Fall 2018 RDAP Webinar

Data Curation Services, Together!

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

2018-11-08

DataCurationNetwork.org

Well curated data are more valuable.

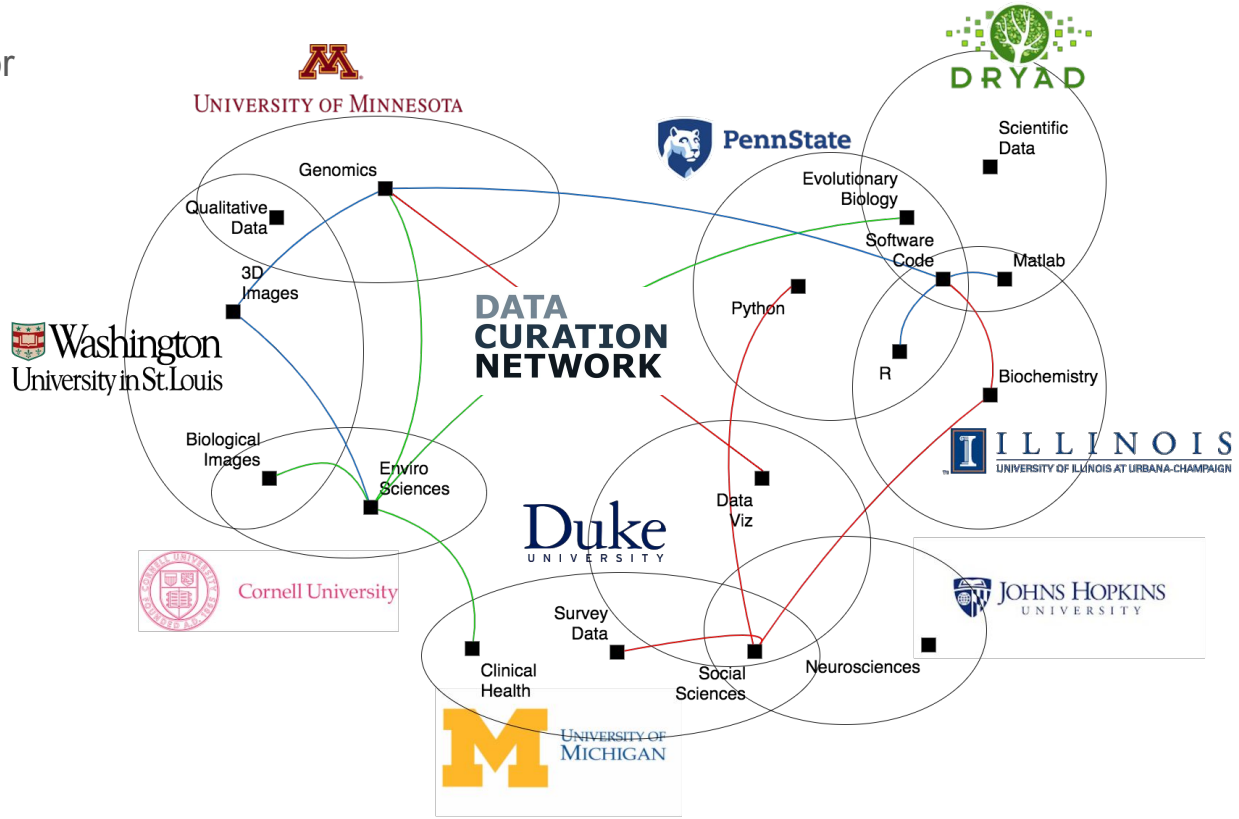
- Easier for fellow scholars and future collaborators to understand
- More likely to be trusted and reused
- The research findings they represent are more likely to be validated and replicated
- More likely to be properly cited

Due to the heterogeneous and multidisciplinary nature of research data generated by the researchers we support, data curation can be a challenge.

The Data Curation Network (DCN) addresses this challenge by collaboratively sharing data curation staff across a network of partner institutions and data repositories.

