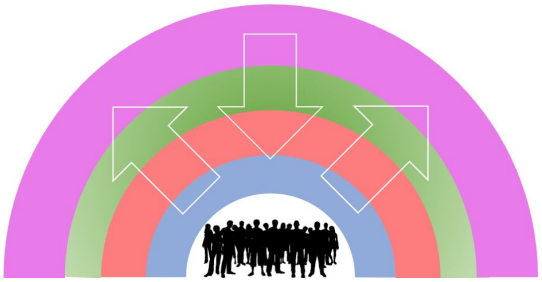
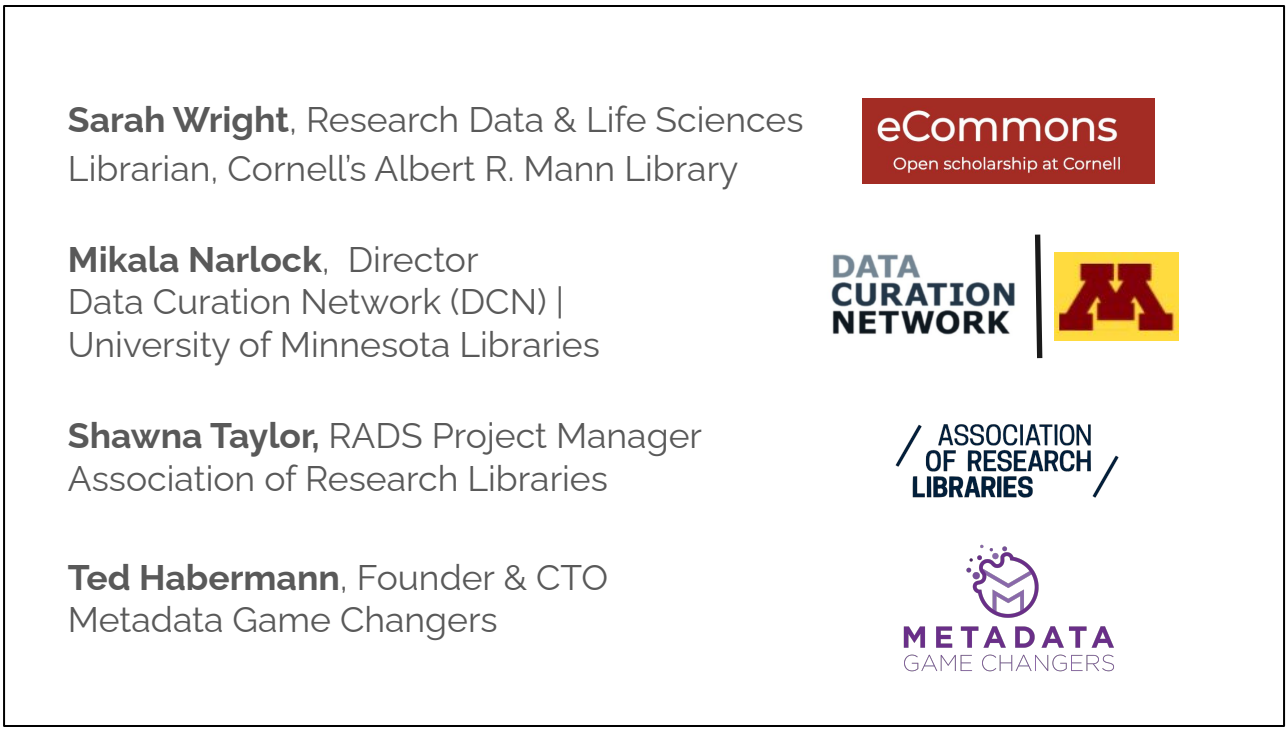
Think Globally, Act Locally: The Importance of Elevating Data Repository Metadata to the Global Infrastructure.



Open Repositories 2022
Panel Session

The title of our presentation today is, “Think Globally, Act Locally: The Importance of Elevating Data Repository Metadata to the Global Infrastructure.” During our session we are going to examine how institutional repositories can adopt measures locally to improve the completeness of their institutions’ research data in the global research infrastructure. This is significant, as improved metadata in the global infrastructure increases the likelihood of findability and reusability of datasets at a larger scale.



My name is Shawna Taylor, I am the Project Manager of the Realities of Academic Data Sharing Initiative at the Association of Research Libraries; If the other panelists could introduce themselves…

- Sarah Wright, Research Data & Life Sciences Librarian, Cornell's Albert R. Mann Library
- Ted Habermann, Founder & CTO, Metadata Game Changers
- Unfortunately at the last minute Mikala Narlock, director of the Data Curation Network, could not join us today, but I will be covering for her and representing her perspective during our panel discussion.

eCommons - Local Institutional Repository

eCommons
Open scholarship at Cornell

- How can we make data as FAIR as possible considering local Institutional Repository (IR) constraints?
  - Busy researchers
  - Gaps in local infrastructure
  - Data-specific metadata and documentation concerns

- How to connect local IR metadata with the global infrastructure (DataCite, Crossref, ORCID, ROR, etc.)?

Our panel brings together four perspectives to best understand how local institutional repository efforts can improve connectivity in the global research infrastructure.

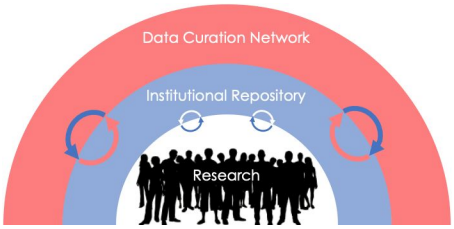
Sarah's perspective is that of the librarian and data curator working directly with a data or institutional repository. She will discuss the challenges in making data FAIR at her IR, and barriers in connecting local metadata with the global infrastructure.

A quick note on the visual here in the left hand corner. As we move through our presentation, the visual will develop, and additional rainbow bands will be added. (NEXT Slide)

DCN - Community Perspective

DATA CURATION NETWORK

- Analyze repository metadata to identify fields and values that are particularly useful for enabling reusability
- Build consensus on best practices



DCN Vision Statement

We strive to be a trusted community-led network of curators **advancing open research** by making data more ethical, reusable, and understandable.

Each band indicates one of our four panelist perspectives and is connected with feedback arrows, indicating dynamic reciprocal relationships.

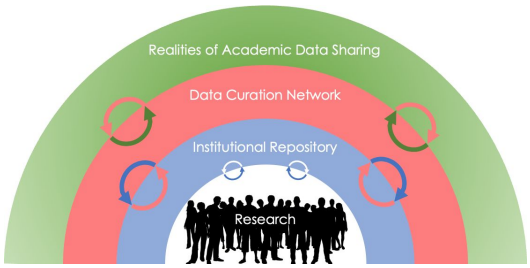
Next is Mikala with the Data Curation Network, and she represents the community perspective. The DCN's mission statement is "to be a trusted community-led network of curators advancing open research by making data more ethical, reusable, and understandable."

Their mission statement is of particular importance here as the DCN can build best practices outward; local repository best practices or general dos and don'ts are shared within their community, and these practices grow, are refined, and ultimately can impact the global research infrastructure. This is one way in which the DCN advances open research.

Realities of Academic Data Sharing (RADS) Initiative

ASSOCIATION OF RESEARCH LIBRARIES

- Assessing metadata quality/completeness at six DCN member institutions.
  - Cornell University
  - University of Michigan
  - Virginia Tech
  - Duke University
  - University of Minnesota
  - Washington University in St. Louis



- Identify opportunities to improve and connect local IR meta(data) to the global research infrastructure.

My perspective represents the Realities of Academic Data Sharing Initiative, an Association of Research Libraries, National Science Foundation-funded project. This initiative, RADS, as we call it, is working with six academic institutions to examine metadata quality of their institutional affiliated research data. These institutions, listed here, are all members of the DCN and are interested in developing communities around data curation.

