Metadata: Key Concepts for Digital Collections

Summer Institute for Digital Scholarship: 6/16/14

Emily Porter, Metadata Analyst
Woodruff Library, Emory University
Session Outline

Overview of important concepts for metadata in digital collections & systems

- Metadata Types, Communities, Contexts
- Schemas: Concepts and Terms
- Implementing: Platforms and Systems
- Dublin Core
- Web Findability/Search Engine Optimization
- Trends and Changing Practices
- References and Resources
Communities and Domains

• Specific schemas and standards across domains
  – Cultural Heritage:
    • Galleries/Libraries/Archives/Museums (GLAM)
  – Scientific
  – Geospatial
  – Media/Broadcasting/Audiovisual
  – Educational/Learning Management
  – Web/Social Media

Seeing Standards: A Visualization of the Metadata Universe (Jenn Riley, Devin Becker)
<table>
<thead>
<tr>
<th>Cataloging/MARC</th>
<th>Digital Objects’ Metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Library cataloging, as a form of metadata, has traditionally had well-defined goals…” (Karen Coyle)</td>
<td>Parallel practices emerged in non-Library domains</td>
</tr>
<tr>
<td>The catalog</td>
<td>Highly contextual and often highly customized:</td>
</tr>
<tr>
<td>Benefits from mature encoding standards, well-defined content standards and system contexts</td>
<td>- Specific collection/content focused</td>
</tr>
<tr>
<td>Bibliographic, published content</td>
<td>- End user/interface needs</td>
</tr>
<tr>
<td>Established infrastructure for large-scale sharing of records</td>
<td>- Wide variety of destination systems/environments</td>
</tr>
<tr>
<td></td>
<td>Granularity/hierarchy of objects in a collection</td>
</tr>
<tr>
<td></td>
<td>- Group photo on a page within a scrapbook</td>
</tr>
<tr>
<td></td>
<td>- Unpublished content / provenance</td>
</tr>
</tbody>
</table>
## Types of Metadata: Cultural Heritage

<table>
<thead>
<tr>
<th>Type</th>
<th>Notes</th>
<th>Schemas/Standards</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Descriptive</strong></td>
<td>Primary focus for today’s overview: search and discovery metadata</td>
<td>Dublin Core; MODS; VRA, PBCore; EAD; MARC/RDA</td>
</tr>
<tr>
<td><strong>Administrative/Technical/Rights</strong></td>
<td>Information about the digital object in a system</td>
<td>Administrative: METS; Repository-generated</td>
</tr>
<tr>
<td>(Repository/digitization)</td>
<td>How it was generated</td>
<td>Technical: MIX; PBCore</td>
</tr>
<tr>
<td></td>
<td>How it can be used/accessed/modified</td>
<td>Rights: PREMIS; METS</td>
</tr>
<tr>
<td><strong>Structural</strong></td>
<td>How to reconstruct a complex digital object</td>
<td>METS/ALTO; Repository-generated</td>
</tr>
<tr>
<td>(Repository/digitization)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Preservation</strong></td>
<td>Tracks changes to a digital object over time</td>
<td>PREMIS; Repository-generated (audit trail)</td>
</tr>
</tbody>
</table>
Considerations: Planning for a Project

- **Which schema should I use?**
  - Depends on system constraints, local practices, content type needs

- **Who will create the metadata?**
  - Does metadata already exist? Can it be re-used?
  - Data entry resources; training and support; usability of interface
  - Consistency and standards
  - Uniqueness of resource/preservation needs – is this the only record?

- **What do the users need?**
  - Display labels (vs. schema element names)
  - Search result screens
  - Search, browse, filter, sort
  - Special navigation/presentation needs
  - Knowledge domain/vocabulary needs
  - Basic web discovery, or library catalog integration?
  - Individual item descriptions, or project/collection level?

- **Documentation**
Schemas:
Concepts and Terms
What’s a “Schema”? 

