Linked Data for Archives

JINFANG NIU

RÉSUMÉ En se basant sur l'examen d'un nombre de projets qui ont mis en œuvre les données liées pour le matériel d'archives, cet article présente l'état actuel des données liées archivistiques et explore leur impact sur la description et sur la découverte de l'information archivistique. L'auteur soutient que malgré le fait que la communauté archivistique est toujours aux tout débuts de la mise en œuvre des données liées, son utilisation fait preuve d'un potentiel important dans l'amélioration de la description archivistique et la découverte de l'information. Plus spécifiquement, l'auteur affirme que les données liées enrichiront la description archivistique en la rendant davantage interopérable et granulaire, faisant en sorte que la découverte de l'information archivistique deviendra plus puissante par sa capacité de répondre directement aux questions des utilisateurs.

ABSTRACT Based on the investigation of a number of projects that have implemented linked data for archival materials, this article reports on the current status of archival linked data and discusses the impact of linked data on archival description and archival information discovery. The author argues that although the archival community is still in the early stage of linked data implementation, the usage of linked data demonstrates great potential for improving archival description and information discovery. Specifically, the author argues that linked data will enrich archival description by making it increasingly interoperable and granular and make archival information discovery more powerful through the ability to directly answer user questions.

Introduction

Linked data is the foundation of the semantic web. According to Tim Berners-Lee,[1] the four rules for producing linked open data are: 1) "use URIs as names for things"; 2) "use HTTP URIs so that people can look up those names";

1 Tim Berners-Lee, "Linked Data," 18 June 2009, https://www.w3.org/DesignIssues/LinkedData.html.

3) "when someone looks up a URI, provide useful information using standards" such as the Resource Description Framework (RDF)[2] and SPARQL[3]; and 4) "include links to other URIs" so that users "can discover more things." As the network of linked data expands, more structured data encoded in open standard format will be published on the Web. Semantic links, which designate specific relationships using RDF predicates that are more meaningful than hyperlinks,[4] will proliferate on the Web. In addition, links can be created between various kinds of entities, not only online documents. These new features will deeply affect all kinds of resources presented on the Web, including those created and/or curated by the LAM (library, archives, museum) community. Linked open data promises many benefits to LAMs for resource description and information discovery. Its openness makes resource descriptions (metadata) available for use in endless and unexpected ways. Its linked nature allows resource descriptions to be produced in a decentralized way by different institutions and then aggregated in a global graph simply through semantic links.[5] The searching mechanism for linked data, more specifically SPARQL queries, allows users to formulate complex queries that are not possible in traditional search interfaces. In addition, SPARQL queries return direct answers, similar to those provided by Google Knowledge Graph. This might fundamentally change the nature of information discovery in LAMs and suggest a redefinition of the user tasks described in the Functional Requirements for Bibliographic Records (FRBR) model for bibliographic description, which assumes that users search for bibliographic resources rather than direct answers.

2 RDF is a conceptual model for making statements about web resources. Each statement is called a triple because it includes three parts: subject, predicate, and object. Multiple triples (statements) may be connected to each other and form an RDF graph. RDF triples and graphs are conceptual. To make them processable by computers, they need to be written in XML, Turtle, N3, or other formats. For more information about RDF, please refer to the RDF primer: Guus Schreiber and Yves Raimond, eds., "RDF 1.1 Primer: W3C Working Group Note 25 February 2014," https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/.

3 SPARQL is a recursive acronym for SPARQL Protocol and RDF Query Language. It is a query language for searching data in RDF triple stores, similar to the way that the SQL query language is used for searching data in relational databases.

4 A hyperlink between two web pages only shows that the two web pages are connected; it does not tell people in what way those two web pages are connected. In contrast, in an RDF triple, the predicate shows exactly how the subject and the object are connected, as in this triple: Michael Chen (subject) was born in (predicate) Ann Arbor, Michigan (object).

5 In linked data, everything is assigned a URI, which is globally unique. To connect any two things on the web of linked data, the only thing you need to do is create an RDF triple and include those two things as the subject and object respectively. Thus, we can integrate the linked data sets produced by different institutions through adding RDF triples to each data set. This is simpler than harvesting data sets from various institutions and storing them in one database.

The LAM community has started several initiatives to research, experiment with, and implement linked data for resource description and information discovery. LODLAM (Linked Open Data in Libraries, Archives and Museums), a network of enthusiasts, technicians, and professionals interested in applying linked data to LAMs, was established in order to share resources and form partnerships. The Schema Bib Extend Community Group was convened to discuss and prepare proposals for extending Schema.org vocabularies to represent library resources (www.w3.org/community/schemabibex). Major knowledge organizational tools used by the LAM community have been converted into linked data format, such as the Library of Congress Authorities, Library of Congress Classification, Dewey Decimal Classification (DDC), Virtual International Authorities File (VIAF), and Faceted Application of Subject Terminology (FAST). The Library of Congress has created the Bibliographic Framework (bibframe.org), or BIBFRAME, a linked data–based vocabulary for describing library resources. BIBFRAME will replace MARC, which is outdated and has been criticized for many years.

Various books, reports, and scholarly articles have appeared to help LAM professionals learn linked data and discuss technical and theoretical issues related to linked data implementation. For example, Seth Van Hooland and Ruben Verborgh published a practical handbook that teaches LAM professionals how to convert existing metadata and generate new metadata in linked data format.[6] W3C provided a user guide for Simple Knowledge Organization System (SKOS), an important tool for converting controlled vocabularies, classifications, and other knowledge organization tools into linked data format.[7] Ed Summers and Dorothea Salo discussed difficulties for the cultural heritage community in implementing linked data and outlined some pragmatic ways to overcome the difficulties.[8] The W3C Library Linked Data Incubator Group,[9] chartered from May 2010 through August 2011, analyzed the benefits of library linked data, discussed technical and legal issues regarding converting and publishing traditional library data, surveyed existing library linked data initiatives, and provided recommendations for next steps. In 2014, OCLC Research conducted a survey on linked data projects in LAMs.[10] This survey

6 Seth van Hooland and Ruben Verborgh, *Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish Your Metadata* (London: Facet Publishing, 2014).

7 Antoine Isaac and Ed Summers, eds., "SKOS Simple Knowledge Organization System Primer: W3C Working Group Note," 18 August 2009, https://www.w3.org/TR/2009/NOTE-skos-primer-20090818/.

8 Ed Summers and Dorothea Salo, "Linking Things on the Web: A Pragmatic Examination of Linked Data for Libraries, Archives and Museums" (Cornell University Library, arXiv.org, 20 June 2013), https://arxiv.org/abs/1302.4591.

9 W3C Incubator Group, "Library Linked Data Incubator Group Final Report," 25 October 2011, https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/.

10 OCLC Research, News & Events, "Results from OCLC Research International Linked Data

provided high-level statistics and revealed that many institutions were still in the planning stage of linked data implementation.

The above initiatives in the larger LAM community include archives or incorporate the needs of the archives community. For example, the BIBFRAME vocabulary includes terms specifically defined for archival materials. However, these initiatives do not focus on the linked data implementation for archival materials. Archival materials are different from library and museum resources in certain aspects. Because of their uniqueness, particular implications of linked data for library and museum resources are not applicable to archival materials. For example, one motivation for libraries to implement linked data is to avoid copy cataloguing.[11] This does not apply to archival materials, which are usually unique to each institution. In addition, archival materials are often described in aggregations, which is very different from library cataloguing practices. These unique features make it necessary to take a close look at linked data practices for archival materials.

This article examines detailed analyses of linked data initiatives for archival materials, discusses the impacts of linked data on archival description and archival information discovery, and identifies outstanding issues in producing and consuming archival linked data. More specifically, this article examines the vocabularies/ontologies used in converting and producing archival linked data, links with external data sets, archival linked data publishing mechanisms, and information discovery services based on archival linked data. The research questions are:

- What kinds of linked data have been produced for archival materials and how?
- How has linked data changed archival description practices and archival information discovery services?
- What kinds of mechanisms are used for publishing and consuming archival linked data?

