Nailing Jello to a Wall: Metrics, Frameworks, & Existing Work for Metadata Assessment

Christina Harlow

asis&t Webinar: Thursday, April 27, 2017

http://bit.ly/JelloToAWall
http://bit.ly/JelloToAWall
About Your Speaker

Metadata Librarian
Cornell University Library

cmh329@cornell.edu
@cm_harlow

http://bit.ly/JelloToAWall
About Your Speaker

Metadata Librarian
Cornell University Libraries

Repository Specialist, Data Operations
Stanford University Libraries

cmh329@cornell.edu
cmharlow@stanford.edu
@cm_harlow

http://bit.ly/JelloToAWall
Topics in Today's Webinar

I. Use Cases for Metadata Assessment

http://bit.ly/JelloToAWall
Topics in Today's Webinar

I. Use Cases for Metadata Assessment
II. Metrics, Context, & “Quality”
Topics in Today's Webinar

I. Use Cases for Metadata Assessment
II. Metrics, Context, & “Quality”
III. Guidelines for Performing Assessment

http://bit.ly/JelloToAWall
Topics in Today's Webinar

I. Use Cases for Metadata Assessment
II. Metrics, Context, & “Quality”
III. Guidelines for Performing Assessment
IV. Examples of Analysis Workflows & Tools

http://bit.ly/JelloToAWall
Topics in Today's Webinar

I. Use Cases for Metadata Assessment
II. Metrics, Context, & “Quality”
III. Guidelines for Performing Assessment
IV. Examples of Analysis Workflows & Tools
V. Further Resources & Engagement

http://bit.ly/JelloToAWall
I. Use Cases for Metadata Assessment

http://bit.ly/JelloToAWall
Moving Beyond Discovery Interfaces Checking as Metadata Assessment
Why Do We Assess Metadata?

- Handling New Object Types
- Impact of Metadata Work
- Migrations & Data Sharing
- Profile Generation
- Standards Choice
- System Design Aid
- Targeted Enhancement
- Validation & Expectations

http://bit.ly/JelloToAWall
Handling New Object Types

Surfacing needs of special or unique types of materials that either are not sufficiently captured for current metadata usage, do not fit well within existing profiles or standards.

<table>
<thead>
<tr>
<th>Field</th>
<th>external NS mapping</th>
<th>expected value</th>
<th>obligation</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>RIFF:ARL</td>
<td></td>
<td>example (default): Cornell U Library</td>
<td>(1,1)</td>
<td>Specified by client organization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>example: Khan Library</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIFF:ART</td>
<td>0</td>
<td></td>
<td>(0,1)</td>
<td>Blank - unless specified by client organization</td>
</tr>
<tr>
<td>RIFF:ICMS</td>
<td></td>
<td>example (default): Cornell U</td>
<td>(0,1)</td>
<td>If not Cornell University, leave blank or use client specified format</td>
</tr>
<tr>
<td>RIFF:IMT</td>
<td>example: ()</td>
<td>damaged cassette, tape rehoused. One minute and twenty seconds of silence between separate interviews removed from file.</td>
<td>(0,1)</td>
<td>Digitization notes unique to file. Usually blank, indicates copy. Notes when circulation copy is captured is assumed otherwise.</td>
</tr>
<tr>
<td>RIFF:ICOP</td>
<td>0</td>
<td></td>
<td>(0,0)</td>
<td>Blank - in catalog record</td>
</tr>
<tr>
<td>RIFF:ICRD</td>
<td>0</td>
<td></td>
<td>(0,0)</td>
<td>Blank - in catalog record</td>
</tr>
<tr>
<td>RIFF:ICRP</td>
<td>0</td>
<td></td>
<td>(0,0)</td>
<td>Blank - not applicable</td>
</tr>
<tr>
<td>RIFF:IDM</td>
<td>0</td>
<td></td>
<td>(0,0)</td>
<td>Blank - not applicable</td>
</tr>
<tr>
<td>RIFF:IDR</td>
<td>0</td>
<td></td>
<td>(0,0)</td>
<td>Blank - not applicable</td>
</tr>
<tr>
<td>RIFF:INEG</td>
<td>example (default): Fike, Karl E.</td>
<td>(1,1)</td>
<td>Supervising Engineer, e.g. person responsible for QC.</td>
<td></td>
</tr>
</tbody>
</table>

http://bit.ly/JelloToAWall
Impact of Metadata Work

Broad area to both measure the impact of metadata in discovery or other systems (through analytics or other), as well as to link metadata assessment to other areas of work, such as training/reskilling.

