NDS/MBDH 2018

Launching the Data Curation Network

Lisa Johnston University of Minnesota
Jake Carlson University of Michigan
Cynthia Hudson-Vitale Penn State 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

7-11-2018

Data Curation Network
http://DataCurationNetwork.org

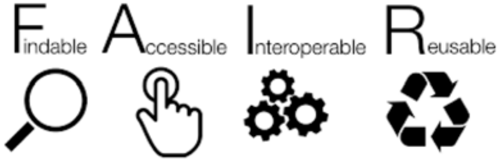
Rise of the Data Sharing Culture

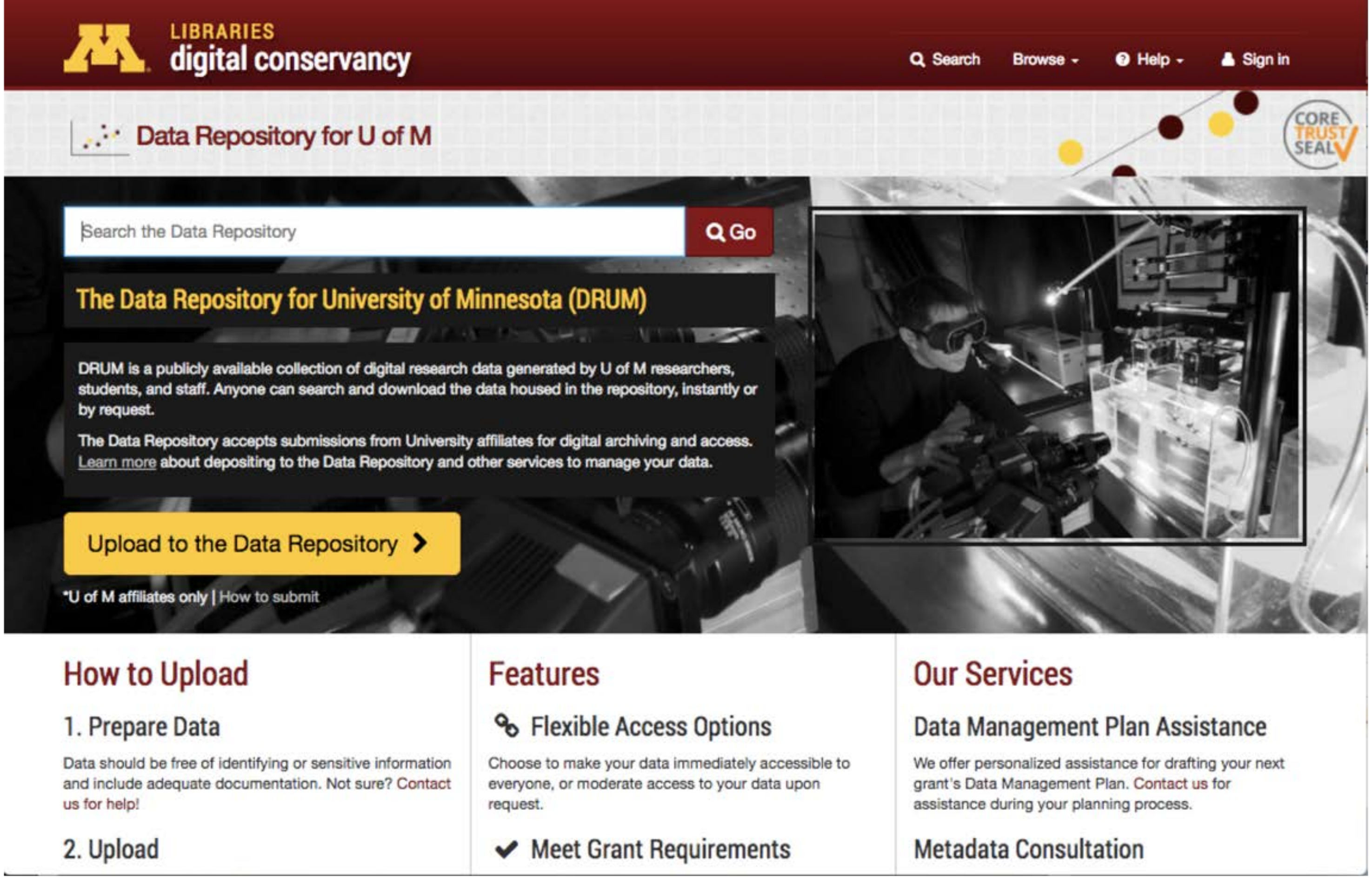
Researchers are increasingly required/incentivised to share data

- Funder data sharing mandates
- Journal data sharing policies
- Disciplinary practices → emphasis on transparency and reproducibility

But! It’s not enough to just share the files, **well-curated data** are more valuable!

Goal of data curation ⇒ Ingest and maintain (trusted digital repositories) in ways that make it findable, accessible, interoperable and reusable.





Data Repository for U of M

The Data Repository for University of Minnesota (DRUM)

DRUM is a publicly available collection of digital research data generated by U of M researchers, students, and staff. Anyone can search and download the data housed in the repository, instantly or by request.

The Data Repository accepts submissions from University affiliates for digital archiving and access. Learn more about depositing to the Data Repository and other services to manage your data.

Upload to the Data Repository

*U of M affiliates only | How to submit

How to Upload

1. Prepare Data

Data should be free of identifying or sensitive information and include adequate documentation. Not sure? Contact us for help!

2. Upload

Features

Flexible Access Options

Choose to make your data immediately accessible to everyone, or moderate access to your data upon request.