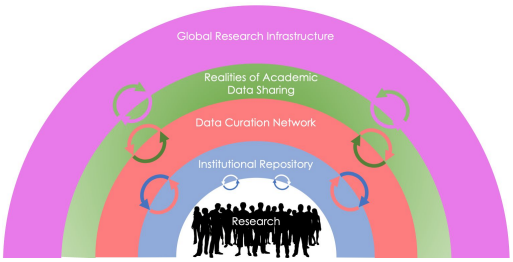
Ted Habermann is the metadata consultant for RADS and he is leading this metadata analysis work in our project.

Metadata Analysis

METADATA GAME CHANGERS

ASSOCIATION OF RESEARCH LIBRARIES

- RADS project metadata analysis
  - DataCite was queried to determine **where** researchers are sharing their research data
  - Metadata **completeness** analyzed using FAIR recommendation for DataCite metadata



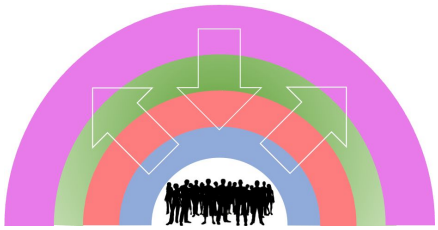
- Serendipitous Improvements
- Can we increase content in the global infrastructure while minimizing impact on the IRs?

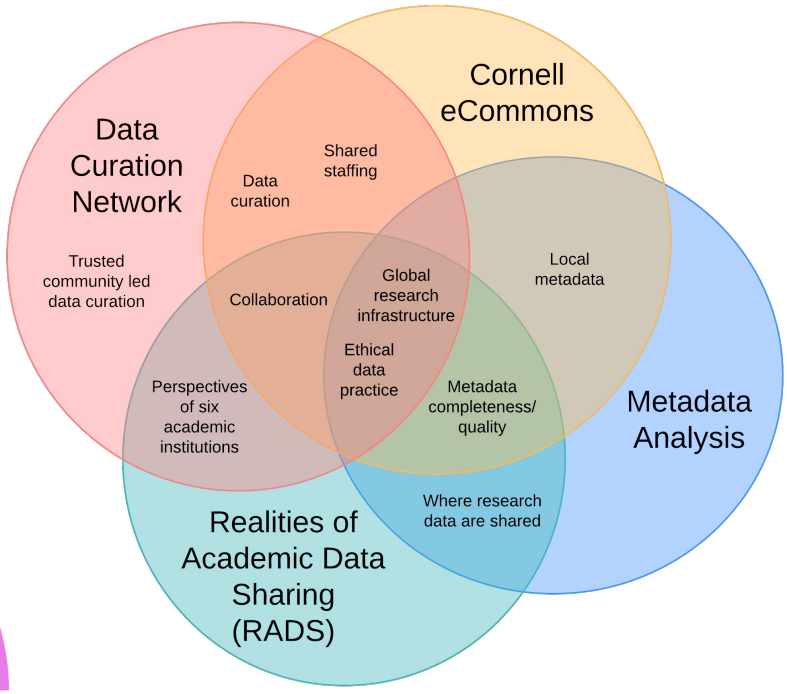
As part of the RADS project, Ted queried DataCite to determine where researchers are sharing their research data and assessed the quality of the metadata. Metadata quality, or completeness, was analyzed using FAIR recommendation for DataCite metadata.

As a result, Ted recognized what we're calling 'Serendipitous Improvements'; meaning, he was able to use local IR metadata to improve upon the metadata of datasets already found in DataCite.

Shared Areas of Interest or Practice

- Cornell eCommons
- Data Curation Network
- RADS Initiative
- Metadata Analysis





This Venn Diagram shows the Shared Areas of Interest or Practice between our four perspectives. As you can see, ‘global research infrastructure’ and ‘ethical data practice’ are at the core of the diagram and are shared by all four of us. We recognize the need to improve findability, discoverability, and connectivity to enable data reuse.

Now to get us started, I will turn it over to Sarah.

Data Curation at Cornell

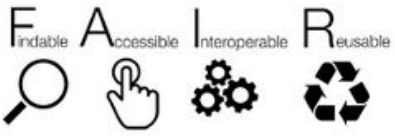
Preserve and share your data in eCommons

- Curatorial review
- Open access
- Persistent identifiers
- Links to publications
- Download statistics

CORNELL UNIVERSITY LIBRARY

eCommons
Open scholarship at Cornell

ORCID
Connecting Research and Researchers



doi

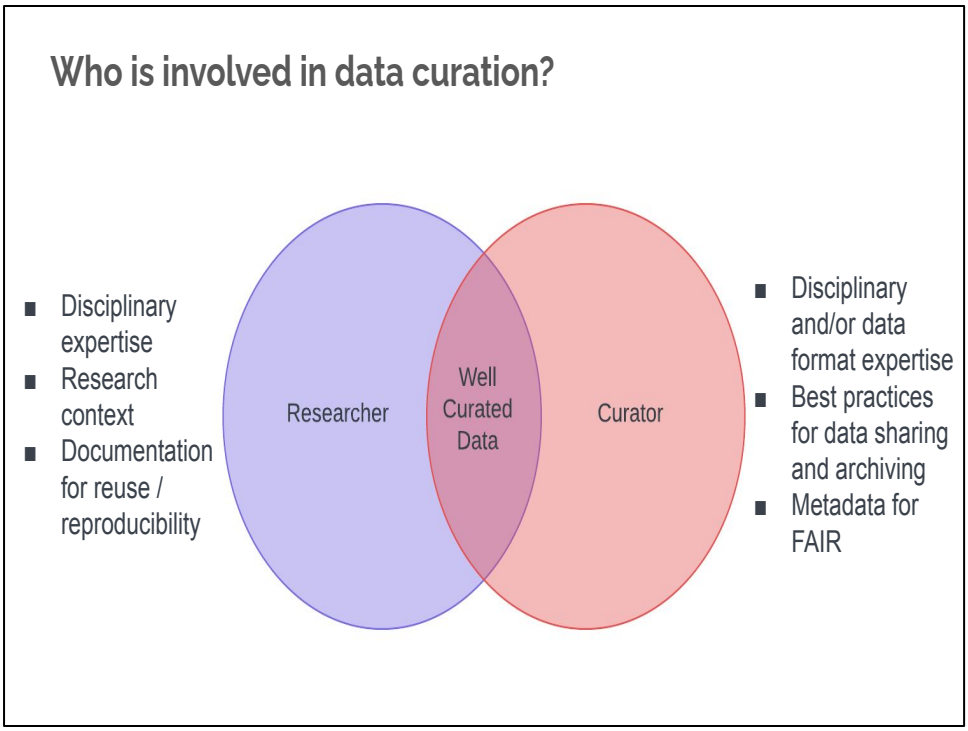
Thanks Shawna! Today I'm going to focus on our local IR, eCommons, and how we work with researchers to preserve and share their data, while keeping in mind the broader implications of our work.
We offer data curation services, to ensure data going into eCommons is as well-described and complete as possible, and in a format and structure that best facilitates long-term access, discovery, and reuse.

Data Curation

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

Data curation actions may require substantial outlays of time and energy, for example detecting and fixing errors in code, reviewing code annotation, or even reviewing documentation for large collections of files; other curation actions may be more limited in scope, like transforming file formats for long term access or reviewing and augmenting metadata. There are different levels of curation work that may require different levels of involvement from the researcher and from the curators.



Data curation is a collaboration between researchers and curators, both doing their part to ensure that data are well prepared and meaningful, including the necessary contextual information for reuse and reproducibility, and repositories work to ensure that data are findable, accessible, and preserved for the long term.
So to go back to the list of curation actions on the last slide, I, the curator might identify missing files or documentation for re-use, and will then work with the researcher to add the missing files and information.

What's in it for the researcher?

So we've established that good curation takes time and effort, both on the part of the researchers and on the part of curators and repository staff. What's in it for the researchers?

