DATA CURATION NETWORK

Data Curation Fundamentals Workshop

DATA CURATION NETWORK

Mission

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

Ethical. Reusable. Better.

Code of Conduct

The Data Curation Workshop is dedicated to providing a harassment-free experience for everyone, regardless of gender, gender identity and expression, sexual orientation, disability, physical appearance, body size, race, age or religion. We do not tolerate harassment of participants in any form. Sexual language and imagery is not appropriate for any conference venue, including talks. Participants violating these rules may be sanctioned or expelled from the workshop at the discretion of the conference organizers.

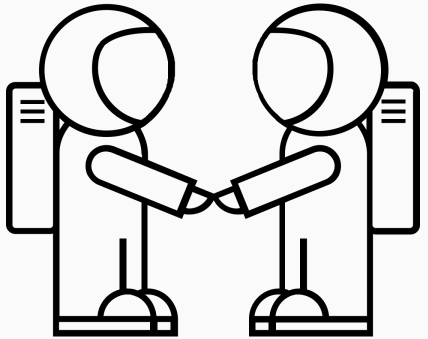
Our code of conduct and anti-harassment policy can be found at:

https://datacurationnetwork.org/about/code-of-conduct/

Code of Conduct (cont)

Examples of behavior that contribute to creating a positive environment include:

- Using welcoming and inclusive language
- Being respectful of differing viewpoints and experiences
- Gracefully accepting constructive criticism
- Focusing on what is best for the community
- Showing empathy towards other community members



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 nationwide.
3. Meet like-minded colleagues who are interested in building and extending curation practices at their institutions.

The Value of Curation

Introduction to the workshop

Reasons for data sharing


Open data movement

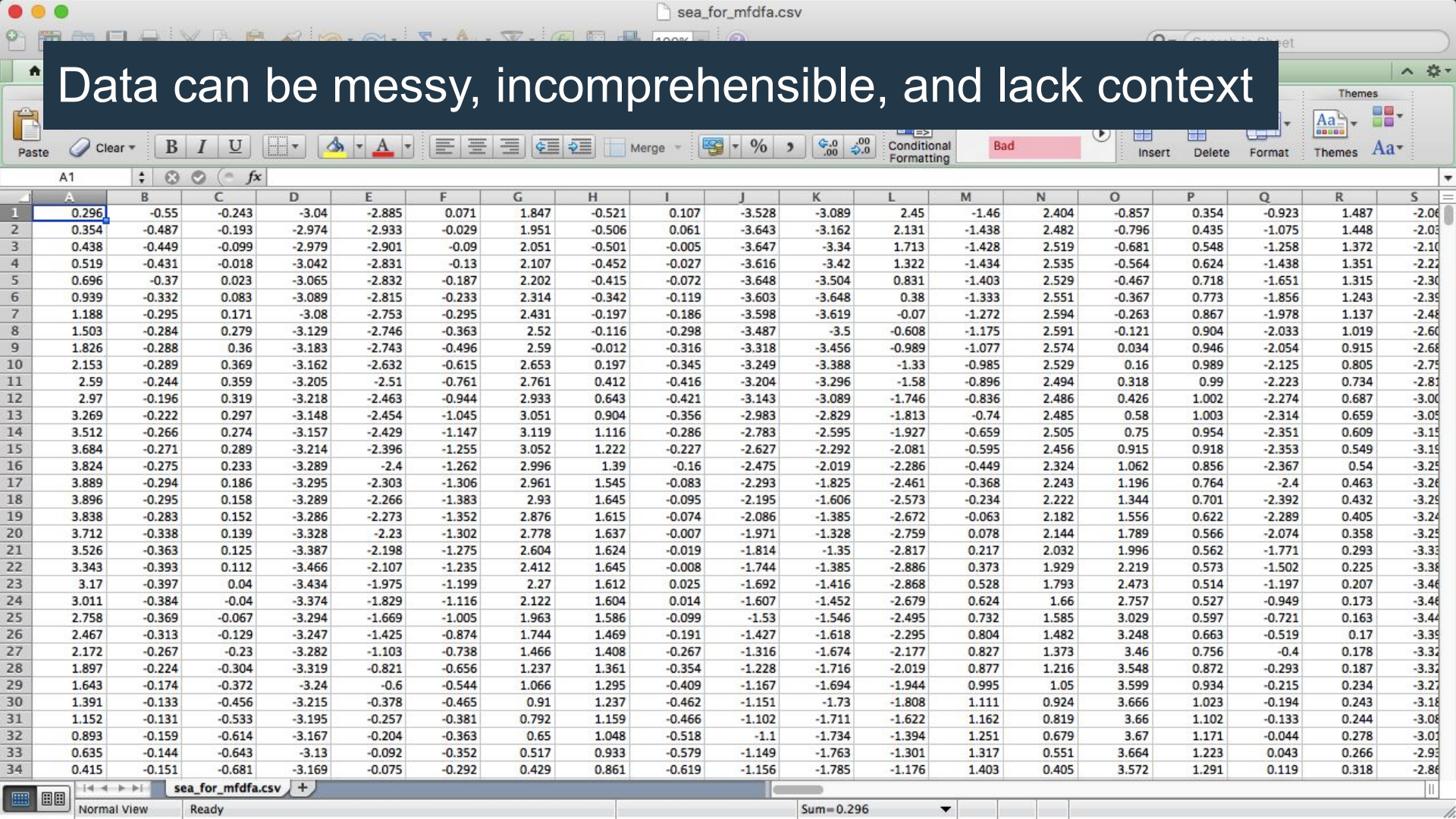

Reproducibility


Funding agency mandates


Publisher policies

Research data have value beyond their original purpose.

But….


sea_for_mfdfa.csv
Data can be messy, incomprehensible, and lack context
Paste
Clear
Merge
Conditional Formatting
Bad
Insert
Delete
Format
Themes
A1
sea_for_mfdfa.csv
Normal View
Ready
Sum=0.296

Digital file formats are constantly at risk


Most data never leave a laptop ⇒ benign neglect
Rainbow 100
CP/M-86/80 Operating System
190

What is data curation?

“Data curation is the active and ongoing management of data through its lifecycle and interest and usefulness to scholarship, science, and education; curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time.”

- University of Illinois’ Graduate School of Library and Information Science

What is data curation?

“Data curation is akin to work performed by an art or museum curator. Through the curation process, data are organized, described, cleaned, enhanced, and preserved for public use, much like the work done on paintings or rare books to make the works accessible to the public now and in the future.”

- ICPSR

Who assists in curation?