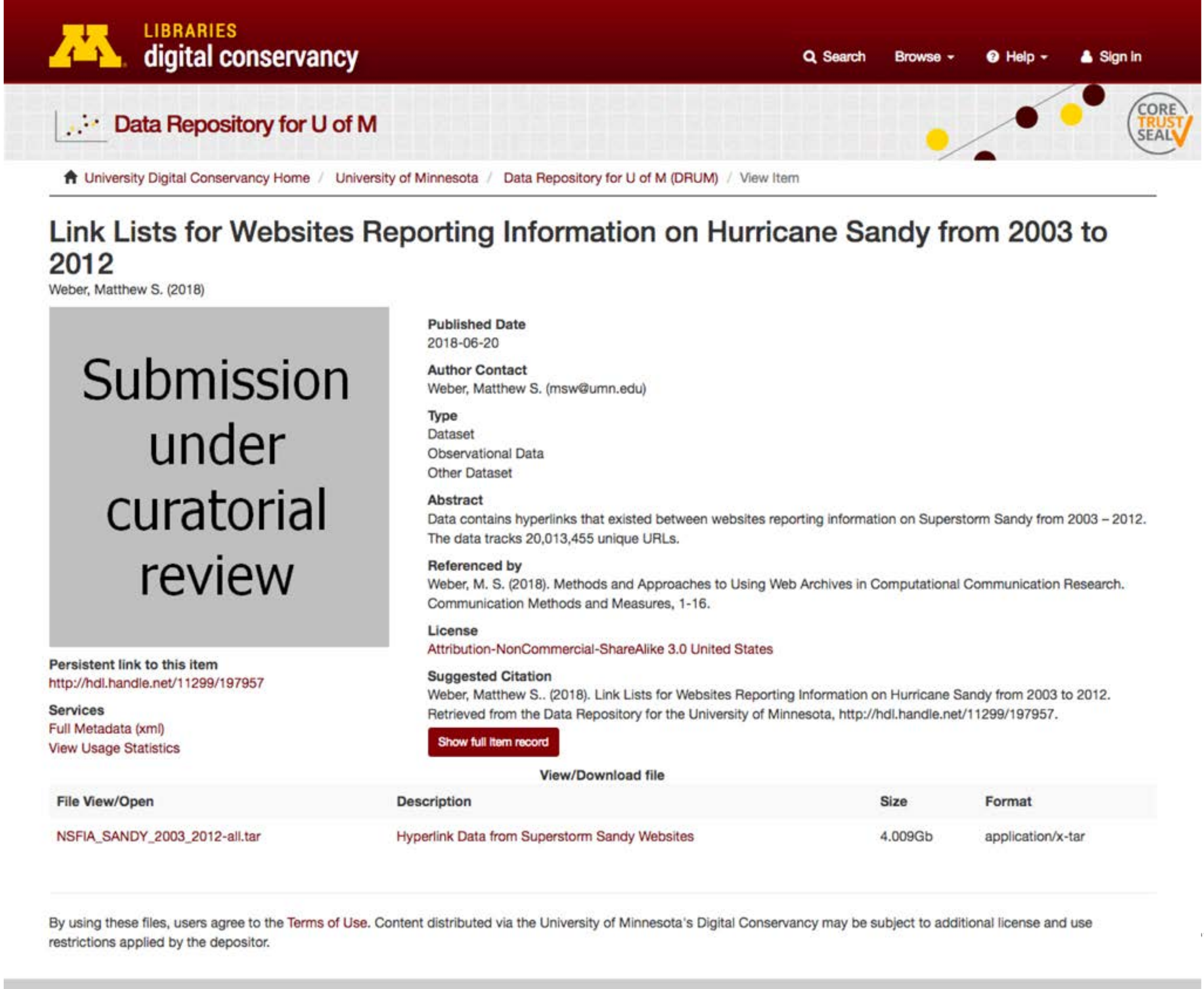
✓ Meet Grant Requirements

Our Services

Data Management Plan Assistance

We offer personalized assistance for drafting your next grant's Data Management Plan. Contact us for assistance during your planning process.

Metadata Consultation



Data Repository for U of M

University Digital Conservancy Home / University of Minnesota / Data Repository for U of M (DRUM) / View Item

Link Lists for Websites Reporting Information on Hurricane Sandy from 2003 to 2012

Weber, Matthew S. (2018)

Submission under curatorial review

Published Date
2018-06-20

Author Contact
Weber, Matthew S. (msw@umn.edu)

Type
Dataset
Observational Data
Other Dataset

Abstract
Data contains hyperlinks that existed between websites reporting information on Superstorm Sandy from 2003 – 2012. The data tracks 20,013,455 unique URLs.

Referenced by
Weber, M. S. (2018). Methods and Approaches to Using Web Archives in Computational Communication Research. Communication Methods and Measures, 1-16.

License
Attribution-NonCommercial-ShareAlike 3.0 United States

Suggested Citation
Weber, Matthew S.. (2018). Link Lists for Websites Reporting Information on Hurricane Sandy from 2003 to 2012. Retrieved from the Data Repository for the University of Minnesota, http://hdl.handle.net/11299/197957.

