

LC's Authorities and Vocabularies Web Service : experimenting with Linked Data

Rebecca Guenther
Network Development &
MARC Standards Office,
Library of Congress
rgue@loc.gov

**LC booth/ALA Midwinter
Jan. 9, 2011**



Outline of presentation

- Types of controlled vocabularies
- Vocabularies maintained at LC
- How to encode controlled vocabularies
- What is SKOS?
- What is Linked Data?
- id.loc.gov vocabulary services
- Some examples of concept schemes
- Next steps

ALA
Midwinter
2011



Why establish controlled vocabularies?

- Control values that occur in metadata
- Reduce ambiguity
- Control synonyms
- Document and publish for reuse
- Test and validate terms
- Establish formal relationships among terms (where appropriate)

ALA
Midwinter
2011



Types of Controlled Vocabularies used in metadata standards

- Lists of enumerated values
- Code lists (e.g. language, country)
- Taxonomies
- Formal Thesauri
- Locally controlled enumerated lists

ALA
Midwinter
2011



Enumerated lists

- Simple list of terms used in a pull-down menu or Web site pick list
- Values enumerated in an XML schema
- Little additional information or structure about each value
- Examples:
 - Enumerated value “MD5” for METS CHECKSUMTYPE
 - Enumerated value “born digital” in MODS digitalOrigin
 - Code and value from a MARC 21 fixed field, e.g. code “e” in Leader/06 is “cartographic material”

ALA
Midwinter
2011



Code lists

- Some established as ISO standards and used worldwide in many communities for many purposes
- The standard standardizes the code, not a particular name for it
- Codes are used as identifiers
- Some examples:
 - ISO 639 family (language codes)
 - MARC relator codes
 - MARC country codes
 - ISO 3166 country codes

ALA
Midwinter
2011



Thesauri

- A *thesaurus* is a controlled vocabulary with multiple types of relationships

Example:

Rice

UF Paddy

BT Cereals

BT Plant products

NT Brown rice

RT Rice straw

ALA
Midwinter
2011



Standards maintained at LC that contain controlled vocabularies

- LCSH/NAF
- Thesaurus of Graphic Materials
- MARC Code lists: GACs, countries, languages, relators
- ISO 639-2 and ISO 639-5 (language codes)
- Other MARC controlled lists
- Enumerated lists in XML schemas
 - MODS enumerated values
 - METS enumerated values
 - MIX (Technical metadata for digital still images)
- PREMIS controlled vocabularies
- Others...

ALA
Midwinter
2011



How to encode information about controlled values

- Established metadata formats, e.g. MARC Authority format
- XML schemas (as enumerated values)
- RDF/XML and RDFS (Resource Description Framework)
- SKOS (Simple Knowledge Organization System)
- MADS (Metadata Authority Description Schema)

ALA
Midwinter
2011



About SKOS

- Simple Knowledge Organization System
- RDF application used to express knowledge organization systems such as classifications, thesauri, taxonomies, and the concepts within.
- Allows distributed, decentralized management of KOS through Linked Data-inspired application.
- All concepts and schemes require a URI

ALA
Midwinter
2011



The SKOS data model (Classes)

- Concept Schemes (e.g., published vocabularies, thesauri, code lists, etc.)
- **Concepts** (individual entries or terms within the larger vocabulary)
- Collections (logical groupings of Concepts)



SKOS concepts

- Labeling properties: **prefLabel**, **altLabel**, **hiddenLabel**, **notation**
- Annotation properties: **note**, **historyNote**, **scopeNote**, **changeNote**, **editorialNote**, **example**, **definition**
- Associative properties: **broader**, **narrower**, **related**, **sameAs**, **broadMatch**, **narrowMatch**, **closeMatch**, **exactMatch**, **minorMatch**, **majorMatch**

ALA
Midwinter
2011



Advantages to using SKOS

- SKOS has a defined element set which is particularly relevant for controlled vocabularies
- Relationships between entries in a concept scheme can be expressed (broader, narrower, etc.)
- Relationships between entries in different concept schemes can be expressed (exactMatch, related)
- Having a dereferencable URI for concepts and their concept schemes enhances the ability to provide web services for consumers of these standards

ALA
Midwinter
2011



“Linked Data”

- A feature of the “Semantic Web” where links are made between resources
- Goes beyond hypertext links (i.e. between web pages) but between any kind of object or concept
- From Wikipedia: "a term used to describe a method of exposing, sharing, and connecting data via dereferenceable URIs on the Web"
- Users can use links to find similar resources and aggregate results

ALA
Midwinter
2011



Reasons for developing a web service for vocabularies

- Expose LC developed vocabularies to wider communities
- Make controlled lists openly available
- Provide comprehensive information about controlled terms
- Experiment with semantic web technologies and linked data
- Facilitate development and maintenance process for vocabularies

