Giving Data Context:

a comparison study of institutional repositories that apply varying degrees of curation

Amy Koshoffer, University of Cincinnati
Amy Neeser, University of California, Berkeley
Lisa R. Johnston, University of Minnesota
Linda Newman, University of Cincinnati
Research Questions

1. How do the metadata elements vary for each institution?

2. How complete is the metadata submission for datasets in each institution repository given the type of curation?

3. Are curated datasets more likely to have documentation associated with the work?

4. Does the number of datasets with DOIs vary given the type of curation?

5. What is the difference in number of keywords associated with each dataset?
What Do We Mean By Curation?

Data Curation steps:

- appraisal/selection
- check/run files
  - includes code review
  - review for sensitive information,
  - licensing and rights management checks etc.
- working authors to collect missing files and to create custom documentation (e.g., readme.txt files)
- metadata augmentation (i.e., curators supplement the author-supplied metadata)
- file format transformations

(Johnston, 2017)
What Do We Mean By Curation?

Cincinnati
- **no formal curation** process
- handle issues as they arose
- can mint a DOI for a work as a step in the submission process

Michigan
- **selective curation**
- in cases where researchers willingly participated either before or after deposit
- could mint a DOI for a work as an optional step after the submission process

Minnesota
- **post-ingest curation** by team of 6 domain-based data curators
- suggested levels of standard description and documentation (a required component)
- DOIs were manually assigned to submissions only after minimum curation standards were met

Oregon State
- **pre-ingest curation**
- standard levels of description and documentation, and
- made datasets public only when they met curation standards.
- DOIs were automatically assigned to submissions.
## Participating Institutions

**Time frame of Study:** 80 data sets = 20 Data sets sampled from each institution as of December 31, 2016

<table>
<thead>
<tr>
<th>Institution</th>
<th>Repository Name</th>
<th>Repository Type</th>
<th>Curation Type</th>
<th>Repository Software</th>
<th>Start Date of Repo</th>
<th>Total # of datasets as of 10-17-2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Cincinnati</td>
<td>Scholar@UC</td>
<td>General IR</td>
<td>No curation</td>
<td>Hydra Fedora (Samvera)</td>
<td>September 2015</td>
<td>48</td>
</tr>
<tr>
<td>University of Michigan</td>
<td>Deep Blue Data</td>
<td>Data-only IR</td>
<td>Selective curation</td>
<td>Hydra Fedora (Samvera)</td>
<td>September 2016</td>
<td>85</td>
</tr>
<tr>
<td>University of Minnesota</td>
<td>Data Repository for the University of Minnesota (DRUM)</td>
<td>Data-only IR</td>
<td>Post-ingest curation</td>
<td>DSpace</td>
<td>March 2015</td>
<td>148</td>
</tr>
<tr>
<td>Oregon State University</td>
<td>ScholarsArchive@OSU</td>
<td>General IR</td>
<td>Pre-ingest curation</td>
<td>DSpace</td>
<td>February 2005</td>
<td>70</td>
</tr>
</tbody>
</table>
1. How do the metadata elements vary for each institution?

<table>
<thead>
<tr>
<th>Institution</th>
<th>Minimum % Completeness</th>
<th>Required Fields</th>
<th>Optional Fields</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cincinnati</td>
<td>35%</td>
<td>Title, Creator(s), College, Department or Program, Description, License, Access rights</td>
<td>Publisher, Required software, DO, Date created, Alternate title, Subject, Geographic subject, Time period, Language, Citation, Note, External link, Related Works</td>
</tr>
<tr>
<td>Michigan</td>
<td>60%</td>
<td>Title, Creator, Method, Description, CC License, Discipline</td>
<td>Date Coverage, Keyword, Language, Citation to Related Work(s)</td>
</tr>
<tr>
<td>Minnesota</td>
<td>16%</td>
<td>Title, Contact</td>
<td>Author(s), Group Author, Subject Keywords, Abstract, Description, DOI, Fundex Information, Date of Collection - start, Date of Collection - end, Date Completed, Citation to Related Paper(s), Time Period, Geographic, Area/Coordinates, Source Information, Source Data URL, License Type</td>
</tr>
<tr>
<td>Oregon State</td>
<td>0%</td>
<td>Title, Contact Email</td>
<td>Title, License, Authors, ORCID, Abstract, Subject(s) or Keyword(s), Contributor(s), Date(s), Sponsorship, Related materials, Format of data, Version, Geolocation, Affiliations, Contact name, Contact email address, Username, Embargo</td>
</tr>
</tbody>
</table>
### Table 3: Detailed Comparison of Metadata Elements for Each Institution

