Trends in Linked Data adoption

Erik Mitchell PhD
Associate University Librarian
University of California, Berkeley

Agenda

- Discuss a model for evaluating metadata systems

- High-point comparative analysis for BIBFRAME, DPLA and Europeana

- Discuss methods for metadata exploration and analysis

1. A model for metadata research

Elings & Waibel Terminology	Schema examples	Common terminology
	RDF, Entity-rel, graph, key-value	Data model
Data content	CCO, RDA, DACS	Content rules, cataloging principle
Data structure	CDWA, MARC, ONIX, OWL	Metadata schema
Data format	XML, JSON, JSON-LD, RDFa	Encoding, serialization
Data exchange	OAI, Z39.50, SPARQL	

Metadata: Conceptual | Structural | Digital

Data model

Metadata schema

Vocabularies

Serialization

Exchange

Content rules

Metadata: Conceptual Structural Digital

Metadata Building Block	Definition
Data model	A data model is the way in which relationships between resources and their metadata and among resources are documented. In essence, the data model is the foundation on top of which the other components are built.
Content rules	Content rules govern how information is extracted or generated from resources and used to create a representation. Examples of content rules include RDA, CCO, DACS, and the IFLA *Statement of International Cataloguing Principles.*
Metadata schema/ vocabularies	Data structures and schemas govern how information extracted from the resource is described and stored in a metadata object.
Data serialization	Data formats and serializations are used to record the generated metadata and typically refer to some digital form of encoding.
Data exchange	Data exchange standards govern the sharing of metadata between systems.

A decoupled approach to metadata:

- Enables granular design and appropriately-scaled systems

- Follows a 'plug-in' architecture in sync with our information systems

- Supports direct data access and use - in sync with data science techniques

Data model
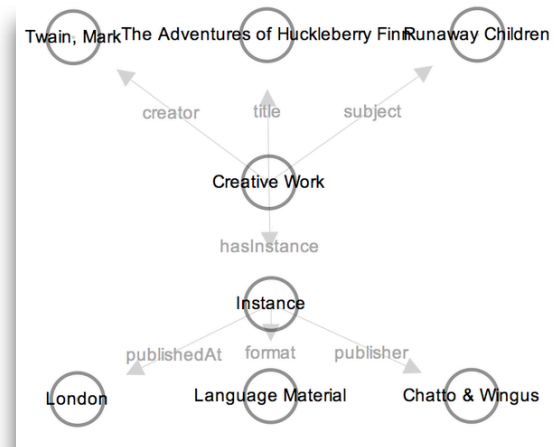
Vocabularies and serialization

```
1       <?xml version="1.0"?>

2       <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"

3       xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:bibo="http://purl.org/ontology/bibo/">

4       <rdf:Description rdf:about="http://lccn.loc.gov/35020965">

5       <dc:title>The Adventures of Huckleberry Finn</dc:title>

6       <dc:publisher
        rdf:resource="http://id.loc.gov/authorities/names/n85242407"/>

7       <dc:subject
        rdf:resource="http://id.loc.gov/authorities/subjects/sh2008110345"
        />

8       <rdf:type rdf:resource="bibo:Book"/>

9       </rdf:Description>

10         </rdf:RDF>
```

2. Comparative analysis: BIBFRAME, DPLA, Europeana

| Metadata Building Block	DPLA	BIBFRAME	Europeana
Broad organizational goal	"Brings together the riches of America's libraries, archives, and museums."	Develop a transition to metadata work and services that support engagement and querying of a network of data.	Bring together collections across European libraries under a common metadata schema and using a common licensing, indexing and dissemination platform.
Commonalities	Broad cross-community focus (LAM+P), open licensing, open data, API foundation, data cleaning tools		
Unique features	18 partners, 2 million records, unified discovery - distributed access	Metadata exchange focus, model development.	Coalition building through object archiving and metadata normalization
Licensing	Open data but no object aggregation	License agnostic	Open licensing, digital object aggregation

Metadata environment comparison

Metadata Building Block	DPLA	BIBFRAME	Europeana
Data model	RDF, faceted	RDF, FRBR-inspired	RDF, versioned
Content rules	Focus on locally defined faceted elements; embraces "big umbrella" approach to metadata aggregation.	RDA and other bibliographic standard foundation, but claims to be extensible.	Defines relationships between resources but is not overly prescriptive for content;
Metadata schema/ vocabularies	Simplified EDM model with DPLA-specific elements geared towards faceted browsing.	Relies heavily on the BIBFRAME vocabularies and Library of Congress endpoints.	EDM is highly prescriptive with structure, extensive use of external endpoints where possible.
Data serialization	JSON-LD	RDF/XML, JSON	RDF/XML
Data exchange	Primarily API-based, data downloads	Data transformation tools, no aggregation, dissemination tools.	SPARQL, APIs, data downloads

