Conformance to DLF/Aquifer MODS Implementation Guidelines

DLF Aquifer Fall Forum 2008
Colin Koteles
koteles@cod.edu
Study Goals

- Snapshot/measure "current" (Fall, 2007) conformance (in practice) to DLF/Aquifer MODS implementation guidelines

- Quantify "...the requirements and recommendations set forth here are not currently met by most current and potential Aquifer participants."
Study Goals

- Recommendations for streamlining guidelines
  - Automated service provider processes
  - Manual data provider processes

- Are guidelines a viable method for improving shareability?
What:
- Guidelines for using MODS schema to create metadata
  - For digital cultural heritage and humanities-based scholarly resources
  - Intended to be shared and aggregated
- Developed by Aquifer Metadata WG
- Version 1.0, November, 2006
- Based on MODS 3.2
- Includes MODS element use rules, examples, and DC mappings
Why

- Guidelines for creating "rich, shareable metadata that is coherent and consistent"

- Enhance metadata **shareability** vs. DC

  - Defines expressive element set -> Richness
  - Defines explicit declarations -> Consistency & interoperability (machines & humans)
  - Useful in multiple views and non-local contexts
MODS elements are required, recommended, or optional

Study focuses on nine explicitly required elements:

- `<titleInfo><title>`
- `<typeOfResource>`
- `<originInfo>`
- `<language>`
- `<physicalDescription>`
- `<subject>`
- `<location>`
- `<accessCondition>`
- `<recordInfo>`
DLF/Aquifer MODS Implementation Guidelines: Required elements

Concise:

<titleInfo><title>

- Each record must include at least one <titleInfo> element with one <title> subelement
- <titleInfo><title> instances are repeatable
DLF/Aquifer MODS Implementation Guidelines: Required elements

Verobse:

<recordInfo>

- Each record requires one and only one <recordInfo> element that includes one and only one <languageOfCataloging> subelement.
- Each <languageOfCataloging> subelement requires a pair of <languageTerm> subelements.
  - One <languageTerm> subelement that includes a type="text" attribute/value pair.
    - The content of this <languageTerm> subelement should appear in the MARC Code List for Languages.
  - One <languageTerm> subelement that includes both type="code" and authority="iso629-2b" attribute/value pairs.
    - The content of this <languageTerm> subelement should be valid ISO 639-2 content.
Method: Test Set

1. Harvest MODS records

• 10 data providers from 9 institutions (as listed on [DLF MODS Portal website](https://www.dlf.org/modsportal) - 8/2007)
• Harvested between Aug. 30, 2007 and Oct. 27, 2007
• 343,529 MODS records

<table>
<thead>
<tr>
<th>Repository ID</th>
<th>Repository Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A Celebration of Women Writers</td>
</tr>
<tr>
<td>2</td>
<td>OCLC Research Publications</td>
</tr>
<tr>
<td>3</td>
<td>University of Tennessee</td>
</tr>
<tr>
<td>4</td>
<td>Southern Spaces</td>
</tr>
<tr>
<td>5</td>
<td>Digital Books from UIUC and the Open Content Alliance</td>
</tr>
<tr>
<td>6</td>
<td>University of Chicago Metadata Repository</td>
</tr>
<tr>
<td>7</td>
<td>Indiana University Library Cushman Collection</td>
</tr>
<tr>
<td>8</td>
<td>Deep Blue at the University of Michigan</td>
</tr>
<tr>
<td>9</td>
<td>Library of Congress Memory Collection</td>
</tr>
<tr>
<td>10</td>
<td>University of Michigan. University Library. Digital Library Production Service</td>
</tr>
</tbody>
</table>
Method: Data Extraction

2. Extract elements, element content, and attribute/value pairs from records

- **IndexReap**
  - VBScript-based tool developed by UIUC OAI/PMH
  - Extract metadata elements, element content, and attribute/value pairs into RDB
  - Configured with hand generated **Repository** table

- Resulted in two tables:
  - **Records** table with row for each MODS record
  - **Metadata** table with row for each metadata element (13,770,392 elements)
Method: Table structures

Repositories table

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Data Type</th>
<th>Field Size</th>
<th>Required</th>
<th>Indexed</th>
<th>New Values</th>
<th>Decimal Places</th>
<th>Allow Zero Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>repoid</td>
<td>AutoNumber</td>
<td>Long Integer</td>
<td></td>
<td>Yes (No duplicates)</td>
<td>Increment</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>baseURL</td>
<td>Text</td>
<td>255</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>setSpec</td>
<td>Text</td>
<td>50</td>
<td>No</td>
<td>No</td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>whenHarvested</td>
<td>Date/Time</td>
<td></td>
<td>No</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>repositoryName</td>
<td>Text</td>
<td>255</td>
<td>No</td>
<td>No</td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>
Method: Table structures