http://bit.ly/JelloToAWall
Migrations & Data Sharing

Assessment work done to support or enable the sharing, lossless conversion, or migration of metadata and data between data systems, standards, and repositories.

http://bit.ly/JelloToAWall
Profile Generation

Metadata Application Profile: resource that defines the expected, recommended, & optional fields, as well as proposed values sources & standards, for metadata in particular application.

http://bit.ly/JelloToAWall
Standards Choice

Decision of which standards- metavocabs, controlled vocabularies, encoding, formats, or other - best fit the current needs, the proposed needs, and the existing & proposed instance metadata.

http://bit.ly/JelloToAWall
Targeted Enhancement

Assessing metadata for areas of work at intersection of most impactful according to context, but also most efficient to perform normalization or enhancement work with given resources.

http://bit.ly/JelloToAWall
Validation & Expectations

Checking metadata follows a certain standard, profile, schema, or other meta-vocabulary, &/or conforms to the defined structure, usage, & expectations.

http://bit.ly/JelloToAWall
Metadata Assessment & Systems

http://bit.ly/JelloToAWall
Other Reasons for Assessment...

Metadata “Quality”
Alternate Discovery?
Metadata Assessment as Research

http://bit.ly/JelloToAWall
Metadata Assessment First Involves Setting Context & Scope
Otherwise...

Nailing Jello to a Wall: U.S. English idiom that describes a task that is difficult because the parameters keep changing (like how Jello/Jell-o moves).

http://bit.ly/JelloToAWall
II. Metrics, Context, & “Quality”
Some Writing & Research...

- Zavalina, Oksana; Kizhakkethil, Priya; et al. (2015). Building a Framework of Metadata Change to Support Knowledge Management.
Some Practice...


http://bit.ly/JelloToAWall
Some Proposed Metadata Quality Metrics

Accessibility

Accuracy

Availability (Technical)

Completeness

Conciseness

Conformance to expectations

Consistency & Coherence

Interlinking

Interoperability

Licensing

Normalization & Enhancement

Performance

Provenance

Timeliness

http://bit.ly/JelloToAWall
Metadata allows multiple access points via language, shared understanding of concepts, indication of accessibility, or other.
Correct use of the field;
Appropriate values captured;
Correctness of metadata.

Accuracy

http://bit.ly/JelloToAWall
Data server response;
Presence of data dumps;
Correct content types.

Availability

http://bit.ly/JelloToAWall
Completeness

Obligations of fields;

Required or recommended;

Data retrieval & capture in fields.

http://bit.ly/JelloToAWall
Avoid redundancy of fields, whether through multiple fields usage that have same meaning, or through annotations & schema usage.

http://bit.ly/JelloToAWall
Conformance to expectations

Use of standards and standard data formatting;
Obligations for fields are fulfilled.
Consistency & Coherence

Field values are normalized as applicable;

Fields are used consistently across instance data.

Yes

- A property not used by any other data
- A specific instance of a property that is used multiple times (i.e. first or last instance) that is consistently found in EVERY RECORD
- In the same property or small subset of properties in EVERY RECORD (including attribute variations)

In other words, something that can be logically predicted.

NO

- Must be parsed out of a data value (e.g. all the ones that start with “http://… etc.)
- Sometimes occurs in a specific instance of a repeated field but not in EVERY RECORD
- Occurs in a variety of properties, or in the same property with a variety of attributes

In other words, something that requires human intelligence or sophisticated logic to find.

http://bit.ly/JelloToAWall
Good quality interlinks;

Links to external datasets, data publishers;

Check for link rot.

http://bit.ly/JelloToAWall
Interoperability

Reuse of external schema, terms, vocabularies;

Clear indication of source of terms & fields.

http://bit.ly/JelloToAWall
Presence of license;
License assigned is machine-readable;
Assigned license is correct.
Normalization & Enhancement

Previous cleanup, enhancement, or normalization jobs have been run on the metadata; Values or scores present from enhancements.

http://bit.ly/JelloToAWall
Low latency where applicable;

High throughput (able to handle many HTTP requests);

Scalability of data publication.
Provenance

History of metadata creation/edits;

Originating source of metadata & metadata additions.

http://bit.ly/JelloToAWall
Currency of the data captured;
Connection between changing resources & updated metadata.

http://bit.ly/JelloToAWall
More Diverse, Interconnected Metadata Require Defining of Edges for Assessment
Metadata Assessment Also Includes Data Management Practices Review
III. Guidelines for Performing Assessment
Metadata Application Profiles

1. What are you describing with this metadata?
2. What do you intend to do with this metadata?
   a. Share with or generate from other systems?
   b. Enable some sort of discovery, lookup, resource management, or other functionality?
   c. Use within a particular system?
3. How will this metadata be generated, managed, and exposed? By whom or what processes?

