Applying Digital Library Metadata Standards

Jenn Riley

Metadata Librarian

IU Digital Library Program

Everything you need to know about metadata on one slide

- Metadata is one *view* of a resource
- It's more than cataloging
- It's an essential part of any digital library project
- The planning is as important as the implementation
- Must choose the right tool for the job
- There are lots of acronyms ☺

Well, not really…

- “Metadata” means many different things
- It involves applying traditional library principles to new environments
- These new environments are diverse
- Many of them have been developed for specific knowledge domains
- Metadata planning requires thinking abstractly
- There is *always* more to learn

Many definitions of metadata

- “Data about data”
- “Structured information about an information resource of any media type or format.” (Caplan)
- “Any data used to aid the identification, description and location of networked electronic resources.” (IFLA)
- …etc.

More definition

- Other characteristics
  - Structure
  - Control
- Origin
  - Machine-generated
  - Human-generated
- Data vs. metadata vs. meta-metadata
- Used in many different environments

Metadata and cataloging

- Depends on what you mean by metadata and cataloging!
- But, in general:
  - Metadata is broader in scope than cataloging
  - Much metadata creation takes place outside of libraries
  - Good metadata practitioners use fundamental cataloging principles in non-MARC environments
  - Metadata created for many different types of materials
- Metadata is NOT only for Internet resources!

Some uses of metadata

- By information specialists
  - Describing non-traditional materials
  - Cataloging Web sites
  - Navigating digital objects
  - Managing digital objects long-term
  - Managing corporate assets
- By novices
  - Preparing Web sites for search engines
  - Describing Eprints
  - Managing personal CD collections

Building “Good digital collections”*

- Interoperable – with the important goal of cross-collection searching
- Persistent – reliably accessible
- Re-usable – repositories of digital objects that can be used for multiple purposes

*Institute for Museum and Library Services. A Framework of Guidance for Building Good Digital Collections. Washington, D.C.: Institute for Museum and Library Services, November 2001. http://www.niso.org/framework/Framework2.html

Building “Good digital collections”

- Interoperable – with the important goal of cross-collection searching
- Persistent – reliably accessible
- Re-usable – repositories of digital objects that can be used for multiple purposes

Good metadata promotes good digital collections.