Methodology

The intention of the research project was to study the implementation of linked data for the description and discovery of archival materials. Thus, it focused on linked data projects conducted by archival institutions and projects that include significant amounts of archival materials. A number of projects were identified through a literature review, online searching, and

Survey for Implementers Now Available," 19 September 2014, http://www.oclc.org/research/news/2014/09-19.html.

11 If a library publishes the metadata for a book as linked open data, another library with the same book only needs to add a triple saying that it has this book. It does not need to copy the metadata into its own database.

citation chasing. In addition, many were identified from the three concentrated sources of linked data projects: the LODLAM website (lodlam.net), the W3C Library Linked Data Incubator Group final report,[12] and the Datahub (datahub.io). Only projects that produced tangible linked data were included.[13] Purely exploratory initiatives were excluded, such as the Linked Archival Materials (LiAM) project conducted by Tufts University Libraries (sites.tufts.edu/liam), the Civil War Data 150 project (www.civilwardata150.net), and Karen F. Gracy's study of the mapping between Encoded Archival Description (EAD) and MARC tags.[14] Also excluded were several popular linked data vocabularies, including the DBpedia Ontology, Schema.org vocabularies, Friend of a Friend (FOAF) vocabulary, Linking Open Descriptions of Events (LODE) Ontology, and GeoName Ontology.

For each project identified, various related information resources were gathered and examined, such as project documentation, websites, related publications, and presentations. Some projects do not have comprehensive documentation or are not documented in English, such as the Australian War Memorial linked data project,[15] the Victoria Semantic Wiki (datahub.io/dataset/public-record-office-victoria-semantic-wiki), and the linked data project of Calames (datahub.io/dataset/calames), which is the union catalogue of archives and manuscripts in French university and research libraries. These three projects were not included in the analyses.

In total, about 23 linked data projects were discovered during this research project. Three were excluded because they did not produce real linked data, and another three were excluded owing to a lack of documentation. Therefore, 17 projects were included in the analyses. Some of them are large-scale initiatives conducted by major metadata aggregators or national libraries, such as OCLC's WorldCat, Europeana, the Digital Public Library of America (DPLA), and Data.bnf.fr, created by the Bibliothèque nationale de France (BNF). Others are smaller projects conducted by individual university libraries or small research groups; they include SALDA, Linked Jazz, and the

12 W3C Incubator Group, "Library Linked Data Incubator Group Final Report."

13 Tangible linked data means linked data that has actual physical existence. In contrast with projects that produce tangible linked data, purely exploratory projects study what linked data is, propose ideas about how to implement linked data, study how linked data might affect current practices, or compare traditional metadata formats with linked data vocabularies, but they do not produce real linked data.

14 Karen F. Gracy, "Archival Description and Linked Data: A Preliminary Study of Opportunities and Implementation Challenges," *Archival Science* 15, no. 3 (2015): 239–94.

15 Web Directions, Videos, "Adam Bell & David Peterson – Bringing History Alive: Telling Stories with Linked Data and Open Source Tools" (Sydney: Web Directions South, 13 October 2011), http://www.webdirections.org/resources/adam-bell-david-peterson-bringing-history-alive-telling-stories-with-linked-data-and-open-source-tools/.

Cultural Repositories & Information Systems (CURIOS), which developed the software platform for the Hebridean Connections cultural repository. See the appendix for a complete list of the projects.

Findings and Discussion

Most Projects Primarily Converted Existing Descriptions instead of Creating Original Linked Data

Four types of linked data have been generated for archival materials: 1) archival descriptions; 2) archival authority files for corporate bodies, persons, and families; 3) controlled vocabularies for subject indexing; and 4) content annotations. Most archival descriptions in linked data format have been created by converting existing EAD, MARC, or other descriptions. Some of these conversions are incomplete or even inaccurate. For example, the LOCAH project developed a detailed data model for mapping from EAD to linked data, but the data model was not finalized. As stated on the project's blog:

> For all of the following, the object is simply a copy of the XML element content from the EAD document as an XML Literal. This is a rather "dumb" and probably not terribly useful "translation" from the EAD; in a future iteration of the transform, we hope to extract further useful triples from this part of the EAD data, and we will probably remove some of these triples.[16]

In addition, the LOCAH project converted only a subset of the EAD finding aids of Archives Hub. At Data.bnf.fr, only seven elements in EAD were mapped to the BNF data model for linked data.[17] OCLC published Schema .org tags for all WorldCat records. Although this is a commendable accomplishment, some of the mappings do not seem accurate. For example, archival materials were mapped to "schema.CreativeWork," but many records may not qualify as creative works; photos, meeting minutes, and diaries capture real-time activities but may not necessarily be the expression of intellectual or creative ideas.

In a number of projects, original metadata was added during the conversion process. In the case of the 20th Century Archive project, newspapers were digitized and existing newspaper descriptions converted.

16 LOCAH Project, "Vocabulary," 2011, http://locah.archiveshub.ac.uk/tag/vocabulary/.

17 Bibliothèque nationale de France (BNF), "Semantic Web and Data Model: Presentation of the BnF Ontology (bnf-onto)," accessed 5 November 2015, http://data.bnf.fr/en/semanticweb#Ancre6.

The archivists also added metadata to the dossier level and document level.[18] For the LOCAH project, some metadata elements were added to converted EAD finding aids, such as "gn:postalCode," "gn:locatedIn," and "postcode: postcode."[19] The Digital Archives of Italian Psychology extended EAD finding aids with event information in order to support browsing and searching based on activities and events.[20] Europeana and DPLA automatically generate geo codes for places and allow users to browse resources based on maps. Only one institution, the Norwegian University of Science and Technology (NTNU), has created original archival descriptions entirely in linked data format and has made linked data generation part of its routine process. NTNU's special collections are catalogued and presented in a workflow based purely on linked open data.[21]

Far fewer projects converted archival authority files[22] or generated controlled vocabularies for indexing archival materials. The ReLOAD project created an ontology for EAC-CPF and mapped archival authority records into linked data format.[23] Those involved with the World War I as Linked Open Data project created a specialized controlled vocabulary in linked data format for indexing WWI collections.[24] They created this controlled vocabulary because existing generic controlled vocabularies such as Library of Congress Subject Headings (LCSH) are not adequate for indexing WWI collections.[25] The Linked Jazz project primarily used URIs from existing linked open data sets for names identified from archival records. For names not in the existing

18 Joachim Neubert, "The 20th Century Press Archives as Linked Data Application," accessed 5 November 2015, http://challenge.semanticweb.org/submissions/swc2010_submission_6.pdf.

19 LOCAH Project, "Vocabulary."

20 Claudio Cortese and Glauco Mantegari, "Extending the Digital Archives of Italian Psychology with Semantic Data," Lombard Interuniversity Consortium for Automatic Computation (CILEA) Segrate, Italy, 2011, accessed 5 November 2015, http://www-e.uni-magdeburg.de/predoiu/sda2011/sda2011_04.pdf.

21 Rurik Thomas Greenall, "NTNU University Library – a Linked Open Data Hub," accessed 19 September 2016 at Internet Archive Wayback Machine, https://web.archive.org/web/20160513172952/http://openbiblio.net/2011/09/08.ntnu/.

22 These archival authority files do not include authority files converted by libraries. Converting authority files is commonly done by libraries but not by archives. Archival authority files are different from library authority files.

23 Summit 2013: Linked Open Data in Libraries, Archives and Museums, 19–20 June 2013, Montréal, Québec, "Challenge Entry: ReLOAD – Repository for Linked Open Archival Data," 1 December 2012, http://summit2013.lodlam.net/2012/12/01/challenge-entry-ReLOAD-repository-for-linked-open-archival-data.

24 Eetu Mäkelä, Juha Törnroos, Thea Lindquist, and Eero Hyvönen, "World War I as Linked Open Data," 2013, accessed 5 November 2015, http://www.semantic-web-journal.net/system/files/swj716.pdf.

