What's Next for Metadata at CUL?

Dean B. Krafft
December 19, 2008
Ground Rules

• I am not a metadata expert
• I’m counting on you to point out areas of critical importance that I’ve overlooked
• Questions and discussion are strongly encouraged
• I reserve the right to limit discussion on any particular topic to ensure broad coverage
Overview

- Framing the problem
- Unique resources
- User created metadata
- Research data
- Implicit, intentional, and programmatic
- Discussion
What is Metadata for?

- Discovery and access - How do I find the thing I want?
- Description, comprehension, and context - Once I’ve found the object, how can I understand it and explore the knowledge framework in which it exists?
Discovery of full-text materials can happen independent of metadata
Even image materials are starting to be discoverable without metadata
More important than discovery and access is context – Metadata that helps users understand the meaning of a resource
In this case, the image is associated with a major news event
What kind of Metadata are we discussing?

- Professional - controlled vocabularies, ontologies, expert understanding
- Contributed - tags, annotation, context, unique knowledge, requires community
- Programmatic - requires analysis & programming, scales well
- Intentional/implicit - requires large amounts of usage data, rich get richer

Lorcan Dempsey categorization of metadata
What will be the focus for Metadata in five years?

• Dealing mostly with digital objects

• Primarily for non-printed-text resources: image, video, datasets, sound, handwritten, physical/conceptual surrogates

• Focused on unique resources & special collections

• Metadata is for computers

Dean B. Krafft
December 19, 2008
Overview

- Framing the problem
- Unique resources
- User created metadata
- Research data
- Implicit, intentional, and programmatic
- Discussion

Where do we need professional metadata?
Commodity resources will get metadata from many sources
Unique, non-text Cornell resources are the prime target for professional Cornell metadata. We have unique knowledge about these artifacts.
There are also non-unique resources for which Cornell can provide unique expertise and insight.
and they don’t even have to be traditional Library–held materials
What Metadata?

- ARTstor platform challenge
- Primary focus on VRA Core, accepts key/value attributes
- Expanding the platform to cover all image materials
- What services are dependent on common metadata?

Specific challenge for images stored using the ARTstor platform
What metadata formats do we use to describe these materials?
Overview

- Framing the problem
- Unique resources
- User created metadata
- Research data
- Implicit, intentional, and programmatic
- Discussion
You really can get volunteers to create structures of lasting value
Huge number of web examples of users contributing information resources together with important metadata
1930s-40s in Color

Operating a hand drill at Vultee-Nashville, woman is working on a "Vengeance" dive bomber, Tennessee (LOC)

At the Vermont state fair, Rutland (LOC)

Retiring a locomotive driver wheel in the Atchison, Topeka, and Santa Fe railway locomotive shops, Shopton, near Fort Madison, Iowa. The tire is heated by means of gas until it can be slipped over the wheel. Contraction on cooling will hold it firmly in place.

Flickr Commons
Supports tags, comments, and image annotations
LoC recently released an initial report on their Flickr project, which they regard as quite successful.
LibraryThing is a major aggregation of user-contributed metadata on individual books and libraries.
Extensive tagging
Provides interesting context and allows exploring the library space in many ways
Users provide ratings, cover images, tags, reviews, and discussions
Authors also exist as resources. Moreover, LibraryThing is now creating a crowdsourced author authority list.
Existing author records can be combined, and “Common Knowledge” provided.
Binghamton
Sept 1st 1868

Hon Ezra Cornell,

Dear Sir, Will you please to accept for the University Library, the accompanying volumes, as memorials of one who was entertained by you the most sincere regard, and who appreciated your early efforts in the great cause in which you have already accrued so much, and who would have rejoiced in your success, had he been lived to witness it.

Yours truly,

Ezra Cornell
Lydin Dickinson

Ezra Cornell Correspondence
1867-1868
August 29, 1868 - September 17, 1868: [57 digital images]
Box 31, Folder 15
TOPICS:
Cascadilla Place; Wisconsin lands; Cornell University: establishment of Library, natural history collection.

http://historical.library.cornell.edu/cgi-bin/cul.ezra/docviewer?
did=000031&seq=723&frames=0&view=50

A Cornell resource that’s a prime candidate for contributed metadata – originally cataloged only at the folder level
A BELLE OF FORMER DAYS

DANIEL S. DICKINSON'S WIDOW.

