DATA CURATION NETWORK

Data Curation Network: Collaboratively Enhancing Capacity for Research Data Support

CNI

December 13-14, 2021

Wendy Kozlowski
Cornell University Library

Lisa Johnston
University of Minnesota Libraries

Outline

- What is the Data Curation Network?
- Importance of data curation in scholarly communication landscape
- DCN project updates for:

    #1 Shared curation to enable ethical/FAIR data sharing

    #2 Thriving community of practice for curators

    #3 Unique platform for exploration and research

    #4 Sustainable organization that advocates for profession

- Takeaways and discussion

DATA CURATION NETWORK

University of Minnesota
Illinois
PennState
Cornell University
University of Michigan
Duke University
NYU
Johns Hopkins University
Dryad
University Libraries Virginia Tech
The Michael J. Fox Foundation for Parkinson's Research
University of Nebraska Lincoln
UC Santa Barbara

datacurationnetwork.org

Trusted, community-led network of curators advancing open research by making data

Ethical. Reusable. Better.

ALFRED P. SLOAN FOUNDATION

Mission

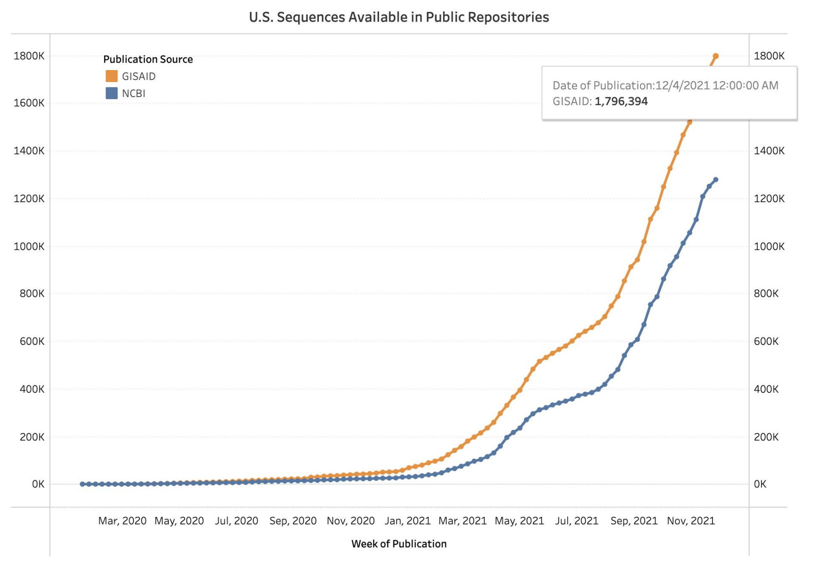
- Curate data ethically and FAIR
- Advance data curation best practices
- Develop the data curation profession
- Grow sustainably and responsively

Importance of data in scholarly communication landscape

Data sharing

Data are critical for reproducibility, increasing public trust and engagement, and enabling new discoveries to be made from re-analysis, re-combining, and reuse.

Rigor and Reuse



CDC. Published SARS-CoV-2 Sequences, as of Dec 12, 2021.
https://covid.cdc.gov/covid-data-tracker/#published-sars-cov-2-sequences

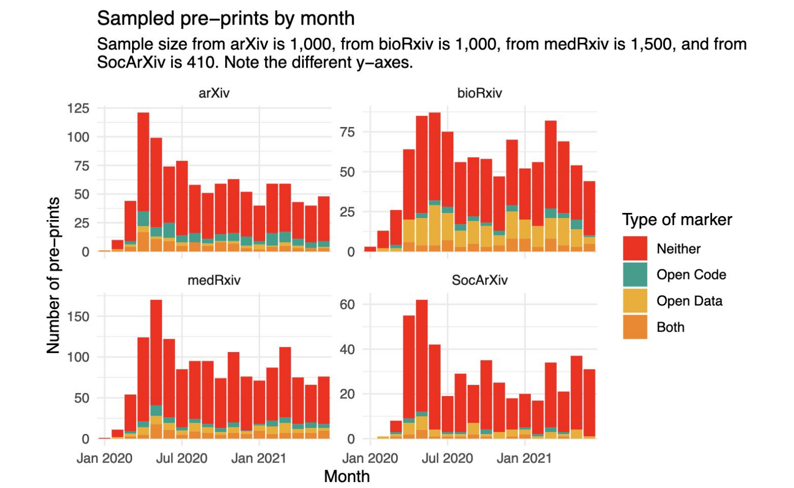
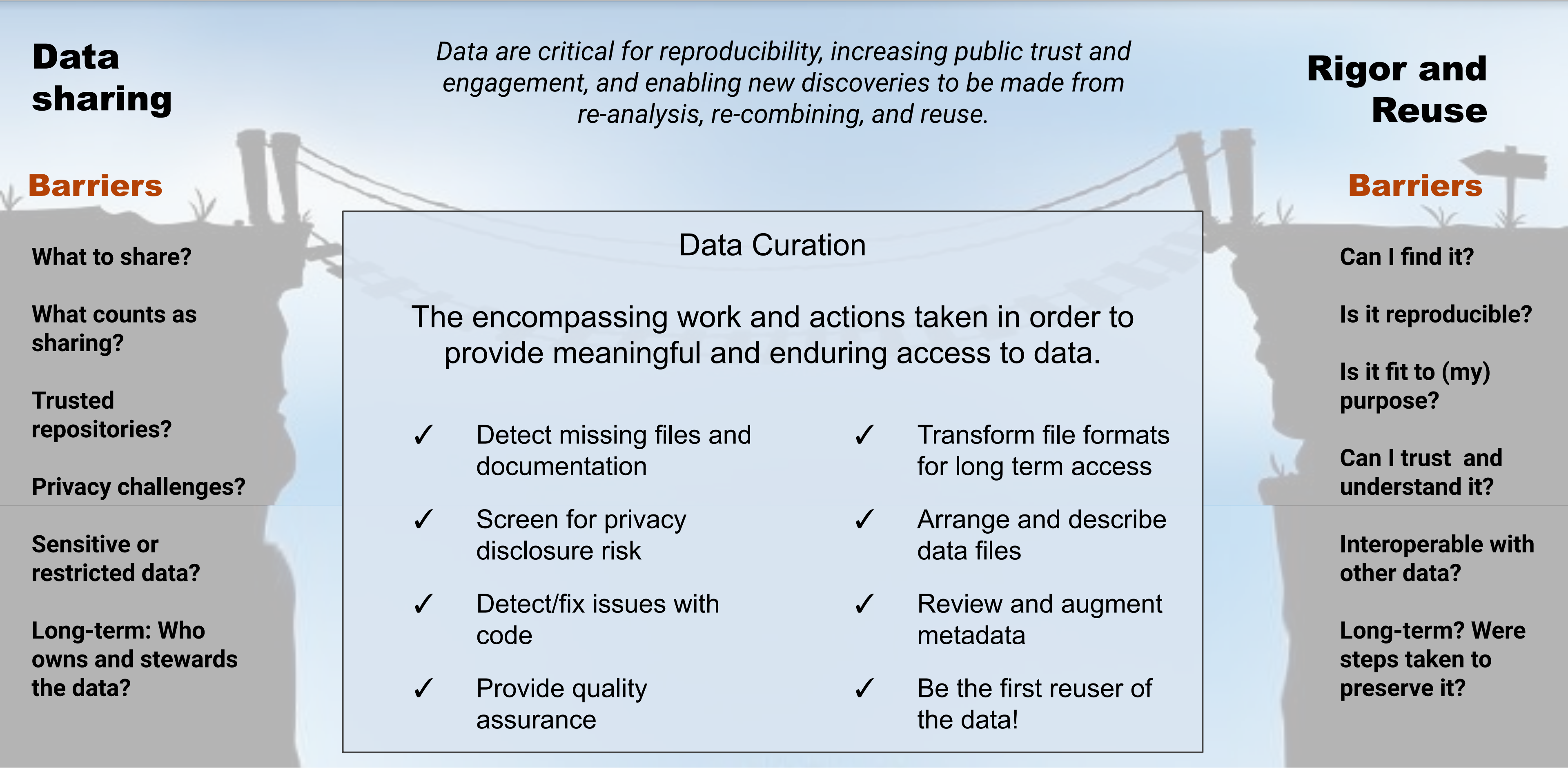


Figure 3: Number of sampled pre-prints related to COVID-19, distinguished by presence of open data or code markers

Annie Collins and Rohan Alexander. Reproducibility of COVID-19 pre-prints. 2021. arXiv: 2107.10724.

Importance of data in scholarly communication landscape



Data sharing

Barriers

What to share?

What counts as sharing?

Trusted repositories?

Privacy challenges?

Sensitive or restricted data?

Long-term: Who owns and stewards the data?

Data are critical for reproducibility, increasing public trust and engagement, and enabling new discoveries to be made from re-analysis, re-combining, and reuse.

Data Curation

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

- ✓ Detect missing files and documentation
- ✓ Screen for privacy disclosure risk
- ✓ Detect/fix issues with code
- ✓ Provide quality assurance
- ✓ Transform file formats for long term access
- ✓ Arrange and describe data files
- ✓ Review and augment metadata
- ✓ Be the first reuser of the data!

Rigor and Reuse

Barriers

Can I find it?

Is it reproducible?

Is it fit to (my) purpose?

Can I trust and understand it?

Interoperable with other data?

Long-term? Were steps taken to preserve it?

