DATA CURATION NETWORK

Leveraging the DCN

St. Louis Regional Library Network

03/03/2021

Presented by

Jennifer Moore
Head of Data Services
Washington University in St. Louis

Why data curation?

Most data is less than good

(Initially)

We generate a lot of data in our research. Many societies, funders, and publishers encourage data sharing, but...

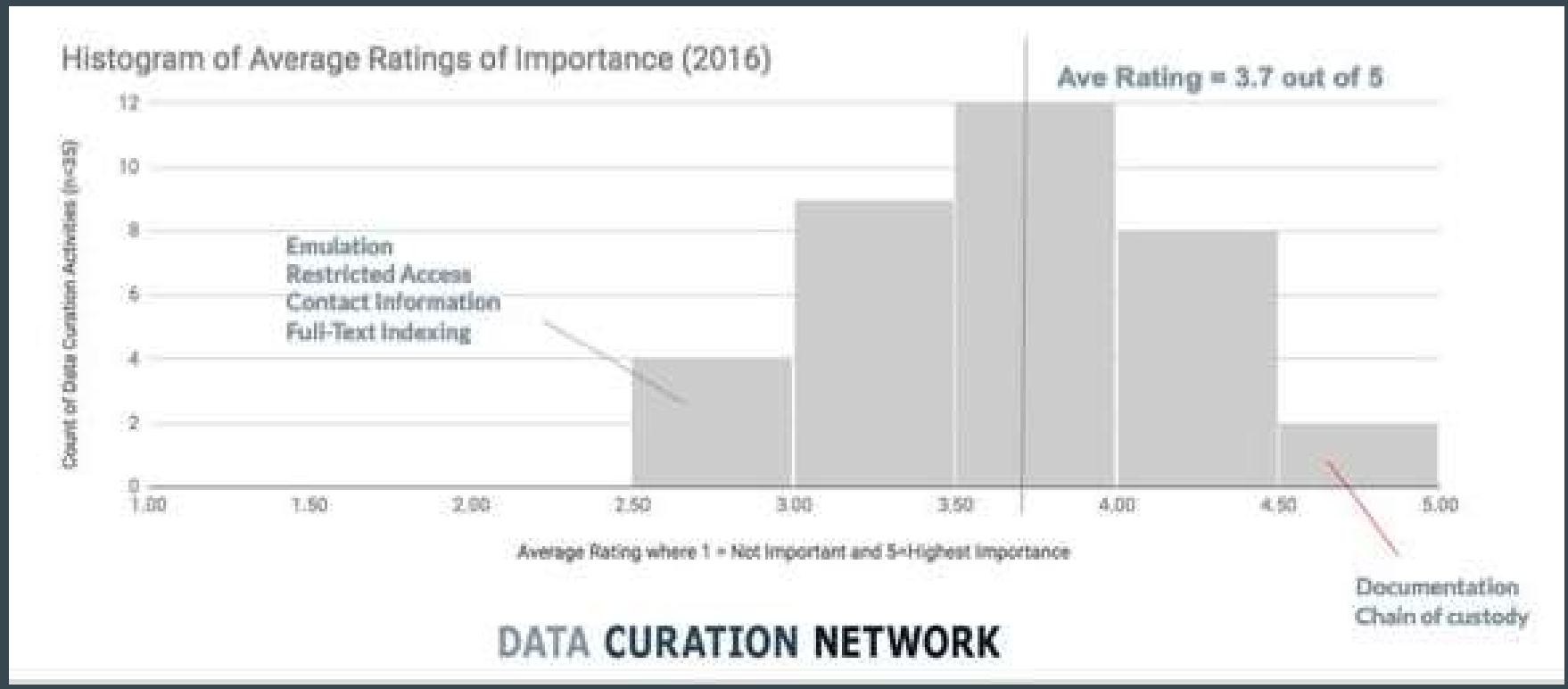
- Data are messy (lack context!)
- Digital file formats are constantly at risk
- Most data never leave their author's laptop ⇒ benign neglect

47%

Data sets with no documentation*

*2017 study of 175 data sets across 6 academic repositories in report: “Data Curation Network: A Cross-Institutional Staffing Model for Curating Research Data” (2017), http://hdl.handle.net/11299/188654.

Focus Group Results (n=91 across 6 institutions)



Data Curation Network. (2018). How Important is Data Curation? Gaps and Opportunities for Academic Libraries. *Journal of Librarianship and Scholarly Communication, 6* (1), eP2198. http://doi.org/10.7710/2162-3309.2198.

Focus Group Results (n=91 across 6 institutions)

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)

Data Curation Network. (2018). How Important is Data Curation? Gaps and Opportunities for Academic Libraries. *Journal of Librarianship and Scholarly Communication, 6* (1), eP2198. http://doi.org/10.7710/2162-3309.2198.

Focus Group Results (n=91 across 6 institutions)

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)

Does this happen for your data?

- Persistent Identifier (37% happens)
- Software Registry (41% happens)
- File Audit (16% happens)
- Contextualization (38% happens)
- Code Review (38% happens)

Data Curation Network. (2018). How Important is Data Curation? Gaps and Opportunities for Academic Libraries. *Journal of Librarianship and Scholarly Communication, 6* (1), eP2198. http://doi.org/10.7710/2162-3309.2198.

Focus Group Results (n=91 across 6 institutions)

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)

If so, are you satisfied with the result?

- 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)

Data Curation Network. (2018). How Important is Data Curation? Gaps and Opportunities for Academic Libraries. *Journal of Librarianship and Scholarly Communication, 6* (1), eP2198. http://doi.org/10.7710/2162-3309.2198.

Data Curation

The encompassing work and actions taken in order to provide meaningful and enduring access to data.

- ✓ Finding and adding missing files and documentation
- ✓ Screening for privacy disclosure risk
- ✓ Detecting and fixing code and other quality assurance issues
- ✓ Transforming file formats for long term access
- ✓ Arranging and describing files
- ✓ Reviewing and augmenting metadata

Impact of Data Curation

Public

1

Everyone benefits from accessible, transparent, and reproducible research.

Researchers

2

Scholars and researchers trust professionally-curated data that are findable accessible, interoperable, and reusable (FAIR).

Repositories

3

Data repositories provide technical access and preservation services to publish high-quality data sets.

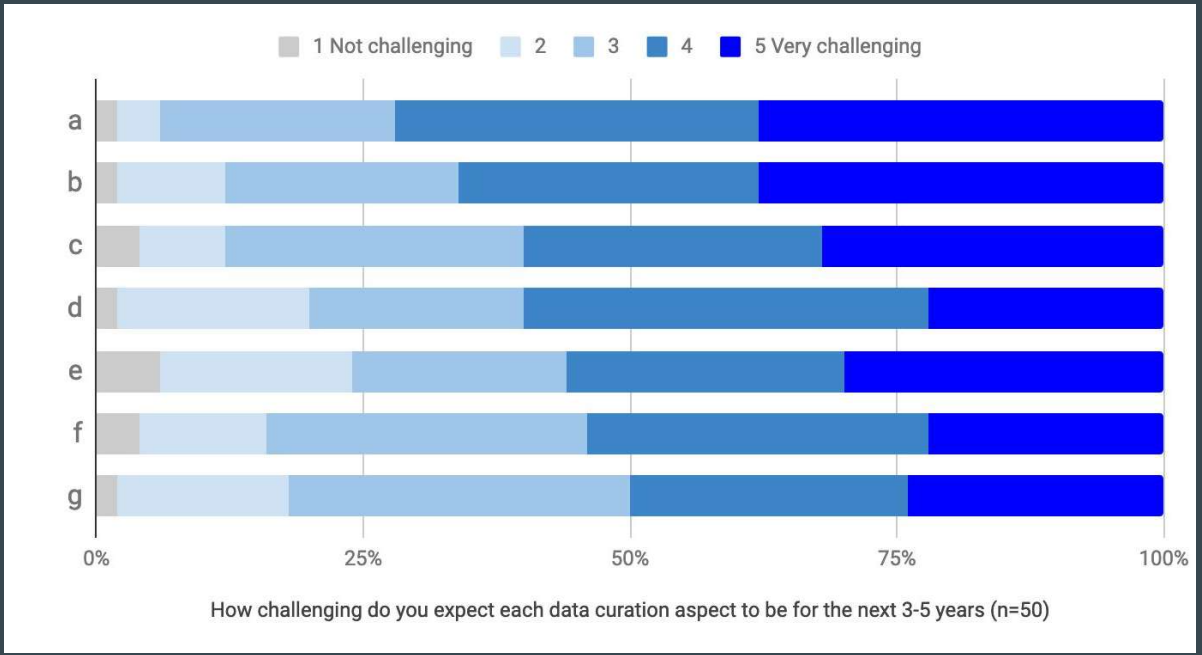
Curators

4

The actions taken by a data curator result in ethical, reusable, and better datasets for research and education.