• **Element**
  – Single, distinct metadata “field” or property in a resource description  
  – Generally implemented as name/value pairs (sometimes with labels)
    • Creator: Doe, John  

• **Element Set**
  – Set of metadata elements which together compose a larger content standard or schema  

• **Schema**
  – Formally structured, machine-processable Element Set (often XML)  
  – Collection of elements with documented usage requirements, data types, and data entry conventions  
  – Typically require technical validation against XML, DTD/XSD  

• **Application Profile**
  – Organizational or consortial interpretation of a schema  
  – Local practices and business rules for how to utilize a schema
    • Dublin Core Library Application Profile  

• **Data Dictionary**
  – Technical documentation for a schema or element set
Metadata Encoding

- **Element Data Types**
  - Specify what type of information can be stored in a field
    - String (text – but can be alphanumeric)
    - Date/Time (conforms to a date encoding scheme)
    - Number (can be manipulated as a number)
    - URI (URL)

- **Element Cardinality**
  - Required, optional
  - Repeatable (and minimum/maximum entries)

- **Data Formats**
  - The storage and encoding format for metadata
    - Database
    - Flat file (tab-delimited, CSV)
    - XML
    - HTML

- **Validation**
  - Technical error-checking of the markup or values

- **Crosswalks/Transformations**
  - Mapping elements from one standard to another (documentation)
  - Technical transformation of data elements from one standard to another (e.g. XSLT)
Populating Elements in a Schema

• **Content standards**
  – Set of guidelines for populating fields
  – Recommended values, element usage (which may be more specific than the schema indicates), punctuation/style guide
  – Often evolve around particular content types
    • Examples: RDA, DACS, CCO, AACR2, CSDGM

• **Value standards** (see DC cheat sheet)
  – Rules/constraints for how the data can be entered
  – Re-usable across schemas/standards
  – May correlate to data types
    • Date formats (e.g. W3CDTF: YYYY-MM-DD)
  – **Controlled vocabularies and authorities**
    • Restrict entries to a controlled list of terms
    • Enhance consistency and data entry entry (e.g. consistent names)
    • Maintained by formal bodies with subject matter expertise
    • Enable batch updates of values over time
    • Examples: Library of Congress Subject Headings, Name Authority File
Implementing: Platforms and Systems
# Metadata in System Environments

<table>
<thead>
<tr>
<th>Environment</th>
<th>Product Examples</th>
<th>Metadata Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILS/Cataloging</td>
<td>ExLibris Voyager, ExLibris Alma</td>
<td>• Mostly MARC focused&lt;br&gt;• Some DC/MODS&lt;br&gt;• Controlled vocabulary integration</td>
</tr>
<tr>
<td>Content/Asset Management</td>
<td>Drupal, Cascade Server, Extensis Portfolio Server, ContentDM, WordPress, Omeka</td>
<td>• HTML metadata&lt;br&gt;• Some Dublin Core&lt;br&gt;• Some custom metadata</td>
</tr>
<tr>
<td>(CMS/DAMS)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discovery</td>
<td>ExLibris Primo*, EBSCO Discovery Service</td>
<td>• OAI ingest&lt;br&gt;• Map incoming data (misc. XML*; Dublin Core*; MARC/MARCXML)</td>
</tr>
<tr>
<td>Presentation</td>
<td>LUNA, ArtStor, ContentDM, Omeka</td>
<td>• Dublin Core&lt;br&gt;• Some custom schemas&lt;br&gt;• Some VRA&lt;br&gt;• Some OAI export</td>
</tr>
<tr>
<td>Repository</td>
<td>DSPACE, ContentDM, Rosetta, Fedora, ePrints</td>
<td>• Package/create desc. schemas&lt;br&gt;• Custom pres. schemas (PREMIS, METS, other)&lt;br&gt;• OAI ingest/export</td>
</tr>
</tbody>
</table>
ECDS Institute Tools

• **Tour app builder**
  – HTML titles and descriptions: stop pages, homepage, splash page
  – Basic image/video metadata
  – Twitter and Facebook metadata for “stops”

• **WordPress**
  – Limited metadata entries without plugins (.com account)
  – Add Meta Tags Plugin enables qualified DC, Facebook, Twitter, Schema.org
  – Standard WordPress themes are optimized for general SEO best practices

• **Omeka**
  – Support for Dublin Core, other schemas (if self-hosting)
  – Metadata-oriented plugins (.net account):
    • CSV import
    • Google Analytics
    • Library of Congress Suggest
    • COinS
    • OAI harvester
    • Shared Shelf
    • (additional plugins for self-hosting)
  – Tags as categorization/discovery options
METADATA

Dublin Core
Dublin Core 101

• Conceived in 1995 in as a “core set of semantics for web-based resources” to enhance search and retrieval in response to rapidly growing web

• Benefits:
  – Universal translator: understood by many systems; common mapping point for metadata across schemas
  – Relatively small element set (easy to populate)
  – Flexible and easy to encode as XML, embed in HTML, or store in a spreadsheet
  – Elements utilized by other schema (PB Core, Darwin Core)
Dublin Core: Challenges

• 1:1 principle - what are you describing?
  – Digital copy vs. analog original vs. born digital original
  – Strict meaning: only describe the *digital* instance
    • Digital Library Federation best practice: use 1:1 “when practical”
    • Consider context for end users; local practices; consistency across collections
      • Is this the canonical metadata record?