25 Generic controlled vocabularies are created for many different kinds of disciplines and subjects. Because of their broad coverage, they often lack the depth needed for indexing specialized and focused collections, such as a collection of WWI records.

data sets, it generated URIs and thus contributed to name authorities for jazz artists. Similarly, the Out of the Trenches project created authorities for concepts and events that did not exist in published authorities.[26]

The Linked Jazz and WWI projects used annotation software to identify entities from record content and then link to internal and external data sets.[27] The ReLOAD project annotated entities in finding aids. Compared with converting or generating archival descriptions, in-depth indexing and linking through content annotation are time-consuming and labour intensive, especially when human intervention is needed. Thus, they are only practical for small-scale projects or when crowdsourcing is utilized for annotation. The Linked Jazz project allows any user to access its annotation tool online to annotate archival records. The WWI project only annotated one WWI collection at the University of Colorado Boulder, even though the specialized controlled vocabulary it created can be used to annotate any WWI collection. Table 1 summarizes the different approaches for linked data generation discussed in this section.

Table 1: Approaches to linked data generation[28]

Approaches	**Projects**
Convert archival descriptions	Recollection, Chronicling America, ReLOAD, SALDA, LOCAH, 20th Century Press Archives, Digital Archives of Italian Psychology, Europeana, Bibliothèque nationale de France (BNF), Out of the Trenches, Digital Public Library of America (DPLA), Cantabria Cultural Heritage project, WWI as Linked Open Data
Convert archival authority files	ReLOAD

26 Pan-Canadian Documentary Heritage Network (PCDHN), "Linked Open Data (LOD) Visualization 'Proof-of-Concept' – Out of the Trenches: Linked Open Data of the First World War, Final Report," accessed 5 August 2016, http://www.canadiana.ca/sites/pub.canadiana.ca/files/PCDHN%20Proof-of-concept_Final-Report-ENG_0_0.pdf.

27 Trevor Owens, "WWI Linked Open Data: An Interview with Thea Lindquist," Library of Congress, *The Signal* (blog), 29 July 2013, http://blogs.loc.gov/digitalpreservation/2013/07/linked-open-wwi-data-an-interview-with-thea-lindquist/.

28 Some projects do not have sufficient documentation in English. So information in all the tables in this article shows what was found, which may not be identical to what actually existed.

Generate original archival descriptions in linked data	Norwegian University of Science and Technology (NTNU)
Generate controlled vocabularies	WWI as Linked Open Data
Annotate record content	WWI as Linked Open Data, Linked Jazz
Annotate content of finding aids	ReLOAD

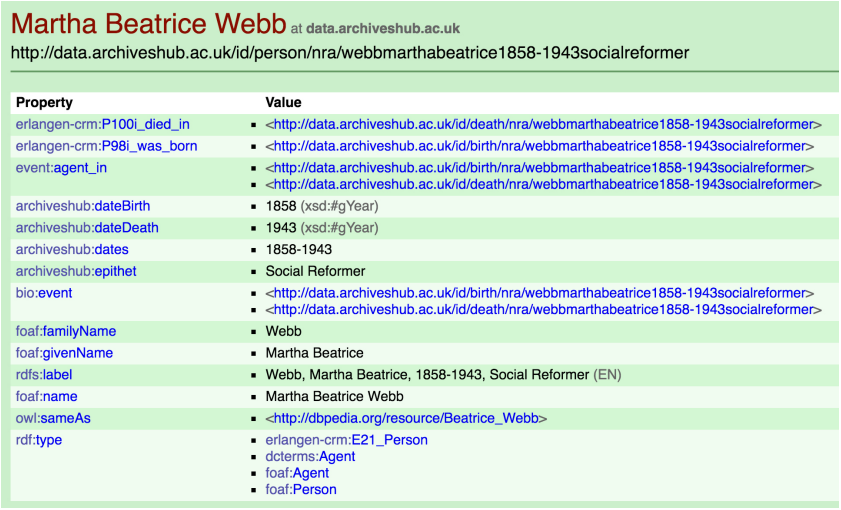
Linked Data Significantly Changes Archival Description

Converting to linked data is not simply a matter of crosswalking to another metadata format. It is a process of enhancing and enriching archival description via linking to external vocabularies and data sets. During the conversion, links are generated with external linked open vocabularies and linked open data sets.[29] Links between the local vocabulary and external linked open vocabularies can be created through mapping relationships, such as "owl: equivalentClass," "owl:equivalentProperty," "SKOS:closeMatch," and "SKOS: exactMatch." These types of links can be used to broaden searches by including equivalent or similar search terms. For example, if the local ontology uses the term "creator," which is linked to "author" in an external vocabulary through "owl:equivalentProperty," then the information retrieval system would be able to return resources described using both "creator" and "author."

Some links to external data sets are created directly by using the URIs maintained by the data sets. For example, URIs of the Library of Congress authorities and Virtual International Authority File (VIAF) are widely used to index archival resources. Other links with external data sets are commonly created by using "owl:sameAs." For example, in figure 1, the person "Martha Beatrice Webb (data.archiveshub.ac.uk/page/person/nra/webbmarthabeatrice1858-1943socialreformer)" at data.archiveshub.ac.uk is

29 Linked open vocabularies are those (metadata schemas, controlled vocabularies, and ontologies) encoded in RDF format and published online for public consumption. Common metadata schemas, ontologies, and controlled vocabularies, such as dcterms, FOAF, Event ontology, and Library of Congress Subject Headings, have all been published as linked open vocabularies. For more information about linked open vocabularies, please refer to Linked Open Vocabularies (LOV), http://lov.okfn.org/dataset/lov. Linked open vocabularies are a type of linked open data set, which also include resource descriptions (finding aids, MARC catalogues, etc.) encoded in linked data format and published online.

linked to Beatrice Webb (dbpedia.org/resource/Beatrice_Webb) in DBpedia. In addition, through the "rdf:type" property, this data set is connected to three linked open vocabularies, including erlangen-crm, dcterms, and FOAF. A user is able to follow these links to discover more information. This enriches archival description with external information.



Martha Beatrice Webb at data.archiveshub.ac.uk

http://data.archiveshub.ac.uk/id/person/nra/webbmarthabeatrice1858-1943socialreformer

Property	Value
erlangen-crm:P100i_died_in	<http://data.archiveshub.ac.uk/id/death/nra/webbmarthabeatrice1858-1943socialreformer>
erlangen-crm:P98i_was_born	<http://data.archiveshub.ac.uk/id/birth/nra/webbmarthabeatrice1858-1943socialreformer>
event:agent_in	<http://data.archiveshub.ac.uk/id/birth/nra/webbmarthabeatrice1858-1943socialreformer> <http://data.archiveshub.ac.uk/id/death/nra/webbmarthabeatrice1858-1943socialreformer>
archiveshub:dateBirth	1858 (xsd:#gYear)
archiveshub:dateDeath	1943 (xsd:#gYear)
archiveshub:dates	1858-1943
archiveshub:epithet	Social Reformer
bio:event	<http://data.archiveshub.ac.uk/id/birth/nra/webbmarthabeatrice1858-1943socialreformer> <http://data.archiveshub.ac.uk/id/death/nra/webbmarthabeatrice1858-1943socialreformer>
foaf:familyName	Webb
foaf:givenName	Martha Beatrice
rdfs:label	Webb, Martha Beatrice, 1858-1943, Social Reformer (EN)
foaf:name	Martha Beatrice Webb
owl:sameAs	<http://dbpedia.org/resource/Beatrice_Webb>
rdf:type	erlangen-crm:E21_Person dcterms:Agent foaf:Agent foaf:Person

Figure 1: An HTML view of linked data

Links with external data sets are generated automatically or semi-automatically. For example, DPLA automatically generates links between its places names and GeoNames.[30] The Linked Jazz project used a transcript analyzer to automatically identify personal names from interview transcripts and then linked them to DBpedia, the Library of Congress Name Authority File, and VIAF through an annotation tool (linkedjazz.org/tools/). Most of the links between GeoNames and the WWI data set were established automatically, but some required manual intervention: "For instance, there were multiple geographic types associated with 'Somme' – an administrative district, a river, etc. – so human intervention was required to determine which to associate with the event 'Battle of the Somme, 1916.'"[31] Table 2 provides an overview of the data sets used by the projects examined in this study.