Show full item record

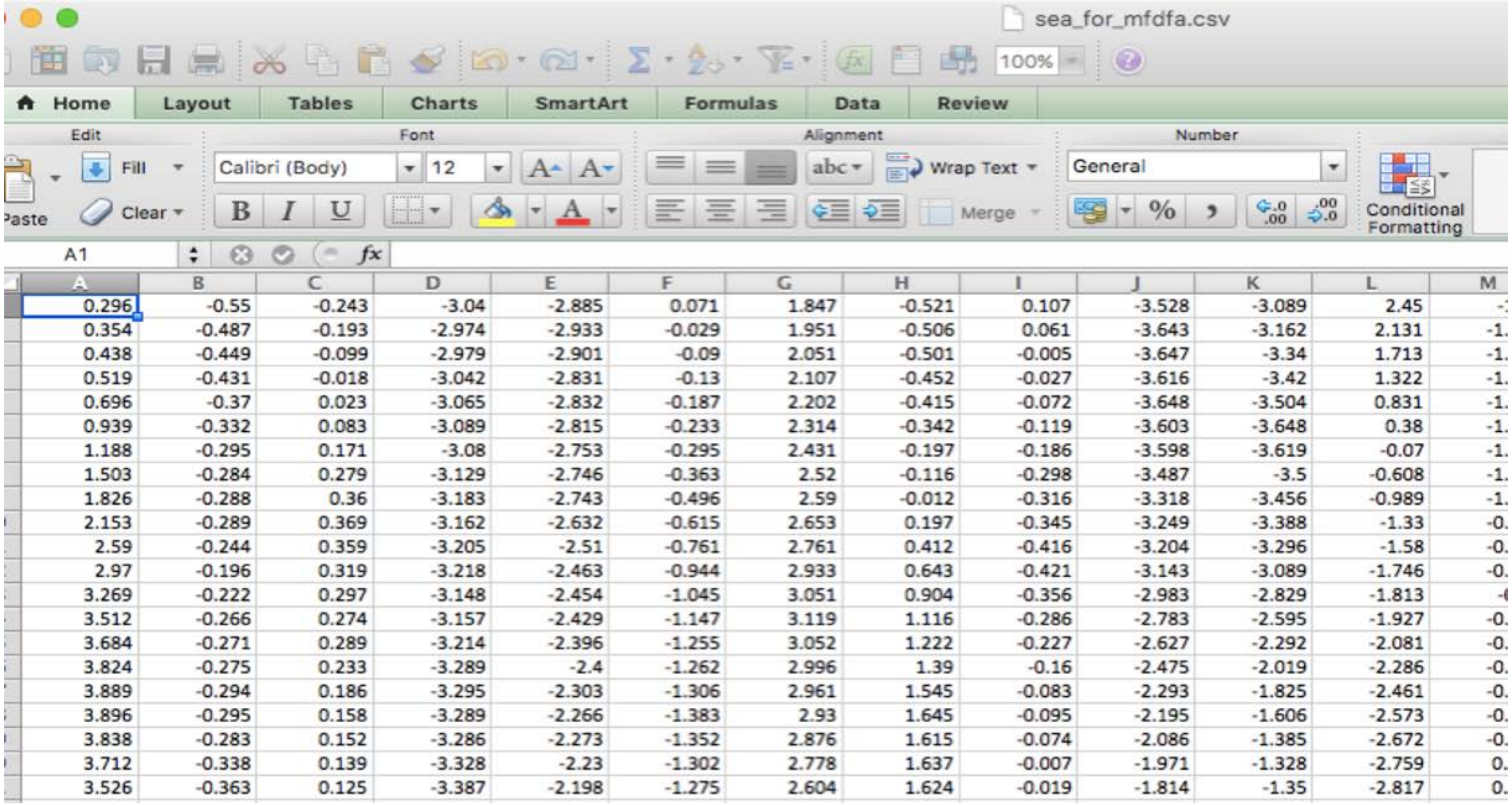
Persistent link to this item
http://hdl.handle.net/11299/197957

Services
Full Metadata (xml)
View Usage Statistics

View/Download file

File View/Open	Description	Size	Format
NSFIA_SANDY_2003_2012-all.tar	Hyperlink Data from Superstorm Sandy Websites	4.009Gb	application/x-tar

By using these files, users agree to the Terms of Use. Content distributed via the University of Minnesota's Digital Conservancy may be subject to additional license and use restrictions applied by the depositor.



