Linked Data Applications for Library Collections

Elk Oswood
oswoo003@umn.edu
Division of Science and Mathematics
University of Minnesota, Morris
Morris, Minnesota, USA

Abstract

A discussion of Linked Data and its potential benefits for the Library Sciences. There is a brief overview of what Linked Data is, along with relevant definitions for related web technologies. Then, through a series of papers detailing potential uses of Linked Data in library archives, there is an exploration of solutions, benefits, and the future of the applications.

Keywords: Linked Data

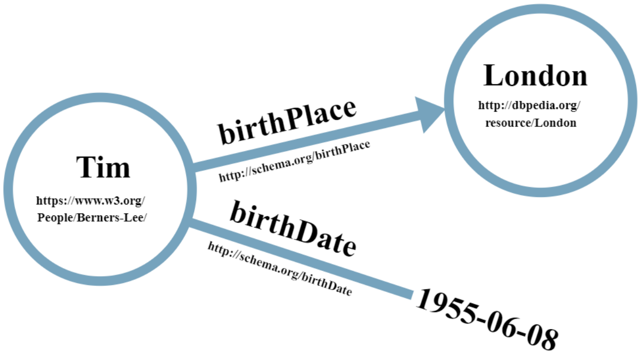

Figure 1. Temp Example of Linked Data Structure

1 Introduction

Information professionals, such as librarians, have moved beyond managing only physical collections and records. With the addition of digital collections comes new opportunities for library collections to both be exposed to a greater audience and to facilitate easier access in search queries. One way to take advantage of the opportunities that come with web based databases of collections is to use Linked Data.

Without utilizing Linked Data, libraries are limiting their collections' reach on the Web. Linked Data is a set of guideline for publishing data on the web. This gives data a standard format that is designed to be both machine readable and human user friendly. This allows a greater scope for searching and organizing information. In addition to making their own collections more easily available on the web, libraries would also be able to search more web pages and external catalogues, which enhances librarian and patron experience.

Furthermore, by utilizing a type of Linked Data called Linked Open Data, libraries would also be able to keep information free to all users with an Internet connection. The concept of Linked Open Data is to make all data free from paywalls and copyright claims, which opens up more of the library collection's information and data to searches and researchers.

In this paper, I first define what Linked Data is, along with relevant technologies. There is more than one way to apply Linked Data to a library collection. Then I discuss the studies of two different research groups and their respective applications of Linked Data for library use. Finally, I discuss the benefits and difficulties of using Linked Data.

2 Background

2.1 Linked Data

The Semantic Web project is an extension of the current World Wide Web. Its goal is to make Internet data machine-readable [1]. When the internet is machine readable, searches on the web are easier and more accurate. At the 2006 World Wide Web Consortium (W3C), Tim Berners-Lee first spoke about Linked Data. Linked Data is a set of practices that dictate the publishing and interlinking of structured data [4].

Web technologies such as Hypertext Transfer Protocol (HTTP), Uniform Resource Identifiers (URIs), and Resource Description Framework (RDF) are used in the creation of Linked Data. While previous uses of these technologies served web pages for human readers, with Linked Data these technologies are extended in a way that allows machines to read web page URIs automatically and extract important information such as the name of the literary work, the genre of the work, or the author's name.

HTTP works as a request-response protocol between a client and server. A client submits an HTTP request message to the server. The server returns a response message to the client, which provides resources to the client.

A URI is a unique sequence of characters that describe a resource that is either physical or logical [6]. They can be used to identify anything, and can return useful information on the object. They may also be a unique name for the object.

The RDF model is used as a standard data format for Linked Data. The standard use for RDF is the RDF triple, which is illustrated in Figure 1. Figure 1 illustrates how

Linked Data works with an RDF triple of the form *subject-predicate-object* [7]. In this case, the information is "Tim's birthplace is London." The subject is Tim, the predicate is about the birthplace with the statement "was born in", and the object is London. In grammar, the predicate designates a property or relation. In Linked Data, this definition is mirrored. The predicate component of an RDF describes the relation of the subject and object, which in turn determines the nature of the interlinks. Each component is named starting with HTTP, following the Linked Data principles [2].