DATA CURATION NETWORK

- Collaborative staffing model for curating research data across academic and general data repositories
- Funded by Alfred P. Sloan Foundation
- Implementation phase (2018-2021) is piloting the model with eight partner institutions
- Goal is to expand to all users in 2020

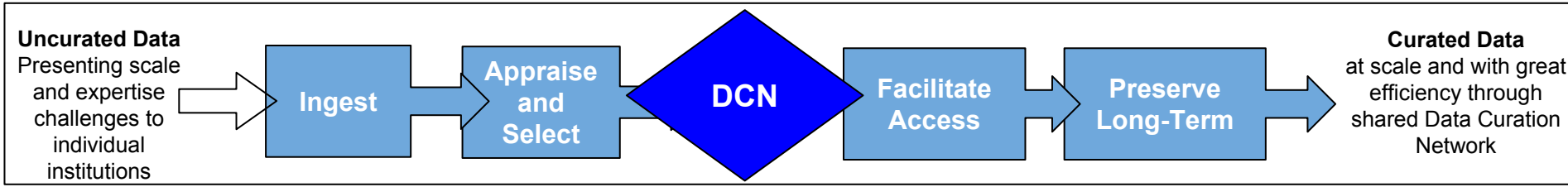


Unique Perspectives and Infrastructure

Institution	Data Repository	Launch Date	Dataset Holdings (as of April 2018)	Technology Platform
Cornell University	eCommons	Fall 2002	120 datasets	DSpace 6.2
Dryad Digital Repository	Dryad Digital Repository	2009	21,293 data packages; 67,907 individual files	DSpace moving to CDL's Merritt
Duke University	Duke Digital Repository (DDR)	Jan 2017	34 data deposits total: 8,120 files (most average 50-100)	Fedora 3.8, and Hyrax
Johns Hopkins University	JHU Data Archive	2011	61 datasets or dataverses; 448 files	Dataverse
Penn State University	ScholarSphere	Fall 2012	966 files	Hydra/Fedora (soon to be Sufia 7)
University of Illinois	Illinois Data Bank	May 2016	77 datasets	Ruby on Rails interface with our preservation repository (Medusa)
University of Michigan	Deep Blue Data	Sep 2016	114 datasets	Samvera/Fedora Hyrax
University of Minnesota	Data Repository for the U of M (DRUM)	Mar 2015	199 in DRUM, ~600 in IR	DSpace 5.5

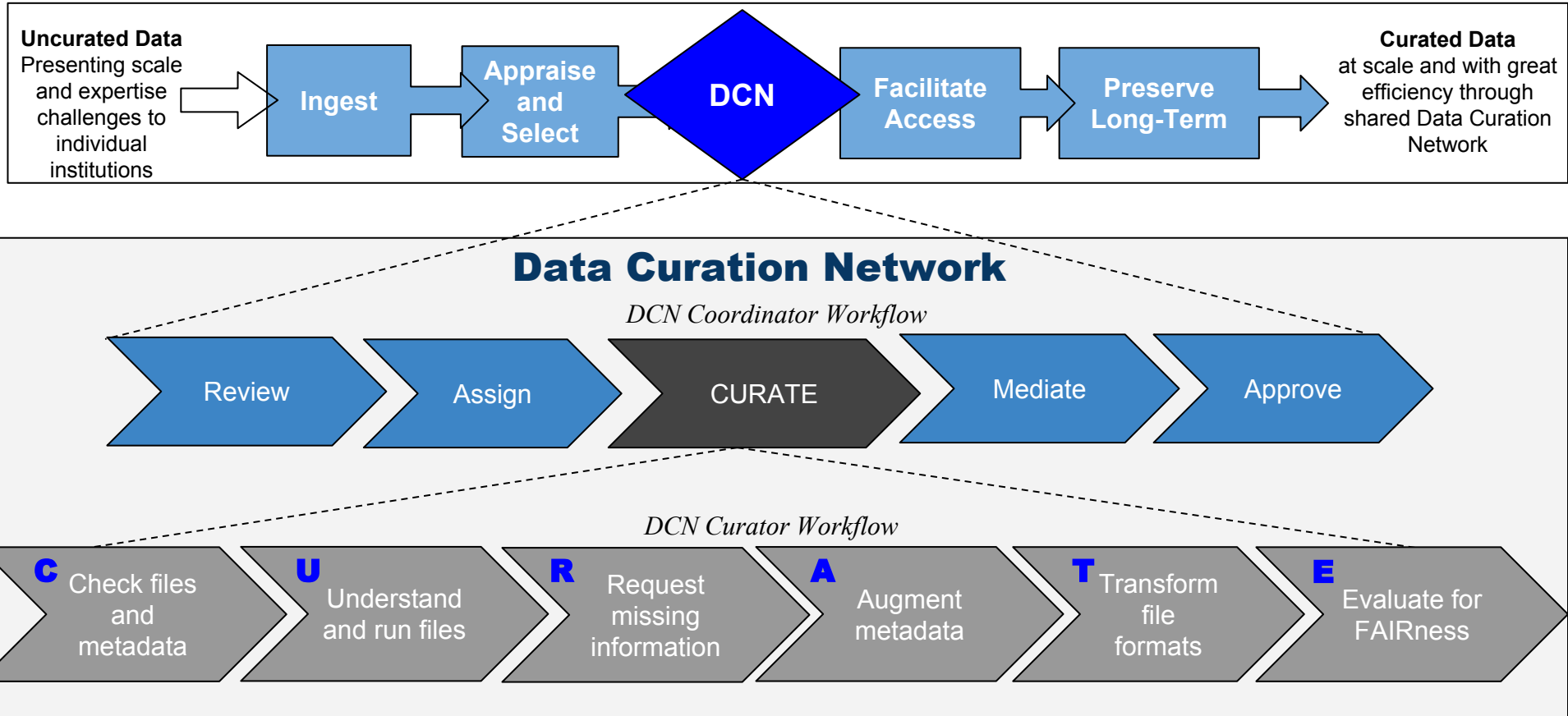
The Data Curation Network (DCN) 3-year implementation phase launched June 2018 with the goal to expand to new members in 2020.