ALA
Midwinter
2011



Introducing id.loc.gov

- Allows both human-oriented and programmatic access to LC-promulgated authorities and vocabularies.
- First offering was Library of Congress Subject Headings in April 2009, additional vocabularies added April 2010 and Jan. 2011
- Offers bulk data downloads in several RDF serializations
- Only serves data values: authority and vocabulary data, not bibliographic

ALA
Midwinter
2011



Available vocabularies

- LCSH
- Thesaurus for Graphic Materials
- MARC Code Lists for Relators, Languages, Countries, Geographic Areas
- ISO 639 language codes
- Preservation metadata vocabularies
 - Preservation events
 - Cryptographic hash functions (METS, PREMIS and MIX)
 - Preservation level role

ALA
Midwinter
2011



URIs in id.loc.gov

- Interaction with any given individual term and vocabulary is with its URI
- Some examples of URIs:

<http://id.loc.gov/vocabulary/relators/art>

<http://id.loc.gov/vocabulary/graphicMaterials/tgm005222>

<http://id.loc.gov/vocabulary/preservationEvents/migration>

<http://id.loc.gov/authorities/sh85063136>

- Known-label searches: use when you know the label but not the identifier

<http://id.loc.gov/vocabulary/relators/label/artist>

<http://id.loc.gov/authorities/label/hunting%20dogs>

ALA
Midwinter
2011



Other features of id.loc.gov

- Can search terms in one or multiple vocabularies: Arabic in ISO 639-1, 639-2 and MARC languages
- Visualizations
- Links to similar concepts in other vocabularies (e.g. Rameau)
- Suggest terminology (for LCSH)
- Can download vocabulary in several formats

ALA
Midwinter
2011

ID: how it's being used

- **John Ocklerbloom, University of Pennsylvania, Online Books**
 - Used LCSH bulk download, manipulated for local needs
 - Integrated with OPAC records
 - Broadens user search/discovery experience
- **Ethan Gruber, University of Virginia**
 - Data used in EAD editor as autosuggest feature
- **Freebase.com**
 - Bulk data loaded to freebase.com
- **National Library of Sweden**
 - Unreleased/published Linked Data component of catalog
 - Links to LCSH
- **RAMEAU**
 - Part of MACS project
 - Determined skos:closeMatch resources between RAMEAU and LCSH
 - Both RAMEAU and LCSH cross linked



University of Pennsylvania

Medicine, Botanic

Here are entered works on a 19th century system of medicine developed by Samuel Thomson and based on the use of plant remedies.

Broader terms:

- [Alternative medicine](#)
- [Medicine](#)

Used for:

- Botanic medicine
- Thomsonianism

Filed under: [Medicine, Botanic](#)

① [A Guide to Health, Being an Exposition of the Principles of the Thomsonian System of Practice, and Their Mode of Application in the Cure of Every Form of Disease \(1846 edition\)](#), by Benjamin Colby (text and PDF with commentary at swsbn.com)

① [Life and Medical Discoveries of Samuel Thomson, and a History of the Thomsonian Materia Medica, As Shown in "The New Guide to Health" \(1835\), and the Literature of That Day \(1909\)](#), by John Uri Lloyd and Samuel Thomson (PDF at swsbn.com)

Items below (if any) are from related and broader terms.

Filed under: [Alternative medicine](#)

① [Natural Liberty: Rediscovering Self-Induced Abortion Methods \(Creative Commons licensed edition, c2008\)](#), by Sage-Femme Collective (Javascript-dependent Flash at scribd.com)



ALA
Midwinter
2011



Rameau

Concept information

URI	http://stitch.cs.vu.nl/vocabularies/rameau/ark:/12148/cb14521343b	
prefLabel	x-notation	FRBNF145213438
	fr	Web sémantique
note	fr	Domaine : 621
inScheme	Rameau	
	Rameau - Noms Communs	
broader	Web	
related	Ontologies (informatique)	
	Services Web	

Mappings (simple SKOS statements)

Mapping Relation	Concept
closeMatch	http://id.loc.gov/authorities/sh2002000569#concept

.A
...dwinter
2011



Example: MARC Code List for Relators

- Available in HTML and XML from MARC website
- LC is maintenance agency
- URIs established for concept scheme and individual concepts
- Statements made to assert relationship to Dublin Core Contributor