Community of 40+ data curators addressing sharing challenges

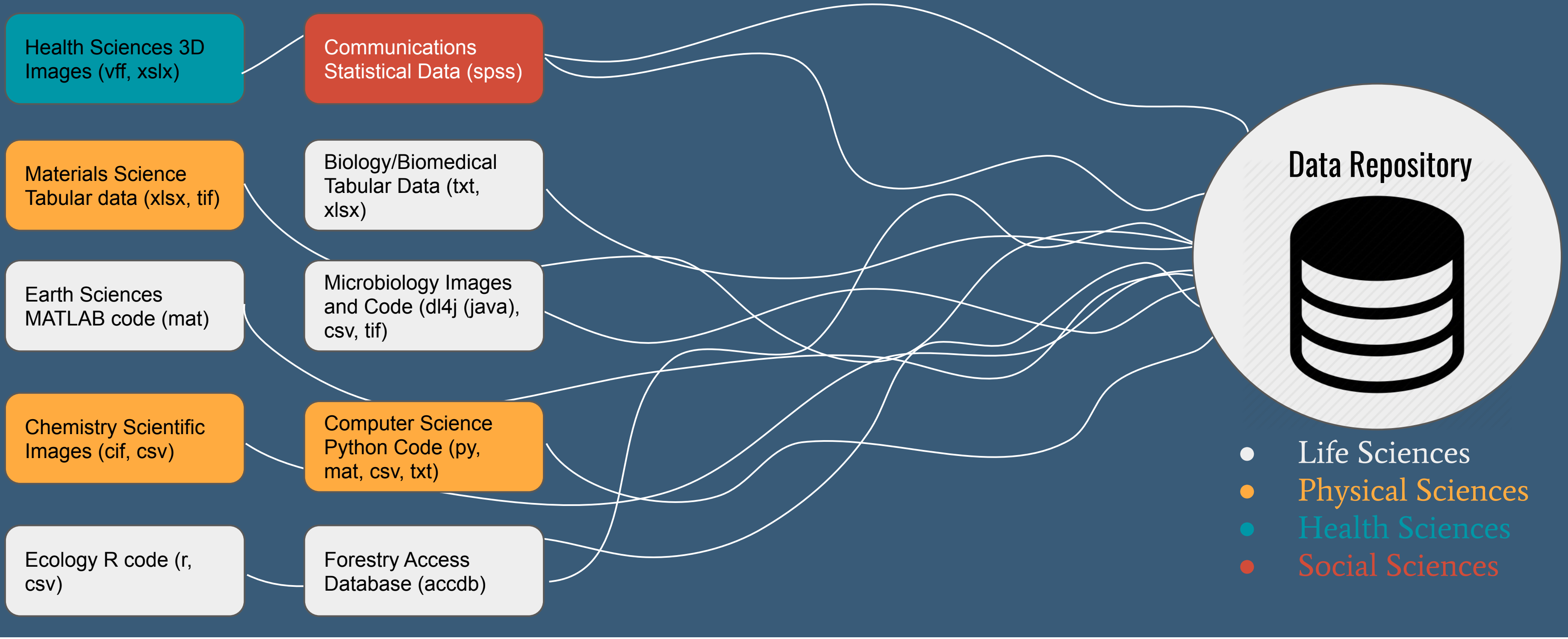
DATA CURATION NETWORK

Alfred P. Sloan
FOUNDATION

DCN highlight #1

Shared curation to enable ethical/FAIR data sharing

Most data repositories ingest multidisciplinary datasets



Health Sciences 3D Images (vff, xslx)

Communications Statistical Data (spss)

Materials Science Tabular data (xlsx, tif)

Biology/Biomedical Tabular Data (txt, xlsx)

Earth Sciences MATLAB code (mat)

Microbiology Images and Code (dl4j (java), csv, tif)

Chemistry Scientific Images (cif, csv)

Computer Science Python Code (py, mat, csv, txt)

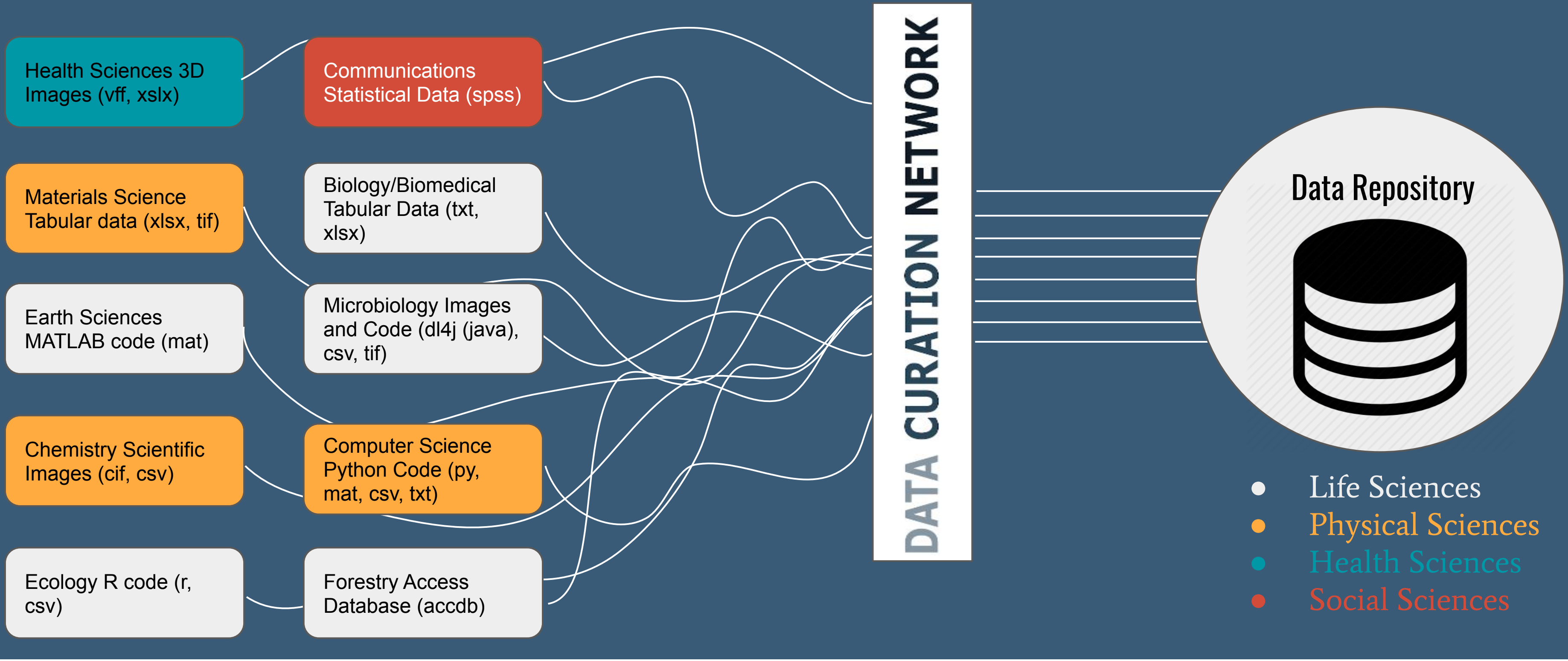
Ecology R code (r, csv)

Forestry Access Database (accdb)

Data Repository

- ● Life Sciences
- ● Physical Sciences
- ● Health Sciences
- ● Social Sciences

Share curation matches trained experts to data sets housed at institutional members of the DCN



Health Sciences 3D Images (vff, xslx)

Materials Science Tabular data (xlsx, tif)

Earth Sciences MATLAB code (mat)

Chemistry Scientific Images (cif, csv)

Ecology R code (r, csv)

Communications Statistical Data (spss)

Biology/Biomedical Tabular Data (txt, xlsx)

Microbiology Images and Code (dl4j (java), csv, tif)

Computer Science Python Code (py, mat, csv, txt)

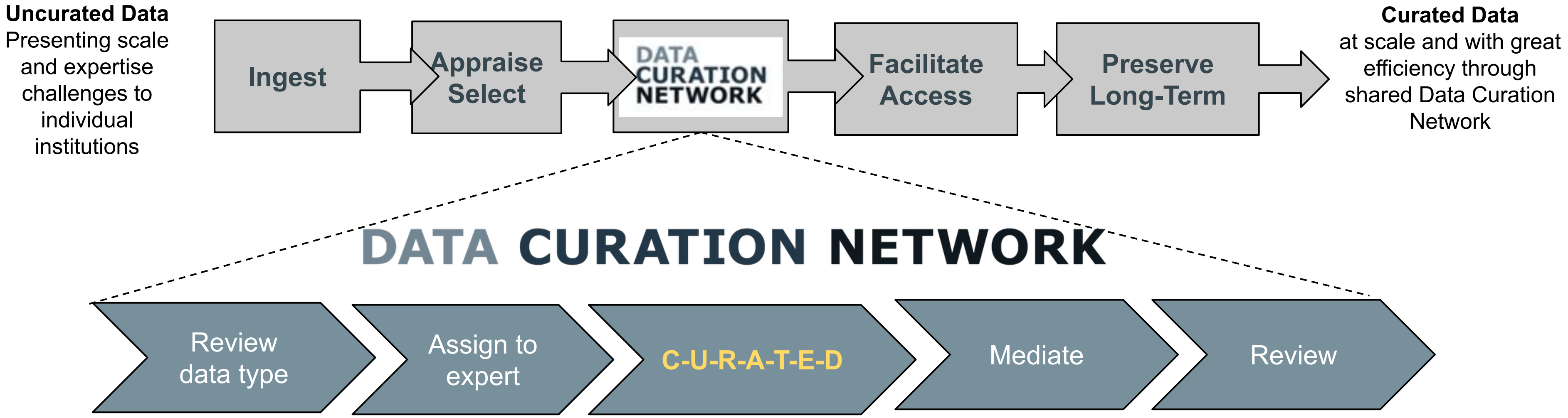
Forestry Access Database (accdb)

DATA CURATION NETWORK

Data Repository

- ● Life Sciences
- ● Physical Sciences
- ● Health Sciences
- ● Social Sciences

Data curation 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.)

The CURATE(D) Steps

Check files

Understand documentation

Request missing information

Augment the submission

Transform the format

Evaluate for FAIRness

Document throughout

https://datacurationnetwork.org/resources/workflows/

Data curation for racial justice and equity

“The DCN is challenged to reimage its CURATE framework for sustainability and inclusive language, actions and processes.”

