Linked Open Data for Cultural Heritage

Evolution of an Information Technology

Julia Marden
Pratt Institute School of
Library Information Science
144 West 14th Street
New York, NY
jmarden@pratt.edu

Carolyn Li-Madeo
Pratt Institute School of
Library Information Science
144 West 14th Street
New York, NY
cmadeo@pratt.edu

Noreen Whysel
Pratt Institute School of
Library Information Science
144 West 14th Street
New York, NY
nwhysel@pratt.edu

Jeff Edelstein
Pratt Institute School of
Library Information Science
144 West 14th Street
New York, NY
jedelstein@pratt.edu

ABSTRACT

Communication design encompasses how information is structured behind the scenes, as much as how the information is shared across networks (Potts & Albers). Information architecture can profoundly alter our perceptions of society and culture (Swarts). Today cultural heritage institutions like libraries, archives, and museums (LAMs) are searching for new ways to engage and educate patrons. This paper examines how linked open data (LOD) can solve the communication design problems that these institutions face and help LAM patrons find new meaning in cultural heritage artifacts.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Theory

Keywords

Linked Open Data, Cultural Heritage, User Experience, User interface

1. INTRODUCTION

Communication design encompasses how information is structured behind the scenes, as much as how the information is shared across networks [22]. Information architecture can profoundly alter our perceptions of society and culture [25]. Today cultural heritage institutions like libraries, archives, and museums (LAMs) are searching for new ways to engage and educate patrons. This paper examines how linked open data (LOD) can solve the communication design problems that these institutions face and help LAM patrons find new meaning in cultural heritage artifacts.

Although nascent in practice, many LAMs are beginning to adopt linked open data as a way to organize and disseminate their catalogs of holdings. Linked open data organizes information using four basic rules:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs, so that they can discover more things.

For example:

Moby Dick (subject) is a book (predicate) written by Herman Melville (object). We can express this using URIs and RDF triples as:

```
<http://dbpedia.org/page/Moby-Dick>
<http://dbpedia.org/page/Herman_Melville>
<http://purl.org/dc/terms/creator>
```

LOD is freely available to access, download, and use. It is distributed on an open license (1-star); as machine-readable, structured data (2-star); in a nonproprietary format (3-star);

made available via World Wide Web Consortium (W3C; 4-star); and is linked to other people's data (5-star). This five-star model, envisioned by Tim Berners-Lee, is the widely accepted framework for evaluating LOD projects[1][28].

While this model does well to evaluate how well the data is structured and shareable, it does not answer the primary question of this paper: how can linked open data help cultural heritage institutions design a communications system that better shares their holdings with the public?

A linked dataset converts a basic catalog of cultural heritage items into RDF triples using a predefined vocabulary[6]. These datasets can then be matched with other RDF triples to offer a richer cultural heritage experience. LOD gives LAMs the opportunity to set their collections free from silos and place them in multiple contexts by pairing them with different LOD sets from around the world. Essentially, LOD allows users to interrelate communication artifacts without needing the interpretation of an archivist, curator or librarian. This ability for users to create their own relationships between artifacts is an important aspect of communication design (What is Communication Design? Clay Spinuzzi)

This paper will examine how LAMs are adopting LOD projects to address five major challenges within communication design:

1. Museums, libraries, and archives often possess specialized or rarefied information. How can they present that siloed information in a way that establishes their collection as a trusted information source?
2. How can these groups combine these siloed collections to create a new and sustainable, high quality datasets on a particular subject?
3. How can these institutions bring their backend development to the forefront and empower other cultural heritage holders to share their collections in a more open network?
4. How can cultural heritage institutions create a better user experience that empowers patrons to draw their own conclusions about cultural heritage artifacts?
5. How can these groups take advantage of the linked open data framework to expand the definitions of what cultural heritage can be?

2. METHODOLOGY

Working from within the framework of Tim Berners-Lee's Five Star model for linked open data, we sought to address common communication design problems faced by cultural heritage institutions, with a focus on how these problems can be resolved through the adoption of linked open data. Our goal is to illustrate how a spirit of openness and an adherence to linked open data web technology standards can benefit both institutions and users.

