of concepts and relationships in a particular area. The term is often used interchangeably with "vocabulary." The W3C's "Ontologies" definition page characterizes the difference in use as one that indicates the level of complexity and constraint, with "ontology" being the more formal of the two, https://www.w3.org/standards/semanticweb/ontology.html, captured at https://perma.cc/XW5N-7XMG. Although multiple ontologies may be used together, one should take care to avoid what Corcho et al. describe as "Frankenstein ontologies," in which terms from ontologies are integrated into other ontologies without sufficient research into whether their areas of focus are appropriate. Oscar Corcho,

María Poveda-Villalón, and Asunción Gómez-Pérez, "Ontology Engineering in the Era of Linked Data," *Bulletin of the Association for Information Science and Technology*, 41, no.4 (2015): 15. A simple example would be "title." In the Dublin Core vocabulary, it means "A name given to the resource." In the Friend of a Friend vocabulary, it means an honorific with the examples "Mr., Mrs, Ms, Dr. etc."

[17] This query can currently be performed in Wikidata, although its results only include records that contain "archives at" statements.

[18] More extensive overviews of linked data in archives can be found in Karen F. Gracy, "Archival Description and Linked Data: A Preliminary Study of Opportunities and Implementation Challenges," *Archival Science*, 15 no.3 (2015): 239–94, http://doi.org/10.1007/s10502-014-9216-2; and Jinfang Niu, "Linked Data for Archives," *Archivaria*, 83 (2016): 83–110, https://archivaria.ca/index.php/archivaria/article/view/13582.

[19] Data from these projects was last updated in 2013 but is still available for use at http://data.archiveshub.ac.uk.

[20] Jane Stevenson, "Linking Lives Evaluation Report," *Linking Lives* (blog), November 15, 2012, http://linkinglives.archiveshub.ac.uk/2012/11/15/linking-lives-evaluation-report. Data from these projects were last updated in 2013 but are still available for use at http://data.archiveshub.ac.uk.

[21] In 2016, the EAD3 Study Group on Discovery, a working group of the SAA EAD Roundtable, released "Implementing EAD3: Search and Exploration," a report that outlines many strategies for incorporating linked data vocabularies and concepts in EAD, https://www2.archivists.org/sites/all/files/EAD3_Study_Group_on_Discovery_Recommendations_20160719.pdf, captured at https://perma.cc/N84F-4AEF.

[22] Belfast Group Poetry, https://belfastgroup.digitalscholarship.emory.edu.

[23] Linked Jazz, https://linkedjazz.org.

[24] Carnegie Hall Data Lab, https://carnegiehall.github.io/datalab.

[25] An early example of such an adaptation is Tufts's proposed experimentation with wholesale reuse of the Open Archives Initiative Object Reuse and Exchange (OAI-ORE), a standard created to describe aggregations of web resources, "Archival Description in OAI-ORE," *Journal of Digital Information* 12, no. 2 (2011), https://journals.tdl.org/jodi/index.php/jodi/article/view/1814.

[26] Jean Godby et al., *Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage* (Dublin, OH: OCLC Research, 2019), 10, 12–13, https://doi.org/10.25333/faq3-ax08.

[27] Godby et al., *Creating Library Linked Data with Wikibase*, 23–25.

[28] Michelle Durocher, "Key Stats and Themes about Pilot Participants and Their Projects," August 27, 2020, https://wiki.lyrasis.org/display/pccidmgt/Wikidata+Pilot+Kick-off+meetings?preview=/187174063/194117730/Wikidata%20Pilot%20meeting%20--%20Key%20stats%20and%20themes.mp4#.

[29] Gregory Wiedeman, "The Historical Hazards of Finding Aids," *American Archivist* 82, no. 2 (2019): 6, https://doi.org/10.17723/aarc-82-02-20.

[30] Wiedeman, "The Historical Hazards of Finding Aids," 20. Also Ciaran B. Trace and Andrew Dillon, "The Evolution of the Finding Aid in the United States: From Physical to Digital Document Genre," *Archival Science* 12, no. 4 (2012): 506, https://doi.org/10.1007/s10502-012-9190-5, as cited by Wiedeman, 20.

[31] Wiedeman, "The Historical Hazards of Finding Aids," 2.

