Linked Data for Cultural Heritage

Edited by
Ed Jones and Michele Seikel



Z
666.73
.L56
L555
2016

INTRODUCTION

Ed Jones

The Web that we experience is designed for people: documents to be read, music, lectures, and so on to be heard; photos, videos, and so on to be viewed; files (spreadsheets, databases, etc.) to be downloaded; and links to be clicked. But of course the Web is also designed for machines, to the extent that machines read the codes and present the results that make it all work.

Over time the Web has evolved some very effective methods for identifying the different parts of a resource using an ever more elaborate hypertext markup language (HTML) and structuring the ways these parts are presented to us visually using cascading style sheets (CSS). The former will tell your browser "this string of text is a main heading" or "this is a bulleted list," while the latter will tell it "main headings should be presented in dark blue bold italic 12-point Times New Roman type" and "bulleted lists appear in dark green 8-point Arial type, with diamonds representing the bullets." And these behind-the-scene methods have become virtually universal on the Web (and much more elaborate than these examples). But introducing semantics to these documents via linked data is like moving from a black-and-white world to a world of color.

Much can be done when machines can process meaning, but that is not possible when the meaning of the documents, and so on, is not understandable by these machines. A machine may know that such-and-such is the main heading in a document, and that this heading is followed by several paragraphs, with images and videos and links to other documents, but that tells the machine nothing about the *meaning* of these things—their *semantics.* What is in the image? Who is the video about? Linked data and its various web standards enable this machine's "understanding" and allow machines to exploit it.

You can see linked data in action today when you perform a Google search for a popular topic. For example, searching for "hunchback of notre dame book"

pulls up a web page with the search results on the left and a panel containing various data about the book on the right. The data has been extracted from a variety of online sources and includes star ratings of the book on websites such as Amazon (with links if you want to order a copy), a summary of the book from Wikipedia, a date of first publication, author, characters, links to various sources for reading the book online, and links to other books by Victor Hugo and other books searched for by people who also searched for the *Hunchback of Notre Dame*. If you click on the link to Victor Hugo, that triggers another search and takes you to search results and a panel of similar data on Hugo, with images, a short description, links to his place of birth, death, and burial, and so on. Clicking on his place of birth (Besançon) pulls up yet another collection, this one including a zoomable map of Besançon, its area and population, current weather, current time, and local points of interest. Clicking on the population brings up a chart created from UN data and an invitation to "explore more" by clicking on a small globe symbol. This in turn expands the chart to explore and compare populations for any place(s) in the world. Google calls this the Knowledge Graph, and while the Google Knowledge Graph makes use of some sophisticated data extraction and categorization techniques as well as linked data, it gives you some idea of the potential of linked data as a technique for establishing semantic relationships across the Web.

While this book will be concerned primarily with the use of linked data in the context of so-called cultural heritage institutions—libraries, archives, and museums (LAM)—you can get an idea of the growing scope of the available datasets by examining the list maintained by the World Wide Web Consortium (W3C) at www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets.

We are still some distance from the world of linked data that Tim Berners-Lee, James Hendler, and Ora Lassila envisaged fifteen years ago when they first proposed a Semantic Web. In their world, one's mobile device would be able to take a verbal request to make a medical appointment and then automatically negotiate with software at the doctor's office to arrive at the absolutely best of all possible appointments, taking into account (and weighting appropriately) a much larger range of relevant variables than you could ever do on your own.[1] Ideally, from a library's point of view, such a world would also enable devices to automatically select the most appropriate copy of a book, article, and so on for you to read, download, or borrow in a physical form, taking into account your preferred format for delivery, your physical location, the institutions where you have borrowing privileges (and where copies are available), commercial

services to which you subscribe, and varying restrictions on access, and so on, that attach to the various available copies. Wouldn't it be lovely?

But such a world represents nirvana. For the time being, we must content ourselves with more modest ambitions for linked open data (LOD). First among these is agreement on standards, for without agreed-upon standards, linking becomes much more problematic and labor-intensive.

FIVE-STAR LINKED OPEN DATA

In 2010 Berners-Lee proposed a five-star scheme for rating linked open data in figure I.1.[2]

It can easily be seen how each additional star makes the data more useful as linked open data. The first star is awarded simply for making the data openly available (even if it's just an image scan of a chart or table, for example). The second star is awarded for giving the data a structure (e.g., spreadsheet or database) that can be read by a program. The third star is awarded for doing so in a nonproprietary format (say, comma-separated values [CSV] for a table).

So far, so easy. The resource usage reports many libraries receive from vendors as part of Project COUNTER would typically be awarded three stars.

But now comes the hard part, where machine-readable structured open data becomes linked open data.

The fourth star is awarded for using open standards from the World Wide Web Consortium (W3C) to give semantic structure to the data, specifically using the resource description framework (RDF) to describe the data and providing a SPARQL endpoint to receive and respond to queries (analogous to a database able to respond to a SQL query). RDF provides a basic structure to

★	Available on the web (whatever format) but with an open license, to be open data
★★	Available as machine-readable structured data (e.g., Excel instead of image scan of a table)
★★★	Same as ★★ but in a non-proprietary format (e.g., CSV instead of Excel)
★★★★	All the above, plus, use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
★★★★★	All the above, plus, link your data to other people's data to provide context

Figure I.1 | **Five-star scheme for rating linked open data**

linked data in the form of subject-predicate-object triples (e.g., <thisResource> <hasTitle> <"A la recherché de ma tante perdue">). This basic structure is often elaborated with other vocabularies such as RDFs and OWL (described in more detail in some of our contributions) and various domain-specific vocabularies. For example, the library domain has elaborated various vocabularies for Resource Description and Access (RDA), Functional Requirements of Bibliographic Records (FRBR), and the International Standard Bibliographic Description (ISBD). When your data has achieved four stars, it is in essence "open for business" (though still possibly in want of customers).

Finally, the fifth star is awarded when you take matters into your own hands and begin linking your data to other people's data. At this point, interesting things start happening, like what we experience with those Google Knowledge Graph panels. But creating these links can be very labor-intensive. It's most cost-effective when machines can be called on to do the matching—at least provisionally—as with the virtual international authority file (VIAF), where a commonality of works is used to identify identifiers for the same author in different vocabularies. (For example, it's highly probable that identifier A in vocabulary M represents the same author as identifier B in vocabulary N if both vocabularies identify them as the authors of works X, Y, and Z.)

How many datasets have achieved five stars? A rough idea can be had by periodically viewing the Linked Open Data Cloud at http://lod-cloud.net, which shows datasets clustered in broad categories and linked to one another with varying success (DBpedia and GeoNames are particularly well-connected).

VERY SIMPLE DESCRIPTION OF LINKED DATA

At this point it will be useful to describe, in very broad terms, just what is involved in getting those fourth and fifth stars. A basic understanding of linked data will also help understand the individual contributions to this volume.

In its simplest form, linked data uses the resource description framework and is expressed as three-part statements called triples, each triple consisting of a subject (what the triple is about), a predicate (describing the relationship of the subject to its object), and the object (describing an attribute of the subject or identifying the subject of another triple to which it is related). The subject and predicate are always represented by persistent HTTP URIs (i.e., uniform resource locators [URLs]), where relevant information will be found about them. Ideally, the object is also represented by an HTTP URI, but failing this

it may be represented by a literal (such as the transcribed title of a resource). If the object in a triple is represented by an HTTP URI and that HTTP URI is maintained by someone other than you, then you are crossing the border into five-star territory. Congratulations! It is the most difficult border to cross, and many abandon the struggle after a few attempts.

In a given RDF triple, the subject typically represents a resource that you control. For example, subjects in a triplestore representing a library catalog would typically contain HTTP URIs that identify items in the collection (e.g., http://lccn.loc.gov/2013005033).

The predicate, characterizing the relationship between the subject and object, would typically be drawn from a published linked data vocabulary (for example, Schema.org (https://schema.org/docs/schemas.html) or the various vocabularies used for expressing data based on the cataloging standards RDA, FRBR, and ISBD as linked data: www.rdaregistry.info for the RDA vocabularies, http://iflastandards.info/ns/fr/ for the FRBR vocabularies, and http://iflastandards.info/ns/isbd/ for the ISBD vocabularies.