Data curation involves

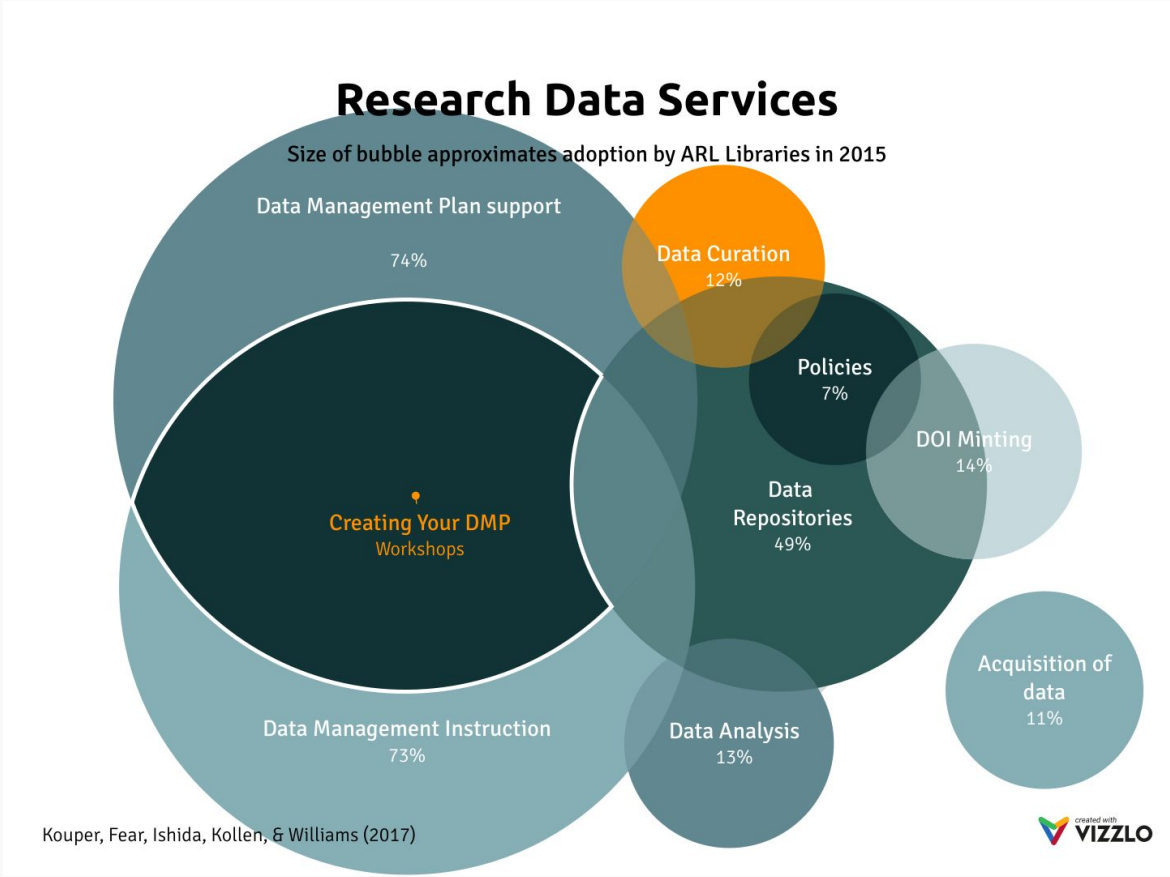
- Automated curation actions taken by **repositories**
- Decisions and actions taken by the **curator**
- Engagement with the **researcher**

Definitions of Data Curation Activities used by the DCN

The value-add

- ★ Easier for fellow scholars and future collaborators to understand
- ★ More likely to be trusted
- ★ The research they represent are more likely to be reproducible
- ★ More likely to be properly cited
- ★ Represent potential cost-savings
- ★ Findable, accessible, interoperable, and reusable (FAIR)

Remember… data curation is just **one piece** of the broader data services landscape



Role of libraries in data curation

Libraries and academic-based data repositories are just one piece of the data repository landscape.

Baker, K. and Duerr, R. (2017). "Data and a diversity of repositories" in *Curating Research Data: A handbook of current practice* (L. R. Johnston, ed.). ACRL press.

TABLE 4.3
Examples of kinds of data repositories found in the United States.

Kind of Repository	Examples
Federally Funded Data Centers	NASA Distributed Active Archives (DAAC), NOAA National Centers for Environmental Information (NCEI), National Snow and Ice Data Center (NSIDC), USGS Earth Resources Observation Systems (EROS) Data Center (EDC)
Federally Funded Research and Development Centers (FFRDC)	National Center for Atmospheric Research (NCAR), Jet Propulsion Lab (JPL), Oak Ridge National Laboratory (ORNL)
National Libraries	National Library of Medicine (NLM), National Agricultural Library (NAL), Library of Congress (LOC)
State and Local Agencies	State geological surveys, County planning offices
Thematic Repository	Long Term Ecological Research Network Information System (LTER NIS), Andrews Forest LTER (AND), National Snow and Ice Data Center (NSIDC), Maria Rogers Oral History Program
Domain Repository	Global Biodiversity Information Facility (GBIF), Inter-university Consortium for Political and Social Research (ICPSR), DataOne, Interdisciplinary Earth Data Alliance (IEDA)
Institutional Repository	Purdue University Research Repository (PURR), Data Repository for the University of Minnesota (DRUM)
Replication Repository	Dryad Digital Repository, Pangaea Data Library
Software Repository	GitHub, SourceForge
Commercial Archives	DigitalGlobe, Aerial photography companies, Resource exploration companies, Figshare
Private Archives	Huntington Library, Getty Research Institute

Data Curation in Practice

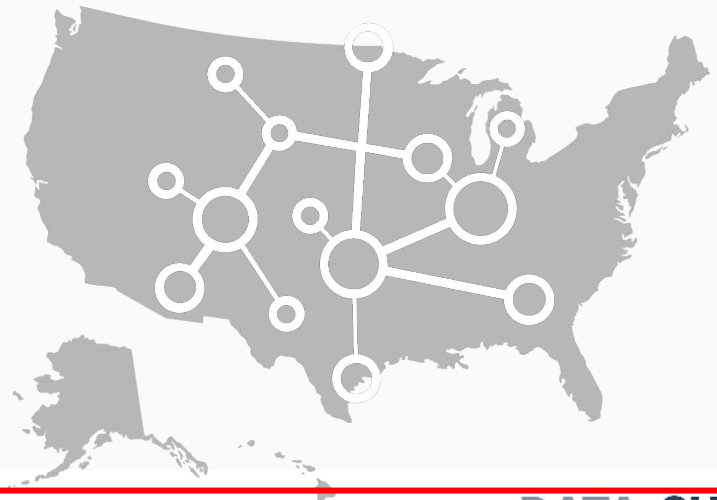
Research performed by the Data Curation Network

DCN Research

- Identification of Data Curation Activities (2016)
- Understanding what data curation activities are important to researchers (2017)
- SPEC Kit for Data Curation Activities in ARL Institutions (2017)
- Value of Curation from the Repository and Depositor Perspective (2021-2022)

What is the Value of Curation?