There are three principles outlined for Linked Data. First, all conceptual elements, such as people, places, or literary works, should have a name starting with HTTP. Second, when looked up, an HTTP name should return useful data in a standard format like RDF. Finally, anything with a relationship to a conceptual entity should also be given a name beginning with HTTP [1].

Data linking involves determining if two URIs can be linked to one another, which indicates that they are in some way related. For Linked Data, these interlinks are called *typed links*. The linking property of the URIs is between the subject URI and the object URI [3].

A common Linked Data interlink is known as an identity link. An identity link describes two URIs that refer to exactly the same thing [3].

2.2 Linked Open Data

Linked Open Data is an extension of Linked Data that is based on the concept of *Open Data*. Open Data refers to the idea that data should be free from restrictions such as paywalls, and should not be hidden by copyright claims or patent limitations[1].

Linked Data that is released under an open license is Linked Open Data. The access to this data is free. While there are many Linked Open Data projects ongoing, this paper will look specifically at one that discusses library collection citations.

3 Uses of Linked Data

The following subsections detail two methods that incorporate Linked Data into library collections. Subsection 3.1 details a project that uses Linked Open Data to create a semi-automated citation catalogue and database. Subsection 3.2 discusses a Linked Data interlinking framework designed for library professionals.

3.1 Linked Open Citation Database

The Linked Open Citation Database (LOC-DB) is a project that curates citation information and also publishes the citations in a Linked Open Data format [3]. This system incorporates the capture of citation data into the standard workflow for new acquisitions. A principle idea of the system is that citations and reference data should be freely available.

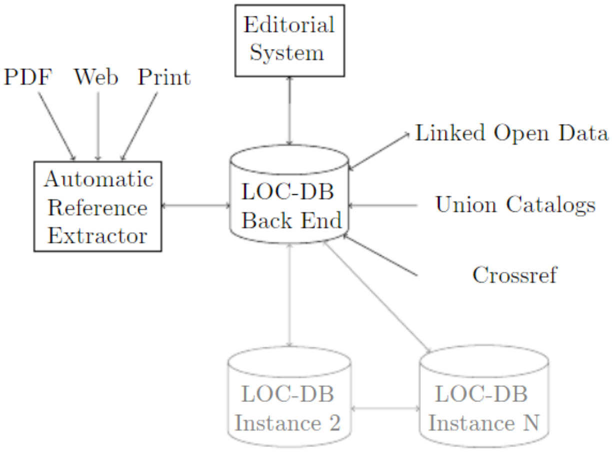

Figure 2. Overview of LOC-DB System Infrastructure

There are existing citation database services, however they do not fully meet the needs of libraries. They are sometimes pay-to-use and are often closed databases. This means the citation data held can only be used in reference to the data also held in this database. As a goal of the Semantic Web is to interconnect all data and turn the Web into a larger database [1], using Linked Open Data principles to create citation data would assist in increasing the reach of the information. In turn, this extended reach of a library's information would benefit the library as it would draw more traffic to their collection.

The LOC-DB system infrastructure is shown in Figure 2 [3]. As shown, it is a semi-automated, distributed system which creates and stores citation data. The first step in the process is to pull the references from a work using the Automatic Reference Extractor.

Shown in Figure 3, the Automatic Reference Extractor is an automated approach to extracting references. A page with references, either digital or scanned print, would be processed via either *text-driven reference extraction* or *layout-driven reference extraction*. Both methods are paired with ParsCit, an open source citation detection and labeling package. ParsCit works to analyze the contents of a reference string and retrieve the citation contexts from that string, working to turn them into metadata that describes what the citation is of and where it was taken, which would be the original document fed to the process.

