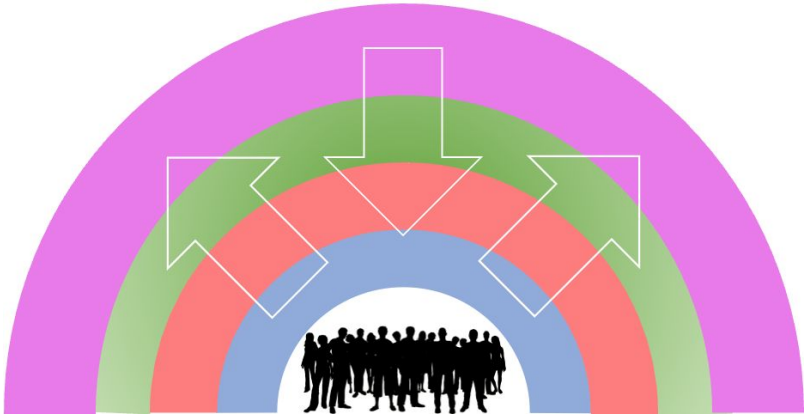
Think Globally, Act Locally: The Importance of Elevating Data Repository Metadata to the Global Infrastructure



Open Repositories 2022

Panel Session

Sarah Wright, Research Data & Life Sciences Librarian, Cornell's Albert R. Mann Library

eCommons
Open scholarship at Cornell

Mikala Narlock, Director
Data Curation Network (DCN) |
University of Minnesota Libraries

DATA CURATION NETWORK

Shawna Taylor, RADS Project Manager
Association of Research Libraries

ASSOCIATION OF RESEARCH LIBRARIES

Ted Habermann, Founder & CTO
Metadata Game Changers

METADATA GAME CHANGERS

eCommons - Local Institutional Repository

eCommons
Open scholarship at Cornell

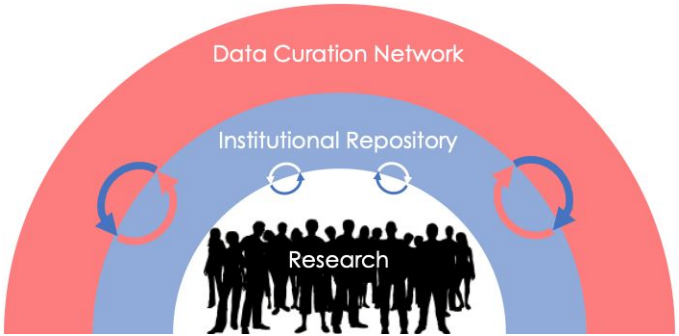
- How can we make data as FAIR as possible considering local Institutional Repository (IR) constraints?
  - Busy researchers
  - Gaps in local infrastructure
  - Data-specific metadata and documentation concerns

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

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

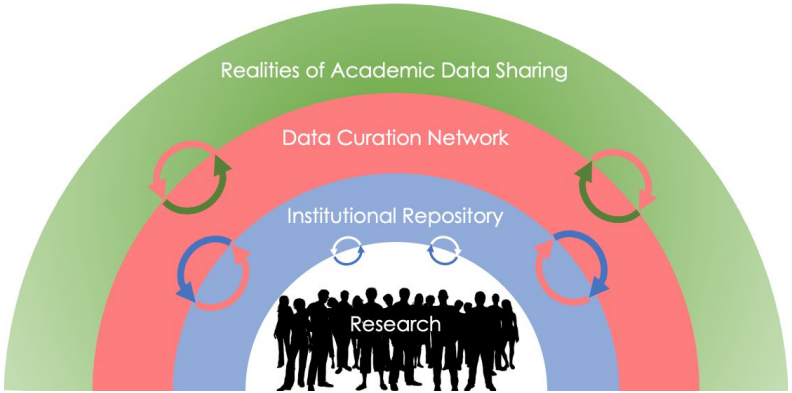
We strive to be a trusted community-led network of curators **advancing open research** by making data more ethical, reusable, and understandable.