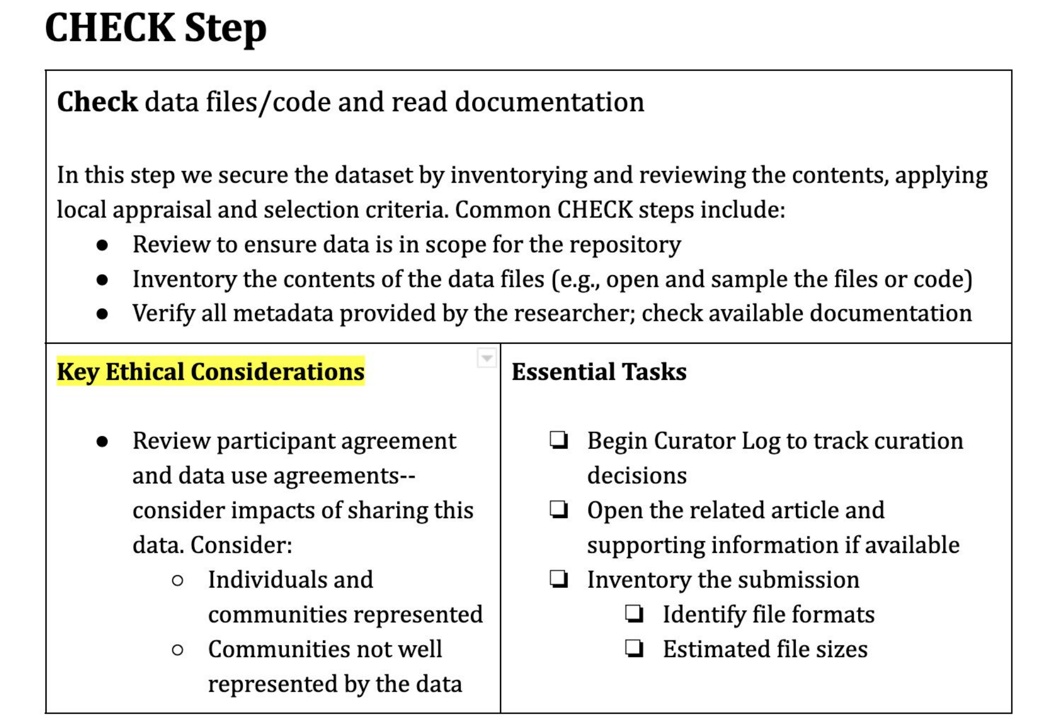
Dr. Fay Cobb Payton
https://cobbpayton.com

Payton, F. C. (2021). Centering Racial Equity in Data Curation. http://datacurationnetwork.org/publications

Activities

- Evaluated our local workflows CURATE(D)
- Identified steps data curators take to ensure data are shared are shared for good
- Looked to the following peer examples:
  - Fairness, Accountability, Transparency & Ethics (FATE) in AI
  - CARE with Collective Benefit, Authority of Control, Responsibility & Ethics
  - ACM Fairness, Accountability & Transparency (FAccT)
  - Principles for Advancing Equitable Data Practice (Urban Institute)

Ethical considerations in CURATE(D) steps



CHECK Step

Check data files/code and read documentation

In this step we secure the dataset by inventorying and reviewing the contents, applying local appraisal and selection criteria. Common CHECK steps include:

- Review to ensure data is in scope for the repository
- Inventory the contents of the data files (e.g., open and sample the files or code)
- Verify all metadata provided by the researcher; check available documentation

Key Ethical Considerations	**Essential Tasks**
● Review participant agreement and data use agreements--consider impacts of sharing this data. Consider: ○ Individuals and communities represented ○ Communities not well represented by the data	❏ Begin Curator Log to track curation decisions ❏ Open the related article and supporting information if available ❏ Inventory the submission ❏ Identify file formats ❏ Estimated file sizes

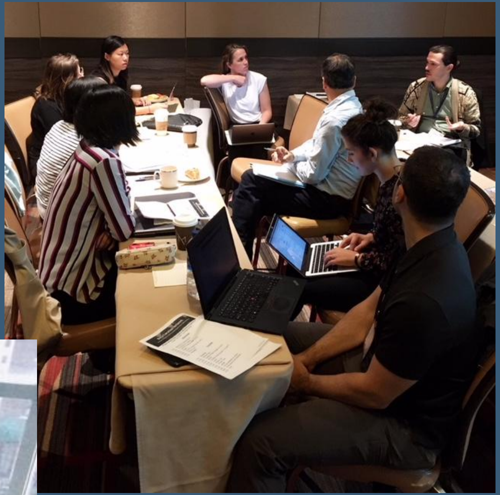
Longer-term recommendations

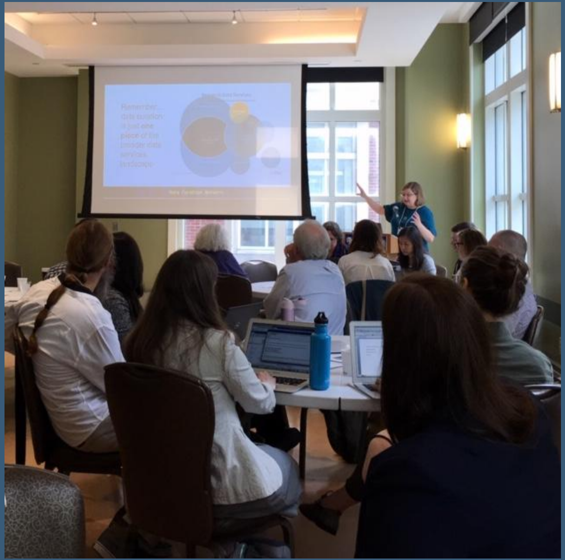
1. Expand collaborative partnerships beyond the current DCN members using participatory design methodologies.
2. Establish cohort communities at DCN institutions; Advocate for data curator’s role to researchers and cohort community.
3. Develop outreach plan for non-member institutions. Engage HBCU, HSI, TCU and other library alliances by leveraging our individual locales & approach with humility.
4. Reach professional associations, special interest groups with a focus on racial justice.

DCN highlight #2

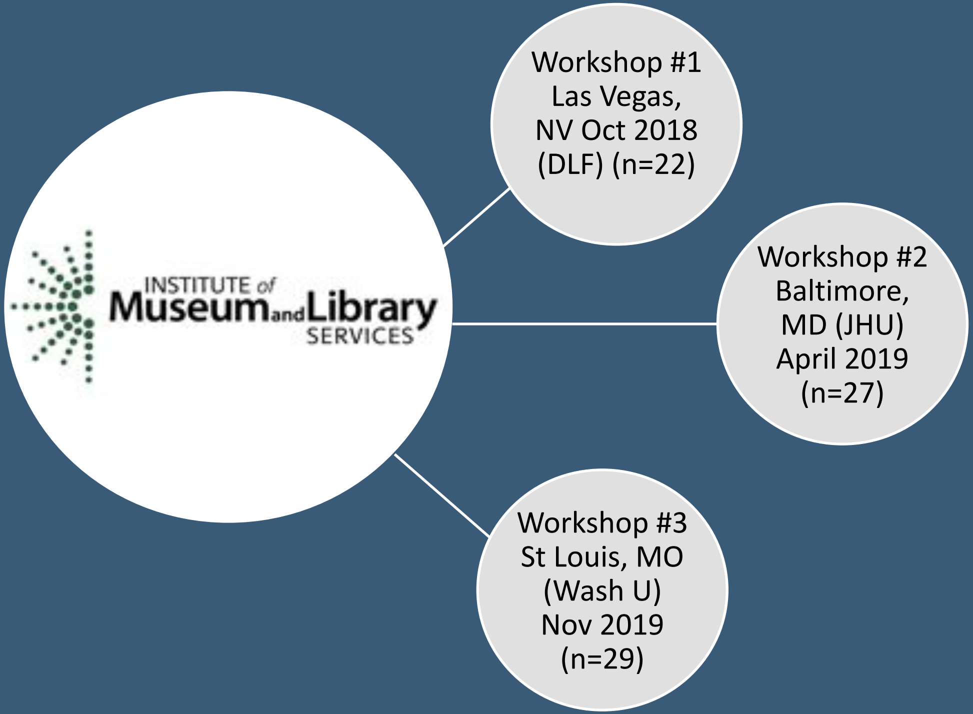
Thriving community of practice for data curation

DCN Workshops





DCN Primers



INSTITUTE of Museum and Library SERVICES

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

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

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

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

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

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

ITHAKA S+R

Data communities

A data community is a fluid and informal network of researchers who share and use a certain type of data.

Most (but not all) data communities are facilitated through information technology

A data community spans disciplines

Cooper, D., & Springer, R. (2019, May 13). Data Communities: A New Model for Supporting STEM Data Sharing. https://doi.org/10.18665/sr.311396. Slide credits: Danielle Cooper.

What do data communities need?

- Help building or identifying existing repository infrastructure
- Technical and policy advice on metadata, vocabularies, preservation, privacy, etc.
- Guidance and advocacy for achieving organizational and financial sustainability
- Help getting the word out to researchers who might be interested in getting involved

Examples of Emergent Data Communities

Stable Isotopes
Researchers who routinely use stable isotope data as part of their work with organic and inorganic samples.

As they continue to build out the IsoBank repository to share data, they need to support in developing **effective metadata framework** to ensure the data can be used across a variety of disciplines.

Phosphorus Sustainability
A growing community of researchers focused on advancing phosphorus sustainability through materials informatics.

They are exploring building a data repository and need help understanding how to **accommodate many heterogeneous data types**, including sensitive industrial data.

Water Treatment
A group of researchers is collecting pre and post treatment data around the removal of a dam on the Maple River in Pellston, Michigan.

As they work to improve their project management, they need help **operationalizing data collection practices** to ensure that it is optimized for data sharing within and beyond their locality.

Using the workshop model to incubate data communities

ITHAKA S+R

DATA CURATION NETWORK

Supported by the National Science Foundation grant No.2103433: Leveraging data communities to advance open science

Pairing Data Communities with Data Curators

14 teams of researchers representing their data community

Matched with 14 data curators with specialized expertise

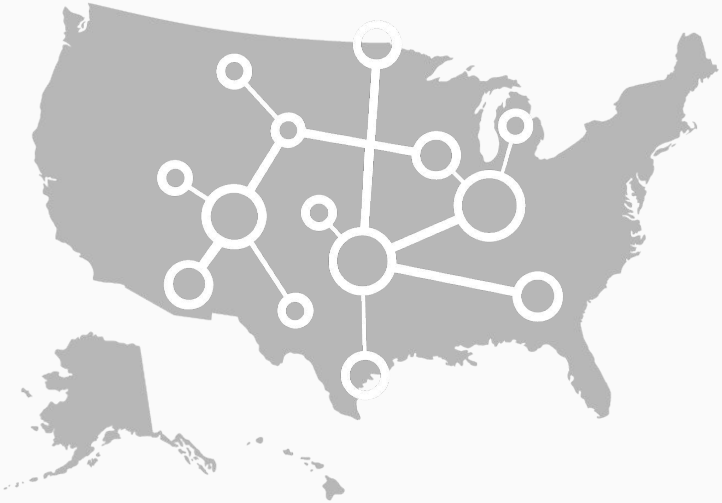
Addressing shared challenges: data formats, accessibility, incentives, engagement, storage, security, ethics, copyright, and more...

DCN highlight #3

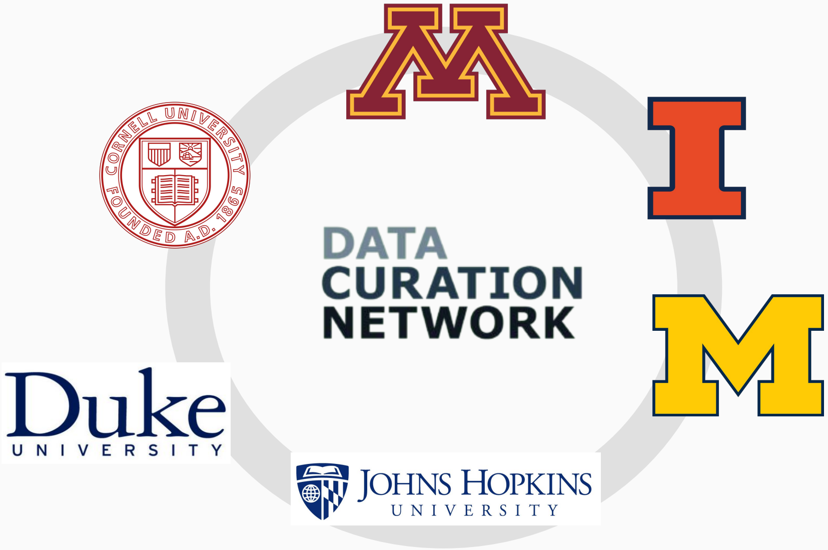
Unique platform for exploration and research

What is the Value of Curation?