Text-driven reference extraction detects bibliographic references using information derived from the text of the pages. This method is mainly used for electronic based source material such as born digital PDFs. Born digital documents are those that originate electronically, such as a typed word document. These only need a text extraction done before being passed to ParsCit. However, text-driven reference extraction can also be used on a scanned document. A pre-processed document is classified into either a single or double column

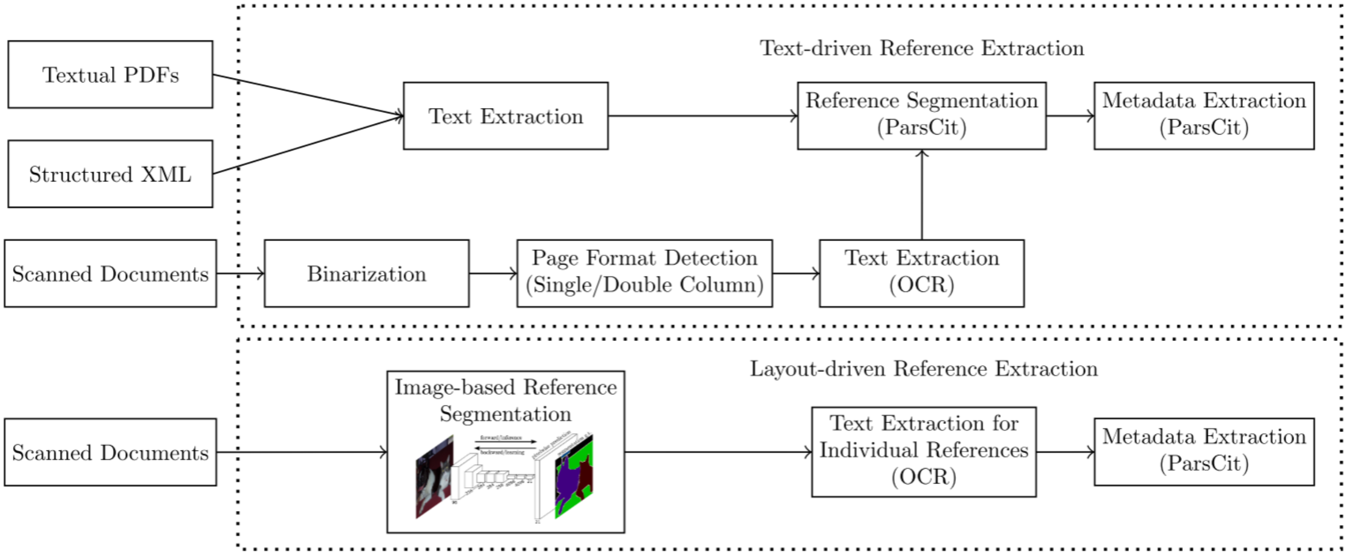
Text-driven Reference Extraction
Textual PDFs
Structured XML
Text Extraction
Reference Segmentation (ParsCit)
Metadata Extraction (ParsCit)
Scanned Documents
Binarization
Page Format Detection (Single/Double Column)
Text Extraction (OCR)
Layout-driven Reference Extraction
Scanned Documents
Image-based Reference Segmentation
Text Extraction for Individual References (OCR)
Metadata Extraction (ParsCit)

Figure 3. Automatic Reference Extraction Pipeline

layout, which is then passed to an optical character recognition (OCR) process for text extraction. In the design, the researchers used an open source OCR software called OCRopus.

According to the developers of the LOC-DB [3], text-driven reference extraction has many drawbacks. The quality of OCRopus meant that sometimes text was incorrectly recognized, which could cause whole references to be missed. Furthermore, they also found issues with ParsCit in being unable to detect some references or merging multiple references together. In order to improve the process, the researchers used layout-driven reference extraction with the results of the text-driven extraction.

Layout-driven reference extraction uses the layout information of documents to localize reference strings in the document image. This method works in a similar manner to how most humans process a document. This method localizes each reference string in the document based on where citation locations are usually found in documents. The deep learning program was trained on papers with marked areas of citations, and used this knowledge to segment the document into images of individual references based on the graphical layout of the given document. Once segmented, the new reference regions are passed through OCRopus and ParsCit in the same manner as the text-driven reference extraction. No additional drawbacks were noted with this method.

Both the works and their extracted references are stored in the database for use in the system. Another component of LOC-DB is the collection and creation of metadata for the citations. This metadata can be found in a variety of sources.

Traditional library cataloging of books, collections, and journals already come with metadata that is populated into a library union catalogue. A library union catalogue is a combined library catalog which describes the collections of multiple libraries. These union catalogues are a useful tool for searching through a larger collection than that of a single library for researchers [8].