Perspective of repositories *doing* curation



Perspective of depositors *receiving* curation

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

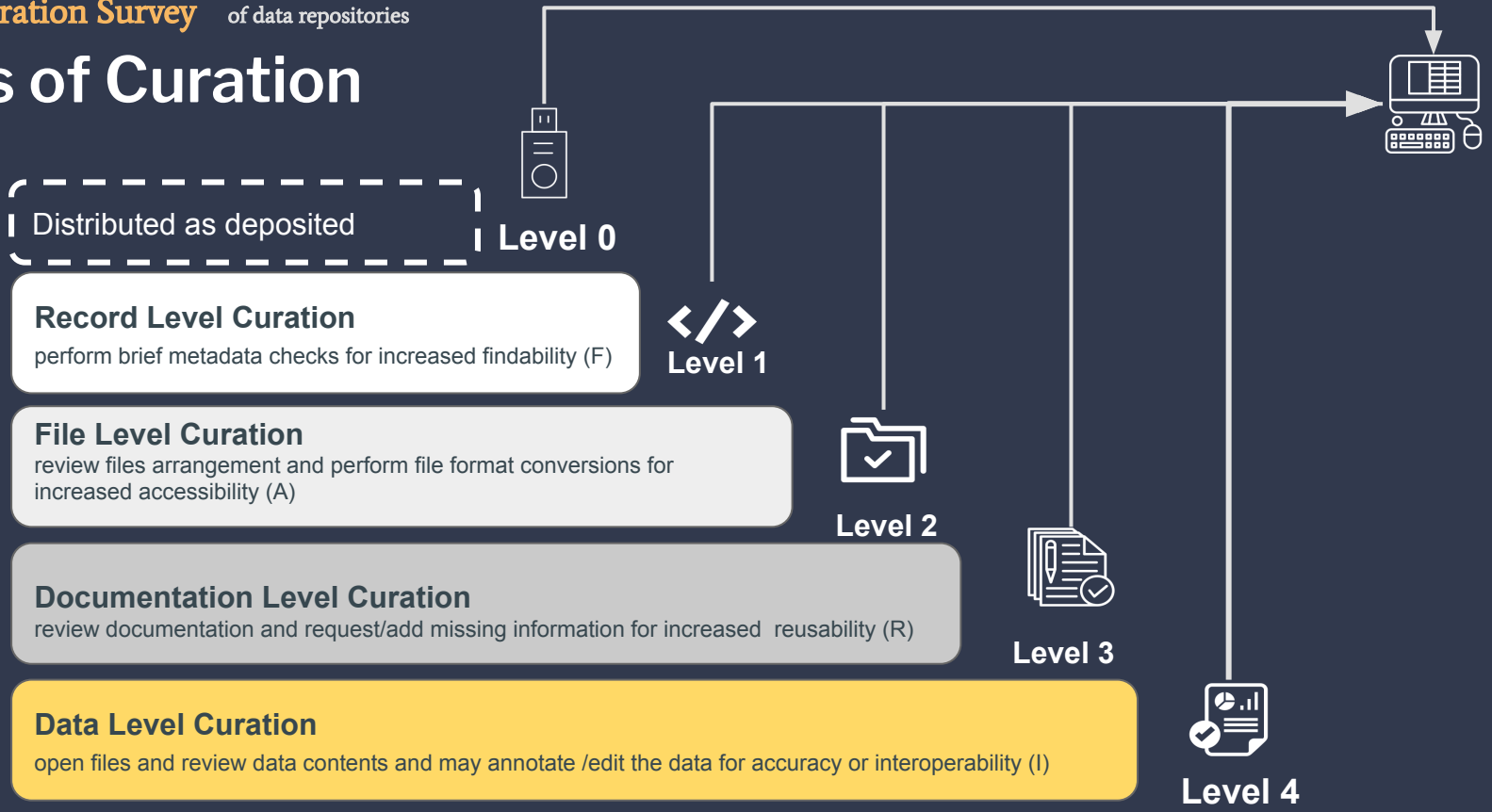
---

59 repositories

US Repositories
from 23 states
- 10 CoreTrustSeal
- 11 DCN

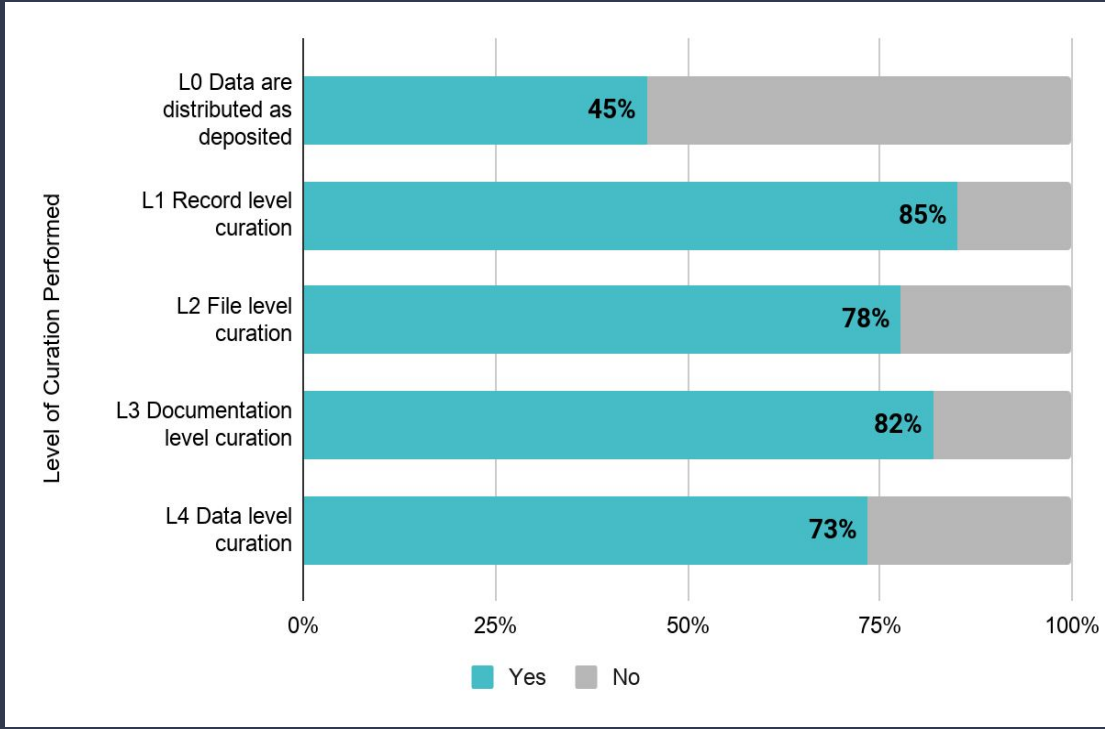
Value of Curation Survey of data repositories

Levels of Curation



Results - Levels of Curation Performed

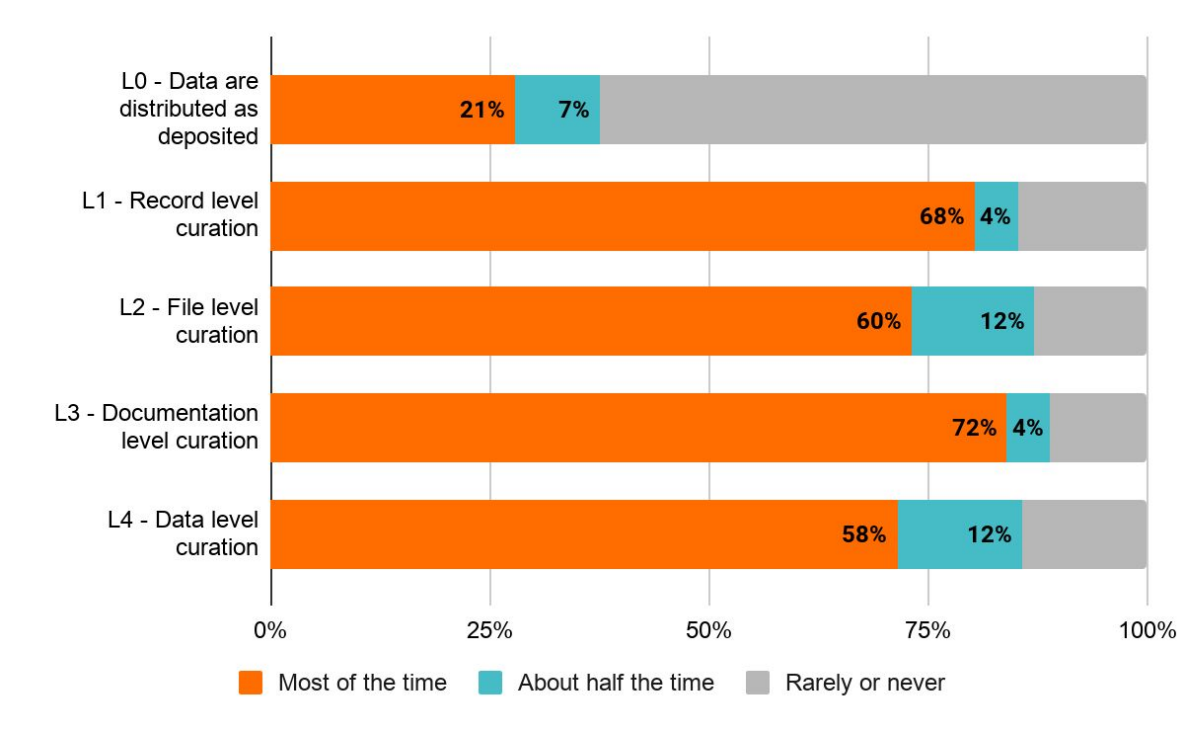
L1 and L3 are the most performed curation levels



*Multiple responses allowed

Value of Curation Survey of data repositories

Results - Levels of Curation (Frequency)



*multiple responses allowed.

