Linked Metadata and New Discoveries

Julienne Pascoe
Canadiana.org

Scholarly and Research Communication
VOLUME 6 / ISSUE 2 / 2015

Abstract
This article examines the opportunities and challenges surrounding the creation and use of linked open data (LOD) for cultural heritage resources in libraries, archives, and museums. With a specific focus on the metadata projects at Canadiana.org, this article explores LOD principles and strategies for implementation within the context of cultural heritage collections, highlighting the significance of the Semantic Web for research and engagement with cultural heritage resources across disciplines and communities.

Keywords
Linked Open Data; Semantic Web; Metadata; Cultural Heritage; Digital Scholarship

Résumé
Cet article étudie les opportunités et les défis présentés par la création et l'application des Données Ouvertes Liées (DOL) dans le cadre des ressources patrimoniales et culturelles détenues par des bibliothèques, archives et musées. En se référant principalement aux projets de métadonnées de Canadiana.org, l'article explore les principes des DOL et les stratégies pour implémenter celles-ci dans le contexte de collections culturelles et patrimoniales, mettant en avant l'importance du Web sémantique pour la recherche et l'engagement avec ces ressources au sein de divers domaines d'études et communautés d'utilisateurs.

Mots clés
Données ouvertes liées; Web Sémantique; Métadonnées; Patrimoine culturel; Recherche numérique

Julienne Pascoe is Metadata Architect at Canadiana.org and teaches the graduate course Digital Applications for Collections Management in the Film and Photographic Preservation and Collections Management program at Ryerson University. Email: julienne.pascoe@canadiana.ca

CISP Press
Scholarly and Research Communication
Volume 6, Issue 2, Article ID 0201218, 8 pages
Journal URL: www.src-online.ca
Received June 1, 2015, Accepted July 13, 2015, Published October 14, 2015

Pascoe, Julienne (2015). Linked Metadata and New Discoveries. *Scholarly and Research Communication, 6*(2): 0201200, 8 pp.

Introduction

Linked data and its semantically enhanced Web environment promises to alter the very fabric of the information landscape, forming new discursive pathways and modes of knowledge while advancing scholarly discourse and research. A key component of this transformation is the use of linked open data (LOD) principles and practices to overcome information silos, which pose a challenge for the discovery, access, and repurposing of data on a global scale. As memory institutions transition from publishing documents to publishing data, the principles and practices of linked data become increasingly interconnected within the larger discourse of the Web and metadata development, as well as the role of cultural heritage institutions in an information society. The theme of the 2015 Implementing New Knowledge Environments (INKE) Conference in Whistler BC, Sustaining Partnerships to Transform Scholarly Production, highlights the critical role partnerships will play in the ongoing digital transformation of knowledge environments. As a membership-based organization dedicated to providing broad access to Canadian documentary heritage, Canadiana.org is in a unique position to connect Canadian cultural heritage resources with the world. Canadiana.org's Héritage Project, a ten-year initiative to digitize and make over 60 million pages of primary-source documents accessible online, presents an opportunity for describing extensive archival resources using Semantic Web principles and then exposing these descriptions using LOD. Toward accomplishing this goal, Canadiana.org has designed a phased approach to building a linked data infrastructure that emphasizes the development of partnerships as well as community-based initiatives. While building a foundation for LOD, Canadiana.org will join other cultural heritage organizations in paving the way for a collaborative, interconnected Semantic Web future.

