IASSIST 2017

Data Curation
Perception and Practice

Lisa Johnston, Jake Carlson, Cynthia Hudson-Vitale, and Wendy Kozlowski
Challenges for Data Curation Services

- How to scale local data curation services across all disciplines?
- How many data curation experts are needed?
  - Types: GIS, spreadsheet/tabular, statistical/survey, software code, video/audio…
  - Disciplines: genomic sequence, chemical spectra, bioinformatics…
- Are there ways to more efficiently curate rare or infrequently generated data types?
- Might our institution specialize in curation skills and represent our academic expertise.
Our Mission

*The Data Curation Network will enable academic institutions to better support researchers that are faced with a growing number of requirements to ethically share their research data. In the next 3-5 years we will...*

1. Develop standards-driven data curation techniques for all types of repository workflows and infrastructure.
2. Expand into a sustainable entity that grows beyond our initial six partner institutions.
3. Datasets curated by the Data Curation Network will be used to advance research and education in ways that are measurably of greater reuse value than non-curated data.
4. Build an innovative community that enriches capacities for data curation writ large.

[https://sites.google.com/site/datacurationnetwork/](https://sites.google.com/site/datacurationnetwork/)
Project Team

Lisa Johnston (PI)
Research Data Management & Curation Lead,
University of Minnesota

Jake Carlson
Research Data Services Manager,
University of Michigan

Robert Olendorf
Science Data Librarian,
Penn State University

Cynthia Hudson-Vitale
Data Services Coordinator,
Washington University in St. Louis

Heidi Imker
Director of the Research Data Service,
University of Illinois

Wendy Kozlowski
Data Curation Specialist,
Cornell University

Claire Stewart
Associate University Librarian for Research and Learning,
University of Minnesota
Model for the Data Curation Network

Uncurated Data
Presenting scale and expertise challenges to individual institutions

Ingest

Appraise and Select

DCN

Facilitate Access

Preserve Long-Term

Curated Data
at scale and with great efficiency through shared Data Curation Network
Model for the Data Curation Network

Local Curator Workflow

Uncurated Data
Presenting scale and expertise challenges to individual institutions

Ingest → Appraise and Select → DCN → Facilitate Access → Preserve Long-Term → Curated Data
at scale and with great efficiency through shared Data Curation Network

DCN Coordinator Workflow

Review → Assign → Curate → Mediate → Approve

DCN Curator Workflow

Check files and metadata → Understand and run files → Request missing information → Augment metadata → Transform file formats → Evaluate for FAIRness
Perception vs Practice for Data Curation

Focus for our panel today:

➢ **Perception**
  - Perceived importance of data curation activities by *researchers*
  - Perceived importance of data curation activities by *libraries*

➢ **Practice**
  - Panel: Four snapshots of institutional data curation services in practice
  - Results of piloting the Data Curation Network

➢ Unpacking the Findings

➢ Discussion with the Audience
Perception
Results of Engagement with Researchers and Librarians
Perceptions of Data Curation

- What data curation activities are important to Researchers?
- What data curation activities are important to Librarians?
- Are there gaps between the perceptions of researchers and librarians?

Source: Flickr - https://flic.kr/p/ccmW5
# Data Curation Activities

<table>
<thead>
<tr>
<th>Data Curation Activity</th>
<th>Data Curation Network Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documentation</td>
<td>Information describing any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text Readme file).</td>
</tr>
<tr>
<td>Chain of custody</td>
<td>Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third-parties.</td>
</tr>
<tr>
<td>Secure Storage</td>
<td>Data files are properly stored in a well-configured (in terms of hardware and software) storage environment that is routinely backed-up and physically protected. Perform routine fixity checks (to detect degradation or loss) and provide recovery services as needed.</td>
</tr>
</tbody>
</table>

Poll: Importance of Data Curation Activities

** Participation implies consent for us to share the deidentified results
POLL: The first IASSIST conference was held in 1984 in Cocoa Beach, FL with 29 representatives from Canada and the US. How many IASSIST conferences have you attended?