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



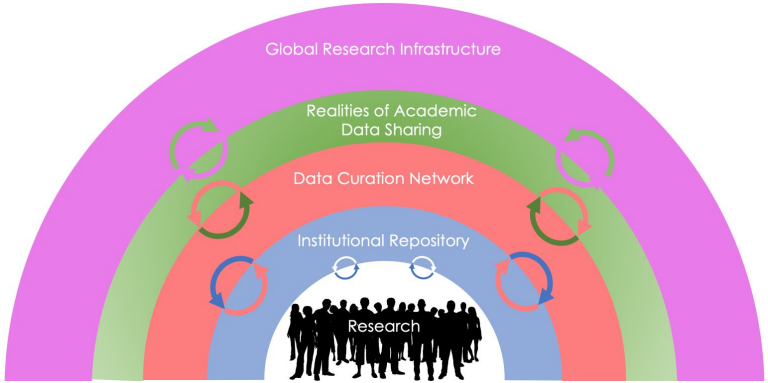
- Identify opportunities to improve and connect local IR meta(data) to the global research infrastructure.

Metadata Analysis

METADATA GAME CHANGERS

ASSOCIATION OF RESEARCH LIBRARIES

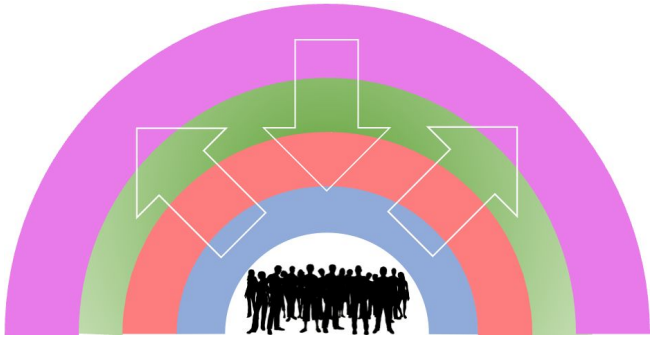
- RADS project metadata analysis
  - DataCite was queried to determine **where** researchers are sharing their research data
  - Metadata **completeness** analyzed using FAIR recommendation for DataCite metadata