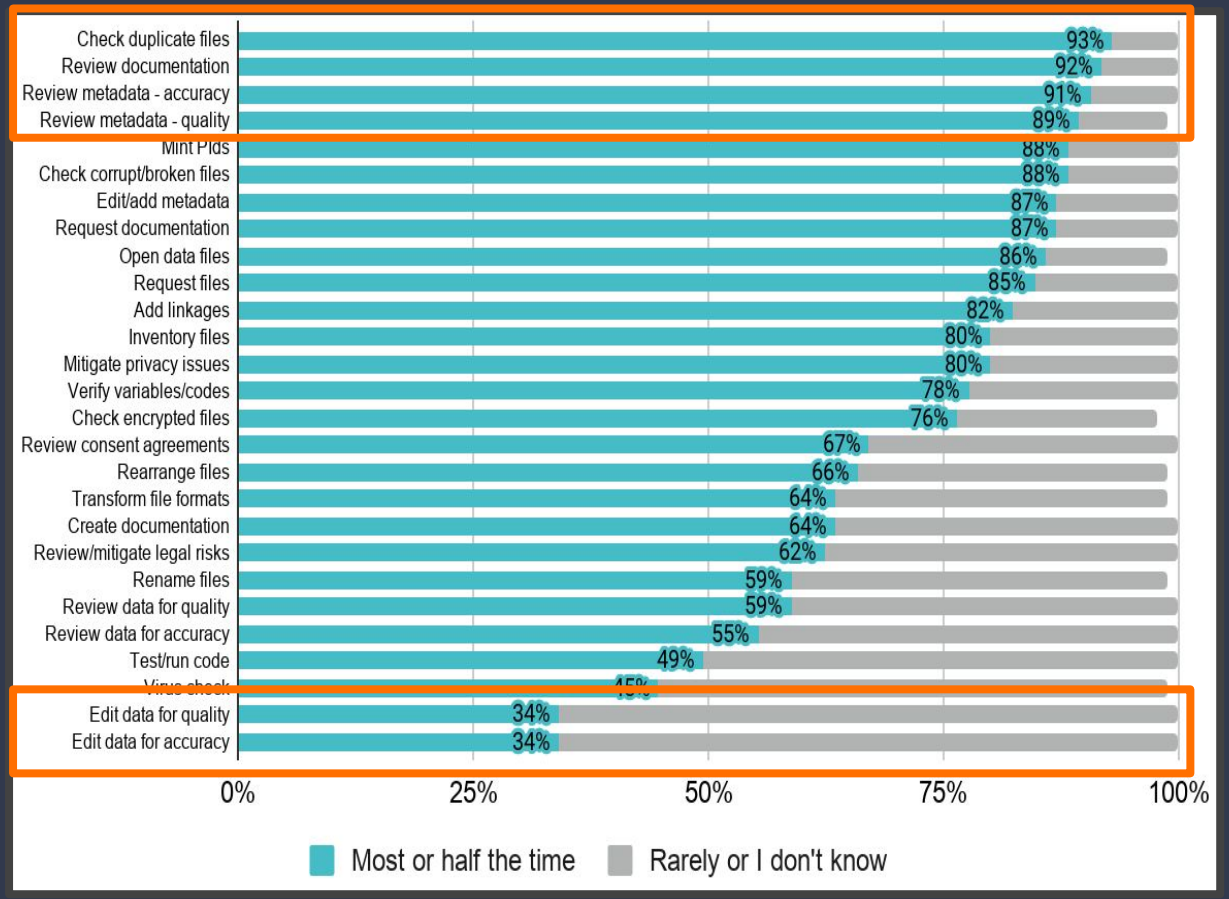
Results - Curation Actions Performed

~90%

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

~35%

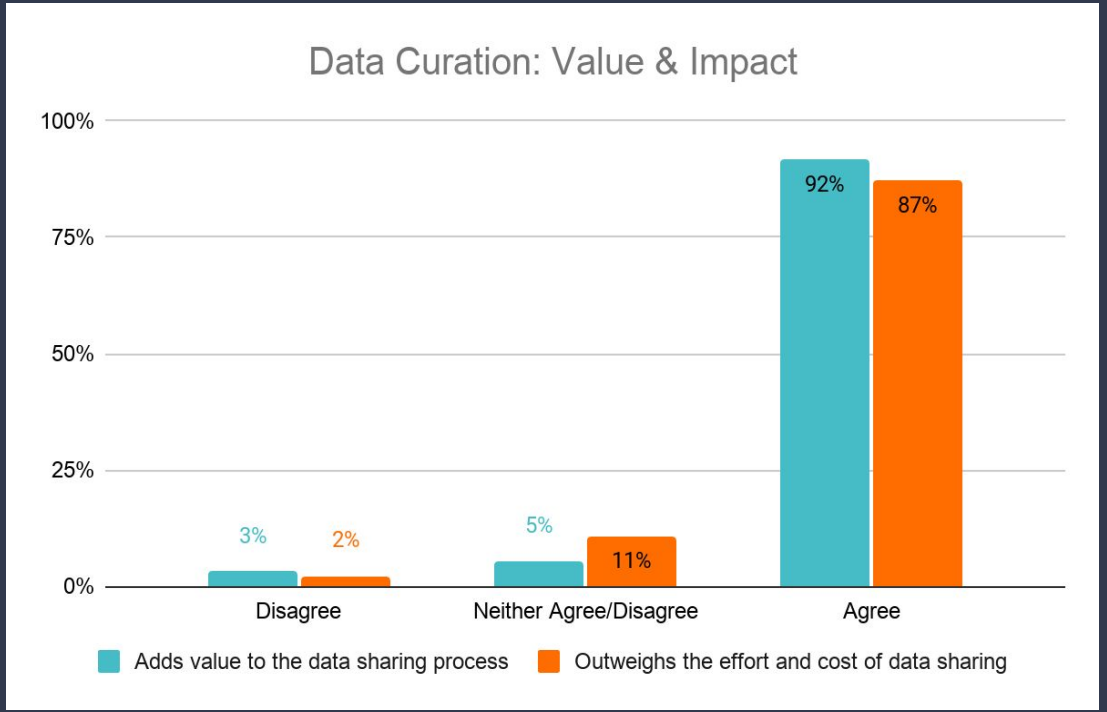
- Edit data for quality and accuracy



Results - Perceived Value & Impact of Curation

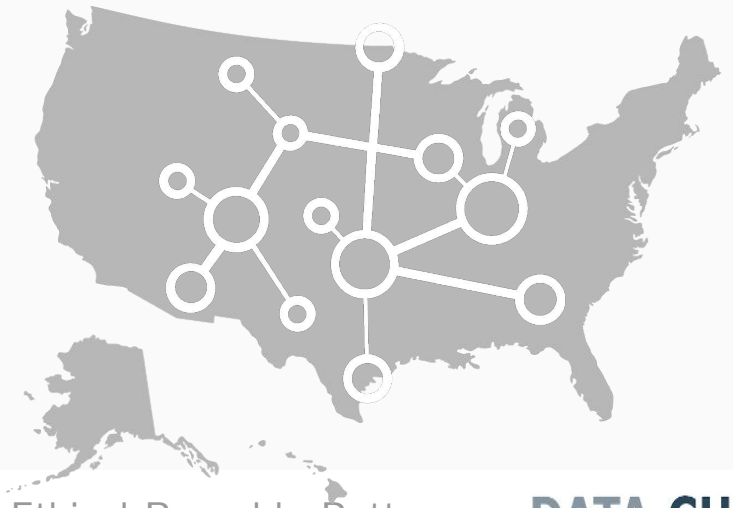
Top ranked impacts of data curation (out of 14)

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



What is the Value of Curation?