For this paper we chose to examine linked open data projects from around the world that had a common goal of improving the cultural heritage experience for their users, typically the citizens of a particular nation. Through the examination of their project documentation, applications, and datasets we first considered their contributions to the linked open data community and then grouped them according to the design problem they excelled at resolving through the use of linked open data.

Reflecting the realities of grant-based project development, this survey of the linked open data for cultural heritage landscape includes projects that are incomplete, unfinished or currently still in development. Although not all of these projects would receive a five-star rating on Berners-Lee's evaluation scale in their current state, their project documentation illustrates a dedication to the linked open data movement and a strong adherence to its standards. Furthermore, each of these projects would rate within the spectrum of the traditional five-star rating system.

The following research reflects the projects we found most exemplary of a particular design problem. It is by no means a complete survey of the linked open data landscape.

3. OPENING SILOED INFORMATION TO ESTABLISH AN INSTITUTION AS A PUBLISHER OF HIGH QUALITY DATA

Many institutions have the potential of turning information that was previously only used for internal purposes – such as cataloging information – into distinctive and informative datasets. These datasets, built off of years of institutional growth and careful work can benefit both the institution and the larger community by expanding the semantic web and establishing an institution as a trusted source of high quality data. The most robust of these datasets have been converted into RDF triples and are shared via an open API or through a SPARQL query endpoint. These linked open data requirements also enable users to have greater accessibility, while this access is lighter and easier to handle for the hosting institution. The Library of Congress and The Hungarian National Library are two national libraries who have released datasets, one to maintain the value of their already well-used cataloging information and the other to promote their more siloed collection to an international audience.

Beginning in 2009, the Library of Congress (LOC) converted its famous subject headings and authority names into linked open data through the Library of Congress Linked Open Data Service porta[12]. This project is part of the larger Bibliographic Framework Plan, an initiative to encourage libraries to transition their collections from MARC records towards RDA and linked open data[15].

The goals of the LOC's Linked Open Data Service are twofold, benefiting both the Library itself as well as human and machine users. With a linked open dataset, users can download authority names and files in bulk, which results in fewer taxing downloads on the LOC's web servers. These users can now also link to the LOC's data values and utilize the LOC's concept and value relationship mapping within their own metadata[20], all at no cost.

For individual human users, the Linked Open Data Service portal's search tool works similarly to the traditional authorities portal but features more information, an updated look, and a simplified results pages. Searching under a related name (e.g., Lady Day for Billie Holiday), users are taken directly to the authority file where popular LOC information can be found; the file's URI, links to alternative formats, and exact matching concepts from other schemes are also provided.

For the LOC, its only logical to coin authoritative and reliable URIs from existing vocabularies and authorities. In order for the LOC to maintain their influence among catalogers it was imperative that they convert their authorities into linked open data.

The success of the LOC's Linked Open Data Services is multifaceted. Through exposing its authority files to linked open data, the LOC has increased the relevancy of its holdings for a new generation of users. Along with updating their dated Authorities portal it connected its holdings to other libraries and alternative schema, therefore making this corpus of knowledge lighter and more flexible for both the LOC's internal use and for its users.

The releasing of linked open datasets can also increase the influence of smaller libraries in the field, as illustrated in the Hungarian National Library's conversion of its authority files, digital library and OPAC to linked open data[16]. Beginning in April 2010, the Hungarian National Library's shared catalog was one of the first successful linked open data projects. The Hungarian National Shared Catalog is part of The European Library, a major provider of cataloging data to the Europeana Project [CITATION]. By sharing their information through the Europeana.eu internet portal this small national library has connected their information with 2,000 other institutions across Europe. The Hungarian National library incorporates RDFDC for bibliographic data, FOAF for name authorities[7],, SKOS for subject and geographical terms, DBpedia name files, CoolURIs and owl:sameAS statements[27], which all help weave its dataset into the fabric of the linked open data community. Furthermore the documentation and creation of their linked open data processes has enabled the library to branch into other interesting projects, and has acted has a guide for other institutions.

Releasing accessible, easy to use datasets is a powerful and meaningful project for institutions with large amounts of information. Although the creation of a dataset may appear to be only the first step towards a larger linked open data project, it can be a standalone project with meaningful results for the end user. Linked open datasets are easier to access as well as to transform them into new information, as examined in the next section.