The object may be a literal, especially when it is transcribed data, or it may, like the subject and predicate, be an HTTP URI. This latter case is certainly the preferred one, since linking—that fifth star—becomes problematic without it. The predicate appearing in a triple typically takes a defined value as its object. For example, the predicate RDA carrier type takes values defined at www.rdaregistry.info/termList/RDACarrierType/.

```
lccn:2013005033 rdam:carrierType rdact:1049
```

This triple says that the resource identified in the subject (lccn:2013005033), a book called "RDA and Serials Cataloging," has a carrier type (rdam:P3001) of "volume" (rdact:1049). The string of characters (prefix) preceding the colon in each part of the triple is an abbreviated human-readable way of representing the full HTTP URI of the namespace from which the label following the colon is taken. In this case, the subject is taken from the Library of Congress control number namespace, the predicate is taken from the RDA manifestation properties namespace, and the object is taken from the RDA carrier type namespace (a value vocabulary).

While there are only a few legal values for the RDA carrier type, value vocabularies can be quite large. This is especially true in the realm of subject analysis, where an object may be drawn from an extensive vocabulary such as the Library of Congress Subject Headings (http://id.loc.gov/authorities/sub

jects) and legal values can run into the thousands. In such cases, the value can link to the thesaurus, and to triples within that thesaurus that link to related values (for instance, broader or narrower terms), and so on. This is where the power of linked data comes into play.

Let us follow one trail (conceivable but still hypothetical at this point, since the links are not all in place):

lccn:2013005033 rdau:author lccn:n2013018475
lccn:n2013018475 skos:exactMatch viaf:105902737
viaf:105902737 skos:exactMatch isni:000000007525185X
isni:000000007525185X skos:exactMatch orcid:0000000279663733
orcid:0000000279663733 rdau:authorOf doi:10.5860/lrts.54n2.77

These five triples take one from the book (the subject of the first triple) to a journal article by the same author (the object of the fifth triple). The intervening triples take one from the identifier of the author in one identification scheme to the corresponding identifier in another (LC NAF, VIAF, ISNI, and finally ORCID) until one has moved from a scheme designed mainly for library materials (LC NAF) to one designed primarily for scholarly articles, research papers, and so on (ORCID).

SPARQL QUERY LANGUAGE

The preceding set of triples can be traversed iteratively by a human searcher using the linked data query language SPARQL or, preferably, by machine with a rather more complex general-purpose SPARQL query built into the discovery system (for example, your library's discovery system or, perhaps, a web search engine) and hidden from the user, a query that, for example, would routinely retrieve all related materials—whether books or scholarly articles—written by an author whenever one clicked on the author's name in a display of bibliographic data.

Of course, this would require that RDF and other linked data standards be implemented much more broadly than they have been at this point. But the potential of linked data for improving the precision and recall of bibliographic searches should be apparent. The crucial element is linking—that fifth star.

CHALLENGES

Among the challenges to linked open data is the fact that while linked data is increasingly common both on the open web and within enterprises—linked cells in different Excel spreadsheets are a form of internal linked data—and increasing amounts of data, especially government data, are freely available on the open web, combining the two goals—linking and openness—has presented challenges. For example, most ISSN metadata—essential for identifying and linking continuing resources such as scholarly journals—lies behind a paywall, and its sale helps fund the ISSN Network.

But the presence of paywalls and similar barriers does not necessarily preclude the linking of open and proprietary data, though it may restrict the availability of certain linked data to those with the proper credentials.[3]

Beyond this, recent research on RDF triples extracted from nearly 400,000 pay-level domains (PLD) suggests that widespread adoption of both schema .org and linked open data will be dependent on a number of factors, including

- a direct business incentive such as improved listing in search engine results;
- good documentation with ready-to-adapt examples;
- implementation in widely deployed platforms such as Drupal; and
- use of a flexible standard that adapts to widespread violation.[4]

And finally the old bugbear of incorrect data will always be with us: just because metadata is properly structured doesn't mean it's correct. In this regard, I seem to be forever notifying websites—including those using linked data—that I am not the same person as the basketball player Ed "Too Tall" Jones (born the same year as me). I expect that after I'm gone, my descendants will still need to carry on this activity.

So linked open data remains very much a work in progress, and much of the progress has taken place within the domain of the cultural heritage institutions: libraries, archives, and museums. For an accessible and clear-eyed manual for implementing linked data for these institutions, I strongly recommend the excellent book by Seth van Hooland and Ruben Verborgh listed at the end of this chapter.[5] There is no question that the structure of linked data and the machine inferencing it supports shows great promise; many very large datasets

have now been made available as RDF, and the SPARQL query language enables sophisticated queries across datasets. The question is, what will be the "killer app" that breaks linked open data out to the wider world and accelerates its uptake? Will it be an incremental extension of schema.org, the linked data vocabulary supported by the major search engines (such as Google and Bing)? Or will there be a thousand flowers blooming and finally achieving a critical mass, as specialized vocabularies enable the optimal exploitation of a variety of domain-specific data sets? Perhaps it will be a project described in this volume.

OUTLINE

In chapter 1, Hilary Thorsen (Stanford University) and M. Cristina Pattuelli (Pratt Institute) survey the use of linked data in significant projects across the cultural heritage domain, including Europeana and the Digital Public Library of America (DPLA), before proceeding to a more detailed description of Linked Jazz, a research project aimed at using linked data technologies "to uncover meaningful connections between documents and data related to the personal and professional lives of jazz artists" and development of related linked data tools and methods.

In chapter 2, Carl Stahmer (University of California, Davis) describes the migration of the renowned English Short Title Catalog (ESTC) from a MARC environment to one of linked data and the possibilities that migration opens up, especially in terms of involving the broader scholarly community in maintaining and enhancing ESTC metadata.

In chapter 3, Allison Jai O'Dell (University of Florida) reviews and reimagines library thesauri, metadata schemas, and information discovery, looking at how controlled vocabularies integrate library practice with linked data and exploring existing practices that are amenable to linked data, as well as areas for expansion of best practices in a linked data environment.

In chapter 4, Iker Huerga (Signifikance) and Michael Lauruhn (Elsevier Labs) examine linked data and authority control from the perspective of STM publishing, describing the role of authority control, identifiers, and vocabularies, including use of the Web Ontology Language (OWL) to add more formal semantics and the use of the SPARQL query language to create mappings between vocabularies.

In chapter 5, Carol Jean Godby (OCLC) describes OCLC's experiments with Schema.org as the foundation for a model of library resource description expressed as linked data, using 900 million catalog records accessible from WorldCat.org.

Godby reports that "OCLC's experiments have shown that Schema.org can be used to define a model . . . which can be expressed in a published standard with institutional backing and potential for widespread adoption."

In chapter 6, Sally McCallum (Library of Congress) relates the development of the Bibliographic Framework Initiative (BIBFRAME) data model, the linked data successor to the data model represented by the MARC 21 formats, describing the fundamental differences between MARC and BIBFRAME. BIBFRAME is designed to be particularly suited as an exchange format for bibliographic data created using Resource Description and Access. The Library of Congress implemented a BIBFRAME pilot in the third quarter of 2015.

As an addendum to the Godby and McCallum contributions (chapters 5 and 6), readers are directed to the executive summary by Godby and Ray Denenberg (Library of Congress) of a technical analysis of the relationship between the LC and OCLC models for library linked data.[6] The technical analysis itself will be released at a later date.

Notes

1. Tim Berners-Lee, James Hendler, and Ora Lassila, "The Semantic Web," *Scientific American* 284, no. 5 (2001): 34–43.
2. Tim Berners-Lee, "Is Your Linked Open Data 5 Star?" in Linked Data, www.w3.org/DesignIssues/LinkedData.html 2010.
3. Carole Jean Goble et al., "Incorporating Commercial and Private Data into an Open Linked Data Platform for Drug Discovery," *Lecture Notes in Computer Science* 8219 (2013): 65–80. doi: 10.1007/978–3–642–41338–4_5.
4. Heiko Paulheim, "What the Adoption of schema.org Tells about Linked Open Data," *JCEUR Workshop Proceedings* 1362 (2015): 84–90. http://ceur-ws.org/Vol-1362/PROFILES2015_paper6.pdf.
5. Seth van Hooland and Ruben Verborgh, *Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish Your Metadata* (Chicago: Neal-Schuman, 2014).
6. Carol Jean Godby and Ray Denenberg, *Common Ground: Exploring Compatibilities between the Linked Data Models of the Library of Congress and OCLC* (Dublin, Ohio: Library of Congress and OCLC Research, 2015), www.oclc.org/content/dam/research/publications/2015/oclcresearch-loc-linked-data-2015-a4.pdf.

Further Reading

Allemang, Dean, and Jim Hendler. *Semantic Web for the Working Ontologist.* Amsterdam: Morgan Kaufmann, 2011.