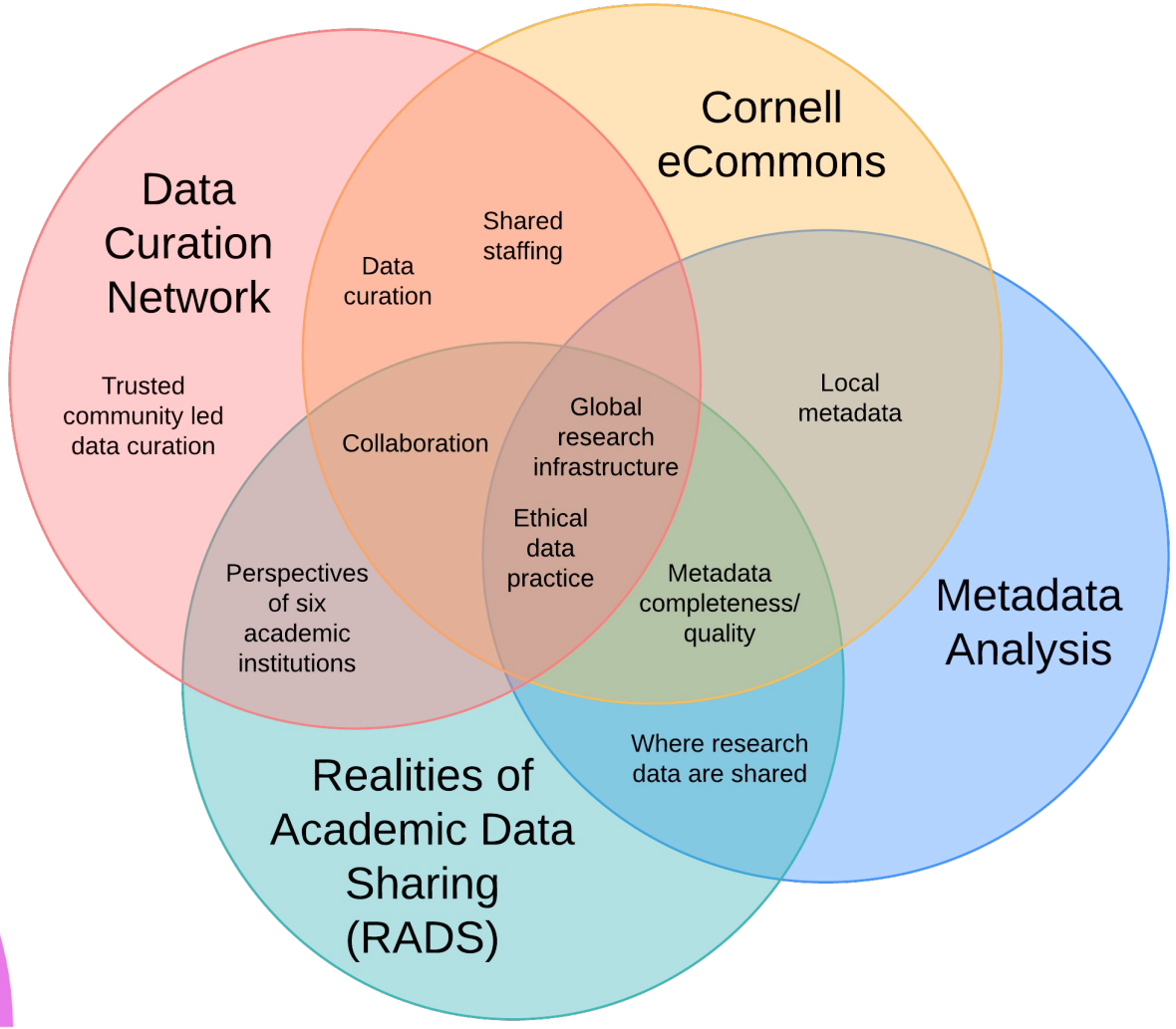


- Serendipitous Improvements
- Can we increase content in the global infrastructure while minimizing impact on the IRs?