DCN Researcher Results 2016 (n=91)

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)**

* Rated by more than one DCN focus group from our 2016 Study

Not Happening for Majority of Researchers

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

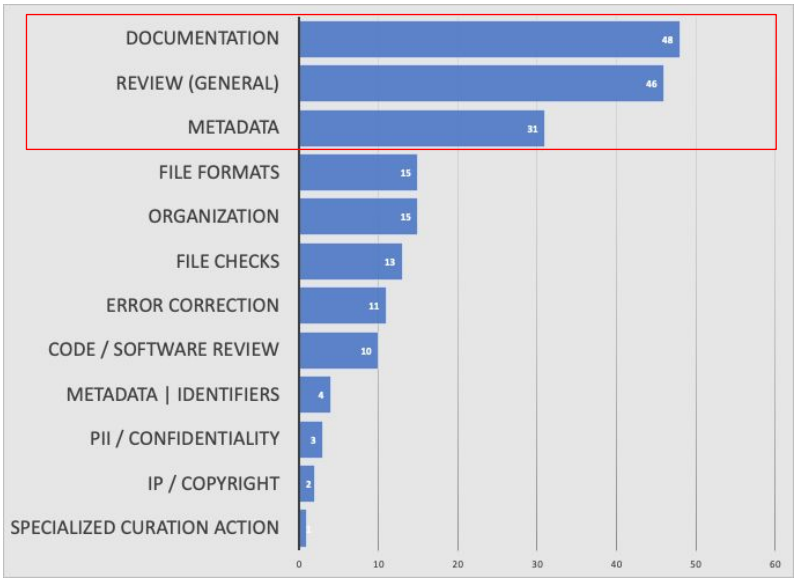
Happening, but not satisfactorily

- **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)

Johnston, L. R. et. al.. (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

Researchers want better metadata, and appreciate curation services - they need our help!

What is the most "value-add" curation action taken by this repository? (n=182, DCN depositor survey 2021)



More recently, we surveyed researchers who had deposited data in our repositories over the past 1.5 years. Documentation and metadata are both very important to researchers!

- Reuse (Funders)
- Reproducibility (Journals)
- Recognition (Researchers)

"You can't keep coming in here and demanding data every two years!"

I think it's important to remember that We are trying to accomplish a lot…
Funders are concerned with ROI,
Journals are concerned with reproducibility,
Researchers care about the data they are putting out there and want it well-described so that it can be re-used appropriately, and they are satisfying funder and publisher requirements, and of course need recognition to continue getting tenure, and grant funding and all of that.
Reminder of how important it is to make all of these connections so that the information can flow among all of these interested parties.

Institutional Data Sharing Requirements

Cornell University Policy Office
policy.cornell.edu

Policy 4.21
Research Data Retention

Accurate and detailed records of research data are an essential component of any research project. This policy defines the shared responsibilities of Cornell University (including Weill Cornell Medicine) and Cornell researchers in collecting, retaining, securing, accessing, publishing, and sharing research data.

1.3.4. University ownership of research data: Cornell … asserts ownership of research data and related property rights arising from the activities of its researchers and others who use university resources...

1.3.7. Ithaca-based faculty – collection and retention of data: Research data is retained for a minimum of three years after the final project closeout. If the primary data and images are used in a subsequent publication, or the initial publication is citied [sic] in a subsequent publication or grant application by the faculty member, the data and images must be available for an additional six years. If specific software or code is required for the University to interpret the data, this software or code should also be deposited with the data, as long as license agreements permit.

https://policy.cornell.edu/sites/default/files/vol4_21.pdf

Example of local impetus for data sharing: Cornell recently released a research data retention policy, so researchers may be looking for a place to safely store their data and satisfy this requirement.

Funder Data Sharing Requirements: NIH example

NIH SCIENTIFIC DATA SHARING

As of 2023, ALL investigators are required to:

•**Submit** a Data Management and Sharing plan outlining how scientific data and any accompanying metadata will be managed and shared, taking into account any potential restrictions or limitations.

•**Comply** with the Data Management and Sharing plan approved by the funding Institute or Center (IC).

And global: Funders are requiring researchers to submit data management and sharing plans, and to report the outputs in interim and final grant reports - this is another example where DOIs and links to sponsorship becomes important to be able to automate or at least simplify reporting, benefitting both the researchers and the funders.

Publisher Requirements

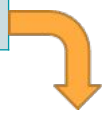
DOI, DOI, DOI

Wilkinson, M. D.. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(160018). doi:10.1038/sdata.2016.18
See also: https://www.force11.org/group/fairgroup/fairprinciples

DOIS: Publisher requirements are a major driver of deposits in our repository - most submitters are asking for a DOI to include in article proofs, in the data availability statement.
Really little to no attention to quality YET…for the majority of publishers, but some check whether the DOI resolves to a published dataset.

Cornell curation service: an idea of scale

Year	Curated Datasets Public in eCommons
2019	35
2020	43
2021	50



Essential Changes (data fundamentally changed)	Major Changes (e.g. file names changed; Readme added)	Minimal Changes (small edits)
14%	54%	32%

I do want to give you an idea of the scale of what we are seeing locally: an increasing, but still rather small, number of datasets being submitted to our repository, and of those submitted >50% need major changes like addition of a readme; 14% need fundamental changes like addition or deletion of files, file transformation, etc. 32% need only minimal changes like our steps to augment metadata to include sponsorship, keywords, or fixing small typos.

Local Roadblocks

- Busy researchers
  - Reluctance to require anything beyond author and title
- Repository infrastructure
  - ORCIDs, RORs not yet supported locally
  - Manual citation generation
- Staff shortages
  - 0 full-time data curators
  - 2 ~10% data curators + DCN membership
- Need for automation and augmentation
  - eCommons → DataCite
  - Augmenting Identifiers and Connections

Busy researchers - is this true??
Staff - focus is on description for reuse - the parts that are hard to figure out without the depositor, and are time consuming or impossible to track down later on. If we can capture as much context as possible in the readme, later with more staff and/or automation, we can circle back and fill in the metadata around sponsorship and funding, ORCIDs, affiliation (RORs), etc.

Local → Global

Metadata	eCommons	DataCite (2016-2020)	DataCite (2021-2022)
*Author	✔	✔	✔
*Title	✔	✔	✔
Abstract	✔		
Funding	✔		✔
Suggested Citation	✔	✔	✔
Keywords	✔		
Author ORCIDs			✔
Author Affiliation (RORs)			✔
Links to Related Content	✔		
License	✔		✔
Resource type	✔	✔	✔
Readme	✔		

This is to give an idea of how we have moved to increase our participation in the global infrastructure as we've developed our data curation service. At the outset we were focused on our local metadata in eCommons, and were really just filling in the minimal information required to mint a DOI. If/when updates are needed, we didn't want to duplicate effort, or end up with discrepancies. However, around 2020 we started discussing leveraging more of the Datacite fields to increase the FAIRness of data in our repository, and in 2021, we started asking researchers for ORCIDs and affiliations, even though we can't yet add them in eCommons.