If the metadata is not already in the database, nor in a linked union catalogue, then the system also searches through Linked Open Data on the Web from sources such as Google Scholar. With these sources pulled in, the user is then able to select all the matching resources, which increases the link coverage of the data they're linking.

All citing resource metadata is given a standardized, structured format that works with Linked Data, and can be shared and reused in the Linked Open Data Cloud. To do this, the system describes citations in the RDF format outlined in Section 2.1 [3].

As the database is built and creates more citations to be shared onto the Linked Open Data cloud, the need to create new citations will diminish, making the process fit better into the workflow of new acquisitions for the collection.

With the system developed and an early user interface prototype implemented, the researchers conducted a user study to find the strengths and weaknesses of their Linked Open Citation Database. Leading informational professionals (who for the purpose of this paper are library professionals) highlighted the fact that many steps in the process of uploading to the LOC-DB are generally handled by different cataloguers. These include scanning in documents and cataloging the document's metadata. Domain experts also valued suggestions from their union catalogs to resolve a citation link, and felt that this should be incorporated in the process.

An advantage of this system is that it opens up citations to a greater audience. The researchers behind this paper believe that citations play an important role in information retrieval, bibliometrics, and scientific discourse. Current systems that create automatic citations either hide important information behind a paywall, or are formatted for use in only one library's citation database. This system would change both of these issues, allowing citations to be easily accessible to all libraries and researchers on the Web.

A downfall of this system could be its estimated labor costs to initialize and maintain LOC-DB in the beginning stages. According to the researchers estimations, the University Library of Mannheim, whom they worked with, would need 6 to 12 full-time staff to maintain LOC-DB only for the social sciences division of the library. However, this estimate is taking into account only one library working on adding and maintaining LOC-DB citations. The researchers are optimistic that if many libraries join the LOC-DB project and collaborate, the labor costs of initializing many of these citations could easily be distributed amongst the libraries involved.

3.2 The Novel Authoritative Interlinking of Schema and Concepts (NAISC)

A group of researchers, Lucy McKena, Cristophe Debruyne, and Declan O'Sullivan, have been researching what library professionals are looking for in their applications of Linked Data [4]. The group conducted a survey in a previous year, asking library professionals about their opinions and wants for Linked Data applications.

According to this survey, library professionals are interested in using Linked Data with their collections. The library professionals were asked for their current Linked Data experience level, their perceived benefits and challenges of using Linked Data, and the usability of current Linked Data interfaces for cataloging. Overall, the feedback collected indicated that library professionals thought Linked Data would benefit their institutions. Linked Data tools would aid in improved resource discoverability. However, current Linked Data tools did not meet their needs, as the tools are difficult to use and not specialized to work with library collections.

The survey concluded that it would be a great help to libraries to create linked data tools designed specifically for library professionals' workflow and expertise [4].

Using the information gained from this study, the research group created NAISC - the Novel Authoritative Interlinking of Schema and Concepts, an interlinking approach designed for the library domain. NAISC aims to facilitate the use of linked data by library professionals [5].

Figure 4 outlines a cyclical, four-step interlinking framework designed for NAISC. I'll discuss the process step-by-step.

In **Step 1**, the user selects both an internal and external Linked Data dataset from which two entities will be linked. The external dataset is also checked for quality to ensure that it is worth linking with the internal entity. The URIs taken from both datasets for interlinking are also checked to make sure they can be interlinked.

As mentioned in subsection 2.1, a majority of linked data interlinks are identity links. However, the researchers believe that creating interlinks to express other relationships may add much value to the Linked Data datasets.

In **Step 2**, the user is guided through the process of creating a typed link. This typed link should accurately describe the internal and external URI relationship. The system then recommends one possible link-type property from a list, based on the kind of relationship between the two URIs. This link type and relationship, as defined by the predicate described in Section 2.1: Linked Data, is called the ontology. The ontology of a link adds to the information provided for data provenance.