Barriers to Well-Curated Data

A. Expertise in domain data
B. Scaling with increased demand
C. Training & retooling existing staff
D. Outreach/marketing
E. Recruiting & retaining staff
F. Keeping up with technology changes
G. Keeping up with data sharing requirement changes



*2017 study of 80 ARL Institutions in: Hudson-Vitale, C, Imker, H, Johnston, LR, Carlson, J, Kozlowski, W, Olendorf, R and Stewart, C. **SPEC Kit #354: Data Curation**. (2017). Association of Research Libraries (ARL). May 2017. https://doi.org/10.29242/spec.354.

Introducing the Data Curation Network

Mission

Trusted community-led network that enables researchers to openly share data in ways that are

Ethical. Reusable. Better.

DATA CURATION NETWORK

Community Led

10 Partners

28 data curators

43 domains

26 speciality file format types

University of Minnesota
Lead: Lisa Johnston, PI
Liza Coburn, Project Coordinator
Curators: Katie Wilson, Alicia Hofelich Mohr, Shanda Hunt, Melinda Kernik, Wanda Marsolek, Alexis Logsdon
Admin: Janice Jaguszewski

Penn State University
Lead: Cynthia Hudson-Vitale
Curators: Xuying Xin, Seth Erickson

University of Illinois
Lead: Hoa Luong
Curator: Ashley Hetrick
Admin: Heidi Imker

Duke University
Lead: Joel Herndon
Curators: Jen Darragh, Sophia Lafferty-Hess
Admin: Timothy M. McGeary

Washington University in St. Louis
Lead: Jennifer Moore
Curator: Dorris Scott

University of Michigan
Lead: Jake Carlson
Curators: Susan Borda, Rachel Woodbrook

Cornell University
Lead: Wendy Kozlowski
Curators: Sarah Wright, Henrik Spoon

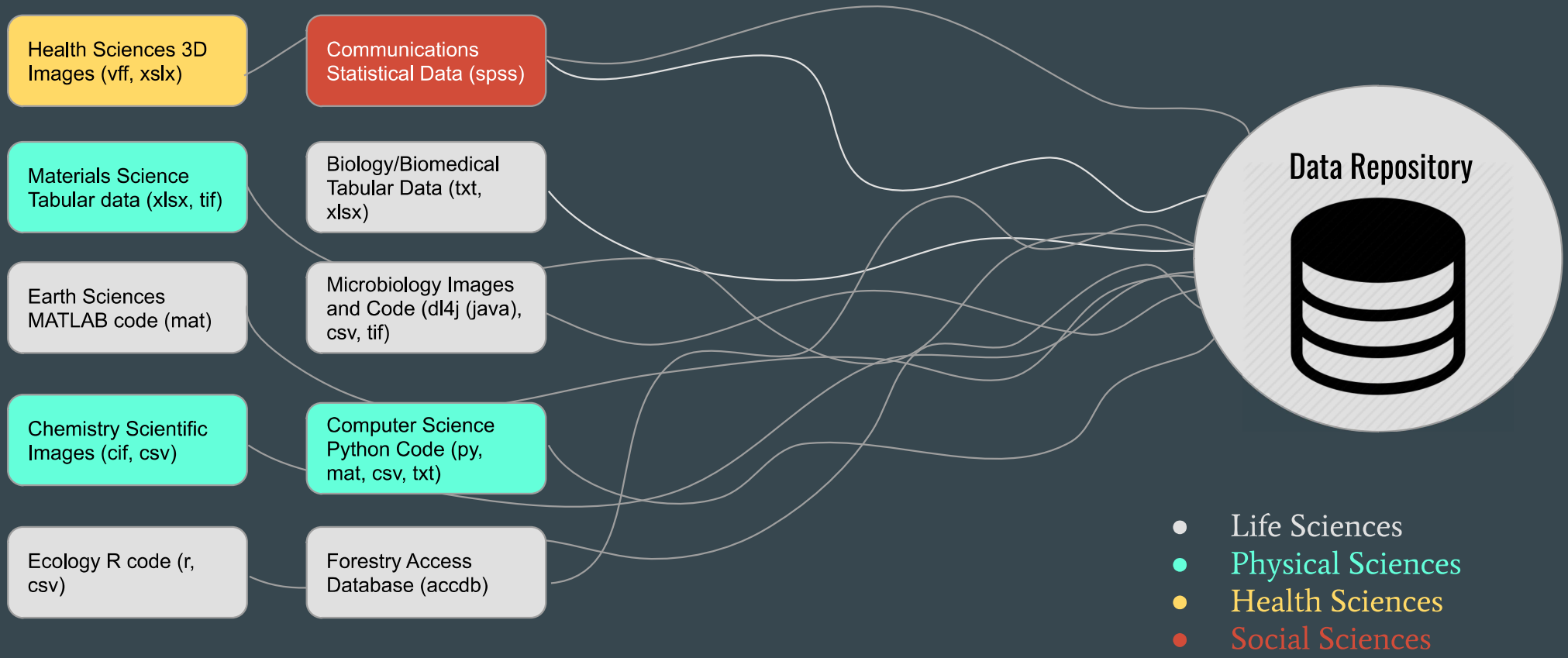
Johns Hopkins University
Lead: Mara Blake, Co-PI
Curators: Chen Chiu, Dave Fearon, Marley Kalt

Dryad Digital Repository
Lead: Elizabeth Hull
Curators: Erin Clary, Debra Fagan, Rich Yaxley

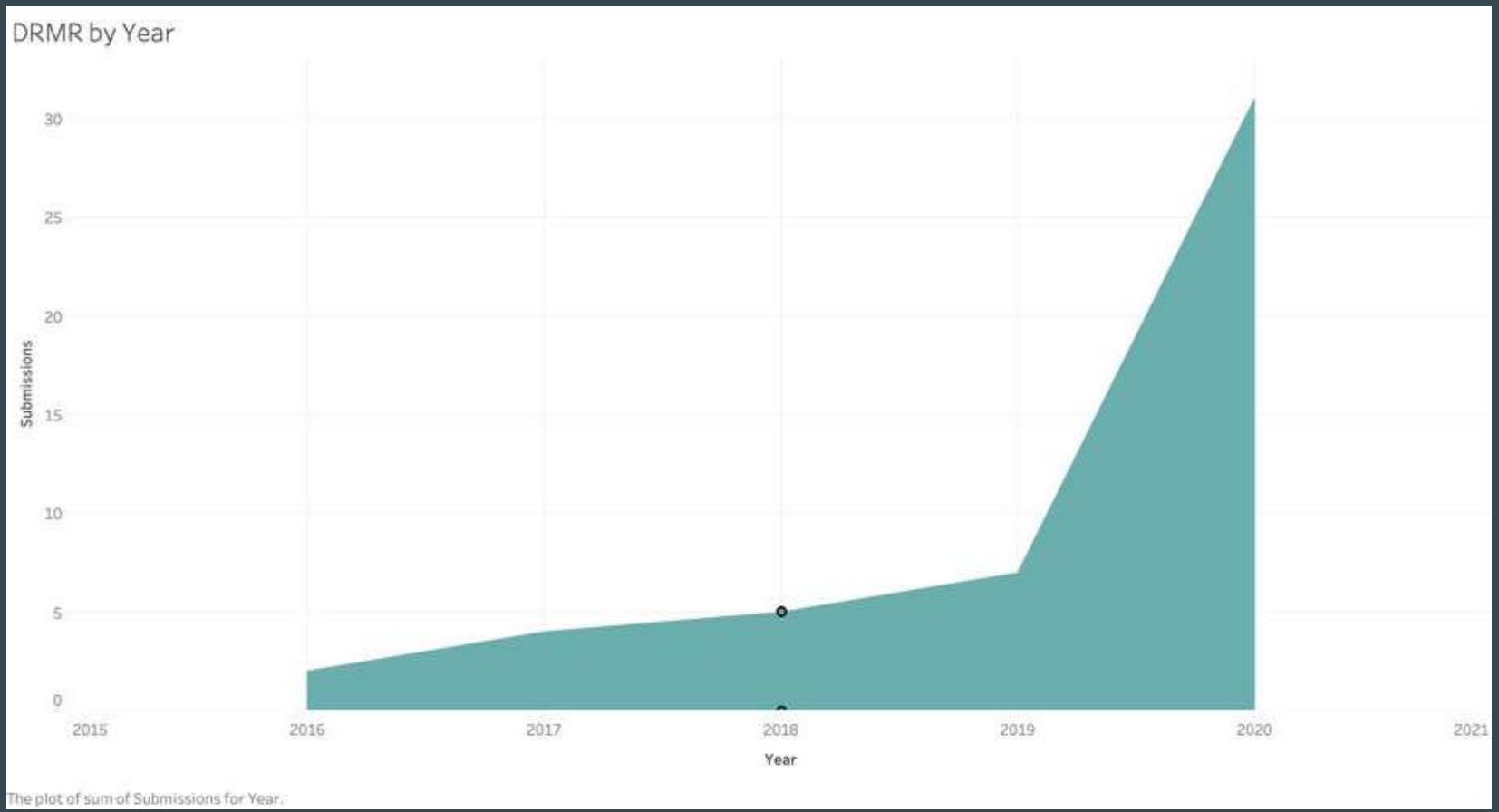
New York University
Lead: Katie Wissel
Curator: Andrew Battista