4. COMBINING SILOED COLLECTIONS TO CREATE NEW AND SUSTAINABLE GROUPS OF CULTURAL HERITAGE ARTIFACTS

Individuals, archival collections, repositories, libraries and other cultural institutions of all sizes and prominence in the field can utilize LOD to combine previously siloed collections. Through the utilization of collaborative knowledge bases and linked open datasets, cultural heritage institutions can enrich their own collections through collaboration, or even foster the creation of a new, authoritative and sustainable subject specific datasets. Alternatively, an existing cultural heritage institution can offer their patrons additional context for understanding their collections by integrating preexisting linked open datasets into their websites and apps, and by encouraging patrons to forge new connections.

A small but ambitious project laid the framework for some of the best practices for the creation and linking of open data. Civil War Data 150 (CWD 150) championed for public engagement, collaborative app development and the growth of a collaborative knowledge bases such DBpedia or in the case of this project, Freebase[8].

A partnership between the Archives of Michigan, the Internet Archive, and Freebase, CWD150 was a multifaceted project that encompassed, and planned to encompass, a number of different data sources, tools, and applications as well as a social media component. Along with promoting the digitization of archival documents from the Civil War, CWD150 championed the adoption of linked open data and the strengthening of Freebase. Libraries, archives, museums, and even individual researchers were encouraged to contribute data to the projec[26].

The issues that led to the termination of the project are not stated anywhere on the project's website, but it can be assumed that this ambitious project did not have the staffing necessary to fulfill all of its goals.

An ongoing project that has received continued funding and growing interest in the Linked Jazz Project. The Linked Jazz Project utilizes linked open data technologies in order to enhance the discovery and understanding of cultural heritage assets. Through processing of archival jazz interview transcripts from disparate institutions the project follows linked open data web standards from the minting of new URIs to the development of RDF triples and the creation of a powerful API. Transcripts are first exposed to natural language processing tools, which pull out full names (entities) and partial names. These entities are then mapped against DBpedia, and previously unrecognized entities have URIs created for them[21].

Utilizing a crowdsourcing tool, these annotated transcripts are then analyzed by users who assign relationships between the interviewee and the names mentioned in the transcript using a linked open data friendly vocabulary. These relationships are then available as RDF triples, an API and a network visualization[1]. Through the analyzing of transcript data and the exposure of this data to linked open data technologies the Linked Jazz Project works to expose the relationships of the jazz community and introduce these relationships to a larger audience.

Finally, the LOCAH project[3] was an effort to publish data from the finding aids of Archives Hub and the catalogs of more than 70 major UK and Irish national libraries[2]. LOCAH, like PCDHN is an example of multiple institutions collaborating to merge their data together in order to create new research paths for their users[4].

[1] http://www.linkedjazz.org

In a brief feature article on the project, Adrian Stevenson describes its value as allowing the development of new channels into the data. Researchers are more likely to discover sources that may materially affect their research outcomes, and the hidden collections of archives and special collections are more likely to be exposed and used[23]. Variations in the data from institutions posed a challenge to end-users; although the libraries and archives providing the data adhered to standards, these standards can be hard to implement uniformly and can interfere with machine-processing [24].

Linking Lives[4]expanded on LOCAH by bringing in more external datasets and creating a model for a Web interface that allowed researchers to search the new joined archives by name-based biographical pages. While in concept phase, Linking Lives illustrates the potential richness of a collection based on the holdings of multiple institutions[24]. Linking Lives focuses on individuals as a way into archival collections as well as other relevant data sources[23]. One goal was to expose archival collections to researchers, who might not be familiar with primary sources or who might not think of searching archival collections when starting biographical research.

When institutions embark on a project to collect or join disparate data sources there are a number of considerations that should be remembered during planning. Institutions must be prepared to face issues that can arise from the merging of datasets of different qualities by planning for data cleaning. Additionally, projects will benefit from extensive funding not only to combat surprise costs such as difficult data merges but also ensure for money to promote and maintain data after it is linked. Contributing to collaborative knowledge bases such as DBPedia or Freebase can also ensure that the project's legacy lives on regardless of funding availability through the coining and publishing of publicly accessible URIs.