Generic MAP Starter Template

http://bit.ly/JelloToAWall
# Metadata Application Profiles

**PCDM:Object > HydraWorks:Work | Article**

Articles that belong to a particular Issue. This seems to be the level that files and OCR sit at (thus nullifying the need for page objects - but need to confirm).

Will require a lot of metadata improvements.

### Descriptive Profile

<table>
<thead>
<tr>
<th>field name</th>
<th>predicate</th>
<th>mapping or collection-wide static value [range]</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>dcterms:identifier</td>
<td>HEADER/FILEDESC/PUBLICATIONSTMT/IDNO [literal]</td>
<td>n/a</td>
</tr>
<tr>
<td>title</td>
<td>dcterms:title</td>
<td>TEXT/BODY/DIV1/DI V2/HEAD/TITLE</td>
<td>n/a</td>
</tr>
<tr>
<td>OCR</td>
<td>dcterms:relations</td>
<td>TEXT/BODY/DIV1/DI V2/P/P</td>
<td>This is at article level?</td>
</tr>
</tbody>
</table>

### Structural Profile

Build Out Your Data Documentation with Your Assessment Tools
Machine-Actionable Mappings

```python
marc_map("020a",isbn.$append)
join_field(isbn," | ")
split_field(isbn, " | ")
marc_map("245ab",title, join: " ")

# Subject Identifiers?
if marc_match('6**0', '\w+')
    set_field("subjectRecon", "exists")
else
    set_field("subjectRecon", "does not exist")
end

# Name Identifiers?
if marc_match('1***0', '\w+')
    set_field("nameRecon", "exists")
else
    set_field("nameRecon", "does not exist")
end
```

```
<ConceptIdentifier>
  rml:logicalSource [
  ];
  rml:referenceFormulation ql:JSONPath;
  rml:iterator "$"
];

rr:subjectMap [
  rr:template "http://vitropilot.internal.library.cor";
];

rr:predicateObjectMap [
  rr:predicate rdfs:label;
  rr:objectMap [ rml:reference "$.conceptDetailInfo.d" ];
];

rr:predicateObjectMap [
  rr:predicate foaf:name;
http://bit.ly/JelloToAWall
]```
Validation Profiles & “Continuous Testing”

```javascript
describe 'bf:originDate from Title, 245 subfield f' do
  context '$f - Inclusive dates', :bf do
    let!(:graph) {
      marcxml = '<record xmlns="http://www.loc.gov/MARC21/slim">
          <leader>00056nam 2200229 45000</leader>
          <controlfield tag="001">catalogKeyID</controlfield>
          <controlfield tag="005">19911001004553.0</controlfield>
          <controlfield tag="007">hd-afa014bacd</controlfield>
          <controlfield tag="008">870660r19761854ctu a 000 0 eng d</controlfield>
          <datafield tag="245" ind1="1" ind2="0">
            <subfield code="a">Diaries</subfield>
            <subfield code="f">1854-1921</subfield>
            <subfield code="h">[Microform]</subfield>
          </datafield>
      </record>'
      self.send(MARC2BF_GRAPH_METHOD, marcxml, '245_subfield_f_title')
    }
  end
end
```

http://bit.ly/JelloToAWall
Semi-Automated / Targeted Human Review

(venv) $ python analysis/oaidc_analysis.py data/carli_bra_jack.oai.qdc.xml -i -p -e 'date' | grep 'False'

    oai:carliliilnois.edu:bra_jack/2200 False
    oai:carliliilnois.edu:bra_jack/2201 False
    oai:carliliilnois.edu:bra_jack/2202 False
    oai:carliliilnois.edu:bra_jack/2203 False
    oai:carliliilnois.edu:bra_jack/2204 False
    oai:carliliilnois.edu:bra_jack/2205 False
    oai:carliliilnois.edu:bra_jack/2206 False
    oai:carliliilnois.edu:bra_jack/2207 False
    oai:carliliilnois.edu:bra_jack/2208 False
    oai:carliliilnois.edu:bra_jack/2209 False
    oai:carliliilnois.edu:bra_jack/2210 False
    oai:carliliilnois.edu:bra_jack/2211 False
    oai:carliliilnois.edu:bra_jack/2212 False
    oai:carliliilnois.edu:bra_jack/2213 False
    oai:carliliilnois.edu:bra_jack/2214 False

...
Metadata Assessment Will Sometimes Require Derivative Datasets
IV. Examples of Analysis Workflows & Tools

http://bit.ly/JelloToAWall
Using the Tools You Got

MARCEdit

http://bit.ly/JelloToAWall
Using the Tools You Got

OpenRefine

http://bit.ly/JelloToAWall
Building Out the Duct Tape You Need

Python Metadata Breakers

http://bit.ly/JelloToAWall
Building Out the Duct Tape You Need

Catmandu Metadata Breakers