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Dublin Core Element</th>
<th>Institution</th>
<th>Submission Form Display Name</th>
<th>Req?</th>
<th>Metadata Elements Used by All Four Institutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title of the Dataset</td>
<td>dc.title</td>
<td>Cincinnati</td>
<td>Title</td>
<td>✓</td>
<td>Title ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Michigan</td>
<td>Title</td>
<td>✓</td>
<td>Title ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minnesota</td>
<td>Title</td>
<td>✓</td>
<td>Title ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oregon State</td>
<td>Title</td>
<td>✓</td>
<td>Title ✓</td>
</tr>
<tr>
<td>Author or Creator of the Dataset</td>
<td>dc.creator</td>
<td>Cincinnati</td>
<td>Creator(s)</td>
<td>✓</td>
<td>Creator ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Michigan</td>
<td>Creator</td>
<td>✓</td>
<td>Creator ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minnesota</td>
<td>Author(s)</td>
<td>✓</td>
<td>Author ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oregon State</td>
<td>Lead Investigator(s)</td>
<td>✓</td>
<td>Lead Investigator ✓</td>
</tr>
<tr>
<td>License applied to the dataset</td>
<td>dc.rights</td>
<td>Cincinnati</td>
<td>License</td>
<td>✓</td>
<td>License ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Michigan</td>
<td>CC License</td>
<td>✓</td>
<td>CC License ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minnesota</td>
<td>License Type</td>
<td>✓</td>
<td>License Type ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oregon State</td>
<td>License</td>
<td>✓</td>
<td>License ✓</td>
</tr>
<tr>
<td>Related works or publications that use or cite the dataset</td>
<td>dc.isReferencedBy(MN) dc.relation.isReferenced by (MN) dc.description (OSU)</td>
<td>Cincinnati</td>
<td>External Link (unmapped)</td>
<td>✓</td>
<td>External Link (unmapped) ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Michigan</td>
<td>Citation to Related Work(s)</td>
<td>✓</td>
<td>Citation to Related Work(s) ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minnesota</td>
<td>Citation to Related Paper(s)</td>
<td>✓</td>
<td>Citation to Related Paper(s) ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oregon State</td>
<td>Related materials</td>
<td>✓</td>
<td>Related materials ✓</td>
</tr>
<tr>
<td>Subject Terms or Keywords that describe the topic of the dataset</td>
<td>dc.subject</td>
<td>Cincinnati</td>
<td>Subject</td>
<td>✓</td>
<td>Subject ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Michigan</td>
<td>Keyword</td>
<td>✓</td>
<td>Keyword ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minnesota</td>
<td>Subject Keywords</td>
<td>✓</td>
<td>Subject Keywords ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oregon State</td>
<td>Keywords</td>
<td>✓</td>
<td>Keywords ✓</td>
</tr>
<tr>
<td>DOI</td>
<td>dc.identifier.doi RDFS.doi (MI)</td>
<td>Cincinnati</td>
<td>DOI (assigned outside of the submission process)</td>
<td>✓</td>
<td>DOI ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Michigan</td>
<td>Persistent Identifier™</td>
<td>✓</td>
<td>Persistent Identifier™ ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Minnesota</td>
<td>DOI</td>
<td>✓</td>
<td>DOI ✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oregon State</td>
<td>DOI</td>
<td>✓</td>
<td>DOI ✓</td>
</tr>
</tbody>
</table>

**Title of the Dataset** `dc.title`

- Cincinnati: Title ✓
- Michigan: Title ✓
- Minnesota: Title ✓
- Oregon State: Title ✓

**Related works or publications that use or cite the dataset**

- Cincinnati: External Link (unmapped)
- Michigan: Citation to Related Work(s)
- Minnesota: Citation to Related Paper(s)
- Oregon State: Related materials
2. How complete is the metadata submission for datasets in each institution repository given the type of curation?

- All datasets had more than the minimum required metadata.
2. How complete is the metadata submission for datasets in each institution repository given the type of curation?

- **Overall average percent complete was 58%** for all 80 data sets 

  (green linear line)
2. How complete is the metadata submission for datasets in each institution repository given the type of curation?

- Datasets fell in a **broad range** of metadata completeness
  - Cincinnati (40-85%)
  - Michigan (60%-100%)
  - Minnesota (42%-95%)
  - Oregon (25%-30%)
2. How complete is the metadata submission for datasets in each institution repository given the type of curation?

- **Average percent completeness for each institution**
  - Cincinnati = 53%
  - Michigan = 78%
  - Minnesota = 70%
  - Oregon State = 29%
2. How complete is the metadata submission for datasets in each institution repository given the type of curation?