<table>
<thead>
<tr>
<th>Number of Conferences Attended</th>
<th>Number of Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>This is my first</td>
<td></td>
</tr>
<tr>
<td>2-3</td>
<td></td>
</tr>
<tr>
<td>3-5</td>
<td></td>
</tr>
<tr>
<td>5-10</td>
<td></td>
</tr>
<tr>
<td>&gt;10</td>
<td></td>
</tr>
</tbody>
</table>
DOCUMENTATION: Information describing necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text "Readme" file).

1 - This activity is not important

2

3

4

Start the presentation to activate live content

If you see this message in presentation mode, install the add-in or get help at PollEv.com/app
SIONING: Provide mechanisms to ingest versions of the data overtime that includes metadata describing the version history and any changes made for each version.

1: This activity is not important

2

3

4

Start the presentation to activate live content
If you see this message in presentation mode, install the add-in or get help at PollEv.com/app
TRICTED ACCESS (Data Enclave): In order to maintain the privacy of research subjects without losing integral components of the data, some data access will be protected and/or mediated to individuals that meet predefined criteria.

1: This activity is not important

Start the presentation to activate live content
CONTENT IDENTIFIER: A URL (or Uniform Resource Locator) is monitored by an authority to ensure a stable web location for consistent citation and long-term discoverability. Provides redirects when necessary (e.g., DOI or handle).

1: This activity is not important

2

3

4

Start the presentation to activate live content

If you see this message in presentation mode, install the add-in or get help at Pollev.com/app
REVIEW: Run and validate computer output (e.g., look for missing files and/or errors) in order to find mistakes overlooked in the initial development phase, improving the overall quality of software.

1: This activity is not important
2
3
4

Start the presentation to activate live content
If you see this message in presentation mode, install the add-in or get help at Poliev.com/app
Researcher Engagement Events

Focus groups (Oct-Nov 2016) at each of the 6 DCN partner institutions asked:

• What data curation activities are important to you?

• What data curation activities are currently being done by you or a 3rd party?

• If the data curation activity is being performed, how satisfied are you with the results?

Source: https://unsplash.com/search/notes?photo=PJzc7LOt2lg
Data Curation Activities

Each institution selected Data Curation Activities from our list of 35 possibilities.

Metadata
Information about a data set that is structured (often in machine-readable format) for purposes of search and retrieval. Metadata elements may include basic information (e.g. title, author, date created, etc.) and/or specific elements inherent to datasets (e.g., spatial coverage, time periods).

Rate how important this activity is to you.
(Write a number 1-5 with 5 = highest importance, 1 = not important)

<table>
<thead>
<tr>
<th>Round 1</th>
<th>Round 2</th>
<th>Round 3</th>
<th>Round 4</th>
</tr>
</thead>
</table>

### Importance of Curation Activities

<table>
<thead>
<tr>
<th>Rank</th>
<th>Curation Activity</th>
<th>CU</th>
<th>WU</th>
<th>IL</th>
<th>PSU</th>
<th>MN</th>
<th>MI</th>
<th># Inst.</th>
<th>Total Responses</th>
<th>1-5 Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>Documentation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>6</td>
<td>91</td>
<td>4.6</td>
</tr>
<tr>
<td>#2</td>
<td>Secure Storage</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>4</td>
<td>60</td>
<td>4.4</td>
</tr>
<tr>
<td>#3</td>
<td>Quality Assurance</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>5</td>
<td>73</td>
<td>4.3</td>
</tr>
<tr>
<td>#4</td>
<td>Persistent Identifier</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>6</td>
<td>91</td>
<td>4.3</td>
</tr>
<tr>
<td>#5</td>
<td>Software Registry</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>2</td>
<td>29</td>
<td>4.1</td>
</tr>
<tr>
<td>#6</td>
<td>Data Visualization</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>2</td>
<td>24</td>
<td>4.0</td>
</tr>
</tbody>
</table>

For activities rated by more than one institution. Ranking on a 1 to 5 scale with 5 = Most Important (CU = Cornell, PSU = Penn State, I = Illinois, WU = Washington University, MI = Michigan, MN = Minnesota).
## Researcher Engagement Events Results

