IDCC 2018

Launching the Data Curation Network

Lisa Johnston University of Minnesota
Jake Carlson University of Michigan
Cynthia Hudson--Vitale Washington Univ.
Heidi Imker University of Illinois
Wendy Kozlowski Cornell University
Robert Olendorf Penn State University
Claire Stewart University of Minnesota
Mara Blake Johns Hopkins University
Joel Herndon Duke University
Elizabeth Hull Dryad Data Repository
Timothy M. McGeary Duke University

02-20-2018
Researchers are faced with a growing number of requirements (and incentives) to ethically share their research data.
Well curated data are more valuable.
The skills and expertise required to curate data cannot be fully automated nor reasonably be provided by a few experts siloed at single institutions.
The Data Curation Network (DCN) addresses this challenge by collaboratively sharing data curation staff across a network of partner institutions and data repositories.
Outline of Talk

1. DCN planning phase research (2016-2017) *(save for end if time)*

2. Implementation pilot → launch the DCN in Spring 2018
   a. DCN staffing model + Advisory panel
   b. DCN training/networking events
   c. DCN workflow and C-U-R-A-T-E steps
   d. Assessment Plan

3. Grow the DCN beyond our grant-funded phase to a sustainable entity
   a. Criteria for new partners
   b. Proposed financial model (alliance curation-as-service)
The Data Curation Network (DCN) 3-year implementation phase is set to launch May 2018.
DCN Pilot Implementation

9 Institutions
- 8 Academic Libraries
- 1 General Data Repository

19 Data Curators

1 Project Coordinator (new hire)

1 Program Director (PI)

8 DCN Representatives (CO-PIs)

2 Admin Leads
DCN Pilot Implementation

Annual DCN “All-Hands” Meeting

Day 1: Business Meeting

- DCN Representatives + Admin
- Advisory Panel (non-partner representatives to help guide our implementation)

Day 2-3: Curator Training/Network

- Train on DCN Workflow
- Set expectations for documentation/best practice
- Intro tracking tool (Jira)
- Format-specific data deep dives
DATA CURATION WORKSHOP

SLIDES AND HANDOUTS

DATE: DECEMBER 11 & 12, 2017
TWEET: #DCW2017
LOCATION: WASHINGTON UNIVERSITY IN ST. LOUIS, MCMILLAN HALL, ST. LOUIS, MO

DESCRIPTION:
This free, 1.5 day workshop is open to all library staff and data professionals who are interested in data curation.
Participants will learn practical, hands-on treatments for data curation based on the Data Curation Network CURATE model.
- C - Check data files and read documentation;
- U - Understand the data (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 - Evaluates and rates the overall submission for FAIRness.

ATTENDEES WILL COME AWAY WITH:
1. A customized, implementable plan to enhance data curation activities at your local institution or organization,
2. Stakeholder focused talking points related to the value of data curation activities,
3. An in-depth understanding of specialized data curation practices in various disciplines, data types, and formats.
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

Data Curation Network
DCN Workflow

- Researchers deposit like normal
DCN Workflow

- Researchers deposit like normal
- DCN functions as a microservice layer (the “human layer in your repository stack”)
DCN Workflow

- Researchers deposit like normal
- DCN functions as a microservice layer (the “human layer in your repository stack”)
- Local institution maintain full responsibility for all technical functionality (e.g., storage) and authority for local decision-making (what to ingest, how long to retain, etc.)
DCN Workflow

Researchers deposit like normal
- DCN functions as a microservice layer (the “human layer in your repository stack”)
- Local institution maintain full responsibility for all technical functionality (eg. storage) and authority for local decision-making (what to ingest, how long to retain, etc.)
- Seamlessly integrates into all repository systems (Samvera, Fedora, DSpace, etc.)
DCN 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

Data Curation Network

DCN Coordinator Workflow

- Review
- Assign
- CURATE
- Mediate
- Approve

DCN Curator Workflow

- C: Check files and metadata
- U: Understand and run files
- R: Request missing information
- A: Augment metadata
- T: Transform file formats
- E: Evaluate for FAIRness
CURATE Steps in DCN Workflow

DCN Curators will take CURATE steps for each data set, that includes:

- **Check** data files and read documentation
- **Understand** the data (try to), if not...
- **Request** missing information or changes
- **Augment** the submission with metadata for findability
- **Transform** file formats for reuse and long-term preservation
- **Evaluate** and rate the overall submission for FAIRness.
# DCN Pilot Implementation (2018-2020)

## Table A1. Draft checklist of DCN CURATE steps and FAIRness scorecard

<table>
<thead>
<tr>
<th>CURATE Actions</th>
<th>Curation Checklist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Check data files and read documentation</td>
<td></td>
</tr>
<tr>
<td>- Review the content of the data files (e.g., open and run the files or code).</td>
<td>□ Files open as expected</td>
</tr>
<tr>
<td>- Verify all metadata provided by the author and review the available documentation.</td>
<td>□ Issues ______</td>
</tr>
<tr>
<td></td>
<td>□ Code runs as expected</td>
</tr>
<tr>
<td></td>
<td>□ Produces minor many errors</td>
</tr>
<tr>
<td></td>
<td>□ Metadata quality is rich, complete</td>
</tr>
<tr>
<td></td>
<td>□ Metadata has issues</td>
</tr>
<tr>
<td></td>
<td>□ Documentation Type (e.g., Readme / Codebook / Data Description)</td>
</tr>
<tr>
<td></td>
<td>□ Other:</td>
</tr>
<tr>
<td></td>
<td>□ Missing/None</td>
</tr>
<tr>
<td></td>
<td>□ Needs work</td>
</tr>
</tbody>
</table>

**Understand** the data (or try to)

- Check for quality assurance and usability issues such as missing

Evaluate and rate the overall data record for FAIRness.²

- 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 (e.g., download link).

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.
- Data are released with clear data usage terms (e.g., a CC License).

---

¹ Format Recommendations, [http://guides.library.cornell.edu/ecommons/formats](http://guides.library.cornell.edu/ecommons/formats)

² Rubric evaluating the FAIR principles are based on the scoring matrix by Dunning, de Smaele, & Böhmer (2017).
DCN Pilot Implementation (2018-2020)

Assessment Plan (two-prong)

Is a networked approach to curating research data more efficient?
- Number of datasets
- Frequency (high-volume time periods, etc.)
- Variety (data file formats; range of disciplines)
- Efficiency (time, costs)

Are curated data are more valuable?
- Track reuse indicators (download counts, citations, alt-metrics)
- Implement a DCN registry
- Apply badges and metadata to signal that data sets curated by the DCN are FAIR.
In Year 3, the DCN will begin transitioning to a self-sustaining service model where institutional and disciplinary partners contribute data curation staff and central operations costs are offset by users of the Network.
Data Curation-as-service

Alliance Model
Institutional Partners contribute staff and fund central coordinator

Data Curation Network

Institutional Partner
Local Representative
Local Curator(s)
roles distributed among 1 or many people at the institution as needed

DCN Coordinator
matches expertise from the pool of local Curators

DCN Curator
roles distributed among 1 or many people at the institution as needed

Individual and Institutional Users gain access to the Network on a fee-for-service basis
### Sustainability and Expansion (2020-)

<table>
<thead>
<tr>
<th>Stakeholder</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Academic libraries with existing data curation services</strong></td>
<td>Gain access to data curation expertise in more disciplines/formats than locally available</td>
</tr>
<tr>
<td><strong>Academic libraries with limited to no resources for data curation services</strong></td>
<td>Are able to provide critical new data curation services when local resources are limited (without needing to hire);</td>
</tr>
<tr>
<td><strong>Disciplinary- and general-subject data repositories</strong></td>
<td>Receive better, more valuable data submissions from DCN partner institutions and customers;</td>
</tr>
<tr>
<td></td>
<td>Have potential to partner with the DCN to expand the scope of curation support for new and/or less frequently encountered data types</td>
</tr>
</tbody>
</table>
## 6-year Roadmap toward Sustainability