Perspective of repositories *doing* curation



Perspective of depositors *receiving* curation

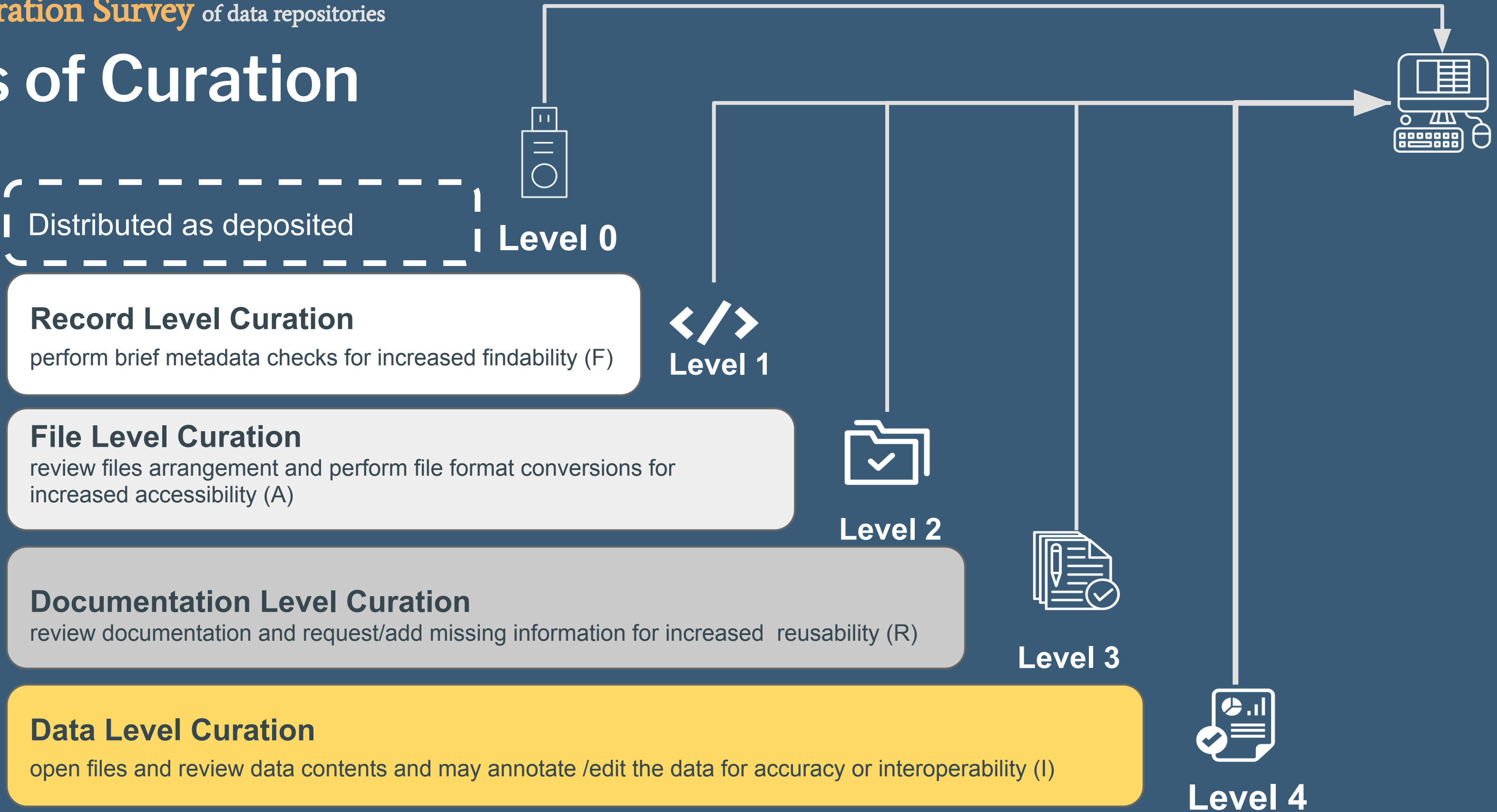


Research Questions

1. What **level(s) of data curation** do data repositories provide?
2. Which aspects of curation **add the most value**?
3. What are the **impacts** of data curation? And what are their importance to the **data sharing** process?
4. Does the value-add of data curation **outweigh the effort and cost**?

Value of Curation Survey of data repositories

Levels of Curation



Distributed as deposited

Level 0

Record Level Curation
perform brief metadata checks for increased findability (F)

Level 1

File Level Curation
review files arrangement and perform file format conversions for increased accessibility (A)

Level 2

Documentation Level Curation
review documentation and request/add missing information for increased reusability (R)

Level 3

Data Level Curation
open files and review data contents and may annotate /edit the data for accuracy or interoperability (I)

Level 4

Methodology

- US-based data repositories
- Non-probabilistic sampling
  - Recruitment: multiple listservs
  - Risk of bias and data skewness towards higher level of curation
  - Limited generalizability
- Open for 3 weeks (Jan. 2021)

Value of Curation Survey of data repositories

Demographics

34	Directors
52	Staff
4	Depositors
5	Users

----

95 respondents

31	Disciplinary
25	Institutional
3	Generalist

----

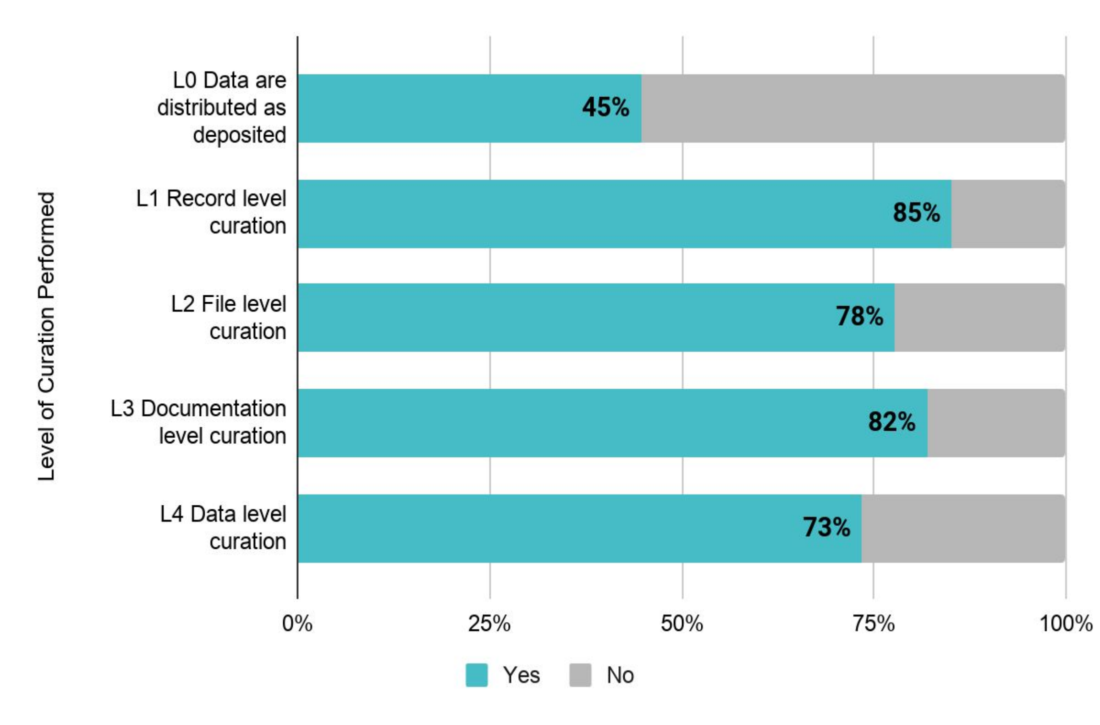
59 repositories

US Repositories from 23 states

- 10 CoreTrustSeal
- 11 DCN

Results - Levels of Curation Performed

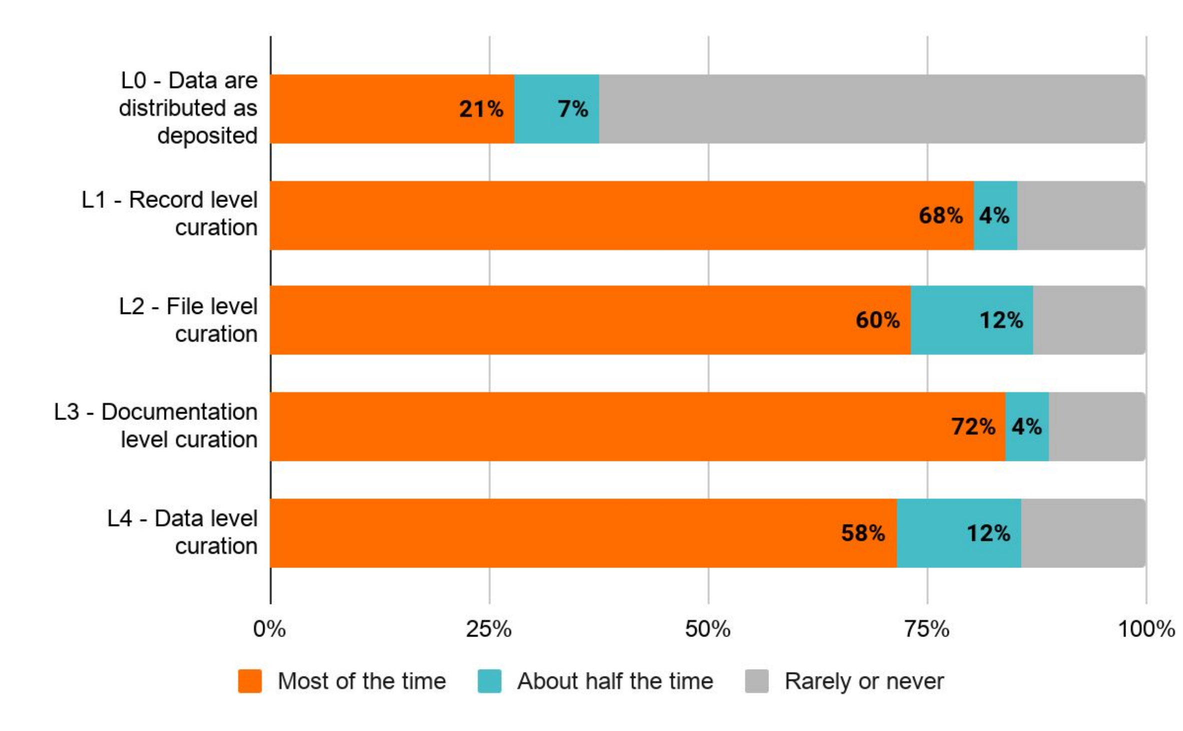
L1 and L3 are the most performed curation levels