Metadata schema and vocabularies

Europeana
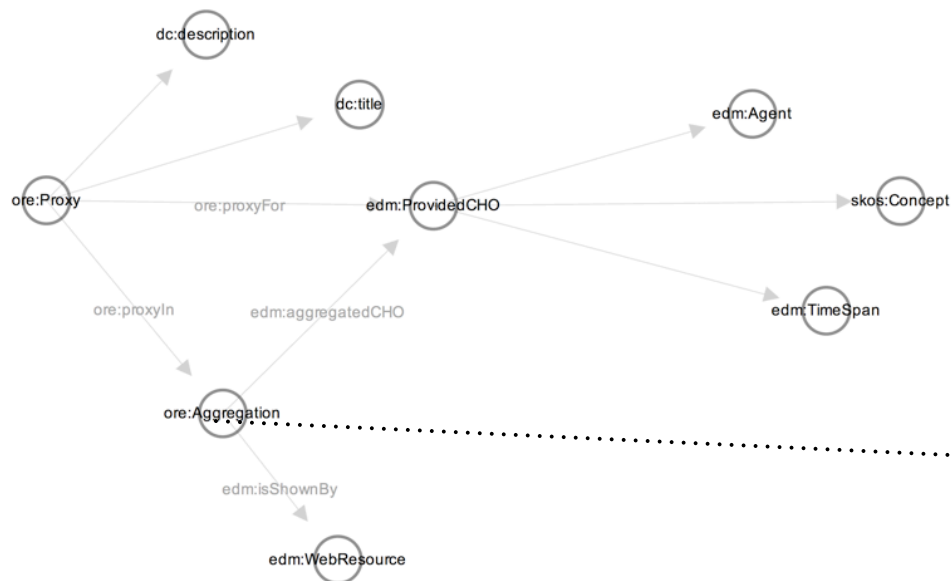
DPLA
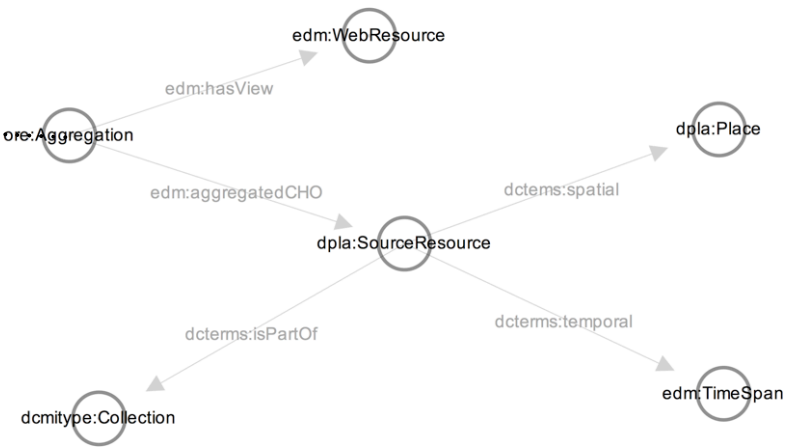
3. Methods for further study

- Community discourse analysis - particularly encouragement of social media capture and subsequent network analysis

- Metadata exploration through API use

- Metadata visualization - particularly for gaps and potential overlap relationships - which requires automated mapping with human review

Discourse: NISO Bibliographic roadmap meeting

Online survey	In person meeting	Notes and minute taking	Post meeting review

Document Name	Length (Words)
Areas for Discussion	2,296
Breakout Group—Business Models	570
Breakout Group—Goals	403
Breakout Group—Interoperability	1,220
Breakout Group—Open/Share	383
Breakout Group—Provenance/Authority	223
Breakout Group—Prototyping	428
Breakout Group—Rules	334
Breakout Group—Users	93
Day 2 Meeting Discussion	4,421
Discussion from Input Survey	922
Other Spare Notes	530

Discourse: Content themes

| Issues	Opportunities	Impact	Metadata Evaluation	Metadata Functions
1. Adoption barriers 2. Business models 3. Cost 4. Implementation 5. Institutional responsibility 6. Literacy issues 7. Migration 8. Open and contractual licensing 9. Organizational issues 10. Original cataloging in LOD 11. Personal privacy 12. Standards compliance 13. Sustainability 14. Technical issues 15. Timeline for implementation 16. Training (see literacy) 17. User needs evaluation	1. Community collaboration 2. Demonstration of value of libraries 3. Innovation 4. LAM collaboration 5. New research methods 6. Open data publishing 7. Patron engagement	1. Adoption 2. Community collaboration 3. Community vision 4. Organizational work 5. Staffing	1. Compatibility 2. Consistency 3. Data integrity/trust 4. Data-centric evaluation 5. Efficiency 6. Metadata value 7. Metrics-based evaluation 8. Provenance/ responsibility 9. Quality assessment 10. Sustainability 11. Use cases 12. User-centric evaluation	1. Aggregation 2. Computation 3. Data publishing 4. Deduplication 5. Discovery 6. Interoperability 7. Mapping 8. Metadata life cycle