Step 3 involves the generation of provenance data. Per the researchers, provenance is the information provided on the processes, resources, people, and institutions involved in creating a piece of data. Furthermore, the data provenance in this case would also keep track of modifications to the dataset and allow for an explanation or justification of the sources and type of links created for resources. The amount of information given by the library (fully encapsulating as much of the creation information as possible) is what creates a rich data provenance. It lends credence to the source, which is an important thing for library data to have, as libraries are considered an authoritative source of information. Publishing the provenance of their interlinks is an important step for libraries, as it aids establishing the origin of data and lends credence to the trustworthiness of libraries. Authoritative interlinks should have rich data provenance.

In **Step 4**, the interlink and provenance Resource Description Framework (RDF) is generated. RDF is used to encode the structured information of a Linked Data dataset. The data generated in Step 4 is stored in a database.

The NAISC system was implemented with a basic GUI and run through a user evaluation from 15 participants who had some prior knowledge of Linked Data and the Semantic Web. These participants were asked to perform six tasks related to creating new linked data. The six tasks are as follows:

- Start a new project.
- Add an internal URI to the link-set.
- Search an external dataset and add a related URI to the link-set.
- Select the link-type that best describes the link-set.
- Generate a Resource Description Framework (RDF) graph for the interlinks created.
- Generate a sample provenance graph that describes how the interlinks were created.

After completing these tasks, the participants were given a post-test interview which asked a series of questions in order

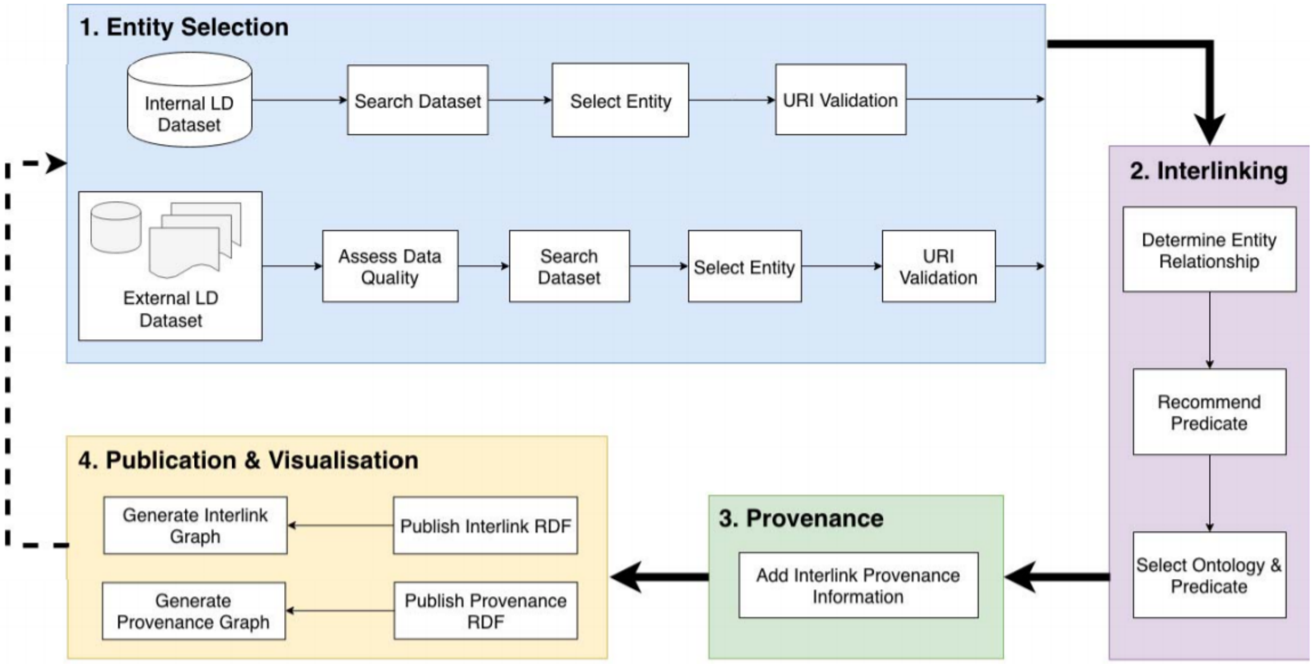

Figure 4. Framework for NAISC Interlinking

to establish thoughts on the GUI, the provenance model, and the interlinking framework.