<table>
<thead>
<tr>
<th>Rank</th>
<th>Curation Activity</th>
<th>CU</th>
<th>WU</th>
<th>IL</th>
<th>PSU</th>
<th>MN</th>
<th>MI</th>
<th># Inst.</th>
<th>Total Responses</th>
<th>1-5 Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>#7</td>
<td>File Audit</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3</td>
<td>49</td>
<td>4.0</td>
</tr>
<tr>
<td>#8</td>
<td>Metadata</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>5</td>
<td>80</td>
<td>4.0</td>
</tr>
<tr>
<td>#9</td>
<td>Versioning</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>6</td>
<td>91</td>
<td>3.9</td>
</tr>
<tr>
<td>#10</td>
<td>Contextualize</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>6</td>
<td>91</td>
<td>3.9</td>
</tr>
<tr>
<td>#11</td>
<td>Code Review</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>6</td>
<td>29</td>
<td>3.9</td>
</tr>
<tr>
<td>#12</td>
<td>File Format Transformations</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>5</td>
<td>73</td>
<td>3.8</td>
</tr>
</tbody>
</table>

For activities rated by more than one institution. Ranking on a 1 to 5 scale with 5 = Most Important (CU = Cornell, PSU = Penn State, I = Illinois, WU = Washington University, MI = Michigan, MN = Minnesota).
Researcher Engagement Events Results

Most Important Activities (4 out of 5)
- (Create) Documentation (4.6)
- Secure Storage (4.4)
- Quality Assurance (4.3)
- Persistent Identifier (4.3)
- Software Registry (4.1)
- Data Visualization (4.0)
- File Audit (4.0)
- (Create) Metadata (4.0)
- Versioning (3.9)
- Contextualization (3.9)
- Code Review (3.9)
- File Format Transformations (3.9)

Not Happening for Majority of Researchers
- Persistent Identifier (37% happens)
- Software Registry (41% happens)
- File Audit (16% happens)
- Contextualization (38% happens)
- Code Review (38% happens)

Happening, but not satisfactorily
- Documentation (26% satisfied),
- Secure storage (38% satisfied),
- Quality Assurance (14% satisfied),
- Data Visualization (12.5% satisfied),
- Metadata (29% satisfied)
- Versioning (13% satisfied)
- File Format Transformations (29% satisfied)
Researcher Engagement Events Results

Most Important Activities (4 out of 5)
- (Create) Documentation (4.6)
- Secure Storage (4.4)
- Quality Assurance (4.3)
- Persistent Identifier (4.3)
- Software Registry (4.1)
- Data Visualization (4.0)
- File Audit (4.0)
- (Create) Metadata (4.0)
- Versioning (3.9)
- Contextualization (3.9)
- Code Review (3.9)
- File Format Transformations (3.9)

Not Happening for Majority of Researchers
- Persistent Identifier (37% happens)
- Software Registry (41% happens)
- File Audit (16% happens)
- Contextualization (38% happens)
- Code Review (38% happens)

Happening, but not satisfactorily
- Documentation (26% satisfied),
- Secure storage (38% satisfied),
- Quality Assurance (14% satisfied),
- Data Visualization (12.5% satisfied),
- Metadata (29% satisfied)
- Versioning (13% Satisfied)
- File Format Transformations (29% satisfied)
Poll: Data Curation Activities Taking Place

** Participation implies consent for us to share the deidentified results**
DOCUMENTATION: Information describing necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text "Readme" file).

- No, this is not happening for my data
- Yes, this is happening but I'm not satisfied with it
- Yes, this is happening and I'm somewhat satisfied with it
- Yes, this is happening and I'm satisfied with it
- I don't know!

Start the presentation to activate live content
N/A If you see this message in presentation mode, install the add-in or get help at PollEv.com/app
SIONING: Provide mechanisms to ingest versions of the data overtime that includes metadata describing the version history and any changes made for each version.

No, this is not happening for my data
Yes, this is happening but I'm not satisfied with it
Yes, this is happening and I'm somewhat satisfied with it
Yes, this is happening and I'm satisfied with it
I don't know!

Start the presentation to activate live content
If you see this message in presentation mode, install the add-in or get help at PolLev.com/app

N/A for me or my data
TRICTED ACCESS (Data Enclave): In order to maintain the privacy of research subjects without losing integral components of the data, some data access will be protected and/or mediated to individuals that meet predefined criteria.