Alfred P. Sloan FOUNDATION

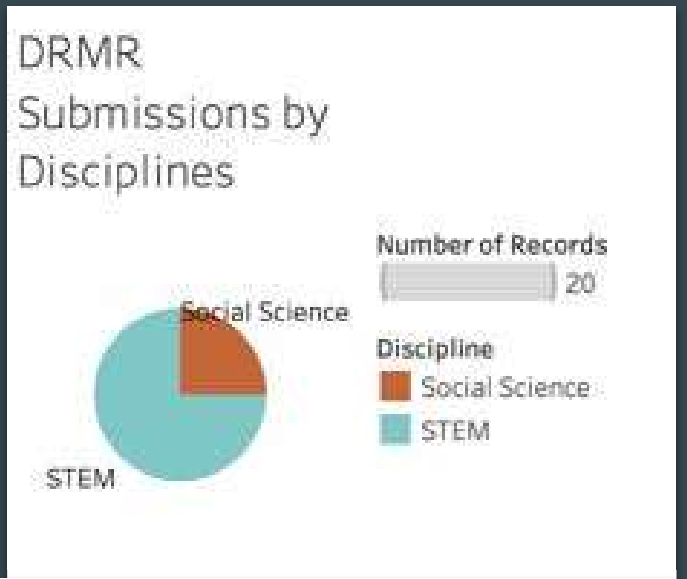
Curation at Scale

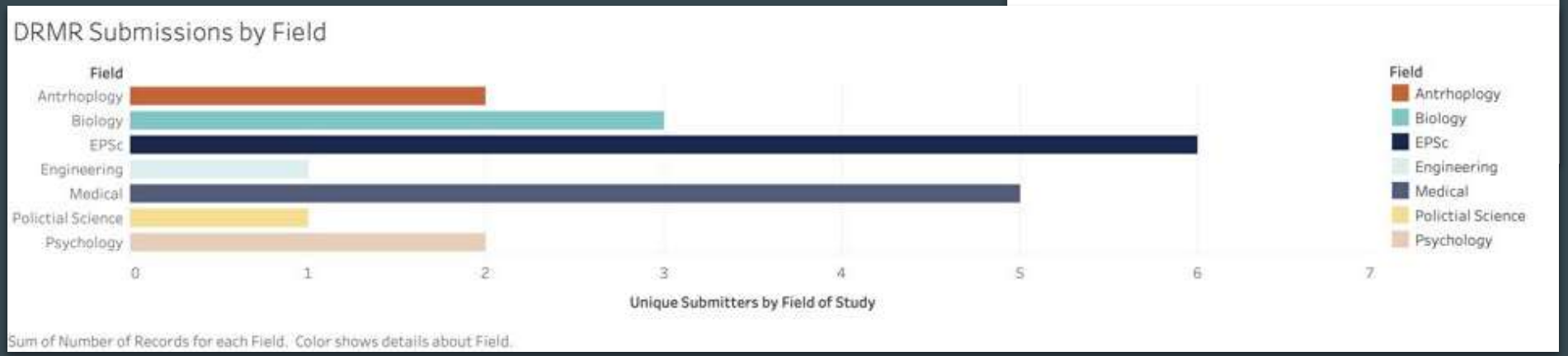


WashU Data Repository Growth

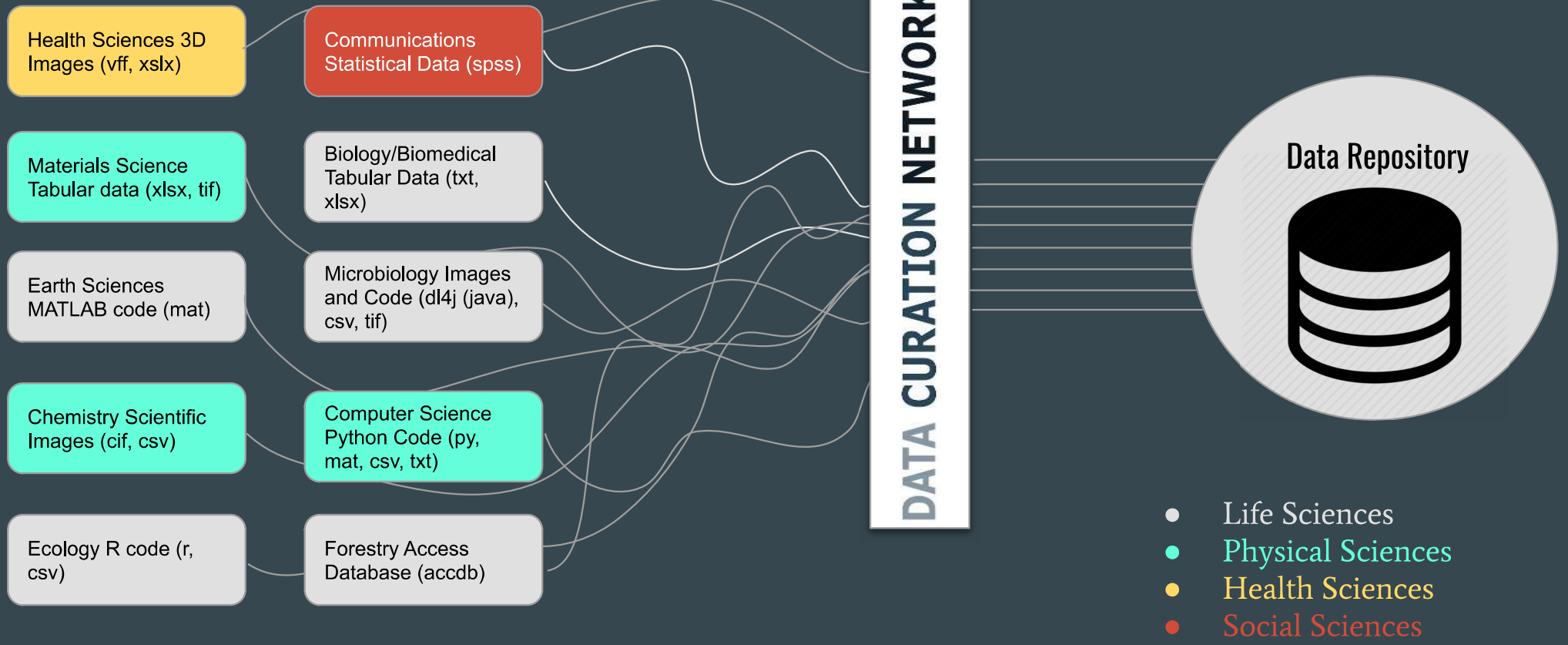


WashU Repository by Disciplines





Curation at Scale



Timeline of the DCN

Planning Phase

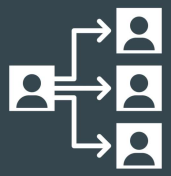
Implementation Phase

Expand

6 partners → 8 partners → 10 partners → (12 partners) → ???

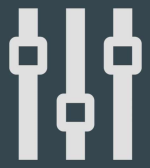
2016	2017	2018	2019	2020	2021
Research and Planning	**Model Published**	**Launching the Network**	**DCN pilot goes Live!**	**Assess and Adjust**	**Sustaining the Network**
Interviewed 91 researchers across 6 institutions about their data curation habits and needs.	Published our "collaborative staffing model for data curation" July 2017 for public comment.	Hired coordinator, trained 25 data curators, and put technology in place.	Partner institutions began curated data across the network Jan 1, 2019. Work to draft detailed sustainability plan.	Continue to improve and adapt workflow to efficiently and expertly curate data. Research value of curation.	Implement new partnership model that grows and sustains the network beyond grant funding phase.