30 Digital Public Library of America (DPLA), "An Introduction to the DPLA Metadata Model," 5 March 2015, http://dp.la/info/wp-content/uploads/2015/03/Intro_to_DPLA_metadata_model.pdf.

31 Summit 2013: Linked Open Data in Libraries, Archives and Museums, 19–20 June 2013, Montréal, Québec, "Challenge Entry: WWI Linked Open Data Project," 1 May 2013, http://summit2013.lodlam.net/2013/05/01/3993/#more-3993.

Table 2: Linked open data sets used by the projects

Data sets	Projects
GeoNames	SALDA, ReLOAD, Recollection, BNF, WWI as Linked Open Data, DPLA, Chronicling America
DBpedia	SALDA, ReLOAD, 20th Century Press Archives, BNF, Linked Jazz, WWI as Linked Open Data, Chronicling America
Freebase	WWI as Linked Open Data
Linked Languages Resources[32]	Chronicling America
VIAF	ReLOAD, SALDA, 20th Century Press Archives, BNF, Linked Jazz, DPLA, Out of the Trenches
Faceted Application of Subject Terminology (FAST)	Out of the Trenches
Agrovoc[33]	BNF
Library of Congress authorities[34]	SALDA, Linked Jazz, Out of the Trenches, BNF, WWI as Linked Open Data, Chronicling America
German Authority Files	20th Century Press Archives
RAMEAU[35]	Out of the Trenches
UK Archival Thesaurus (UKAT) event subject headings	Out of the Trenches
Library and Archives Canada Canadian Subject Headings	Out of the Trenches
Canadiana Name Authorities	Out of the Trenches
Government of Canada Core Subject Thesaurus	Out of the Trenches
French National Archives' Thesaurus	BNF
Muninn project[36]	WWI as Linked Open Data
Data.archiveshub.ac.uk	SALDA
Chronicling America	20th Century Press Archives
Europeana	WWI as Linked Open Data
Out of the Trenches	WWI as Linked Open Data

32 Linked Languages Resources: A Contribution to the Web of Data by Bernard Vatant, Mondeca, accessed 5 November 2015, http://linkedvocabs.org/lingvoj/.

33 AGROVOC is a controlled vocabulary covering all areas of interest of the Food and Agriculture Organization (FAO) of the United Nations; see Datahub, Organizations: Food and Agricultural Organisation, https://datahub.io/dataset/agrovoc-skos.

34 Library of Congress, Linked Data Service: Authorities and Vocabularies, accessed 5 November 2015, http://id.loc.gov.

35 RAMEAU is a subject-indexing language used in France by the National Library of France, university libraries, many public research libraries, and several private organizations.

36 *The Munnin Project* (blog), "About Munnin," accessed 5 November 2015, http://blog.muninn-project.org/node/3.

As shown in table 2, most data sets used by the projects in this study are various kinds of controlled vocabularies. There are also some generic data sets, such as DBpedia and GeoNames. The projects examined also link to each other. For example, WWI as Linked Open Data links to Europeana, and SALDA links to Data.archiveshub.ac.uk, which is created by the LOCAH project.

Linked data furthers the trend of granulating archival description. Archival descriptions were traditionally created as finding aid documents that were, in some cases, many pages long. One finding aid described one record aggregate, usually a fonds, collection, or series. Consequently, archival information systems supported the search for record aggregates, not for individual components within the aggregate. For example, if finding aids were created for the collection level, then the archival information system could only support the search for collections and not individual series with collections. Later, some archival institutions, such as the National Archives and Records Administration (NARA), created separate metadata records for components of record aggregates, thus dissembling one finding aid into several separate and linked metadata records. This approach to archival description is more granular than the traditional finding aid as it makes each component in a record aggregate individually discoverable. Today, linked data makes archival descriptions even more granular. The description of a record aggregate becomes an RDF graph,[37] and each node and edge[38] within this graph is individually discoverable. In addition to supporting the discovery of records, archival information systems based on linked data can potentially support the search for information about any entities within archival descriptions and provide direct answers to queries.

Multiple Approaches to Publishing and Consuming Archival Linked Data

Most projects have followed the rules for linked open data and have published their data sets through one or more of the following approaches: data

37 An archival finding aid is essentially a collection of statements about various aspects of an archival collection, such as its title, creator, extent, and date range. It also includes statements about relevant entities, such as the birth and death dates/places of the creator. Each of these statements in the archival finding aid can be represented as an RDF triple. All the triples in the finding aid are connected to each other and form a RDF graph.

38 Node means subjects and objects. Edge means predicates. SPARQL query is able to search for everything with an RDF graph.

download, URI resolution,[39] RDFa/microdata[40] embedded into web pages, Web Application Programming Interface (API),[41] or SPARQL endpoint. These publishing mechanisms make archival linked data accessible. However, they are not friendly to generic users. Metadata embedded into web pages are for search engines, browsers, and web crawlers to consume. Web APIs, by definition, are created for applications, not humans. Even a tech-savvy user needs to read the detailed instructions in order to learn how to use a specific API. To use a downloaded data set, one needs to know how to analyze or visualize linked data, or how to mash up with other data sets in order to support certain applications. Some projects provide an HTML view of RDF data, which is easier to understand than XML, JSON, or JavaScript format. For example, at Linked Jazz, BNF, NTNU, and Chronicling America, the same URIs can be used to display an HTML page and the RDF data. However, these HTML web pages display URIs instead of labels. Thus, they are still not friendly for human users. Figure 1, which appeared earlier in this article, is an HTML view of linked data from Archives Hub.

SPARQL endpoints[42] are very powerful. They can be used in searching triple stores, and for adding and editing RDF triples and graphs. As mentioned earlier, SPARQL endpoints allow more complicated queries than traditional search engines do, and they provide direct answers. For example, using the SPARQL endpoint of Europeana, you can search for digital objects, agents, concepts, places, and other entities, and you can count digital objects that reference agents from DBpedia.[43] This will greatly improve the capability of archives in satisfying researchers' needs. Archives users want direct answers[44]

39 URI resolution allows people to retrieve linked data through the URI of an entity. For example, type the following URI into the address bar of a browser: dbpedia.org/page/Beatrice_Webb. You will be able to see the HTML view of the RDF data about Martha Beatrice Webb, as shown in figure 1.

40 RDFa and Microdata are two technical methods to embed metadata into linked data format into HTML web pages. HTML is created primarily for displaying web pages and thus limited in providing metadata. Metadata embedded into web pages through RDFa and Microdata can be discovered and used by search engines to index web pages and improve the accuracy of search engine results.

41 API is the interface through which applications can use the linked data published by an organization. For example, many applications have been created to provide various kinds of services based on the linked data of DPLA (http://dp.la/apps).

42 A SPARQL endpoint is an interface through which people or applications can query the underlying triple store using SPARQL language.

43 Europeana LOD Service, "EDM Sample Queries," accessed 5 November 2015, http://europeana.ontotext.com/sparql/queries#Queries_for_the_main_resources_in_Europeana. (Europeana has reconfigured its website since this study was done; some webpages accessed then are no longer accessible.)

44 Wendy M. Duff and Catherine A. Johnson, "A Virtual Expression of Need: An Analysis of E-mail Reference Questions," *American Archivist* 64, no. 1 (Spring/Summer 2001): 43–60.

but end up searching for records, often because of the limitations of traditional archival information systems.

