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

OR2013
Charlottetown, PEI, 9 July 2013
Anne Ward, Independent Consultant
Sharon Farnel, University of Alberta
Agenda

• “Out of the Trenches: A Linked Open Data Project”
  • Project Background
  • Approach
  • Visualizing the Data
  • Lessons Learned
• “Out of the Trenches” and Over the Top
  • New Recruits / Resources
  • Automation: Mapping / Linking
  • The Next Campaign: Opportunities or Challenges?
• Conclusions
“Out of the Trenches: A Linked Open Data Project”

Video, Report, and RDF Data
http://www.canadiana.ca/en/pcdhn-lod
Project Background

(PCDHN) is a consortium of Canadian cultural heritage institutions that includes universities, libraries and archives.

- A proposal was made to the network in 2011 to "develop a national model for linking our metadata in order to improve the discoverability of Canada’s documentary heritage collections and to provide Canadians seamless and integrated access to this content”

- A subset of PCDHN partners agreed to participate/provide First World War resource metadata and undertake a proof-of-concept using linked open data.
Approach

- Metadata records (e.g., MARC, Excel spreadsheets, Dublin Core RDF, MODS XML) provided by partners were analyzed
- Metadata published as RDF by other organizations was reviewed (including Bibliothèque nationale de France (BnF) who included some Resource Description and Access (RDA) elements in their RDF mappings)
- A preliminary set of target ontologies was identified, including commonly used ontologies such as FOAF and DC (for discovery by a broader set of users), as well as ontologies better suited to the actual content / type of resource described (e.g., RDA for published materials such as music recordings and music sheets, OAI/ORE for aggregated resources (for more specificity))
- Example mappings were created and ratified by/with partners ...
  ... using a classification model informed by FRBR/RDA, CIDOC, ISAD/ISAAR ...
approach: the model

- Person
  - may relate to another person(s)
  - may relate to another family(ies)
  - may be a member of one or more families
  - may have an affiliation with one or more organizations

- Family
  - may relate to other organizations, may contain sub-organizations

- Organization
  - may relate to other organizations, may contain sub-organizations

- Concept
  - agent, observer, reporter, etc.
  - born in, dies in, resides in, is associated with, etc.

- Event

- Object

- Place

- Resource
  - is about
  - author, creator, donor, publisher, manufacturer, etc.

- Archival Aggregation
  - contains sub-aggregations, is related to other archival aggregations (e.g., other collections)

- Archival Item
  - may contain items, may relate to other items

- Werk
  - may contain, be about, or relate to other works

- Expression
  - is realized through
  - embodies one or more expressions
  - is embodied in one or more manifestations

- Manifestation
  - manifestation(s) of work(s)
  - is exemplified by one or more items

- Item
  - may contain items (e.g., CD with enumerated songs)

A network of relationships between these concepts allows for a comprehensive understanding of the model's structure and how the various entities interact with each other.
• ... the ratified ontologies (element sets) and vocabularies identified by the partners were applied to the model and used to develop the mappings for individual metadata records ...
Approach: Applying Ontologies and Vocabularies

- Properties using published ontologies (element sets) including RDA, Dublin Core, Friend-of-a-friend (FOAF), event ontology, time ontology, etc.

- Properties using published ontologies (element sets) for “concepts” - SKOS (Simple Knowledge Organization System)

- Value of “place” properties using published URIs including [www.geonames.org](http://www.geonames.org) or geographic based ontologies (element sets) where detailed information is required

- Relationships using published ontologies including RDA, Dublin Core, Friend-of-a-friend (FOAF), event ontology, etc.

- Values of properties using published vocabularies including RDA vocabularies, LCSH, ISBD, RVM, etc.
Approach: Mapping Metadata, Creating “Authorities” and Linking

- Conscious decision was made to include the mapping of individual metadata elements to multiple RDF elements (e.g., “rdaGrp2:preferredNameForThePerson” and “foaf:name”) to address broader community use (i.e., through more commonly understood/used ontologies) and specificity (e.g., RDA for specifying preferred and alternative names, IBSD for more explicit characterization of the content of a resource)
- Authority data (RDF files) was created for the “things” in the metadata records that conformed to the model (e.g., for person, organization, event, resource)
- For some concepts, PCDHN authority data was created (e.g., for French subject headings that do not currently exist in RDF form)
- PCDHN URIs were assigned to the unique “things” represented in the metadata and used to create links between them using the relationships provided by the ontologies (e.g., “dc:creator” / “rdaRole:author” for linking a resource to the person / organization)
Approach: Mapping Metadata, Creating “Authorities” and Linking

• Values for properties such as "dc:subject" were researched in the identified vocabularies (such as LCSH and FAST) and the corresponding URIs used in the RDF for resources