MRS. LYDIA DICKINSON DYING AT HIGH BRIDGE—A SKETCH OF THE LIFE OF A CELEBRATED SOCIETY WOMAN.

Mrs. Lydia Dickinson, wife of the late Daniel S. Dickinson, has for weeks past been lying dangerously ill at the residence of her daughter, Mrs. Samuel J. Courtney, at High Bridge. For weeks past the life of the venerable lady has been despaired of, and yesterday her physician, Dr. Henry Paine, said that she would hardly live the week out, and, in fact, was liable to drop off at any moment. The disease from which she suffers is dropsy, aggravated by other complications incident to old age. In early life Mrs. Dickinson's social and personal accomplishments, added to the high political position and deserved fame of her husband, made her a notable figure in the society of New-York and the national capital. She was born in June, 1803, in the township of Unadilla. Her
The Cornell community provides a unique pool of authenticated contributors, many with known authority and expertise.
Some tools we developed in the NSDL project to support the contribution of resources, metadata, and context
Supports providing simple metadata about the blog post itself, as well as resources within it.
Referenced Resources

<dct:references> in metadata
http://ndr.nsdlo.org/api/get/2200/20070828124324051T/format_nsdlo_dc

...<dct:references xsi:type="dct:URI">
http://earthobservatory.nasa.gov/Library/GlobalWarming/
</dct:references>

<dct:references xsi:type="dct:URI">

Relationships in objects
http://ndr.nsdlo.org/api/get/2200/20070828124324051T

...<relationships>
<nsdl:relatedto>2200/20070828124316248T</nsdl:relatedto>
<nsdl:relatedto>2200/20070828124316248T</nsdl:relatedto>
</relationships>
...
Wiki also supports simple metadata and related resources
Common references create further implied relationships creating a web of context.
We can generalize from the NSDL experience to a discovery/access/context overlay system for the Cornell Library.
User contributed metadata

- Crowdsources importance and context
- Can range from folksonomic tags to structured controlled vocabulary terms
- Leverages passionate communities
- Supports creating context and relationships
Overview

• Framing the problem
• Unique resources
• User created metadata
• Research data
• Implicit, intentional, and programmatic
• Discussion
Supporting eResearch

- Preserving primary digital data tied to intermediate work products and publication
- Ensuring fixity, security and recoverability
- Developing cross-disciplinary ontologies and metadata mappings
- Pulling order and information from data, descriptions, and observations across multiple disciplines

Image courtesy climateandfarming.org
NSF Datanet Challenge

“Integrate library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise to provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering over a decades-long timeline.”

National Science Foundation Solicitation 07-601
The Data Conservancy

• Collaboration of Johns Hopkins, Cornell, NCAR, UCLA, UIUC, and others

• Carl Lagoze, Steve Kelling, Jim Cordes, Mirek Riedewald, Dean Krafft from Cornell

• $4 million/year for five years with possible five year extension

• Just approved by NSB
Summary

- Data growing exponentially, Petabytes/year by 2010
- Explosion is coming from inexpensive sensors and value added data products
- Requires a new model for science
  - Having more data makes it harder to extract knowledge
- Same thing happening in all sciences
  - High energy physics, genomics, cancer research, medical imaging, oceanography, remote sensing, …
- Science with so much data requires a new paradigm
  - Computational methods, algorithmic thinking will come just as naturally as mathematics today
- **eScience**: an emerging new branch of science

Alex Szalay - from [http://www.eresearch.edu.au/docs/270607/Alex_Szalay_v1.pdf](http://www.eresearch.edu.au/docs/270607/Alex_Szalay_v1.pdf)
“All scientific observations share the same semantic template: they consist of an object/event/phenomenon captured via some observing method at a location/time and recorded as some database entry/spectrum/image.”