However, constructing SPARQL queries requires both knowledge about SPARQL language and the underlying data model, which varies for each data set. BNF, WWI, and Europeana provide examples of SPARQL queries that users might be interested in. At Europeana, expert users can contribute and share interesting SPARQL queries. Lay users can ask for certain natural-language queries to be translated into SPARQL queries. Although these mechanisms help people formulate SPARQL queries, search results of SPARQL queries are also returned in machine-readable format, such as JSON, XML, and JavaScript. The SPARQL endpoint of BNF can export search results in HTML, spreadsheet, and comma-separated values (CSV) formats. However, without formatting information provided by a style sheet, the HTML code is not more user-friendly than other machine-readable formats. Please see figure 2 for a spreadsheet output of the BNF SPARQL endpoint. The query is "biographical dates of an author."[45] Thus using a SPARQL endpoint is very difficult for lay users. If no mechanism is used to overcome this barrier, people who benefit from archival linked open data will constitute a much smaller population than even the current user community.

```
<table class="sparql" border="1">
  <tr>
    <th>auteur</th>
    <th>jour</th>
    <th>date1</th>
    <th>date2</th>
    <th>nom</th>
  </tr>
  <tr>
    <td><a href="http://data.bnf.fr/ark:/12148/cb12205345s#foaf:Person">http://data.bnf.fr/ark:/12148/cb12205345s#foaf:Person</a></td>
    <td>"..-.."</td>
    <td>"1330"</td>
    <td>"1394"</td>
    <td>"Luigi Marsili"</td>
  </tr>
  <tr>
    <td><a href="http://data.bnf.fr/ark:/12148/cb122390864#foaf:Person">http://data.bnf.fr/ark:/12148/cb122390864#foaf:Person</a></td>
    <td>"..-.."</td>
    <td>"1650"</td>
    <td>"1711"</td>
    <td>"Jean-Baptiste CramÃ©"</td>
  </tr>
  <tr>
    <td><a href="http://data.bnf.fr/ark:/12148/cb122583299#foaf:Person">http://data.bnf.fr/ark:/12148/cb122583299#foaf:Person</a></td>
    <td>"..-.."</td>
    <td>"1540"</td>
    <td>"15.."</td>
    <td>"Jean de Marconville"</td>
  </tr>
```

Figure 2: Search result of a SPARQL query

To support information discovery, a few projects keep their original user interfaces in addition to publishing linked data. For example, the LOCAH project provides a SPARQL endpoint, but the search interface of Archives Hub remains unchanged. BNF, although it has created a search interface based on linked data, keeps the original catalogues from which it extracts and converts data. BNF and some other projects incorporate linked data features

45 This was translated from French to English by Google Translate (https://translate.google.ca). The original query is "Dates biographiques d'un auteur."

into a conventional user interface. On their interfaces, users can construct conventional user queries, which might be translated into SQL or SPARQL queries under the hood, and search results from the database (relational or triple store) are translated back to human-readable form before being presented to human users. These user interfaces do not require researchers to know anything about linked data technologies. Some user-friendly features of these interfaces include:

- *Search for multiple kinds of entities*. The Hebridean Connections Cultural Repository allows users to search for people, boats, buildings, business, resources, and other kinds of entities. For each search result, it provides a description of the entity and shows semantic links to other associated entities. For example, figure 3 shows the description of a person, and the places, people, and organizations associated with that person. Data.bnf .fr groups all search results into categories based on entity types: people, organizations, works, places, etc. Similar to the Hebridean Connections Cultural Repository, for each entity it displays basic information and various associations. For example, for Victor Hugo (1802–1885), it displays documents about Hugo and other authors related to him. Data.bnf.fr not only links to other linked open data sets, such as VIAF, but it also brings content from external websites and databases, such as Wikipedia and OCLC WorldCat. Although Data.bnf.fr provides only a simple keyword search function, the grouping of search results essentially supports the search for multiple different kinds of entities. The Digital Archives of Italian Psychology is another project that supports the search for multiple kinds of entities.[46]
- *Suggest search terms automatically*. At the 20th Century Press Archives, when a user types the first several letters in the search box, the system automatically searches for possible matches from external authority files published as linked open data and shows several suggestions.
- *Multiple browsing options*. Browsing on a map is a commonly supported function. This is possible because geographic coordinate information from GeoNames is widely used. In addition, DPLA allows users to browse search results based on a timeline. The 20th Century Press Archives supports the browsing of people and organizations in alphabetical order. Europeana and DPLA support browsing through exhibitions, which are organized based on themes.
- *Visualization*. The Linked Jazz project uses LODLIVE, which is an open source tool for linked data visualization. It also created a network visualization tool that can show the connections among jazz artists. The Recollection project provides multiple visualization mechanisms for linked

46 Cortese and Mantegari, "Extending the Digital Archives of Italian Psychology."

data (tag cloud, list, map, pie chart, bar chart, scatter plot, table, timeline, gallery, etc.).[47] DPLA provides a list of user-created APPs that can visualize DPLA data in various ways (dp.la/apps).

The existence of user-friendly features does not mean that the user interfaces of these projects are perfect. The support of the search for multiple kinds of entities is not common. DPLA, ReLOAD, and 20th Century Press Archives allow only simple keyword search. Europeana supports the search based on a number of fields, including title, subjects, creators, dates, and places. All searches lead to digital objects, not other kinds of entities such as people, events, and places. Some user interfaces have evident usability issues. In the search result interface of the Hebridean Connections Cultural Repository, some metadata elements are not labelled appropriately or should not be displayed to users. For example, in figure 3, "Record Type" does not need to be displayed to users, and "Title" for the person should be changed to "Name." Europeana displays URIs instead of labels in search results, and some of its auto-generated tags are not user-friendly. See figure 4 for a screen capture of the metadata for one digital object. The user interfaces of the Out of the Trenches project and the Cantabria Cultural Heritage semantic portal were described in publications and presentations but cannot be found after various efforts. All of these projects have generated RDF data either originally or through conversion. However, some of them do not publish their data or link to external open data sets; these include the Hebridean Connections cultural repository and the Digital Archives of Italian Psychology. The Digital Archives of Italian Psychology creates a further barrier by requiring user registration to see search results.

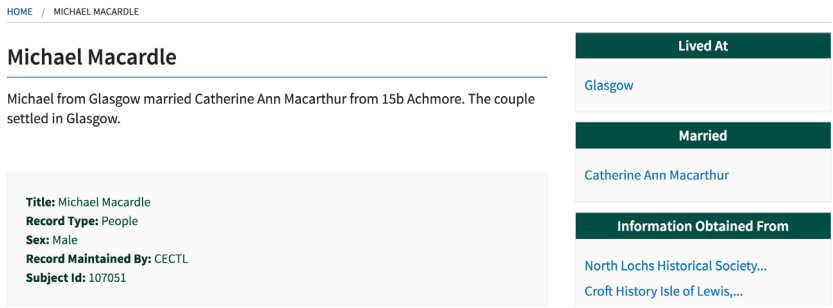


Figure 3: Search result display from the Hebridean Connections cultural repository

47 Eric Miller and David Wood, "Recollection: Building Communities for Distributed Curation and Data Sharing" (presented at Museums and the Web 2010: The International Conference for Culture and Heritage On-line, Denver, CO, 13–17 April 2010), accessed 5 November 2015, http://www.museumsandtheweb.com/mw2010/papers/miller/miller.html.

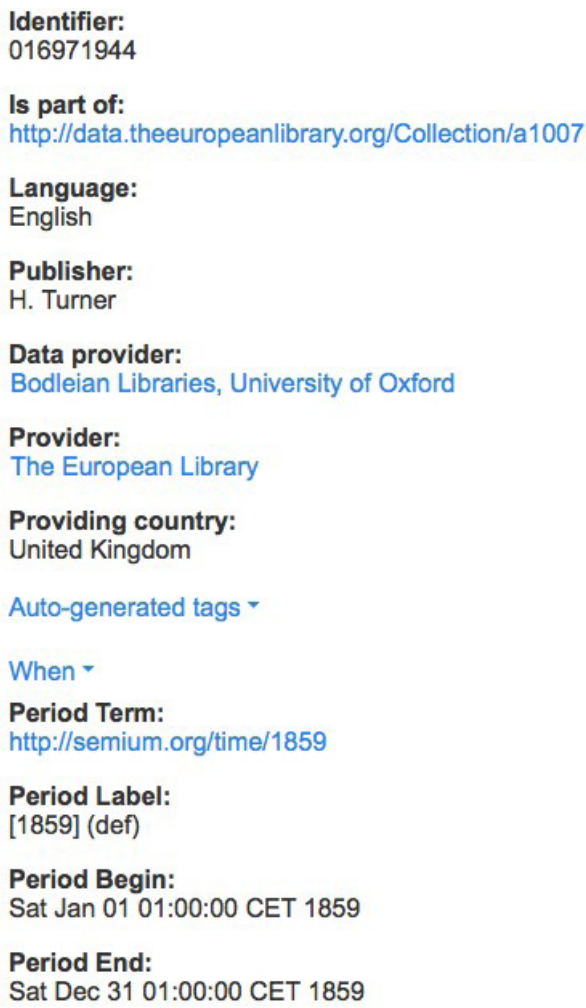


Figure 4: Metadata for a search result in Europeana

Common Use of Existing Ontologies/Vocabularies

In metadata modelling, it has been common practice to reuse existing terms and create new terms (classes and properties in ontologies) only when no existing ones can satisfy the specific needs. This approach avoids reinventing the wheel, reduces the varieties of metadata schemas, and improves interoperability. In fact, all the projects examined in this study reused existing vocabularies, which can be divided into two types. Type I vocabularies describe

various kinds of resources, such as agents, works, places, and events. Type II vocabularies are used for ontology building, including Resource Description Framework Schema (RDFS), Web Ontology Language (OWL), and Simple Knowledge Organization System (SKOS). RDFS defines classes and properties needed in creating basic ontologies. OWL adds more classes, properties, and constraints and can be used to construct more complicated ontologies. SKOS is created for representing knowledge organization systems such as thesauri, classification schemes, subject heading systems, and taxonomies using RDF.