Antoniou, Grigoris, and Frank van Harmelen. *A Semantic Web Primer.* 3rd ed. Cambridge, Mass.: MIT Press, 2012.

Breitman, Karin K., Marco Antonio Casanova, and Walter Truszkowski. *Semantic Web: Concepts, Technologies and Applications.* London: Springer, 2007. doi: 10.1007/978–1–84628–710–7.

Coyle, Karen. *Linked Data Tools: Connecting on the Web.* Chicago: American Library Association, 2012. doi: 10.5860/ltr.48n4.

Godby, Carol Jean, Shenghui Wang, and Jeffrey K. Mixter. *Library Linked Data in the Cloud: OCLC's Experiments with New Models of Resources Description*. San Rafael, CA: Morgan & Claypool Publishers, 2015. doi: 10.2200/S00620ED1V01Y201412WBE012.

Heath, Tom, and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space.* San Rafael, CA: Morgan & Claypool Publishers, 2011. doi: 10.2200/S00334ED1V01Y201102WBE001.

Wood, David, Marsha Zaidman, and Luke Ruth. *Linked Data: Structured Data on the Web.* Shelter Island, NY: Manning, 2014.



LINKED OPEN DATA AND THE CULTURAL HERITAGE LANDSCAPE

Hilary K. Thorsen and M. Cristina Pattuelli

Linked open data is assuming an increasingly central role in how cultural institutions organize and share their data on the Web. The open and unified information environment that the LOD initiative is actively building has enormous potential for addressing some of the enduring issues in the field of cultural heritage, including data interoperability and integration. The "walled gardens" approach to data management in libraries, archives, and museums, as well as other cultural organizations, often has the unintended consequence of limiting access to and use of cultural heritage materials. LOD, on the other hand, with its principle of openness, offers an opportunity to expose data currently siloed in databases or institutional repositories and foster greater collaboration among cultural institutions across domains. Creating this expanded, even boundless scenario requires a radical revision of our approach to traditional data representation, curation, and discovery.

Libraries, museums, and archives have historically developed diverse and often divergent descriptive practices that are tailored to local forms of access and use of resources rooted in the analog world. The new representational and technological framework that linked data provides offers a unifying methodology for creating, sharing, and reusing data and metadata across domains. This model makes it possible for different communities to collaborate without the need for coordination, because the sharing of common representational conventions no longer requires a priori agreements on schemas. For example, vocabularies from different domains, when expressed in RDF, can be recombined and expanded through the mixing and matching of terms. In the broadest sense, LOD brings to libraries, archives, and museums a new way of thinking about the information object and its boundaries centered on entities and their relationships rather than documents. This makes it possible for atomic bits of content to be interlinked from heterogeneous sources, from the humanities to the sciences, in unprecedented ways, creating new and meaningful pathways to discovery and navigation.

Digital cultural heritage is becoming an integral part of an extended and expanded web of data through the work of Linked Open Data in Libraries, Archives, and Museums (LODLAM).[1] LODLAM, an international community of information professionals, researchers, and educators, is actively working to promote the convergence of cultural heritage communities through the implementation of linked open data principles and technology. Emerging in 2011, LODLAM has progressed from a small grassroots effort to a growing movement at the forefront of sowing the culture of linked data in cultural heritage communities through LOD research and implementation. The LODLAM community has expanded and evolved through a series of initiatives and events, from biannual international summits to regional meetings that bring together people from different cultural contexts and organizations to share case studies, experience, and ideas to advance LOD development in cultural heritage.

While we have only begun to envision the potential for cultural organizations to permeate the Web with open cultural heritage data, considerable efforts are being made towards devising appropriate practices and building effective applications needed to create and consume cultural LOD. The LOD landscape for cultural heritage is now not only extensive, but also highly diversified. Led by some of the world's most important national libraries, including the British Library, the Bibliothèque Nationale de France, and the Deutsche National Bibliothek, libraries have paved the way by contributing their legacy data to

the LOD ecosystem and making the LOD pool of data grow exponentially. To provide an overview of the range of LOD activities and to illustrate some of the challenges and opportunities LOD presents for the cultural heritage community, we have identified a representative sample of projects from the arts and humanities, ranging from large-scale resource aggregators to smaller experimental initiatives.

EUROPEANA

Perhaps the best-known linked data cultural heritage project is Europeana, an online platform launched in 2008 that aggregates cultural heritage data from 2,300 small and large libraries, archives, and museums in Europe.[2] Currently it contains descriptions and thumbnails for more than 33 million cultural objects that link to the source objects on the institutions' websites. With a single access point to cultural heritage data, users are not constrained by typical institutional silos, enabling them to search and browse across institutions, collections, formats, and countries situating these objects within a broader cultural context. As an open and authoritative source of cultural heritage data, Europeana began publishing a pilot subset of its data as LOD in 2011 and currently has 20 million cultural objects in its dataset.[3]

Implementing LOD at Europeana required overcoming a number of obstacles. One of the most difficult challenges was the mental shift required of cultural heritage institutions forced to consider how their collections integrate with and contribute to a wider information continuum.[4] Since openness of the data is key to Europeana's linked data vision, they needed to convince cultural heritage institutions to release their data. Their approach was to launch a highly successful online video campaign demonstrating how releasing data benefits users and generates interest among new users as the data is reused in novel ways. Since 2012, the open data released has grown from 2.4 million to 20 million objects.

Europeana has also had to deal with the extremely heterogeneous data from its 2,300 providers, which is described according to a wide variety of formats ranging from international standards to homegrown schemas.[5] The Europeana Data Model (EDM) was developed to make them interoperable. The EDM allows complex, hierarchically structured objects and entities to be represented while attaching provenance to each resource. It successfully

reuses and links to existing vocabularies, such as Dublin Core and SKOS, but has run into problems with provenance and licensing information, critical for users of cultural items and their providers. To associate provenance information with each object, Europeana combined a number of existing means to supply this necessary information as it waits for working groups to adopt a standard approach.

To stimulate and support reuse of its data, Europeana devotes a section of its website to professionals so librarians, archivists, and curators can exchange information, share their projects, and connect. A special Labs portal serves as a place of experimentation for developers to access Europeana's data through its application profile interface (API) and share their work. Serendip-o-matic, which helps users discover sources within Europeana and other cultural portals related to their research interests, and Storyana, which develops stories using Wikipedia articles and Europeana content, are just two examples of the kind of innovation Europeana's open approach has spawned.

DIGITAL PUBLIC LIBRARY OF AMERICA

The U.S. counterpart to Europeana is the Digital Public Library of America (DPLA).[6] The DPLA collects millions of resources from America's libraries, archives, and museums and makes them available through a single access point. The goal of the DPLA is to ensure all types of materials ranging from written text and artwork to scientific datasets and oral histories are freely available, discoverable, and usable. To accomplish this goal, DPLA delivers content through an online portal that can be searched and browsed in a variety of ways. It includes an interactive time line and map as well as a virtual bookshelf. Similar to Europeana, DPLA publishes its data through an API in J-SON-LD, a linked data serialization, so that the data can be used for new purposes, including apps.

DPLA prioritizes demonstrations of what can be done with its openly available linked data dataset. The general public can submit ideas for DPLA apps and tools, which are shared with the community. People can declare that they will work on something themselves or request someone else to develop it, making it easy for developers, programmers, and engineers to know what DPLA users would find most useful for their research. Once apps have been designed that use DPLA data or tools created that assist others in more easily making use of DPLA's data, they are showcased in the App Library.[7]

SOCIAL NETWORKS AND ARCHIVAL CONTEXT PROJECT

Based in the archival community, the Social Networks and Archival Context (SNAC) Project intends to create software and LOD resources to make connections between people and archival descriptions, authority files, and library catalogs.[8] Archival description has traditionally been restricted intellectually and functionally by mixing record and context descriptions. It is the aim of SNAC to facilitate archival research for users by improving their searches for people, families, and organizations and their understanding of the socio-historical context in which documents are situated by exposing the social connections between people and historical records. Archivists will further benefit with an increase in efficiency through the reuse of resource description.[9]

Data for SNAC is extracted from archival source descriptions and used to create records for persons, families, and organizations using the Encoded Archival Context-Corporate Bodies, Persons, and Families (EAC-CPF) standard. In the prototype SNAC interface, users can browse or search for records that contain the preferred form of name, alternate forms of name, occupations, topical subject headings, biographical entries, and links to archival collections and resources created by or related to the entity and other related named entities, which are used to construct a social network.[10] SNAC had 3.7 million entities from 4,000 repositories as of 2015.