[32] A guide to appropriate combination of vocabularies/ontologies is beyond the scope of this article. Later sections on evaluating function in Schema.org and Wikidata provide examples of how one should consider the domain, functions, and purpose of an ontology before reuse.

[33] Michelle Light and Tom Hyry, "Colophons and Annotations: New Directions for the Finding Aid," *American Archivist* 65, no.2 (2002): 216–30, https://doi.org/10.17723/aarc.65.2.l3h27j5x8716586q.

[34] SNAC, "O'Connor, Flannery, 1925–1964," https://snaccooperative.org/view/83773688, captured at https://perma.cc/Q8JS-KR5N.

[35] "SAA Core Values Statement and Code of Ethics," Society of American Archivists, last revised August 2020, https://www2.archivists.org/statements/saa-core-values-statement-and-code-of-ethics, captured at https://perma.cc/63SX-YWKB.

[36] Homosaurus, https://homosaurus.org.

[37] While linked data holds the potential to bring in more voices, it is important to note that it is not an automatic path to diversifying our description. In their 2019 keynote, "Questioning Wikidata," at WikidataCon 2019, race, gender, and power researcher Os Keyes examines how the same power dynamics that marginalize communities in the first place replicate themselves in linked data spheres and may block or discourage members from engaging. They also note the importance of recognizing that the membership in a marginalized community does not inherently lead to consensus on how to describe that community's identity or experiences, https://www.wikidata.org/wiki/Wikidata:WikidataCon_2019/Program/Sessions/Keynote:_Questioning_Wikidata, captured at https://perma.cc/2J2Z-K5FP.

[38] Julie Hardesty, "Transitioning from XML to RDF: Considerations for an Effective Move Towards Linked Data and the Semantic Web," *Information Technology & Libraries* 35, no 1 (2016): 59, https://doi.org/10.6017/ital.v35i1.9182.

[39] An adoption rate of 90 percent was reported in a 2018 survey conducted by the SAA Technical Subcommittee on Encoded Archival Standards (TS-EAS), Wim van Dongen and Kathy Wisser, *EAD3 Implementation Survey Results and Discussion* (2018), https://www.loc.gov/ead/EAD3_Implementation_Survey_Results_and_Discussion_20190320.pdf, captured at https://perma.cc/AE8Z-AEFE. However, this survey represents a self-selecting sample of institutions, and a comprehensive study including smaller repositories would likely show a much smaller rate.

[40] It is not a coincidence that the Mellon-funded LD4 suite of projects experimenting with BIBFRAME implementations is centered at well-resourced institutions such as Columbia, Cornell, Harvard, Princeton, and Stanford Universities.

[41] Carlson et al., *Perplexed Librarian*, 8.

[42] Kate Theimer, "What Is the Meaning of Archives 2.0?," *American Archivist* 74, no. 1 (2011): 58–68, https://doi.org/10.17723/aarc.74.1.h7tn4m4027407666; Mark Greene, "MPLP: It's Not Just for Processing Anymore," *American Archivist* 73, no. 1 (2010): 175–203, https://doi.org/10.17723/aarc.73.1.m577353w31675348; Daniel A. Santamaria, *Extensible Processing for Archives and Special Collections: Reducing Processing Backlogs* (Chicago: American Library Association, 2015).

[43] We take inspiration from the LOUD Principles, "LOUD: Linked Open Usable Data," Linked Art, https://linked.art/loud/index.html, captured at https://perma.cc/5P2D-HKEN, and the Oregon Digital metadata principles as described in Julia Simic and Sarah Seymour, "From Silos to Opaquenamespace: Oregon Digital's Migration to Linked Open Data in Hydra," *Art Documentation: Bulletin of the Art Libraries Society of North America* 35, no.2 (2016): 4, https://doi.org/10.1086/688730. LOUD principles: "The right Abstraction for the audience," "Few Barriers to entry," "Comprehensible by introspection," "Documentation with working examples," and "Few Exceptions, instead many consistent patterns." Oregon Digital metadata principles: "You are not that special"; "You do not know everything"; "If your data isn't reusable, shareable, and machine-readable, then you're not doing good enough work"; and "Use exemplary behavior and reuse from others so that they may also reuse from you."

[44] OCLC Research Archives and Special Collections Linked Data Review Group, *Archives and Special Collections Linked Data: Navigating between Notes and Nodes* (Dublin, OH: OCLC Research, 2020), 7, https://doi.org/10.25333/4gtz-zd88.

