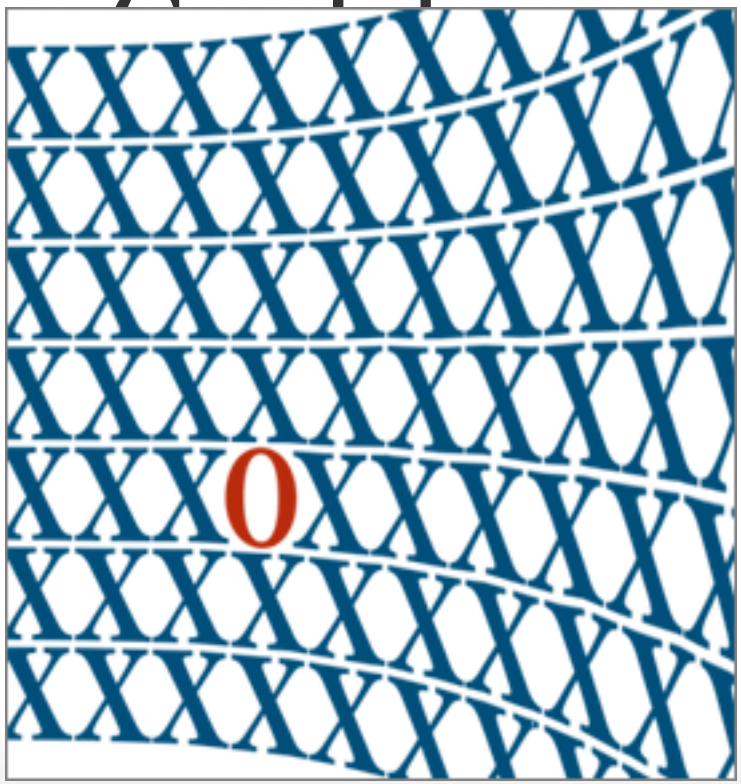
Conformance to DLF/Aquifer MODS Implementation Guidelines



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?

DLF/Aquifer MODS Implementation Guidelines

"DLF/Aquifer MODS Implementation Guidelines for Shareable MODS Records"

- 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

DLF/Aquifer MODS Implementation Guidelines

- 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

DLF/Aquifer MODS Implementation Guidelines: Required elements

- 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 - 8/2007)
- Harvested between Aug. 30, 2007 and Oct. 27, 2007
- 343,529 MODS records

A.1 Repository Names	
Repository ID	**Repository Name**
1	A Celebration of Women Writers
2	OCLC Research Publications
3	University of Tennessee
4	Southern Spaces
5	Digital Books from UIUC and the Open Content Alliance
6	University of Chicago Metadata Repository
7	Indiana University Library Cushman Collection
8	Deep Blue at the University of Michigan
9	Library of Congress Memory Collection
10	University of Michigan. University Library. Digital Library Production Service.

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

B.1 Repositories Table (dbo_Repositories)							
Field Name	Data Type	Field Size	Required	Indexed	New Values	Decimal Places	Allow Zero Length
repoid	AutoNumber	Long Integer		Yes (No duplicates)	Increment		
baseURL	Text	255	Yes	No			Yes
setSpec	Text	50	No	No			Yes
whenHarvested	Date/Time		No	No			
repositoryName	Text	255	No	No			Yes

Method: Table structures

Records table

B.2 Records Table (dbo_Records)							
Field Name	Data Type	Field Size	Required	Indexed	New Values	Decimal Places	Allow Zero Length
repoid	Number	Long Integer	Yes	No		Auto	
recordID	AutoNumber	Long Integer		Yes (No duplicates)	Increment		
OAIdentifier	Text	255	No	No			Yes
OAIDateStamp	Date/Time		No	No			
XMLMetadata	Memo		No	No			Yes

Method: Table structures

Metadata table

B.3 Metadata Table (dbo_Metadata)							
Field Name	Data Type	Field Size	Required	Indexed	New Values	Decimal Places	Allow Zero Length
recordID	Number	Long Integer	Yes	No		Auto	
metaRowID	AutoNumber	Long Integer		Yes (No Duplicates)	Increment		
propertyName	Text	50	No	No			Yes
propertyNS	Text	255	No	No			Yes
parent_propName	Text	50	No	No			Yes
parent_propNS	Text	255	No	No			Yes
parent_metaRowID	Number	Long Integer	No	No		Auto	
propText	Text	255	No	No			Yes
propTextOverflow	Memo		No	No			Yes
a_type	Text	100	No	No			Yes
a_authority	Text	100	No	No			Yes
a_encoding	Text	100	No	No			Yes
a_href	Text	100	No	No			Yes
a_displayLabel	Text	100	No	No			Yes
a_keyDate	Text	50	No	No			Yes
a_usage	Text	100	No	No			Yes

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

```
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…

3.1 Summary of Number of Records with Required MODS Elements				
A	B	C	D	E
<titleInfo>	305386	**99.97%**	305386	**99.97%**
<typeOfResource>	281297	**92.09%**	281297	**92.09%**
<originInfo>	9170	**3.00%**		
<language>	366	**0.12%**		
<physicalDescription>	24356	**7.97%**		
<subject>	201162	**65.85%**		
<location>	1130	**0.37%**		
<accessCondition>	179058	**58.62%**		
<recordInfo>	308	**0.10%**		

Wait for it...

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

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 indivdual 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				
A	B	C	D	E
<titleInfo>	305386	**99.97%**	305386	**99.97%**
<typeOfResource>	281297	**92.09%**	281297	**92.09%**
<originInfo>	9170	**3.00%**	118825	**38.90%**
<language>	366	**0.12%**	165496	**54.18%**
<physicalDescription>	24356	**7.97%**	24356	**7.97%**
<subject>	201162	**65.85%**	201162	**65.85%**
<location>	1130	**0.37%**	281748	**92.23%**
<accessCondition>	179058	**58.62%**	305468	**100.00%**
<recordInfo>	308	**0.10%**	305468	**100.00%**

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