Perspective of repositories *doing* curation

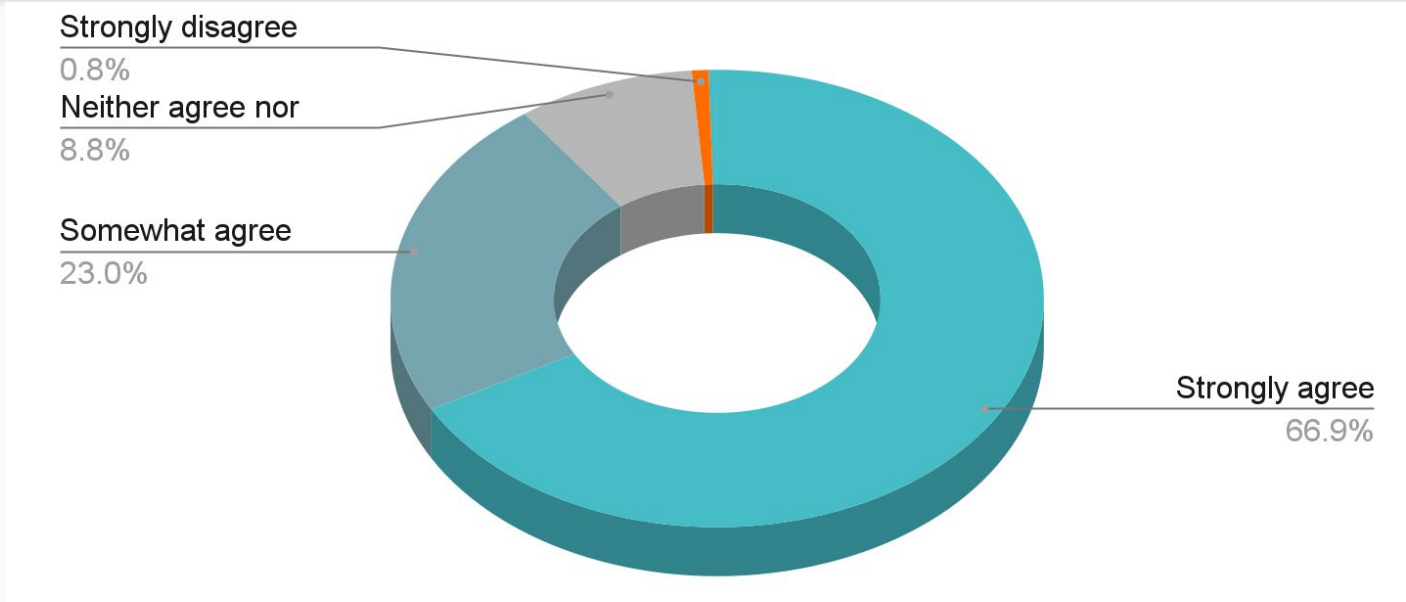


Perspective of depositors *receiving* curation

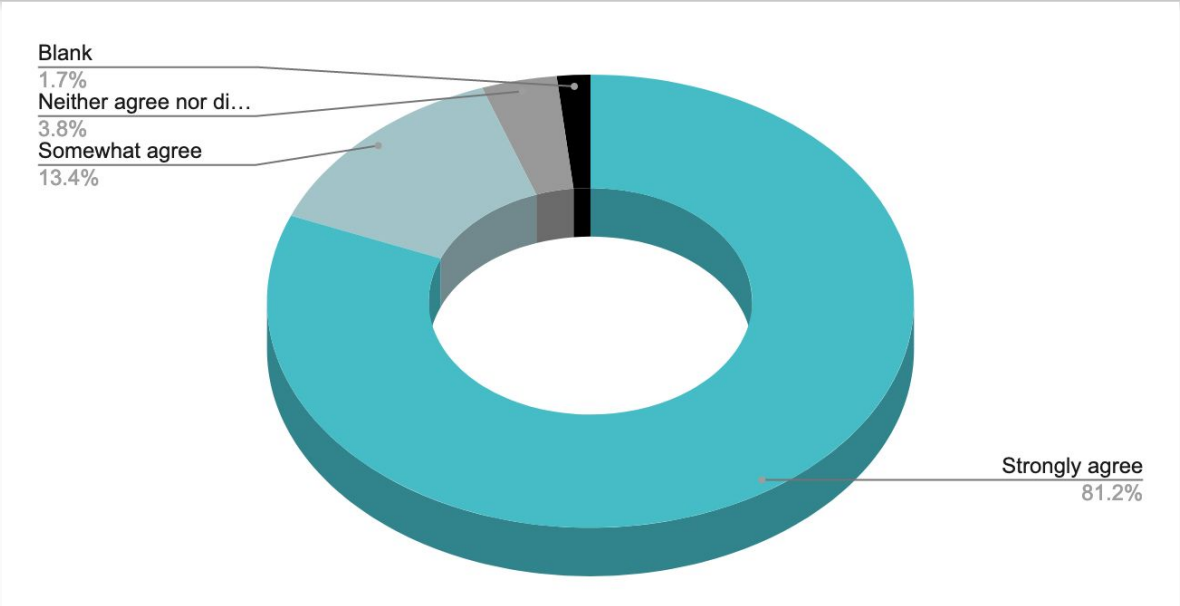
Survey Response rate = 40% (n=227)

	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%**

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



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.

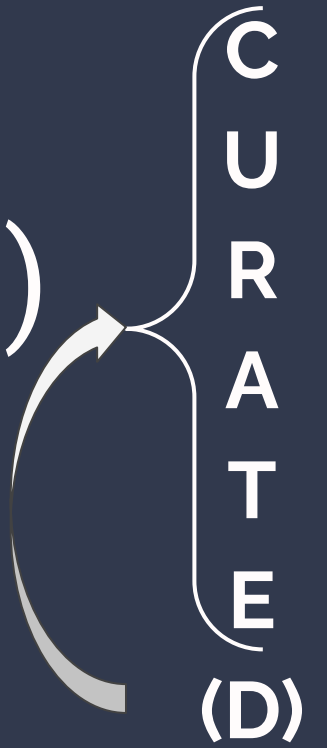
The curation process makes the data more accessible and readable for other scientists. This amplifies the impact of our research.

What do researchers value?

"The data curation was helpful in that it forced me to consider the **viewpoint of the end user** of the data. The curators nudged me to add pieces that would help third parties navigate the data (e.g. readme file) months or years in the future. **In retrospect** this is a good thing; at the time I was focused on getting my thesis chapter out and published, and didn't pay attention to this aspect as much as I needed to."

CURATE(D) Model

CURATE(D) Steps



Check files and read documentation.

Understand the data (or try to), if not…

Request missing information or changes.

Augment the submission.

Transform file formats for reuse.

Evaluate for FAIRness.