The results of the evaluation indicated that library professionals found the system easy to use and helpful in creating linked data for their collections. The majority of issues found in the system were from usability issues with the GUI created for NAISC, however overall framework of the NAISC system held up.

4 Discussion

These two studies on the application of Linked Data for library use show promise in the field. They also show the range of ways that linked data can be used successfully, as they both address different aspects of library collections. LOC-DB focuses on curating and publishing citation data for a given work, while NAISC is interested in publishing data for the given work. They also display different levels of automation in the process, which highlights the strengths of both systems.

In terms of quality, the more user-involved process of NAISC may have a higher accuracy and quality rate. But the time-consuming nature of heavily involving the user in the creating of the interlinks may intimidate a library interested in publishing Linked Data, as it would take many employees a large amount of time to get through that library's collection.

In contrast, LOC-DB had more automation in its process. It still involved the user, but relied more on automated-processes with its reference extraction and data linking suggestions. The failing of the automated reference extraction was addressed successfully in the introduction of a second method of reference extraction which increased accuracy.

Both studies show the potential of high-labor cost, which is a major concern at this time. Using a lot of time and manpower to publish an entire collection's worth of Linked Data is a daunting project for any library collection, especially for larger libraries. After processing an entire collection, the library would only have to work with new acquisitions, which can be put into the current workflow. However, there is concern about adding Linked Data tools into the current workflow for library professionals. LOC-DB intends for one user to follow the entire process, however the components of that process are typically done by different employees. And NAISC takes a non-trivial amount of time to use, which could add substantially to the workload of the library professional handling the new acquisition. These issues would have to be addressed in later iterations of the studies.

5 Conclusion

As the Semantic Web project grows in scope and reach, the benefits of using linked data in a variety of organizations has grown. This includes libraries, where information professionals in charge of the digitized collections work to use Linked Data in a meaningful and efficient way.

These studies are still in earlier stages of development, and will likely get more attention and refinement before being the tools can be fully released to the public. Once refined, Linked Data tools created for library use will become very valuable assets. As already discussed, Linked Data is very promising for library collections for searching and cataloguing information, and investing the time and resources into Linked Data tools and technology will ultimately enhance a library's digital presence.

6 Acknowledgements

Thank you to Elena Machkasova for being a mentor and critic, and for guiding me in the research process.

Thank you to Jocelyn Bayer and Peter Dolan for giving valuable feedback and guidance.

References

[1] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 2009.

[2] A. Haller, J. D. Fernández, M. R. Kamdar, and A. Polleres. What Are Links in Linked Open Data? A Characterization and Evaluation of Links between Knowledge Graphs on the Web. *J. Data and Information Quality*, 12(2), May 2020.

[3] A. Lauscher, K. Eckert, L. Galke, A. Scherp, S. T. R. Rizvi, S. Ahmed, A. Dengel, P. Zumstein, and A. Klein. Linked Open Citation Database: Enabling Libraries to Contribute to an Open and Interconnected Citation Graph. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, page 109–118, New York, NY, USA, 2018. Association for Computing Machinery.

[4] L. McKenna, C. Debruyne, and D. O'Sullivan. Understanding the Position of Information Professionals with Regards to Linked Data: A Survey of Libraries, Archives and Museums. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, page 7–16, New York, NY, USA, 2018. Association for Computing Machinery.

[5] L. McKenna, C. Debruyne, and D. O'Sullivan. NAISC: An Authoritative Linked Data Interlinking Approach for the Library Domain. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 11–20, 2019.

[6] Wikipedia contributors. Hypertext Transfer Protocol — Wikipedia, The Free Encyclopedia, 2021. [Online; accessed 17-March-2021].

[7] Wikipedia contributors. Resource Description Framework — Wikipedia, The Free Encyclopedia, 2021. [Online; accessed 17-March-2021].

[8] Wikipedia contributors. Union catalog — Wikipedia, The Free Encyclopedia, 2021. [Online; accessed 4-April-2021].