Alfred P. Sloan FOUNDATION



DCN Curation

DCN Education

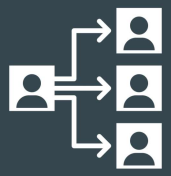


DCN Resources

DCN R&D

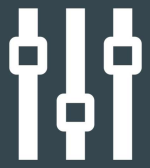
DCN Sustainability

Vision for the Data Curation Network



DCN Curation → Provide expert data curation services for network partners

DCN Education → Offer professional development opportunities for an emerging data curator professional community



DCN Resources → Create and openly share data curation best practices

DCN R&D → Demonstrate that curated datasets are measurably of greater reuse value than non-curated data

DCN Sustainability → Expand into a sustainable entity that grows beyond our initial partner institutions

Data Curation

The DCN provides the training, coordination, and technical infrastructure to seamlessly connect expert data curators across the network with all types of data sets for robust curation.

http://datacurationnetwork.org

DCN CURATE Steps

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

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

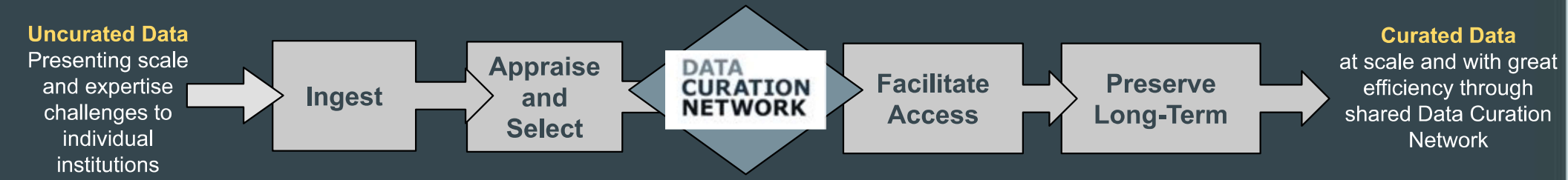
T **Transform** file formats for reuse and long-term preservation

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

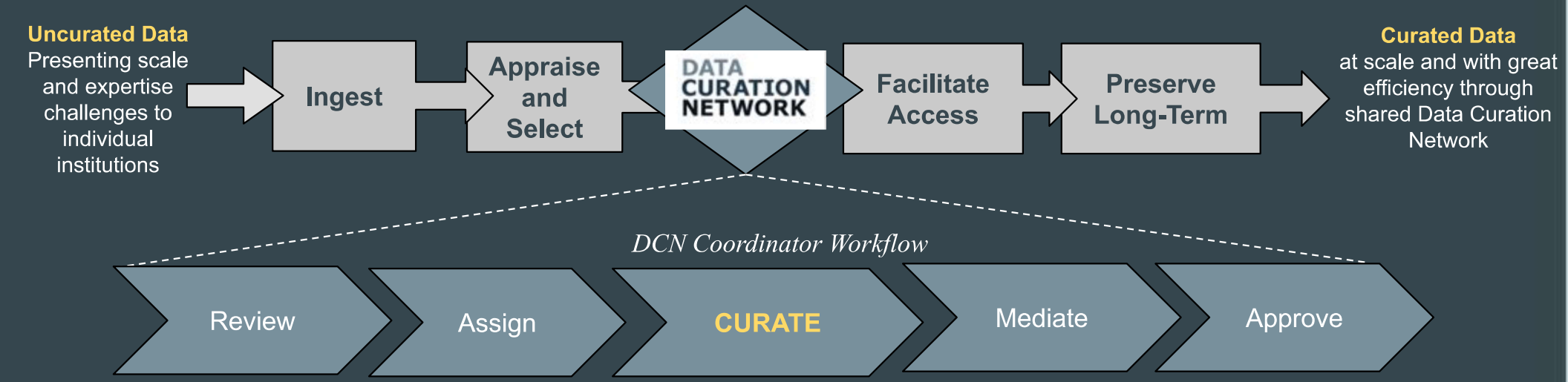
Table A1. Draft checklist of DCN CURATE steps and FAIRness scorecard

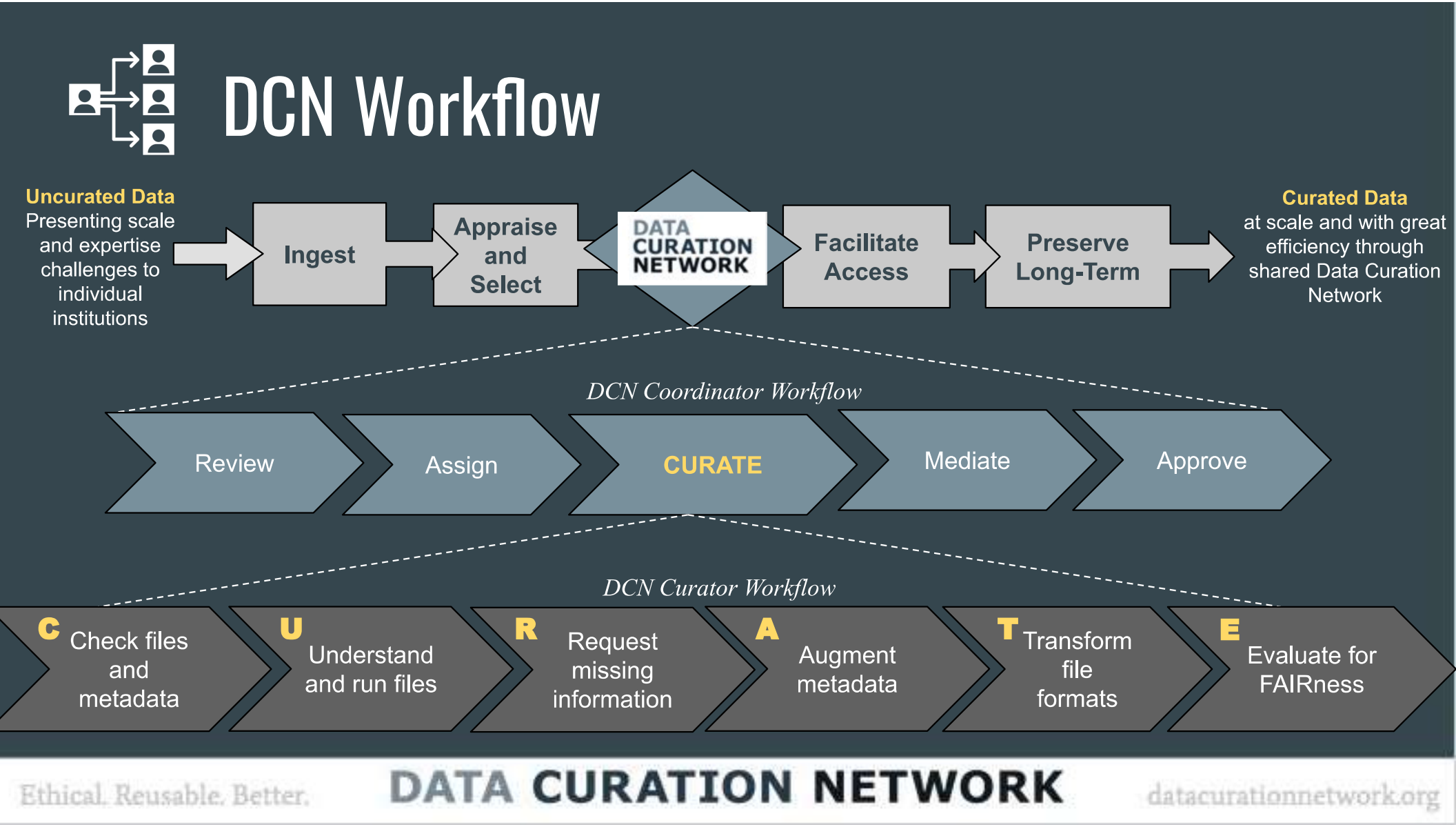
CURATE Actions	Curation Checklist
Check data files and read documentation ▪ Review the content of the data files (e.g., open and run the files or code). ▪ Verify all metadata provided by the author and review the available documentation.	☐ Files open as expected ☐ Issues _____ ☐ Code runs as expected ☐ Produces minor errors ☐ Does not run and/or produces many errors ☐ Metadata quality is rich, accurate, and complete ☐ Metadata has issues _____ ☐ Documentation Type (*circle*) Readme / Codebook / Data Dictionary / 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 subject domain. For example....*