[45] Building from DACS Principle 1: "Archival description expresses professional ethics and values," we also recognize that sometimes the ethics, values, and standards of our profession fall short and that this is on us to fix. Society of American Archivists' Technical Subcommittee, "Statement of Principles."

[46] What about EAD/EAC-CPF and the like? These encoding standards provide one method of encoding the elements outlined in the content standards. They serve a particular function: to encode an indexable finding aid or archival authority record. One may find it practical to create mappings from these standards into linked data as a means of reusing one's data. We recommend, however, that, when performing such work, one use content standards as the basis to allow for re-envisioning how elements might be expressed.

[47] "Ethics in Linked Data Affinity Group," https://wiki.lyrasis.org/display/LD4P2/Ethics+in+Linked+Data+Affinity+Group, captured at https://perma.cc/F9GB-5QTR.

[48] Kelly J. Thompson, "More Than a Name: A Content Analysis of Name Authority Records for Authors Who Self-Identify as Trans," *Library Resources & Technical Services*, 60 no. 3 (2016): 141–42, https://doi.org/10.5860/lrts.60n3.140; Ruth K. Tillman, "Barriers to Ethical Name Modeling in Current Linked Data Encoding Practices," in *Ethical Questions in Name Authority Control*, ed. Jane Sandberg (Sacramento: Library Juice Press, 2019), 241–57.

[49] Jennifer Douglas, "Toward More Honest Description," *American Archivist* 79 no. 1 (2016): 26–55, https://doi.org/10.17723/0360-9081.79.1.26.

[50] Michelle Caswell and Marika Cifor, "From Human Rights to Feminist Ethics: Radical Empathy in the Archives," *Archivaria* 81 (2016): 23–43, https://archivaria.ca/index.php/archivaria/article/view/13557.

[51] Ed Summers, "Designing for Consent," *Documenting DocNow* (blog), May 20, 2019, https://news.docnow.io/designing-for-consent-2f9e9cb2ab4f, captured at https://perma.cc/DG6L-EDDU.

[52] Stacy Wood et al., "Mobilizing Records: Re-Framing Archival Description to Support Human Rights," *Archival Science* 14, no. 3 (2014): 397–419, http://dx.doi.org/10.1007/s10502-014-9233-1.

[53] The University of Alberta Libraries have made a series of presentations on their work, including Sharon Farnel and Sheila Laroque, "Decolonizing Description at the University of Alberta Libraries" (webinar, presented as part of SLA Western Chapter webinar series, April 24, 2018), https://doi.org/10.7939/R3FQ9QM1S.

[54] Local Contexts, "TK Labels," https://localcontexts.org/labels/traditional-knowledge-labels.

[55] Amber Billey, "Just Because We Can, Doesn't Mean We Should: An Argument for Simplicity and Data Privacy with Name Authority Work in the Linked Data Environment," *Journal of Library Metadata* 19 nos. 1–2 (2019): 1–17, https://doi.org/10.1080/19386389.2019.1589684.

[56] Society of American Archivists' Technical Subcommittee, "Statement of Principles."

[57] For a fuller discussion of the barriers to linked data implementations in archives, see OCLC Research Archives and Special Collections Linked Data Review Group, *Archives and Special Collections Linked Data*.

[58] See Society of American Archivists, "SAA Core Values Statement and Code of Ethics." This may mean, for instance, that some institutions encode their data within complex ontologies such as BIBFRAME or RiC-O, while others transform record-based metadata in EAD or MARC or use another platform such as Wikidata/Wikibase. Some may mint entities locally, while others depend upon external name authority systems.

[59] Kyle Banerjee, "The Linked Data Myth," *Library Journal*, August 13, 2020, https://www.libraryjournal.com/?detailStory=the-linked-data-myth, captured at https://perma.cc/9D6G-UW87.

[60] DACS Principle 2: "Users are the fundamental reason for archival description." Society of American Archivists' Technical Subcommittee, "Statement of Principles."

[61] The Archives and Linked Data Interest Group was an independent, unaffiliated group of archivists initiated by Elizabeth Russey Roke in 2017. It was formed to generate ideas, build skill sets in linked data among archivists, work through linked data implementation problems, and collaborate on shared solutions. Much of its work explored the alignment of existing linked data vocabularies and projects with standards for archival description.