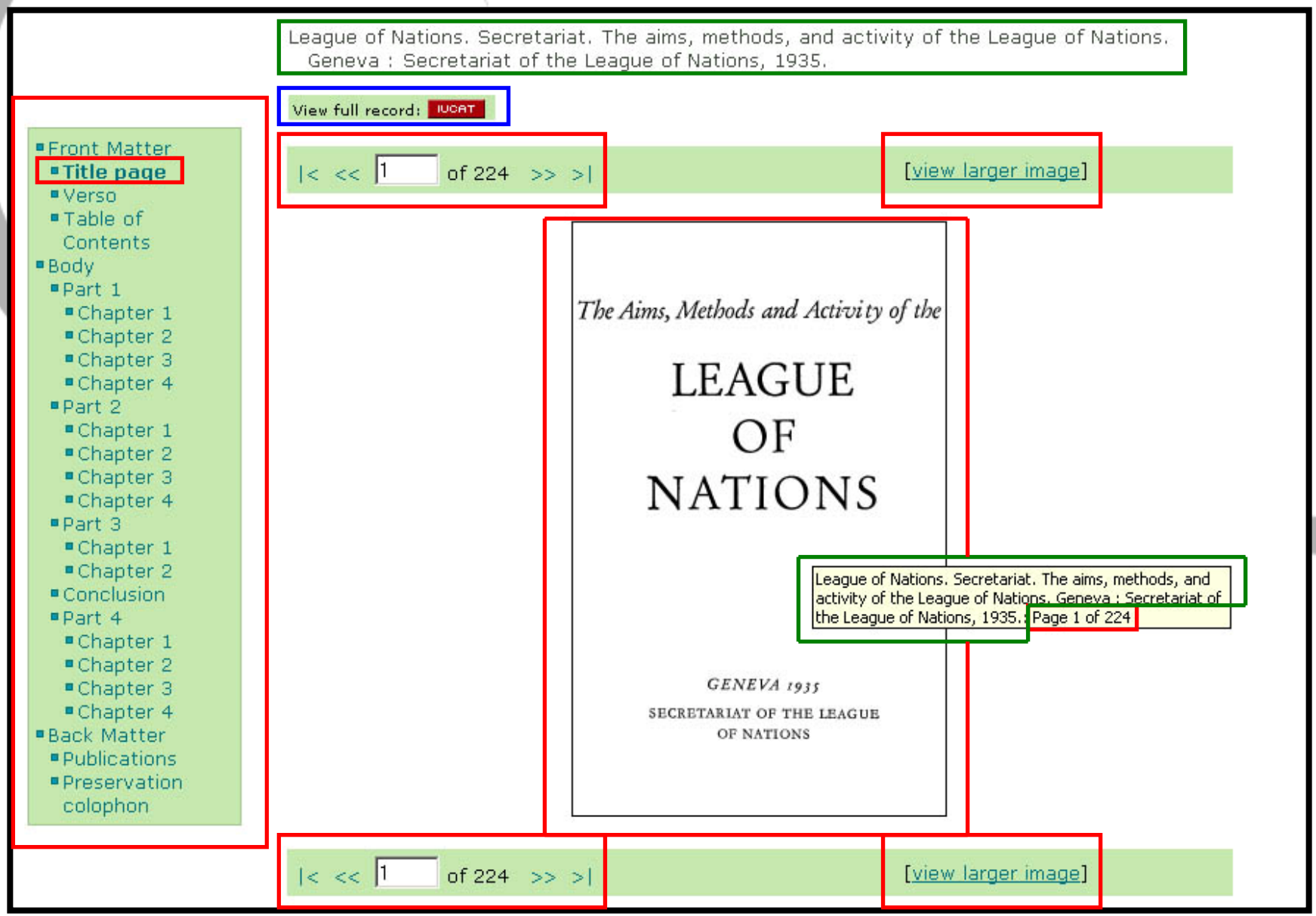
Metadata in digital library projects

- Searching
- Browsing
- Display for users
- Interoperability
- Management of digital objects
- Navigation

Some types of metadata

Type	Use
Descriptive metadata	Searching Browsing Display Interoperability
Technical metadata	Interoperability Digital object management
Structural metadata	Navigation

Metadata in practice



Creating descriptive metadata

- Digital library content management systems
  - ContentDM
  - ExLibris Digitool
  - Greenstone
- Library catalogs
- Spreadsheets & databases
- XML

Creating other types of metadata

- Technical
  - Stored in content management system
  - Stored in separate Excel spreadsheet
- Structural
  - Created and stored in content management system
  - METS XML
- GIS
  - Using specialized software
- Content markup
  - In XML

Implementing metadata in digital library projects

- Levels of control
- Planning your project
- Choosing standards
- Best practices
- Thinking about interoperability

Levels of control

- Data structure standards (e.g., MARC)
  - “Buckets” of information (fields)
  - Both label and scope important
- Data content standards (e.g., AACR2)
  - Selection, structure and formatting of value within a field
- There are others as well
- Standards don’t always fall neatly into one category

When there's no data content standard…

http://digital.library.ucla.edu - Sheet Music OAI Service Provider Item Display - Microsoft...

Sheet Music Consortium — Hosted by UCLA

You are not logged in

- Title : **Anaconda Two-Step. Souvenir Brockton Fair 1901.**

Creators :	By Sylvester B. Grant. Fisher
Publisher :	Brockton, Mass.: Compliments of Old Colony Piano Co., 4 Main St.
Date :	1901
Subjects :	Horses Horse racing
Description :	Respectfully Dedicated to mrs. A.E. Rice, Boston, Mass. ads on back cover for Old Colony Piano Co.
Source :	Johns Hopkins University, Levy Sheet Music Collection, Box 028, Item 003
Collection :	Johns Hopkins University Levy Collection
Add a Note :	

add - add/edit a note for this record

[access online] [add - add to virtual collection]

Done — Internet

http://digital.library.ucla.edu - Sheet Music OAI Service Provider Item Display - Microsoft...