• Lack of specificity; ambiguity of some elements
• Looseness of requirements (all optional, repeatable)
• Loss of data when mapping from richer schemas
Dublin Core: Simple vs. Qualified

• Simple Dublin Core
  – 15 “core” elements (no attributes or qualifiers)
  – Issues of interpretation: date, relation, coverage
  – No way to specify attributes for controlled vocabulary names/encoding schemes

• Qualified Dublin Core
  – Refinements that add context to core elements
    • date.created
    • relation.isPartOf
    • coverage.spatial
    • format.extent
  – Qualified Dublin Core is “deprecated”; evolved into DCMI Terms Vocabulary
Dublin Core: Elements (aka “Simple”)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Title</td>
</tr>
<tr>
<td>2.</td>
<td>Description</td>
</tr>
<tr>
<td>3.</td>
<td>Type</td>
</tr>
<tr>
<td>4.</td>
<td>Subject</td>
</tr>
<tr>
<td>5.</td>
<td>Coverage</td>
</tr>
<tr>
<td>6.</td>
<td>Relation</td>
</tr>
<tr>
<td>7.</td>
<td>Source</td>
</tr>
<tr>
<td>8.</td>
<td>Creator</td>
</tr>
<tr>
<td>9.</td>
<td>Contributor</td>
</tr>
<tr>
<td>10.</td>
<td>Publisher</td>
</tr>
<tr>
<td>11.</td>
<td>Rights</td>
</tr>
<tr>
<td>12.</td>
<td>Date</td>
</tr>
<tr>
<td>13.</td>
<td>Format</td>
</tr>
<tr>
<td>14.</td>
<td>Identifier</td>
</tr>
<tr>
<td>15.</td>
<td>Language</td>
</tr>
</tbody>
</table>

See Dublin Core Cheat Sheet for more detail
Dublin Core: Elements + Qualifiers

- abstract
- accessRights
- accrualMethod
- accrualPeriodicity
- accrualPolicy
- alternative
- audience
- available
- bibliographicCitation
- conformsTo
- contributor
- coverage
- created
- creator
- date
- dateAccepted
- dateCopyrighted
- dateSubmitted
- description
- educationLevel
- extent
- format
- hasFormat
- hasPart
- hasVersion
- identifier
- instructionalMethod
- isFormatOf
- isPartOf
- isReferencedBy
- isReplacedBy
- isRequiredBy
- issued
- isVersionOf
- language
- license
- mediator
- medium
- modified
- provenance
- publisher
- references
- relation
- replaces
- requires
- rights
- rightsHolder
- source
- spatial
- subject
- tableOfContents
- temporal
- title
- type
- valid
Dublin Core: Points of Confusion

• Type vs. Format
  – Type = original content
  – Format = physical or digital file characteristics
  – Digitized page of a letter
    • Type = text
    • Format = image/tiff

• Source vs. Relation
  – Source = original source (generally analog) from which digital copy/excerpt was made
  – Relation = multiple types of relationships

• Creator vs. Contributor
  – Primary vs. secondary role in creation

• Date vs. Coverage
  – Date related to content lifecycle vs. date within content
Dublin Core: Missing Concepts

Elements/concepts that are hard to express in DC, but available in other schemas:

- Roles for creators/contributors: editor, photographer, cinematographer, advisor, donor
- Audience
- Edition or version
- Place of publication
- Location (museum/repository/library name)
- Style/genre/material technique
- Notes
- Metadata record information
Sample Record: Digitized Map

**Identifier:** [http://digitalgallery.emory.edu/luna/servlet/detail/EMORYUL~3~3~2187~100303](http://digitalgallery.emory.edu/luna/servlet/detail/EMORYUL~3~3~2187~100303)

**Relation:** Atlas of Atlanta and Vicinity, 1928

**Coverage:** Atlanta

**Coverage:** United States

**Coverage:** Fulton County; Cobb County

**Coverage:** Georgia

**Creator:** U.S. Coast and Geodetic Survey

**Date:** 1930

**Description:** Color map showing Pleasant Hill Church and Chattahoochee River.

**Format:** image/jpeg

**Title:** City of Atlanta: Sheet 51. Construction Department, William A. Hansell, Chief; S.P. Floore, Topographic Engineer in charge. Topography by E.J. Essick and J.B. Leachman. Control by U.S. Coast and Geodetic Survey and City of Atlanta Mapping Division. Surveyed in 1927. Williams & Heintz Co., Wash, D.C.

