Exploration of metadata change in a digital repository

Oksana L. Zavalina, Assistant Professor
Priya Kizhakkethil, PhD student
Ongoing exploratory research into the changes in metadata records over time

- large regional distributed digital library that versions its metadata records

**Findings** regarding:

- metadata elements receiving the most editing
- the most prevalent types of metadata change
- distribution of metadata change types across metadata elements.
Background: Computer Science

- **Change** in texts, strings, files, scripts, etc.

- Mechanisms for identifying change
  - **edit distance** (e.g., Bille, 2005)

- **File comparison tools** (DIFF, COMM, PRETTY DIFF, PROMPTDIFF, etc.) for isolating differences between:
  - files
  - programs
  - applications
  - ontologies
  - multiple versions of the same entities.

  (e.g., Cheney, 2010; Horwitz, 1990; Noy et al., 2004)
Metadata quality research:
- suggested the link between metadata change and metadata quality
- emphasized the need to measure the metadata change and its outcomes for the users

Almost no published research identifying and measuring metadata change
- In part due to unavailability – until recently – of open-source (or inexpensive proprietary) information systems that allow metadata versioning.

(Stvilia et al., 2004; Stvilia & Gasser, 2008)
Background: Information Science: evaluating metadata change

- **Zavalina et al., 2008**: small-scale qualitative analysis as part of metadata quality study in IMLS DCC aggregation (140+ collection-level records):
  - frequency of revisions in collection-level IMLS DCC metadata records.
  - 2 broad categories of metadata revisions: change & addition

- **Tarver et. al., 2014**: “big data” quantitative analysis of metadata change in UNT Digital Collections over a period of 4 years (600,000+ item-level records):
  - overall frequency distribution of metadata change events
  - change in metadata record length, access status, etc.
Gaps to address

- Lack of studies identifying and measuring metadata change in information science research

- Broad scope of existing studies (macro-analysis)

Need for deeper, more granular investigation (micro-analysis) of metadata change

- at the level of individual records, metadata elements, and data values.
Data collection: main sample

Dublin Core records created by human metadata creators

- created between October 1, 2009, & December 31, 2012
- visible to end-users both at the time of creation & at the time of data collection (April 2014)
- representing different collections in repository, and describing different kinds/genres of information objects
- edited at least once, with last editing in January-April 2014

157 records X 2 (initial & latest versions) = 314
Dublin Core records created by human metadata creators:

- created between October 1, 2009, & December 31, 2012.
- visible to end-users at the time of creation & at the time of data collection (April 2014).
- representing different collections in the repository, and describing different kinds/genres of information objects.
- edited at least once, with the last editing in January-April 2014.

# of Record Versions:

- Initial: 22%
- Intermediate: 17%
- Latest: 13%
- 2 versions: 10%
- 4 versions: 9%
- 5 versions: 4%
- 6 versions: 2%
- 7 versions: 1.3%
- 8 versions: 0.6%
- 9 versions: 0.6%
- 10 versions: 0.6%
- 11 versions: 0.6%
- 15 versions: 0.6%
- 17 versions: 0.6%

O.L.Zavalina, & P.Kizhakkethil © 2015
Data collection: sample #2

Subsample (expanded) from the main sample

- 33% of all records edited 3 times

11 records × 4 (all versions) = 44

Initial

1st revision

2nd revision

Latest (3rd revision)
Research questions

How and when do metadata records change in a digital repository?

- What **categories of change** can be identified & what is the relative frequency of their occurrence?
- In which **metadata fields** does **change occur** the most often?
- How is **metadata change** related to:
  - record age?
  - number of editing events?
  - fluctuations in the record length?
- How is the metadata **change distributed across editing events (over time)**?
## Metadata change: basic statistics

