Metadata Capital in a Data Repository

Dublin Core 2013 – Lisbon Portugal
4 September 2013

- Jane Greenberg, Univ. of North Carolina, Metadata Research Center
- Shea Swauger, Colorado State University Libraries
- Elena M. Feinstein, Duke University
Overview

- ‘Capital’ and metadata capital
- **Dryad Repository**
  - Research questions
  - Methods, sample, and procedures
- Results and conclusions
- Q & A
‘Capital’ and metadata capital

- An economic concept (Weber, 1905; Smith’s, 1776)
  - Business and operations (net gains or losses)
  - Finances, goods and services, and public needs
- Intellectual capital (Marr, 2005)
- Social capital

*a tangible result, value increase*
‘Capital’ and metadata capital

- Metadata as an asset, a product
  - Reuse of **good quality metadata** increase value of initial investment
    - Poor quality may reduce metadata capital
  - Metadata reuse prevalence
    - Cooperative cataloging, CIP, ISBD, MARC, FRBR, LCC, VIAF, OAI-PMH, CrossRef, PubMed, Zotero, BibTex, DataCite. Linked data/Semantic Web, PIDs, etc.
“a curated general-purpose repository that makes the data underlying scientific publications discoverable, freely reusable, and citable.

enables scientists to validate published findings, repurpose data, etc.
Research Questions

- Does metadata reuse within Dryad’s curation workflow build metadata capital?
- Where is metadata reuse most common or lacking?
- How might Dryad’s curation workflow and/or repository design be modified to enable greater metadata capital and efficiency?

**SAMPLE:** 20 Cases/100 instantiations

**PROCEDURES**
- Capture
- Archive
- Track
- Flag
- Calculate
Flagging – *metadata change*

- Addition (metadata property content added)
- Deletion (metadata removed)
- Modification (metadata edited)
- Reuse (metadata property content was not flagged as changed, it was recorded as reuse)
FIGURE 2: PHASE 1 - CASE A & B COMBINED

- **dc:contributor.author**
  - Reused: 74
  - Added: 0
  - Deleted: 7
  - Modified: 0
  - 91% Reuse

- **correspondingAuthor**
  - Reused: 8
  - Added: 2
  - Deleted: 0
  - Modified: 0
  - 90% Reuse

- **dc:coverage.spatial**
  - Reused: 12
  - Added: 0
  - Deleted: 0
  - Modified: 0
  - 0% Reuse

- **dc:coverage.temporal**
  - Reused: 7
  - Added: 0
  - Deleted: 0
  - Modified: 0
  - 0% Reuse

- **dc:date.issued**
  - Reused: 1
  - Added: 4
  - Deleted: 5
  - Modified: 0
  - 10% Reuse

- **dc:description**
  - Reused: 8
  - Added: 8
  - Deleted: 2
  - Modified: 0
  - 40% Reuse

- **relation.isreferencedby**
  - Reused: 10
  - Added: 0
  - Deleted: 0
  - Modified: 0
  - 100% Reuse

- **dc:subject**
  - Reused: 92
  - Added: 0
  - Deleted: 7
  - Modified: 1
  - 92% Reuse

- **dc:title**
  - Reused: 14
  - Added: 2
  - Deleted: 4
  - Modified: 0
  - 70% Reuse

- **dwc:ScientificName**
  - Reused: 22
  - Added: 0
  - Deleted: 0
  - Modified: 0
  - 0% Reuse

- **prism:publicationName**
  - Reused: 20
  - Added: 0
  - Deleted: 0
  - Modified: 0
  - 100% Reuse
### FIGURE 4: TOTAL METADATA WORKFLOW
**PHASES 1 & 2 - CASES A & B**