This article examines the issues surrounding the creation and use of rich and linked metadata from a variety of perspectives, including strategies for the bulk automated creation of metadata, crowdsourcing, the re-purposing of pre-existing metadata, and the creation and sharing of linked data by end users. Through exploring LOD principles, the article argues for the significance of the Semantic Web for research and engagement with cultural heritage resources across disciplines and communities, as well as defines the opportunities and challenges that exist within the current linked data movement. While LOD seems to be a moving target for many institutions, the opportunities for enhanced research, user engagement, multimodal discovery, and the inter-linking of cultural heritage resources provide incentives for exploring the latest discoveries in generating and publishing linked data. Using the Web's inherent characteristics, data publishers and users can work together on the design of user-centred interfaces and collaborative tools that will shape the global information space of tomorrow. The discourse of linked data is framed within the larger narratives of the development of the Web as an information space, as well as the history of metadata standards and applications within cultural heritage communities. The first of these areas involves the transformation of the Web as a network of connected documents to one of connected data sets: "The vision behind the semantic web was born out of the frustration of having only human-readable information on the web, which restricts the ways in which software can help us find information" (van Hooland & Verbogh, 2014, p. 44). To counter this escalating problem, Tim Berners-Lee drafted a set of best

Pascoe, Julienne (2015). Linked Metadata and New Discoveries. *Scholarly and Research Communication, 6*(2): 0201200, 8 pp.

practices for publishing and linking data on the Web, known as the "linked data principles." In his influential Web architecture paper on LOD, Tim Berners-Lee (2006) introduced the four basic principles: 1) use Uniform Resource Identifiers (URIs) as names for things, 2) use Hypertext Transfer Protocol (HTTP) URIs, so that people can look up those names, 3) when someone looks up a URI, provide useful information using the standards, Resource Description Framework (RDF) and SPARQL, 4) include links to other URIs, so they can discover more things. These guidelines harness the architecture of the Web and its existing standards, including HTTP and URIs, to evolve the Web from a global document space to a global data space. In *Linked Data: Evolving the Web into a Global Data Space,* Tom Heath and Christian Bizer (2011) highlight this extension of Web standards in creating a single data space:

> Just as hyperlinks in the classic Web connect documents into a single global information space, Linked Data enables links to be set between items in different data sources and therefore connect these sources into a single global data space. The use of Web standards and a common data model make it possible to implement generic applications that operate over the complete data space. This is the essence of Linked Data. (p.4)

Defining linked open data

Linked data is not a specific technology or standard, rather it is a set of best practices for the publication of structured data on the Web. At the heart of LOD is the RDF data model. RDF is a simplified, node-edge graph model that consists of single data statements about a resource using triples. Triples are data statements about a resource consisting of two nodes (subject and object) and an edge (predicate) that indicates a relationship (predicate) between a resource (subject) and another resource (object). Each statement is complete and does not depend on other statements or the context of a record. These statements correlate to the field/subfield and value model employed in record-based metadata models (Mitchell, 2013). The model is highly flexible and extendable. It allows additional data statements to be added as needed, in which objects and predicates can themselves become subjects with their own predicates and objects, creating a graph representation of a single resource's complex relationships with itself and external resources. For example:

> The Mona Lisa (subject) is created (predicate) by Leonardo Da Vinci (object)
> Leonardo Da Vinci (subject) was born (predicate) in Vinci, Italy (object)

In addition, identifiers are used for subjects, predicates, and objects, to create unique meaning as opposed to locally defined literal values. All of the semantics are made explicit in the statement itself, and statements can be linked with other statements outside your information system, "allowing heterogeneous data to connect and interact" (van Hooland & Verbogh, 2014, p. 44). Serialization of RDF includes RDF-XML (Extensible Markup Language), Turtle, and JavaScript Object Notation for Linked Data (JSON-LD). The RDF Schema (RDFS) is an extension of the basic RDF vocabulary, providing "an internal vocabulary to help establish the rules and structure of the assertions made about a resource" (Mitchell, 2013, ch. 2). The exchange of linked data is facilitated by technologies including the SPARQL Protocol, a query language for RDF

data sets. SPARQL is a W3C recommendation consisting of a set of specifications that facilitate querying and manipulating RDF graph content on the Web (W3C, 2013). SPARQL endpoints, the mechanism for conducting data queries, allows for deep graph searching across LOD sources, returning answers in the form of RDF data that can then become new LOD data sources. These building blocks of metadata – the data model (RDF), the data schema (RDF Schema), the vocabularies (Simple Knowledge Organization System [SKOS], Web Ontology Language [OWL]), the serializations (Turtle, JSON-LD), and the exchange format (SPARQL) – create the linked data structures that enable the Semantic Web: "By using these structures, we can enhance a computer's ability to infer relationships between resources, helping us to bridge from resource description to knowledge representation" (Mitchell, 2013, ch. 2).