SNAC faced significant challenges with identities, as multiple data sources made it possible for multiple URIs and records to be created for the same entity.[11] Given the vast number of records, the reconciliation of these names was and is a time-consuming and daunting process. Improvement in the quality of results has been further complicated by the heterogeneous nature of the archival descriptions, but it is hoped that future research and development will identify new automated solutions. In response, a national archival authorities program cooperative that describes and enhances authorities and manages and disambiguates identities is in development.[12]

REGESTA.EXE

The Italian company, regesta.exe, works on projects to improve the technologies used to make digital content available on the Web. They recently began experimenting with linked data on a couple of archival projects. The Repository for Linked Open Archival Data (RELOAD) project explores the methodology

needed to expose archival resources from various archival collections as linked data.[13] Using data from the Istituto dei Beni Culturali Regione Emilia Romagna and Archivio Centrale della Stato, it plans to become a central repository for archival descriptions that can be easily accessed and shared. In order to transform archival descriptions into linked data and to model classes and properties of archival data, it created an ontology, Ontology of Archival Description (OAD), using the Web Ontology Language (OWL). OAD was used to transform Encoded Archival Description (EAD) files to RDF/XML files using an XSL stylesheet. For records of people, corporate bodies, and families, the EAC-CPF Ontology was used to transform the EAC-CPF records to linked data.[14] The archival descriptions can be searched, browsed via facets, or viewed in graph form.

A second archival project involves the records of the Contemporary Jewish Documentation Center Foundation (CDEC). It holds the largest collection of Holocaust documents in Italy. Regesta.exe is working to create a digital library with LOD as the information structure.[15] As part of this goal, the Shoah Ontology was created to link CDEC's database of 8,000 names of Jewish victims of the Holocaust in Italy with resources and names in other archives and linked open datasets.[16] The ontology relates the "persecution" class, which describes arrest, imprisonment, and deportation, and the "person" class, which contains the biographical information of the victims. While the ontology describes the persecution and deportation of Jews in Italy between 1943 and 1945, it is hoped that it will eventually expand to represent the Shoah as a whole.

AMSTERDAM MUSEUM

Museum data has traditionally been the least exposed of cultural heritage data. Licensing and provenance information are extremely important for museums for maintaining their credibility and trustworthiness within the art world, so museum data must be perfect before it can be published online. However, more recently, museums have begun to experiment with LOD to expose their data as a means of increasing the visibility of their collections on the Web. The Amsterdam Museum was an early adopter of LOD. Because of physical limitations, it is only possible to display 20 percent of the museum's collection at any given time, so in 2010 the museum made its collection available online under a Creative Commons license.[17] The collection can be browsed online, with its metadata retrievable through an API. Utilizing the Europeana Data

Model, the data has been converted into a five-star Linked Data set, which means the HTTP URIs in the dataset provide useful information and link to other URIs to enable further discovery.

Often when an institution's data is mapped to a standard that increases its interoperability, it loses some of the richness and detail that may have been present in the original. Fortunately, the museum did not lose this robust descriptive content because the museum's thesaurus of 28,000 concepts and person authority file with almost 67,000 persons in addition to the over 70,000 object descriptions were converted to preserve internal links within the dataset.[18] Links to external datasets included GeoNames for geographical concepts, the Dutch Art and Architecture thesaurus for non-geographical concepts, and the Getty Union List of Artist Names (ULAN) for people. The museum's LOD dataset is available through Europeana and will serve as a model within Europeana for future metadata conversions.

SMITHSONIAN MUSEUM OF AMERICAN ART

More recently, the Smithsonian Museum of American Art converted its dataset of 41,000 objects and 8,000 artists to linked data.[19] Because much of the data museums collect and retain must remain private, especially donor and financial information, publishing the data openly can be difficult. For this project, the Smithsonian converted only the data for the collection that is already published on the website. Because of its complicated and rich hierarchical metadata, choosing the ontologies for mapping the data was especially challenging. Ultimately, the museum selected the Dublin Core, FOAF, SKOS, RDA Group Elements, schema.org, and EDM ontologies for mapping their data. Data was then linked to external sources, including DBPedia for artist biographies, the *New York Times* for obituaries, exhibition and publication reviews, and auction results, as well as the Getty ULAN identifiers.[20]

As a result of the process, the museum developed tools that anyone can use to convert their data to linked data. One is Karma, which uses an automated process to integrate data from various sources based on an ontology and publish it in RDF.[21] Data can then be linked externally with a high level of accuracy and experts can review and curate links to ensure the data meets the high standards that museums require. With its LOD, the museum intends to continue developing new applications, including a virtual museum where works held at various museums are brought together as well as a tool to create multimedia

stories about art. The museum will also lead the American Art Collaborative, a consortium of fourteen American museums that will use LOD to bring greater access, research and scholarship, and collaboration to American art.[22]

LINKED JAZZ

Similar to other cultural heritage projects, the authors' project, Linked Jazz, has evolved in an experimental way and through an iterative process, facing some of the challenges common to this pioneering phase of LOD development. For this reason, Linked Jazz provides a representative case study grounded in a real-world application scenario. Given the familiarity with the project, we will provide an extensive overview of Linked Jazz, highlighting its various components and phases of development to share outcomes as well as lessons learned.

Linked Jazz is an ongoing project based at Pratt Institute's School of Information and Library Science that explores the application of LOD principles and technology to digital cultural heritage materials with the goal of enhancing their discovery, visibility, and ultimately interpretation and use.[23] Started in 2011 thanks to an OCLC/ALISE Research Grant, the project has focused on using oral histories from digital archives of jazz history to provide jazz researchers, educators, and the general public with new ways to discover and engage with archival content.

The immediate goal of the project was to leverage LOD technology to represent and visualize the complex network of relationships held among jazz musicians as described in oral histories. The unprecedented availability of digitized text offers researchers a unique opportunity to conduct network analysis on a large scale. Using LOD technology to represent network data from primary sources is a novel approach that offers remarkable advantages to researchers, including making this data machine-processable and freely available and reusable in various contexts and for different applications and data services. The set of network data can then be interlinked with external sources that exist in the open linked data cloud so that new associations between data and between resources can be identified. Applying LOD to archival documents and special collections makes information that is otherwise contained, if not hidden, in primary sources and typically scattered across collections and institutions more easily available. Researchers can access and navigate this content in a unified and integrated way that is useful for enhancing discovery and facilitating analysis. Even though the Web has made archival content more accessible, the

research process required for locating resources and collecting data continues to often be hard and labor-intensive. For example, the task of identifying and presenting the relationships that exist within a social group such as the jazz community as a cohesive set of data would require digging into several digital libraries and archives and sifting through documents that even experienced researchers can find difficult to locate.

This approach, however, required the development of a whole array of methods and tools needed to create linked data from textual resources and to expose and visualize it. The project has progressed in an iterative way through a number of phases including

1. the creation of RDF triples from textual documents;
2. the curation of data to ensure data quality;
3. the visualization of the dataset; and
4. the semantic enrichment of the data.

While most of the current LOD projects in the cultural heritage domain are concerned with converting legacy metadata into RDF triples (for instance, cultural heritage data aggregators Europeana and DPLA as well as single-institution projects including the Smithsonian Museum of American Art), Linked Jazz focuses on the content of the document as the source of linked data rather than its description. The possibility to leverage unstructured textual data, in addition to bibliographic metadata, can augment the reach of LOD technology exponentially. It requires, however, a different development process and a suitable set of tools.