sea_for_mfdfa.csv

Zoom: 100%

Font: Calibri (Body), 12

Number: General

A1

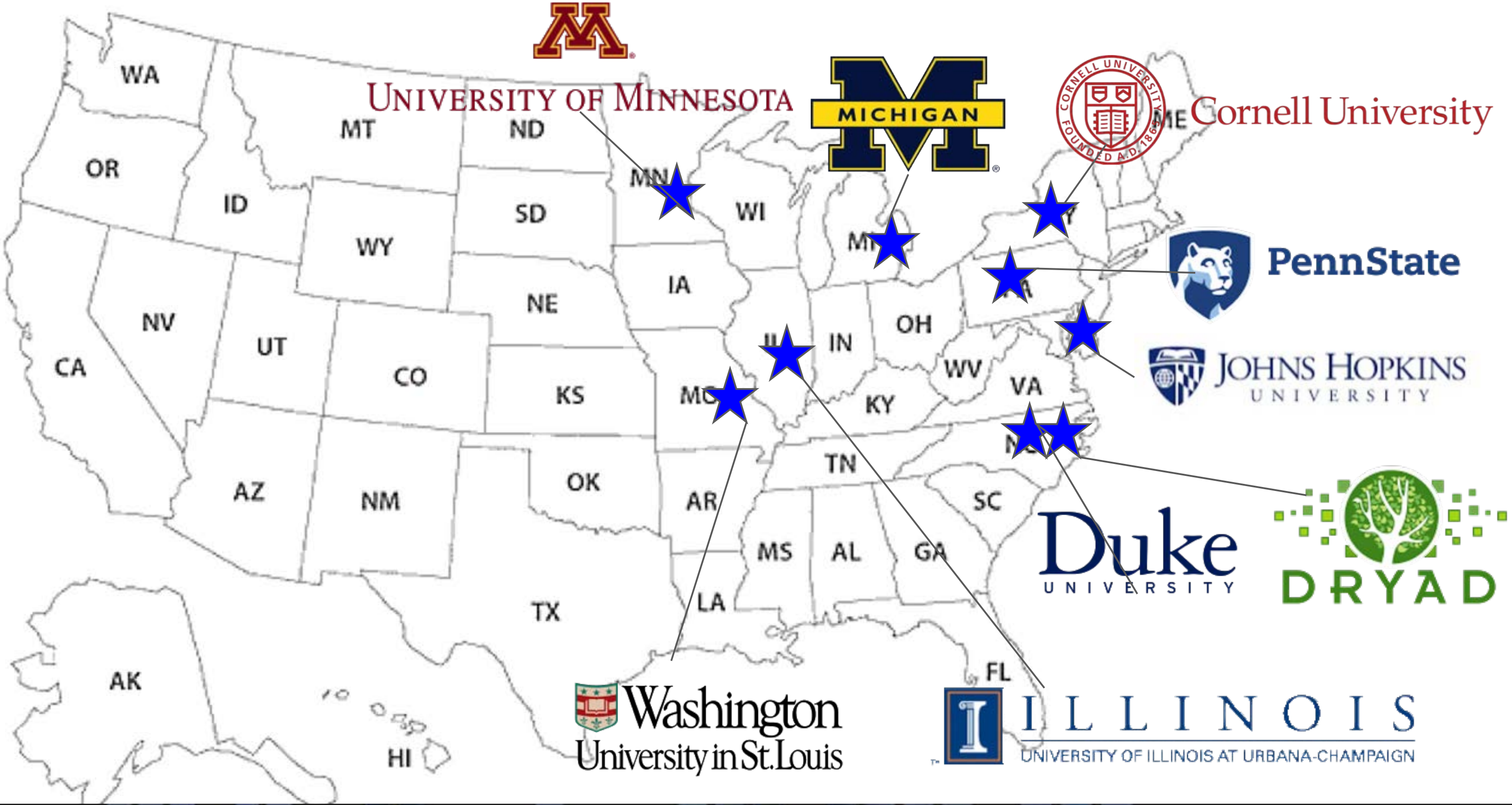
A	B	C	D	E	F	G	H	I	J	K	L	M
0.296	-0.55	-0.243	-3.04	-2.885	0.071	1.847	-0.521	0.107	-3.528	-3.089	2.45	-
0.354	-0.487	-0.193	-2.974	-2.933	-0.029	1.951	-0.506	0.061	-3.643	-3.162	2.131	-1.
0.438	-0.449	-0.099	-2.979	-2.901	-0.09	2.051	-0.501	-0.005	-3.647	-3.34	1.713	-1.
0.519	-0.431	-0.018	-3.042	-2.831	-0.13	2.107	-0.452	-0.027	-3.616	-3.42	1.322	-1.
0.696	-0.37	0.023	-3.065	-2.832	-0.187	2.202	-0.415	-0.072	-3.648	-3.504	0.831	-1.
0.939	-0.332	0.083	-3.089	-2.815	-0.233	2.314	-0.342	-0.119	-3.603	-3.648	0.38	-1.
1.188	-0.295	0.171	-3.08	-2.753	-0.295	2.431	-0.197	-0.186	-3.598	-3.619	-0.07	-1.
1.503	-0.284	0.279	-3.129	-2.746	-0.363	2.52	-0.116	-0.298	-3.487	-3.5	-0.608	-1.
1.826	-0.288	0.36	-3.183	-2.743	-0.496	2.59	-0.012	-0.316	-3.318	-3.456	-0.989	-1.
2.153	-0.289	0.369	-3.162	-2.632	-0.615	2.653	0.197	-0.345	-3.249	-3.388	-1.33	-0.
2.59	-0.244	0.359	-3.205	-2.51	-0.761	2.761	0.412	-0.416	-3.204	-3.296	-1.58	-0.
2.97	-0.196	0.319	-3.218	-2.463	-0.944	2.933	0.643	-0.421	-3.143	-3.089	-1.746	-0.
3.269	-0.222	0.297	-3.148	-2.454	-1.045	3.051	0.904	-0.356	-2.983	-2.829	-1.813	-
3.512	-0.266	0.274	-3.157	-2.429	-1.147	3.119	1.116	-0.286	-2.783	-2.595	-1.927	-0.
3.684	-0.271	0.289	-3.214	-2.396	-1.255	3.052	1.222	-0.227	-2.627	-2.292	-2.081	-0.
3.824	-0.275	0.233	-3.289	-2.4	-1.262	2.996	1.39	-0.16	-2.475	-2.019	-2.286	-0.
3.889	-0.294	0.186	-3.295	-2.303	-1.306	2.961	1.545	-0.083	-2.293	-1.825	-2.461	-0.
3.896	-0.295	0.158	-3.289	-2.266	-1.383	2.93	1.645	-0.095	-2.195	-1.606	-2.573	-0.
3.838	-0.283	0.152	-3.286	-2.273	-1.352	2.876	1.615	-0.074	-2.086	-1.385	-2.672	-0.
3.712	-0.338	0.139	-3.328	-2.23	-1.302	2.778	1.637	-0.007	-1.971	-1.328	-2.759	0.
3.526	-0.363	0.125	-3.387	-2.198	-1.275	2.604	1.624	-0.019	-1.814	-1.35	-2.817	0.

Challenges for IR 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 institutions specialize in curation skills? Represent our academic expertise?

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

The Data Curation Network (DCN) 3-year implementation phase launched June 2018 with funding from the Alfred P Sloan Foundation.



Data Curation Network

The Data Curation Network (DCN) serves as the "human layer" in the data repository stack that provides specialized dataset curation and professional development training for an emerging data curator community.