Since type II vocabularies are generic and not specifically related to archival description, they are not examined in this study. Table 3 shows the type I vocabularies used by the projects examined in this study.

Table 3: Use of existing vocabularies

Vocabularies/ontologies	Projects
Dublin Core Metadata Element Set	ReLOAD, Cantabria Cultural Heritage project, BNF, DPLA
Dublin Core terms	SALDA, Chronicling America, LOCAH, Europeana, Out of the Trenches, BNF, DPLA
dcmi-box[48]	BNF
FOAF	ReLOAD, SALDA, Out of the Trenches, Chronicling America, LOCAH, Europeana, Linked Jazz, BNF, WWI as Linked Open Data
Schema.org and its extensions	OCLC WorldCat, BNF, Linked Jazz, WWI as Linked Open Data
Open Graph Protocol	BNF
Open Archives Initiative Object Reuse and Exchange (OAI-ORE)	SALDA, Europeana, DPLA, Chronicling America, LOCAH, 20th Century Press Archives, Europeana, Out of the Trenches
Linking Open Descriptions of Events (LODE)	SALDA, LOCAH
Event Ontology	LOCAH, Linked Jazz, Out of the Trenches
BIO (A vocabulary for biographical information)	ReLOAD, LOCAH, Out of the Trenches, BNF
Relationship Vocabulary[49]	Linked Jazz, WWI as Linked Open Data
DBpedia Ontology	Linked Jazz

48 Dublin Core Metadata Initiative, Metadata Innovation, "DCMI Box Encoding Scheme: Specification of the Spatial Limits of a Place, and Methods for Encoding This in a Text String," http://dublincore.org/documents/dcmi-box/.

49 "Relationship: A Vocabulary for Describing Relationships between People," accessed 5 November 2015, http://vocab.org/relationship.

Vocabularies/ontologies	Projects
Timeline Ontology	LOCAH
Postcode Ontology[50]	LOCAH
GeoNames Ontology[51]	BNF, LOCAH
geo[52]	BNF, WWI as Linked Open Data
GeoRSS[53]	WWI as Linked Open Data
Data Cube[54]	WWI as Linked Open Data
ChangeSet Ontology[55]	Out of the Trenches
E-mail Message Ontology	Out of the Trenches
ign[56]	BNF
insee[57]	BNF
Music Ontology	Linked Jazz
British Broadcasting Corporation (BBC) Ontology	Out of the Trenches
Bibliographic Ontology[58]	Chronicling America, LOCAH, Out of the Trenches, BNF
Marcrel[59]	BNF
International Standard Bibliographic Description (ISBD) Ontology	Out of the Trenches
Resource Discovery and Access (RDA) Element Sets[60]	Out of the Trenches, BNF

50 Ordnance Survey Linked Data Platform, Ontologies: Postcode Ontology, http://data.ordnancesurvey.co.uk/ontology/postcode/.

51 GeoNames ontology (www.geonames.org/ontology/documentation.html) and GeoNames data set mean different things.

52 W3C Semantic Web Interest Group: Basic Geo (WGS84 lat/long) Vocabulary, accessed 1 August 2016, https://www.w3.org/2003/01/geo/.

53 GeoRSS: Geographically Encoded Objects for RSS Feeds, "GeoRSS in RDF," http://www.georss.org/rdf_rss1.html.

54 W3C, "The RDF Data Cube Vocabulary: W3C Recommendation 16 January 2014," https://www.w3.org/TR/vocab-data-cube/.

55 http://vocab.org/changeset/.

56 http://data.ign.fr/ontology/topo.owl#.

57 "Publication de données géographiques au format RDF," http://rdf.insee.fr/geo/.

58 The Bibliographic Ontology, "Bibliographic Ontology Specification Document – 4 November 2009, http://bibliontology.com/.

59 Library of Congress, Linked Data Service, MARC Relators, http://id.loc.gov/vocabulary/relators.html.

60 Open Metadata Registry: Supporting Metadata Interoperability, "The RDA (Resource Description and Access) Vocabularies," http://rdvocab.info/.

Vocabularies/ontologies	Projects
BIBFRAME	Linked Jazz, WWI as Linked Open Data
Europeana Data Model (EDM)	DPLA
FRBRoo	Cantabria Cultural Heritage project
CIDOC Conceptual Reference Model (CIDOC-CRM)	WWI as Linked Open Data, LOCAH, Digital Archives of Italian Psychology, Cantabria Cultural Heritage project
BNF ontology	Out of the Trenches
EAC-CPF ontology of the ReLOAD project	Ontology for Archival Description (OAD) of ReLOAD
ISAD(G) ontology	ReLOAD
LOCAH	SALDA

Ontologies/vocabularies in table 3 roughly fall into three groups: Group I includes generic vocabularies used across communities, such as Dublin Core (DC) Metadata Element Set and DC terms, FOAF, Event Ontology, Relationship Ontology, and BIO. Group II, as indicated in the shaded cells, includes specialized vocabularies created outside of the LAM community, such as BBC ontology and Music Ontology.[61] Group III includes ontologies/vocabularies created for the LAM community, such as CIDOC-CRM and BIBFRAME. The bottom three ontologies in table 3 are created specifically for archival materials, including the LOCAH vocabulary, the *General International Standard Archival Description (ISAD(G))* ontology, the EAC-CPF ontology, and OAD by the ReLOAD project.

The vocabularies used in each project were selected based on careful consideration. The Out of the Trenches project selected existing vocabularies that have stable meanings and are supported by major organizations, rather than ontologies specific to a domain and supported by a small/relatively unknown group.[62] Similarly, the LOCAH project asked the following questions when selecting vocabularies: "Is the vocabulary stable or still being developed? Is it described following 'modern' good practice for RDF vocabularies? Is it being managed/curated? By an individual/institution/community? Does it have the support of a community of users?"[63]

The Open Graph Protocol and Schema.org vocabularies were used because of the specific functionalities they support. BNF uses Open Graph Protocol to make its web pages shareable on social media sites. OCLC WorldCat

61 The boundary between generic and specialized vocabularies is not always clear-cut.
62 PCDHN, "Linked Open Data (LOD) Visualization 'Proof-of-Concept.'"
63 LOCAH Project, "Vocabulary."

uses Schema.org vocabularies to make its catalogue records discoverable by major search engines, which is important for archival materials because today most searches for information "start not in a library, or even in a Web-accessible library catalog, but elsewhere on the Internet."[64] Archival finding aids, catalogues, and authority files are created and curated by professionals, and regarded as accurate and reliable by scholars. Yet this valuable knowledge is not visible to the general public and is underutilized by major search engines. For example, Google Knowledge Graph gathers data from Wikipedia, the Central Intelligence Agency (CIA) World Factbook, and Freebase.[65] It does not gather data from or link to the authority files and catalogues of libraries and archives. Instead, it chooses to link its search results to Wikipedia, which is not considered by scholars to be an authoritative and quality-checked resource. Using Open Graph Protocol and Schema.org vocabularies will make archival materials more visible to general Web users.