The first step was to perform the extraction of proper names of musicians from interview transcripts and associate these names in order to create a social network. The underlying assumption behind this method is such that if the subject of an interview mentions an individual, we can infer that some type of connection, even a very basic familiarity, should exist with that person. To process oral histories, the *Transcript Analyzer* was developed to perform named entity recognition (NER) to locate instances of personal names in text. More specifically, this tool extracts proper names and identifies basic relationships between people based on their mention in the text. See figure 1.1.[24]

To detect proper names the Transcript Analyzer relies on a directory of jazz musician names that serves to support the matching of the name instances. Such a domain-specific name vocabulary had to be created from scratch and was derived from DBpedia. While the details of its creation are described elsewhere,

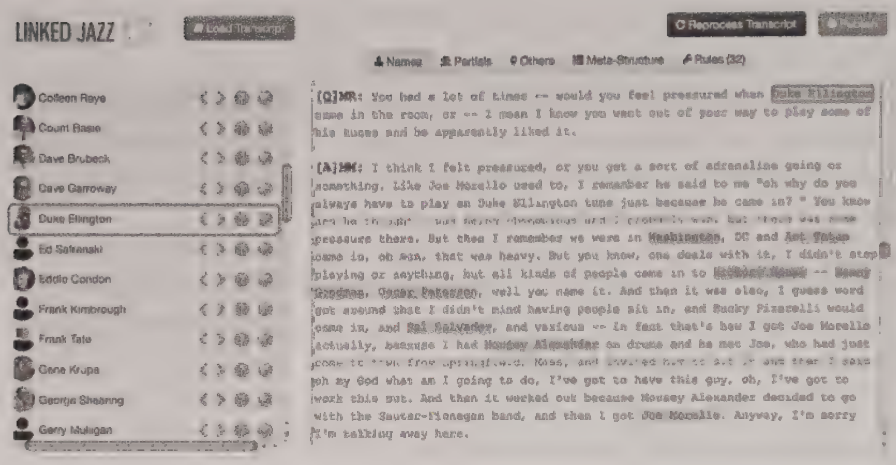

Figure 1.1 | **The Transcript Analyzer tool**

it is worthwhile to mention the need to address issues of data quality, as is often the case when dealing with large user-generated datasets.[25] While DBpedia is a rich source of linked data resulting from the extraction of structured data from Wikipedia, the data also contains a good deal of noise. Proper names of artists, with their array of aliases and stage names, are particularly prone to inconsistency and ambiguity, including homonyms, different spellings, and misspellings. For example, *Duke Ellington* presents six alternate names including spelling variants and nicknames—*Ellington, Edward Kennedy; Ellington, Diuk; Turner, Joe; Greer, Sonny; Duke, Obie; and Ellington, Obie Duke*—as per the Library of Congress name authority record.[26]

Curatorial work is an essential component in LOD development that is required to address various issues of data quality.[27] To make our jazz name vocabulary as inclusive and accurate as possible, we mapped individuals' URIs onto the Library of Congress Name Authority File and Virtual International Authority File (VIAF) to disambiguate and enrich the data with name variants from authority records. Figure 1.2 gives an integrated view of Duke Ellington's preferred and alternate names included in the Linked Jazz name directory.

The mapping process was supported by a dedicated application, now in the process of being redesigned, that delivered results grouped by degrees of confidence.[28] See figure 1.3. A perfect match is counted when both name strings and corresponding birth and death dates overlap. Results with a lower level of confidence can then be manually revised to disambiguate homonyms as well as to validate or remove uncertain matches.

Preferred Name	<skos:prefLabel>	Ellington, Duke, 1899-1974
Alternate Name	<skos:altLabel>	Ellington, Edward Kennedy, 1899-1974
Alternate Name	<skos:altLabel>	Ellington, Diuk, 1899-1974
Alternate Name	<skos:altLabel>	Turner, Joe, 1899-1974
Alternate Name	<skos:altLabel>	Greer, Sonny, 1899-1974
Alternate Name	<skos:altLabel>	Ellington, Obie Duke, 1889-1974
Alternate Name	<skos:altLabel>	Duke, Obie, 1889-1974

Figure 1.2 | **Duke Ellington's preferred and alternate names included in the Linked Jazz name directory**

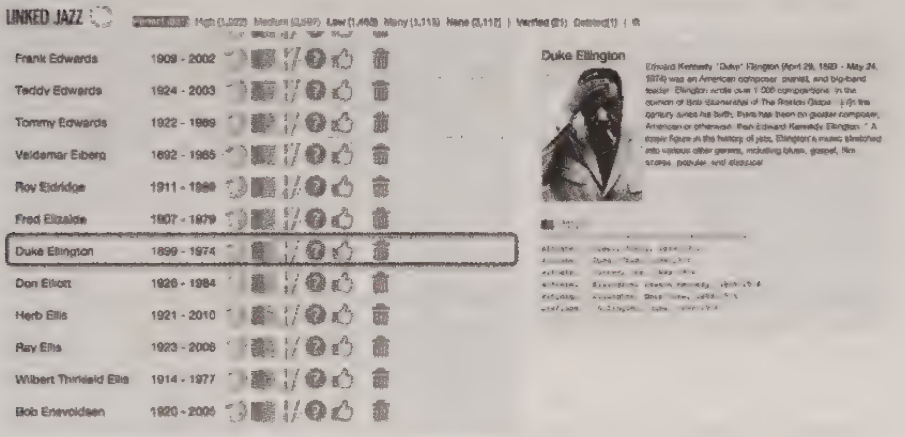

Figure 1.3 | **The name mapping and curation tool**

More than fifty digitized interview transcripts from special collections and jazz archives in the United States have now been processed through the Transcript Analyzer resulting in a dataset of approximately 15,000 RDF triples.[29] Each triple represents a connection between musicians expressed through the predicate rel:knowsOf, shown here:

<http://linkedjazz.org/resource/Stanley_Kay>
<http://purl.org/vocab/relationship/knowsOf>
<http://linkedjazz.org/resource/Duke_Ellington>.

It is often noted that the benefits of linked data cannot be conveyed in a vacuum. As an effective way to make the Linked Jazz data easily accessible and

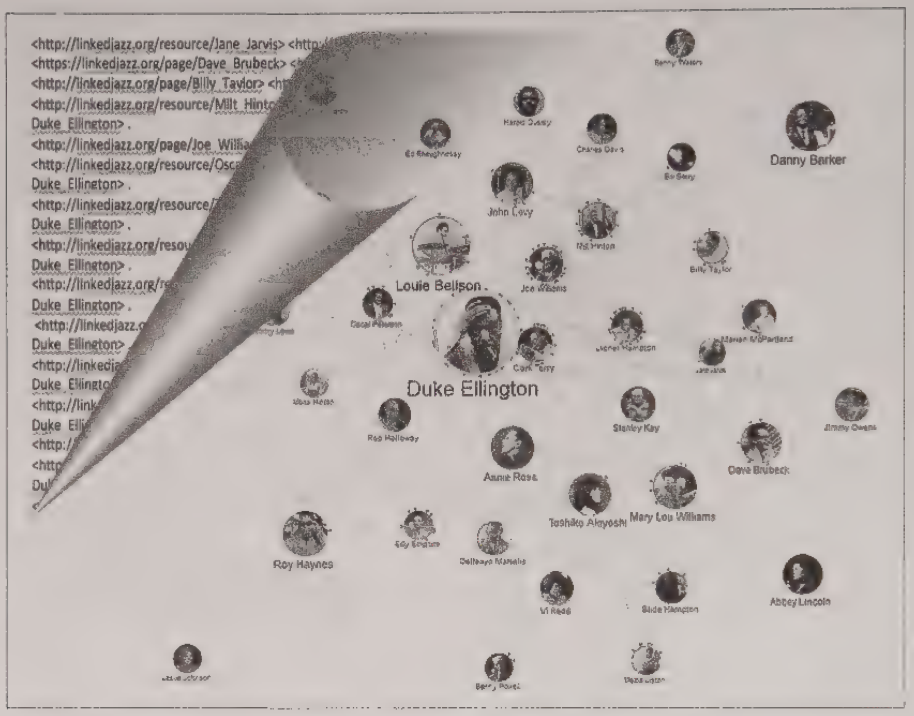

Figure 1.4 | **The RDF triples power the network visualization tool to reveal Duke Ellington's social network**

usable, we developed a visualization tool that displays the web of relationships existing in our dataset.[30] See figure 1.4. The nodes in the network represent jazz artists and the edges denote their relations. Nodes and edges are labeled with a Unique Resource Identifier (URI) as the application is driven by our linked open dataset. Figure 1.5 provides a sample of RDF triples representing the relationships between musicians at the back end of the network visualization tool.

This is a highly interactive tool that allows users to explore and analyze the network through different views. For example, users can navigate the network dynamically by drilling down to a musician's personal circles or ego-network. Each musician-node is associated with their biographical sketch derived from Wikipedia and even a sample of their work through an audio clip. Individuals can also be clustered into subgroups based on the relationships they share, or selected from a list to verify whether a connection between them exists and reveal other shared connections. See figure 1.6. Another distinctive feature of the tool is the possibility to easily and directly review the source of the connection

<http://linkedjazz.org/resource/Jane_Jarvis><http://purl.org/vocab/relationship/knowsOf><http://linkedjazz.org/resource/Duke_Ellington>.

<https://linkedjazz.org/page/Dave_Brubeck><http://purl.org/vocab/relationship/knowsOf><http://linkedjazz.org/resource/Duke_Ellington>.

<http://linkedjazz.org/page/Billy_Taylor><http://purl.org/vocab/relationship/knowsOf><http://linkedjazz.org/resource/Duke_Ellington>.