Document your curation activities

C⇛ U⇛ R⇛ A⇛ T⇛ E⇛(D)

⇓

Check files

CHECK step - action

- Review and inventory the content of the data files (e.g., open the files)
- Verify all metadata provided by the author and review the available documentation
- Ensure content of dataset is in scope
- Look for obvious ethical/sharing red flags

CHECK step - curator checklist

- ❑ Begin Curator Log to track curation decisions
- ❑ Open the related article and supporting information if available
- ❑ Inventory the dataset
  - ❑ Identify file formats
  - ❑ Review file organization, hierarchy, and naming convention(s)
  - ❑ Extract zip files when possible)
- ❑ Create working copy of files for formal inventory and testing
- ❑ Examine code for obvious errors/missing components, etc.
- ❑ Check that metadata quality is rich, accurate, and complete to institutional requirements.
- ❑ Check documentation
  - ❑ Complete
  - ❑ Needs work
  - ❑ If missing, document for the “Request” step
- ❑ Check whether human subject data. If so,
  - ❑ Request consent form / participation agreement if not present

CHECK step - curator checklist

- ❏ Check the accessibility of all files
  - ❏ Ensure there are robust descriptions in plain text of data files and any images.
- ❏ Check whether any visualization(s) of data are easily accessible
  - ❏ Review alt-text and visualization descriptions. Ensure these describe, but do not interpret, associated visualizations.
- ❏ Check data visualizations follow accessible color contrast guidelines
- ❏ Recommend graphical representation

  _____
- ❏ Recommend web-accessible surrogate

  _____

CHECK step – key ethical considerations

Review participant agreement and data use agreements; examine potential impacts of sharing this data. Consider:

- Individuals and communities represented
- Representativeness of diverse human populations
- Protection or endangerment status of species
- Geographic locations (e.g., specificity, contested boundaries, historical and current political situations)
- Animal research ethics and approval

Is it possible that the data deposit may impact a specific group?

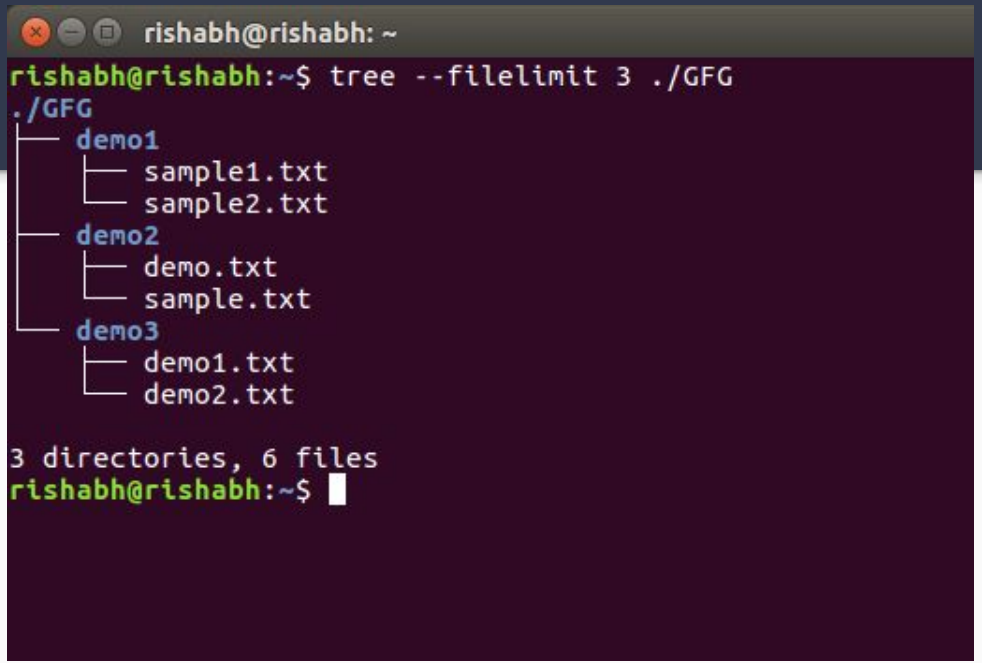
Does this data deposit follow compliance and institutional policy?

CURATED Training:
https://datacurationnetwork.github.io/CURATED/

Software Roundup

1. Appraisal Tools
    a. **Tree and Print Window**

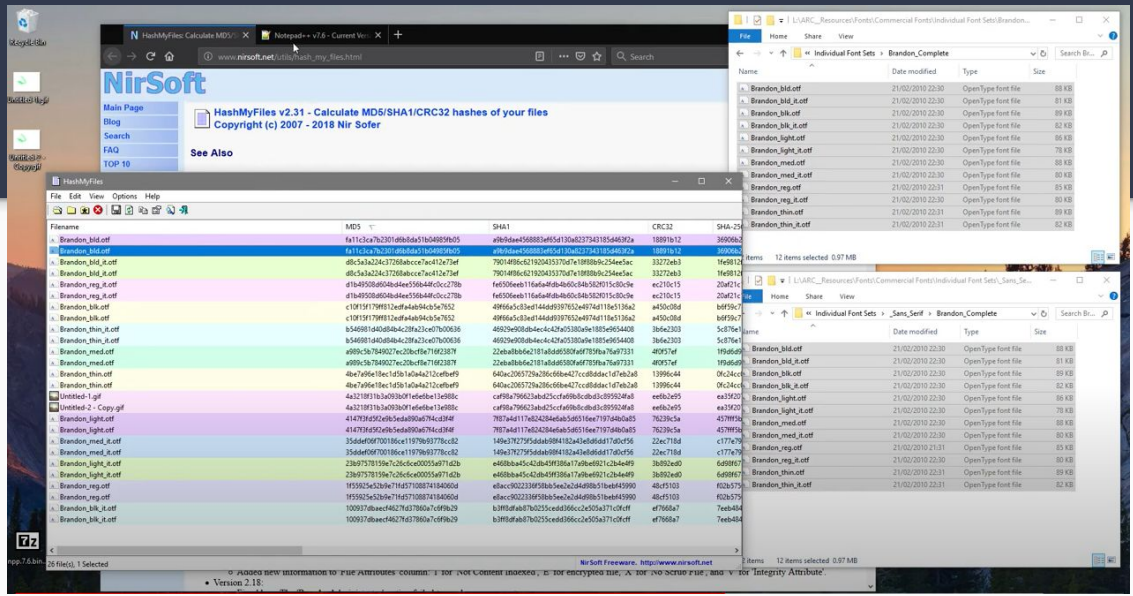
Tree /f /a > FileToPrint



Software Roundup

1. Appraisal Tools
   a. Tree command
   b. **HashMyFiles**

Download HashMyFiles: https://www.nirsoft.net/utils/hash_my_files.html



Screenshot from [HashMyFiles tutorial on YouTube]

Software Roundup

2. Speciality Tools (Processing and Review)

   a. ArcGIS, QGIS, GeoNetwork
   b. RStudio
   c. Matlab
   d. AliView (genomics data viewer)
   e. Omero (microscopic images)
   f. ChemDraw
   g. MZmine - mass spec data
   h. Jena - 3d molecular structure/Crystallographic files

Discussion
What tools do you use in curation?

C⇛ **U**⇛ R⇛ A⇛ T⇛ E⇛ (D)

Understand the data

Understand step - action

- Check for quality assurance and usability issues
- Try to detect and extract any “hidden documentation”
- Determine if the documentation of the data is sufficient for a user with similar qualifications to the author’s
- Assess whether the data requires a deeper dive due to the sensitivity of the data

Understand step - curator checklist

- ❏ Examine files, organization, and documentation more thoroughly. Are there changes that could enhance the dataset?
  - ❏ Are there missing data?
  - ❏ Could a user with similar qualifications to the author's understand and reuse these data and reproduce the results?
  - ❏ Are the data, documentation and/or metadata presented in a way that aids in interpretation? (e.g., readme)
- ❏ Record all questions and concerns in a curation log.

