11-7-2012

Linked Data Demystified: Practical Efforts to Transform CONTENTdm Metadata for the Linked Data Cloud

Silvia B. Southwick
University of Nevada, Las Vegas, silvia.southwick@unlv.edu

Cory K. Lampert
University of Nevada, Las Vegas, cory.lampert@unlv.edu

Follow this and additional works at: http://digitalscholarship.unlv.edu/libfacpresentation
Part of the Cataloging and Metadata Commons, and the Databases and Information Systems Commons

Repository Citation

Available at: http://digitalscholarship.unlv.edu/libfacpresentation/99
LINKED DATA DEMYSTIFIED

PRACTICAL EFFORTS TO TRANSFORM CONTENTDM METADATA INTO LINKED DATA
PRESENTERS

Silvia Southwick
Digital Collections Metadata Librarian
UNLV Libraries

Cory Lampert
Head of the Digital Collections Department
UNLV Libraries
OUTLINE

• Why should I care?

• What is it?
  • Defining Linked Data / Introduction to Linked Data Concepts / Linked Data Principles
  • Technologies & Standards for Linked Data
  • The Linked Data Cloud

• How?
  • Applying these concepts to digital collection records
  • Anticipated challenges working with CONTENTdm

• The UNLV Libraries Linked Data Project

• How could you start working with Linked Data?
LINKED DATA MYTHS

My collections are already visible through Google; so who cares

This is a topic for catalogers

It’s too technical / complicated / boring

Actually ...

Linked data is the future of the Web

Data will no longer be in silos (catalog, CONTENTdm)

Relationships are powerful and worth the effort
HOW DO WE CURRENTLY CREATE OUR DIGITAL COLLECTIONS?

Data (or metadata) are encapsulated in records
Records are contained in collections
Very few links are created within and/or across collections
Links have to be manually created
Existing links do not specify the nature of the relationships among records

This structure hides potential links within and across collections – DATA IS TRAPPED!
Example: A search on “water” in the OCLC collection of collections resulted in 26 collections that are not cross-linked

<table>
<thead>
<tr>
<th>Digital Collections containing records on “water”</th>
</tr>
</thead>
<tbody>
<tr>
<td>California Water Documents</td>
</tr>
<tr>
<td>Western Waters Digital Library</td>
</tr>
<tr>
<td>Bear River Watershed Historical Collection</td>
</tr>
<tr>
<td>The Historic Landscape of Nevada: Development, Water, and Natural Environment</td>
</tr>
<tr>
<td>Seattle Power Water Supply Collection</td>
</tr>
<tr>
<td>Western Waters Digital Library: The Columbia River Basin in Oregon</td>
</tr>
</tbody>
</table>
EXPOSED DATA RELATIONSHIPS

POWERFUL, RELATED DATA
Example: Google Knowledge Graph

http://youtu.be/mmQl6VGvX-c
A LEGO METAPHOR FOR CREATING LINKED DATA

This is the Data Model
Transforming records into data

Publishing data

Linking data as you search or browse
DEFINING LINKED DATA

Linked Data

*refers to a set of best practices for publishing and interlinking data on the Web*

- Data needs to be machine-readable
- Linked data (Web of Data) is an expansion of the Web we know (Web of documents)
WEB IN TRANSITION

1. Two types of data:
   1. Human-readable documents (email, brochure, report)
   2. Machine-readable data (calendar, playlist, spreadsheet)

2. Shopping example
   1. A web page ad (document) says “dress”, “color”, “price”, “designer”
   2. But machines cannot extract data to re-use in another application (e.g., spreadsheet to compare prices)

3. RDF – new way to specify relationships and transfer context with data across applications: reusable data

4. The time is now to start to evolve our documents into data
TECHNOLOGIES FOR LINKED DATA

Linked data is built in the Web architecture (HTTP, URIs)

RDF is a data model (not a format)

Most common serializations:
- RDF/XML
- RDFa

RDF is based on triples/statements

SPARQL - **SPARQL Protocol and RDF Query Language** -- is an query language able to retrieve and manipulate data stored in RDF.
WHAT ARE TRIPLES?

Triples are expressed as:

subject – predicate – object

Examples:

Frank Sinatra -- is an -- entertainer
Frank Sinatra -- knows -- Jack Entratter
Introduction to RDF at http://www.linkeddatatools.com/introducing-rdf
PRINCIPLES OF LINKED DATA

1. Use **URIs** as names for things (people, organizations, artifacts, abstract concepts, etc.)

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 URI so that they users discover other related items (note: RDF Links have types)
THE LINKED DATA CLOUD
CREATING LINKED DATA FROM ORIGINAL RECORD VS. HARVESTED RECORD
Title: Café Monico menu, February 19, 1903
Category: regular services
Restaurant Name: Café Monico (London, England)
Additional Information: Advertisement on back and around edges if the menu. Insert lists Indian curries as special on Mondays and Thursdays
Graphic Elements: Borders(Ornament areas); Buildings; Photographs
Enclosures: daily specials; advertisements
Type of restaurant: Non-specialized restaurant
Type of menu: `a la carte
Meals served: dinner; lunch
City: London