Shared Areas of Interest or Practice

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





Data Curation at Cornell

Preserve and share your data in eCommons

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

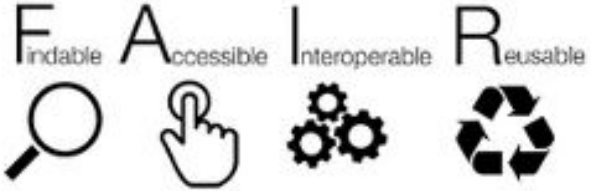
CORNELL UNIVERSITY LIBRARY

eCommons

Open scholarship at Cornell

ORCID

Connecting Research and Researchers



doi

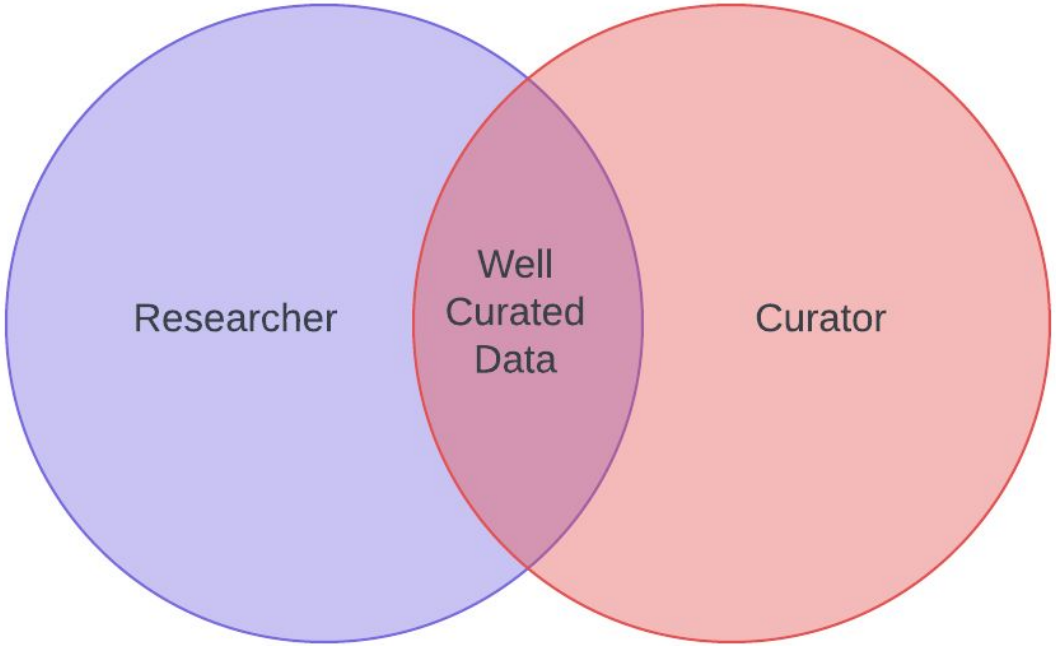
Data Curation

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

Who is involved in data curation?

- Disciplinary expertise
- Research context
- Documentation for reuse / reproducibility



- Disciplinary and/or data format expertise
- Best practices for data sharing and archiving
- Metadata for FAIR

What's in it for the researcher?

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

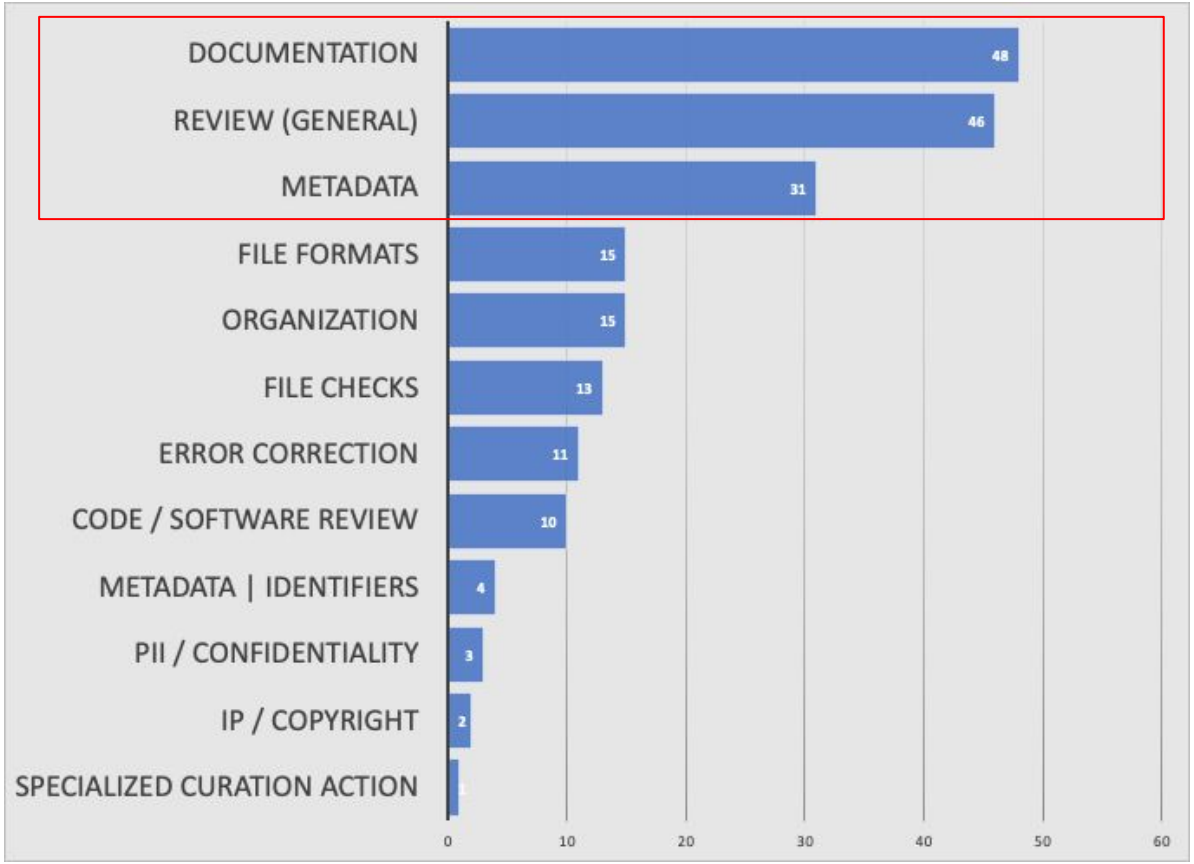
- **Persistent Identifier** (37% happens)
- **Software Registry** (41% happens)
- **File Audit** (16% happens)
- **Contextualization** (38% happens)
- **Code Review** (38% happens)

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

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



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

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

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

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).