However, even with the financial and data challenges, smaller institutions can offer their patrons a much richer experience by linking their datasets with other related LAMs to provide a richer database for research.

5. BRINGING BACKEND DEVELOPMENT TO THE FOREFRONT IN ORDER TO EMPOWER OTHER CULTURAL HERITAGE HOLDERS

With linked open data in its infancy, a spirit of openness fostered by successful linked open data projects can help to improve user experience, define best practices and foster interest in the technology. The documentation and dissemination of backend development is key to demystifying linked open data for both users and potential creators by explicitly outlining the development of and potential uses for powerful linked open data applications. Linked open data projects such as the projects created by The New York Times exemplify the creation and stewardship of linked open data's future in cultural heritage.

The New York Times has adopted linked open data to maintain and share the newspaper's extensive holdings and is actively encouraging reuse via public APIs[17]. The datasets are based in large part on the newspaper's 150-year-old controlled vocabulary, The New York Times Index, an authoritative, cross-referenced index of all of the names, articles, and items that appear in the newspaper.

As of the spring of 2013, the New York Times has released fifteen APIs, ranging from Movie Reviews to the TimesTags API, which matches queries to the New York Times controlled vocabulary. The documentation for the suite of APIs is hosted in the Developer section of the New York Times website[17], which includes a glance view of the API as well as suggested uses for each API and a forum for users and developers. All of the New York Times APIs are available in a JSON response format and a smaller subset is available as XML or serialized PHP.

The New York Times publicizes its projects through its blog, Open: All the News Thats Fit to print(f)[18]. In addition to creating prototype tools such as Who Went Where, a search engine that enables users to search for recent Times coverage of the alumnae of a university or college, the New York Times also promotes the use of its APIs and source code. In a blogpost introducing the search engine the step-by-step process behind the creation of a API based application is also explained. Open has been a regularly updated blog since the New York Times Company began its foray into the use and promotion of open source software in 2007.

Who Went Where showcases the value of the New York Times and its APIs, as an elegant example of a straightforward application of these LOD APIs. Who Went Where is a JQuery application that queries DBpedia's SPARQL endpoint. The power of this tool is amplified by the documentation surrounding it, including the source code, which is freely available.

6. USING LINKED OPEN DATA TO IMPROVE CULTURAL HERITAGE USER EXPERIENCE

LAMs have a vested interest in improving user experience for their patrons in order to compete with the other major technological influences on culture– smart phones and the internet. A backend that runs on linked open data can radically alter the traditional library, archive, or museum experience. Several organizations and informal groups have made headway in conceptualizing user interfaces that expand users ability to experience and interact with cultural heritage. Many of these projects are still at a proposal stage, but highlight what can happen when linked open data is integrated with a cultural heritage website or application.

EUscreen, is Europeana's main aggregator for audiovisual media. Building on a network of content providers, standardization bodies, television research partners, and specific user groups, EUscreen provides multilingual and multicultural access to European essential components of European heritage, collective memory, and identity. By its nature, audiovisual media, particularly analog recordings, such as pre-digital television, radio, sound recordings and film, are difficult to access. EUscreen's linked open data pilot was created to address the need to make these artifacts openly accessible to a wide audience of users[10].

EUscreen's content selection policy and metadata framework borrows from existing standards such as the metadata schema of the European Broadcasting Union to tag a multiplicity of content through Europe and encourage exploration, comparative study, and serendipitous discovery. The different metadata models of the contributing institutions (XML, RDF, EBU Core ontology, 4store triple store repository, and SPARQL query)[9] are aggregated into a single EBU Core metadata structure and published to the EUscreen portal[11]. From there Europeana aggregates the content and makes it available through its website. Users can take advantage of this rich linked open data backend to find digital media from dozens of countries, in multiple languages and genres, dating back to the beginning of the Twentieth Century.
On a more conceptual level, the Agora project is a collaborative effort involving several Dutch cultural heritage institutions concerned with historical context and methods of manipulating and redefining context through social media platforms[5].. One major aim of the project is to shift the viewpoint of historical narrative from that of the curator or institution to that of the viewer.