Katie Wilson
Scientific Data Curator

Valerie Collins
DRUM coordinator

Shanda Hunt
Public Health Data Curator

UNIVERSITY OF MINNESOTA

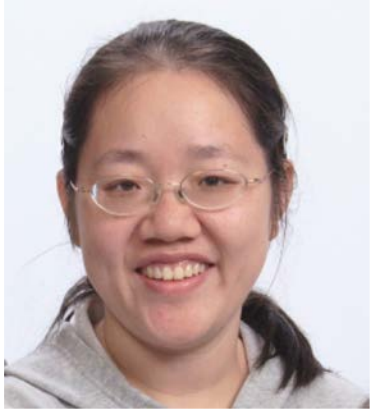
Ben Wiggins
Digital Arts and Humanities
Curator

Melinda Kernik
GIS/Spatial Data Curator

Alicia Hofelich Mohr
College of Liberal Arts
Data Management
Specialist

Mara Blake
Data Services Manager
Johns Hopkins University

JOHNS HOPKINS UNIVERSITY



Chen Chui
Data Management Consultant
Johns Hopkins University

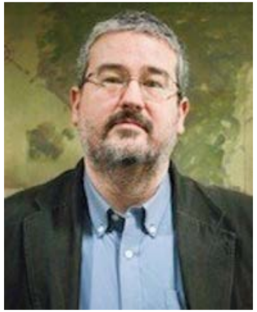
Dave Fearon
Data Management Consultant
Johns Hopkins University

Jake Carlson
Research Data Services Manager
University of Michigan

Susan Borda
Data Workflows
Specialist

Rachel Woodbrook
Data Curation Librarian

Robert Olendorf, PhD
Science Data Librarian
Pennsylvania State University



John Russell
Associate Director Center
for Humanities and Information

Cynthia Hudson-Vitale
Head, Digital Scholarship and Data Services
Pennsylvania State University Libraries

Nathan Piekielek
Geospatial Services Librarian

Joel Herndon
Head of Data and Visualization Services
Duke University Libraries

Duke
UNIVERSITY

Jen Darragh
RDM Consultant
Duke Libraries

Tim McGeary
Associate University Librarian for Digital Strategies and Technology
Duke University Libraries

Sophia Lafferty-Hess
RDM Consultant
Duke Libraries

Elizabeth Hull
Operations Manager
Dryad Digital Repository

DRYAD

Erin Clary
Senior Curator

Debra Fagan
Curation and Technical Specialist

Wendy Kozlowski
Data Curation Specialist
Cornell University Library

Cornell University

Erica Johns
Research Data and Environmental Sciences Librarian

Henrik Spoon
Physics, Astronomy and Math Librarian

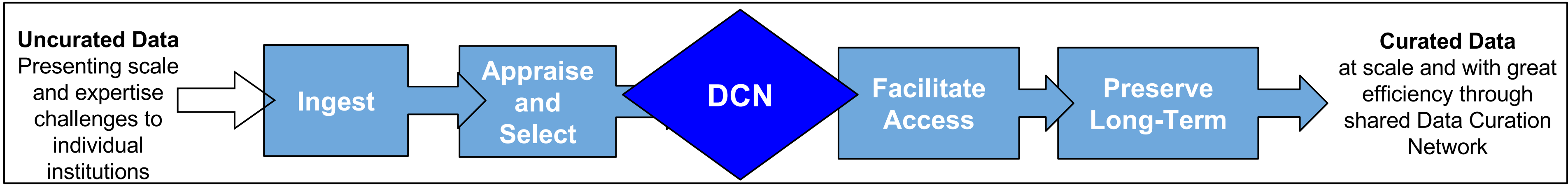
ILLINOIS

Heidi Imker
Director, Research Data Service
University of Illinois at Urbana-Champaign

Hoa Luong
Research Data Specialist
University of Illinois at Urbana-Champaign

Ashley Hetrick
Research Data Specialist
University of Illinois at Urbana-Champaign

DCN Workflow



Uncurated Data
Presenting scale and expertise challenges to individual institutions

Ingest

Appraise and Select

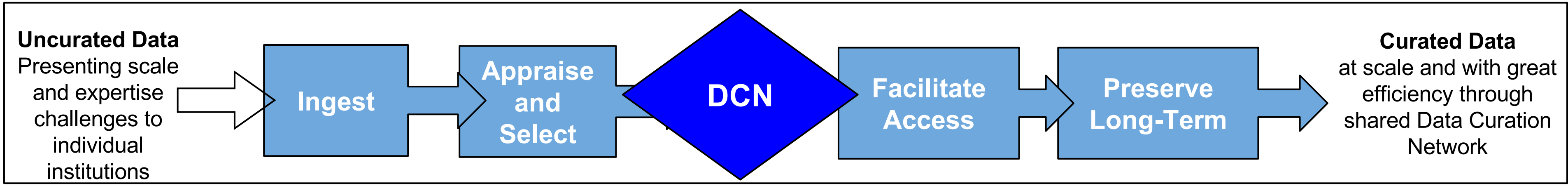
DCN

Facilitate Access

Preserve Long-Term

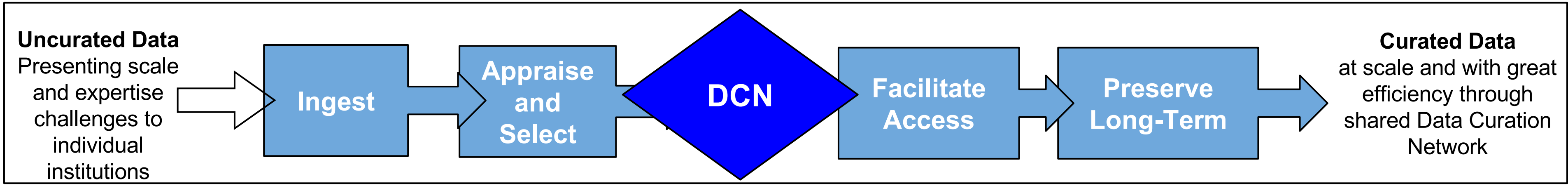
Curated Data
at scale and with great efficiency through shared Data Curation Network