<http://linkedjazz.org/resource/Milt_Hinton><http://purl.org/vocab/relationship/knowsOf><http://linkedjazz.org/resource/Duke_Ellington>.

<http://linkedjazz.org/page/Joe_Williams><http://purl.org/vocab/relationship/knowsOf><http://linkedjazz.org/resource/Duke_Ellington>.

<http://linkedjazz.org/resource/Oscar_Peterson><http://purl.org/vocab/relationship/knowsOf><http://linkedjazz.org/resource/Duke_Ellington>.

Figure 1.5 | **Sample of RDF triples that drive the network visualization tool**

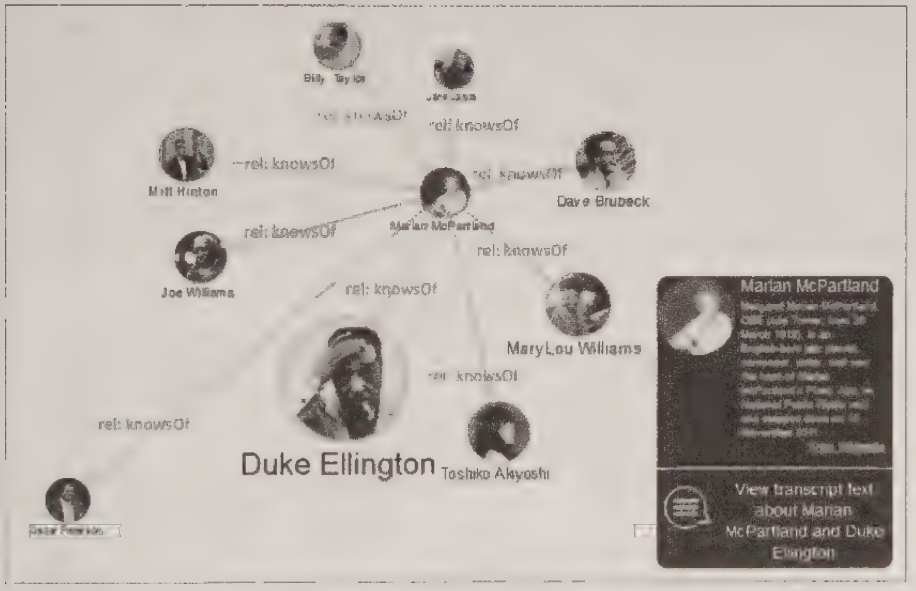

Figure 1.6 | **A view of the shared relationships between Duke Ellington and Marian McPartland**

by mousing-over an edge. The interview passage that mentions the relationship pops up providing context to the data and provenance.

The Linked Jazz network reveals how the jazz community is highly interconnected. As our dataset continues to grow, the size and accuracy of the social

network will also increase, supporting more comprehensive analyses of its structure and dynamics. To dig deeper into the nuanced nature of the human relationships expressed by a social network, a whole new set of methods and tools are needed to be able to determine the specific meaning of these relationships and assign them granular semantics.

Traditional natural language processing (NLP) methods, such as the ones employed in our Transcript Analyzer, are of limited help for identifying the semantics of human relationships with its degree of granularity and nuances. In general, performing complex semantic analysis of text through automated methods is computationally extremely arduous. To classify the relationships detected with the Transcript Analyzer, our automated approach was complemented by a human-driven one based on crowd-sourcing. To this end, we developed a tool, *Linked Jazz 52nd Street,* which engages jazz experts and enthusiasts with helping to interpret the relationships in the transcripts.[31] This is a web-based application that asks contributors to label the relationships between two jazz musicians by choosing from a list of options. See figure 1.7. The pool of possible relationships is derived from well-established RDF vocabularies and the user's selection is automatically converted into an RDF triple.

To compile the list of relationships, suitable RDF vocabularies were identified including the Relationship vocabulary, FOAF, and the Music Ontology.[32] Candidate relationships were selected by analyzing and mapping person-related predicates (Pattuelli, 2011).[33] The final set covers lower degrees of personal closeness (e.g., knows_of, has_met) as well as deeper levels of social

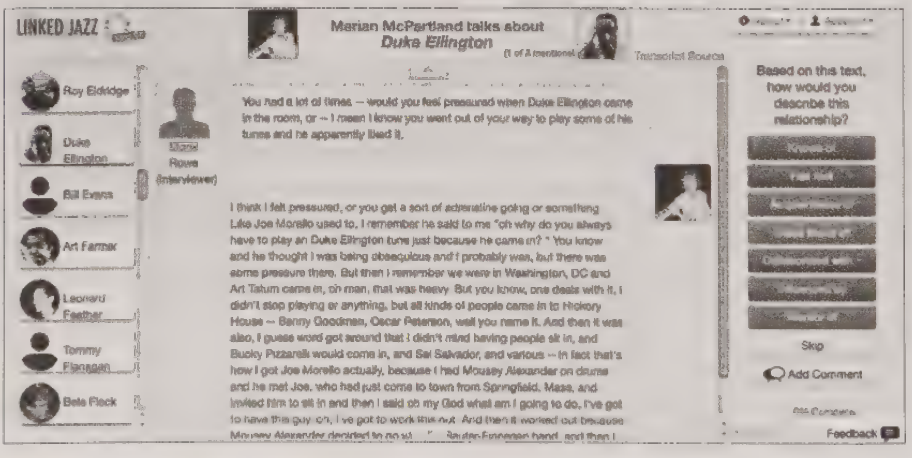

Figure 1.7 | **The Linked Jazz 52nd Street classifier interface**

and professional ties (e.g., collaborated_with, influenced_by, and mentor_of). Aligning vocabulary terms helps to ensure that terms in existing vocabularies are reused where applicable and that new terms are consistently integrated with the existing ones. Jazz experts were also consulted to determine which of these relationships would be most appropriate and useful to them and the broader community of users of jazz archives.

Linked Jazz 52nd Street participants are presented with segments of the interview transcript where a musician is mentioned and are prompted to choose the most suitable label, as discussed earlier. The output of these tasks is in the form of RDF triples that feed back into the project's set of linked data and thus will semantically enhance the social network. See figures 1.8 and 1.9.

In addition to supporting a sophisticated data service specific to the project, this tool has the potential to engage the community of jazz researchers and enthusiasts as well as the general public with primary sources and archival collections, increasing their visibility and use. Designing crowd-sourcing tools is challenging, because many complex elements need to be considered to make the system successful and sustainable. For example, different types of motivational factors are at the root of active and continued engagement by contributors.[34]

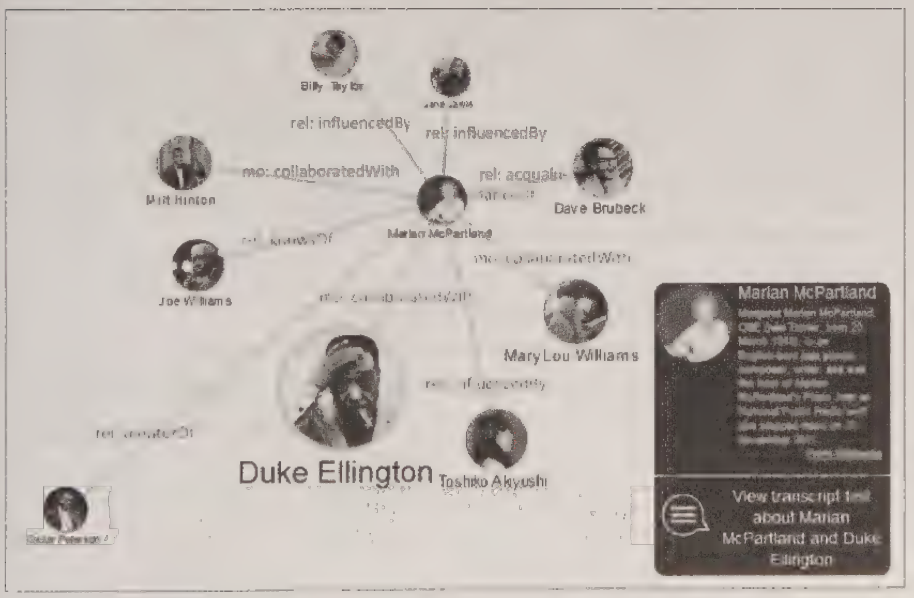

Figure 1.8 | **Example of shared relationships between Duke Ellington and Marian McPartland semantically enhanced via crowd-sourcing**

<http://linkedjazz.org/resource/Marian_McPartland><http://purl.org/vocab/relationship/influencedBy><http://linkedjazz.org/resource/Jane_Jarvis>.