- No, this is not happening for my data
- Yes, this is happening but I'm not satisfied with it
- Yes, this is happening and I'm somewhat satisfied with it
- Yes, this is happening and I'm satisfied with it
- I don't know!

**Start the presentation to activate live content**

N/A If you see this message in presentation mode, install the add-in or get help at PollEv.com/app
PERSISTENT IDENTIFIER: A URL (or Uniform Resource Locator) that is monitored by an authority to ensure a stable web location for consistent citation and long-term discoverability. Provides redirection when necessary (e.g., DOI or handle).
DE REVIEW: Run and validate computer debugging (e.g., look for missing files and/or errors) in order to find mistakes overlooked in the initial development phase, improving the overall quality of software.

No, this is not happening for my data
Yes, this is happening but I'm not satisfied with it
Yes, this is happening and I'm somewhat satisfied with it
Yes, this is happening and I'm satisfied with it
I don't know!

Start the presentation to activate live content

N/A If you see this message in presentation mode, install the add-in or get help at PollEv.com/app
Survey of the 124 members of the Association of Research Libraries

Distributed in January 2017

80 libraries responded (65%)
• 51 currently provide services
• 13 are developing services

http://publications.arl.org/Data-Curation-SPEC-Kit-354/
### Data Curation @ ARL Libraries

<table>
<thead>
<tr>
<th>Rschr. Rank</th>
<th>Activity</th>
<th>Offer</th>
<th>Plan to Offer</th>
<th>Percentage of Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>Documentation</td>
<td>36</td>
<td>3</td>
<td>80%</td>
</tr>
<tr>
<td>#2</td>
<td>Secure Storage</td>
<td>39</td>
<td>2</td>
<td>84%</td>
</tr>
<tr>
<td>#3</td>
<td>Quality Assurance</td>
<td>22</td>
<td>1</td>
<td>47%</td>
</tr>
<tr>
<td>#4</td>
<td>Persistent Identifier</td>
<td>40</td>
<td>2</td>
<td>86%</td>
</tr>
<tr>
<td>#5</td>
<td>Software Registry</td>
<td>4</td>
<td>2</td>
<td>12%</td>
</tr>
<tr>
<td>#6</td>
<td>Data Visualization</td>
<td>14</td>
<td>4</td>
<td>37%</td>
</tr>
</tbody>
</table>

Number of ARL institutions = 49
# Data Curation @ ARL Libraries

<table>
<thead>
<tr>
<th>Rschr. Rank</th>
<th>Activity</th>
<th>Offer</th>
<th>Plan to Offer</th>
<th>Percentage of Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>#7</td>
<td>File Audit</td>
<td>21</td>
<td>7</td>
<td>57%</td>
</tr>
<tr>
<td>#8</td>
<td>Metadata</td>
<td>43</td>
<td>1</td>
<td>90%</td>
</tr>
<tr>
<td>#9</td>
<td>Versioning</td>
<td>24</td>
<td>3</td>
<td>55%</td>
</tr>
<tr>
<td>#10</td>
<td>Contextualization</td>
<td>28</td>
<td>4</td>
<td>65%</td>
</tr>
<tr>
<td>#11</td>
<td>Code Review</td>
<td>4</td>
<td>1</td>
<td>10%</td>
</tr>
<tr>
<td>#12</td>
<td>File Format Transformation</td>
<td>25</td>
<td>5</td>
<td>61%</td>
</tr>
</tbody>
</table>

Number of ARL institutions = 49
## Libraries Plan to Support ... (or do we?)

<table>
<thead>
<tr>
<th>Rsch. Rank</th>
<th>Data Curation Activity</th>
<th>Currently providing</th>
<th>Plan or would like to provide</th>
<th>No Interest</th>
<th>Unsure</th>
</tr>
</thead>
<tbody>
<tr>
<td>#31</td>
<td>Repository Certification</td>
<td>3</td>
<td>30</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>#11</td>
<td>Code Review</td>
<td>4</td>
<td>29</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>#32</td>
<td>Emulation</td>
<td>1</td>
<td>26</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>#23</td>
<td>Peer Review</td>
<td>1</td>
<td>22</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>#5</td>
<td>Software Registry</td>
<td>4</td>
<td>23</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>#30</td>
<td>Deidentification</td>
<td>8</td>
<td>25</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>#17</td>
<td>Interoperability</td>
<td>11</td>
<td>28</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>
Practice
Institutional Data Curation Service
Case Studies
DCN Baseline Assessment