<table>
<thead>
<tr>
<th>Metadata Field</th>
<th>Reused 1</th>
<th>Added 1</th>
<th>Deleted 1</th>
<th>Modified 1</th>
<th>Reused 2</th>
<th>Added 2</th>
<th>Deleted 2</th>
<th>Modified 2</th>
<th>Total Reuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc:contributor.author</td>
<td>74</td>
<td></td>
<td></td>
<td></td>
<td>74</td>
<td></td>
<td></td>
<td></td>
<td>93%</td>
</tr>
<tr>
<td>dc:correspondingAuthor</td>
<td>8</td>
<td>2</td>
<td>16</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>16</td>
<td>4</td>
<td>80%</td>
</tr>
<tr>
<td>dc:coverage.spatial</td>
<td>12</td>
<td>12</td>
<td></td>
<td></td>
<td>12</td>
<td>12</td>
<td></td>
<td></td>
<td>46%</td>
</tr>
<tr>
<td>dc:coverage.temporal</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td></td>
<td>7</td>
<td>7</td>
<td>1</td>
<td></td>
<td>46%</td>
</tr>
<tr>
<td>dc:date.issued</td>
<td>14</td>
<td>5</td>
<td>10</td>
<td>9</td>
<td>14</td>
<td>5</td>
<td>10</td>
<td>9</td>
<td>6%</td>
</tr>
<tr>
<td>dc:description</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>17</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>17</td>
<td>65%</td>
</tr>
<tr>
<td>dc:identifier.citation</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>dc:relation.isReferencedBy</td>
<td>10</td>
<td>10</td>
<td></td>
<td></td>
<td>10</td>
<td>10</td>
<td></td>
<td></td>
<td>50%</td>
</tr>
<tr>
<td>dc:subject</td>
<td>92</td>
<td></td>
<td></td>
<td></td>
<td>92</td>
<td></td>
<td></td>
<td></td>
<td>95%</td>
</tr>
<tr>
<td>dc:title</td>
<td>14</td>
<td>2</td>
<td>4</td>
<td>13</td>
<td>14</td>
<td>2</td>
<td>4</td>
<td>13</td>
<td>67%</td>
</tr>
<tr>
<td>dwc:ScientificName</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td>48%</td>
</tr>
<tr>
<td>prism:publicationName</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td>95%</td>
</tr>
</tbody>
</table>
Package metadata harvested from email

DCContributor

Contr. 101 (gr. 99%, bl. 1%)

Pkg metadata (exact harvest)

DDICorresp

Pkg metadata (some editing)

DCDescription

Pkg metadata (not from email)

DCTitle

Email metadata (not used)

DCSubject

Subj. 177 (gr. 97%, rd. 2%, bl. 1%)

DCSpatial

Spat. 35

DCTemporal

Temp. 2

DwCSci.Name

DwCSci. 26
File metadata harvested from package metadata

Contr. 100 (gr. 93%, bl. 7%)

DCContributor

Subj. 185 (gr. 83 %, or. 1%, red 4 %, bl. 12 %)

DCSubject

Contr. 100 (gr. 93%, bl. 7%)

DCSpatial

File metadata (inherit exactly)

DCTemporal

File metadata (some editing)

DwCSci.Name

File metadata (created, not inherited)

Pkg metadata not used for file

Subj. 177 (gr. 97%, rd. 2%, bl. 1%)
Results/summary

- Reuse for nearly all of the properties
  - 8 of 12 metadata properties had reuse @ 50% or greater
  - 5 of 8 confirmed reuse at 80% or higher.
  - Reuse more common w/basic bib. properties,
  - Lacking/non-existent for more complex, scientific properties (spatial, taxonomic, spatial, + temporal)

Dryad’s workflow...likely builds capital

- Areas for improving workflow/techniques
  - Complex scientific properties
  - HIVE – Controlled terminologies (consistency + quality)
  - Identifiers, referenced by, etc. (automate)
Conclusions

- Limitation: size, cases, etc.
- Generating **quality metadata** at time of initial creation is paramount
  - Sweet spot (e.g., title article/data title change)
- Reuse not novel; context may spark interest
  - First articulation of the notion of metadata capital
  - Methodology provides a framework
  - Baseline for future analyses
- Outside comfort zone
- Formulas being explored - HIVE/US DataNets
Acknowledgments and more information

- **SILS-UNC-CH** <Metadata Research Center> (ls.unc.edu/mrc/)
  

- Dryad Consortium Board, journal partners, and data authors ([datadryad.org](http://datadryad.org); blog.datadryad.org; datadryad.org/wiki)

- NESCent

- NCSU Digital Libraries: and Yale/TreeBASE

- DataONE

- British Library and Oxford University

National Science Foundation (NSF), Award number: 1147166
Dryad: scalable and sustainable infrastructure for the publication of data