*Multiple responses allowed

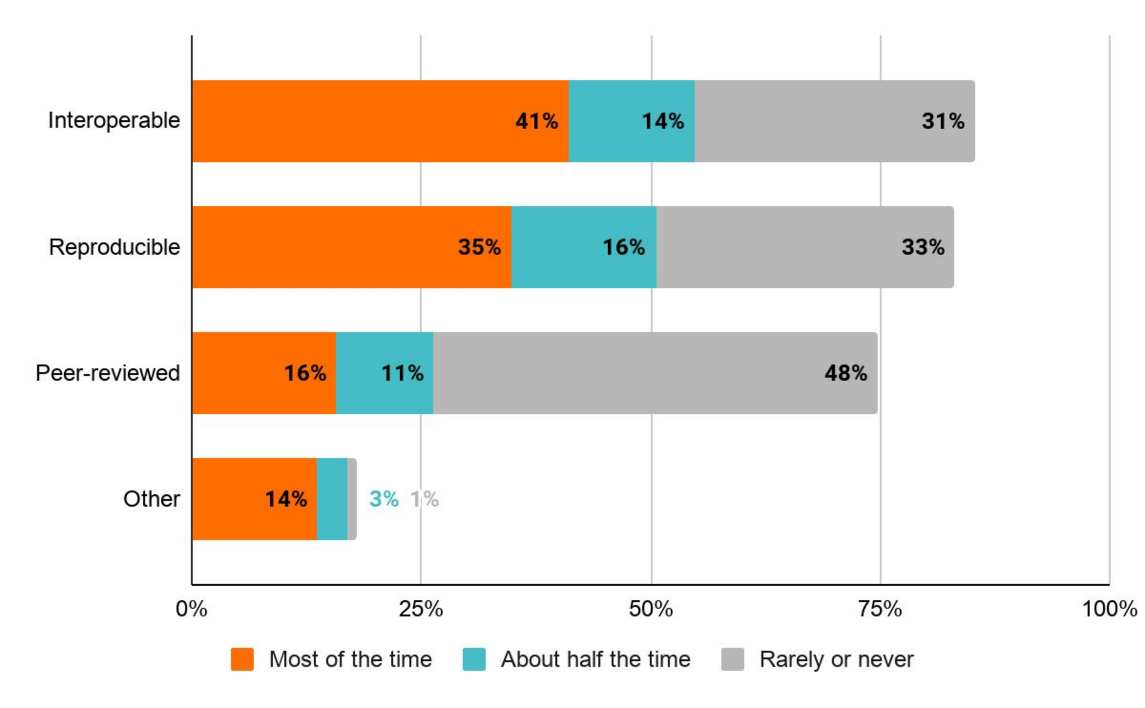
Value of Curation Survey of data repositories

Results - Levels of Curation (Frequency)



*multiple responses allowed.

Results - Additional standards for curation



Interoperable: data are harmonized/normalized/enhanced to provide a rich resource for meta-level investigations across datasets.

Reproducible: data and code reproduce the results presented in an associated scholarly output.

Peer-reviewed: a researcher from the same domain reviews the data.

Other: Discoverable, author-reviewed, reusable, findable, accessible, validated against schema.

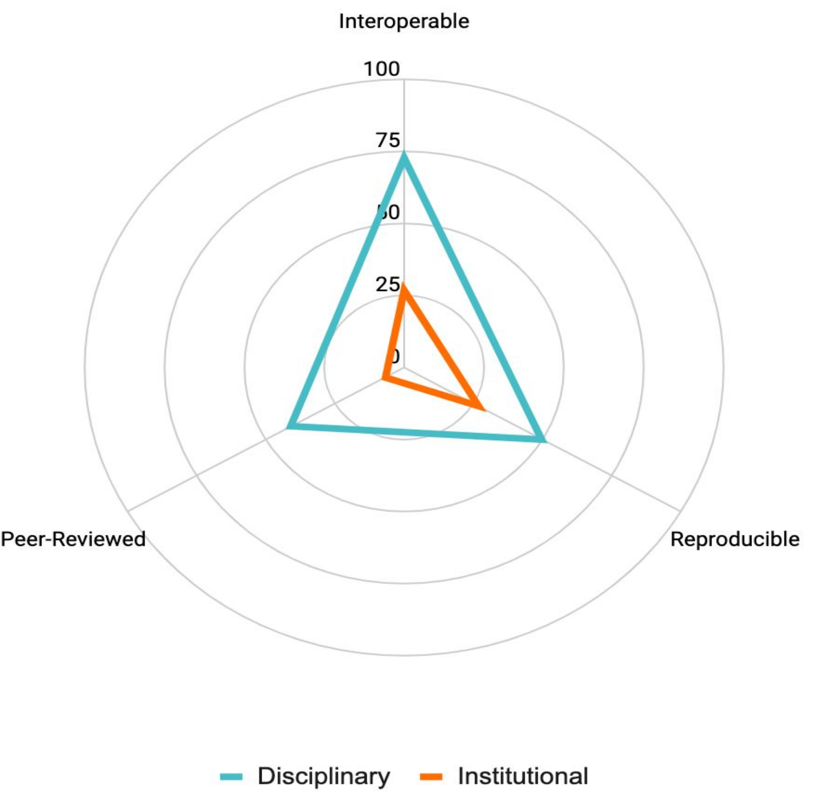
*multiple responses allowed.

Results - Additional Standards vs. Repo Type

Does this repository aim to ensure that data are interoperable, reproducible, peer-reviewed?

RepoType/Actions	Disciplinary	Institutional
Interoperable	73%	27%
Reproducible	50%	27%
Peer-Reviewed	41%	7%

Relative to total "most often" cases for each repo type.



Value of Curation Survey of data repositories

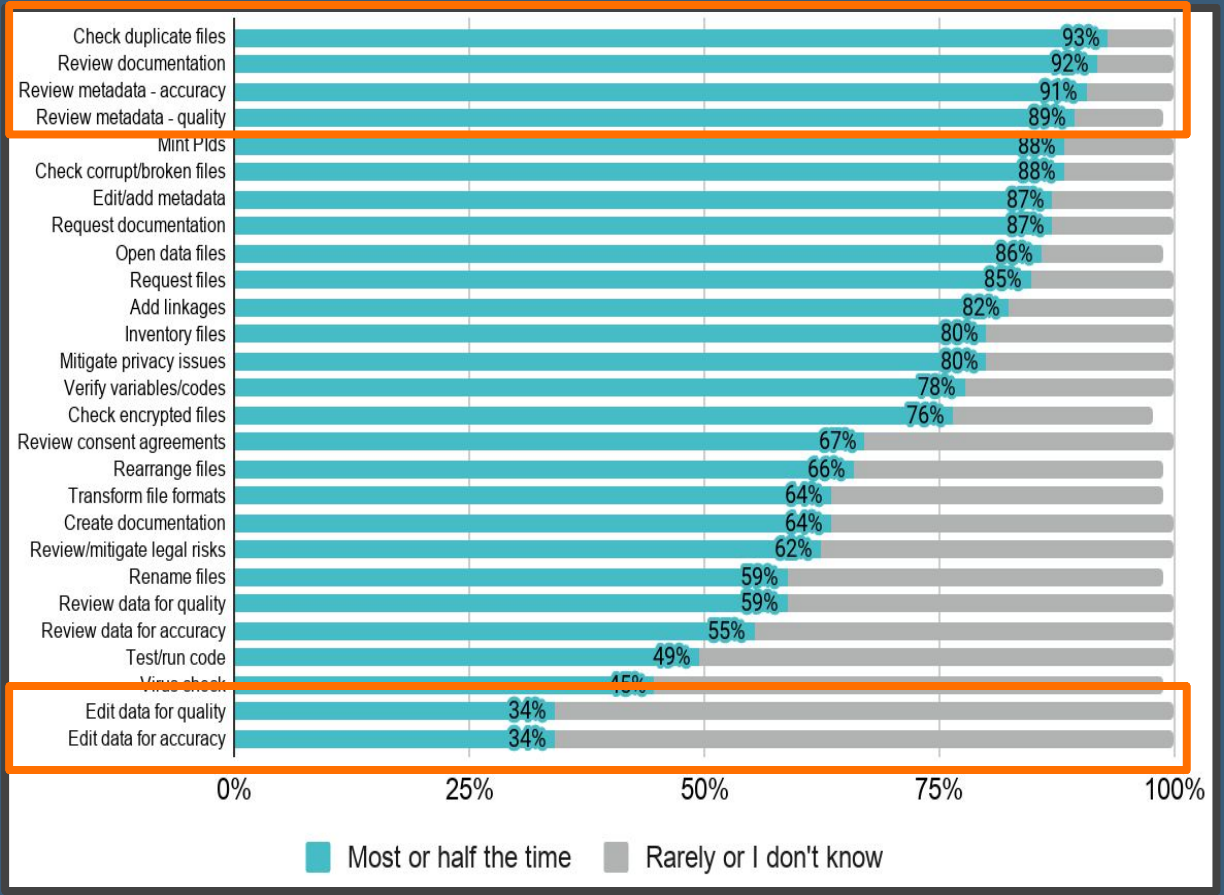
Results - Curation Actions Performed

~90%

- Check for duplicate files
- Review documentation and metadata for accuracy and quality

~35%

- Edit data for quality and accuracy



Results - Curation Actions by Repo Type

Most curation actions at data level are more often performed by disciplinary repositories

- Edit/Review data for Quality
- Edit/Review data for Accuracy
- Check for Locked/Encrypted Files
- Rearrange Files
- Rename Files
- Transform files to alternative formats

Disciplinary
Institutional

$p \leq 0.05$

Value of Curation Survey of data repositories

Results - Perceived Value & Impact of Curation

Top ranked impacts of data curation (out of 14)

The ability for others to...