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



CURATE Steps in DCN Workflow

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

- **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** **Evaluate** and rate the overall submission for FAIRness.

DCN Pilot Implementation (2018-2020)

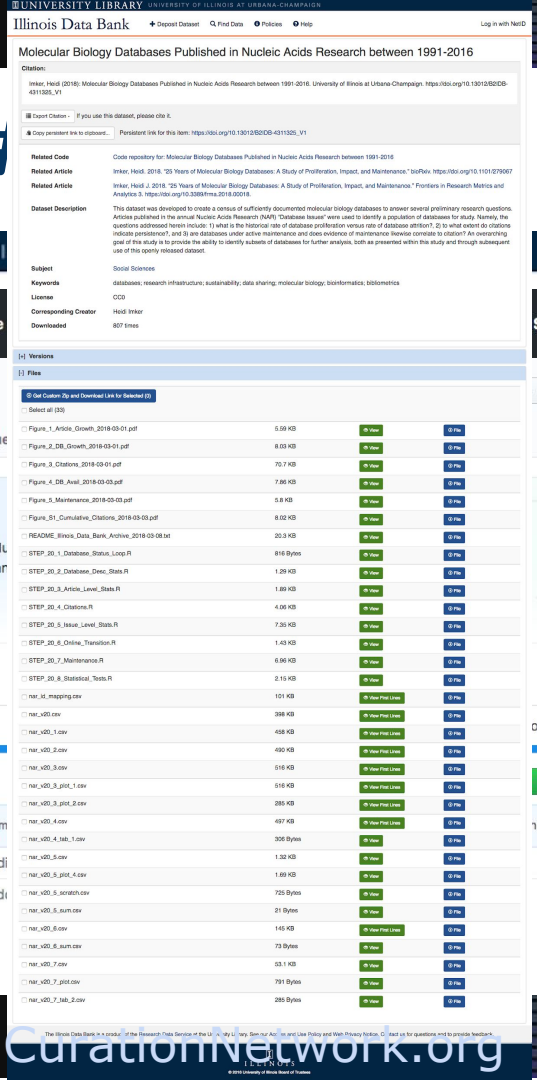
Table A1. Draft checklist of DCN CURATE steps and FAIRness scorecard

CURATE Actions	Curation Checklist
Check data files and read documentation ▪ Review the content of the data files (e.g., open and run the files or code). ▪ Verify all metadata provided by the author and review the available documentation.	☐ Files open as expected ☐ Issues _____ ☐ Code runs as expected ☐ Produces minor ☐ Does not run an many errors ☐ Metadata quality is rich, complete ☐ Metadata has iss ☐ Documentation Type (*ci* Readme / Codebook / D Other: _____ ☐ Missing/None ☐ Needs work
Understand the data (or try to) ▪ Check for quality assurance and usability issues such as missing	*Varies based on file formats and s* *example….*

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
[2] Rubric evaluating the FAIR principles are based on the scoring matrix by Dunning, de Smaele, & Böhmer (2017).