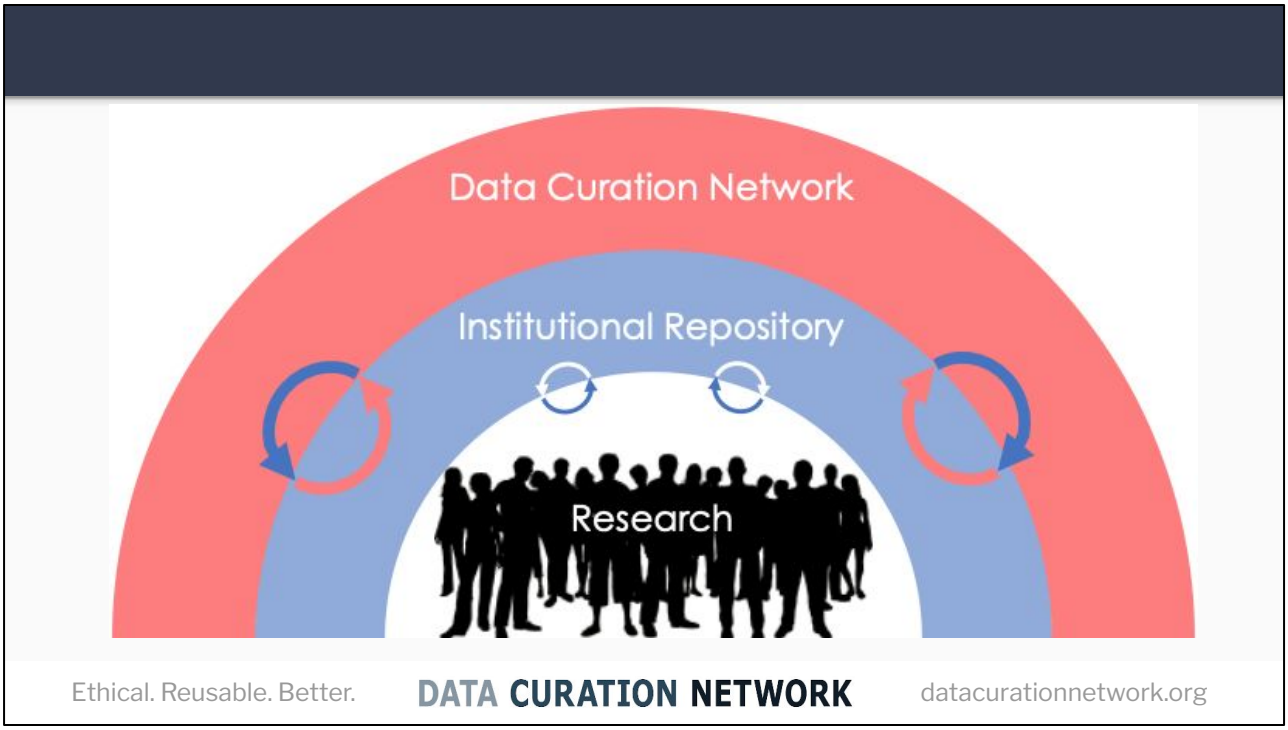
Recent Improvements (2021 - present)

- DataCite
  - ORCIDs, RORs (important for funders and institutions to track ROI and for researchers to get credit for their contributions)
  - Prioritize adding more metadata
- Repository infrastructure
  - New deposit form collects MORE METADATA!

So to summarize:
Why are ORCIDs and RORs important? Enables institutions and funders to track ROI (and for researchers to cite their data and get credit for it!)

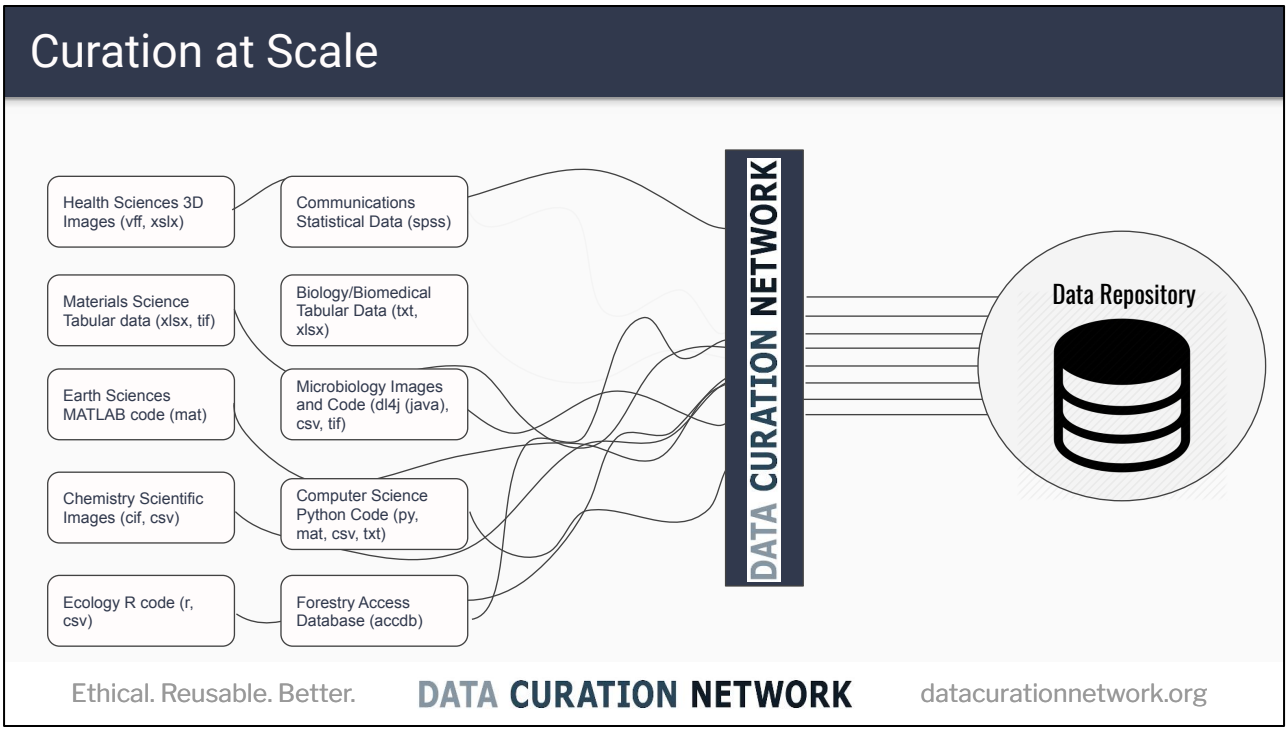
Future steps

- Data curation service
  - Working on promotion and improving local workflow
  - More staff?
- More FAIR (Findable, Accessible, Interoperable, Reusable)
  - We are improving the datasets, but still could be more machine readable, more metadata, etc.
- Making datasets in eCommons more discoverable
  - ORCIDs, RORs



As I mentioned earlier, Mikala is the director of the DCN and she represents the community perspective of our panel. Turning back to the visual, Cornell University is a member of the Data Curation Network, hence our rainbow's two connected bands.

The DCN is a community-led network of curators who advance open research by making data more ethical, reusable, and understandable. This is achieved through a shared curation model, education and training opportunities, and through research and advocacy. The latter two points, research and advocacy, is why the DCN is represented today, but before we dig into those two areas, I want to take a moment to share a little bit about the DCN, and why the DCN community is the most important part of their work.



One of the biggest benefits of the DCN, in addition to the active community, is the support of curation at scale. As many of you know, institutional and generalist data repositories are seeing increasing deposits from a wide variety of disciplines in various formats. Through the DCN, institutional members can ask for a DCN expert to curate these different datasets or work with experts to learn how to curate them for future reference. In other words, the DCN community is wonderful for both addressing any immediate issues, as well as for teaching and empowering local curators.

The CURATE(D) Workflow

C **Check** files and read documentation.

U **Understand** the data (or try to), if not...

R **Request** missing information or changes.

A **Augment** metadata for FAIR.

T **Transform** file formats for reuse.