```
$ catmandu convert MARC to Breaker --handler marc < t/camel.usmarc > result.breaker
$ catmandu breaker result.breaker

<table>
<thead>
<tr>
<th>name</th>
<th>count</th>
<th>zeros</th>
<th>zeros%</th>
<th>min</th>
<th>max</th>
<th>mean</th>
<th>median</th>
<th>mode</th>
<th>variance</th>
<th>stdev</th>
<th>uniq</th>
<th>entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>10</td>
<td>0</td>
<td>0.0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>10/3.3</td>
</tr>
<tr>
<td>003</td>
<td>10</td>
<td>0</td>
<td>0.0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>10/3.3</td>
</tr>
<tr>
<td>005</td>
<td>10</td>
<td>0</td>
<td>0.0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>10/3.3</td>
</tr>
<tr>
<td>008</td>
<td>10</td>
<td>0</td>
<td>0.0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>10/3.3</td>
</tr>
<tr>
<td>010a</td>
<td>10</td>
<td>0</td>
<td>0.0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>10/3.3</td>
</tr>
<tr>
<td>020a</td>
<td>9</td>
<td>1</td>
<td>10.0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.97</td>
<td>9</td>
<td>9.0/3.3</td>
</tr>
<tr>
<td>040a</td>
<td>10</td>
<td>0</td>
<td>0.0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>10/3.3</td>
</tr>
<tr>
<td>040c</td>
<td>10</td>
<td>0</td>
<td>0.0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>10/3.3</td>
</tr>
<tr>
<td>040d</td>
<td>5</td>
<td>5</td>
<td>50.0</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>[0, 1]</td>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>1.0/3.3</td>
</tr>
<tr>
<td>042a</td>
<td>10</td>
<td>0</td>
<td>0.0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>10/3.3</td>
</tr>
<tr>
<td>050a</td>
<td>10</td>
<td>0</td>
<td>0.0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>10/3.3</td>
</tr>
<tr>
<td>050b</td>
<td>10</td>
<td>0</td>
<td>0.0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>10/3.3</td>
</tr>
<tr>
<td>0822</td>
<td>10</td>
<td>0</td>
<td>0.0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>10/3.3</td>
</tr>
<tr>
<td>082a</td>
<td>10</td>
<td>0</td>
<td>0.0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>10/3.3</td>
</tr>
<tr>
<td>100a</td>
<td>9</td>
<td>1</td>
<td>10.0</td>
<td>0</td>
<td>1</td>
<td>0.09</td>
<td>0.3</td>
<td>8</td>
<td>3.1</td>
<td>3.3</td>
<td>0.9</td>
<td>3.0/3.3</td>
</tr>
<tr>
<td>100d</td>
<td>1</td>
<td>9</td>
<td>90.0</td>
<td>0</td>
<td>1</td>
<td>0.1</td>
<td>0</td>
<td>0.09</td>
<td>0.3</td>
<td>1</td>
<td>0.5</td>
<td>3.3/3.3</td>
</tr>
</tbody>
</table>
```

http://bit.ly/JelloToAWall
Selective Querying

SQL/SPARQL & Response Checks