Sheet Music Consortium — Hosted by UCLA

You are not logged in

- Title : **Anaconda : two-step for piano**

Creators :	Grant, Sylvester B. [composer] Fisher, F.
Publisher :	New York : Old Colony Piano Co.
Date :	c1901
Subjects :	Piano music
Type :	Sheet music
Description :	For piano At head of title: Souvenir Brockton Fair 1901. Respectfully dedicated to Mrs. A. E. Rice, Boston, Mass Illustration by F. Fisher Includes publisher's advertising
Source :	Lilly Library (Indiana University, Bloomington) Starr Sheet Music Collection. Call number: M1 .S8 II Animals Horses
Relation :	IsPartOf http://www.indiana.edu/~liblilly/starr.shtml
Rights :	http://purl.dlib.indiana.edu/iudl/lilly/starr/rights
Collection :	Indiana University
Add a Note :	

Done — Internet

Planning your project

- Work collaboratively with technical staff
- Think beyond your local system
- Commit to do it right the first time
- Assess
  - Materials
  - Currently existing metadata
  - Capabilities of your software

Assessing materials for ease of metadata creation

- Number of items?
- Homogeneity of items?
- Foreign language?
- Published or unpublished?
- Specialist needed?
- How much information is known?
- Any existing metadata?

Assessing currently existing metadata

- Machine-readable?
- Divided into fields?
- What format?
- What content standards?
- Complete?

Assessing software capabilities

- Are there templates for standard metadata formats?
- Can you add/remove fields to a template?
- Can you create new templates?
- Can you add additional clarifying information without creating a separate field?
  - Personal vs. corporate names
  - Subject vocabulary used
- Is there an XML export? Does it produce valid records?

Some factors to consider when choosing metadata standards

- Genre of materials being described
- Format of materials being described
- Nature of holding institution
- Robustness needed for the given materials and target users
- What others in your community are doing
- Formats supported by your delivery software
- Dublin Core can be a good choice, but consider all options
- More information on handout

Descriptive metadata schemas

- Purpose
  - Description
  - Discovery
- Some common general schemas
  - MARC
  - MODS
  - Dublin Core
- MANY domain-specific schemas

Comparison of major players

	MARC [book example]	MODS [book example] [photo example]	DC [book example] [photo example]
Record format	Binary (ISO 2709)	XML	XML and others
Field labels	Numeric	Text	Text
Reliance on AACR	Strong	Implied	None
Common method of creation	By specialists	By specialists and by derivation	By specialists and by novices

MODS

- “Metadata Object Description Schema”
- Developed and maintained by the Library of Congress
- For encoding bibliographic information
- Influenced by MARC, but not equivalent
- Much more robust than Dublin Core
- Quickly gaining adoption
  - Base metadata format for DLF Aquifer

Dublin Core (1)

- National and international standard
  - 2001: Released as ANSI/NISO Z39.85
  - 2003: Released as ISO 15836
- Maintained by the Dublin Core Metadata Initiative (DCMI)
- Some important DCMI groups
  - DCMI Working Groups
  - DC Usage Board

Dublin Core (2)

- 15-element set
- “Core” across all knowledge domains
- No element required
- All elements repeatable
- Extensible
- 1:1 principle

Dublin Core (3)

- Two “flavors”
  - Unqualified
  - Qualified
    - Additional elements
    - Element refinements
    - Encoding schemes (vocabulary and syntax)
    - All qualifiers must follow “dumb-down” principle
- Most digital library software uses qualified DC
- Unqualified DC required for sharing metadata via the Open Archives Initiative

Data content standards

- Some types of data content standards
  - Robust sets of rules (e.g., AACR2)
  - Vocabulary encoding schemes
  - Syntax encoding schemes
- Sometimes suggested by data structure standard
- Some for individual elements, some for multiple elements

Vocabulary encoding schemes

- TGM I
- TGM II
- TGN
- GeoNet
- AAT
- LCSH
- LCNAF
- DCMI Type
- MIME Types
- …etc.

Syntax encoding schemes

- ISO8601
- W3CDTF
- URI
- AACR2r
- …etc.