<table>
<thead>
<tr>
<th>Type of change</th>
<th>Total no. of metadata change instances observed</th>
<th>Number of fields with change per record</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Min</td>
</tr>
<tr>
<td>addition</td>
<td>254</td>
<td>0</td>
</tr>
<tr>
<td>deletion</td>
<td>534</td>
<td>0</td>
</tr>
<tr>
<td>modification</td>
<td>475</td>
<td>0</td>
</tr>
<tr>
<td>change overall</td>
<td>1263</td>
<td>1</td>
</tr>
</tbody>
</table>
Metadata change types by record field (% of records)
Metadata change examples
(iConference 2012 paper record)
No change vs. multiple change

<table>
<thead>
<tr>
<th>Field</th>
<th>No Change</th>
<th>Multiple Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>94%</td>
<td>71%</td>
</tr>
<tr>
<td>Creator</td>
<td>35%</td>
<td>32%</td>
</tr>
<tr>
<td>Contributor</td>
<td>25%</td>
<td>50%</td>
</tr>
<tr>
<td>Publisher</td>
<td>46%</td>
<td>0.6%</td>
</tr>
<tr>
<td>Date</td>
<td>50%</td>
<td>0%</td>
</tr>
<tr>
<td>Language</td>
<td>50%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Description</td>
<td>68%</td>
<td>7.0%</td>
</tr>
<tr>
<td>Subject</td>
<td>23%</td>
<td>1.9%</td>
</tr>
<tr>
<td>Primary Source</td>
<td>71%</td>
<td>24.2%</td>
</tr>
<tr>
<td>Coverage</td>
<td>81%</td>
<td>36.9%</td>
</tr>
<tr>
<td>Source</td>
<td>50%</td>
<td>0%</td>
</tr>
<tr>
<td>Citation</td>
<td>96%</td>
<td>0%</td>
</tr>
<tr>
<td>Relation</td>
<td>96%</td>
<td>35.7%</td>
</tr>
<tr>
<td>Collection</td>
<td>97%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Institution</td>
<td>98%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Rights</td>
<td>99%</td>
<td>0.0%</td>
</tr>
<tr>
<td>ResourceType</td>
<td>96%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Format</td>
<td>97%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Identifier</td>
<td>98%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Degree</td>
<td>99%</td>
<td>0.6%</td>
</tr>
<tr>
<td>Note</td>
<td>96%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

O.L.Zavalina, & P.Kizhakkethil © 2015
Change in the number of field instances (% of records)

results in substantial change — mostly increase, sometimes decrease — in the length of the record
## Metadata change: correlations

<table>
<thead>
<tr>
<th>Pearson’s r</th>
<th>Record age</th>
<th># of record versions</th>
<th>record length increase/decrease: initial to latest version</th>
<th># of edited fields in record</th>
</tr>
</thead>
<tbody>
<tr>
<td>Record age</td>
<td>X</td>
<td>-0.23944</td>
<td>0.49472</td>
<td>-0.248</td>
</tr>
<tr>
<td># of record versions</td>
<td>-0.23944</td>
<td>X</td>
<td>0.2099</td>
<td>0.7741</td>
</tr>
<tr>
<td>record length increase/decrease: initial to latest version</td>
<td>0.49472</td>
<td>0.2099</td>
<td>X</td>
<td>0.1167</td>
</tr>
<tr>
<td># of edited fields in record</td>
<td>-0.248</td>
<td>0.7741</td>
<td>0.1167</td>
<td>X</td>
</tr>
</tbody>
</table>
When does metadata change?

No. of instances in 11 records changed 3 times

- **addition**
- **deletion**
- **modification**

<table>
<thead>
<tr>
<th>Editing events</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>14</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

O.L. Zavalina, & P. Kizhakkethil © 2015
Implications

Digital library/repository development:

• improving and maintaining metadata quality

• strategically distributing scarce resources
Future/concurrent research

Similar studies in various environments:

• digital libraries
• institutional repositories:
  • research **products** (publications, presentations etc.)
  • research **data**
• bibliographic databases (e.g., WorldCat)

**which use other metadata schemes**

& enable metadata **versioning**
Works cited

THANK YOU!

Questions?
Comments?
Ideas?

Your feedback is very welcome!

Oksana L. Zavalina: Oksana.Zavaina@unt.edu