- From telescopes mapping the galaxies to citizen bird counts to Antarctic ice cores
- Cross-disciplinary model that supports discipline-specific elements
- Graph-based data model with semantic (ontology-derived) attributes
Describing and combining cross-disciplinary data
Neurocommons Project

- Courtesy John Wilbanks’ CNI presentation
- Goal: Enable neurobiology scientists to integrate research data from many different databases into new research and publications
- Licensing challenges are daunting
- Ontology/naming challenges equally daunting
Courtesy John Wilbanks – MIT coffee drinking ontology
The NeuroCommons project makes heavy use of semantic web technologies and tools: RDF, OWL, SPARQL to enable cross project and database usage of neuroscience research materials.
Example of an article listing the databases that it relies on

**Supplementary Table 1. NAR Molecular Biology Database Collection: 2008 update**

<table>
<thead>
<tr>
<th>No</th>
<th>Database name</th>
<th>Full name and/or description</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DDBJ - DNA Data Bank of Japan</td>
<td>All known nucleotide and protein sequences</td>
<td><a href="http://www.ddbj.nig.ac.jp">http://www.ddbj.nig.ac.jp</a></td>
</tr>
<tr>
<td>2</td>
<td>EMBL Nucleotide Sequence Database</td>
<td>All known nucleotide and protein sequences</td>
<td><a href="http://www.ebi.ac.uk/embl.html">http://www.ebi.ac.uk/embl.html</a></td>
</tr>
<tr>
<td>403</td>
<td>ACLAME</td>
<td>A classification of genetic mobile elements</td>
<td><a href="http://aclame.ulb.ac.be/">http://aclame.ulb.ac.be/</a></td>
</tr>
<tr>
<td>337</td>
<td>CORG</td>
<td>Comparative Regulatory Genomics: conserved non-coding blocks in vertebrate species</td>
<td><a href="http://corg.molgen.mpg.de">http://corg.molgen.mpg.de</a></td>
</tr>
<tr>
<td>30</td>
<td>CUTG</td>
<td>Codon Usage Tabulated from GenBank</td>
<td><a href="http://www.kazusa.or.jp/codon/">http://www.kazusa.or.jp/codon/</a></td>
</tr>
</tbody>
</table>
They make use of a foundry of relevant ontologies and mappings
Example of Amphibian Ontology
Common Naming

1. Each URI must have a clear intended referent.

2. URI documentation must be available via HTTP and provided in RDF.

3. Provision of URI documentation must be an ongoing concern outliving the original database.

4. The provider of the URI documentation must be responsive to community needs.

5. URI documentation must be open so that it can be replicated and reused.

Wilbanks identified common naming as critical for making computational connections between ontologies and databases, and allowing cross-database queries.
Here’s an example of a page “behind” on of the URIs. Note that the format and content of this page needs to be appropriate for human interpretation – need to decide what object/concept the URI really represents.
a. it’s very hard work to use the semantic web right now
b. it’s worth it if you have the cognitive overload problem.
c. none of it works without an open knowledge approach
Local project that makes use of a variety of metadata schema and ontologies – primarily for documentation, discovery, and access – not direct cross-database query
Data Set Citation


Data Tables, Images, and Other Entities:

Metadata download: Ecological Metadata Language (EML) File
- Data Table: GN_Survey_Sum.txt (View Metadata | Download File)
- Data Table: GN_Spec_Codes.txt (View Metadata | Download File)
- Data Table: Gillnet_locations.txt (View Metadata | Download File)
- Data Table: GN_CPUE_Sum.csv (View Metadata | Download File)
- Data Table: Year_age.csv (View Metadata | Download File)

Data Set Owner(s):

Individual: Lars Rudstam
Organization: Cornell Biological Field Station
Position: Associate Director
Address: 900 Shackelton Point Road, Bridgeport, NY 13030 USA
Phone: (315) 633-9243 (voice)
Email Address: lgr1@cornell.edu
Web Address: http://www.dnr.cornell.edu/people/faculty/profiles/rudstam.html

Individual: James R. Jackson
Organization: Cornell Biological Field Station
Position: Senior Research Associate
Address: 900 Shackelton Point Road, Bridgeport, NY 13030 USA
Phone: (315) 633-9243 (voice)
Email Address: jrz26@cornell.edu

Associated Party:

Individual: Edward Mills
Organization: Cornell Biological Field Station
Position: Director
Address: 900 Shackelton Point Road, Bridgeport, NY 13030
Role: Principal Investigator
**Entity Description**

**Data Set Citation**

**Online Distribution Info:**
Download File: **GN_Survey_Sum.txt**

**Physical Structure Description:**
- **Object Name:** GN_Survey_Sum.txt
- **Text Format:**
  - Maximum Record Length: column
  - Simple Delimited: Field Delimeter: ,

**Attribute(s) Info:**

<table>
<thead>
<tr>
<th>Attribute Description (label and definition)</th>
<th>Measurement Type and Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Date</strong> - The date that sampling occurred.</td>
<td><strong>datetime</strong></td>
</tr>
<tr>
<td><strong>Site</strong> - Name of site where gill net survey occurred</td>
<td><strong>nominal</strong></td>
</tr>
<tr>
<td><strong>Standard</strong> - Indicates whether sampling was part of standard effort; value=yes for all records in this table</td>
<td><strong>nominal</strong></td>
</tr>
<tr>
<td><strong>Gear</strong> - Code indicating the sampling protocol used to collect data; all values = GN for this table.</td>
<td><strong>nominal</strong></td>
</tr>
<tr>
<td><strong>Gear Type</strong> - Code indicating the equipment used to collect samples</td>
<td><strong>nominal</strong></td>
</tr>
<tr>
<td><strong>WE Walleye</strong> - Number of walleye in the sample</td>
<td><strong>ratio</strong></td>
</tr>
</tbody>
</table>

**Domain Info**

**Format** yyyyMMdd
**Precision** 1
**Unit** number
**Type** integer
Good metadata is essential to discovering and reusing data: you can’t just Google a table of numbers the way you can with text on a website or in a book. And you need to know precisely what the numbers represent.

Creating metadata can be a lot of work.

Library staff who have happen to have expertise in a subject area can assist in creating metadata, but:

In many cases, someone from the research group who thoroughly understands the data will need to create the metadata.
A URL (Uniform Resource Locator) from which this resource can be downloaded or additional information can be obtained. If accessing the URL would directly return the data stream, then the "function" attribute should be set to "download". If the URL provides further information about downloading the object but does not directly return the data stream, then the "function" attribute should be set to "information". If the "function" attribute is omitted, then "download" is implied for the URL function. In more complex cases where a non-standard connection must be established that complies with application specific procedures beyond what can be described in the simple URL, then the "connection" element should be used instead of the URL element.
Overview

- Framing the problem
- Unique resources
- User created metadata
- Research data
- Implicit, intentional, and programmatic
- Discussion
The king of programmatic and intentional analysis and discovery of information resources
We have an opportunity to make use of our LDC to both do research and create unique finding aids across 3–4 million monograph volumes.
This kind of programmatic analysis is not limited to text – 3D reconstruction of user-contributed images
LibraryThing also has a huge database of user-contributed information
Obvious – recommendations
UnSuggester takes "people who like this also like that" and turns it on its head. It analyzes the thirty million books LibraryThing members have recorded as owned or read, and comes back with books least likely to share a library with the book you suggest. The unsuggestions come from LibraryThing data, not from Amazon. LibraryThing also produces great suggestions.

UnSuggestions for The Grand Sophy by Georgette Heyer

676 members (4,603 more popular); 16 reviews; average rating 4.28 stars. Members with the book have have a total of 883,735 books in their libraries (see good suggestions).

People with this book also have... (more obscure)

1. Survivor: A Novel by Chuck Palahniuk (expected 39.9, found 0)
2. The Road by Cormac McCarthy (expected 36.5, found 0)
3. Diary: A Novel by Chuck Palahniuk (expected 34.8, found 0)
4. How to Win Friends & Influence People by Dale Carnegie (expected 25, found 0)
5. Ishmael: An Adventure of the Mind and Spirit by Daniel Quinn (expected 24.5, found 0)
6. Good to Great: Why Some Companies Make the Leap... and Others

Less obvious – maximum distance
25 books people can't agree on

Must have 50 raters. The numbers are the standard deviation of rating—how much they differ from the average.

