

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

Well curated data are more valuable.

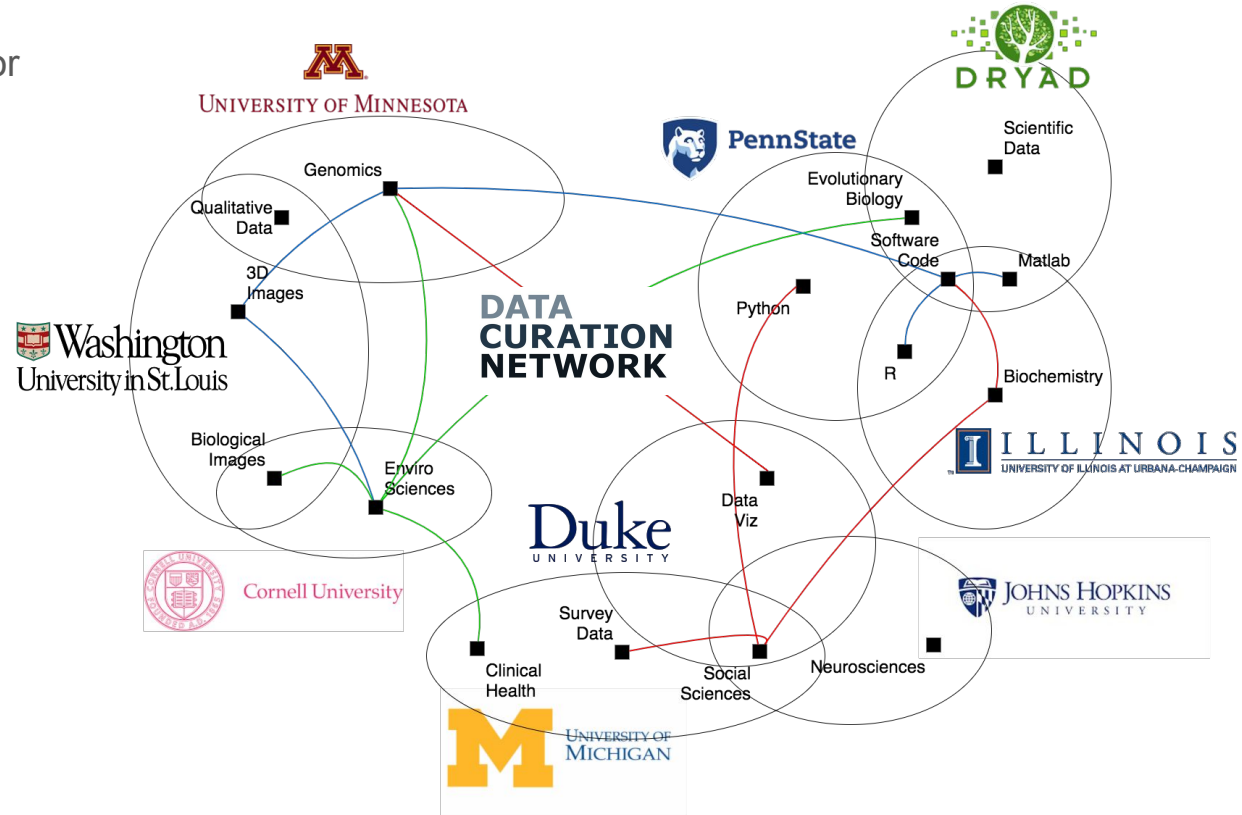
- Easier for fellow scholars and future collaborators to understand
- More likely to be trusted and reused
- The research findings they represent are more likely to be validated and replicated
- More likely to be properly cited

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



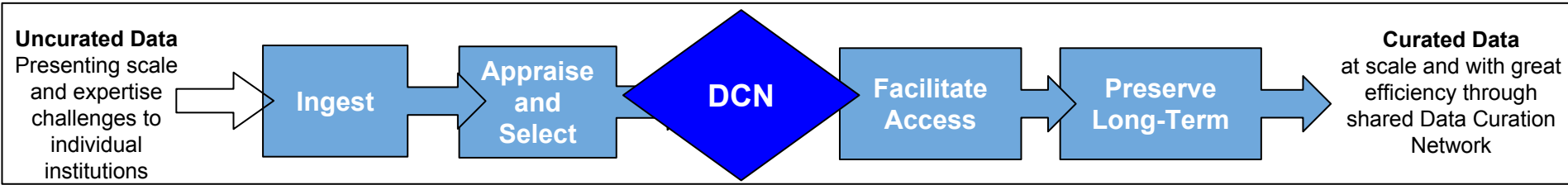
Unique Perspectives and Infrastructure



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

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:

- 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		
<p>Check data files and read documentation</p> <ul style="list-style-type: none"> ▪ 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. 	<ul style="list-style-type: none"> <input type="checkbox"/> Files open as expected <ul style="list-style-type: none"> <input type="checkbox"/> Issues _____ <input type="checkbox"/> Code runs as expected <ul style="list-style-type: none"> <input type="checkbox"/> Produces minor errors <input type="checkbox"/> Does not run and many errors <input type="checkbox"/> Metadata quality is rich, complete <ul style="list-style-type: none"> <input type="checkbox"/> Metadata has issues <input type="checkbox"/> Documentation Type (cite Readme / Codebook / Data) Other: _____ <ul style="list-style-type: none"> <input type="checkbox"/> Missing/None <input type="checkbox"/> Needs work 	<p>Evaluate and rate the overall data record for FAIRness.²</p> <ul style="list-style-type: none"> ▪ Score the dataset and recommend ways to increase the FAIRness of the data and become “DCN approved.” 	<p>Findable -</p> <ul style="list-style-type: none"> <input type="checkbox"/> Metadata exceeds author/ title/ date, Unique PID (DOI, Handle, PURL, etc.). <input type="checkbox"/> Discoverable via web search engines like Google. <p>Accessible -</p> <ul style="list-style-type: none"> <input type="checkbox"/> Retrievable via a standard protocol (e.g., HTTP). <input type="checkbox"/> Free, open (e.g., download link). <p>Interoperable -</p> <ul style="list-style-type: none"> <input type="checkbox"/> Metadata formatted in a standard schema (e.g., Dublin Core). <input type="checkbox"/> Metadata provided in machine-readable format (OAI feed). <p>Reusable -</p> <ul style="list-style-type: none"> <input type="checkbox"/> Data include sufficient metadata about the data characteristics to reuse without the direct assistance of the author. <input type="checkbox"/> <u>Clear indicators of who created, owns, and stewards the data.</u> <input type="checkbox"/> Data are released with clear data usage terms (e.g., a CC License).
<p>Understand the data (or try to)</p> <ul style="list-style-type: none"> ▪ Check for quality assurance and usability issues such as missing 	<p><i>Varies based on file formats and example....</i></p>		

¹ Format Recommendations, <http://guides.library.cornell.edu/ecommonsw/formats>

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

DCN Test Case

UNIVERSITY LIBRARY UNIVERSITY OF ILLINOIS

Illinois Data Commons

Features Business Explore

Molecular Biol

Citation:

Imker, Heidi (2018): Mol
4311325_V1

1heidi / nar_persistent

Code Issues Pull request



Work in progress

24 commits

Branch: master New pull request

1heidi Merge pull request #3 from olendorf/m

.gitignore edi

Figure_1_Funders.TIFF adi

Molecular Data Databases Published in Nucleic Acids Research between 1991-2016

Citation:
Imker, Heidi (2018): Molecular Biology Databases Published in Nucleic Acids Research between 1991-2016. University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-4311325_V1

Related Code
Code repository for "Molecular Biology Databases Published in Nucleic Acids Research between 1991-2016"
Related Article
Imker, Heidi. 2018. "25 Years of Molecular Biology Databases: A Study of Proliferation, Impact, and Maintenance." *bioRxiv*. <https://doi.org/10.1101/279507>
Imker, Heidi. 2018. "25 Years of Molecular Biology Databases: A Study of Proliferation, Impact, and Maintenance." *Frontiers in Research Metrics and Analytics* 3. <https://doi.org/10.3389/fmea.2018.00018>

Dataset Description
This dataset was developed to create a census of sufficiently documented molecular biology databases to answer several preliminary research questions. Articles published in the annual Nucleic Acids Research (NAR) "Database Issue" were used to identify a population of databases for study. Namely, the questions addressed herein include: 1) what is the historical rate of database proliferation versus rate of database attrition? 2) to what extent do citations indicate persistence? and 3) are databases under active maintenance and does evidence of maintenance behavior correlate to citation? An overarching goal of this study is to provide the ability to identify subsets of databases for further analysis, such as presented within this study and through subsequent use of this openly released dataset.