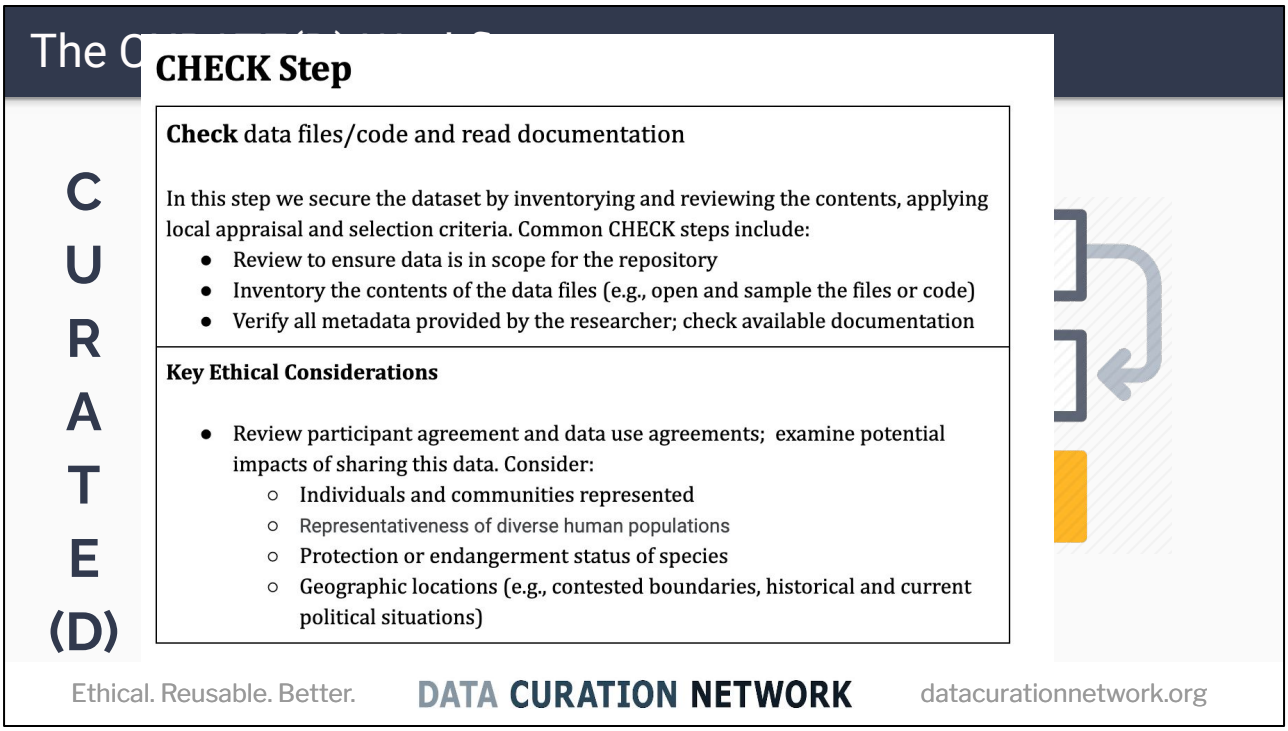
E **Evaluate** for FAIRness.

(D) **Document** your curation activities

Ethical. Reusable. Better. DATA CURATION NETWORK datacurationnetwork.org

Next, is the DCN CURATE(D) workflow– all DCN curators are trained to follow these steps when curating datasets. This workflow is a training tool to onboard new curators and is also used by DCN curators in both the shared curation model, and at partner institutions.

This workflow is flexible enough to allow for nuances in file formats and disciplines, but rigorous enough to be practical and useful.



DCN members recently updated this critical workflow and teaching tool to include key ethical considerations. These ethical considerations were incorporated based on feedback from the DCN as well as the wider Research Data Management community. We needed to be explicit about data curation nuances as we train the next generation of data curators and continue to learn.

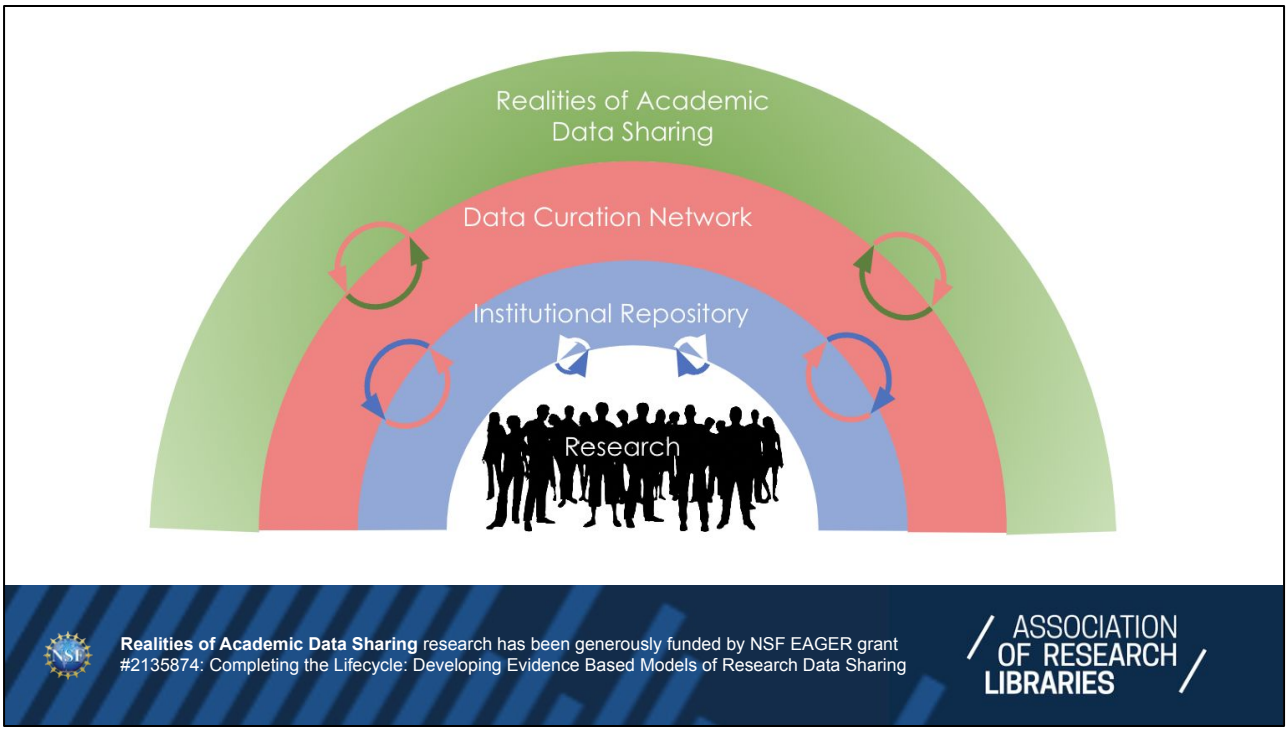
(Current) DCN Community

- 15 institutions & organizations
- 45 data curators
- 15 representatives
- 1 beta-test member
- 1 full time director

Ethical. Reusable. Better. DATA CURATION NETWORK datacurationnetwork.org

As you can probably determine from my brief introduction to the DCN, the community of curators and representatives are the foundation of all that the DCN does. This includes the various research projects DCN Community members undertake. Mikala, and this community, collectively use this information to improve services, advocate for curation, and make data more ethical, reusable, and better.

Research goals, as you can imagine, are DCN Community driven. They might start, for example, from peer to peer discussions around projects or efforts that have been successful, or not. DCN institutions provide opportunities to learn from one another, and the community as a whole can share pain points and challenges to identify shared opportunities for research and development.



This is why the DCN embarked on this collaborative project with the Association of Research Libraries and the RADS Initiative – to learn and to ensure that the systems and metadata we are creating, managing, and sharing improve the FAIRness of our data and metadata.

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 EAGER grant #2135874: Completing the Lifecycle: Developing Evidence Based Models of Research Data Sharing

ASSOCIATION OF RESEARCH LIBRARIES

This brings us to my project, the Realities of Academic Data Sharing Initiative, working with six academic institutions, all DCN members. We are answering three core research questions. Pertinent to our discussion today is the project's first research question, "Where are funded researchers sharing their data and what is the quality of that metadata?"

RADS: What is the Quality of the Metadata?