.....
**OCLC WorldCat Linked Data**

**SAME RECORD (HARVESTED)**

<table>
<thead>
<tr>
<th>schema:publisher</th>
<th>rdf:type</th>
<th>schema:Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>schema:name</td>
<td></td>
<td>&quot;University of Nevada, Las Vegas University Libraries&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>rdf:type</th>
<th>library:Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>library:holdingsCount</td>
<td>&quot;1&quot;</td>
</tr>
<tr>
<td>library:oclcnum</td>
<td>&quot;729486897&quot;</td>
</tr>
<tr>
<td>schema:inLanguage</td>
<td>&quot;en&quot;</td>
</tr>
<tr>
<td>schema:name</td>
<td>&quot;Cafe Monico menu, February 19, 1903&quot;</td>
</tr>
</tbody>
</table>
HOW CAN WE ADDRESS THIS PROBLEM?

Create a **complementary** data structure that would allow dynamic interlinking among data

**How?**

- Export records from the collections
- Deconstruct these records by extracting data from them
- Apply vocabularies
- Adopt a common model to express data
- Publish data in a data space (Linked Data Cloud) where links among data are created automatically
EXAMPLES OF RECORDS

Showgirls

Menus

Dreaming the Skyline
What are possible triples for this photo?

<this photo> <created by> <Las Vegas News Bureau>
< this photo > <is a> <photograph>
< this photo > <is a> <print>
< this photo > <depicts> <Frank Sinatra>
< this photo > <depicts> <Jack Entratter>

----------

<Frank Sinatra> <knows> <Jack Entratter>
<Jack Entratter> <knows> <Frank Sinatra>
<Frank Sinatra> <is an> <entertainer>
<Jack Entratter> <is a> <theatrical producer>

----------
GRAPHICAL REPRESENTATION
OF THE PHOTO TRIPLES

entertainer
is an
Frank Sinatra

knows
knows

depicts

Jack Entratter

is a

theatrical producer

is a
photographic print
created by
Las Vegas News Bureau
What are the URIs for subjects, predicates and objects?
VOCAULARIES

ALERT:
Finally a place in the presentation we feel at home!

------------

Vocabularies are specific terms used in RDF statements to describe specific resources.

-------------

Vocabulary examples in linked data (Linked Open Vocabulary):
DCMI Type Vocabulary
Friend of a Friend Vocabulary
Geonames
MARC Code List for Relators
Creative Commons Rights Expression vocabulary
Schema.org
Many more at: http://lov.okfn.org/dataset/lov/
UNLV LINKED DATA PROJECT

Goals:

Study the feasibility of developing a single process that would allow the conversion of our collection records into linked data preserving their original expressivity and richness.

Publish data from our collections in the Linked Data Cloud to improve discoverability and connections with other related data sets on the Web.
PHASES OF THE PROJECT

Literature Review

Evaluating Technologies
  • Research existing technologies and best practices
  • Develop small experiments with technologies
  • Make decisions of which technologies to adopt, adapt or develop

Data preparation
  • Select and prepare records from digital collections to participate in the project

Run process for data transformation

Publish on the Linked Data Cloud

Assess results
DATA PREPARATION

- Defining vocabularies that will be adopted for predicates
- Defining types of triples to be created (literal, outgoing links, incoming links, triples that describe related resources, triples that link to descriptions, triples that indicate provenance of the data, etc.)
- Specifying URIs for new “things”
- Identifying potential URIs for external links (e.g., things that already have URIs)
- Describing data sets that will be published in the linked data cloud
TECHNOLOGY OPTIONS FOR DATA TRANSFORMATION
Type of Data

Structured Data (CONTENTdm)

Data Preparation

RDF-izers for Excel or XML

Data Storage

Drupal DB

Data Source API

Linked Data Wrapper

RDF Store

Linked Data Interface

RDF Files

Web Server

Data Publication

Drupal RDFa

Linked Data on the Web

Adapted from *Linked Data Evolving the Web into a Global Data Space* by Heath and Bizer
ANTICIPATED CHALLENGES

- Developing of a single process for transforming records into data because digital collections adopt different metadata schema

- Creating URIs for all our unique materials

- Finding ways to associate URIs to “things” in CONTENTdm

- Adopting linked data while it is in early stage of development
TIPS TO CONSIDER WHEN CREATING DIGITAL COLLECTIONS METADATA

- Avoid mixing different types of data in metadata fields
- Avoid creating aggregated data fields
- Record URIs whenever available
- Reinforce use of controlled vocabularies
- Monitor how CMS are adopting linked data technologies
HOW WE STARTED

- Created a study group in the Library (members from various areas of the library)
- Watched webinars on the topic and have discussions after the webinars
- Created an internal wiki with linked data resources
- Participated in linked data interest groups
- Follow the literature on this topic
QUESTIONS?

Contact Information:

Silvia Southwick
Silvia.Southwick@unlv.edu

Cory Lampert
Cory.Lampert@unlv.edu

Department of Digital Collections
UNLV Libraries