- **Mann-Whitney U test** showed significant difference in completeness profiles in pairwise comparisons (i.e. UC to UM, UC to UMN, etc.)
2. How complete is the metadata submission for datasets in each institution repository given the type of curation?

- No remarkable increase in optional metadata fields completed in repo’s w/ curation support (Oregon, Minnesota) over repo’s w/o consistent curation support (Cincinnati, Michigan).
2. How complete is the metadata submission for datasets in each institution repository given the type of curation?

- **Minnesota** (which employed post-ingest curation for all datasets) saw the greatest increase (54%) from the percent completeness of required fields (16%) to the average percent completeness (70%), which could be attributed, at least in part, to curation.

54% jump in ave completeness over min
2. How complete is the metadata submission for datasets in each institution repository given the type of curation?

- Michigan (which did not routinely curate author-submitted metadata for the datasets) had the overall highest average percent completeness of 78% benefiting from its requiring 60% of its metadata fields and by having fewer metadata fields available.
2. How complete is the metadata submission for datasets in each institution repository given the type of curation?

- **Oregon** which used a pre-ingest curation method and has no required fields, **averaged only 29% completeness** among optional metadata fields.

29% jump in ave completeness over min
2. How complete is the metadata submission for datasets in each institution repository given the type of curation?

- Our findings are inconclusive to directly link curation with metadata completeness and it is not possible to conclude that curator intervention will result in more completion of metadata beyond the minimum required fields.

Does curation impact metadata completeness?
3. Are curated datasets more likely to have documentation associated with the work?

- 20 out of 20 datasets at Minnesota (post-ingest curation) reported documentation
3. Are curated datasets more likely to have documentation associated with the work?

- 15 of the 20 submissions at Oregon State (pre-ingest curation) included documentation

---

Additional Dataset Documentation Submitted

<table>
<thead>
<tr>
<th>Institution</th>
<th>Number of Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cincinnati</td>
<td>20</td>
</tr>
<tr>
<td>Michigan</td>
<td>15</td>
</tr>
<tr>
<td>Minnesota</td>
<td>10</td>
</tr>
<tr>
<td>Oregon State</td>
<td>25</td>
</tr>
</tbody>
</table>

Legend:
- Other
- Codebook
- Data Dictionary
- Readme
- None
3. Are curated datasets more likely to have documentation associated with the work?

- Data sets at Michigan (selective curation) and Cincinnati (no curation) had very low numbers of documentation files.
4. Does the number of datasets with DOIs vary given the type of curation?

- Each institution supported Digital Object Identifiers (DOIs)
4. Does the number of datasets with DOIs vary given the type of curation?

- **100% of the datasets** from the two **institutions** with curation have **DOIs**.
4. Does the number of datasets with DOIs vary given the type of curation?

- **100% of the datasets from the two institutions with curation have DOIs.**

- The high number of DOI’s for datasets at Michigan suggests factors other than curation apply.
5. What is the difference in number of keywords associated with each dataset?

- **No institutions required keywords** or used a controlled vocabulary list, i.e. Library of Congress Subject Headings or Medical Subject Heading terms.
5. What is the difference in number of keywords associated with each dataset?

- The majority of the datasets had at least **5 keywords** added per dataset in three of the four institutions.
- The average number of keywords overall was 4.35.
5. What is the difference in number of keywords associated with each dataset?

- Cincinnati was the outlier as most datasets had no keywords.
- This may be due to the input form format which hid keywords.
Conclusions

The **curation process** may have had a measurable impact on the metadata captured. However, there are institutional factors that may be more impactful.

Curation **did result in more documentation**, especially the inclusion of **readme files**, with a dataset submission.

Additionally there were more keywords and DOI’s for datasets that were curated, but differences could not confidently be attributed to curation.
Recommendations

1. Consider institutional factors when evaluating curation.

1. Consider curation practices and offer best practice guidelines for various scenarios.

1. Metadata schema should be standardized to promote interoperability between IRs.

1. The community should evaluate the differences between schemas and develop a minimum requirement for metadata for datasets in IRs.
References

1. Title page and bottom of slide photo credit (crop of larger image): Park Güell, Barcelona, https://www.flickr.com/photos/seba_sofariu/9534760477, CC BY 2.0


1. Closing slide photo credit - Barcelona by oh-barcelona.com

1. Thank you to Steven Van Tyl for contributing the data for Oregon State University
Thank you, *Gracias*!

Multi-coloured, Recycled Decorations at the Fiesta de Gracia