Some specialized metadata standards

- Metadata formats
  - Art & architecture
  - Learning materials
- Markup languages
  - Archival materials
  - Full-text markup

Art & architecture

- Visual Resources Association (VRA) Core
  - Data structure standard
  - From Visual Resources Association
  - Separates Work from Image
  - Library focus
  - Inspiration from Dublin Core
- Categories for the Description of Works of Art (CDWA) Lite
  - Data structure standard
  - From J. Paul Getty Trust
  - Museum focus
  - Conceived for record sharing
- Cataloging Cultural Objects (CCO)
  - Data content standard
  - From Visual Resources Association
  - Leadership from both museums and libraries

Learning materials

- Gateway to Educational Materials (GEM)
  - From the U.S. Department of Education
  - Based on Qualified Dublin Core
  - Adds elements for instructional level, instructional method, etc.
  - “GEM's goal is to improve the organization and accessibility of the substantial collections of materials that are already available on various federal, state, university, non-profit, and commercial Internet sites.”*
- IEEE Learning Object Metadata (LOM)
  - Elements for technical and descriptive metadata about learning resources

* From <http://www.thegateway.org/about/documentation/schemas>

Archival materials

- Encoded Archival Description (EAD)
  - For encoding full text of archival finding aids
  - Requires specialized search engine
  - Delivery requires specialized software or offline conversion to HTML
- Describing Archives: A Content Standard (DACS)
  - Replaces APPM
  - Can be used with EAD, MARC, etc…

Full-text markup

- Text Encoding Initiative (TEI)
- TEI in Libraries
- For encoding full texts of documents
  - Literary texts
  - Letters
  - Transcripts
  - …etc.
- Requires specialized search engine
- Delivery requires specialized software or offline conversion to HTML

Other decisions to make

- Required?
- Repeatable?
- Field lengths
- Public vs. private information
- Unique, persistent identifiers

No, *really*, how do I pick?

- It depends. Sorry.
- Be as robust as you can afford
- Plan for future uses of the metadata you create
- Leverage existing expertise as much as possible
- Focus on content and value standards as much as possible

Good practices for metadata

- Use library cataloging principles whenever possible
- Enter one value per field; repeat fields when necessary
- Clearly describe original vs. digitized item
- Create clear relationships between records
- Plan for interoperability

Best practices for classes of metadata elements

- Titles
- Names
- Dates
- Subjects/Topics
- Language
- Geographic places
- Identifiers
- Rights
- Types of resources

Titles

- Provide a title in every record; supply one if necessary, according to established standards
- Express multiple titles in repeated fields
- Make the distinction between title and sub-title clear through the metadata format used or through standard punctuation

Names

- Include all known names expected by your community of practice
- Format names consistently within a collection, according to authority files or standards expected by your community of practice
- Provide as granular an encoding of a name as possible in the metadata schema being used
- Express multiple names in repeated fields

Dates

- Date elements should contain values important for discovery of the resource by end-users
- When providing multiple dates, clearly indicate the relationship of each to the resource, and repeat the relevant date element for each date
- Include only easily-parsable values in date elements
- Present dates in a consistent format, according to established machine-readable standards

Subjects/Topics

- Choose subject values from relevant controlled vocabularies consistently and explicitly
- Repeat subject information in more specific fields when they are available in the metadata format being used
- Express multiple subjects in repeated fields

Language

- Supply a language element when relevant to the resource
- Format the value of the language element according to the rules of the metadata format in use
- Express multiple titles in repeated fields
- Supply the language of the metadata record only in a metadata element specifically designed for this purpose

Geographic places

- Choose geographic place values from relevant controlled vocabularies consistently and explicitly
- Provide an indication of a hierarchy of geographic places when possible

Identifiers

- Include recognized standard identifiers when available
- Include a URI or DOI linking to the resource when available
- Explicitly encode the nature of an identifier provided
- Identifier must be unique within the repository context
- Ensure persistence of the identifier
- Express multiple identifiers in repeated fields

Rights

- Include rights information about a resource in the most granular format possible
- State rights information in plain language intended for the end-user of a resource
- Supply rights over the metadata record only in a metadata element specifically designed for this purpose

Types/Genres of resources

- Present format and type/genre information in all records.
- Choose type values from relevant controlled vocabularies consistently and explicitly
- Express multiple type/genre terms in repeated fields

Thinking about interoperability

- Metadata as one *view* of a resource
- Have a stranger review a record out of context
- Working towards “shareable metadata”

Metadata as a *view* of the resource

- There is no monolithic, one-size-fits-all metadata record
- Metadata for the same thing is different depending on *use* and *audience*
- Harry Potter as represented by…
  - a public library
  - an online bookstore
  - a fan site

Choice of vocabularies as a *view*

- Names
  - LCNAF: Michelangelo Buonarroti, 1475-1564
  - ULAN: Buonarroti, Michelangelo
- Places
  - LCSH: Jakarta (Indonesia)
  - TGN: Jakarta
- Subjects
  - LCSH: Neo-impressionism (Art)
  - AAT: Pointillism

What does this record describe?