Records table

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Data Type</th>
<th>Field Size</th>
<th>Required</th>
<th>Indexed</th>
<th>New Values</th>
<th>Decimal Places</th>
<th>Allow Zero Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>repoid</td>
<td>Number</td>
<td>Long Integer</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Auto</td>
<td>No</td>
</tr>
<tr>
<td>recordID</td>
<td>AutoNumber</td>
<td>Long Integer</td>
<td>Yes</td>
<td>Yes (No duplicates)</td>
<td>Yes</td>
<td>Increment</td>
<td>No</td>
</tr>
<tr>
<td>OAIIdentifier</td>
<td>Text</td>
<td>255</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>OAIDateStamp</td>
<td>Date/Time</td>
<td></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>XMLMetadata</td>
<td>Memo</td>
<td></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>
Method: Table structures

Metadata table

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Data Type</th>
<th>Field Size</th>
<th>Required</th>
<th>Indexed</th>
<th>New Values</th>
<th>Decimal Places</th>
<th>Allow Zero Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>recordID</td>
<td>Number</td>
<td>Long Integer</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Auto</td>
<td></td>
</tr>
<tr>
<td>metaRowID</td>
<td>AutoNumber</td>
<td>Long Integer</td>
<td>Yes (No Duplicates)</td>
<td>Increment</td>
<td>No</td>
<td>Auto</td>
<td></td>
</tr>
<tr>
<td>propertyName</td>
<td>Text</td>
<td>50</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Auto</td>
<td>Yes</td>
</tr>
<tr>
<td>propertyNS</td>
<td>Text</td>
<td>255</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Auto</td>
<td>Yes</td>
</tr>
<tr>
<td>parent_propName</td>
<td>Text</td>
<td>50</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Auto</td>
<td>Yes</td>
</tr>
<tr>
<td>parent_propNS</td>
<td>Text</td>
<td>255</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Auto</td>
<td>Yes</td>
</tr>
<tr>
<td>parent_metaRowID</td>
<td>Number</td>
<td>Long Integer</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Auto</td>
<td></td>
</tr>
<tr>
<td>propText</td>
<td>Text</td>
<td>255</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Auto</td>
<td>Yes</td>
</tr>
<tr>
<td>propTextOverflow</td>
<td>Memo</td>
<td></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Auto</td>
<td>Yes</td>
</tr>
<tr>
<td>a_type</td>
<td>Text</td>
<td>100</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Auto</td>
<td>Yes</td>
</tr>
<tr>
<td>a_authority</td>
<td>Text</td>
<td>100</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Auto</td>
<td>Yes</td>
</tr>
<tr>
<td>a_encoding</td>
<td>Text</td>
<td>100</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Auto</td>
<td>Yes</td>
</tr>
<tr>
<td>a_href</td>
<td>Text</td>
<td>100</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Auto</td>
<td>Yes</td>
</tr>
<tr>
<td>a_displayLabel</td>
<td>Text</td>
<td>100</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Auto</td>
<td>Yes</td>
</tr>
<tr>
<td>a_keyDate</td>
<td>Text</td>
<td>50</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Auto</td>
<td>Yes</td>
</tr>
<tr>
<td>a_usage</td>
<td>Text</td>
<td>100</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Auto</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Method: SQL Queries

3. Apply SQL queries formulated to test requirements of the nine explicitly required elements in OAI/MODS

- Anywhere from four to 17 queries per requirement
- Applied to Records and Metadata tables

```sql
SELECT dbo_Records.recordID, dbo_Records.repoID, CLng([dbo_Metadata.metaRowID]) AS metaRowID, dbo_Metadata.propText, dbo_Metadata.a_type, dbo_Metadata.parent_MetaRowID, dbo_Metadata.a_authority INTO language_languageTerm_temp
FROM dbo_Records INNER JOIN dbo_Metadata ON dbo_Records.recordID=dbo_Metadata.recordID
WHERE (((propertyName="languageTerm") And (parent_propName="language")) And (propText Is Not Null));
```
Results

Initially, results showed substantial non-conformance...
Results: the Good

<titleInfo><title> (99.97%)

- Most conformance
- 7 repositories average ≈2 titles
- Further analysis of other subelements, attribute/values, repeated content

<typeOfResource> (92.09%)

