PROPOSAL for Open Repositories 2013

Case Study: Using Linked Data to Integrate Resources from Cultural Heritage Institutions across Canada

PAM ARMSTRONG, MARG STEWART, and ANNE WARD
Independent Consultants, Ottawa, ON, Canada

A consortium of Canadian universities, libraries and archives, undertook a project to demonstrate the use of semantic web technologies and linked data in showcasing the wealth of digital resources in their collections. The presenters will review the process adopted by the project, the ontologies/element sets and vocabularies used to integrate the resource metadata, and will share with the audience the lessons learned by the project.

Keywords: metadata re-use, linked data, resource description framework, resource description and access, RDF ontologies, RDF element sets, RDF vocabularies, semantic web

Background

Why Linked Open Data?

The major reasons for the adoption of Linked Open Data (LOD) by libraries, archives and museums are that:

- As cultural and heritage institutions, libraries, archives and museums are in the business of providing access to resources by the public;
- Opening (meta)data enables the reuse of this (meta)data for purposes unforeseen by today's information professionals;
- Linking open (meta)data enables the ultimate connection of resources from a broad community of institutions to create a semantically rich experience for the public;
- Linked open (meta)data provides the means for libraries, archives and museums of showcasing our resources in a context that crosses organizational and subject boundaries.

The Linked Open Data Landscape

Today, libraries, archives and museums throughout the world are:

- Exposing (meta)data from catalogues and controlled vocabularies as Linked Open Data (LOD) sets;
- Promoting broader and strengthened discovery of resources through LOD;
- Breaking out of institutional silos by joining this (meta)data with a wide variety of knowledge domains through LOD; and
- Positioning their organizations to take full advantage of the power of the intelligent web.

A consortium of Canadian universities, libraries and archives recognized the potential in this approach for:

- Enhancing discovery of Canada's documentary heritage for and by Canadians; and
- Making Canada's cultural and heritage collections more accessible and meaningful for generations to come.

Approach

The overall approach to the project was to make use of existing metadata about the resources and repurpose it without loss of context and meaning.

Rather than reduce the metadata to a common subset, the approach was to maximize its use by moving to the semantic web's "web of data" concept, and expressing the contributed metadata as Resource Description Framework (RDF) "triples", that is, in the form of:

<subject> <predicate> <object>

For example:

"The Handmaid's Tale" "has author" "Margaret Atwood"

"The Handmaid's Tale" "was published in" "Canada"

"The Handmaid's Tale" "was published in" "1985"

"The Handmaid's Tale" "was published by" "McLelland and Stewart"

In effect taking a "record" about a resource and transforming it into a series of "data statements" (or facts) about the resource that can be combined in different and unforeseen ways. In addition, rather than simply keeping the textual values of the <subject> <predicate> <object> data facts, Uniform Resource Identifiers (URIs) were defined for the <subject> and <object>, and existing/published RDF ontologies/element sets were used for expressing the <predicate> (thereby promoting a common understanding of the expressed "data statement" by users outside the project).

The Metadata

Metadata for resources that related to a selected topic (i.e., the First World War) was contributed to the project using a range of encoding formats, including: MARC records, spreadsheets, MODS XML, and Dublin Core RDF.

Putting the Metadata in Context

The metadata for the contributed resources was organized in accordance with a FRBR/RDA-informed model, specifically using the Group 2/3 entities of concept, object, event, person, family, organization (corporate body), and geography (place); and the relationships between these entities and the resources that the metadata described. As well, a subset of the resources (specifically "published" resources) were described in accordance with the work-expression-manifestation-item model.

Expressing the Metadata in RDF

The contributed metadata was mapped to a selection of ontologies/element sets, including mapping to multiple elements sets to address both specificity (e.g., the use of the IFLA ISBD element set / vocabulary for the characteristics of resources) and broader user community use (e.g., the inclusion of Dublin Core, FOAF, and BIO which represent commonly used ontologies / element sets).

Linking Data

"Authority data" was created for all instances of persons and organizations, and their URIs used for linking between project resources: in addition links to existing "authority data" (e.g., VIAF) were established where a match was found. For new concepts (e.g., concepts that did not exist in published authorities) and events, "authorities" were also created (using the RDA Group 3 Elements and/or SKOS as deemed appropriate), and their URIs used for linking.

Published vocabularies were used in establishing the values for the “object” of the “data statements”, specifically for subjects, type of content (e.g., photograph, music), characteristics of the resources, and profession/field of activity of persons/organizations; as well as to enable linking to similar resources in other collections on the web (e.g., to the “Trenches to Triples” initiative in the UK, to resources categorized using Rameau) and discovery by other initiatives making use of these vocabularies.