DCN Workflow

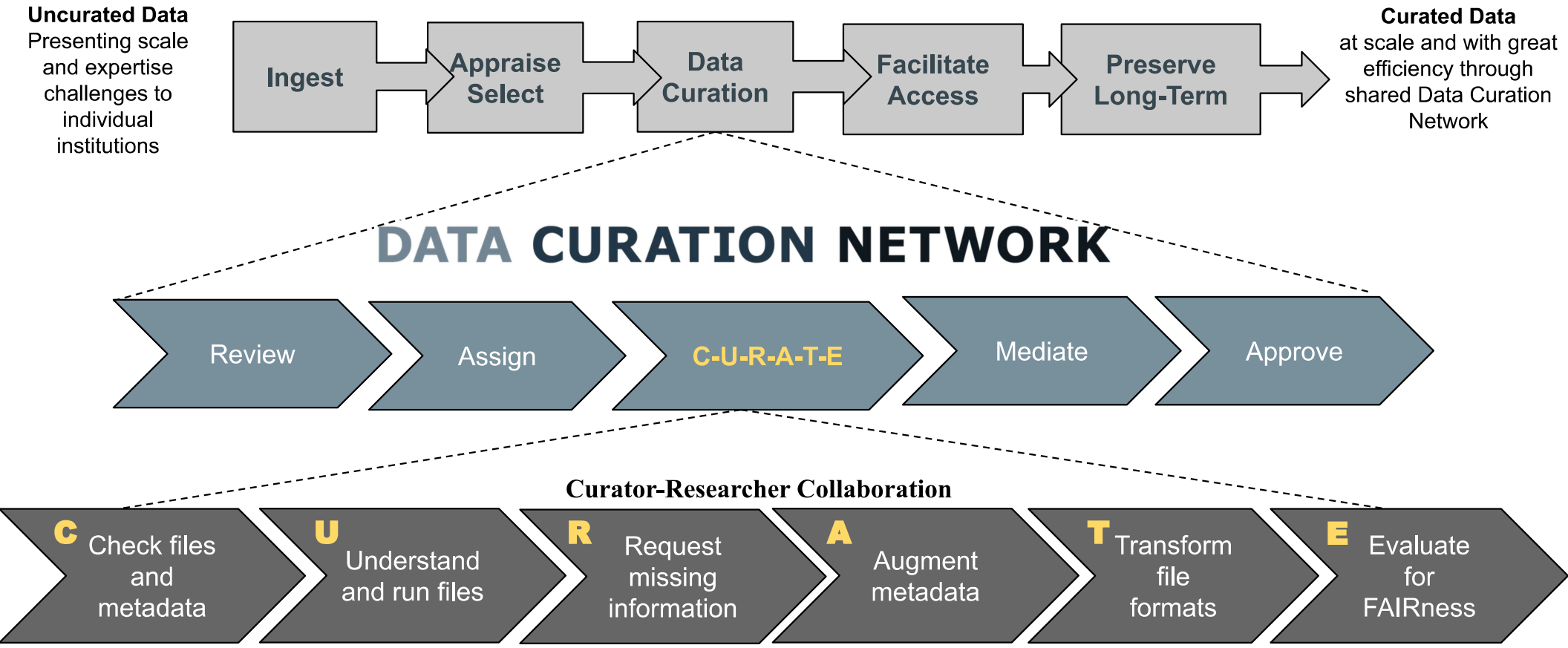


DCN Workflow




DCN Workflow
Uncurated Data
Presenting scale and expertise challenges to individual institutions
Ingest
Appraise and Select
DATA CURATION NETWORK
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
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
Ethical. Reusable. Better.
DATA CURATION NETWORK
datacurationnetwork.org

CURATE Steps used by the Data Curation Network



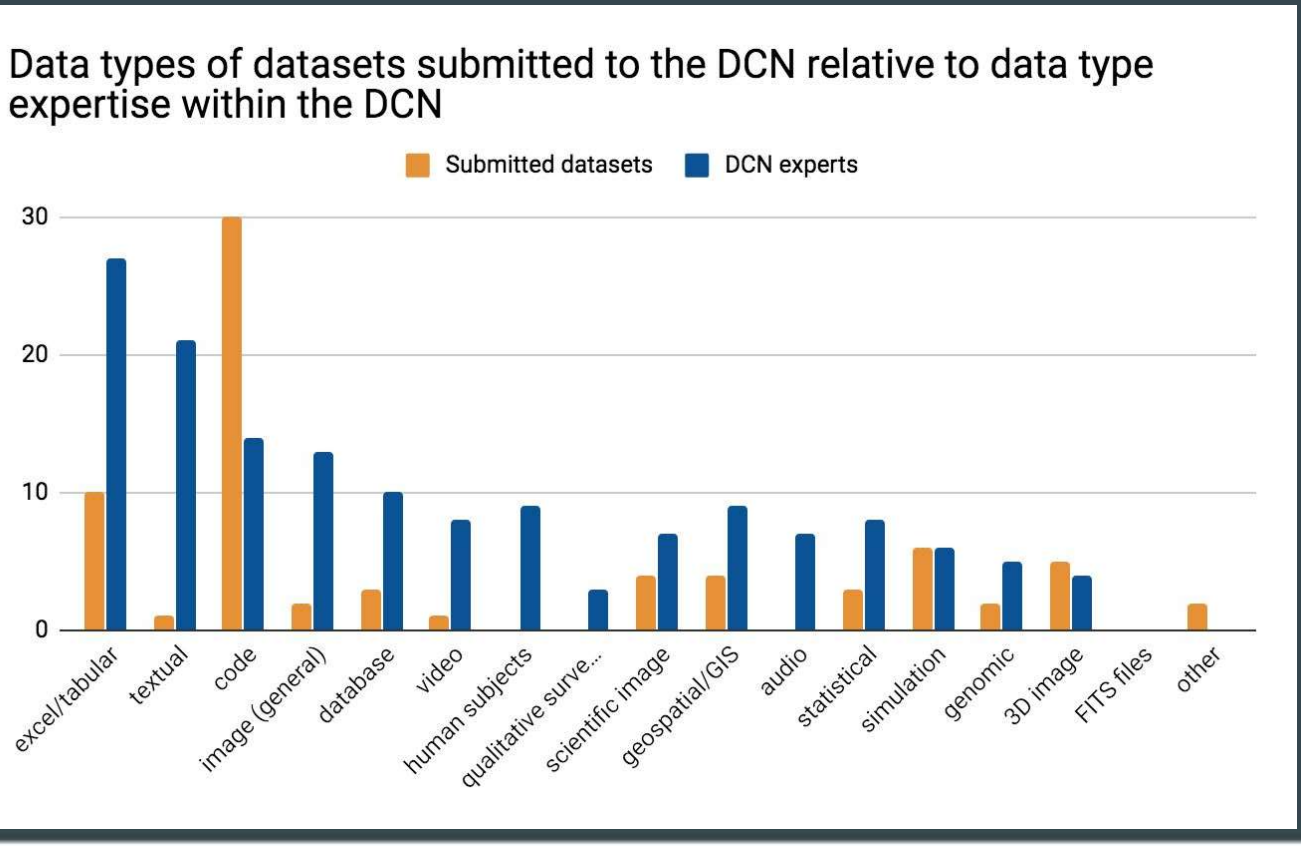
Tools we use to run the Network

- Jira time-tracking software[1]
- Survey to capture curator expertise
- Annual training and networking opportunities
- Slack, listserv for ongoing community

WordPress Jira Software slack

1. Kozlowski, Wendy, Elizabeth Coburn and Mara Blake. Walk it Like you Talk it: Jira as a tool for documenting the curation process. RDA 13th Plenary Meeting; 2019 April 2-4; Philadelphia, PA.