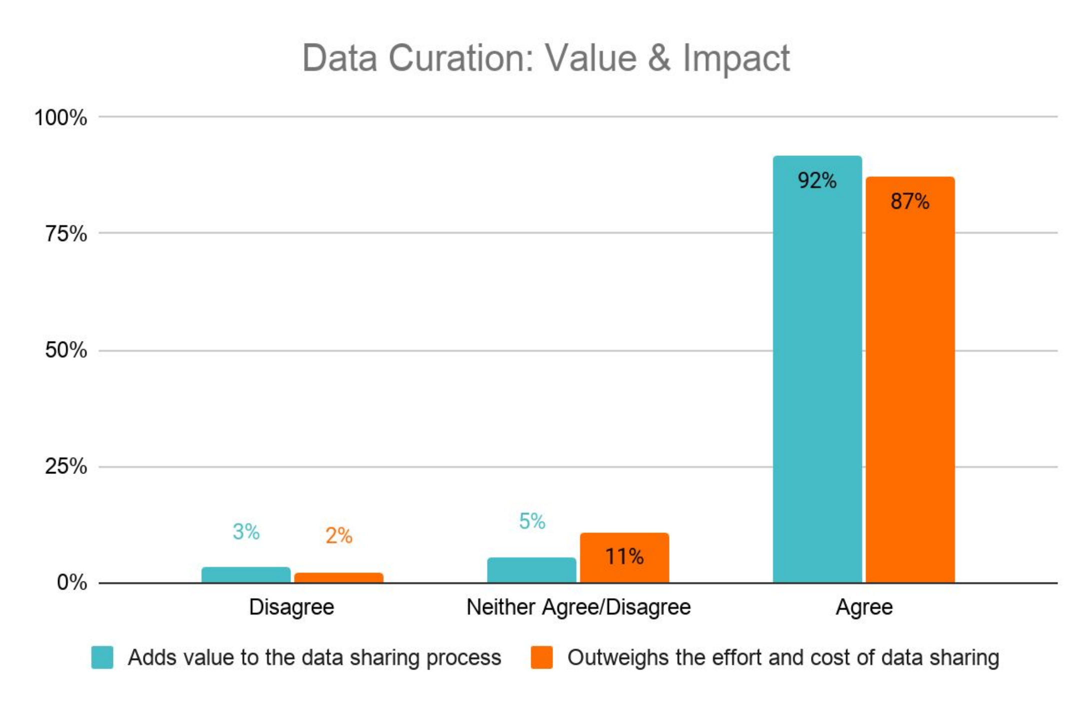
#1 **find** the data

#2 **understand** the data

#3 **use** the data

#4 **access** the data

#5 **preserve** the data



Some Takeaways

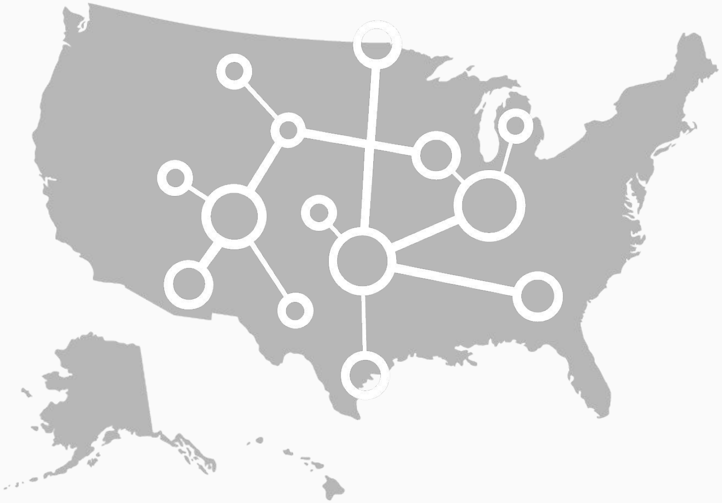
- Most respondents (73%) reported there repository providing **“level 4” (data level) curation**
- Most agreed that curation **adds value**, specifically, impacting ability for others to **find/understand/use** the data.

However,

- **Perceptions** are not equivalent to attitudes, nor actual behaviors
- Responses for the same repository did not always agree: Is there such a thing as an **authoritative answer**?

What is the Value of Curation?

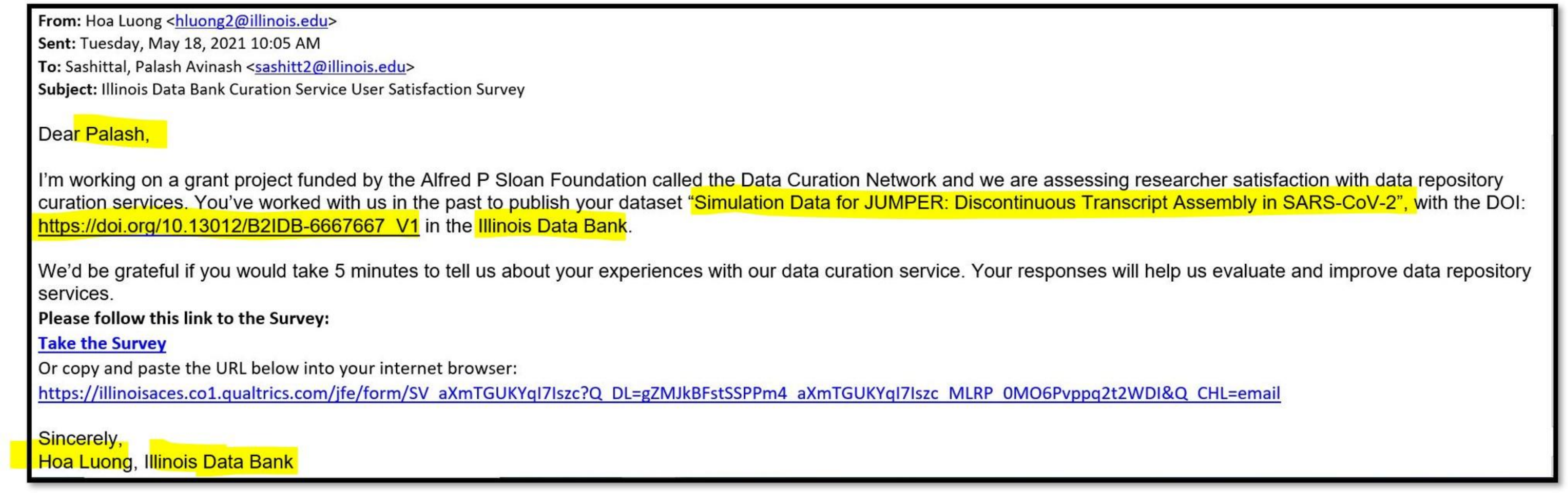
Perspective of repositories *doing* curation



Perspective of depositors *receiving* curation

CORNELL UNIVERSITY FOUNDED A.D. 1865
Duke UNIVERSITY
JOHNS HOPKINS UNIVERSITY

Invitation email (Apr - Jun 2021)



From: Hoa Luong <hluong2@illinois.edu>
Sent: Tuesday, May 18, 2021 10:05 AM
To: Sashittal, Palash Avinash <sashitt2@illinois.edu>
Subject: Illinois Data Bank Curation Service User Satisfaction Survey

Dear Palash,

I'm working on a grant project funded by the Alfred P Sloan Foundation called the Data Curation Network and we are assessing researcher satisfaction with data repository curation services. You've worked with us in the past to publish your dataset "Simulation Data for JUMPER: Discontinuous Transcript Assembly in SARS-CoV-2", with the DOI: https://doi.org/10.13012/B2IDB-6667667_V1 in the Illinois Data Bank.

We'd be grateful if you would take 5 minutes to tell us about your experiences with our data curation service. Your responses will help us evaluate and improve data repository services.

Please follow this link to the Survey:

Take the Survey

Or copy and paste the URL below into your internet browser:

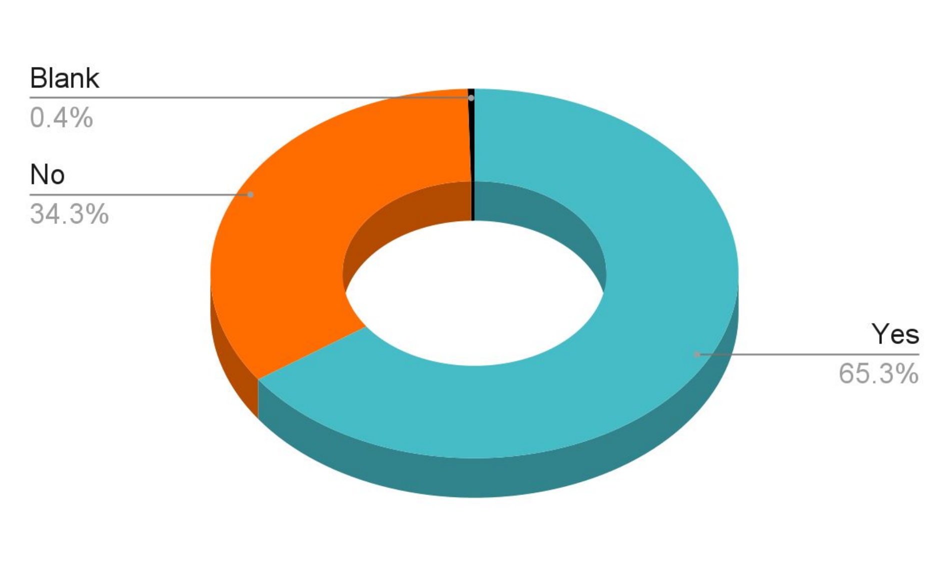
https://illinoisaces.co1.qualtrics.com/jfe/form/SV_aXmTGUKYqI7Iszc?Q_DL=gZMJkBFstSSPPm4_aXmTGUKYqI7Iszc_MLRP_0MO6Pvppq2t2WDI&Q_CHL=email

Sincerely,
Hoa Luong, Illinois Data Bank

Survey Response rate = 40% (n=227)