<http://linkedjazz.org/resource/Marian_McPartland><http://purl.org/vocab/relationship/acquaintanceOf><https://linkedjazz.org/resource/Dave_Brubeck>.

<http://linkedjazz.org/resource/Marian_McPartland><http://purl.org/vocab/relationship/influencedBy><http://linkedjazz.org/resource/Billy_Taylor>.

<http://linkedjazz.org/resource/Marian_McPartland><http://purl.org/ontology/mo/collaborated_with><http://linkedjazz.org/resource/Milt_Hinton>.

<http://linkedjazz.org/resource/Marian_McPartland><http://purl.org/vocab/relationship/knowsOf><http://linkedjazz.org/resource/Joe_Williams>.

<http://linkedjazz.org/resource/Marian_McPartland><http://purl.org/vocab/relationship/mentorOf><http://linkedjazz.org/resource/Oscar_Peterson>.

<http://linkedjazz.org/resource/Marian_McPartland><http://purl.org/ontology/mo/collaborated_with><http://linkedjazz.org/resource/Duke_Ellington>.

<http://linkedjazz.org/resource/Marian_McPartland><http://purl.org/vocab/relationship/influencedBy><http://linkedjazz.org/resource/Toshiko_Akiyoshi>.

<http://linkedjazz.org/resource/Marian_McPartland><http://purl.org/ontology/mo/collaborated_with><http://linkedjazz.org/resource/Mary_Lou_Williams>.

<http://linkedjazz.org/resource/Jane_Jarvis><http://purl.org/vocab/relationship/knowsOf><http://linkedjazz.org/resource/Duke_Ellington>.

<https://linkedjazz.org/page/Dave_Brubeck><http://purl.org/vocab/relationship/mentorOf><http://linkedjazz.org/resource/Duke_Ellington>.

<http://linkedjazz.org/page/Billy_Taylor><http://purl.org/vocab/relationship/knowsOf><http://linkedjazz.org/resource/Duke_Ellington>.

<http://linkedjazz.org/resource/Milt_Hinton><http://purl.org/vocab/relationship/knowsOf><http://linkedjazz.org/resource/Duke_Ellington>.

<http://linkedjazz.org/page/Joe_Williams><http://purl.org/vocab/relationship/knowsOf><http://linkedjazz.org/resource/Duke_Ellington>.

<http://linkedjazz.org/resource/Oscar_Peterson><http://purl.org/vocab/relationship/mentorOf><http://linkedjazz.org/resource/Duke_Ellington>.

<http://linkedjazz.org/resource/Toshiko_Akiyoshi><http://purl.org/vocab/relationship/knowsOf><http://linkedjazz.org/resource/Duke_Ellington>.

<http://linkedjazz.org/resource/Mary_Lou_Williams><http://purl.org/ontology/mo/collaborated_with><http://linkedjazz.org/resource/Duke_Ellington>.

Figure 1.9 | **RDF triples of the shared relationships between Duke Ellington and Marian McPartland**

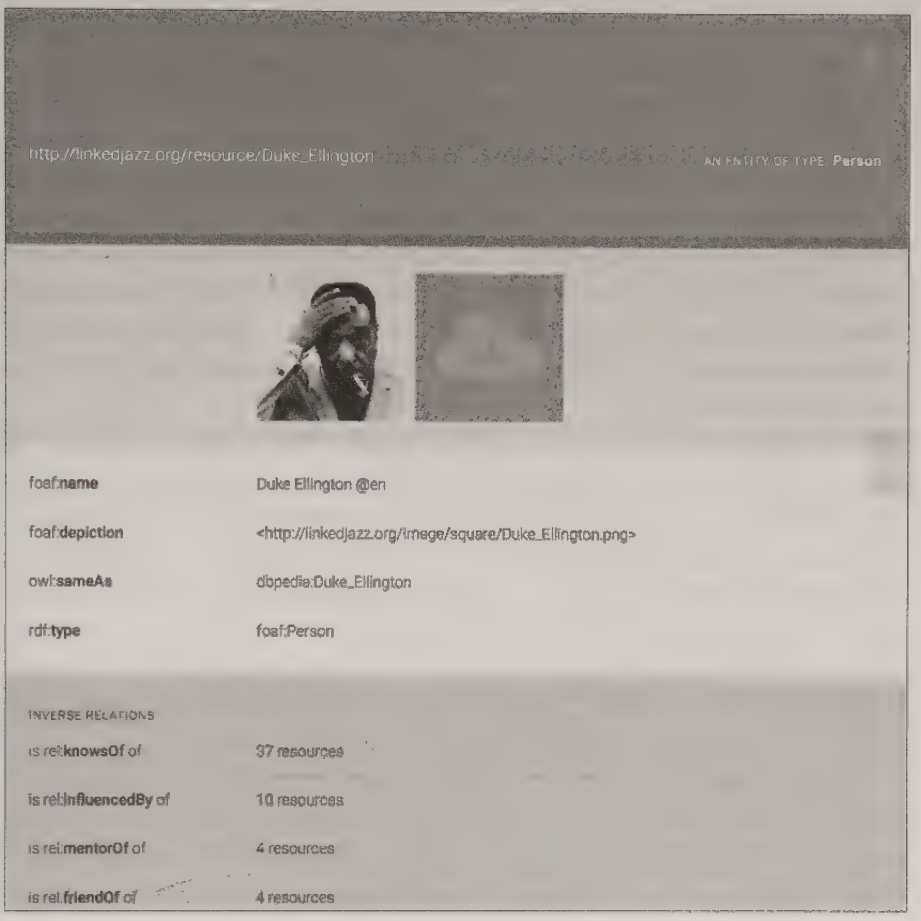

Figure 1.10 | **Duke Ellington's LodView page**

A number of features were included in the design of Linked Jazz 52nd Street in order to maximize crowd participation over time. A dynamic ego network and a progress bar were among the features introduced to provide immediate visual feedback on the user's progress with the intent of making their work recognizable and transparent. New iterations of testing and refinements are planned to further the development of this tool. As of today, Linked Jazz 52nd Street has been used by approximately 500 registered users and has resulted in the classification of more than 15,000 relationships.

The Linked Jazz resources, essentially person entities and their predicates, can also be viewed with LodView, an open-source publishing platform for LOD datasets.[35] See figure 1.10. LodView displays LOD resources as web

pages, in line with one of the key principles of linked data that requires that each URI is "dereferenceable," meaning that it resolves to a web page that can be looked up by humans and machines.

In line with Linked Open Data's principles of data sharing and reuse and best practices, the Linked Jazz dataset is accessible via an API, as well as a SPARQL endpoint.[36] SPARQL Protocol and RDF Query Language (SPARQL) is a query language for RDF. Queries to the Linked Jazz dataset through its SPARQL end point would, for example, return all of the relationships between jazz musicians, the relationships between specific musicians, all of the relationships of a specific musician, or all of the musicians with a specific type of relationship.

Linked Jazz is entering a new phase in which the current dataset will be interlinked with external LOD sources to further increase its richness. Mashups with triples representing discographies, musician photographs, performances, and library holdings will expand traditional access to archival content with the ultimate goal of making it part of a global and integrated discovery environment. Bringing together a broad range of spatial and temporal data, including time periods, dates, events, and geographic locations, and music-specific data, including professional roles, instruments, recordings, and music venues, will offer new opportunities for information discovery and analysis.

An example is offered by the first data mashup which is under way using data from two of Tulane University's jazz photography collections, the Hogan Jazz Archive Photography collection and the Ralston Crawford Collection of Jazz Photography. Combined, the collections contain more than 1,700 images in which more than 2,700 individuals are depicted, 681 being unique individuals. Based on the assumption that people depicted in the same photo are likely to know each other, triples are created that express exactly this type of relationship through the predicate foaf:knows.

In this instance, our source of data was the metadata describing the photographs. As is to be expected with most cultural heritage resources, Tulane's metadata did not come in the form of linked data and thus it needed to be converted into LOD. The process consisted of cleaning and normalizing the data, including mapping the proper names of the musicians depicted to name authorities. As an example of how the process works, one of the photographs in the Hogan Jazz Archive's Collection depicts Duke Ellington and Louie Armstrong at a party held in their honor along with other guests in Chicago on February 14, 1935. In the original metadata record, specifically in the subject field, their names appear as Ellington, Duke and Armstrong, Louis, 1901–1971.