Measures of Success



Data curation stats (viz)

Satisfaction surveys

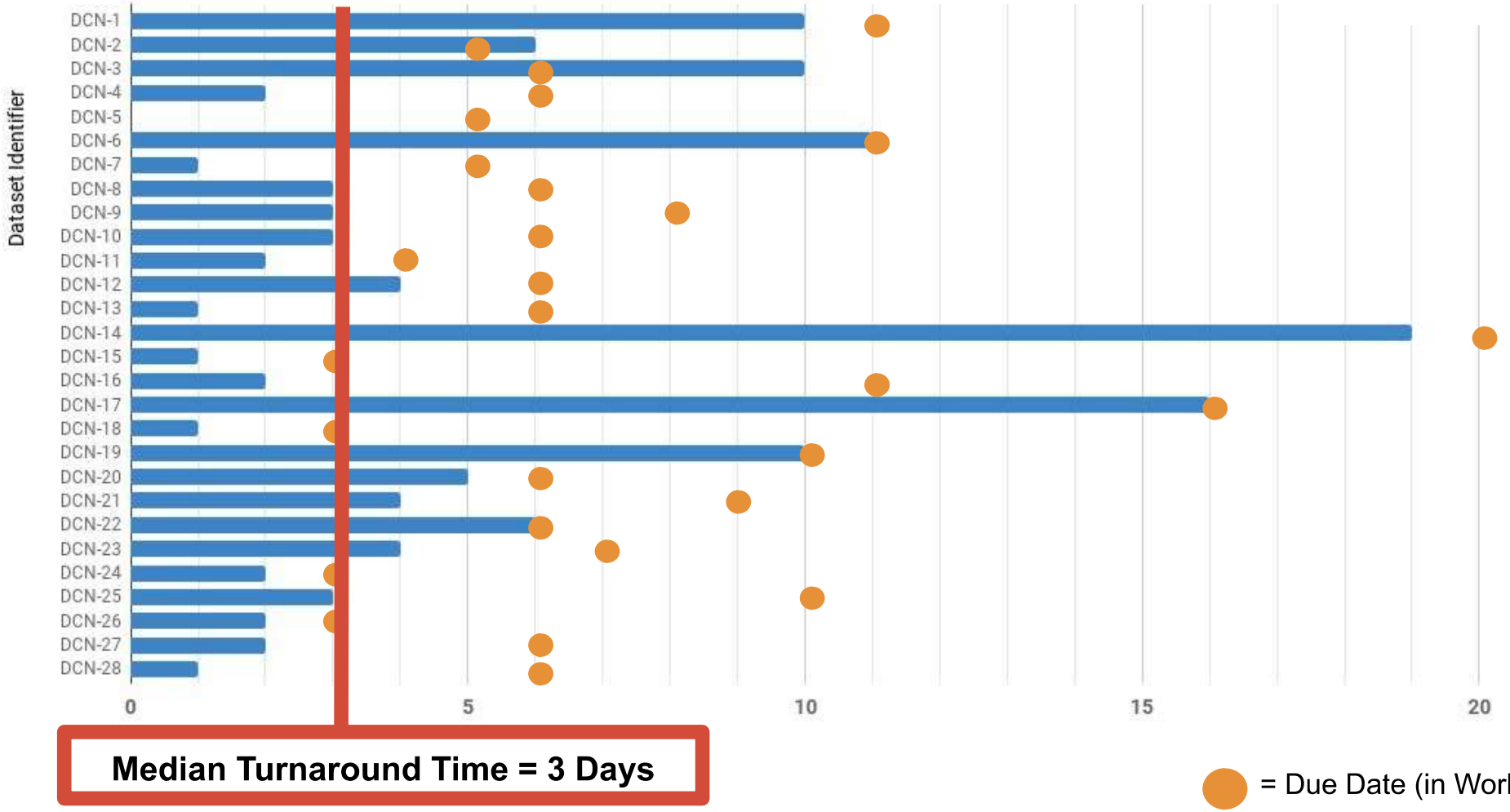
Efficiencies gained

- 58 data sets curated
- Averaging 2.5 hours/dataset
- Assignments made within 24 hours

DATA CURATION NETWORK

DCN Dataset Turnaround Time in Working Days (M-F)

Number of Working Days from Submission to Curation Completion



Ashley Hetrick

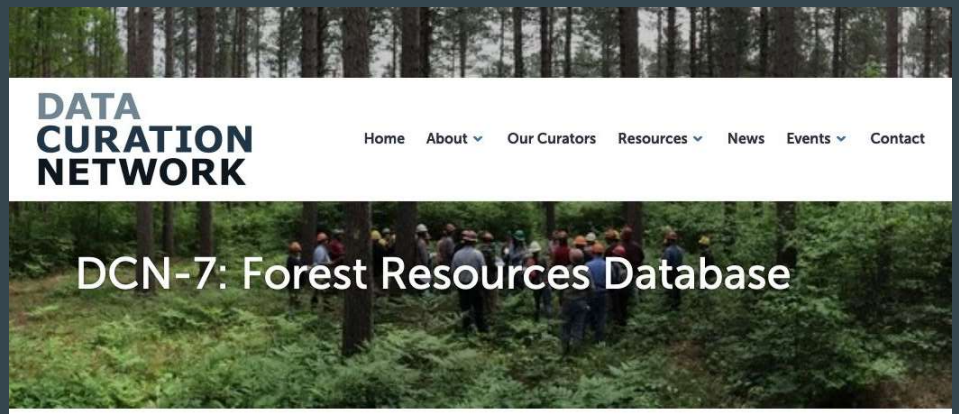
Assistant Director for Research Data Engagement and Education
University of Illinois

Hetrick leads the development and delivery of data management work
well as data management plan (DMP) reviews, for the Research Data Service (RDS).
Prior to this position, Hetrick worked in various roles within Illinois' central
Information Technology (IT) group, including social media analytics, IT
communications, and leading an AV/IT help desk service.

Follow

Data sets curated by Ashley

DCN-7: Forest Resources Database



Data set citation

"Cloquet Forestry Center Continuous Forest Inventory (1959-2014)" available at the Data Repository for the University of Minnesota, https://doi.org/10.13020/096z-kg59.

Curated by Lisa Johnston at the Data Repository for the University of Minnesota and Ashley Hetrick at the Illinois Data Bank.

Curation actions

DISCOVERY SERVICES

DOCUMENTATION

EVALUATE FAIRNESS

Follow

DCN Education

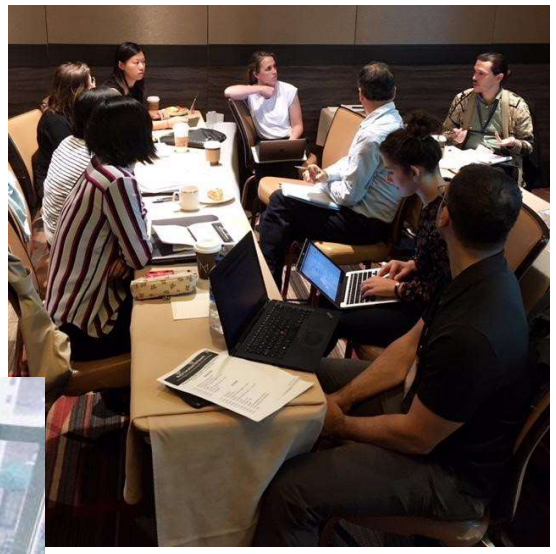
We offer professional development opportunities for an emerging data curator professional community

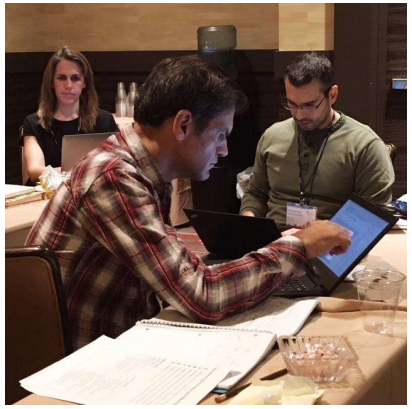
https://sites.psu.edu/dcnworkshops

Enhancing Expertise throughout the Broader Community

Specialized Data Curation Workshop @JHU 2019

DCN Education



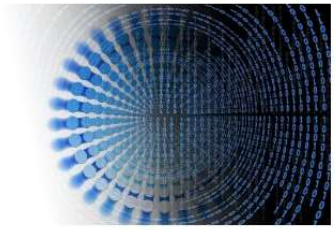


https://sites.psu.edu/dcnworkshops/

Learning Outcomes

1. Increase understanding of data curation practices and tools in various disciplines, data types, and formats.
2. Share expertise and enhance curation capacity for curation nationwide.
3. Meet like-minded colleagues who are interested in building and extending curation practices.

INSTITUTE of Museum and Library SERVICES



DATA CURATION NETWORK

Specialized Data Curation Workshop Agenda

April 17th & 18th ◆ Johns Hopkins University ◆ Baltimore, Maryland

Wednesday

9:00	Welcome & Breakfast
9:30	The Value of Curation
10:00	Curation Deep Dive #1: C Step
10:30	Break
10:45	Curation Deep Dive #1: U Step
12pm	Lunch
1:00	Primer Timer → pitch idea of primer topics
1:30	Curation Deep Dive #2: R & A Steps
2:30	Break
3:00	Curation Deep Dive #2: R & A Steps continued
4:00	End of Day One
5:30	Reception

Thursday

9:00	Breakfast
9:30	Coffee with Data
10:15	Review Day 1
10:30	Curation Deep Dive #3: T Step
11:30	Lunch
12:15	Curation Deep Dive #3: E & D Step
1:15	Primer Time 2
2:00	Group feedback on primers
2:15	Wrap up
2:30	Everyone Disperses

Check files
Understand or try to
Request missing information
Augment the submission
Transform the format
Evaluate for FAIRness
Document throughout

Pictured: Group activity at the DCN Specialized Data Curation Workshop, co-located at the DLF Forum on October 17-18, 2018.

www.datacurationnetwork.org

Our curriculum engages attendees with lectures, group activities and demonstrations.