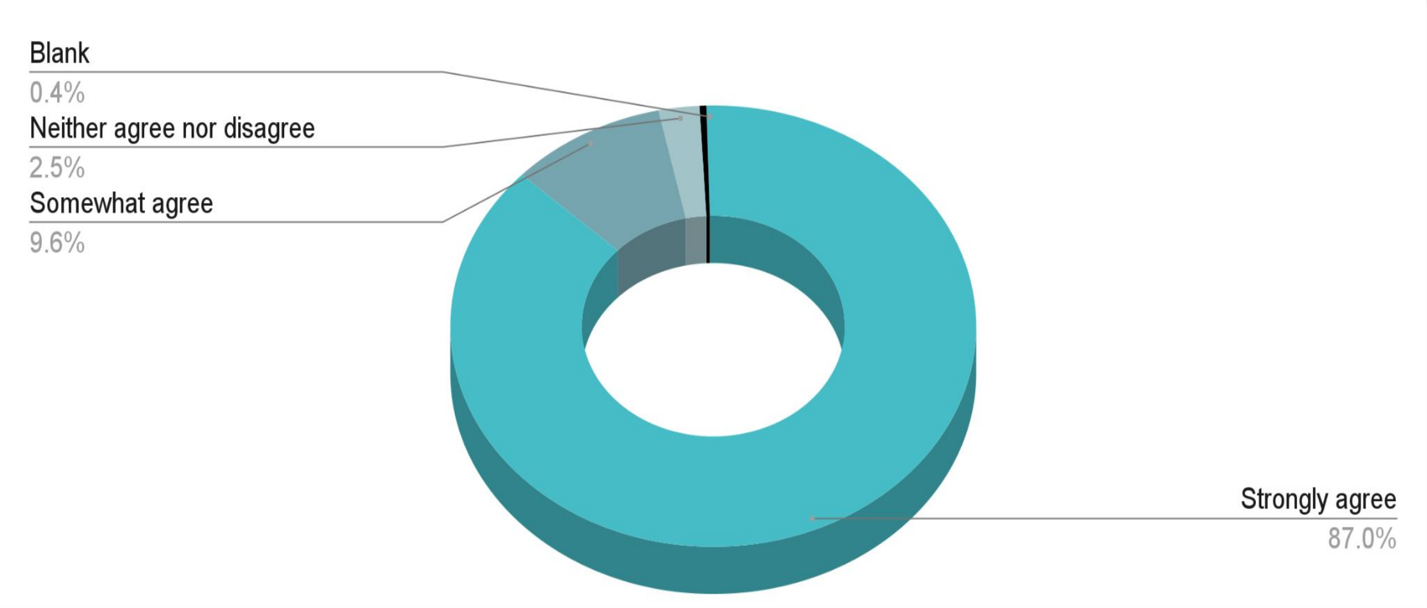
	Date	Distribution Count	Response Count	Response Rate
Minnesota	4-26-21	197	82	42%
Cornell	4-26-21	34	18	53%
Michigan	5-11-21	130	63	48%
Duke	5-05-21	54	19	35%
Illinois	5-18-21	121	45	37%
Johns Hopkins	6-17-21	32	12	38%
Total		**568**	**227**	**40%**

Did you expect repository staff to curate your dataset? N = 227

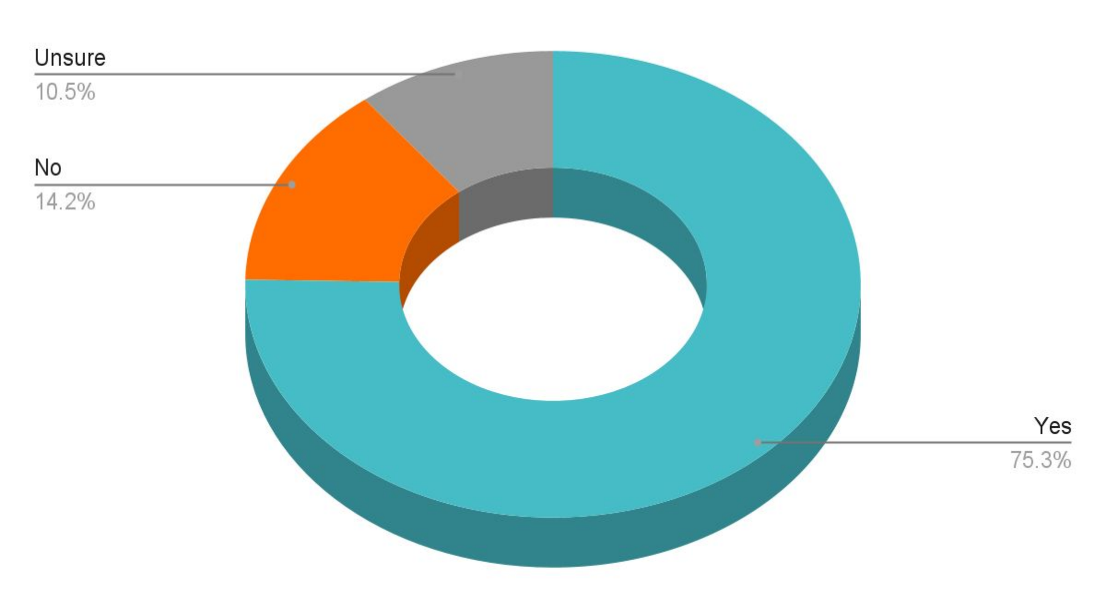


Value of Curation Survey of end-user depositors

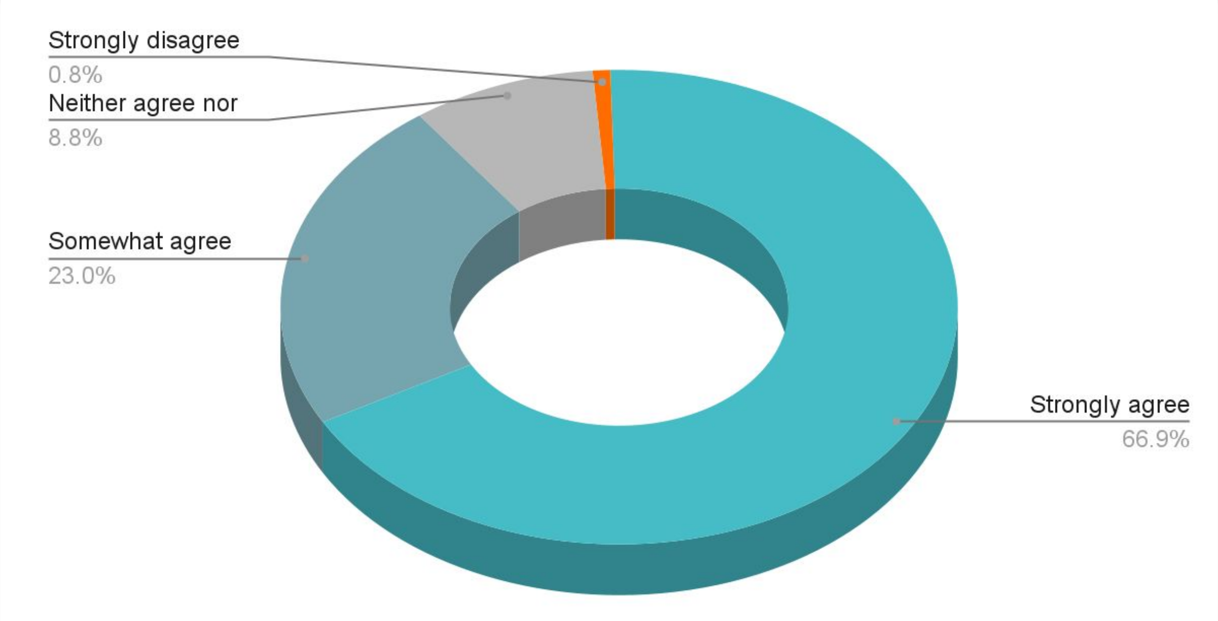
I was satisfied with the curatorial review my data received. N = 227



Were any changes made to your data submission due to the curatorial review? N = 227

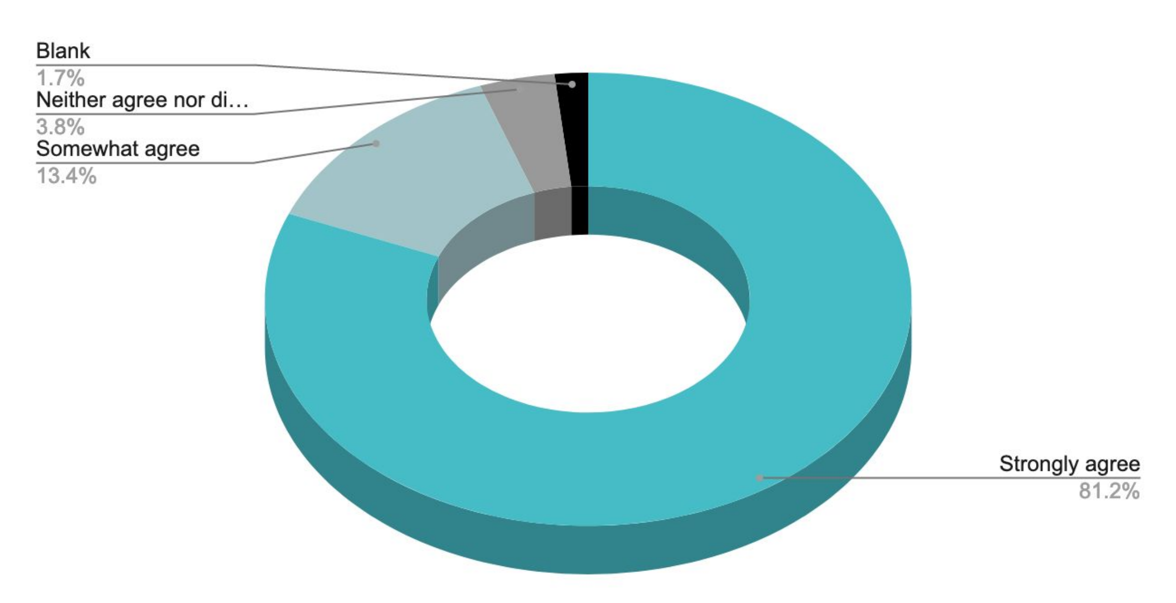


Due to the curation process I felt more confident sharing my data. (N = 227)



Value of Curation Survey of end-user depositors

Data curation by this repository adds value to the data sharing process. (N = 227)



What is the most "value-add" curation action taken by this repository? (N = 172)

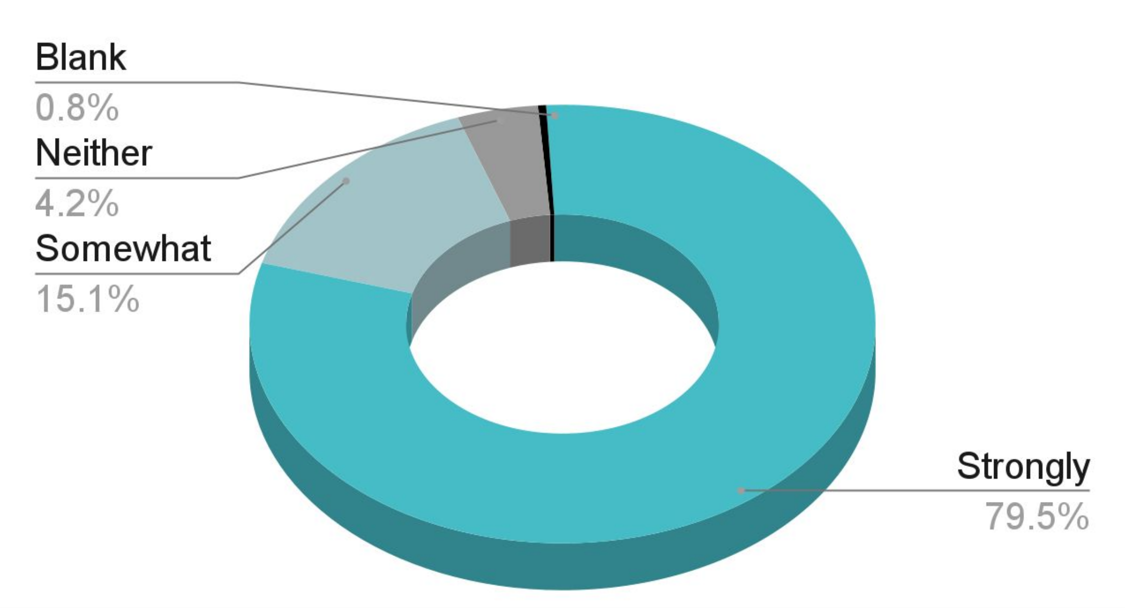
caught mistakes - increased dataset quality!