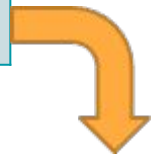
Publisher Requirements

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

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%

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

Local → Global

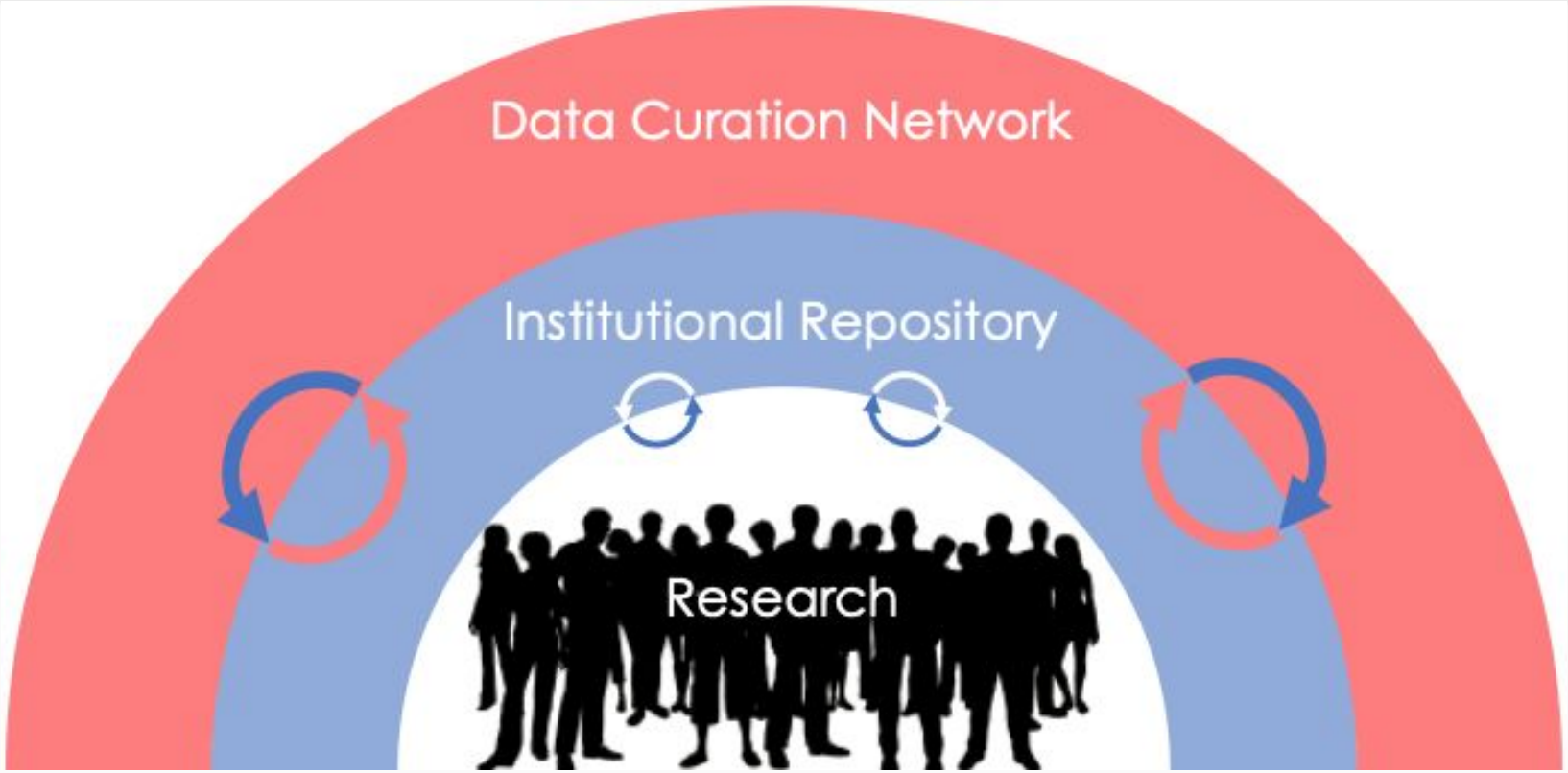
Metadata	eCommons	DataCite (2016-2020)	DataCite (2021-2022)
*Author	✔	✔	✔
*Title	✔	✔	✔
Abstract	✔		
Funding	✔		✔
Suggested Citation	✔	✔	✔
Keywords	✔		
Author ORCIDs			✔
Author Affiliation (RORs)			✔
Links to Related Content	✔		
License	✔		✔
Resource type	✔	✔	✔
Readme	✔		