Hands-on data curation activities

Created by Peter van Driel
from Noun Project

- Survey Data
- Tabular Data
- Code
- Image Data
- Geospatial Data

Data Curation Assignment: Images (Penn State)

Title: S'Urachi Site-Based Archaeological Survey 2015

Author: Victor T. Hail

Discipline: Archeology

Date: 2015

Access: Public

Reason for deposit: Connect to published article and report

Data Curation Network

DCN Workshops

Workshop #1
Las Vegas, NV Oct 2018 (DLF) (n=22)

- Geodatabases
- Microsoft Excel
- Jupyter Notebooks
- Microsoft Access
- netCDF files
- Wordpress
- SPSS

Workshop #2
Baltimore, MD (JHU) April 2019 (n=27)

- Atlas.ti
- Confocal microscopy
- GeoJSON
- Google Docs
- Lidar Point Clouds
- NVivo
- PDF
- R
- STL files
- Tableau
- Text/character encoding

Workshop #3
St Louis, MO (Wash U) Nov 2019 (n=29)

- Shape files
- ISO Images
- GEOTiff
- DarwinCore
- NIFTI BIDs
- NVIVO
- Oral Histories
- SAS

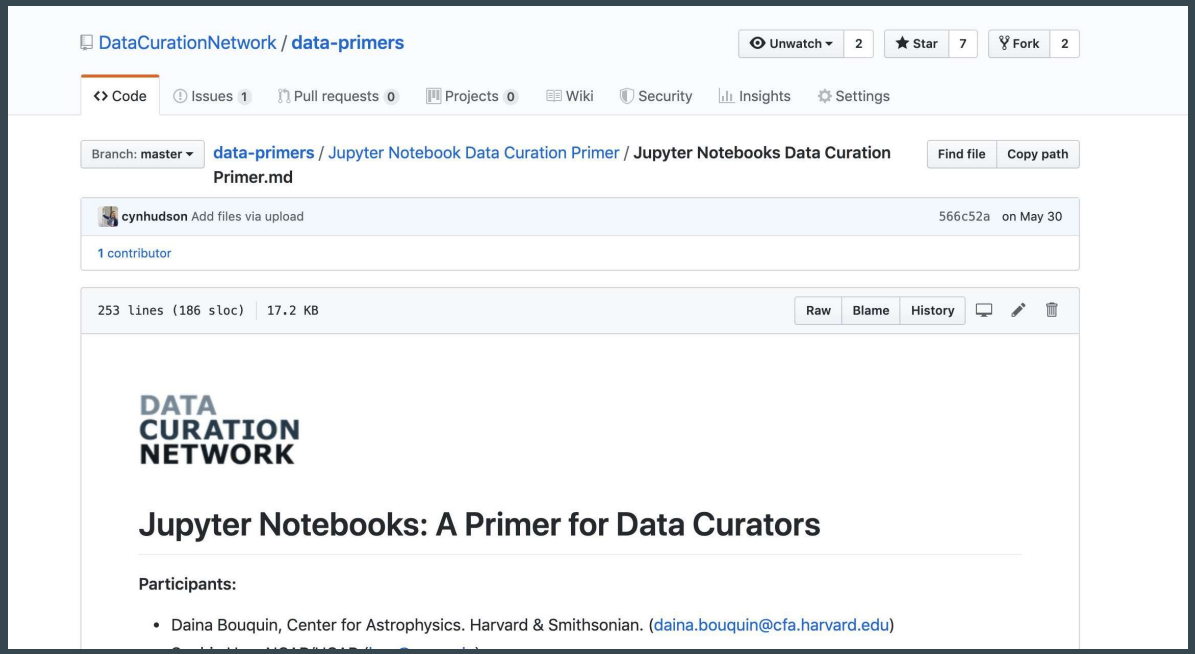
Data Curation Primers

A platform for the community to create and openly share data curation best practices

https://github.com/DataCurationNetwork/data-primers

Publication

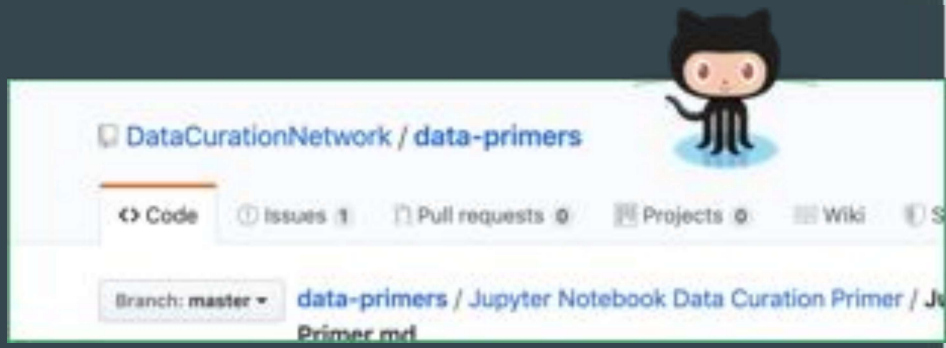
https://github.com/DataCurationNetwork/data-primers



- Published on GitHub
- Primers are expected to grow from their original version
- The community may suggest revisions

Primer creation process:

- Attendees select topic (2-3 people)
- Working session (Roadmap)
- DCN mentor
- Meet monthly (6 months)
- Peer review; co-occurring with internal webinar
- Revisions
- Publish to Github / IR

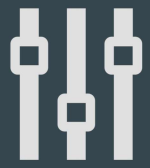


SPSS

Authors: Joshua Dull, Sai Deng, Shahira Khair & Jeanine Finn

DCN Mentor: Sophia Lafferty-Hess

https://github.com/DataCurationNetwork/data-primers



Key Curatorial Considerations:

- Preservation actions
  - Save as .por? To ASCII or not to ASCII?
  - Preservation recommendations
    - ICPSR, LOC and others
  - Suggested software for converting & reviewing SPSS files
- Further considerations
  - SPSS Version
  - Researcher feedback
    - Which files do researchers save?
- Other highlights
  - SPSS Tutorials
  - Bibliography for more curation resources

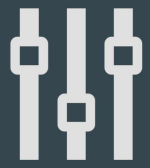
Data Curation Network. (2019). RDAP Primerpalooza: Introducing Data Curation Primers [SPSS Primer slides by Joshua Dull, Sai Deng, Shahira Khair & Jeanine Finn]. Retrieved from: https://vimeo.com/350235467

Microsoft Access

Author: Fernando Rios & Dave Fearon

DCN Mentor: Dave Fearon

https://github.com/DataCurationNetwork/data-primers



Key Curatorial Considerations:

What is the complexity of the database?

- Simple DBs (few tables, no forms, queries, macros) could be curated like a spreadsheet

As a base level for preservation:

- Keep original files + export tables to flat CSVs
- Screenshot the Relationships Diagram
- Run the Database Documenter and save the report alongside the DB
- Check for linked tables
- Other objects (SQL, forms, VB)?

Need help from creator

- Table relations, meaning of column names, how data is to be queried

---

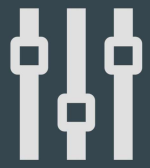
Data Curation Network. (2019). RDAP Primerpalooza: Introducing Data Curation Primers [Microsoft Access Primer slides by Fernando Rios & Dave Fearon]. Retrieved from: https://vimeo.com/350235467

Microsoft Excel

Authors: Ho Jung Yoo, Sandra Sawchuk & Greg Janée

DCN Mentor: Wendy Kozlowski

https://github.com/DataCurationNetwork/data-primers



Key Curatorial Considerations:

There are no metadata standards for Microsoft Excel, so detailed documentation from the depositor is encouraged. Documentation should contain info about:

- Context of the original study
- Description of each file
- Description of each worksheet (ideally one table per worksheet)
- Revisions of the data
- Description of each variable in the files

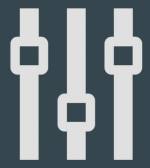
Data Curation Network. (2019). RDAP Primerpalooza: Introducing Data Curation Primers [Microsoft Excel Primer slides by Ho Jung Yoo, Sandra Sawchuk & Greg Janée]. Retrieved from: https://vimeo.com/350235467

Jupyter Notebooks

Authors: Daina Bouquin, Matthew Benzing, Sophie Hou & Lee Wilson

DCN Mentor: Susan Borda

https://github.com/DataCurationNetwork/data-primers



Code is not Data!