The Book of Mormon (1.82), The Bondage Breaker (1.65), Asterix and the Falling Sky (1.62), Intercourse (1.54), Worlds in Collision (1.53), Sins of a Duke (1.51), The Intellectual Foundation of Information Organization (1.51), The Principia: Mathematical Principles of Natural Philosophy (1.48), Skye O'Malley (1.48), The Late Great Planet Earth (1.48), My little blue dress: a novel (1.47), The Devil's Labyrinth: A Novel (1.47), Minion (1.47), Leadership and the New Science: Learning About Organization from an Orderly Universe (1.46), The Openness of God: A Biblical Challenge to the Traditional Understanding of God (1.46), The New Dare to Discipline (1.46), Created to Be His Help Meet: Discover How God Can Make Your Marriage Glorious (1.45), The

(10,765), Wicked: The Life and Times of the Wicked Witch of the West (10,764)

Top Books by LibraryThing Authors


Less obvious – identifying strong disagreement
Sitting on a Gold Mine

- Univ. of Huddersfield just released a major portion of their book circulation and recommendation database
- 80,000 titles, 3 million records, 13 years
- Open Data Commons/CC0 license
- Aggregated/anonymized - no privacy issues
<description>
This file contains usage data for over 80,000 items borrowed from the libraries at the University of Huddersfield. Data has only been included where usage has been significant and when an ISBN is available. Usage breakdowns are provided by year and, where usage has been significant, by academic school and by academic course. See separate files for ID mappings for schools and courses. Total loans and the current number of copies have been included for each item. ISBNs included in the MARC record are shown, with further mappings available in the isbnlookup.xml file.
</description>

<licence>
<type>CC0 / Open Data Commons</type>
<statement>To the extent possible under law, Computing and Library Services, University of Huddersfield, UK has waived all copyright, moral rights, database rights, and any other rights that might be asserted over the data contained within this file.</statement>
<url>http://labs.creativecommons.org/licenses/zero/1.0/</url>
<url>http://wiki.creativecommons.org/CC0</url>
<url>http://www.opendatacommons.org</url>
</licence>

<items>
<item id="51353">
<title>Schools for the boys? : co-education reassessed</title>
<copies>2</copies>
<isbn>0091608414</isbn>
<url>http://library.hud.ac.uk/catlink/bib/51353</url>
<loan_history>
<total>11</total>
<schools>
<school id="D">3</school>
<school id="H">7</school>
</schools>
</loan_history>
</item>
</items>
Suggested Books

This page is a quick demo of slicing and dicing the circulation data provided by the University of Huddersfield [here](http://example.com). Please see my blog post about this demo [here](http://example.com) to read more about it and to leave comments, suggestions, bug reports, etc.

Daughters of decadence: women writers of the Fin de Siècle /

Total loans: 116

RDF Book Mashup (requires semantic web browser/plugin)

Library Page

<table>
<thead>
<tr>
<th>Title</th>
<th>Times checked out before above item</th>
<th>Times checked out after above item</th>
<th>Times check out at same time as above item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature and gender /</td>
<td>14</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Literature and gender /</td>
<td>14</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Literature and gender /</td>
<td>14</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Literature and gender /</td>
<td>14</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Literature and gender /</td>
<td>14</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Making a difference: feminist literary criticism /</td>
<td>13</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Making a difference: feminist literary criticism /</td>
<td>13</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Modernism/postmodernism /</td>
<td>20</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Modernism/postmodernism /</td>
<td>20</td>
<td>7</td>
<td>0</td>
</tr>
</tbody>
</table>

Supports recommendations based on common usage
Should Cornell be making use of the data we have about our resources and usage?
Discussion

...then he said, “I can’t feel my legs” and then I said, “Stay with me Joe!” but it was too late. He was gone. It was the PowerPoint.

Courtesy Kathy Sierra/WickedlySmart.com
Questions

• Where should we focus our limited professional metadata resources?

• When and how is it appropriate to use contributed metadata?

• Can we make use of Cornell community usage patterns, and at what level and with whom should we aggregate?

• What should we do here, where should we partner, and what should we let others do?