[62] Schema.org, "Getting Started with Schema.org Using Microdata," last updated December 4, 2020, https://schema.org/docs/gs.html, captured at https://perma.cc/65X2-GG7N.

[63] Through the efforts of Zepheira (now a part of EBSCO), the adoption of Schema.org structured data for bibliographic materials has led to the integration of library catalogs into Google's Knowledge Graph and the ability to discover (and borrow) materials directly in a Google search without needing to use a local library OPAC.

[64] Schema.org, "About Schema.org," https://schema.org/docs/about.html, captured at https://perma.cc/X8S4-F9HZ.

[65] Google's developer documentation explicitly points out that web pages using structured data markup like Schema.org will enable "special search results and enhancements." Google Search Central, "Understand How Structured Data Works," last updated February 16, 2021, https://

developers.google.com/search/docs/guides/intro-structured-data, captured at https://perma.cc/S7JB-GR3L.

[66] Schema.org, "Schema Architypes Community Group," https://www.w3.org/community/architypes, captured at https://perma.cc/3R4V-8LQG.

[67] Mark Matienzo, Elizabeth Russey Roke, and Scott Carlson, "Creating a Linked Data-Friendly Metadata Application Profile for Archival Description" (poster, International Conference on Dublin Core and Metadata Applications, Washington, DC, December 12, 2017), https://dcevents.dublincore.org/IntConf/dc-2017/paper/view/506.html.

[68] It is worth noting that Schema.org is controlled by Google and other commercial search engines so is not a neutral entity. As cultural heritage organizations begin to share their data using the language of the (commercial) Web, we must remain vigilant to these biases and protect against harm. For one critique of Schema.org, see Michael Andrews, "Who Benefits from Schema.org?," *Story Needle* (blog), August 23, 2020, https://storyneedle.com/who-benefits-from-schema-org, captured at https://perma.cc/H742-J6AC.

[69] A visualization of these mappings is available in Matienzo et al., "Creating a Linked Data-Friendly Metadata Application Profile for Archival Description," https://dcevents.dublincore.org/IntConf/dc-2017/paper/view/506/632.html

[70] Richard Wallis, "Archives and Their Collections" (GitHub issue opened September 28, 2017, and closed April 10, 2019), https://github.com/schemaorg/schemaorg/issues/1758.

[71] Robert Sanderson, "It's 2020 . . . Where Is My Flying Car and Cultural Heritage Research Data Ecosystem?" (keynote, Coalition for Networked Information, online, March 30, 2020), https://www.cni.org/events/membership-meetings/past-meetings/spring-2020/plenary-sessions-s20#opening. This principle has become part of the IIIF Design Principles to which he contributed, International Image Interoperability Framework, https://iiif.io/api/annex/notes/design_patterns/#select-solutions-that-are-as-simple-as-possible-and-no-simpler, captured at https://perma.cc/6SY2-MV2L.

[72] This project demonstrated ways that the cultural heritage community can improve and contribute to wider structured data initiatives to improve discovery on the web. Some of the newly proposed vocabulary terms, such as the property for conditions of access, were quickly identified as reusable by those outside the field (Pragmatic Principle 4). See https://github.com/schemaorg/schemaorg/issues/2173.

[73] ArchiveGrid, "About ArchiveGrid," OCLC, https://researchworks.oclc.org/archivegrid/about, captured at https://perma.cc/AX2X-E3GR.

[74] Wikidata, "Wikidata:Introduction," https://www.wikidata.org/wiki/Wikidata:Introduction, captured at https://perma.cc/3HWH-WHM3.

[75] Wikidata, "Wikidata:Notability," https://www.wikidata.org/wiki/Wikidata:Notability, captured at https://perma.cc/6RWN-29J5.

[76] It is worth remembering that Wikidata functions much as other wiki tools, sharing the benefits and downsides of a decentralized, constantly changing environment. While some safeguards are in place (such as recording the IP address of anyone creating or updating content without an account to identify and ban bad-faith editors) inaccurate, misleading, or malicious data are an unavoidable reality in this model.

[77] On December 10, 2020, for instance, the Wikidata SPARQL endpoint was queried 6.85 million times. "Wikidata Query Service Usage Dashboard," Discovery Dashboards, https://discovery.wmflabs.org/wdqs/#wdqs_usage.