A roadmap for linked data

While linked data certainly offers the promise of an integrated, enhanced data exchange landscape – an idealized webscape driven by open access to information that transcends institutional and disciplinary boundaries – there are challenges and obstacles to overcome. Debates surrounding the development of new systems and tools supporting linked data-driven applications generate renewed discussion about metadata value and quality, community assessment and needs, as well as institutional impact and services. In addition there is an increasingly complex world of information objects, with which new models and standards attempt to reconcile.[1] The question of how we will accomplish this is one that many cultural heritage organizations are asking.[2] The goal is to strike a balance between the new approaches to metadata (Web scale, interoperability, resource discovery, user-centred design) with ongoing struggles for metadata quality, richness, sustainability, and other issues involving persistence and provenance. Taking into considerations the larger goals of linked data within the context of institutional objectives and collections, libraries, archives, and museums (LAMs) have begun to pave technical development paths for implementing LOD systems. Mitchell (2013) outlines one such path: "[1] Define an LOD model; [2] aggregate data; [3] publish data for user and computational access; [4] enhance LOD endpoint integration; and [5] disseminate via SPARQL endpoints" (ch. 4). Based on linked data principles and technologies, as well as its own position as a provider of Canadian documentary heritage, Canadiana.org has developed a customized, phased roadmap to achieving a linked open data vision. This approach consists of five key stages including

1. Lay the foundation: describe and link Canadiana resources;
2. Expose the foundation: publish and visualize Canadiana resources;
3. Build on the foundation: link with and visualize partner resources;
4. Enrich the foundation: establish the framework for cooperative description; and
5. Link the foundation to the world (Ward, 2014).

Each stage incorporates strategies for implementing LOD standards and technologies as well as outlining key partnerships and collaborations with the private sector, academic communities, memory institutions, and the public. Toward this end, critical strategies and methods for approaching the challenges of rich, reliable, and consistent metadata that can be published as linked data are being explored, and many of them

Pascoe, Julienne (2015). Linked Metadata and New Discoveries. *Scholarly and Research Communication, 6*(2): 0201200, 8 pp.

utilize the very capabilities and characteristics that continue to develop the Web as a distributed, collaborative, and diverse information network.

The "how" of linked data is intricately tied to the challenges and opportunities of metadata and Web technologies afforded in twenty-first century information organization and cultural heritage management. As Mitchell (2013) highlights: "One of the enduring values of the Web that made it central to how people engage with information is the notion that information in the digital world is not bounded because of scale, authority, or cost because the efficiencies, communities, and economies of the Web changed how people engaged with and valued information" (ch. 4). The Web's inherent qualities that make it a driving force of innovative, open source development, and community-based initiatives can be utilized to advance metadata extraction, enhancement, and the aggregated dissemination of digital resources. Canadiana.org's linked data strategy incorporates Web-centred initiatives as well as key partnerships with diverse stakeholders and users in order to publish, link, enhance, and reuse Canadiana resources. In addition, Canadiana.org will employ the tools and techniques of LOD to publish and connect its collections across the Web, using the inherent qualities of the linked data technologies to link Canadiana.org's extensive documentary collections to the world.