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!

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


Data Curation Network
Institutional Repository
Research

DATA CURATION NETWORK

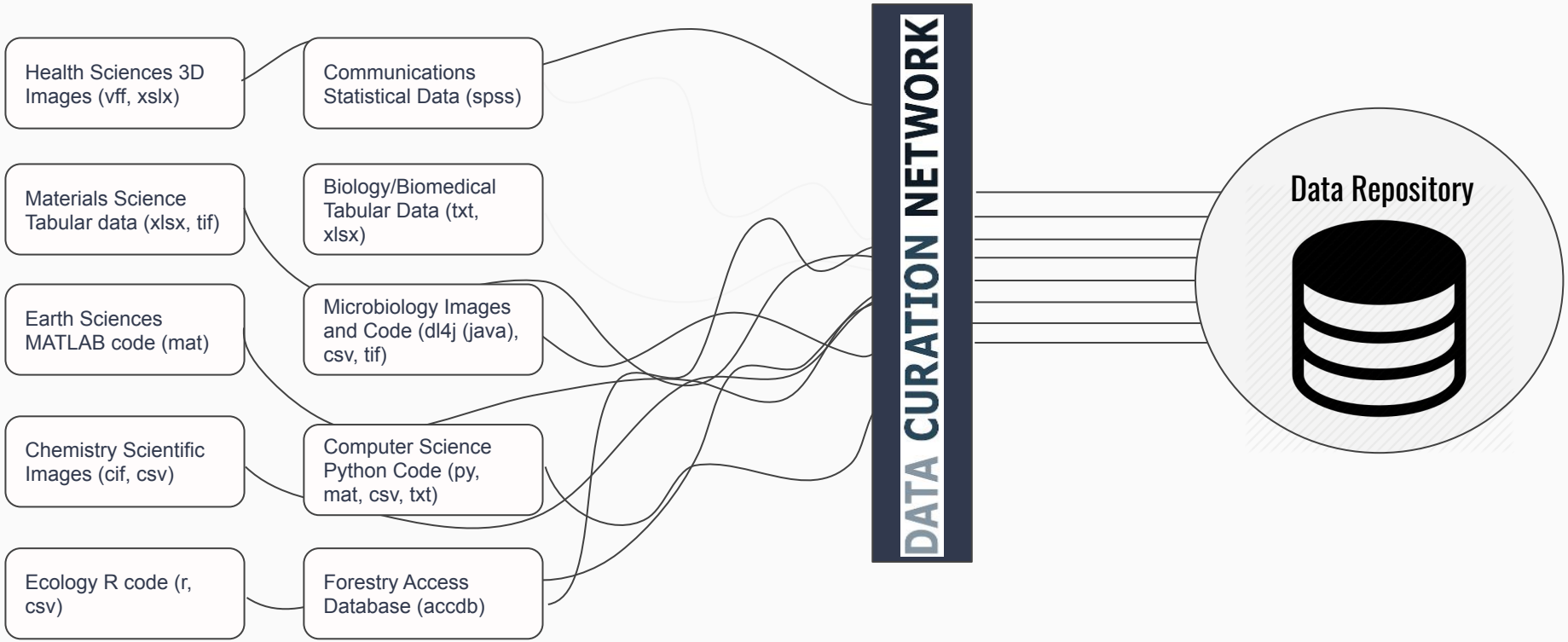
Mission

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

Ethical. Reusable. Better.