Table 1: Comparison of the data curation workflows at the six institutions

<table>
<thead>
<tr>
<th>Workflow Steps by Institution</th>
<th>Pre-ingest Curation?</th>
<th>Mediated vs Self-deposit?</th>
<th>Accept/Reject Stage?</th>
<th>Public</th>
<th>Post-ingest curation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consult only</td>
<td>Consult only</td>
<td>Staging Area for deposit</td>
<td>Mediated deposit</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Staging Area for deposit</td>
<td>Mediated deposit</td>
<td>Self-deposit</td>
<td>Approval to accept or reject</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Self-deposit</td>
<td>Auto Accept</td>
<td>Go Live Here</td>
<td>As needed</td>
<td>Review metadata only</td>
<td>Review files and metadata</td>
</tr>
<tr>
<td>Minnesota</td>
<td>X</td>
<td>X*</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Cornell</td>
<td>X</td>
<td>X*</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Illinois</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Michigan</td>
<td>X</td>
<td>X*</td>
<td>X</td>
<td>X</td>
<td>X*</td>
</tr>
<tr>
<td>Michigan</td>
<td>X</td>
<td>X*</td>
<td>X</td>
<td>X</td>
<td>X*</td>
</tr>
<tr>
<td>Penn State</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Wash U</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

* On request

https://doi.org/10.7191/jeslib.2017.1102
Local Practice: Cornell University

eCommons
digital repository
ecommons.cornell.edu

Launched: 2002
First dataset: 2003 / 2005
Datasets published: 111

Features:
● DSpace 5.5 (Dublin Core)
● Curation encouraged
● Versioning encouraged
● Handles for all items, and DataCite DOI’s upon request
● “Embargos” for delayed publication needs
● Links to associated publications
Local Practice: Cornell University

Staffing

IR Administration and Development

Gail Steinhart
Repository Manager
Chloe McLaren
Admin and Policy Development
Mira Basara
Admin and Batch Upload Support
George Kozak
System Maintenance and Development

Data-specific Support and Curation

Wendy Kozlowski
Data Curation Specialist
Erica Johns
Research Data and Environmental Sciences

Data expertise from across Cornell - available for consultation on specific issues

Re:search data management service group
data.research.cornell.edu
Local Practice: University of Michigan

Deep Blue Data
https://deepblue.lib.umich.edu/data

Soft Launch: Feb 2016
Official Launch: Sept 2016

96 entries
- 60 Works
- 36 Collections

Features:
- Hydra Fedora (Dublin Core)
- DataCite DOIs
- Extension of Deep Blue (UM’s IR)
Local Practice: University of Michigan

Roles in Research Data Services

- **RDS Core Team**
  Oversee services including Deep Blue Data and the curation process

- **Liaisons**
  The front line in offering RDS. Consult on data deposit and curation

- **Specialists**
  Apply expertise as needed
University of Minnesota
Data Repository for U of M (DRUM)
Launched: March 2015
124 data sets published
Curation: Required

Features:
- DSpace 6.0 (Dublin Core)
- DataCite DOIs
- Versioning
- Web of Science Indexing
- Documentation assistance (e.g., readme.txt template)
Local Practice: Washington University

Digital Research Materials Repository

Launched -
Soft launch 2015

Features -
● BePress Platform
● DataCite DOI’s
● Restricted access
● 10 year preservation then review by subject liaison
Local Practice: Washington University

Staffing

Jennifer Moore, GIS & Data Projects Manager
Lauren Todd, Engineering Librarian
Daria Carson-Dussan, Romance Languages Librarian
Micah Zeller, Copyright Specialist
Emily Stenberg, Repository Manager
Cynthia Hudson-Vitale, Data Services Coordinator
Practice
Data Curation Pilots
DCN Curation Pilots

Fall of 2016 → conducted two rounds of controlled data curation pilots among 16 institutional data curation staff to:

1. Identify actual and individual curation practices taken at partner DCN institutions (compare).
2. Establish training needs of DCN curators.
3. Identify any issues, misaligned expectations, and/or conflicts with the goals of the project.