- ✓ Quality = FAIR complete
- ✓ Using rubric developed by Ted Habermann
- ✓ Classified metadata elements as **essential** or **supporting** for each component of FAIR
- ✓ Analyzed quality/completeness in each local IR and institutional affiliated (meta)data in DataCite

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

ASSOCIATION OF RESEARCH LIBRARIES

To answer the question, “What is the quality of the metadata?” of the research data from our six institutions, Ted, as RADS’s metadata consultant, first had to determine how he was going to define quality. Using a rubric he had previously developed, we take quality here to mean complete, or more specifically, FAIR complete. We don’t have much time to dig into the details of the rubric today, but, in sum, metadata elements were classified as either essential or supporting for each component of FAIR. For example, Findable Essential elements include: Abstract, Date Created, Keyword, Resource Author; whereas, Findable Supporting elements include Date Submitted, Keyword URI, etc.

The quality or completeness of the metadata was analyzed from three different repositories.

Data Repository for the University of Minnesota (DRUM): Metadata Comparison

Metadata Element	DRUM	DRUM@DataCite	Other Repositories
dc.description	95%	0.30%	10%
dc.description.abstract	80%	12%	95%
dc.subject	85%	2%	63%
dc.relation.isreferencedby	78%	0%	0.6%
dc.description.sponsorship	72%	1%	12%
Average	**82%**	**3%**	**36%**

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

ASSOCIATION OF RESEARCH LIBRARIES

Analysis from eCommons at Cornell is still in progress, so instead we have used information from the Data Repository for University of Minnesota (DRUM). The second column shows metadata completeness of datasets found locally in DRUM. The third column shows metadata completeness of DRUM datasets entered directly into DataCite. And the fourth column shows metadata completeness of University of Minnesota affiliated datasets entered into other repositories, such as Dryad or Zenodo. Typically, these other repositories then transfer metadata automatically into DataCite.

So, regardless if metadata is transferred into DataCite by DRUM staff or by these “other repositories”, it’s clear that metadata does exist, but is being lost during the transfers. This is clearly shown when looking across the rows in this DRUM table.

RADS: Serendipitous Improvements

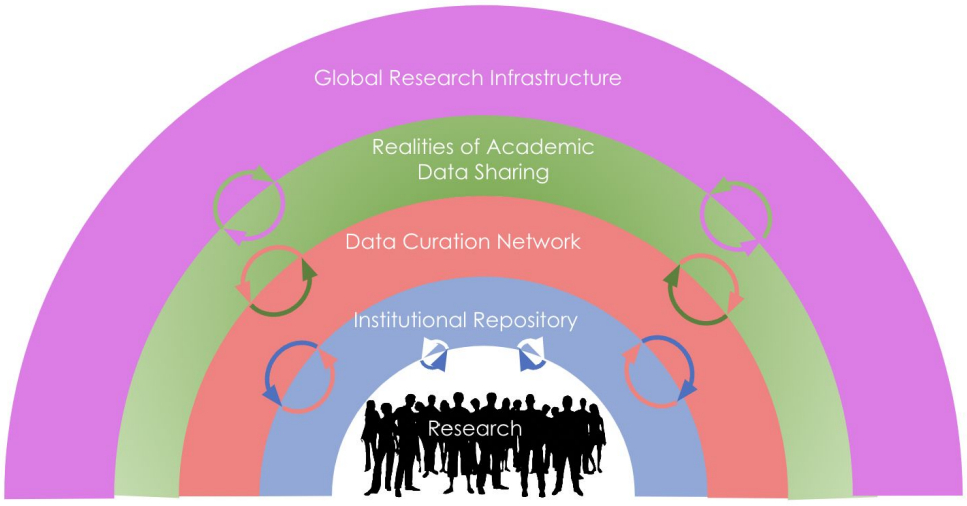
- ✓ Used DataCite's API to transfer metadata from local repositories to DataCite
- ✓ No new metadata content created
- ✓ Can we make further improvements in the local to global transfer using PIDs or documentation such as README files?

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

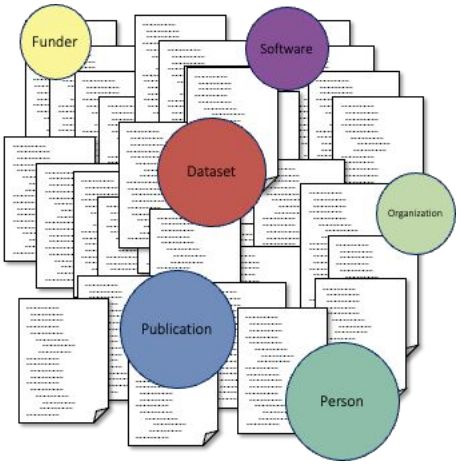
ASSOCIATION OF RESEARCH LIBRARIES

As a result of analyzing metadata completeness in these three ways, Ted figured there must be a way to take existing metadata from the local IR or these "other repositories", to improve DataCite metadata. To do this, he used DataCite's API to transfer metadata from local repositories to DataCite. Using this method, no new metadata had to be created and this left us wondering if further improvements could be made along these lines to enhance connections within the global research infrastructure.

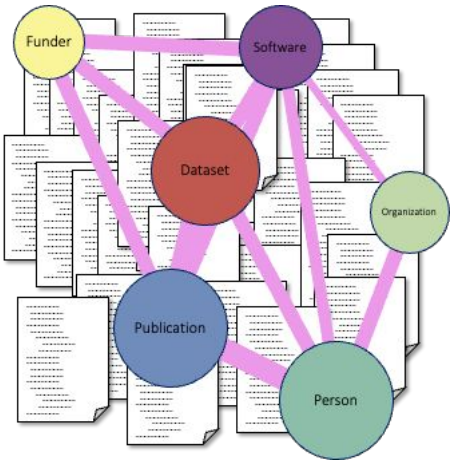
With that in mind, I'll turn it over to Ted.


Global Research Infrastructure
Realities of Academic Data Sharing
Data Curation Network
Institutional Repository
Research