**Publisher:** U.S. Coast and Geodetic Survey and City of Atlanta Mapping Division

**Rights:** The City of Atlanta has granted Emory University, Woodruff Library, permission to digitize, distribute, display and geo-reference maps produced by the U.S. Coast Guard and Geodetic Survey and the City of Atlanta Mapping Division in a 1928 survey published as the Atlas of Atlanta and Vicinity. Emory may digitize, display, and georeference the maps in electronic formats, including free public access to maps on the web. The City of Atlanta does not attest to the accuracy of the image. The Maps Content, including all images and text, are for personal, educational, and non-commercial use only.

**Type:** Atlas Map

**Type:** Image
Sample Record: Digitized Photo

Identifier: MSS1218_B001_I117_P0002
Title: Surveying crew staking out a grove, 1936
Description: Recto: Surveying crew staking out a grove 1936, [Tung Grove Development Co., Florida]
Date: 1936
Subject: African American men.
Subject: Employees.
Subject: Surveying.
Type: Image
Type: Black-and-white photographs
Format: 05.61 x 07.47 inches
Related: Robert Langmuir African American Photograph Collection, MSS1218, Manuscript, Archives, and Rare Book Library, Emory University
Rights: Emory University does not control copyright for this image. This image is made available for individual viewing and reference for educational purposes only such as personal study, preparation for teaching, and research. Your reproduction, distribution, public display or other re-use of any content beyond a fair use as codified in section 107 of US Copyright Law is at your own risk. We are always interested in learning more about our collections. If you have information regarding this photograph, please contact marbl@emory.edu.
Sample Record: Web Application

Title: Battle of Atlanta
Creator: Pollock, Daniel
Subject: American Civil War (1861-1865)
Subject: Battlefields
Subject: Battle of Atlanta
Description: This website is a resource for getting in touch with the Battle of Atlanta in history and in memory. It is a means of engaging with the recorded past and the remembered past of a particularly fierce fight between North and South on July 22, 1864. This site combines a narrative of battlefield events with images from the Cyclorama and other visual and textual artifacts into a digital tour guide with contextual links.
Publisher: Emory Center for Digital Scholarship, Emory University, Atlanta, GA
Contributor: Tullos, Allen
Contributor: Varner, Jay
Contributor: Croxall, Brian
Date: 2013
Type: InteractiveResource
Format: text/html
Identifier: http://dev.emorydisc.org/battleofatlanta/tour/battle-of-atlanta/
Language: English
Coverage: 1864
Populating Dublin Core – General Tips

• Use structured entries over free-text when possible
  – Controlled vocabularies, URLs, date formats etc.

• If an element’s value is unknown, don’t create it (avoid “Unknown” or blank element entries)
  – Publisher: unknown

• Create separate element instances for multiple entries
  – <subject>Term 1</subject>
  – <subject>Term 2</subject>
  – Vs. <subject>Term 1; Term 2</subject>

• (See the Dublin Core cheat sheet for more guidelines)
Harvesting and Sharing with OAI-PMH

- Open Archives Initiative Protocol for Metadata Harvesting
- Standard data mechanism for importing/exporting records across systems or organizations (XML)
- Many systems automate OAI
- Simple Dublin Core is always required within OAI
  - You can also transmit your records in additional formats, too (e.g. supply both MODS and DC versions of records)

```xml
OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI
<responseDate>2014-06-15T10:41:24Z</responseDate>
<request verb="ListRecords" set="EMORYUL~3~3" metadataPrefix>
- <ListRecords>
- <record>
  - <header>
    - <identifier>
oai:artimages.service.emory.edu:EMORYUL~3~3~2762~:
    </identifier>
  - <datestamp>2012-10-12T11:48:12Z</datestamp>
  - <setSpec>EMORYUL~3~3</setSpec>
  </header>
  - <metadata>
    - <oai_dc:dc xsi:schemaLocation="http://www.openarchives
      - <dc:identifier>
        http://digitalgallery.emory.edu/luna/servlet/detail/EMOF
      </dc:identifier>
      - <dc:identifier>
        http://artimages.service.emory.edu:8086/MediaManager
      </dc:identifier>
      - <dc:relation>
      </dc:relation>
      <dc:coverage>United States</dc:coverage>
```
Web Find-ability: Search Engine Optimization, Social Media Metadata
Guidelines for SEO

• **Filenames**
  – Short and sweet: human and machine readable
  – Google prefers hyphens to separate words