Strategies for metadata generation and discovery

The demand for rich online content that can be repurposed has led to strategies and applications for the automation of metadata, both in creation and harvesting. A useful tool for the extraction of searchable metadata is Optical Character Recognition (OCR). Using algorithms for pattern recognition, OCR converts files, scans, and other documents into searchable documents, enabling the full-text searching of digital documents. OCR paves the way for computation analysis such as text mining as well as further metadata extraction for keywords and indexing. Quality OCR output is dependent on the image provided, which sometimes requires human intervention or the design of algorithms that manipulate the image to facilitate the machine recognition of characters and reduce the rate of error. Canadiana.org uses adaptive algorithms to optimize images as well as the open source OCR software Tesseract to provide full-text access to its digitized documents. Automated metadata extraction will become an increasingly important tool for creating large amounts of metadata for subsequent discovery and analysis. While there has been success with converting typed materials, character recognition of handwritten documents remains an elusive target. Describing the archival collections in the Héritage Project, most of which contain handwritten documents, will require partnering with online communities to provide metadata transcriptions and descriptions of Canadiana content.

User-centred design and engagement is a vital component of the development of tools and interfaces for metadata creation, discovery, and reuse. Design, however, does not need to be an isolated phenomenon but can be integrated into developing tools for community building around collections development, supporting user engagement along with metadata enrichment. This bottom-up, user-driven, collections-building approach is featured in the set of activities known as crowdsourcing. Mia Ridge from The Open University and the Association for Computers and the Humanities

characterized crowdsourcing as "asking the public to undertake meaningful tasks related to cultural heritage collections in an environment where the activities and/or goals provide inherent rewards for participations. The project should contribute to a shared, significant goal or research interest" (OCLC Web Junction, 2014). The result of such a partnership provides users with informative access to and engagement involving cultural heritage resources while assisting institutions in describing their collections. Crowdsourcing has the added benefit of building communities of volunteers around collections, reflecting the potential of the Web for distributed, collaborative social networking around centralized philanthropic missions. The success of such initiatives depends on the user-centred design and conceptualization of the crowdsourcing project, which will ultimately request users to commit cumulative time and energy to the analysis and description of diverse cultural heritage projects. With the Héritage Project, Canadiana.org is developing crowdsourcing projects that work with both targeted user-groups, such as genealogical organizations, as well as the public to enhance the metadata description of Canadiana's extensive archival collections. Canadiana's tools will support both the transcription of records and the tagging of key features in archival documents. Tapping into the drive to contribute, collaborate, and discover, crowdsourcing provides a Web and user-centred solution to metadata challenges of resource discovery, engagement, and scale, while developing a mutually beneficial partnership between diverse audiences and institutions.

Linked data and the Semantic Web propose to connect vast data sets across the islands of information they are currently organized in, creating aggregations of large-scale amounts of information. As indicated by Mitchell (2013): "The transformation of bibliographic and digital collection metadata to LOD and LOV environments … opens up new opportunities for researchers to work with collections and metadata using computational and cross-repository techniques" (ch. 4). These cross-domain data sets can be processed and analyzed by the design of algorithms, data mining, and visualization tools shaped by methods such as distant reading and emerging disciplines such as cultural analytics. Distant reading, as defined by Anne Burdick, Johanna Drucker, Peter Lunenfeld, Todd Presner, and Jeffrey Schnapp (2012) in *Digital_Humanities*, is "a form of analysis that focuses on larger units and few elements in order to reveal patterns and interconnections through shapes, relations, models and structures" (ch. 2) Aided by the use of computational techniques and analysis, distant reading uses the vast and growing data sets made available on the Web to cross-examine the cultural record. Cultural analytics continues this tradition of distant reading through the dissection of large-scale cultural data sets using computational analysis and data visualization to reveal insight as well as enhance literary and historical scholarship (Burdick et al., 2012). Visualization and data design have become critical analytical tools for both reading and authoring visual interpretations and annotations of the historical record (Burdick et al., 2012). These tools provide further points of engagement with distributed data sets and linked cultural heritage, asking new questions of the data through complex data querying, visualizations, and graphical interfaces (Mitchell, 2013). At the foundation of these applications are the underlying partnerships between research communities and digital repositories. In building a linked data framework, Canadiana.org will seek partners within the academic community to design advanced visualization and interpretive tools that employ

Pascoe, Julienne (2015). Linked Metadata and New Discoveries. *Scholarly and Research Communication, 6*(2): 0201200, 8 pp.

computational techniques to extend, interpret, and reuse metadata from Canadiana's published digital resources.