<table>
<thead>
<tr>
<th>Curators</th>
<th>Minnesota</th>
<th>Cornell</th>
<th>Penn State</th>
<th>Illinois</th>
<th>Michigan</th>
<th>Wash U</th>
</tr>
</thead>
</table>
**DCN Curation Pilots**

<table>
<thead>
<tr>
<th>Rsch. Rank</th>
<th>Data Curation Activities*</th>
<th>Round 1 (n=6)</th>
<th>Round 2 (n=9)</th>
<th>Total (n=15)</th>
<th>% Did This</th>
</tr>
</thead>
<tbody>
<tr>
<td>#3</td>
<td>Quality Assurance</td>
<td>6</td>
<td>8</td>
<td>14</td>
<td>93%</td>
</tr>
<tr>
<td>#1</td>
<td>(Create) Documentation</td>
<td>6</td>
<td>5</td>
<td>11</td>
<td>73%</td>
</tr>
<tr>
<td>#34</td>
<td>Correspondence (with author)</td>
<td>6</td>
<td>5</td>
<td>11</td>
<td>73%</td>
</tr>
<tr>
<td>#12</td>
<td>File Format Transformations</td>
<td>6</td>
<td>2</td>
<td>8</td>
<td>53%</td>
</tr>
<tr>
<td>#8</td>
<td>(Create) Metadata</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>53%</td>
</tr>
<tr>
<td>#10</td>
<td>Contextualization</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>47%</td>
</tr>
<tr>
<td>#28</td>
<td>File Inventory or Manifest</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>40%</td>
</tr>
<tr>
<td>#21</td>
<td>Risk Management</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>33%</td>
</tr>
</tbody>
</table>

*Not ranked by researchers, curators also Inspected Files, Inspected Metadata and (33%) created a working copy.*
DCN Curation Pilots

- **Recommendation 1:** Assignments to curators prioritize file format and software expertise over discipline when necessary.

- **Recommendation 3:** Centralize all DCN correspondence and perform routine checks on all submissions before assigning to DCN curator.

- **Recommendation 5:** Create levels of curator criteria for curators to aim for rather than allowing curators to fall into the “never ending” quest for high standards.

- **Recommendation 7:** Data curation activities taken should differentiate between the role of the local repository curators versus the role of the Data Curation Network curator.

https://sites.google.com/site/datacurationnetwork/
Our Findings
Data Curation Network
CURATE Steps

**Appraise**
Local curator appraises submission and determines if the data should be submitted to the DCN for curation.

**Review + Assign**
DCN Coordinator reviews submission and assigns to appropriate DCN curator based on file and discipline expertise.

**Curate**
DCN Curators performs expert curatorial review and reports back to DCN coordinator.

C - Check data files and read documentation
U - Understand the data (or try to), if not...
R - Request missing information or changes
A - Augment the submission with metadata for findability
T - Transform file formats for reuse and long-term preservation.
E - Evaluate and rate the overall submission for FAIRness.
# DCN Draft Procedures and Checklist

<table>
<thead>
<tr>
<th>CURATE Actions</th>
<th>Curation Checklist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Check data files and read documentation</td>
<td>☐ Files open as expected</td>
</tr>
<tr>
<td>• Review the content of the data files (e.g., open and run the files or code).</td>
<td>☐ Issues ________</td>
</tr>
<tr>
<td>• Verify all metadata provided by the author and review the available documentation.</td>
<td>☐ Code runs as expected</td>
</tr>
<tr>
<td></td>
<td>☐ Produces minor errors</td>
</tr>
<tr>
<td></td>
<td>☐ Does not run and/or produces many errors</td>
</tr>
<tr>
<td></td>
<td>☐ Metadata quality is rich, accurate, and complete</td>
</tr>
<tr>
<td></td>
<td>☐ Metadata has issues ________</td>
</tr>
<tr>
<td></td>
<td>☐ Documentation Type (circle)</td>
</tr>
<tr>
<td></td>
<td>Readme / Codebook / Data Dictionary / Other:</td>
</tr>
<tr>
<td></td>
<td>☐ Missing/None</td>
</tr>
</tbody>
</table>
Local Control

- **DCN Coordination**: DCN coordinator communicates needed actions back to the local curator.
- **Take Action**: Local curator assists local researcher with any actions (in person, via email, etc.).
- **Re-submit**: Researcher addresses curatorial issues and resubmits.
- **Finalize**: Local curator finalized data submission and notifies DCN coordinator.
- **Assess**: DCN curator reviews final data publication to access if DCN standards were met.