Example courtesy of Sarah Shreeves, University of Illinois at Urbana-Champaign

identifier: http://name.university.edu/IC-FISH3IC-X0802]1004_112

publisher: Museum of Zoology, Fish Field Notes

format: jpeg

rights: These pages may be freely searched and displayed. Permission must be received for subsequent distribution in print or electronically.

type: image

subject: 1926-05-18; 1926; 0812; 18; Trib. to Sixteen Cr. Trib. Pine River, Manistee R.; JAM26-460; 05; 1926/05/18; R10W; S26; S27; T21N

language: UND

source: Michigan 1926 Metzelaar, 1926--1926;

description: Flora and Fauna of the Great Lakes Region

Orig. No. 460 Sta. No. .....

Locality Michigan: Trib. to Sixteen Cr., Sec. 26-27, T. 21 N. R. 10 W., Wexford Co. Trib. Pine River → Manistee

Water: mostly spring-fed; some surface clear; spring floods rather ~~not as bad~~

Vegetation: ~~not much~~ absolutely nothing

Bottom: sand; no mud, rather deep Temp. 47° air 62°

Shore: mostly cleared → farms; some brush Current: ~~swift~~

Distance from shore: av. width 4 ft Tide: .....

Depth of capture: ..... Depth of water: ~~1 ft~~ 1 ft.

Method of capture: .....

Collected by Metzelaar Date: V:18:1926

Orig. preserv. ..... Time: .....

Animal life subnormal

Shareable metadata defined

- Metadata for aggregation with records from other institutions
- Promotes search interoperability - “the ability to perform a search over diverse sets of metadata records and obtain meaningful results” (Priscilla Caplan)
- Is human understandable outside of its local context
- Is *useful* outside of its local context
- Preferably is machine processable

Why share metadata?

- Benefits to users
  - One-stop searching
  - Aggregation of subject-specific resources
- Benefits to institutions
  - Increased exposure for collections
  - Broader user base
  - Bringing together of distributed collections

Don’t expect users will know about your collection and remember to visit it.

Finding the right balance

- Metadata providers know the materials
  - Document encoding schemes and controlled vocabularies
  - Document practices
  - Ensure record validity
- Aggregators have the processing power
  - Format conversion
  - Reconcile known vocabularies
  - Normalize data
  - Batch metadata enhancement

Five C's of shareable metadata

- Consistency
- Coherence
- Content
- Context
- Conformance

Consistency

- Records in a set should all reflect the same practice
  - Fields used
  - Vocabularies
  - Syntax encoding schemes
- Allows aggregators to apply same enhancement logic to an entire group of records

Coherence

- Record should be self-explanatory
- Values must appear in appropriate elements
- Repeat fields instead of “packing” to explicitly indicate where one value ends and another begins

Content

- Choose appropriate vocabularies
- Choose appropriate granularity
- Make it obvious what to display
- Exclude unnecessary “filler”
- Make it clear what links point to

Context

- Include information not used locally
- Exclude information *only* used locally
- Current safe assumptions
  - Users discover material through shared record
  - User then delivered to your environment for full context
- Context driven by intended use

Conformance

- To standards
  - Metadata standards (and not just DC)
  - Vocabulary and encoding standards
  - Descriptive content standards (AACR2, CCO, DACS)
  - Technical standards (XML, Character encoding, etc)

A final word on interoperability

- We can no longer afford to only think about our local users
- Creating shareable metadata will require more work on your part
- Indiana is moving toward a portal of Indiana-related digital content – you should be planning for this *now*

Putting it all into practice

- Develop written documentation
- Develop a quality control workflow for metadata creation
- Share your findings with others

For more information

- Indiana Digital Library home page: <http://www.statelib.lib.in.us/www/isl/diglibin/>
- These presentation slides and handouts: <http://www.dlib.indiana.edu/~jenlrile/presentations/palni2006/>
- jenlrile@indiana.edu