... a conscious decision was also made to include multiple authorities for a given concept (to achieve maximum discoverability) as well as retain its textual value (for use in an application to showcase the data)

• Common subjects (e.g., all First World War subject headings) were injected into the RDF for the resources

• The content of the resource was also characterized using vocabularies such as ISBD, LCSH, TGM, and FAST using the corresponding URIs (linking to RDF datasets in the LOD cloud)

• Mappings (mapping rules) for each individual set of metadata records were developed along with mappings of textual values to vocabularies (for subject and content type)

(see Appendix E in the report at PCDHN No Proof-of-concept Final Report)
Approach: Visualizing the Data

- The partners also decided to “visualize” the data in the RDF files generated from the mappings rather than simply implement another “search” application.
- Inspired by initiatives such as:
  - Tim Wray’s Canvas ([http://timwray.net/2011/12/canvas/](http://timwray.net/2011/12/canvas/))
  - The Real Face of White Australia ([http://invisibleaustralians.org/faces/](http://invisibleaustralians.org/faces/))
- the partners elected to tell the story of an individual soldier ... and along the way ...
Visualizing the Data

... dimensions for the visualization were derived from the characteristics of the resources/entities described, including:

- the type of entity (e.g., person, event, organization)
- their relationships (e.g., subject of, affiliations to organizations, participation in events)
- the type of resource (e.g., films, photographs, postcards, posters, songs, newspapers)
Visualizing the Data

• The scope of the visualization was limited to:
  • exploring two dimensions: a single soldier’s story (Mike Foxhead), and the events of the First World War
  • using the links to resources within the PCDHN only (although links to external resources are provided (e.g., links to published vocabularies such as OCLC Fast and LCSH, and authorities such as VIAF))
  • querying data along a limited set of pre-defined paths
Visualizing the Data: The Story

authority data
(RDF for Mike)
a depiction
biographical information

places associated with the person (place of birth, death)

“geonames” used for places and provide maps

“dc:subject” used for resources linked to places
Visualizing the Data: The Story

“dc:subject” used for resources linked to Mike
thumbnails specified in RDF or generated from source resource

“dc:subject” used for resources linked to the group / organization

“event:isAgentIn” used for events Mike participated in
“dc:subject” used for resources linked to the event

group/organization Mike was associated with (“rdaGrp2:affiliation”)
Visualizing the Data: An Event

authority data for an event

“event:isAgentIn” used for identifying who participated in the event

place associated with the event

“geonames” used for place and provide map

“dc:subject” used for resources linked to places

links to other resources for the event
Visualizing the Data: An Event

“dc:subject” used for resources linked to the event
Lessons Learned: Project Observations

• ... and of course as many in the cultural heritage community have realized, “RDF and LOD are an elegant approach for integrating resource discovery across different domains, institutions, and services”
• ... RDF and LOD “enable web users and third party organizations to integrate PCDHN resources with their own resources to create their own “stories” and “virtual exhibitions””
• ... RDF and LOD “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”
Lessons Learned: Project Observations

• A large and rich set of RDF ontologies / element sets exists today
• Multiple ontologies / element sets can be used together: no longer must resource descriptions be constrained to a “one-size-fits-all” nor to a “one-vocabulary-to-rule-them-all” model
• Through the work of the community, equivalences between ontologies / element sets can be expected in the future, for example by linking terms in different element sets ... ... as well as between vocabularies (FAST and LCSH, LCSH and Rameau / German subject headings)

Figure 3.1
Specific titles from RDA and FRBR can link to Dublin Core for more compatibility with more generalized metadata.
Source: Karen Coyle, Chapter 3, Linked Data Tools: Connecting on the Web, Library Technology Reports, alatechsource.org, May/June 2012

1 RDF enables “the replacement of attempts at one-size-fits-all schema, rules and other international/global standards with what might be termed an all-sizes-fit-one approach”
2 Term used by Owen Stephens in a e-mail exchange on the “Bibframe” Listserv
Lessons Learned: Project Observations

• Some gaps exist, for example:
  • the “vocabulary” or “classification structure” for archival aggregations that is in use today (e.g., “collection”, “fonds” / “series” / “sub-series”, etc.) - these may be established as an RDF vocabulary (used as the “object” for a “predicate” of “rdaElement:systemOfOrganization”) or as a class in an ontology (used as the “rdf:type” for an archival resource)
  • addressing the “time” dimension of properties such as for the name of an individual without going into describing the events that led to name changes over time - although this could be modelled by mapping the property as follows (borrowing from examples in the vcard (http://www.w3.org/2006/vcard/ns#) ontology):

    <rdaGrp2:preferredNameForThePerson>
    <rdf:Description>
      <rdf:value>"name of the person"</rdf:value>
      <time:intervalStarts>"date"</time:intervalStarts>
      <time:intervalFinishes>"date"</time:intervalFinishes>
    </rdf:Description>
    </rdaGrp2:preferredNameForThePerson>
Lessons Learned: Project Observations

- Care needs to be applied to the selection of target ontologies (longevity, community support, stability), although mappings can be revised or “better” data statements can be added.