• **Use good document structure**
  – Use heading tags (<h1>, <h2> etc.)
  – Make sure your content level title is formatted as a Heading 1/</h1>
  – Google looks at the document as a whole

• **Title <title>**
  – Shown as link text in search results (~70 chars)
  – Provide unique titles for all pages
  – Best practice – include web site name appended to end of individual resource title

• **Description <meta name="description">**
  – Google uses this as the excerpt in search results (~160 chars)

• **HTML 5**
  – More restrictive on <meta> tags’ name attributes ([new registry](#))

• **Google webmaster account**
  – Google sitemap – XML outline of your site (helps Google discover your content)
  – Track and refresh Google crawls of your site
  – Test structured data for rich snippets

• **Google Analytics account**
  – Track visitor demographics, content usage
  – Configure site search for your account to see search terms

• **Maturity of site; incoming links**
  – The longer and more established your site is, the better
  – When re-designing a site, make sure to provide redirects to your most visited pages
Social Media Metadata

- Emerging, proprietary standards offer additional context for sharing/discovery in social media platforms
  - Override and enhance default values
  - Some re-usability across platforms
  - Can be automated
- **Twitter Cards**: customize thumbnail, description, title, creator information, type of content
  - Tour app integration
- **Facebook OpenGraph**: customize thumbnail, description, title, language/locale, type of content
  - Tour app integration
- **Pinterest – Rich Pins**: product information, related links
Search Engine Initiatives

- Shift from `<meta>` tags, keywords, Dublin Core in `<head>`
- Schema.org
  - Google, Bing, Yahoo
  - Creative Work definition
  - Person definition
- Microdata, microformats, RDFa
  - Embedded structured metadata as attributes within your HTML content
- Rich snippets
  - More contextual display in search results
- Google Structured Data tool (for testing)
Google Scholar – Tactics for Inclusion

• Very specific markup requirements
• Prefers <meta> tags in one of the following formats:
  – Highwire Press
  – Eprints
  – BE Press
  – PRISM
• Recommends *against* use of Dublin Core (not granular enough)
• Crawling/indexing challenges when mixing scholarly articles with other types of content
Trends and Evolving Practices
Trends

- **RDF/Linked Data**
  - Web model: making connections across distributed data vs. creating one-off records
  - Opening up closed systems
  - Shifting from unstructured text to ids/URIs: “use URIs as names for things”
  - Mixing metadata elements and vocabularies
  - Enabling new types of search engines/queries, machine learning

- **Crowdsourcing**
  - Metadata Games
  - Ancient Lives/Zooniverse
  - Langmuir African American Photographs

- **Automation**
  - Extraction of technical metadata
  - Embedding metadata
  - Text analysis
  - Image analysis

*General shift toward distributed, role-based creation of metadata*
Metadata – General Overviews

• **Understanding Metadata (NISO)**

• **Metadata for digital collections: a how-to-do-it manual**

• **Digital Library Federation: Best Practices for Shareable Metadata**
  – Also in print as *Best Practices for OAI PMH Data Provider Implementations and Shareable Metadata* (2007)

• **Glossary of Metadata Standards**

• **An Introduction to Metadata (JISC)**
  – [http://www.jiscdigitalmedia.ac.uk/guide/an-introduction-to-metadata](http://www.jiscdigitalmedia.ac.uk/guide/an-introduction-to-metadata)

• **Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files** (Federal Agencies Digitization Initiative (FADGI) - Still Image Working Group)
Search Engine/Social Media Resources

- Twitter Cards metadata tags reference:
  - [https://dev.twitter.com/docs/cards/markup-reference](https://dev.twitter.com/docs/cards/markup-reference)

- Facebook OpenGraph metadata tags reference:
  - [https://developers.facebook.com/docs/opengraph/howtos/maximizing-distribution-media-content](https://developers.facebook.com/docs/opengraph/howtos/maximizing-distribution-media-content)

- WordPress - Add Meta Tags plugin:
  - [http://wordpress.org/plugins/add-meta-tags/](http://wordpress.org/plugins/add-meta-tags/)

- Google Webmaster:

- Google Analytics:
  - [http://www.google.com/analytics/](http://www.google.com/analytics/) (account setup)
  - [https://support.google.com/analytics/answer/1012264?hl=en](https://support.google.com/analytics/answer/1012264?hl=en) (site search)

- Google Scholar inclusion reference:

- Google: Rich Snippets
  - [https://support.google.com/webmasters/answer/99170?hl=en](https://support.google.com/webmasters/answer/99170?hl=en)