### Transition from planning phase to sustaining phase

<table>
<thead>
<tr>
<th>Year</th>
<th>Support</th>
<th>Timing</th>
<th>Phase</th>
<th>Partners</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Sloan Grant</td>
<td>2016-17</td>
<td>Planning</td>
<td>6 academic institutions</td>
</tr>
<tr>
<td>1</td>
<td>Grant Funded (Y1-Y2)</td>
<td>2017-19</td>
<td>Implementation</td>
<td>8 academic institutions and 2 disciplinary partners</td>
</tr>
<tr>
<td>2</td>
<td>transition to partnership model (Y3)</td>
<td>2020-22</td>
<td>Transition</td>
<td>Recruit new partners as use and demand dictate</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>Sustaining</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Mission:** With a proven and appealing value-proposition, the Data Curation Network will expand into a sustainable entity that grows beyond our initial partner institutions.
The Data Curation Network (DCN) planning phase ran from 2016-2017 with six institutions and support from the Alfred P. Sloan Foundation.
Planning Phase (2016-2017)

1. **Compared local policy**, technologies, and workflows across the 6 planning phase institutions;
2. **Held six focus groups** with researchers on what data curation activities were important;
3. **Ran controlled pilots** of data curation workflows with 17 data curators to ID issues;
4. **Surveyed the 124** ARL institutions to gauge support for data curation services;
5. **Researched cost recovery** models for sustainable data curation and repository services;
6. **Held information exchanges** with leaders of successful collaboration projects;
7. **Analyzed one-year** of data types, disciplines, frequency, and curation levels needed vs taken).
Planning Phase (2016-2017)

How would we deal with conflicting policy issues?

What do researchers actually need our help with? Will they care if curation is distributed?

Can I trust someone else to curate our data? What about quality control?

What skills do we need? What types of data sets are deposited into our data repositories? How long does curation take?

Baseline Assessment

<table>
<thead>
<tr>
<th>Workflow Steps by Institution</th>
<th>Pre-ingest Curation?</th>
<th>Mediated vs Self-deposit?</th>
<th>Post-ingest curation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Consult only</td>
<td>Staging Area for deposit</td>
<td>Mediated deposit</td>
</tr>
<tr>
<td>Minnesota</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Cornell</td>
<td>X</td>
<td></td>
<td>X*</td>
</tr>
<tr>
<td>Illinois</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Michigan</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Penn State</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wash U</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Planning Phase (2016-2017)

How would we deal with conflicting policy issues?

What do researchers actually need our help with? Will they care if curation is distributed?

Can I trust someone else to curate our data? What about quality control?

What skills do we need? What types of data sets are deposited into our data repositories? How long does curation take?

Researcher Focus Groups (n=91)

Ave Rating = 3.7 out of 5

Secure Storage  
Quality Assurance  
Persistent IDs  
Software Registry  
Data Visualization  
Metadata  
Emulation  
Restricted Access  
Contact Information  
Full-Text Indexing  

Planning Phase (2016-2017)

How would we deal with conflicting policy issues?

What do researchers actually need our help with? Will they care if curation is distributed?

Can I trust someone else to curate our data? What about quality control?

What skills do we need? What types of data sets are deposited into our data repositories? How long does curation take?

ARL Institutions Survey (n=80)

Planning Phase (2016-2017)

How would we deal with conflicting policy issues?

What do researchers actually need our help with? Will they care if curation is distributed?

Can I trust someone else to curate our data? What about quality control?

What skills do we need? What types of data sets are deposited into our data repositories? How long does curation take?
Planning Phase (2016-2017)

How would we deal with conflicting policy issues?

What do researchers actually need our help with? Will they care if curation is distributed?

Can I trust someone else to curate our data? What about quality control?

What skills do we need? What types of data sets are deposited into our data repositories? How long does curation take?

One Year Metrics Tracking (n=175)

Planning Phase (2016-2017)

What are the measure of success?

How can we grow and sustain the Network beyond the grant-funding period?
Planning Phase (2016-2017)

What are the measure of success?

How can we grow and sustain the Network beyond the grant-funding period?

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-Kind</td>
<td>Institutions ally together to form in-kind network of data curators.</td>
</tr>
<tr>
<td>Membership</td>
<td>Institutions pay annual membership fee to support centralized staff and operations.</td>
</tr>
<tr>
<td>Fee-for-Service</td>
<td>Institutions pay for curation services for each dataset.</td>
</tr>
</tbody>
</table>
Thanks!

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

Twitter #DataCurationNetwork