- Element sets and vocabularies can be used in multiple contexts (rather than creating single purpose properties / vocabularies): for example, “rdaRole:author” can be used as a “predicate” to link an author and a resource, or as the “object” to the “rdaGrp2:professionOrOccupation” predicate; or an LCSH heading for “soldiers” can be used for the subject of a resource or as the value for a group affiliation (rdaGrp2:affiliation) or profession (rdaGrp2:professionOrOccupation).

- Metadata mapping is essentially done once (although it may be updated as new information / better ontologies are identified): the most resource-intensive work is in defining the appropriate mapping rules - after that it is a matter of developing the tools to apply the rules on a larger scale.
Lessons Learned: Project Observations

- Integrating resources across separate and distinct collections is easy and no negotiation is required due to the RDF model itself - everything is expressed in the same manner (triples) and using the data statements is a matter of understanding the intent of the ontologies used and the fact that these ontologies exist / are published on the web.

- Manually creating RDF for a small set of records is manageable but not sustainable for large volumes.

- Data queries can yield a very high volume of results (as demonstrated by the application when loading all resources for the First World War event): intelligent applications will be required to better organize results for users as well as address any duplication in results (for example, with resources fully described at an individual resource level as well as at the level of an aggregation, that share common characteristics such as subject).
Lessons Learned: Project Recommendations

• Specific recommendations endorsed by the partners:
  • develop “visualizations” for other dimensions in the data, including “geographic” and “time” dimensions for events
  • “productionize” the visualization application
  • provide more intelligence in parsing the results
  • provide the ability to retrieve/filter by multiple dimensions
  • add additional resources from group members and interested parties
  • develop programmatic support for applying the mapping rules

... and as we will see next, the University of Alberta has taken up the challenge ...
“Out of the Trenches” and Over the Top
New Recruits / Resources

Memorial University of Newfoundland
WWI Newfoundlandia

Experiences in France
By Wilfred T. Grnspell, M.D.

On the Gallipoli front last summer, where for a quarter of a century he has been conducting a great work of philanthropy, by the healing of wounds and the spreading of knowledge, Dr. Grnspell, the foundation of the Memorial University of Newfoundland, is now daily aiding the wounded soldiers of France to gain back thelost capacities in the theatre of war. He prays to our readers, and is now chief of his work, to the"process" to the"healing" of their wounds.

O

If we return from France you ask.

It is not my intention to speak about the medical condition of the men in the trenches. A large portion of the men have returned to their homes and the wounded soldiers have returned to their homes. The men are in their own homes and are being cared for by their own families. I will mention only the experiences of the wounded soldiers in the trenches.

Here is the first of a series of articles that will be published in the Memorial University of Newfoundland Review. It is hoped that these articles will give you an insight into the life of the men in the trenches.

A Pair of Grey Socks

Lovingly dedicated to the men of the Newfoundland Regiment and to every woman who has assisted in their work.

By Mrs. T. J. Delee

Memorial University of Newfoundland

WWI Newfoundlandia
Investigating options for programmatic metadata mappings and conversions
- XSL stylesheets based on mappings to date
- Existing generic conversion tools + some customization
- Enhancement tools like SILK, OpenRefine
The Next Campaign: Opportunities or Challenges?

• New content, new partners
• Improved processes and workflows
• Enhanced application, new tools
• More exposure for the project and the data
The Next Campaign: Opportunities or Challenges?

• And ...
  • Starting small is ok
  • Think curation
  • Focus on unique content
Conclusions

The future of resource description is not about making resources discoverable on the web, it is about “description that happens on, in, and as part of the web and the networked world we live in” (Library of Congress Bibliographic Framework Initiative) ...

... it is about using semantic web technologies to describe resources ...

... it is about linking resources to other resources (on the web) ...

... the future of resource description is happening now ...

... it is happening in many domains ...

... and it is not about adopting a “one-size-fits-all” model nor conforming to “one-vocabulary-to-rule-them-all” ...

... but about finding links between concepts expressed in ontologies / vocabularies and between instances of “things” expressed in these ontologies
Thank you!

Questions?

Anne Ward
eaward@computer.org

Sharon Farnel
sharon.farnel@ualberta.ca