The project's tagline, Eventing History, plays on the concept of inventing history. The project aims to put the power of defining what constitutes a historical event into the hands of app users. Members of the project have expressed the desire to do away with the conventional version of history by creating applications that connect artifacts in disparate collections and allowing users to link and discuss artifacts, locations, and events as they see them[5].

The demo, which is geared for touchscreen devices, allows users to unite objects from multiple collections based on a common historical event or actor. Artifacts which wound up in different collections over the years can now be regarded in the same frame of reference.

The project is ambitious in ideology and scope but technological documentation is not a strong point. Development of the project is partially fueled by the dissertation research of graduate students and the fate of the project beyond participant graduation is uncertain. Although Agora has begun some user interface development, its mark may be more philosophical than as the producer of a usable application.

Rich linked open data empowers users to better navigate cultural heritage collections and draw their own conclusions about the meaning and significance of artifacts.

7. EXPANDING THE DEFINITION OF CULTURAL HERITAGE

In addition to the technical requirements, LOD projects are executed with a spirit of openness and collaboration, that can not only simplify but also redefine the cultural heritage user experience.

We think of the traditional cultural heritage user experience as a consumer experience. Libraries, archives, and museums preserve and curate a cultural heritage experience. Patrons come to each institution to consume that pre-packaged experience. A linked open data project can remove the barriers between curator and cultural consumer.

More and more, governments and private citizens are taking on a role of promoting use and reuse of open datasets. In September 2011, the Dutch Heritage Innovators Network[13] began the Open Cultuur Data [2] initiative to encourage cultural institutions to release their data under open standards and encourage users to develop new uses for this data. They facilitated the creation of datasets from eight organizations: the Rijksmuseum, Amsterdam Museum, EYE Film Institute Netherlands, National Archives, the Netherlands Institute for Sound and Vision, and National Heritage Sites of the Netherlands[19].

The datasets were made public in time to be relevant to developers entering the Apps for the Netherland contest, a government-sponsored nationwide contest encouraging developers to create smartphone apps that would engage users with the rich heritage of the Netherlands. INE hosted several hackathons before the contest deadline, creating a supportive environment for developers to use the new open cultural heritage datasets in the creation of cultural heritage apps with a strong user interface. Thirteen apps were created during the initial contest, including three award-winners:

- Rijksmonumenten.info [3]. This app that allows users to browse more than 61,000 buildings in the Netherlands and take geotagged photos of each building to share via Wikimedia. It won an education award.
- ConnectedCollection [4]. This app that is targeted more toward the cultural heritage organizations themselves, allowing them to install a widget on their site that shows users related objects from partner institutions. They won funding to continue development.
- Vistory [5] This project used a linked open dataset of images and video. Users can discover historical films shot near their location, and contribute to the geotagging of historic videos.

Each of these apps redefine who is a creator and who is a preserver of cultural heritage. The developers take on a preservation role, and the users gain the ability to draw new meaning and create new understandings of cultural heritage.

However, although these projects exemplify the philosophical intent behind linked open data, many of the datasets used were in unlinked formats. Among government agencies and developers, the spirit of open data is catching on much more quickly than the technical specifications for linking.

Japan has become a leader in linked open government data, hosting the 2011 and 2012 Linked Open Data Challenge Japan [6] along with nonprofit and business leaders. Winners in 2012 developed linked open data apps to improve

[2] http://www.opencultuurdata.nl
[3] http://rijksmonumenten.info
[4] http://www.opencultuurdata.nl/?p=583
[5] http://www.vistory.nl/what-is-vistory.shtml
[6] http://lod.sfc.keio.ac.jp/challenge2012/

user experience and discovery. One app tracked the spending of tax dollars in local government; another helped users to explore photos related to the history of Hakodate, the first Japanese port opened to foreign trade.

Japan demonstrates that giving people access to linked open data sets can blur the lines between national and cultural heritage identities. Users can track tax dollars to see how much is spent on a local museum, or gain easier access to primary source knowledge about important periods in history. Their apps have richer potential because the datasets are compatible and reusable using linked data standards.

Looking at these examples, we find that just as the open government data movement could benefit from adopting linked open data standards, cultural heritage LOD projects could benefit from adopting contest models that encourage users to access datasets and transform them into meaningful new experiences for other users.