Conclusion

The Semantic Web vision promises to connect heterogeneous collections through the linking of data statements, while using Web architecture and standards to enable machines or software to interpret and process the data. As Heath and Bizer (2011) emphasize, the Semantic Web "presents a revolutionary opportunity for deriving insight and value from data" (p. 5). How memory institutions respond to this vision will transform access to cultural heritage collections. Canadiana.org is in a unique position to implement linked data principles while forming collaborations with users in designing the tools that will unlock the vast potential of its documentary collections, redefining and reshaping the knowledge environment for Canadian heritage. The opportunities and challenges for this transformation require the establishment of a partnership-driven infrastructure, one in which memory institutions focus on access, discovery, user-centred design, and engagement with diverse communities, while users design tools and interfaces to query and interpret connected data sets. Using the principles of linked data and the inherent transformative characteristics of the Web, the cultural heritage community can overcome the isolated collections of the past, connecting not only to each other's resources, but also to the global data space. In doing so, these organizations will provide the distributed infrastructure for linking repositories that will support researchers in the development of sophisticated tools and software, and contribute to the transformation of scholarly production. Throughout this process the foundational role information institutions play in society, as well as their responses to the technical and conceptual questions of linked data, will shape the knowledge environment through which scholarly discourse and production investigates and interprets the historical record.

Notes

1. Erik Mitchell (2013) discusses the attempt to accommodate the increasing complexity of information objects in the development of the Resource Description and Access (RDA) standard: "The RDA community, for example, has spent considerable effort in building out a new series of cataloguing rules geared toward accommodating an increasingly complex world of information objects" (ch. 1).

2. This question was raised throughout The National Information Standards Organization (NISO) Bibliographic Roadmap Project meeting in April 2013 and found to be a common theme in discussions within the library, archive, and museum (LAM) communities (Mitchell, 2013).

Websites

Canadiana.org, http://www.canadiana.ca/en/home

Tesseract, https://code.google.com/p/tesseract-ocr/

References

Berners-Lee, Tim. (2006). Linked data – design issues. *W3C.org*. URL: http://www.w3.org/DesignIssues/LinkedData.html [December 12, 2014].

Burdick, Anne, Drucker, Johanna, Lunenfeld, Peter, Presner, Todd, & Schnapp, Jeffrey. (2012). *Digital_humanities* (Kindle Edition). Cambridge, MA: MIT Press.

Heath, Tom, & Bizer, Christian. (2011). *Linked data: Evolving the Web into a global data space*. San Rafael, CA: Morgan & Claypool Publishers.

Mitchell, Erik T. (2013). Library linked data: Research and adoption. *Library Technology Reports* (Kindle Edition), *49*(5). Chicago, IL: American Library Association.

Morris, Liz. (2014). Something for everyone: How crowdsourcing creates community, preserves knowledge, and promotes discovery. *OCLC WebJunction*. URL: http://www.webjunction.org/news/webjunction/something-for-everyone-how-crowdsourcing-creates-community.html [December 19, 2014].

van Hooland, Seth, & Verbogh, Ruben. (2014). *Linked Data for libraries, archives, and museums*. Chicago, IL: Neal-Schuman.

Ward, Anne. (2014). Héritage project - our vision: Describing and linking Canadiana resources with linked open data (LOD). *White paper report*.

W3C. (2013). *SPARQL 1.1 Overview*. URL: http://www.w3.org/TR/sparql11-overview/ [December 19, 2014].