```
SELECT DISTINCT ?obj (COUNT(DISTINCT ?subject) as ?count)
WHERE {
    ?subject fedorasys:hasModel "Book" ;
}
GROUP by ?obj
```

http://bit.ly/JelloToAWall
Field frequency of 'Proxy/dcterm:alternative' per data providers

Metadata Quality Assurance Framework

Select field: Proxy/dcterm:alternative

This chart shows the frequency of the analyzed fields in all records. 100% means that the field is available in every record, 0% means that this field is never available. The numbers are rounded to 2 decimals.

<table>
<thead>
<tr>
<th>Provider</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Österreichische Nationalbibliothek - Austrian National Library</td>
<td>79.93%</td>
</tr>
<tr>
<td>National Library of the Netherlands</td>
<td>70.05%</td>
</tr>
<tr>
<td>Biblioteca Virtual de la Presa Historica</td>
<td>65.85%</td>
</tr>
<tr>
<td>Bayerische Staatsbibliothek</td>
<td>54.45%</td>
</tr>
<tr>
<td>Bildarchiv Foto Marburg</td>
<td>38.26%</td>
</tr>
<tr>
<td>National Library of Spain</td>
<td>36.26%</td>
</tr>
<tr>
<td>National Library of the Netherlands - Koninklijke Bibliotheek</td>
<td>35.84%</td>
</tr>
<tr>
<td>National Library of Finland</td>
<td>35.84%</td>
</tr>
<tr>
<td>Rijksdienst voor het Cultureel Erfgoed</td>
<td>35.84%</td>
</tr>
<tr>
<td>National Library of Denmark</td>
<td>35.84%</td>
</tr>
<tr>
<td>Royal Botanic Gardens, Kew</td>
<td>35.84%</td>
</tr>
<tr>
<td>Cineteca Luce S.p.A.</td>
<td>35.84%</td>
</tr>
<tr>
<td>Bodleian Libraries, University of Oxford</td>
<td>35.84%</td>
</tr>
<tr>
<td>The Danish Agency for Culture</td>
<td>35.84%</td>
</tr>
<tr>
<td>CERES Red Digital de Colecciones de museos de España</td>
<td>35.84%</td>
</tr>
<tr>
<td>Complutense University Library of Madrid</td>
<td>35.84%</td>
</tr>
<tr>
<td>Bibliothek der Friedrich-Ebert-Stiftung</td>
<td>35.84%</td>
</tr>
<tr>
<td>Europeana 1900-1930</td>
<td>35.84%</td>
</tr>
<tr>
<td>Stiftelsen Nordiska museet</td>
<td>35.84%</td>
</tr>
<tr>
<td>The British Library</td>
<td>35.84%</td>
</tr>
<tr>
<td>Ghent University Library</td>
<td>35.84%</td>
</tr>
<tr>
<td>Galiciana Biblioteca Digital de Galicia</td>
<td>35.84%</td>
</tr>
<tr>
<td>National Library of Wales</td>
<td>35.84%</td>
</tr>
<tr>
<td>Staatsbibliothek zu Berlin - Preußischer Kulturbesitz</td>
<td>35.84%</td>
</tr>
<tr>
<td>Botanic Garden and Botanical Museum Berlin-Dahlem</td>
<td>35.84%</td>
</tr>
<tr>
<td>Arhim Republika Slovenija</td>
<td>35.84%</td>
</tr>
<tr>
<td>National and University Library of Slovenia</td>
<td>35.84%</td>
</tr>
</tbody>
</table>
| [Link](http://bit.ly/JelloToAWall)
V. Further Resources & Engagement

http://bit.ly/JelloToAWall
DLF AIG
Metadata Working Group
(Digital Library Federation Assessment Interest Group)

dlfmetadataassessment.github.io
www.zotero.org/groups/metadata_assessment

http://bit.ly/JelloToAWall
Europeana Task Force on Metadata Quality

http://bit.ly/JelloToAWall
DPLA Quality Assessment Working Group

(Digital Public Library of America)
Metadata Assessment Needs Your Involvement!
Acknowledgements

Members of DLF AIG Metadata Working Group

Members of Europeana / DPLA QA Efforts, Special Nod to Péter Király, Antoine Isaacs, & Gretchen Gueguen

Members of Open Library Technology Development Communities, especially Mark Phillips, Corey Harper & Patrick Hochstenbach

Everyone who has sat through my evolving set of workshops around this topic

http://bit.ly/JelloToAWall
Nailing Jello to a Wall: Metrics, Frameworks, & Existing Work for Metadata Assessment

Christina Harlow

asis&t Webinar, Thursday, April 27, 2017

http://bit.ly/JelloToAWall