We believe linked open data has the potential not just to preserve cultural heritage for users, but to offer users new opportunities to understand, manipulate, and recreate cultural heritage experiences. Embracing the philosophy behind open government data, that citizens have a right to access and contribute to data, we believe users have the same right to contribute to their cultural heritage experience.

8. CONCLUSION

As Hart-Davidson and Grabill put it, "Technology drives change because it alters culture." [14] Certainly we've seen this with the advent of mobile devices, but perhaps we haven't paid as much attention to the ways that information architecture has changed our culture. Linked open data offers a new way for cultural heritage institutions to share their holdings with a wider audience, and to change the traditional relationship between the holder of knowledge, the interpreter of knowledge, and the consumer of knowledge. With a strong user interface built upon a linked open data set, users with all levels of expertise can access and analyze information once siloed in many different LAMs. This new way to interpret and access cultural heritage information might allow us to update how we define cultural heritage itself.

9. REFERENCES

[1] Linked data, June 2009.

[2] Archives Hub. Archives hub, n.d.

[3] Archives Hub. Linked open copac and archives hub (locah), n.d.

[4] Archives Hub. Linking lives: Using linked data to created biographical resources, n.d.

[5] L. Aroyo. Agora creating the historic fabric for & providing web-enabled access to objects in dynamic historical sequences - isab 2012 site visit, Dec. 2012.

[6] C. Bizer, R. Cyganiak, and T. Heath. How to publish linked data on the web., 2007.

[7] D. Brickley and L. Miller. Foaf vocabulary specification 0.98, Aug. 2010.

[8] Digital Library Federation. Civil war 150: Notes toward a linked data case study, 2011.

[9] European Broadcasting Union (EBU). Metadata specifications, n.d.

[10] EUscreen. About euscreen, n.d.

[11] EUscreen. Euscreen linked open data pilot, n.d.

[12] T. Gheen. Library of congress launches beta release of linked data classification, July 2012.

[13] B. Grob, L. Baltussen, L. Heijmans, R. Kits, P. Lemmens, E. Schreurs, N. Timmermans, and E. van Tuijin. Why reinvent the wheel over and over again? how an offline platform stimulates online innovation. paper presented at museums and the web 2011, philadelphia, pa, Apr. 2011.

[14] W. Hart-Davidson and J. Grabill. The value of computing, ambient data, ubiquitous connectivity for changing the work of communication designers. *Communication Design Quarterly*, 1(1):16–22, September 2012.

[15] Library of Congress. Library of congress linked data service: About, n.d.

[16] National Szechenyi Library. Hungarian national library opac and digital library published as linked data, nd.

[17] New York Times. Api documentation and tools. developer network beta, 2013.

[18] New York Times. Open [web log site]., 2013.

[19] J. Oomen, L. Baltussen, and M. van Erp. Sharing cultural heritage the linked open data way: Why you should sign up. paper presented at museums and the web 2012, san diego, ca., 2012.

[20] Open Metadata Registry. International standard bibliographic description (isbd) elements., n.d.

[21] M. C. Pattuelli, M. Miller, L. Lange, S. Fitzell, and C. Li-Madeo. Crafting linked open data for cultural heritage: Mapping and curation tools for the linked jazz project. *Code 4 Lib*, 21, July 2013.

[22] L. Potts and M. Albers. Defining the design of communication. *Communication Design Quarterly*, 1(1):3–7, September 2012.

[23] B. Ruddock and J. Stevenson. Creating linked open data for library and archive description. multimedia information & technology, 37(4): 19-20. stevenson, j. (2012). linking lives: Creating an end-user interface using linked data. information standards quarterly, 24: 14-2, 2011.

[24] J. Stevenson and A. Stevenson. Lifting the lid on linked data: Linked data and the locah project. presentation at european library automation group (elag) conference, prague, czech republic., May 2011.

[25] J. Swarts. Communication design. *Communication Design Quarterly*, 1(1):12–15, September 2012.

[26] Use case Civil War Data 150. World wide web consortium.

[27] World Wide Web Consortium. Owl web ontologylanguage reference., 2009.

[28] World Wide Web Consortium. World wide web consortium., Sept. 2011.