Helping me prepare the data/metadata for long term access. The curator's expertise in this area was very helpful as I was not as familiar with what metadata was required to ensure easy, long term utility/access of my dataset.

The curation process makes the data more accessible and readable for other scientists. This amplifies the impact of our research.

Value of Curation Survey of end-user depositors

The impact that data curation has on the data sharing process is worth the effort. (N= 227)



Value of Curation Survey of end-user depositors

> I love it! I want to share more data this way! Get ready for some cool stuff!

> I try HARD to make my data good before it ever gets to the repository and every single time there's been curation, there's always been something that I've missed. I'm very grateful for such careful and helpful curation. It definitely increases my comfort and even pride in the dataset.

> So helpful! I was expecting nothing, but got a lot of personal help and kindness. Much appreciated.

DCN highlight #3...a bit more

Unique platform for exploration and research

Realities of Academic Data Sharing (RADS)

ASSOCIATION OF RESEARCH LIBRARIES

PI: **Cynthia Hudson-Vitale,** Director of Scholars and Scholarship, Association of Research Libraries (PI)

Advisors: COGR, AAU, APLU, university offices

UNIVERSITY OF MINNESOTA

Duke UNIVERSITY

MICHIGAN

DATA CURATION NETWORK

Cornell University

Washington University in St. Louis

VIRGINIA TECH

Realities of Academic Data Sharing research has been generously funded by NSF EAGAR grant #2135874: Completing the Lifecycle: Developing Evidence Based Models of Research Data Sharing

Realities of Academic Data Sharing (RADS)

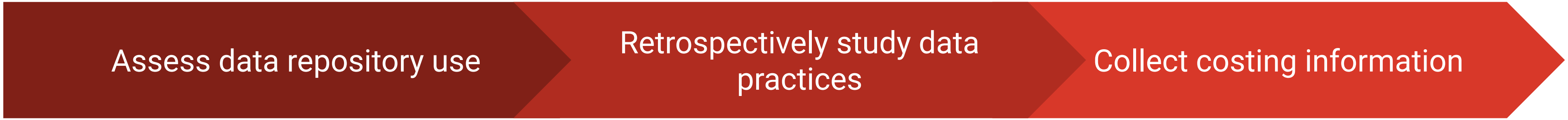
Research Questions

- Where are funded researchers sharing their data and what is the quality of that metadata?
- How are researchers making decisions about why and how to share research data?
- What is the cost to the institution to implement federally mandated public access to research data policies?

Realities of Academic Data Sharing research has been generously funded by NSF EAGAR grant #2135874: Completing the Lifecycle: Developing Evidence Based Models of Research Data Sharing

Realities of Academic Data Sharing (RADS)

Work Plan



Assess data repository use

Retrospectively study data practices

Collect costing information

5 specific disciplines:

- **environmental science**
- **materials science**
- **psychology**
- **biomedical sciences**
- **physics**

Realities of Academic Data Sharing research has been generously funded by NSF EAGAR grant #2135874: Completing the Lifecycle: Developing Evidence Based Models of Research Data Sharing

Realities of Academic Data Sharing (RADS)

Expected Outcomes

- Data and information about where funded researchers are sharing their research data – and a workflow for other institutions
- Models for institutional support for public access to research data
- Disciplinary case studies and decision-making factors influencing public access to research data
- $ Data, information, and case studies on costs for public access to research data and possible differentiators to those expenses

NSF

Realities of Academic Data Sharing research has been generously funded by NSF EAGAR grant #2135874: Completing the Lifecycle: Developing Evidence Based Models of Research Data Sharing

DCN highlight #4

Sustainable organization that advocates for the profession

Toward Sustainability

Planning Phase Grant

Implementation Phase Grant

Sustaining the DCN

Shared curation

6 partners → 8 partners → 10 partners → 12 partners → Steady growth

2016

Research and planning

- Researcher focus groups “Most important curation activities”
- Pilots
- Comparison of curation workflows
- Fiscal models

2017

Model published

(2017) A cross-institutional staffing model for curating research data. IJDC 13(1)

2018

Launching the Network

- Hired 1FTE coordinator
- Trained first curator cohort
- Put technology/ workflows into place.

2019

Data Curation Network goes live Jan 1, 2019 with 10 partner institutions!

2020

Assess and Adjust

- Special interest groups
- Research value of curation
- Sustainability planning

2021

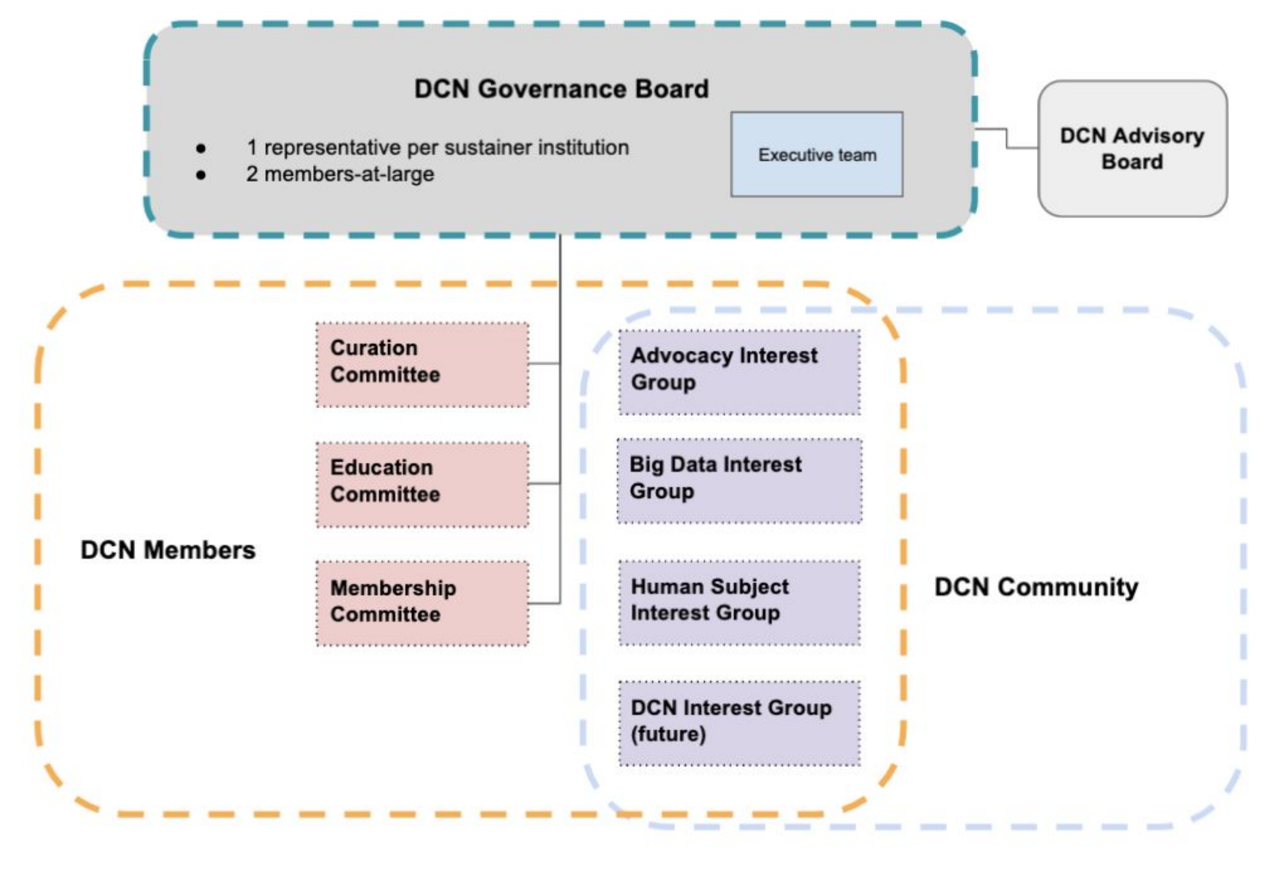
Transition to membership model

- Establish membership model
- Launch Governance Board

Governance Model

- Shared operating costs ~ (salary, annual event costs, travel to AHM)
- University of Minnesota serves as fiscal home (e.g., HR, staff benefits, legal, taxes, technology admin, etc.)

University of Minnesota



DCN Governance Board

- 1 representative per sustainer institution
- 2 members-at-large

Executive team

DCN Advisory Board

DCN Members

Curation Committee

Education Committee

Membership Committee

Advocacy Interest Group

Big Data Interest Group

Human Subject Interest Group

DCN Interest Group (future)

DCN Community

Membership Model

	Sponsor	Ambassador	Member	Sustainer
	$1,000	$2,500	Beta testing	$10,000 + in-kind
● Logo on website ● Invited to present virtually at the AHM	✔	✔	✔	✔
● Host a DCN workshop for regional attendees		✔	✔	✔
● Join DCN community of practice ● Participate in special interest groups and primer cohorts ● Attend the AHM		✔	✔	✔
● Travel support to attend the AHM ● Receive up to 200 hours of curation service ● Representation in the DCN Governance and Advisory Board(s)			?	✔

Get involved

- Recommend your institution join the DCN as a member (applications open Mar/Apr 2022)
- Invite DCN instructors to teach a “Specialized data curation” workshop in your local region
- Special interest groups are open to all

Takeaways and Discussion

- Collaboration across institutions makes data sharing and reuse more successful and effective
- Shared outcomes drive collaboration - - for us, that means better research data (more understandable, reusable, and always ethically shared) regardless of institutional origin
- What are your challenges? How can we help?

Please contact us:

dcn-team@googlegroups.com

http://datacurationnetwork.org

Ethical. Reusable. Better.

Thank You!

DATA
CURATION
NETWORK