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

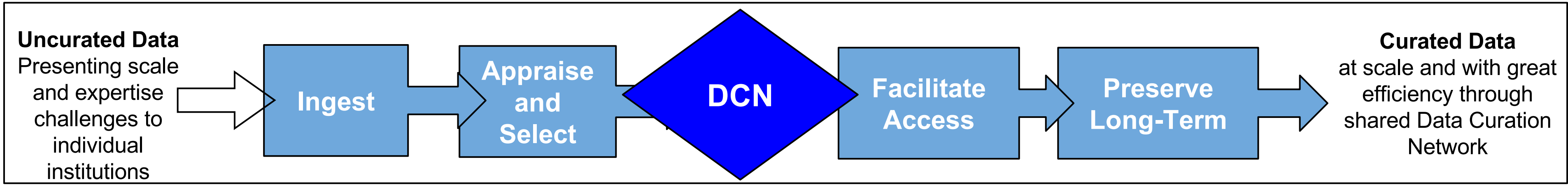
- Researchers deposit like normal

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

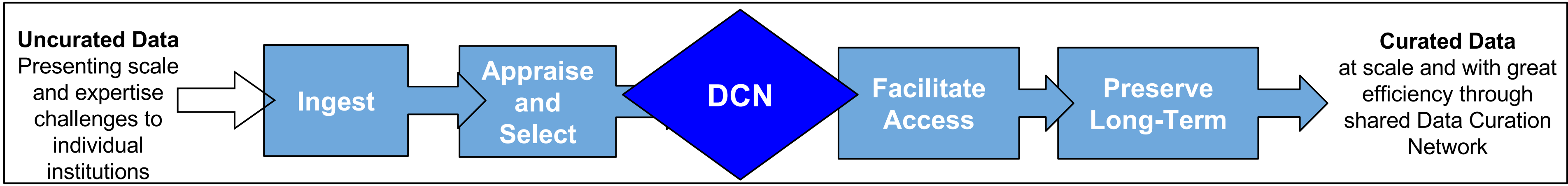
- Researchers deposit like normal
- DCN functions as a microservice layer (the “human layer in your repository stack”)

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

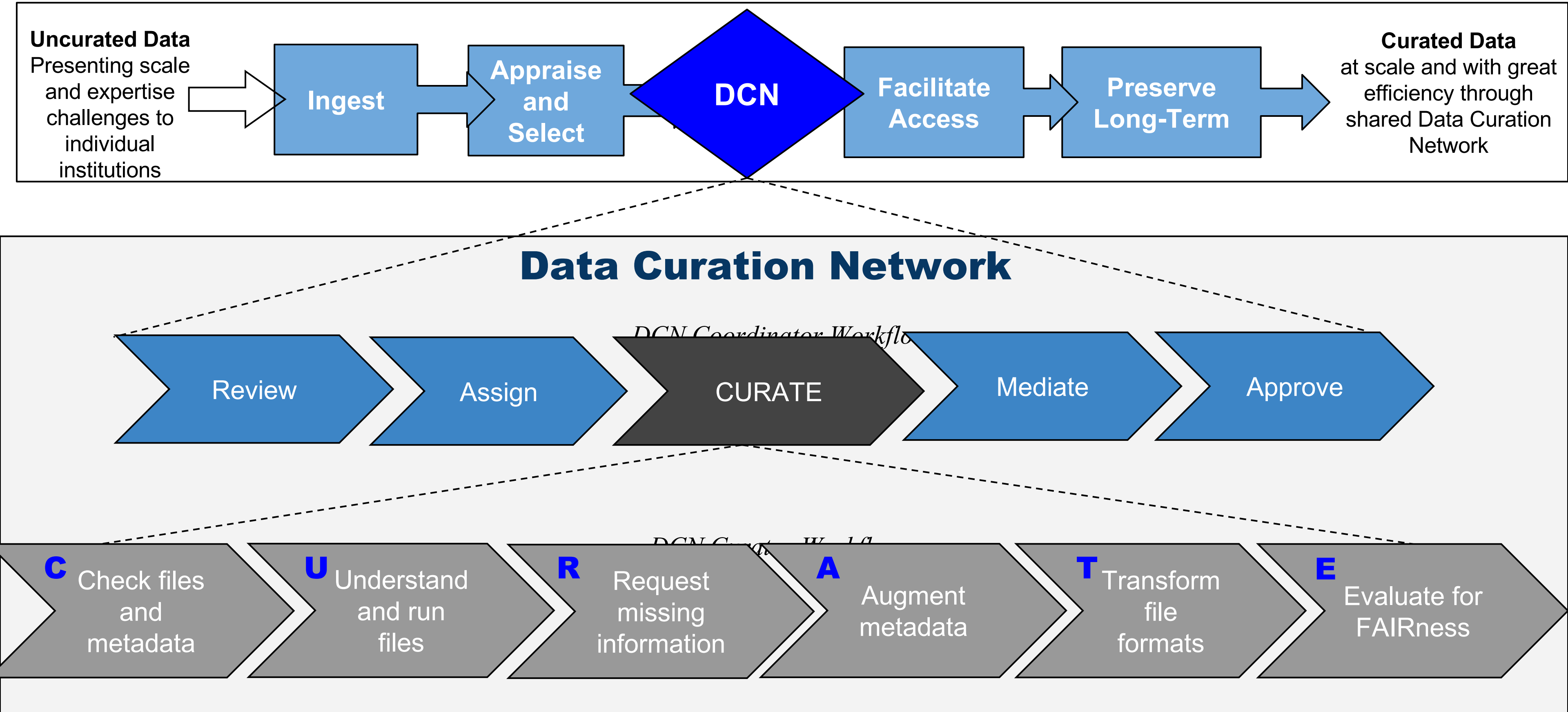
- Researchers deposit like normal
- DCN functions as a microservice layer (the “human layer in your repository stack”)
- Local institution maintain full responsibility for all technical functionality (eg. storage) and authority for local decision-making (what to ingest, how long to retain, etc.)