**DCN Stamp of Approval For FAIRness**
DCN Draft Procedures and Checklist

Evaluate and the overall data record for FAIRness.
- Score the dataset and recommend ways to increase the FAIRness of the data and become “DCN approved.”

**Findable**
- Metadata exceeds author/title/date,
- Unique PID (DOI, Handle, PURL, etc.)
- Discoverable via web search engines like Google

**Accessible**
- retrievable via a standard protocol (e.g., HTTP)
- Free, open download.

**Interoperable**
- Metadata formatted in a standard schema (e.g., Dublin Core)
- Metadata provided in machine-readable format (OAI feed)

**Reusable**
- Data include sufficient metadata about the data characteristics to reuse without the direct assistance of the author
- Clear indicators of who created, owns, and stewards the data help ensure trust
- Data are released with clear data usage
Data Curation Network Outcomes

1. Standards-driven data curation techniques for all types of repository workflows and infrastructure.
2. A sustainable entity that grows beyond our initial six partner institutions.
3. Datasets curated by the Data Curation Network will be used to advance research and education in ways that are measurably of greater reuse value than non-curated data.
4. An innovative community that enriches capacities for data curation writ large.
# Results

## Poll 1: Curation Activity Importance

<table>
<thead>
<tr>
<th>Activity</th>
<th>Average @ DCN Engagement Events</th>
<th>Average @ IASSIST 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documentation</td>
<td>4.6</td>
<td>4.7</td>
</tr>
<tr>
<td>Persistent Identifier</td>
<td>4.3</td>
<td>4.3</td>
</tr>
<tr>
<td>Versioning</td>
<td>3.9</td>
<td>4.1</td>
</tr>
<tr>
<td>Code Review</td>
<td>3.9</td>
<td>3.6</td>
</tr>
<tr>
<td>Restricted Access</td>
<td>2.6</td>
<td>4.0</td>
</tr>
</tbody>
</table>
# Results (partial)

## Poll 2: Curation Activity Engagement and Satisfaction

<table>
<thead>
<tr>
<th>Activity</th>
<th>Happening @ DCN Events</th>
<th>Happening @ IASSIST 2017</th>
<th>Satisfied / Somewhat Satisfied @ DCN events</th>
<th>Satisfied / Somewhat Satisfied @ IASSIST 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documentation</td>
<td>80%</td>
<td>89%</td>
<td>26 / 46 (72%)</td>
<td>15 / 54 (69%)</td>
</tr>
<tr>
<td>Persistent Identifier</td>
<td>37%</td>
<td>74%</td>
<td>19 / 33 (52%)</td>
<td>8 / 15 (23%)</td>
</tr>
<tr>
<td>Versioning</td>
<td>56%</td>
<td>68%</td>
<td>13 / 37 (50%)</td>
<td>25 / 20 (45%)</td>
</tr>
<tr>
<td>Code Review</td>
<td>39%</td>
<td>31%</td>
<td>22 / 14 (36%)</td>
<td>10 / 13 (23%)</td>
</tr>
<tr>
<td>Restricted Access</td>
<td>38%</td>
<td>67%</td>
<td>21 / 4 (25%)</td>
<td>12 / 26 (38%)</td>
</tr>
</tbody>
</table>
Discussion
Data Curation Perception vs Practice
Discussion Question

● What do we value in terms of curation and how does that align with our approach?

● How do we measure the value/impact of data curation?

● How do you know when data are “well curated”? What does “done” look like? Is it a matter of time spend, effort made, perfection?
Thanks!

Web: https://sites.google.com/site/DataCurationNetwork

Twitter #DataCurationNetwork
Planning a network of expertise model for curating research data in academic libraries

2016-2017

Planning the Data Curation Network project is supported by a grant from the ALFRED P. SLOAN FOUNDATION.