[78] The work done in SNAC has created one of the largest CPF-type bodies of records, which are also licensed for reuse. However, its use remains niche within the archival community, and it cannot be directly edited.

[79] Wikidata, "Wikidata:WikiProject Archival Description," https://www.wikidata.org/wiki/Wikidata:WikiProject_Archival_Description, captured at https://perma.cc/E76D-VB6J.

[80] While the openness of Wikidata encourages all to participate, the database that results is difficult to query in its entirety, with inconsistent, ever-changing data models and millions of properties available for use. The proliferation of properties and classes (in Wikidata and domain-centric ontologies) is an issue that the linked data community will need to contend

with broadly if we want to share metadata between domains. However, we believe that Wikidata is a step in the right direction that rejects gatekeeping and diversifies the linked data community beyond a small group of experts.

[81] Wikidata, "Wikidata: Administrators," https://www.wikidata.org/wiki/Wikidata:Administrators, captured at https://perma.cc/4L3C-7HU5.

[82] Wikidata, "Wikidata:Tools," https://www.wikidata.org/wiki/Wikidata:Tools, captured at https://perma.cc/8Z7N-HWKF.

[83] Wikidata, Wididata:WikiProject LD4 Wikidata Affinity Group, https://www.wikidata.org/wiki/Wikidata:WikiProject_LD4_Wikidata_Affinity_Group, captured at https://perma.cc/BQ4C-KVUF.

[84] Cradle templates are particularly suited for beginners and those who wish to ensure that the records they create use a specific set of fields. Because they do not allow users to add the qualifying statements and references necessary to most records, their use must be supplemented with additional edits in the Wikidata interface or using another tool.

[85] The introduction to our guidelines for describing persons and organizations also links to the Wikidata "list of properties" page for properties about people and organizations. This list includes many properties that may be of use in creating fuller descriptions but do not directly map to a DACS/ISAAR-CPF field or describe a relation/related entity in such a niche way that they might apply to only one person in an entire repository. They are as generic as "manner of death" or as specific as "astronaut mission."

[86] Some may be a textual string, others must be a certain kind of Wikidata entity, and still others must come from a controlled list of possible values.

[87] Karen Smith-Yoshimura, "Experimentations with Wikidata/Wikibase," *Hanging Together: The OCLC Research Blog*, June 18, 2020, https://hangingtogether.org/?p=8002, captured at https://perma.cc/A4NV-KBYW.

[88] LD4 Community Site, https://sites.google.com/stanford.edu/ld4-community-site/home.

[89] For those interested in working with Wikidata, one place to get started is the "50 Things" list created by the Wikidata/Archives exploration group for the Lighting the Way project, https://www.wikidata.org/wiki/Wikidata:WikiProject_Archives_Linked_Data_Interest_Group/50_Things, captured at https://perma.cc/3YNU-XMMP.

[90] We offer special thanks to the members of the Archives and Linked Data (2017–2019) interest group (https://archival.github.io) for their contributions to these projects and to Jennie Knies, Mark Matienzo, and Rob Sanderson for reading and providing comments on drafts of this article.

ABOUT THE AUTHORS

Elizabeth Russey Roke is the Discovery and Metadata Archivist at the Stuart A. Rose Library of Emory University. Primarily focused on description, discovery, and access to special collections materials, she works on a variety of technology projects and initiatives related to metadata standards, linked data, archival descriptive practice, and discovery. She currently is a member of the ArchivesSpace Technical Advisory Council, SAA's Technical Subcommittee on Encoded Archival Standards (TS-EAS), and the LD4 Community Steering Committee. She is also an adjunct instructor at Dominican University. Roke holds a dual MA/MLS from the University of Maryland.

Ruth Kitchin Tillman is the Sally W. Kalin Librarian for Technological Innovation at the Penn State University Libraries. She has written and presented on metadata encoding standards, library discovery, linked data, institutional repositories, and labor. She coleads her library's Program for Cooperative Cataloging Wikidata Pilot Group and has contributed to the Art and Rare Materials BIBFRAME Extension, NISO's 2020 Linked Data Focus Group, and to other linked data and metadata efforts. Tillman holds an MLS from the University of Maryland.