Subject: Social Sciences
Keywords: databases; research infrastructure; sustainability; data sharing; molecular biology; bioinformatics; bibliometrics
License: CC0
Corresponding Creator: Heidi Imker
Downloaded: 807 times

File	Size	View	File
Figure_1_Article_Growth_2018-09-01.pdf	5.69 KB	View	File
Figure_2_DB_Growth_2018-09-01.pdf	8.02 KB	View	File
Figure_3_Citations_2018-09-01.pdf	79.7 KB	View	File
Figure_4_DB_Avail_2018-09-03.pdf	7.86 KB	View	File
Figure_5_Maintenance_2018-09-03.pdf	5.9 KB	View	File
Figure_6_Cumulative_Citations_2018-09-03.pdf	8.02 KB	View	File
README_BioRx_Data_Bank_Active_2018-09-08.txt	20.3 KB	View	File
STEP_00_1_Database_Status_Loop.R	916 Bytes	View	File
STEP_00_2_Database_Devel_Status.R	1.29 KB	View	File
STEP_00_3_Article_Level_Status.R	1.89 KB	View	File
STEP_00_4_Citations.R	4.06 KB	View	File
STEP_00_5_Issue_Level_Status.R	7.35 KB	View	File
STEP_00_6_Change_Transition.R	1.43 KB	View	File
STEP_00_7_Maintenance.R	6.96 KB	View	File
STEP_00_8_Statistical_Tests.R	2.15 KB	View	File
nar_k_mapping.csv	124 KB	View	File
nar_x00.csv	388 KB	View	File
nar_x00_1.csv	458 KB	View	File
nar_x00_2.csv	490 KB	View	File
nar_x00_3.csv	516 KB	View	File
nar_x00_3_p01_1.csv	518 KB	View	File
nar_x00_3_p01_2.csv	285 KB	View	File
nar_x00_4.csv	497 KB	View	File
nar_x00_4_hab_1.csv	306 Bytes	View	File
nar_x00_5.csv	1.32 KB	View	File
nar_x00_5_p01_4.csv	1.89 KB	View	File
nar_x00_5_search.csv	729 Bytes	View	File
nar_x00_5_sum.csv	21 Bytes	View	File
nar_x00_6.csv	145 KB	View	File
nar_x00_6_sum.csv	73 Bytes	View	File
nar_x00_7.csv	53.1 KB	View	File
nar_x00_7_p01.csv	791 Bytes	View	File
nar_x00_7_hab_2.csv	285 Bytes	View	File

Log in with NetID

Sign in or Sign up

0 Fork 1

Dismiss

https://doi.org/10.13012/B2IDB-4311325_V1

Contributors

Clone or download

hit 337ce08 on Aug 27

3 months ago

3 months ago

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

Species	Model Output (%)
Hyena (brown)	99.78 %
Aardwolf	0.24 %
Wild dog	0 %
Hyena (spotted)	0 %
Bat-eared fox	0 %

Species	Model Output (%)
Warthog	99.44 %
Monkeybaboon	0.18 %
Rhino	0.11 %
Bushpig	0.11 %
Elephant	0.06 %

DCN Test Case #3



eCommons

Cornell's digital repository

Home / College of Arts and Sciences / Neurobiology and Behavior

Neurobiology and Behavior

Browse by

By Issue Date Authors Titles Subjects Types

Search within this community and its collections:



The Graduate Program in Neurobiology and Behavior provides incoming graduate students with a strong foundation in neurobiology and behavior at the biophysical, molecular, cellular, neural network, and behavioral levels. If you are interested in ion channel function, neural development, synaptic transmission, or chemical ecology, we have individualized programs of study to meet your needs.

For more information, go to the [Department of Neurobiology and Behavior Home Page](#).

DCN Test Case #3



eCommons
Cornell's digital repository

Home / College of Arts and Sciences / Neurobiology and Behavior

Neurobiology and Behavior

Browse by

By Issue Date Author Titles Subjects Types

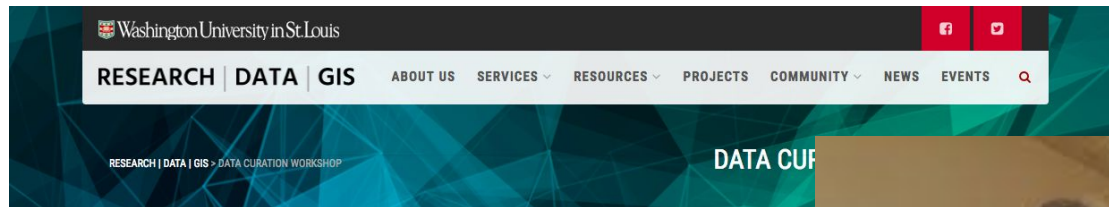
Search within this community and its collections:

 Go

The Graduate Program in Neurobiology and Behavior provides incoming graduate students with a wide range of research opportunities in neurobiology and behavior at the biophysical, molecular, cellular, neural network, or chemical ecology, we have individualized programs of study to

For more information, go to the [Department of Neurobiology and Behavior Home Page](#).

DCN Specialized Workshops



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** – 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
<i>Academic libraries with existing data curation services</i>	Gain access to data curation expertise in more disciplines/formats than locally available
<i>Academic libraries with limited to no resources for data curation services</i>	Are able to provide critical new data curation services when local resources are limited (without needing to hire);
<i>Disciplinary- and general-subject data repositories</i>	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