Metadata Madness: Quality Issues in Metadata Management

Daniel Gelaw Alemneh
University of North Texas Libraries,
Digital Projects Unit, Denton, Texas
Dalemneh@library.unt.edu

TLA, April 11, 2007
Quality Issues

The two aspects of digital library data quality:

1. The quality of the data in the objects themselves
2. The quality of the metadata associated with the objects

Maintaining usable, flexible, interoperable, and sustainable digital collections necessitates maintaining high quality metadata about those digital objects.
Metadata Quality Issues

- **In terms of Typographical Errors:**
  - Letter transposition e.g. 9198 for 1998
  - Letter omission, e.g. Omt for omit
  - Letter insertion, e.g. asnd for and
  - Letter substitution or misstrokes, e.g. likw for like

- **No omissions**
  - Null values for mandatory elements
  - Incomplete information

- **Non ambiguous**
  - Inconsistency eg. multiple spellings, multiple possible meanings, mixed cases, initials, etc
Factors Influencing Metadata Quality

- **Resource types**
  - Heterogeneity

- **Local requirements**
  - Different functionality and granularity

- **Collaborators requirements**
  - Diverse and conflicting requirements

- **Cost**
  - Resource limitations (CBA)
Factors Influencing Metadata Quality: Resource Types Heterogeneity

What type of objects will the repository contain?
- Museum objects,
- Archives and historical documents,
- Wide format items
- Scholarly documents, etc.
Factors Influencing Metadata Quality: *Resource Types Heterogeneity* …

- **How will they be described?**
  - Levels of details

- **How will they be used?**
  - Functionality required

- **By whom?**
  - Users category
Factors Influencing Metadata Quality: Local/Collaborators Requirements

- What functionality is required locally? By collaborating institutions?
- What entry points will be required locally? And or by collaborating institutions?
  - The type of access,
  - Type of templates,
  - Type of interfaces, etc.
Factors Influencing Metadata Quality: Collaborators Requirements

- How does the information-seeking behavior of the diverse users differ?
  - Genealogists
  - Historians
  - Students
  - Researchers, etc.

- How best their need can be met?
Factors Influencing Metadata Quality: Collaborator Requirements

What are the associated digital rights issues?
- Content packaging,
- Repackaging
- Repurposing

Does participation in the wider community impose specific requirements?
- Are requirements formal or informal?
- Will access restrictions be imposed?
Factors Influencing Metadata Quality: Collaborators Requirements

- Will metadata be meaningful within aggregations of various kinds?
  - Mapping and Crosswalks

- What is required for interoperability?
  - Structure
  - Semantics
  - Syntax
Factors Influencing Metadata Quality: *Cost Issues*

- Are resources sufficient to produce the required metadata quality?
  - Available expertise, etc.

- If not, what are the priorities?
  - Cost Benefit Analysis
Managing Metadata Quality

- **Identify the 'right' metadata**
  - Avoid large schemas

- **Create at the right time**
  - At creation vs. at other points in object life-cycle
  - Human vs. automatic

- **Determine level of quality required**
  - Collaborators may have diverse and sometimes conflicting metadata requirements.
Managing Metadata Quality

- Determine nature of gap and how to close
  - Effectiveness
  - Efficiency
  - Practicability
  - Scalability

- Produce the required metadata quality
  - Quality Assurance mechanisms
Managing Metadata Quality …

- **Compromise**
  - One size does not fit all!

- **Prioritize**
  - Resources unlikely to be available to meet all requirements

- **Test the workflow**
  - The quality cycle
Quality Assurance Lifecycle

1. Determine purpose of metadata
2. Determine required metadata quality
3. Determine target metadata quality
4. Design & implement workflow
5. Refine workflow
6. Review

Local environment
Wider environment

UNT Libraries’ Quality Assurance for Mandatory Elements

Web Template Creator

Javascript Validation

Rules Processing

edit

download

mandatory elements
- title
- collection
- institution
- description
- language
- subject
- resource-type
- format

fail

pass
### Metadata Analysis: NULL

<table>
<thead>
<tr>
<th>DC Element</th>
<th>Number that are NULL</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC Title</td>
<td>0</td>
</tr>
<tr>
<td>DC Creator</td>
<td>10329</td>
</tr>
<tr>
<td>DC Subject</td>
<td>0</td>
</tr>
<tr>
<td>DC Description</td>
<td>0</td>
</tr>
<tr>
<td>DC Contributor</td>
<td>16369</td>
</tr>
<tr>
<td>DC Publisher</td>
<td>18522</td>
</tr>
<tr>
<td>DC Identifier</td>
<td>1037</td>
</tr>
<tr>
<td>DC Source</td>
<td>13803</td>
</tr>
<tr>
<td>DC Relation</td>
<td>18255</td>
</tr>
<tr>
<td>DC Language</td>
<td>0</td>
</tr>
<tr>
<td>DC Coverage</td>
<td>0</td>
</tr>
<tr>
<td>DC Date</td>
<td>2790</td>
</tr>
<tr>
<td>DC Resource Type</td>
<td>0</td>
</tr>
<tr>
<td>DC Format</td>
<td>0</td>
</tr>
<tr>
<td>DC Rights</td>
<td>19156</td>
</tr>
</tbody>
</table>
UNT Libraries Quality Assurance Lifecycle

Created by automated means

Created/Enhanced by resources owners

Created/Enhanced by metadata team

UNTIL Metadata Records
UNT Libraries’ Quality Assurance Lifecycle Loop

Measure Quality and Usefulness of UNTL metadata
Challenges and Scenarios: Summaries

- At what “level” is metadata assigned?
- How much of the process can be automated?
- What kind of quality assurance mechanisms implemented?
- How will the metadata be maintained?
- What kind of skills will be required?