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

- 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, Dataverse, 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 Workflo

Review

Assign

CURATE

Mediate

Approve

DCN Curat Workfl

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 CURATE Steps

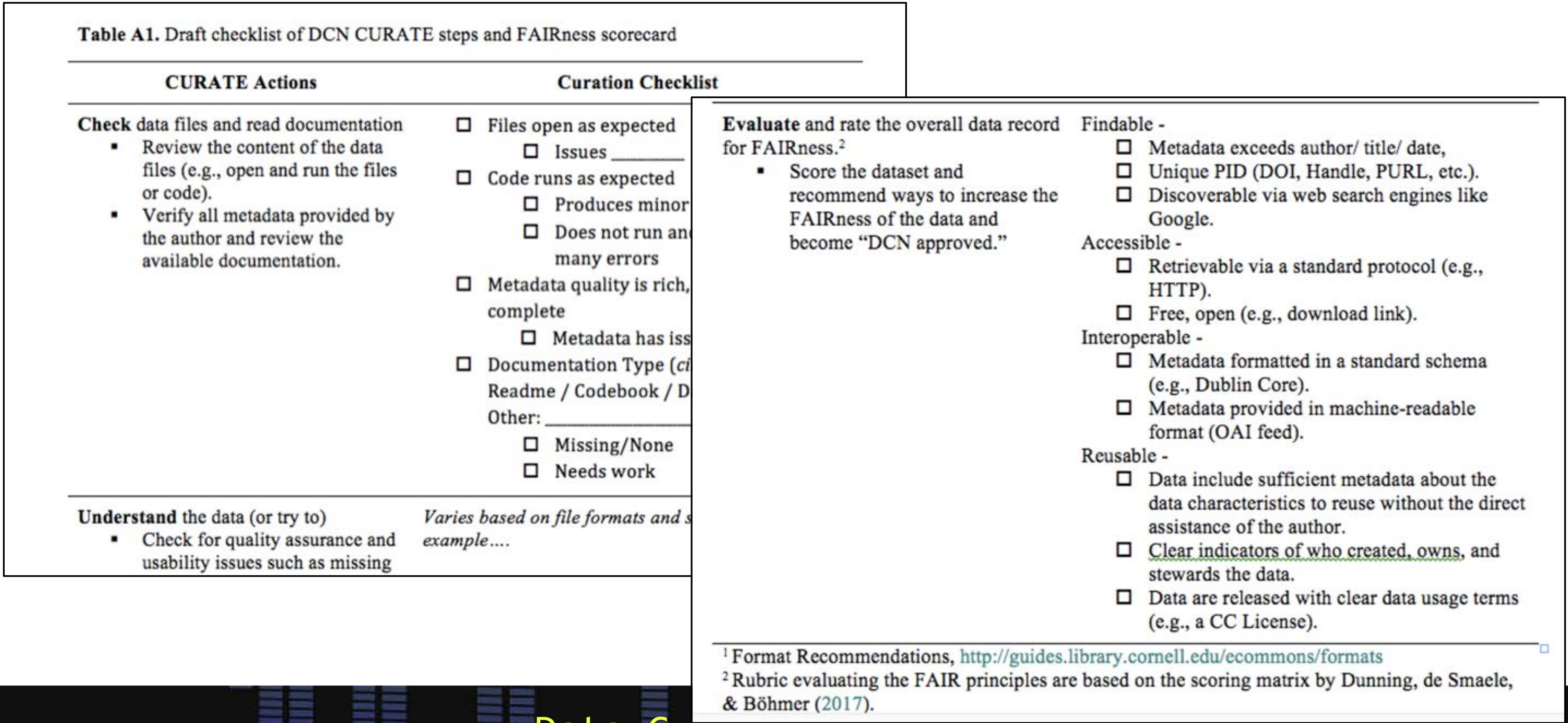


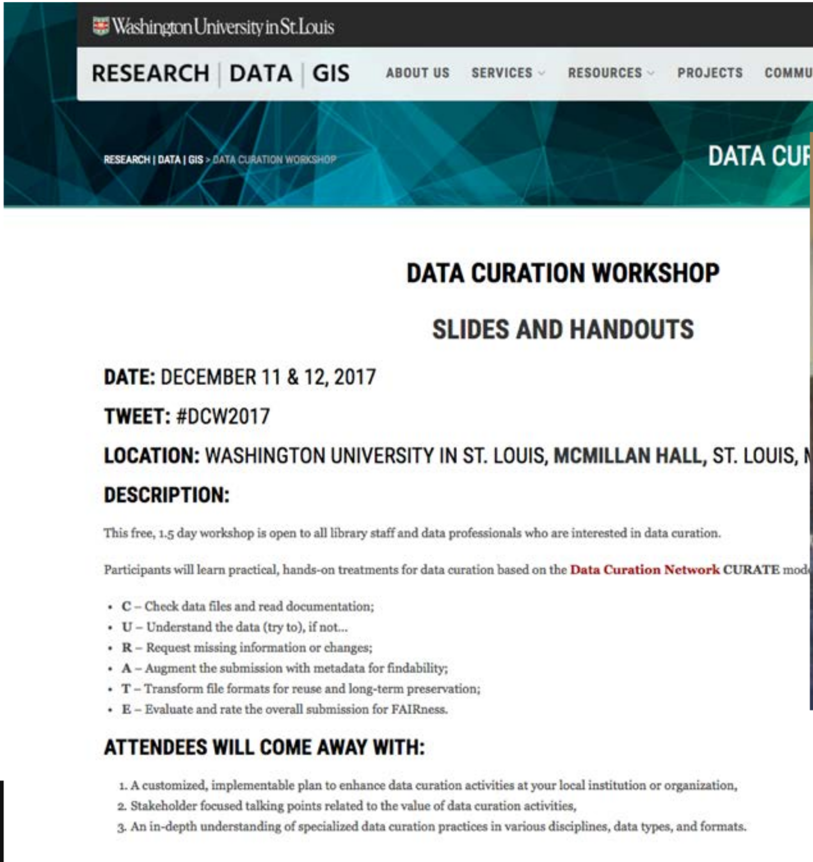
Table A1. Draft checklist of DCN CURATE steps and FAIRness scorecard

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).

Specialized Curation Training (2018-2020)



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:

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.

DCN Implementation (2018-2021)

Alfred P. Sloan FOUNDATION

Assessment Plan (two-prong)

Is a networked approach to curating research data more efficient?

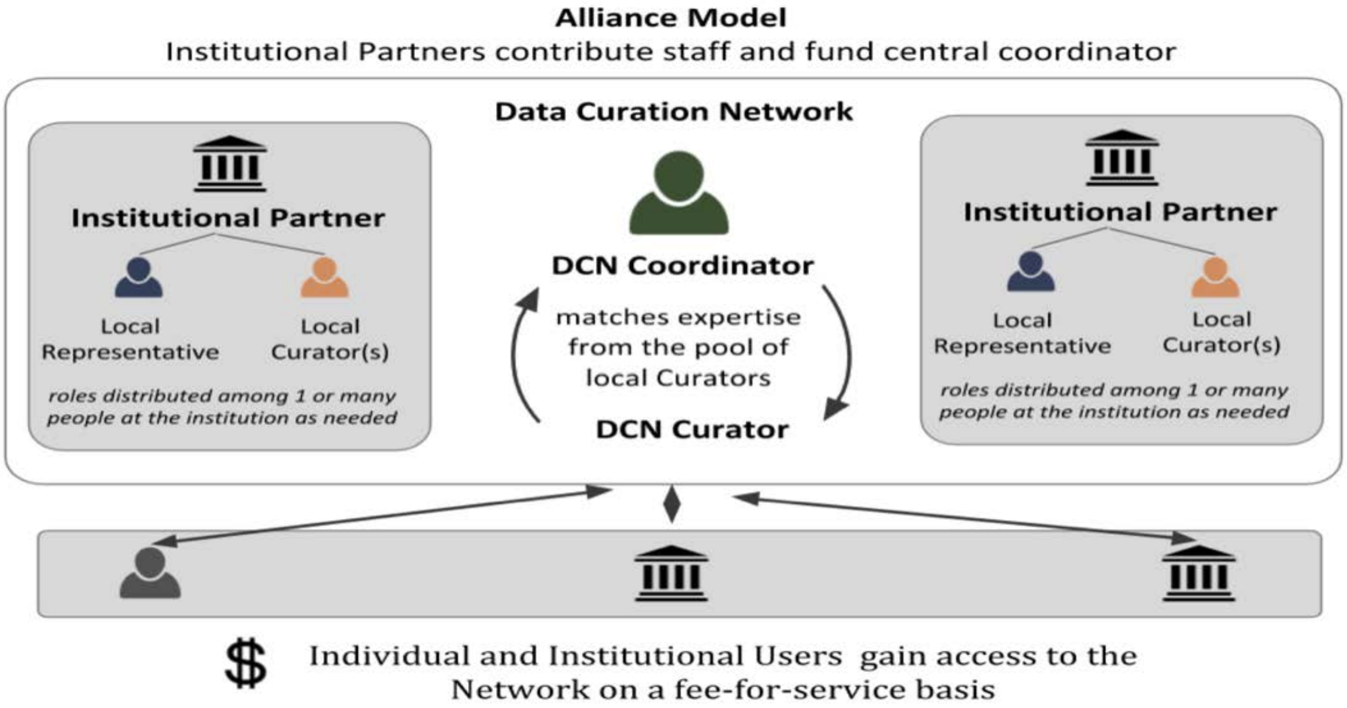
- Number of datasets
- Frequency (high-volume time periods, etc.)
- Variety (data file formats; range of disciplines)
- Efficiency (time, costs)

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

Institutional Partner

Local Representative

Local Curator(s)

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-)

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

6-year Roadmap toward Sustainability

Transition from planning phase to sustaining phase

	Year 0	Year 1 Year 2	Year 3	Year 4	Year 5 Year 6
Support	Sloan Grant	Grant Funded (Y1-Y2) transition to partnership model (Y3)		Curation-as-service (Y4-6)	
Timing	2016-17	2018-19	2020-21		2022-2023
Phase	Planning	Implementation	Transition		Sustaining
Partners	6 academic institutions	8 academic institutions and 2 disciplinary partners		Recruit new partners as use and demand dictate	

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.

Thanks!

https://DataCurationNetwork.org

Twitter #DataCurationNetwork