By matching these names to DBpedia, triples were created that, for example, would describe Ellington as a foaf:person with a foaf:name associated with it. Ellington was then linked to the photograph using the predicate foaf:depiction and to Armstrong using foaf:knows. Finally, the date and the place the photograph was taken in were also linked to the photograph using the predicates dc:created and dc:spatial. Figure 1.11 shows the triple structure in the left column and instances from the specific example concerning Ellington and Armstrong in the right.

As the phase of data conversion and triple creation is concluded, the next step will be to integrate the Tulane data with the existing dataset and enhance

<personURI> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person>.	<http://dbpedia.org/resource/Duke_Ellington> <http://www.w3.org/1999/02/22-rdf-syntax-ns #type> <http://xmlns.com/foaf/0.1/Person>.
<personURI> <http://xmlns.comfoaf/0.1/name> "First Last"@en.	<http://dbpedia.org/resource/Duke_Ellington> <http://xmlns.com/foaf/0.1/name> "Duke Ellington"@en.
<personURI> <http://xmlns.com/foaf/0.1/depiction> <photoURI>.	<http://dbpedia.org/resource/Duke_Ellington> <http://xmlns.com/foaf/0.1/depiction> <http://louisdl.louislibraries.org/cdm/singleitem/collection/p16313coll33/id/1689/rec/1>.
<person1URI> <http://xmlns.com/foaf/0.1/knows> <person2URI>.	<http://dbpedia.org/resource/Duke_Ellington> <http://xmlns.com/foaf/0.1/knows> <http://dbpedia.org/resource/Louie_Armstrong>.
<photoURI> <http://purl.org/dc/terms/created> "YYYY-MM-DD"^^<http://www.w3.org/2001/XMLSchema #date>.	<http://louisdl.louislibraries.org/cdm/singleitem/collection/p16313coll33/id/1689/rec/1> <http://purl.org/dc/terms/created> "1935-02-14"^^<http://www.w3.org/2001/XMLSchema#date>.
<photoURI> <http://purl.org/dc/terms/spatial> <geonamesURI>.	<http://louisdl.louislibraries.org/cdm/singleitem/collection/p16313coll33/id/1689/rec/1> <http://purl.org/dc/terms/spatial> <http://www.geonames.org/4887398/chicago.html>.

Figure 1.11 | **RDF triples representing Duke Ellington and Louis Armstrong**

the Linked Jazz dataset with links to the actual photographs and their source archives.[37]

The suite of tools developed for Linked Jazz is freely available on GitHub.[38] Following prototyping development methodology, these tools will continue to be improved through iterations of testing. Plans are under way to continue the development of the chain of tools to reach their full implementation. The goal is to increase their versatility so that they can be applied to different contexts of use and maximize their usability, considering librarians and archivists as the primary users. Lessons learned are shared through documentation intended to contribute to establishing best practices in LOD development for libraries, archives, and museums.

Notes

1. "LODLAM: Linked Open Data in Libraries, Archives, and Museums," http://lodlam.net.
2. Europeana, http://europeana.eu.
3. "Europeana Linked Open Data," Europeana Labs, http://labs.europeana.eu/api/linked-open-data/introduction/.
4. Cesare Concordia, Stefan Gradmann, and Siberd Siebinga, "Not Just Another Portal, Not Just Another Digital Library: A Portrait of Europeana as an Application Program Interface," *IFLA Journal* 36 (2010): 61.
5. Antoine Isaac, Robina Clayphan, and Bernhard Haslhofer, "Europeana: Moving to Linked Data," *Information Standards Quarterly* 24, no. 213 (2012): 35.
6. Digital Public Library of America, http://dp.la.
7. "Apps Library," Digital Public Library of America, http://dp.la/apps.
8. "SNAC: Social Networks and Archival Context," http://socialarchive.iath.virginia.edu.
9. Daniel V. Pitti, "Social Networks and Archival Context Project: Archival Authority Control," Social Networks and Archival Context Project, http://ica2012.ica.org/files/pdf/Full%20papers_%20upload/ica12Fina100310.pdf.
10. Ray R. Larson and Krishna Janakiraman, "Connecting Archival Collections: The Social Networks and Archival Context Project," *Theory and Practice of Digital Libraries* 2011 (2011): 5; SNAC: Social Networks and Archival Context, http://socialarchive.iath.virginia.edu/scope.html.
11. Ibid., 6.
12. Pitti, "Social Networks and Archival Context Project," http://ica2012.ica.org/files/pdf/Full%20papers_%20upload/ica12Fina100310.pdf.
13. "Il progetto," RELOAD, http://labs.regesta.com/progettoReload/en.
14. "L'ontologia EAC-CPF," RELOAD, http://labs.regesta.com/progettoReload/en/lontologia-eac-cpf/; "OAD Ontology," RELOAD, http://labs.regesta.com/progettoReload/en/oad-ontology/.

15. "Dati," Contemporary Jewish Documentation Center Foundation, http://dati.cdec.it/indiceEN.html.
16. "Shoa Vocabulary Specification Beta Version," Contemporary Jewish Documentation Center Foundation, http://dati.cdec.it/lod/shoah/reference-document.html.
17. Victor de Boer Jan Wielemaker, Judith van Gent, Marijke Oosterbroek, Michiel Hildebrand, Antoine Isaac, Jacco van Ossenbrugen, and Guus Schreiber, "Amsterdam Museum Linked Open Data," *Semantic Web* 4, no. 3 (2013): 1.
18. Ibid., 2.
19. Pedro Szekely, Craig A. Knoblock, Fengyu Yang, Xuming Zhu, Eleanor E. Fink, Rachel Allen, and Georgina Goodlander, "Connecting the Smithsonian American Art Museum to the Linked Data Cloud," *Extended Semantic Web Conference* 2013 (2013): 593.
20. Ibid., 606.
21. "Karma: A Data Integration Tool," www.isi.edu/integration/karma/.
22. "Linked Data: Smithsonian American Art Museum Announces Formation of American Art Collaborative," *Library Journal*, www.infodocket.com/2015/02/03/linked-data-smithsonian-american-art-museum-announces-formation-of-american-art-collaborative/.
23. Linked Jazz, http://linkedjazz.org.
24. "Transcript Analyzer," Linked Jazz, http://linkedjazz.org/tools/transcript-analyzer/.
25. M. Cristina Pattuelli, Matt Miller, Leanora Lange, Sean Fitzell and Carolyn Li-Madeo, "Crafting Linked Open Data for Cultural Heritage: Mapping and Curation Tools for the Linked Jazz Project," *Code4Lib Journal* no. 21, http://journal.code41ib.org/articles/8670.
26. "Ellington, Duke, 1899–1974," Library of Congress Authorities, http://authorities.loc.gov/cgi-bin/Pwebrecon.cgi?AuthRecID=2337811&v1=1&HC=9&SEQ=20140929170541&PID=Vt5H8Yq48ZNcd22rCN_hTOC9JlZGf.
27. 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), 71–72.
28. "Mapping," Linked Jazz, http://linkedjazz.org/mapping/.
29. Sources of interviews include Rutgers University, Hamilton College, the Smithsonian Institution, University of Michigan, and UCLA (University of California, Los Angeles).
30. "Linked Jazz Network," Linked Jazz, http://linkedjazz.org/network/.
31. "52nd Street," Linked Jazz, http://linkedjazz.org/52ndStreet/.
32. "Relationship: A Vocabulary for Describing Relationships between People," Relationship, http://vocab.org/relationship/.html; "FOAF Vocabulary Specification 0.99," FOAF, http://xmlns.com/foaf/spec/; "MO: The Music Ontology," Music Ontology, http://musicontology.com.

33. M. C. Pattuelli, "Mapping People-Centered Properties for Linked Open Data," *Knowledge Organization* 38, no. 4 (2011): 352–59.
34. Tim Causer, Justin Tonra, and Valerie Wallace, "Transcription Maximized; Expense Minimized? Crowdsourcing and Editing *The Collected Works of Jeremy Bentham*," *Literary and Linguistic Computing* 27, no. 2 (2012): 127.
35. "https://linkedjazz.org/page/Duke_Ellington," Linked Jazz LodView, https://linkedjazz.org/page/Duke_Ellington.
36. "Linked Jazz API Documentation," Linked Jazz, http://linkedjazz.org/api/.
37. W. J. Levay, "A Semantic Makeover for CMS Data," Presentation at Code4Lib conference, February 10, 2015, Portland, Oregon, https://www.youtube.com/watch?v=gCfpQgXcpTE&feature=youtu.be&t=5m18s.
38. "Matt Miller," GitHub, https://github.com/thisismattmiller?tab=repositories.