Understand step - curator checklist

Tabular data sub-checklist (e.g., Microsoft Excel)

- ❑ Organization of data well-structured
  - ❑ Not rectangular
  - ❑ Split tables into separate tabs
- ❑ Headers/codes clearly defined
  - ❑ Define headers
  - ❑ Clarify codes used _____
  - ❑ Clarify use of “blanks”
  - ❑ Clarify units of measurement
- ❑ Quality control clearly defined
  - ❑ Unclear quality control
  - ❑ Update/add Methodology

Full checklist

UNDERSTAND step – key ethical considerations

If working with human data, is this research done with and not on communities and populations involved? (You may wish to review data sources, researchers, and their connections to the communities and subjects they are serving to facilitate further conversation with researcher(s).)

- Are there authoritative group representatives who should be contacted in the next (request) step?

Are there labels or other descriptive indicators that could be applied to better represent or protect an identified group of people impacted by this dataset? (Example: TK labels)

CURATED Training:
https://datacurationnetwork.github.io/CURATED/

Documentation can be for...

- assessing data for reuse
- assistance in actual reuse
- explaining and/or maintaining consistency of data
- training new people
- efficiency in archiving

README Template

```
README TEMPLATE

Dataset Title:

Author(s):

Description:

Methods and Processing:

Funding Source:

Supplements:
```

```
<?xml version="1.0"?>

<metadata
  xmlns="http://example.org/myapp/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://example.org/myapp/ http://example.org/myapp/schema.xsd"
  xmlns:dc="http://purl.org/dc/elements/1.1/">

  <dc:title>
    UKOLN
  </dc:title>
  <dc:description>
    UKOLN is a national focus of expertise in digital information
    management. It provides policy, research and awareness services
    to the UK library, information and cultural heritage communities.
    UKOLN is based at the University of Bath.
  </dc:description>
  <dc:publisher>
    UKOLN, University of Bath
  </dc:publisher>
  <dc:identifier>
    http://www.ukoln.ac.uk/
  </dc:identifier>

</metadata>
```

Note that the http://example.org/myapp/schema.xsd XML schema does not exist - this is a ficticious example.

Informal ReadMe → **Formal Schema**

Lower-Barrier	Lower-Potential	Higher-Potential	Higher-Barrier
Fast	Irregular	Standardized	Slow
Easy	Incomplete	Machine actionable	Skilled

Readme files as documentation

Recommended content

- Data and file overview
- Sharing and access information
- Methodological information
- Data-specific information
- General information

https://data.research.cornell.edu/content/readme

Some issues with **CHECK**

- Can’t open the files/code
- Too many files to open
- Can open the files, but don’t know what I’m looking for
- No (or not enough) documentation

Some issues with **UNDERSTAND**

- Not enough experience to evaluate these data
- Documentation is over my head
- Still can’t tell what’s typical in this field
- Required software is too hard to find/use/justify $$$

C⇛ U⇛ **R**⇛ A⇛ T⇛ E⇛ (D)

Request missing information

The “Request” Step

What information do you need to address the gaps that you have identified?

How will you get this information?

Goals for the “Request” Step

Photo by rawpixel on Unsplash

- Goal #1 - Get a response
- Goal #2 - Establish a rapport
- Goal #3 - Get the information that you need

Determine what information is needed

Image from Taskworld Blog:
https://medium.com/taskworld-blog/build-on-top-of-taskworld-with-public-api-fca3d9e110d2

- Be as specific as possible
- Prioritize your needs
  - Essential
  - Important
  - Supplemental
- Do you need information or permission to take action?

Advice: Make it Easy

Image from Staples website: https://www.staples.com/Staples-Easy-Button/product_606396

- Limit to 4 asks
- Be specific, but try to keep it short
- Keep it as simple as possible (Ideally, response could be “yes or no”)
- Provide resources where useful
- If possible, do the work yourself and ask for approval

Request step - action

- Triage and prioritize issues. Highlight those with the highest data reuse implications
- Convey a sense of urgency, as it becomes more difficult to get responses from researchers as time passes
- Collaborate with the researcher(s) to make necessary changes
- Communicate any changes you, the curator, will make on their behalf
- Pause and consider how best to frame and communicate requests
- This should be the start of a conversation

REQUEST step - key ethical considerations

- Consider asking researchers if their participants will be notified that their data are being shared
- If you feel uncomfortable about sharing the data in its current state and/or it does not meet your institution's requirements, reserve the right not to publish
- Consider asking researcher(s) if there are limitations to how data could/should be used. Include any limitations in the documentation

Rapport: It may take more than one email

Created by Vectors Market
from Noun Project

- May be able to ask for additional information once rapport is established
- Offer to meet in person if needed
- Put the depositor first
  - What are their needs in depositing?
  - A little empathy goes a long way

REQUEST step - curator checklist

- ❑ Ask about additional data contributors, beyond publication authors. Consider using the Contributor Roles Taxonomy to communicate this: https://casrai.org/credit/

- ❑ Summarize conversations / outreach efforts in Curator Log

Long version at: Request Step

Request information

“I'm impressed with how well you seem to understand the data and how through your review was!”

- Researcher response to Data Repository for the U of Minnesota (DRUM) curator, Feb 2017

C⇛ U⇛ R⇛ **A**⇛ T⇛ E⇛ (D)

Augment the submission

Augment step - action

Image: pxfuel
https://www.pxfuel.com/en/free-photo-qfeat

- Apply the information that you have received from the depositor
- Add metadata as appropriate to enhance discovery and access
- Improve documentation to make data more understandable, interoperable and reusable

Areas to consider

Photo by Louis Reed on Unsplash

- Discoverability
  - Full text indexing
  - File Ordering
  - Data Set Description
  - File Description
  - Keywords
- Access
  - File / Folder Compression

Areas to consider

- Linkages
  - Articles & other outputs
  - Other sources or inputs
  - Related datasets
- Persistent Identifiers
  - DOIs, ORCIDs, RORs

Photo by Chris Leipelt on Unsplash

Common metadata elements needed for datasets

- Author(s)
- PID (DOIs)
- Title
- Keywords
- Abstract / Description
- Language
- Citation
- Link(s) to related publications
- Licensing information
- Funder information

Brief recap of “C” & “U”

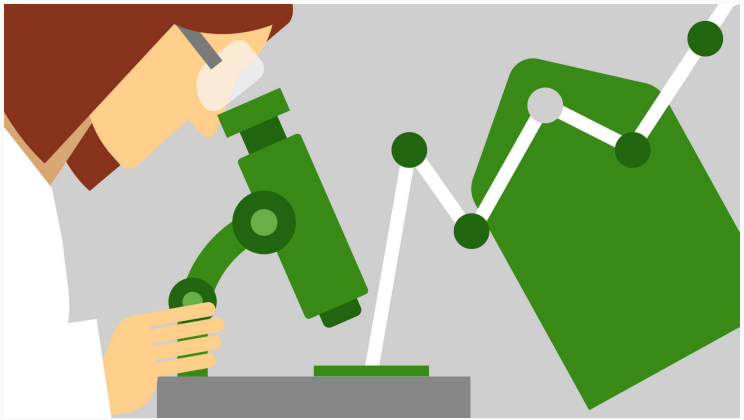
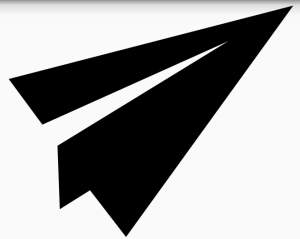


Image Source: https://www.linkedin.com/learning/the-science-of-sales

These are the steps taken to learn about the data set

- What is it you have?
- What should (or could) it be?
- What are the gaps between what you have and what you want?