Global Infrastructure: Role Evolves



Identification and Citation

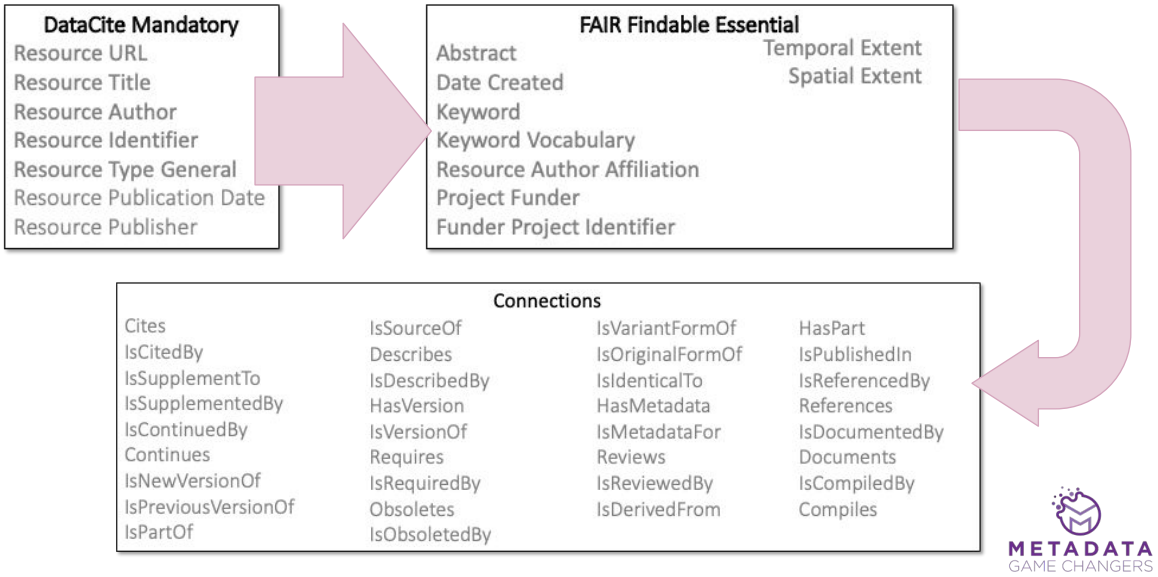


Connection

METADATA
GAME CHANGERS

Metadata Content Evolves

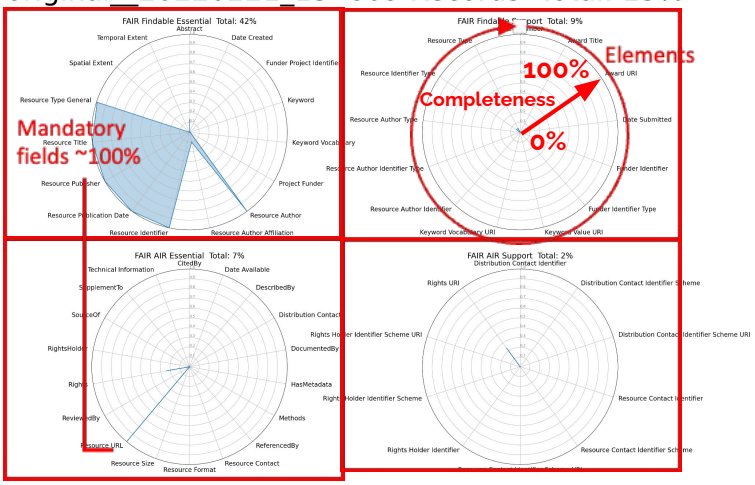
Starting Point



METADATA
GAME CHANGERS

Metadata Starting Point: Mandatory Fields

original__20220221_15 609 Records Total: 15%



*DRUM Repository University of Minnesota

** Accessible, Interoperable, Reusable

Completeness of DataCite metadata* in four categories:

Findable Essential

Findable Supporting

AIR** Essential

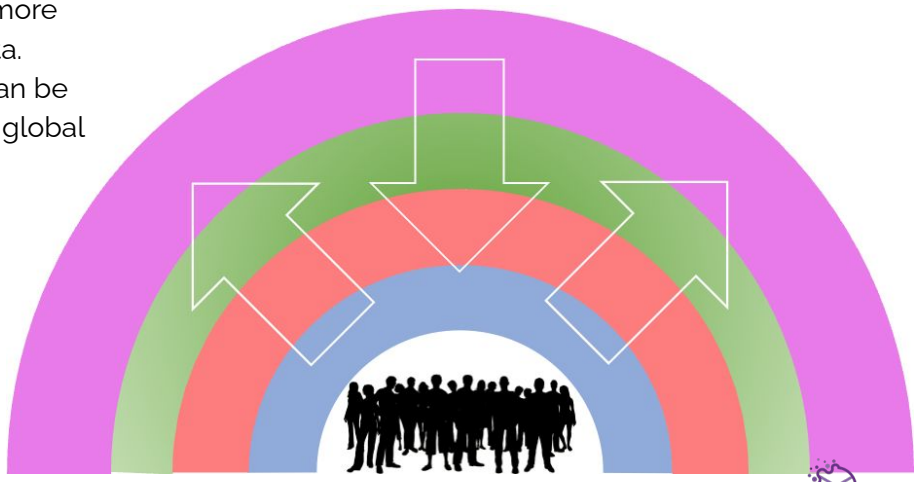
AIR** Supporting

The observations clearly indicate that DataCite metadata is currently dominated by fields required for identification and citation, the mandatory fields.

METADATA GAME CHANGERS

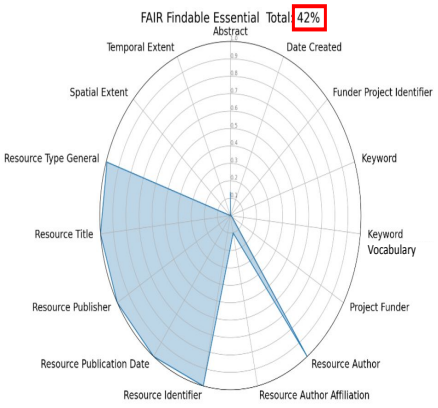
Complete Metadata Flows Out

We know that the institutional repositories have more complete metadata. These metadata can be used to enrich the global infrastructure.