ALA
Midwinter
2011

```

<rdf:Description
  rdf:about="http://id.loc.gov/vocabulary/relators/art">
  <rdf:type
    rdf:resource="http://www.w3.org/2004/02/skos/core#Concept" />
    <rdf:type rdf:resource="http://www.w3.org/
      1999/02/22-rdf-syntax-ns#Property"/>
    <owl:sameAs rdf:resource="info:lc/vocabulary/relators/art" />
    <skos:prefLabel xml:lang="en">Artist</skos:prefLabel>
    <skos:notation xml:lang="zxx">art</skos:notation>
    <skos:scopeNote xml:lang="en">Use for a person (e.g., a
painter) or organization who conceives, and perhaps also
implements, an original graphic design or work of art, if
specific codes (e.g., [egr], [etr]) are not desired. For book
illustrators, prefer Illustrator [ill].
  </skos:scopeNote>
  <skos:inScheme
    rdf:resource="http://id.loc.gov/vocabulary/relators" />
    <vs:term_status>stable</vs:term_status>
    <skos:altLabel xml:lang="en">Graphic
technician</skos:altLabel>
    <skos:note xml:lang="en">Relator term "Graphic technician"

```



Additional vocabularies coming

- Other PREMIS controlled vocabularies
- Name authorities
- MARC Code List for Sources
- MARC Code List for Organizations
- Others

ALA
Midwinter
2011



Next steps

- Enhance existing vocabularies to show relationships
 - Broader/narrower relator terms
 - Matches to other vocabulary terms (e.g. MARC vs. ISO 3166 country codes, language codes to subject headings, etc.)
 - Add additional vocabularies
- MADS OWL Schema to identify facets within name and subject authorities
- Ontology is out for review at: <http://www.loc.gov/standards/mads/rdf/>

ALA
Midwinter
2011

Current markup in SKOS/XML:

<http://id.loc.gov/authorities/sh85140205>

```
<rdf:Description rdf:about="http://id.loc.gov/authorities/sh85140205#concept">
  <skos:inScheme rdf:resource="http://id.loc.gov/authorities#conceptScheme"/>
  <skos:inScheme rdf:resource="http://id.loc.gov/authorities#geographicNames"/>
  <skos:narrower rdf:resource="http://id.loc.gov/authorities/sh85092617#concept"/>
  <skos:narrower rdf:resource="http://id.loc.gov/authorities/sh2010014605#concept"/>
  <skos:narrower rdf:resource="http://id.loc.gov/authorities/sh94003150#concept"/>
  <skos:prefLabel xml:lang="en">United States--History--Civil War, 1861-1865</skos:prefLabel>
  <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
  <skos:altLabel xml:lang="en">Civil War, U. S., 1861-1865</skos:altLabel>
  <skos:altLabel xml:lang="en">War between the States, 1861-1865</skos:altLabel>
  <skos:altLabel xml:lang="en">War of Secession, U.S., 1861-1865</skos:altLabel>
  <skos:altLabel xml:lang="en">War of the Rebellion, 1861-1865</skos:altLabel>
  <skos:altLabel xml:lang="en">American Civil War, 1861-1865</skos:altLabel>
  <skos:altLabel xml:lang="en">Southern States--History--Civil War, 1861-1865</skos:altLabel>
  - <dcterms:source xml:lang="en">
    United States. War Dept. War of the Rebellion, 1880-1901.
  </dcterms:source>
  <owl:sameAs rdf:resource="info:lc/authorities/sh85140205"/>
  <skos:example xml:lang="en">Example under Civil war</skos:example>
  <dcterms:modified rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2000-02-25T13:10:51-04:00</dcterms:modified>
  <dcterms:created rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">1999-12-13T00:00:00-04:00</dcterms:created>
</rdf:Description>
```

Enhanced markup in MADS/RDF:

Example: ComplexSubject Authority

```
<madsrdf:Authority rdf:about="http://United_States--History--Civil_War,_1861-1865">
  <rdf:type rdf:resource="http://id.loc.gov/ontologies/mads/2010/11#ComplexSubject"/>
  <madsrdf:authoritativeLabel>United States--History--Civil War, 1861-1865</madsrdf:authoritativeLabel>
  <madsrdf:componentList rdf:parseType="Collection">
    <madsrdf:Geographic rdf:resource="http://United_States"/>
    <madsrdf:Topic rdf:resource="http://History"/>
    <madsrdf:Temporal rdf:resource="http://Civil_War,_1861-1865"/>
  </madsrdf:componentList>
</madsrdf:Authority>
```



Technical infrastructure

- Django (Python) framework
- LCSH
 - MySQL
 - SKOS RDF generated at time of request
 - Operates, more or less, as traditional relational DB
 - MARC mapped to relational DB tables
- Everything else
 - RDFlib (Python library, uses MySQL as triplestore)
 - Runs on triples
 - XML to SKOS RDF/XML before ingest
 - XSL, Xquery used

ALA
Midwinter
2011



Questions?

- Contact rgue@loc.gov
- Join the discussion list:
<http://listserv.loc.gov/listarch/id.html>

ALA
Midwinter
2011