Metadata for the Web
A Necessary Evil?

CS 431 - March 2, 2005
Carl Lagoze - Cornell University
“Metadata is data about data”
Metadata is semi-structured data conforming to commonly agreed upon models, providing operational interoperability in a heterogeneous environment.
Are metadata and data distinguishable?

• Objectivity?
• Intellectual property?
• Structure?
• Aboutness?
Some untested hypotheses

• **Metadata is useful for...**
  - People
  - Machines

• **More metadata is better**

• **Simple metadata created by non-experts (Joe Sixpack and his smarter friends) is useful**
Metadata Quality as function of Creator Expertise

Hoped

Actual?
Some known facts

• Number and variety of metadata vocabularies will continue to increase

• The Tower of Babel is a franchise
  - There is not one common view of reality

• “The one thing I know about metadata is that it is expensive” (Bill Arms)

• “I hate metadata projects because they make every other digital library project more expensive” (Michael Lesk)
Metadata Triage
Metadata Takes Many Forms

- resource discovery
- content rating
- document administration
- security and authentication
- rights management
- archival status
- products and services
- database schemas
- process control or description
Metadata Challenges

• Accommodate multiple varieties of metadata
  - community-specific functionality, creation, administration, access
  - Lowest common denominator

• Tensions
  - functionality and simplicity
  - extensibility and interoperability
  - human and machine creation and use
The fiction of classification

...there is no classification of the universe that is not fictional and conjectural.

Jorge Luis Borges
Lenses and Views

• All classification does and should provide a biased *lens* or *view* of reality
• Each view emphasizes certain characteristics and hides others
Reality is Complex

Created by: Leonardo da Vinci
Created on: 1506

Created by: George Castaldo
Created on: 1994

Relationship?
Objects are Related

IFLA Entity Model

Figure 2: Based on IFLA FRBR Figure 3.1
Haven’t we done metadata already?
What’s wrong with this model?

- **Expensive**
  - Complex (even for its original goal?)
  - Professional intervention (assumes single community of expertise)

- **Monolithic**
  - One size fits all approach
  - Reflects its centralized system origins

- **Bias towards physical artifacts**
  - Fixed resources
  - Incomplete handling of resource evolution and other resource relationships

- **Anglo-centric**
Why hasn’t metadata worked on the Web?

• Its all about trust
• People are lazy
• Metadata is hard
• No perceived benefit
  – “Reverse tragedy of the commons”
• No agreement on one way to describe things

• “Metacrap” -
  http://www.well.com/~doctorow/metacrap.htm
The fifteen Dublin Core Elements

<table>
<thead>
<tr>
<th>Creator</th>
<th>Title</th>
<th>Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contributor</td>
<td>Date</td>
<td>Description</td>
</tr>
<tr>
<td>Publisher</td>
<td>Type</td>
<td>Format</td>
</tr>
<tr>
<td>Coverage</td>
<td>Rights</td>
<td>Relation</td>
</tr>
<tr>
<td>Source</td>
<td>Language</td>
<td>Identifier</td>
</tr>
</tbody>
</table>

http://dublincore.org/usage/terms/dc/current-elements/
http://dublincore.org
A *Pidgin* for Digital Tourists

- *Metadata is language*
- *Dublin Core is a small and simple language -- a pidgin -- for finding resources across domains.*
- *Speakers of different languages naturally "pidginize" to communicate*
  - E.g., tourists using simple phrases to order beer ("zwei Bier bitte" "dva pivo prosim" "biru o san bai kudasi"...)
- *We are all "tourists" on the global Internet.*
What is the Dublin Core (1)

- A simple set of properties to support resource discovery on the web (fuzzy search buckets)?
What is Dublin Core (2)?

- An extensible ontology for resource description?

Greater Functionality & Cost
Progressive Metadata Models: Drill-Down Searching Paradigm

- Moving along a specificity spectrum
- Inter-domain vs. intra-domain terms, models, query mechanisms
Drill-down search paradigm

- Domain Independent view
- Domain Specific View
What is the Dublin Core (3)?

• A cross-domain switchboard for interoperable metadata?
Dublin Core Qualifiers

• From fuzzy buckets to more specific description
• Model of “graceful degradation”
  – Support both simplicity and specificity
  – Intra-domain and inter-domain semantics
Varieties of qualifiers: Element Refinements

• Make the meaning of an element narrower or more specific.
• Narrowing implies an *is a* relationship
  - a "date created" *is a* "date"
  - an "is part of relation" *is a* "relation"
• If your software does not understand the qualifier, you can safely ignore it.
Varieties of Qualifiers: Value Encoding Schemes

• Says that the value is
  - a term from a controlled vocabulary (e.g., Library of Congress Subject Headings)
  - a string formatted in a standard way (e.g., "2001-05-02" means May 3, not February 5)

• Even if a scheme is not known by software, the value should be "appropriate" and usable for resource discovery.
A Grammar of Dublin Core

- http://www.dlib.org/dlib/october00/baker/10baker.html
- By design not as subtle as mother tongues, but easy to learn and extremely useful in practice
- Pidgins: small vocabularies (Dublin Core: fifteen special nouns and lots of optional adjectives)
- Simple grammars: sentences (statements) follow a simple fixed pattern...
Example Dublin Core statements

- Resource has Title 'Grammar of Dublin Core'.
- Resource has Creator 'Tom Baker'.
- Resource has Subject 'Metadata'.
Resource has one of 15 properties.

implied subject

implied verb

property

X

property value (an appropriate literal)
Resource has property with [optional qualifier] one of 15 properties:
- DC:Creator
- DC:Title
- DC:Subject
- DC:Date

X

property value (an appropriate literal)

qualifiers (adjectives)
Dumb-Down Principle for Qualifiers

• The fifteen elements should be usable and understandable with or without the qualifiers

• Qualifiers refine meaning (but may be harder to understand)

• Nouns can stand on their own without adjectives

• If your software encounters an unfamiliar qualifier, look it up -- or just ignore it!

• "has a" relations break the model
  - E.g., a creator has a hair color
Test for “good” qualifiers:

- Does the statement still make sense?
- Is it still correct?

<table>
<thead>
<tr>
<th>Resource</th>
<th>has</th>
<th>Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>&quot;Languages -- Grammar&quot;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LCSH</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resource</th>
<th>has</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>&quot;2000-06-13&quot;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ISO8601</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Revised</td>
</tr>
</tbody>
</table>
"Incorrect" Qualification

Resource has creator / "Cornell University"

Resource has subject / "pre-schoolers"