Jupyter notebooks contain code, incorporate data, and require different considerations

Different metadata for different situations

- Minimal deposit
- Runnable deposit
- Comprehensive deposit

} Consider repository suitability

Data Curation Network. (2019). RDAP Primerpalooza: Introducing Data Curation Primers [Jupyter Notebooks Primer slides by Daina Bouquin, Matthew Benzing, Sophie Hou & Lee Wilson]. Retrieved from: https://vimeo.com/350235467

Primer Topic Preview (coming in January 2020)

- Atlas.ti
- Confocal microscopy
- GeoJSON
- Google Docs
- Lidar Point Clouds
- NVivo
- PDF
- R
- .STL files
- Tableau
- Text/character encoding

Workshop #2 at Johns Hopkins University

Share your expertise

Community Authored Data Curation Primers

https://github.com/DataCurationNetwork/data-primers

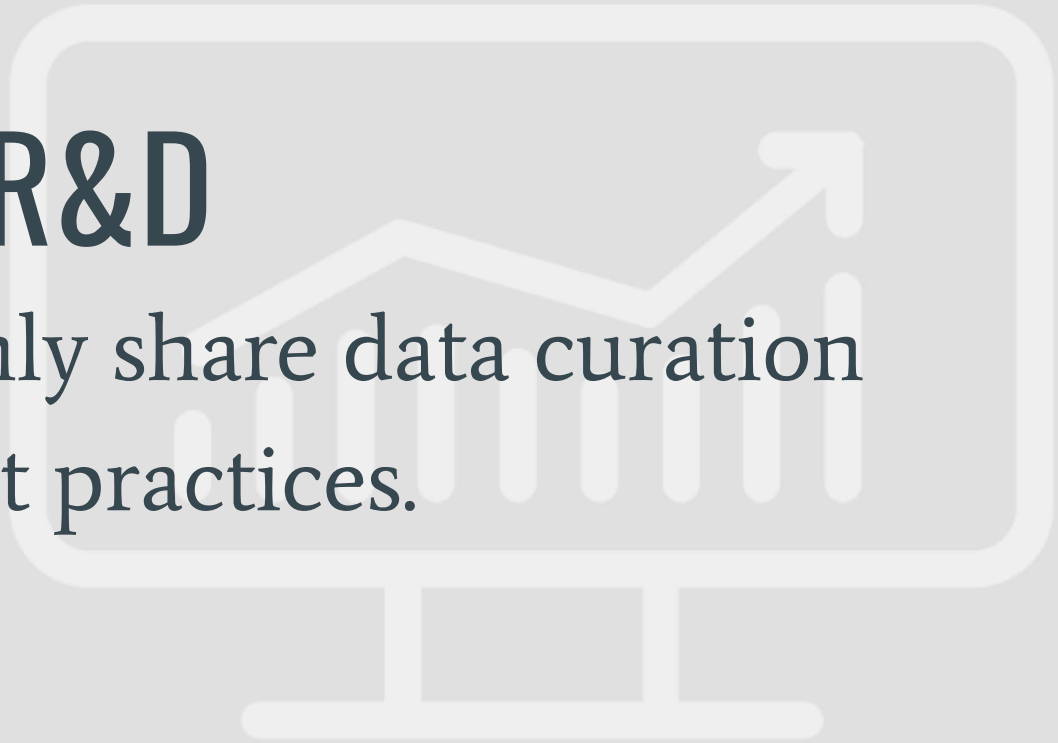
Contribute to these community resources via Github

More opportunities coming soon!

DATA CURATION NETWORK

Data Curation R&D

We create and openly share data curation procedures and best practices.



http://datacurationnetwork.org

Data Curation Network Interest Groups

Value of Curation

Goal: to determine the ROI of data curation

- Surveying library and repository field about the frequency by which curation activities are performed = create a model for curation practices
- Evaluate curated versus uncurated datasets on this model

Big Data

Goal: to collaboratively explore big data curation

- Depositing large data sets (exceed file size limits)
- Integrating local research data repository with Globus
- Archiving data that requires extra security and access restrictions
- Extending data curation for all of the above

Data Curation Network Interest Groups

Institutional Outreach

Goal: to investigate and share stories across institutions related to:

- engagement at the institutional level
- surrounding both data curation and research data management services
- in order to learn from each others' experiences and,
- help increase effectiveness in outreach.

Human Subjects

Goal: Build curation primers for:

- Human subjects data in general
- Consent form review
- De-identification methods

Sustainability

Expand into a sustainable entity that grows beyond our initial partner institutions.

http://datacurationnetwork.org

Growing the Data Curation Network

- Slow, intentional growth
- Y1-2 Recruit 4 new partners
  - 2019: 2 new partners!
  - 2020: Future
- Expand more broadly in 2021

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
Support	Grant Funded (Y1-Y2) transition to partnership model (Y3)			Curation-as-service (Y4-6)		
Timing	2017-19		2020-22		2022-2023	
Phase	Implementation		Transition		Sustaining	
Partners	8 initial partners + 4 more incrementally			Recruit new partners as use and demand dictate		

Sustainability Planning

- 2019 Advisory Panel
- Consultant RFP process
- Lyrasis (June 2019-Dec 2019)
  - Market analysis (focus groups, interviews).
  - Administrative Structures (legal support, not-for-profit)
  - Financial Models (in-kind, membership, fee-for-service, or a hybrid)
  - Community Engagement case studies

DCN 2019 Advisory Panel

Yasmeen Shorish, James Madison

Jeff Spies, 221B

Limor Peer, Yale University

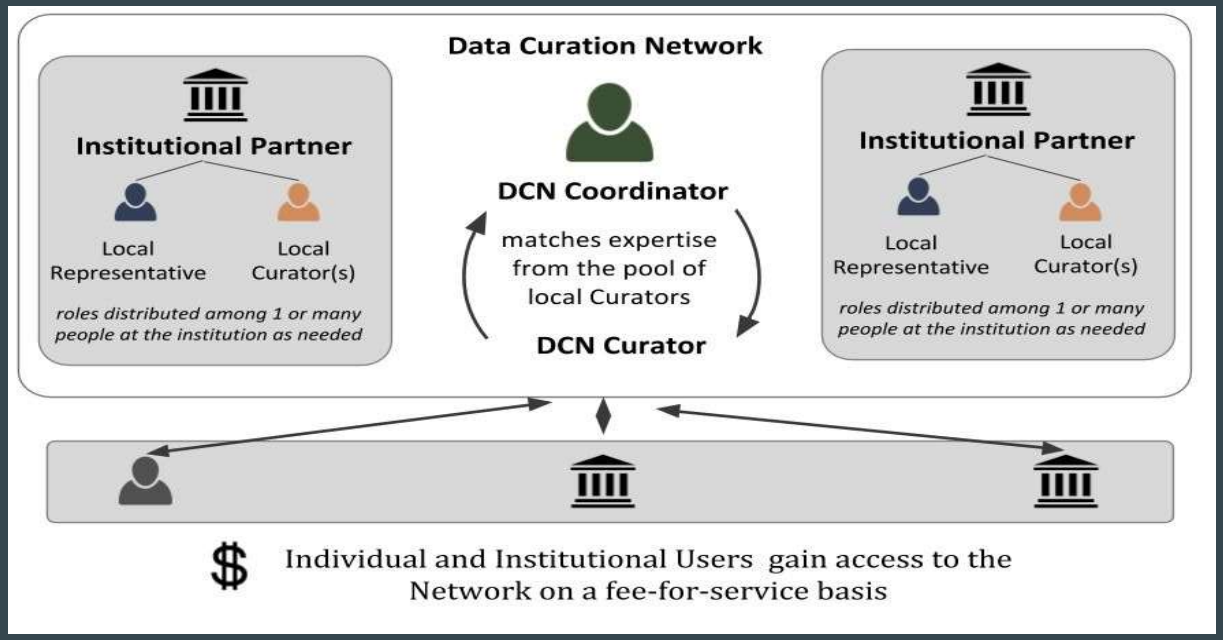
John Chodacki, Univ California

Mike Roy, Middlebury College

Jay Brodeur, McMaster College

DRAFT : Hybrid Model

- Opt 1: In Kind
- Opt 2: Fee for service



What's next?

Future directions

- Advocacy?
- Consultation?
- Domain repositories?
- Professional curator community?

We are happy to work together! Data curation without borders!

Thank you

Contact us!

DATA CURATION NETWORK

dcn-team@googlegroups.com

http://datacurationnetwork.org