Many projects choose to use generic vocabularies for some classes and properties instead of specialized LAM terms, because they recognize the benefits of doing so. Various user studies have shown that archives users, especially novices, find archival terms difficult to understand.[66] Using generic vocabularies makes archival descriptions more understandable for a broader user community and makes it easier to integrate archival materials with other resources described using the same vocabularies. In fact, some libraries have explicitly stated that they purposely avoid library-specific standards, whose complexity does not provide practical benefit. For example, NTNU's website states:

> We try to keep things relatively simple, using common vocabs and adding few properties and classes of our own. We have largely avoided relying on the explicitly library domain vocabularies.… If you're interested in reasons why we avoid the library domain stuff, it's simple: the models are entrenched in (often record-based) approaches that aren't particularly interesting for us … we have avoided wholehearted adoption of, for example, FRBR because – when we tried it – it gave us no real benefits or things that we couldn't achieve in a simpler (non-domain-specific/more trivial) way.[67]

64 OCLC, "OCLC's Work with Library Linked Data Detailed in New Book," news release, 19 June 2015, http://www.oclc.org/news/releases/2015/201519dublin.en.html.

65 Danny Sullivan, "Google Launches Knowledge Graph to Provide Answers, Not Just Links," Search Engine Land, 16 May 2012, http://searchengineland.com/google-launches-knowledge-graph-121585.

66 Elizabeth Yakel, "Encoded Archival Description: Are Finding Aids Boundary Spanners or Barriers for Users?" *Journal of Archival Organization* 2, no. 1–2 (2004): 63–77.

67 Greenall, "NTNU University Library – a Linked Open Data Hub."

BIBFRAME has compressed the four levels of group 1 entities in FRBR to two classes: work and instance. Schema.org makes the definition of "CreativeWork" even simpler, not differentiating between the conceptual and physical aspects of information resources.

Generic vocabularies cannot represent all the nuances needed for archival description. Therefore, some projects also use specialized vocabularies that define terms for archival materials.[68] For example, BIBFRAME and the bibliographic extension of Schema.org define "collection" as a class. Although this class is not unique for archival materials, it can definitely be used for archival collections. BIBFRAME also defines "archival" as a subclass of instance (bibframe.org/vocab-list/#Archival). Other terms in BIBFRAME that address archival needs include "arrangement" and "custodialHistory." "ArchiveMaterials" has also been proposed as a subclass for "CreativeWork" in the experimental library extension of Schema.org vocabularies (www.lov.okfn.org/dataset/lov/vocabs/lib). As explained earlier, this may not be appropriate because many archival materials do not qualify as creative works.

Projects examined in this study also use the ontologies/vocabularies created by each other. For example, DPLA uses the EDM model created by Europeana. SALDA uses the data model and data conversion tool created by LOCAH. The OAD ontology uses the EAC-CPF ontology, which was also created by the ReLOAD project. In addition, some projects define new terms when no existing terms suffice for specific needs; these projects include BNF, DPLA, Chronicling America, Europeana, ReLOAD, and LOCAH.

Data Modelling Based on Existing Archival Description Standards instead of the Current Archival Universe

Many projects design their data models based on existing description standards. In other words, they select existing terms and define new ones based on the needs to convert existing descriptions. The OAD ontology created by the ReLOAD project is a synthesis of commonly used metadata elements in existing archival description standards, including EAD and *ISAD(G)*. The *ISAD(G)* ontology (www.cc.uah.es/ie/ontologies.html) is an OWL mapping of elements defined in *ISAD(G)*. The EAC-CPF ontology created by the ReLOAD project

68 In order to design a data model that can fully reflect archival materials and also be interoperable with other vocabularies, archivists can use a generic vocabulary and then create extensions for archival materials, or they can create a data model specifically for archival materials and then map to generic vocabularies. The appropriate balance between specialization and interoperability might be decided on a case-by-case basis and depends on the purpose of creating the data set. If the data set is created primarily to share across communities, then using generic vocabularies might have more weight. If it is created for use in a particular archival institution, then representing the special features of archival materials might dominate.

defined the class "controlArea" and the class "descriptionArea" following the structure of the EAC-CPF schema.[69] The blog of the LOCAH project states, "We need to think in terms of what an EAD document is 'saying' about 'things in the world' and what sort of questions we want to answer about those 'things.'"[70] The data model of the LOCAH project includes these three classes: finding aid, EAD document, and biography history.[71] These classes (finding aids, EAD document, biography history, controlArea, descriptionArea) are defined specifically for converting EAD and EAC descriptions. Including them in the data model for archival linked data will ensure that information recorded in those elements will not be lost during conversion.

Existing archival description standards were created based on traditional archival description practices and represented community consensus. They intend to represent the hierarchical structure of archival collections, and to help users locate and understand records. Data modelling based on existing archival description standards is easy and quick compared with creating completely new models from scratch. This is a rational choice in the transitional stage when most archival linked data is generated through converting existing archival descriptions. However, existing archival description standards were created in the older technology environment. Finding aids were produced primarily for archival materials in analog formats. Users were separated from records and could only see the finding aids. In addition, archivists were available to assist with or even educate researchers about finding aids. This is not always the case anymore. Today, people expect to find archival records on the open Web using Google and other generic search engines rather than library catalogues and finding aids, and they prefer to discover and use those materials on their own rather than through an archivist as a mediator.[72] Higgins, Hilton, and Dafis pointed out that the digital environment offers new possibilities for archival description, such as search engine discovery, a merging of the description and digital resource, user-generated arrangement and description, tagging, and linkage to existing biographical, historical, and contextual resources.[73] The shifted technology and information environment may mean

69 Silvia Mazzini and Francesca Ricci, "EAC-CPF Ontology and Linked Archival Data," in *Proceedings of the 1st International Workshop on Semantic Digital Archives (SDA), Berlin, 29 September 2011*, 72–81, accessed 5 November 2015, http://ceur-ws.org/Vol-801/paper6.pdf.

70 LOCAH Project, "The 'Things' in EAD: A First Cut at a Model," 28 September 2010, http://locah.archiveshub.ac.uk/2010/09/28/model-a-first-cut/.

71 LOCAH Project, "Modelling," 2011, http://locah.archiveshub.ac.uk/tag/modelling/.

72 Jennifer Schaffner, "The Metadata *Is* the Interface: Better Description for Better Discovery of Archives and Special Collections, Synthesized from User Studies" (Dublin, OH: OCLC Research, 2009), 4, accessed 5 November 2015, http://www.oclc.org/content/dam/research/publications/library/2009/2009-06.pdf.

73 Sarah Higgins, Christopher Hilton, and Lyn Dafis, "Archives Context and Discovery: Rethinking Arrangement and Description for the Digital Age" (presented at the ICA Second

that certain information in traditional finding aids is no longer required or is less important, such as the physical location of digital records. In the meantime, information needed in this new technology environment, such as file formats of digital records, may not exist in traditional finding aids. Data modelling based on existing standards will limit the power of archival linked data to capture what is available in existing archival descriptions, and may not fully meet the needs of archival description today.

Conversion is only a temporary stage during the migration to a new technology environment. Much more archival linked data will be generated originally rather than converted from traditional finding aids. In data modelling for archival linked data, in addition to incorporating useful elements from existing archival description standards, archivists need to consider the needs of representing new types of archival materials in a new technology environment. Archival linked data are organized based on classes, properties, and instances, and are represented as RDF graphs. EAD and EAC-CPF documents will no longer exist, nor will some elements defined in them, such as "controlArea" and "descriptionArea." Although the decision on classes and properties in the data model for archival linked data should be made in light of information needs of today's users, only the WWI project started with a user study, and this project primarily created a controlled vocabulary for annotating records content rather than archival descriptions.

Previous studies on archives users and archival description, although they might be outdated to various degrees, have produced findings informative to data modelling for archival description. For example, archival terminology and multi-level hierarchical structure are unfamiliar and hard to understand, especially for novice users.[74] Large blocks of text, a carry-over from print finding aids, tend to slow users down and cause frustration. Most users prefer item-level search and retrieval rather than having to scan hierarchical structure.[75] Elena et al. found that when researchers are faced with a particular topic, they identify and isolate entities – such as persons, places, or organizations – related to the topic. Then they search for information about each entity using one or more keywords.[76] This means that searching for information about

Annual Conference, 11–15 October 2014, Girona, Italy), accessed 5 September 2016, http://www.girona.cat/web/ica2014/ponents/textos/id174.pdf.