DCN Test Case



DCN Test Case #2

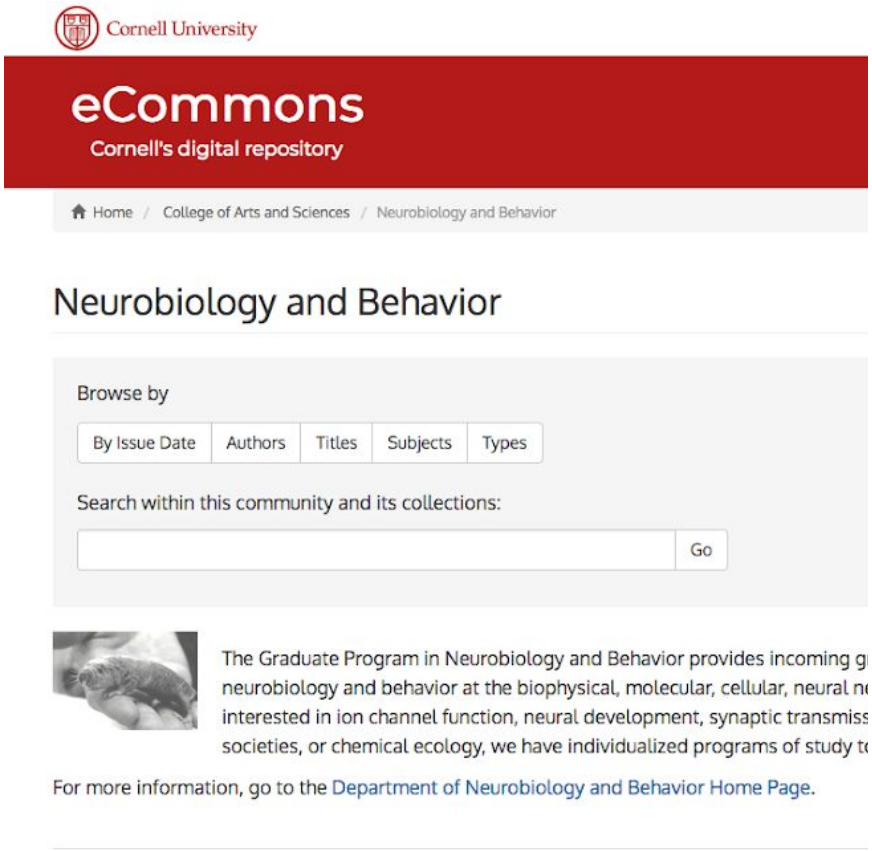
LIBRARIES

View/Download file

File View/Open	Description	Size
meta_data.tar.gz	Metadata (JSON files)	123.5Mb
camera_catalogue_images.tar	Camera CATalogue Images (506,241JPG files)	10.71Gb
snapshot_wisconsin_images.tar	Snapshot Wisconsin Images (497,204 JPG files)	14.31Gb
snapshot_serengeti_images.tar	Snapshot Serengeti Images (6,163,870 JPG files)	123.5Gb
README_image_data.pdf	Description of the Dataset	592.0Kb

Dataset
Programming Software Code
Statistical Computing Software Code

DCN Test Case #3



DCN Test Case #3

Cornell University

eCommons

Cornell's digital repository

Home / College of Arts and Sciences / Neurobiology and Behavior

Neurobiology and Behavior

Sent Via Email

Browse by

By Issue Date | Author | Titles | Subjects | Types

Search within this community and its collections:

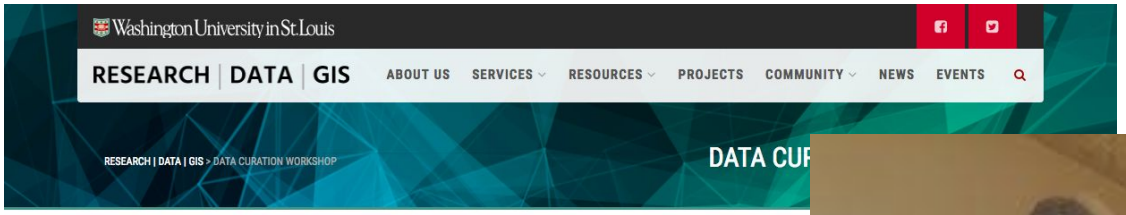
Go

The Graduate Program in Neurobiology and Behavior provides incoming g
neurobiology and behavior at the biophysical, molecular, cellular, neural n
interested in ion channel function, neural development, synaptic transmiss
societies, or chemical ecology, we have individualized programs of study t

For more information, go to the Department of Neurobiology and Behavior Home Page.

DCN Specialized Workshops

INSTITUTE of Museum and Library SERVICES



DATA CURATION WORKSHOP

SLIDES AND HANDOUTS

DATE: DECEMBER 11 & 12, 2017

TWEET: #DCW2017

LOCATION: WASHINGTON UNIVERSITY IN ST. LOUIS, **MCMILLAN HALL,** ST. LOUIS, M

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** mod

- **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** – Evaluate and rate the overall submission for FAIRness.

ATTENDEES WILL COME AWAY WITH:

DCN Research Questions

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)
- Curator incentives?

Are curated data are more valuable?

- Survey researcher satisfaction
- 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 **sustainable service model** where institutional and disciplinary partners join and contribute data curation staff to the Network.*

Sustainability and Expansion (2020-)

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

Thanks!

https://DataCurationNetwork.org

Twitter #DataCurationNetwork