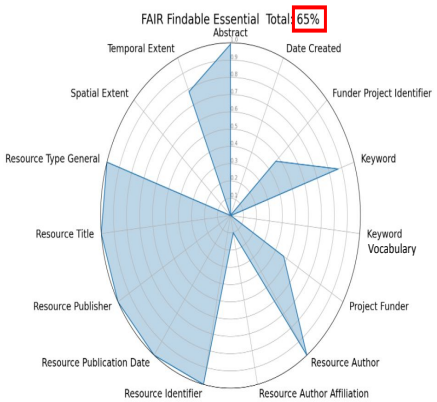


METADATA
GAME CHANGERS

DRUM Metadata: Findable Essential

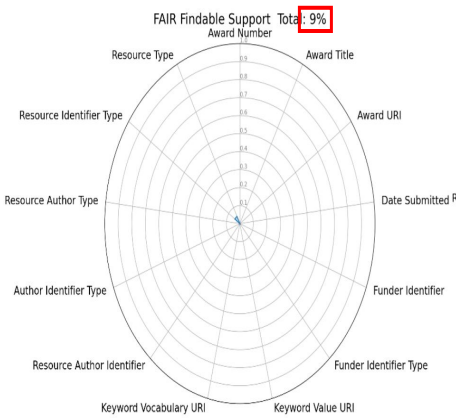


Before

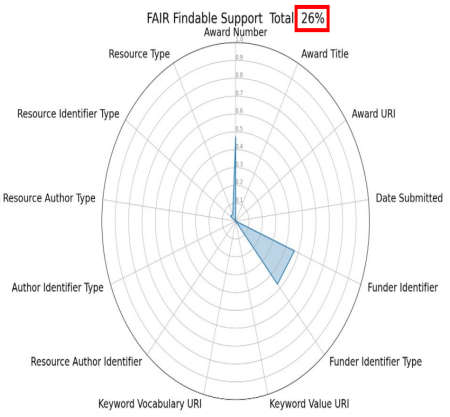


After

DRUM Metadata: Findable Supporting

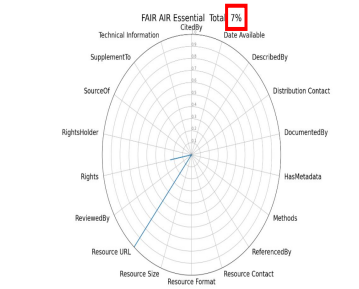


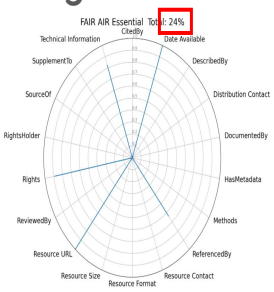
Before

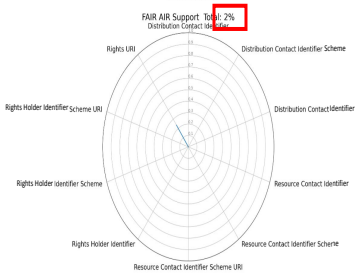


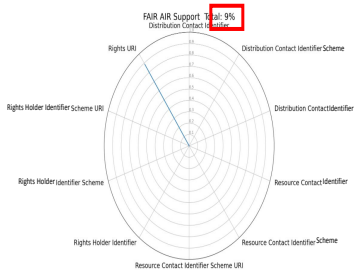
After

DRUM Metadata: Findable Supporting







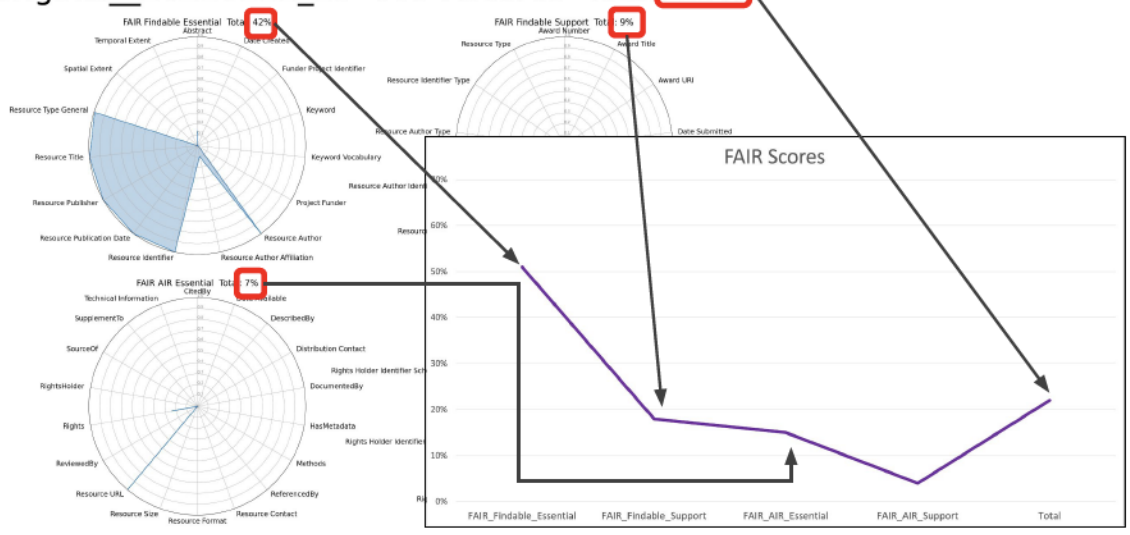


Before

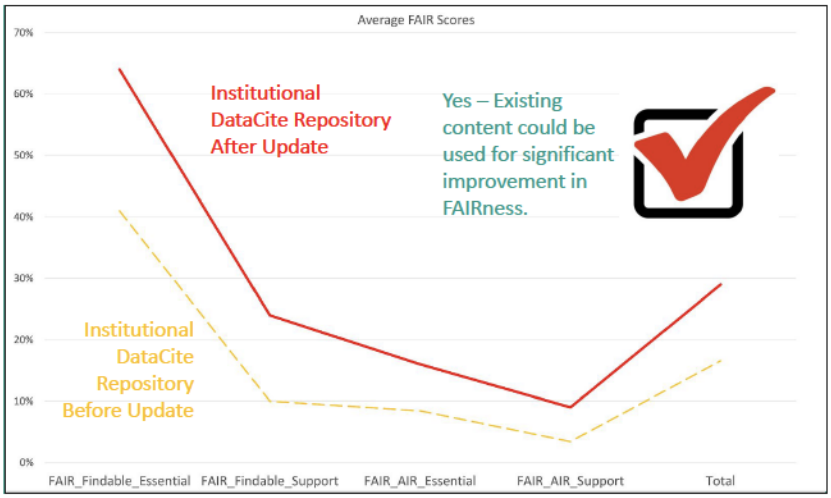
After

Visualizing FAIRness

original__20220221_15 609 Records Total: 15%

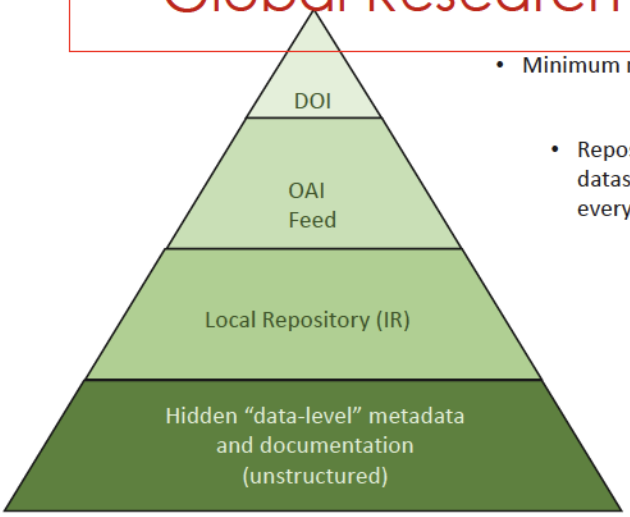


Real-World Improvements in the Global Infrastructure!

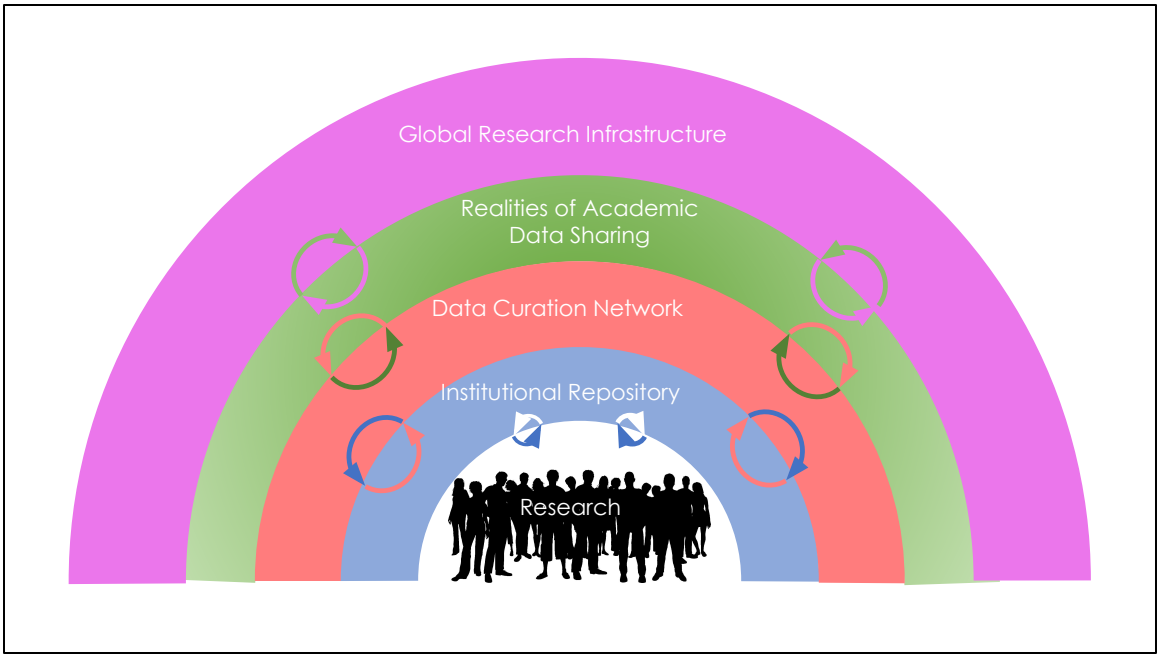


Metadata Mountain

Global Research Infrastructure



- Minimum metadata required for DOI
- Repository streams a subset of metadata about a dataset (more than required for a DOI, but not everything)
- Full set of metadata is collected / maintained within a system (e.g. Dublin Core)
- Description of the data captured within files to support reuse (e.g. readme files)



Act locally, think globally is a mantra for activists worldwide, who recognize the importance of local, community building activities that can have national and global impacts. It is time that institutional repositories, which are by definition locally constrained, similarly adopt this mantra. The spectrum of metadata quality, i.e. completeness, strengthens over time, as users reuse data and contribute to metadata by adding new metadata elements. As local metadata improves and grows over time, users can find and develop connections within data not previously available to them. By feeding local IR metadata into the global data infrastructure, the global infrastructure starts giving back in the form of these connections. Furthermore, there are benefits for stakeholders at every point the metadata spectrum: local IRs engage with community builders such as the DCN, local IRs provide feedback to member organizations such as ARL, and tools for data interoperability and reusability are built for both the local and global infrastructures.

Questions?

Ted Habermann
https://orcid.org/0000-0003-3585-6733
ted@metadatagamechangers.com

Mikala Narlock
https://orcid.org/0000-0002-2730-7542
mnarlock@umn.edu

Shawna Taylor
https://orcid.org/0000-0002-9842-7867
staylor@arl.org

Sarah Wright
https://orcid.org/0000-0002-1502-131X
sjw256@cornell.edu