Brief recap of “R” & “A”



Created by Briyan Design
from Noun Project

Created by NeMaria
from Noun Project

These are the steps taken get more information and improve the metadata

- What are the priority issues?
- What are 4 things you need to know?
- How can you enhance the metadata using what you’ve learned?
- What will make it more findable and useable?

C⇛ U⇛ R⇛ A⇛ **T**⇛ E⇛ (D)

Transform file formats

Example File Format Transformations

Native Software or Format	Suggested Formats or Transformations	Transformation Tools and Notes
CZI (microscope images)	TIFF, JPG, FITS	Use “export”; Omero, Bioformats; WikiData tracks software and file formats for preservation
Microsoft Excel / XLS, XLS	CSV, TSV	Use “save as”; Use Excel Archival Tool to preserve formulas
Chemdraw / CDX	CDXML, MOL, JPG, 001, OPJ, TRI	Retain original. Some conversions will result in loss of information.
PDF	PDF-A	Use “save as”
MP4, MOV, WMV	Uncompressed AVI or MOV + captions	No information is gained going from a “lower” resolution image to a “higher” one, but long-term access may be improved. Use YouTube, Vimeo, Kaltura or other tools for captioning.
.SHP (geocoded xls)	CSV + extracted metadata	Retain .SHP. Use FME Tool or ArcGIS
Webpages	WARC, TIFF	Link to Internet Archive. Provide screen shots.

Transform step - curator checklist

- ❑ Preferred file formats in use
- ❑ Recommend conversion
  - from _____
  - to _____
- ❑ Retain original formats
- ❑ Software needed is readily available
- ❑ Unclear version of software
- ❑ Unclear software used
- ❑ Visualization of data easily accessible
- ❑ Recommend graphical representation _____
- ❑ Recommend web-accessible surrogate _____

C⇛ U⇛ R⇛ A⇛ T⇛ **E**⇛ (D)

Evaluate for…

What do we mean by Evaluate?

C⇛ U⇛ R⇛ A⇛ T⇛ **E**⇛ (D)

Evaluate for curation success

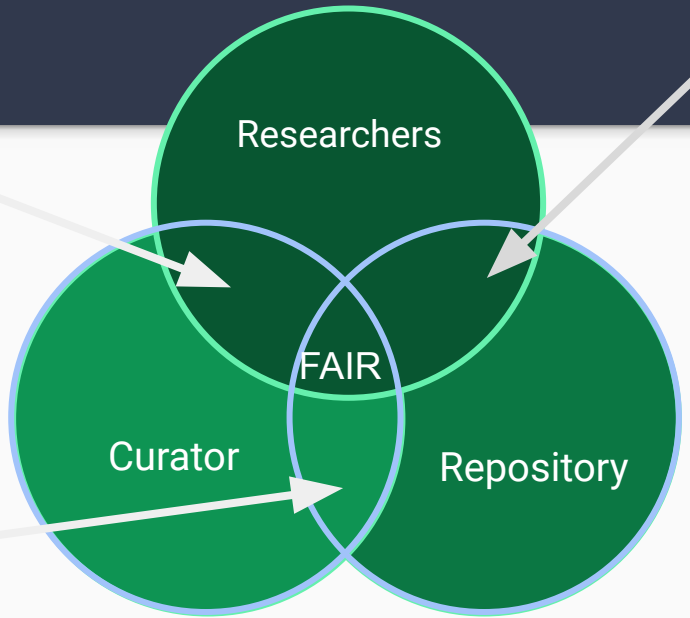
What do we mean by Evaluate?

Did the curation partnership result in FAIR data?

- Communication
- Expertise
- Buy-in?

Will this repository make my data FAIR?

- Services
- Policies
- Transparency



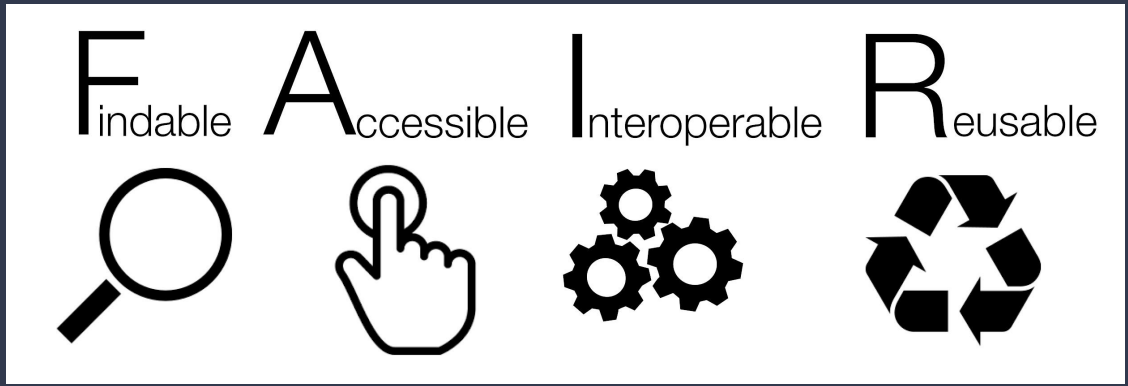
Does the platform work for FAIR data?

- Platform Features
- Technology
- Standards

C⇛ U⇛ R⇛ A⇛ T⇛ **E** ⇛ (D)

⇓

Evaluate for curation FAIRness



Findable

*To be **findable (F)** or discoverable, data and metadata should be richly described to enable attribute-based search*

- ❏ (meta)data are assigned a globally unique and eternally persistent identifier
- ❏ data are described with rich metadata
- ❏ (meta)data are registered or indexed in a searchable resource
- ❏ metadata specify the data identifier

Accessible

*To be broadly **accessible (A),** data and metadata should be retrievable in a variety of formats that are sensible to humans and machines using persistent identifiers*

- ❏ (meta)data are retrievable by their identifier using a standardized communications protocol
- ❏ the protocol is open, free, and universally implementable
- ❏ the protocol allows for an authentication and authorization procedure, where necessary
- ❏ metadata are accessible, even when the data are no longer available

Interoperable

*To be **interoperable (I),** the description of metadata elements should follow community guidelines that use an open, well defined vocabulary.*

- ❏ (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
- ❏ (meta)data use vocabularies that follow FAIR principles
- ❏ (meta)data include qualified references to other (meta)data

Reusable

*To be **reusable (R),** the description of essential, recommended, and optional metadata elements should be machine processable and verifiable, use should be easy and data should be citable to sustain data sharing and recognize the value of data.*

- ❏ meta(data) have a plurality of accurate and relevant attributes
- ❏ (meta)data are released with a clear and accessible data usage license
- ❏ (meta)data are associated with their provenance
- ❏ (meta)data meet domain-relevant community standards

Evaluate step - curator checklist for FAIR

Findable -

- ❏ Metadata exceeds author/ title/ date,
- ❏ Unique PID (DOI, Handle, PURL, etc.).
- ❏ Discoverable via web search engines.

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
- ❏ Contact info displayed if the direct assistance of the author needed.
- ❏ Clear indicators of who created, owns, and stewards the data.
- ❏ Data are released with clear data usage terms (e.g., a CC License).

Checklist based on: http://www.ijdc.net/article/view/567

C⇛ U⇛ R⇛ A⇛ **T**⇛ E⇛ (**D**)

Document your curation activities

Document your curation activities

Using whatever practices / tools in place at your institution, be sure you've documented and recorded everything you've done to curate this dataset

- **What to document**
  - Accessioning and deposit records
  - Repository dataset cataloging metadata
  - Provenance/change logs
  - Service workflow
  - Preservation packaging metadata

Stay in touch!

https://datacurationnetwork.org/
https://datacurationnetwork.org/monthly-newsletter/