- Content is 100% valid (1 of 11 choices)
- Missing elements limited to one repository
Results: the OK

<subject> (65.85%)

- Avg. ≈2 or more <subjects> (one avg. ≈10)
- ≈50% declare subject authority
- 6 repositories >96% of <subject> elements include at least one subelement
- Repeated values
- Further analysis of quality, consistency

<accessCondition> (58.62%)

- 5 repositories have 100%, 4 have 0%, 1 has 64%
- All but one repository uses type="useAndReproduction" when <accessCondition> present
Results: the Not So Good

Five elements with under 8% conformance

- `<originInfo>` (3%)
- `<language>` (0.12%)
- `<physicalDescription>` (7.97%)
- `<location>` (0.37%)
- `<recordInfo>` (0.10%)
Results: One Bad Apple?

LOC Memory Collection accounts for 85% of all harvested records

- Never a significant outlier in element-by-element analysis
- (See full report for by-element and by-repository summaries)
Results: Can we easily fix these problems?

Yes, We Can!

Automatic remediation can substantially help in 5 of 7 cases...
Automatic Remediation

Generate text values from supplied controlled vocabulary values

<language>

- Generate <languageTerm type="text"> values from existing <languageTerm type="code" authority="iso639-2b"> values
- Conformance improves from 0.12% to 53%
Automatic Remediation

Infer attribute/value pairs when not necessary to disambiguate between repeated elements

<originInfo>

- Apply keydate="yes" when only one <originInfo> date-related subelement exists
- 118,225 records

<location>

- Apply usage="primary display" when only one <location><url> exists
- 281,748 records
Semi-Automatic Remediation

Data providers provide blanket content values that can be overridden by individual records

<accessCondition>

- Data providers supply single <accessCondition type="useAndReproduction"> content values for collections

<recordInfo>

- Data providers supply single <languageOfCataloging> content values for collections
Results: Part Deux

Conformance closer than may (first) appear...

### 3.1 Summary of Number of Records with Required MODS Elements

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;titleInfo&gt;</td>
<td>305386</td>
<td>99.97%</td>
<td>305386</td>
<td>99.97%</td>
</tr>
<tr>
<td>&lt;typeOfResource&gt;</td>
<td>281297</td>
<td>92.09%</td>
<td>281297</td>
<td>92.09%</td>
</tr>
<tr>
<td>&lt;originInfo&gt;</td>
<td>9170</td>
<td>3.00%</td>
<td>118825</td>
<td>38.90%</td>
</tr>
<tr>
<td>&lt;language&gt;</td>
<td>366</td>
<td>0.12%</td>
<td>165496</td>
<td>54.18%</td>
</tr>
<tr>
<td>&lt;physicalDescription&gt;</td>
<td>24356</td>
<td>7.97%</td>
<td>24356</td>
<td>7.97%</td>
</tr>
<tr>
<td>&lt;subject&gt;</td>
<td>201162</td>
<td>65.85%</td>
<td>201162</td>
<td>65.85%</td>
</tr>
<tr>
<td>&lt;location&gt;</td>
<td>1130</td>
<td>0.37%</td>
<td>281748</td>
<td>92.23%</td>
</tr>
<tr>
<td>&lt;accessCondition&gt;</td>
<td>179058</td>
<td>58.62%</td>
<td>305468</td>
<td>100.00%</td>
</tr>
<tr>
<td>&lt;recordInfo&gt;</td>
<td>308</td>
<td>0.10%</td>
<td>305468</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

- **A** – Element name
- **B** – Number of records meeting element requirements
- **C** – Percentage of records meeting element requirements
- **D** – Number of records meeting element requirement after safe algorithmic normalization
- **E** – Percentage of records meeting element requirement after safe algorithmic normalization
Conclusions

Create guidelines that let services providers do what they do best:

- Supply easily and accurately inferred content and values
- Supply global content and values

Create guidelines that let data providers do what they do best:

- Reduce redundant burdens (see above)
- Focus on quality of subjective content and values

Education and outreach for data providers focused on providing quality, subjective metadata
Conclusions

Still difficult to gauge effectiveness of guidelines using these results...
Additional Notes

No significant differences in records created before and after release of guidelines
- 2 repositories with records created BG and AG
- 2 repositories all records created BG
- 6 repositories all records created AG

Guideline requirements not in sync with Levels of Adoption
- `<name>` is not a Guideline requirement

Full report available: [https://www.ideals.uiuc.edu/handle/2142/8958](https://www.ideals.uiuc.edu/handle/2142/8958)