“Discovering” Resources from Metadata

The project elected to implement an application to “visualize” the metadata (rather than simply implement a “search” application), using ideas and concepts from:

- Tim Wray’s Canvas (http://timwray.net/2011/12/canvas/);
- The Real Face of White Australia (http://invisibleaustralians.org/faces/);
- Mildenhall's Canberra (http://mildenhall.moadoph.gov.au/); and
- applying ideas from http://discontents.com.au/shoebox/every-story-has-a-beginning to tell a story to pull resources together.

The project identified “interesting” topics to be explored, using the “subjects” of the resources (e.g., person, events, etc.), the “affiliations” of persons to organizations (e.g., soldiers to the Canadian Expeditionary Force (CEF), soldiers to battalions), and the “type of content” of the resources (e.g. war posters, war songs, photographs, films, panoramas, newspapers, postcards, etc.) that appeared in the metadata: these “topics” were used in establishing an initial set of dimensions to be explored.

Using the LOD/RDF approach, a resource can be linked to any other resource by simply adding a “triple” identifying the relationship between the resources: with a traditional search environment, the linkages would be left to the user to make using the results from searches across the separate individual collections. Using a traditional “web resource aggregation” approach, the links between all elements would need to be defined in advance and the protocols for using the metadata from each collection would need to be known.

The “visualization” application provided only a sampling of what is possible: the metadata provides many more things that can be discovered using the ontologies/element sets implemented. For example: songs have composers, lyricists, and performers, some of whom were also soldiers; events occurred in many geographic areas over time.

The potential for discovery is limited only by the imagination of the user of the metadata: “data statements” (or facts) can be added to the “web of data” by anyone, and new “stories” / ”visualizations” can be developed by combining these facts with the project knowledge base.

What the Project Learned

Ontologies/Element Sets

From the analysis of the contributed metadata and the exercise of selecting published RDF ontologies/element sets for mapping, it is apparent that:

- A large and rich set of RDF ontologies / element sets exists today;
- Multiple ontologies / element sets can be used together;
- Some gaps remain to be addressed by the LOD community;
- Care in selecting target ontologies / element sets is still needed.

Power of "Reuse"

When investigating vocabularies, and in some instances element sets (e.g., the RDA "role" element set), for use in establishing the values for "objects" in "data statements", it was interesting to discover that:

- Element sets can be used in multiple contexts;
- Vocabularies can be repurposed;
- Repurposing published vocabularies extends the reach;
- Metadata mapping is performed once.

"Power" of RDF

From the "visualization" application, the power of RDF became very apparent:

- Integrating resources across separate and distinct collections is easy;
- Including resources from external organizations is easy (no negotiation required);
- "No programming" required (in the small);
- Follow the URI to "learn something new";
- Lots of information moving forward - need to organize results.

Summary

RDF and LOD are an elegant approach for integrating resource discovery across different domains, institutions, and services. More specifically, RDF and LOD:

- enable web users and third party organizations to integrate the project's resources with their own resources to create their own "stories" and "virtual exhibitions"[1];
- enable web users/third party organizations to make connections to resources outside the project (through linked URIs in the metadata (e.g., to Rameau, UKAT, LC, OCLC));
- provide access to project's resources freely[2] and at no additional cost (outside of publishing the metadata and making the resources described by the metadata "persistently" accessible)[3];
- effectively remove the constraints of existing web approaches in which the paths followed by web users are explicitly defined by the organization hosting the metadata/ resources.

---

[1] "No negotiation required": there is no need for web users/third party organizations to understand the hosting organizations protocols/schemas for accessing their resources/metadata - the metadata is published using "open" ontologies and "vocabularies".

[2] All project metadata is available under an open license (ODC-PDDL), and resources (or data objects) are available for use in accordance with the policies of the individual participating institutions (e.g., resources are available under open license schemes to a great extent, however some resources are subject to copyright and/or specific attribution requirements): regardless of the resource-specific policies, all resources are freely discoverable.
Please refer to http://www.canadiana.ca/sites/pub.canadiana.ca/files/PCDHN%20Proof-of-concept_Final-Report-ENG_0.pdf for the complete report or http://www.canadiana.ca/en/pcdhn-lod for a video of the visualization application and for accessing the project metadata in RDF.

[3] There is no cost of building an application to search/navigate the metadata/resources: this effectively becomes the domain of the web user/third party organization.