74 Wendy M. Duff and Penka Stoyanova, "Transforming the Crazy Quilt: Archival Displays from a User's Point of View," *Archivaria* 45 (Spring 1998); Morgan G. Daniels and Elizabeth Yakel, "Seek and You May Find: Successful Search in Online Finding Aid Systems," *American Archivist* 73, no. 2 (Fall/Winter 2010): 535–68.

75 J. Gordon Daines III and Cory L. Nimer. "Re-Imagining Archival Display: Creating User-Friendly Finding Aids," *Journal of Archival Organization* 9, no. 1 (2011): 4–31; Ian G. Anderson, "Are You Being Served? Historians and the Search for Primary Sources," *Archivaria* 58 (Fall 2004): 81–129.

76 Torou Elena, Akrivi Katifori, Costas Vassilakis, George Lepouras, and Constantin Halatsis,

entities is the original need of researchers. However, finding aids are not well designed to support the search for specific entities.[77] Thus, researchers have to search using keywords, which results in the loss of the contextual meanings of entities.

Bron, Proffitt, and Washburn analyzed the EAD finding aids harvested by ArchiveGrid. They found that some EAD elements are never used or used very rarely. They also found that EAD 2002 does not support map-based and event-based discovery. To gain such support, new elements need to be added, or the content of some existing elements needs to be structured and more consistent. For example, content in the <geogname> element of EAD 2002 could be recorded so that it supports map-based discovery, and the content in the <extent> element could be recorded to support sorting based on size.[78] EAD3 has been released, and it claims to be supportive of linked data.[79] However, this author has found that it only contains minor adjustments of elements and attributes, not an overhaul of the modelling structure to better suit today's technology environment and user needs.

There exist non-EAD-based tools and methods for archival information organization. Earlier in this article, it was mentioned that some national archives have created catalogues containing multiple separate and linked metadata records, each for one node in the archival hierarchy. As shown in table 3, OAI-ORE, which is based on linked data, has been used for archival description in a number of projects. Many digital repository software tools, such as DSpace, offer a hierarchical way to organize digital records, which can be exported in METS format. Higgins, Hilton, and Dafis have suggested that archivists apply to archival finding aids the knowledge organization features common to popular online services such as Amazon, Facebook, and Flickr.[80] I presented the concept of content-level control for digital archives.[81] I pointed

"Historical Research in Archives: User Methodology and Supporting Tools," *International Journal on Digital Libraries* 11, no. 1 (2010): 25–36.

77 Wendy M. Duff and Catherine A. Johnson, "Accidentally Found on Purpose: Information-Seeking Behavior of Historians in Archives," *Library Quarterly* 72, no. 4 (October 2002): 472–96; Wendy Duff and Catherine Johnson, "Where Is the List with All the Names? Information-Seeking Behavior of Genealogists," *American Archivist* 66, no. 1 (2003): 79–95.

78 Marc Bron, Merrilee Proffitt, and Bruce Washburn, "Thresholds for Discovery: EAD Tag Analysis in ArchiveGrid, and Implications for Discovery Systems," in *Making Archival and Special Collections More Accessible* (Dublin, OH: OCLC Research, 2015), 63, accessed 26 March 2016, http://www.conference-center.oclc.org/content/dam/research/publications/2015/oclcresearch-making-special-collections-accessible-2015-a4.pdf#page=70.

79 Technical Subcommittee for Encoded Archival Description of the Society of American Archivists, *Encoded Archival Description Tag Library: Version EAD3* (Chicago, IL: Society of American Archivists, 2015), accessed 30 March 2016, http://www2.archivists.org/sites/all/files/TagLibrary-VersionEAD3.pdf.

80 Higgins, Hilton, and Dafis, "Archives Context and Discovery."

81 Jinfang Niu, "Archival Intellectual Control in the Digital Age," *Journal of Archival Organization* 12, no. 3–4 (2014): 186–197, DOI:10.1080/15332748.2015.1154747.

out that the intellectual control of digital records often reaches below the item level and extends to the content of records, including the description of record components and individual variables of data sets, as well as the annotation of textual content. These methods and ideas provide food for thought for modelling archival linked data.

Conclusions

The archives community is, for the most part, in the early stages of linked open data implementation. Most archival institutions are converting existing descriptions rather than producing original ones. Archival linked data are made accessible, but not always in an easy-to-use format for researchers. Some data models for archival linked data are created based on existing archival description standards rather than user needs in a shifted technology environment. Notwithstanding these limitations, there have been some encouraging accomplishments. Archival descriptions have been enriched via external linking generated during linked data conversion. The power of SPARQL queries for searching archival linked data has been demonstrated. Generic vocabularies are used in archival linked data, which makes archival description understandable to a broader user community and facilitates interoperability.

Linked data implementation is a complex, multiple-step process. As semantic web technologies mature, and as more archival institutions shift their attention from converting to producing original linked data, and from publishing data to providing user-friendly linked data services, archival description practices will be significantly changed, and archives users will experience better information services without requiring technical knowledge of linked data.

Jinfang Niu *is an assistant professor at the School of Information, University of South Florida. She received her PhD from the University of Michigan, Ann Arbor. Prior to that, she worked as a librarian at the Tsinghua University Library in Beijing for three years. Her current research focuses on information organization, digital curation, and archives management.*

Appendix: List of Archival Linked Data Projects

1. Data.bnf.fr created by Bibliothèque nationale de France (BNF).
2. 20th Century Press Archives created by the German National Library of Economics, http://zbw.eu/beta/p20.
3. Chronicling America newspaper archive created by Library of Congress, http://chroniclingamerica.loc.gov/about/api/#linked-data.
4. Recollection software platform created by Library of Congress, https://zepheira.com/2011/11/library-of-congress-launches-recollection-as-viewshare-org.
5. World War I (WWI) Linked Open Data project conducted by University of Colorado Boulder library and the Semantic Computing Research Group at Aalto University and University of Helsinki, Finland, http://www.seco.tkk.fi/u/juhtornr/lodlam.
6. Norwegian University of Science and Technology (NTNU) special collections catalog, http://www.ntnu.no/ub/digital/document/ntnu22.
7. Sussex Archive Linked Data Application (SALDA) Project conducted by University of Sussex libraries, http://blogs.sussex.ac.uk/salda/about.
8. Cantabria Cultural Heritage ontology and Semantic Portal created by University of Cantabria libraries [no website was found].
9. Digital Public Library of America (DPLA), http://dp.la.
10. OCLC WorldCat, https://www.worldcat.org.
11. Europeana digital library created by the Europeana Foundation, http://www.europeana.eu/portal.
12. Cultural Repositories & Information Systems (CURIOS) project conducted by a UK research group that consists of professors from sociology, computer science and informatics. This project developed the software platform for the Hebridean Connections cultural repository, http://curiosproject.abdn.ac.uk.
13. Linked Jazz project led by Cristina Pattuelli, associate professor at the School of Information at the Pratt Institute, New York, https://linkedjazz.org.
14. Digital archives of Italian Psychology created by a group of psychology professors in collaboration with the library of the University of Milan-Biococca, http://aspi.promemoriagroup.com.
15. Archives Hub Linked Data (LOCAH) project conducted by UK Office for Library and Information Networking (UKOLN) and Mimas, which is part of the Digital Resources Division at Joint Information Systems Committee (JISC); http://locah.archiveshub.ac.uk.
16. ReLOAD (Repository for Linked Open Archival Data) project sponsored by Archivio Centrale dello Stato (ACS), Istituto Beni Culturali Regione Emilia Romagna (IBC), and Regesta.exe, http://labs.regesta.com/progettoReload/en.

17. Out of the Trenches project conducted by the Pan-Canadian Documentary Heritage Network (PCDHN); http://www.canadiana.ca/en/pcdhn-lod.