API exploration

- Europeana

  - Rest-based, signed

  - Returns subset of database in RDF/XML, JSON, downloads

- DPLA

  - Rest-based, signed

  - JSON-LD

- BIBFRAME

  - Python and xquery tools

  - HTML browse able and JSON

```
▼ sourceResource {12}
      title : Club House, Otranto, front
   ▼ spatial [2]
      ▼ 0 {5}
            county : Charleston County
            name : Charleston County (S.C.)
            state : South Carolina
            coordinates : 32.8285713196, -79.8656082153
            country : United States
      ▶ 1 {4}
      description : Albumen ; 12 x 20 cm on 13.5 x 22 cm mount.
   ▼ subject [9]
      ▼ 0 {1}
            name : Buildings--Earthquake effects--South Carolina--Charleston--
                   Photographs
      ▶ 1 {1}
      ▶ 2 {1}
      ▶ 3 {1}
      ▶ 4 {1}
      ▶ 5 {1}
      ▶ 6 {1}
      ▶ 7 {1}
      ▶ 8 {1}
      rights : Digital image copyright 2010, The University of South Carolina.
               All rights reserved. For more information contact The South
               Caroliniana Library, USC, Columbia, SC 29208.
      relation : Charleston Earthquake 1886
   ▶ language [1]
      format : Images
   ▶ collection {3}
   ▶ date {3}
      type : image
      creator : Cook, Geo. L. (George L.), photographer
   ingestType : item
```

curl -G "http://api.dp.la/v2/collections?api_key=privatekey" > collout.json

Data visualization

- Gephi

  - Rdf, JSON, csv importers

  - Network analysis and visualization

- W3C RDF Validator

  - RDF/XML focus

  - Graph display

- Approaches

  - Manual modeling of small sets

  - Large scale network analysis
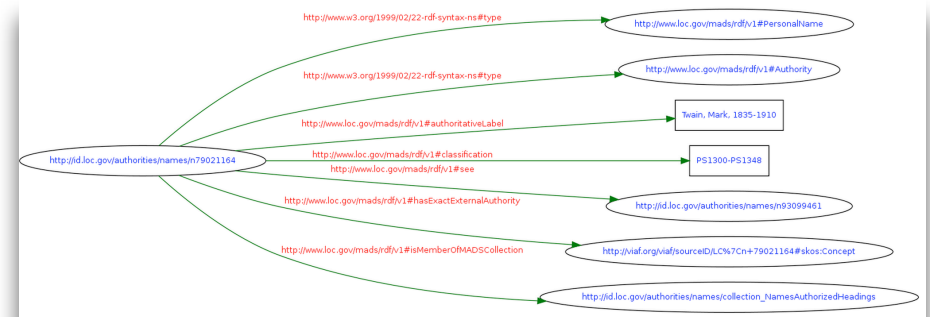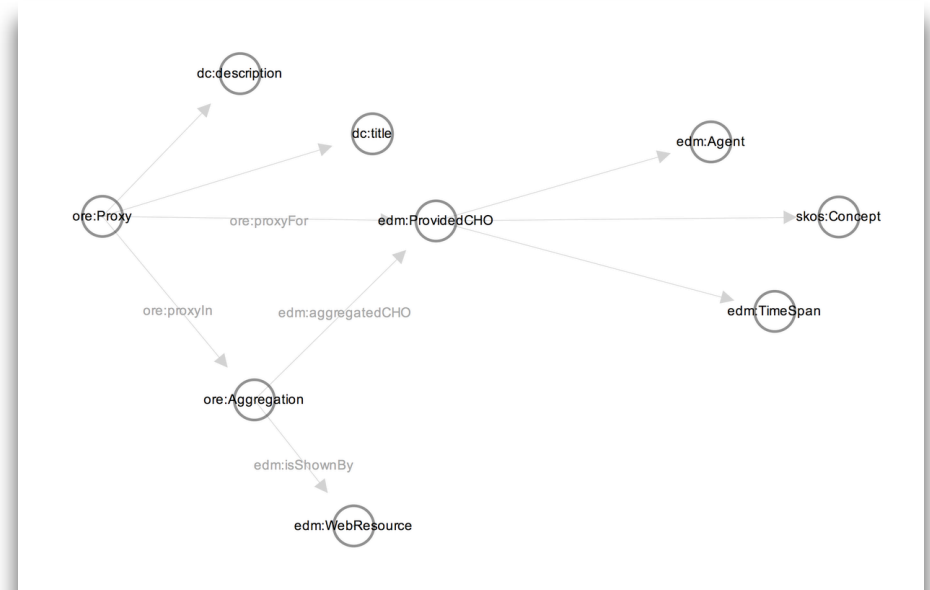
Thoughts and next steps

- Community discourse analysis

  - Themes in NISO exploration showed broad community involvement and impact

  - Potential for social media, listserv content capture and subsequent network analysis

- Metadata exploration through API use

  - Converging serialization and exchange standards are simplifying cross-community analysis

  - Tutorial development (check github https://github.com/mitcheet/ltr)

- Metadata visualization

  - First pass showed explicit and implicit connections

  - gaps and potential overlap relationships

- Early in Linked Data adoption curve

Erik Mitchell
mitcheet@gmail.com
http://erikmitchell.info
@mitcheet