ALFRED P. SLOAN FOUNDATION

Curation at Scale



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

C U R A T E (D)

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:

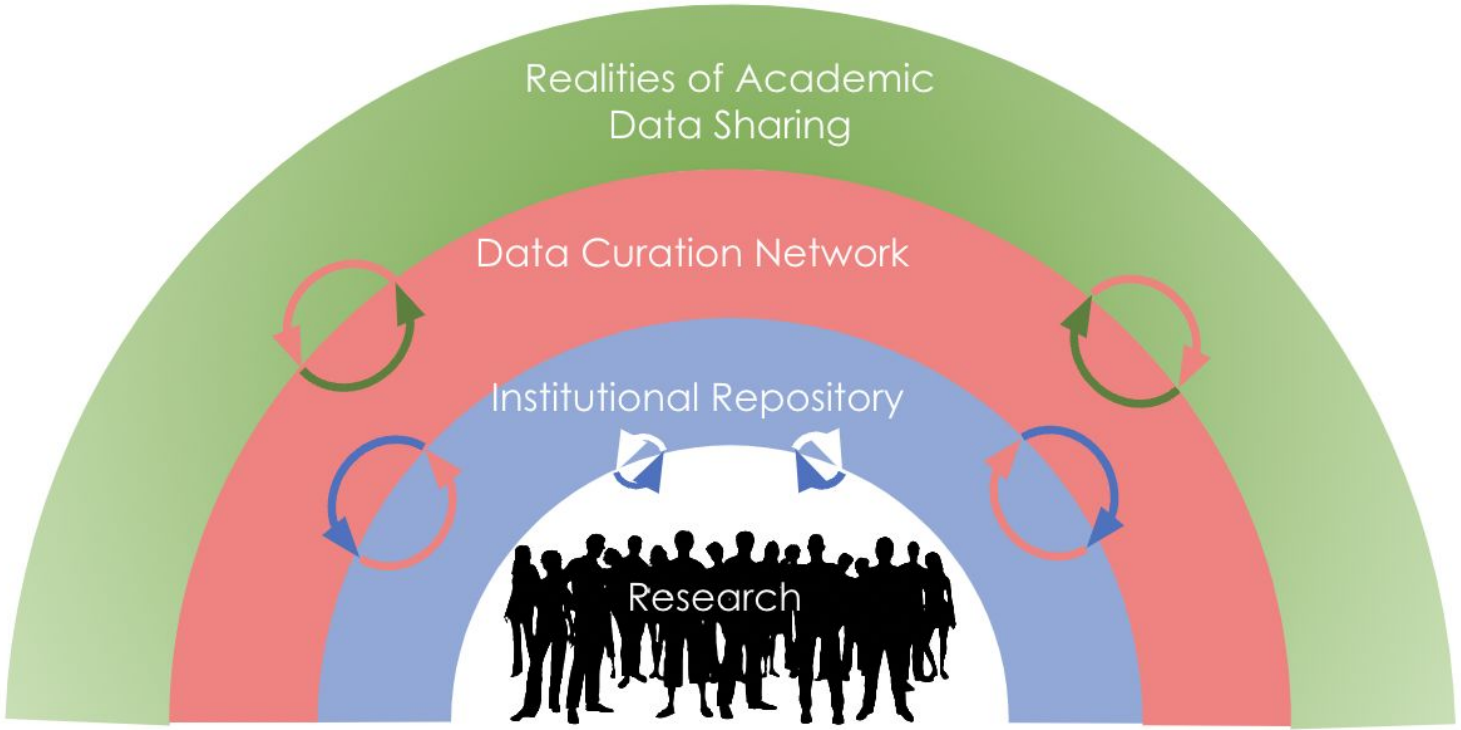
- Review to ensure data is in scope for the repository
- Inventory the contents of the data files (e.g., open and sample the files or code)
- Verify all metadata provided by the researcher; check available documentation

Key Ethical Considerations

- 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., contested boundaries, historical and current political situations)

(Current) DCN Community

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



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

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

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

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

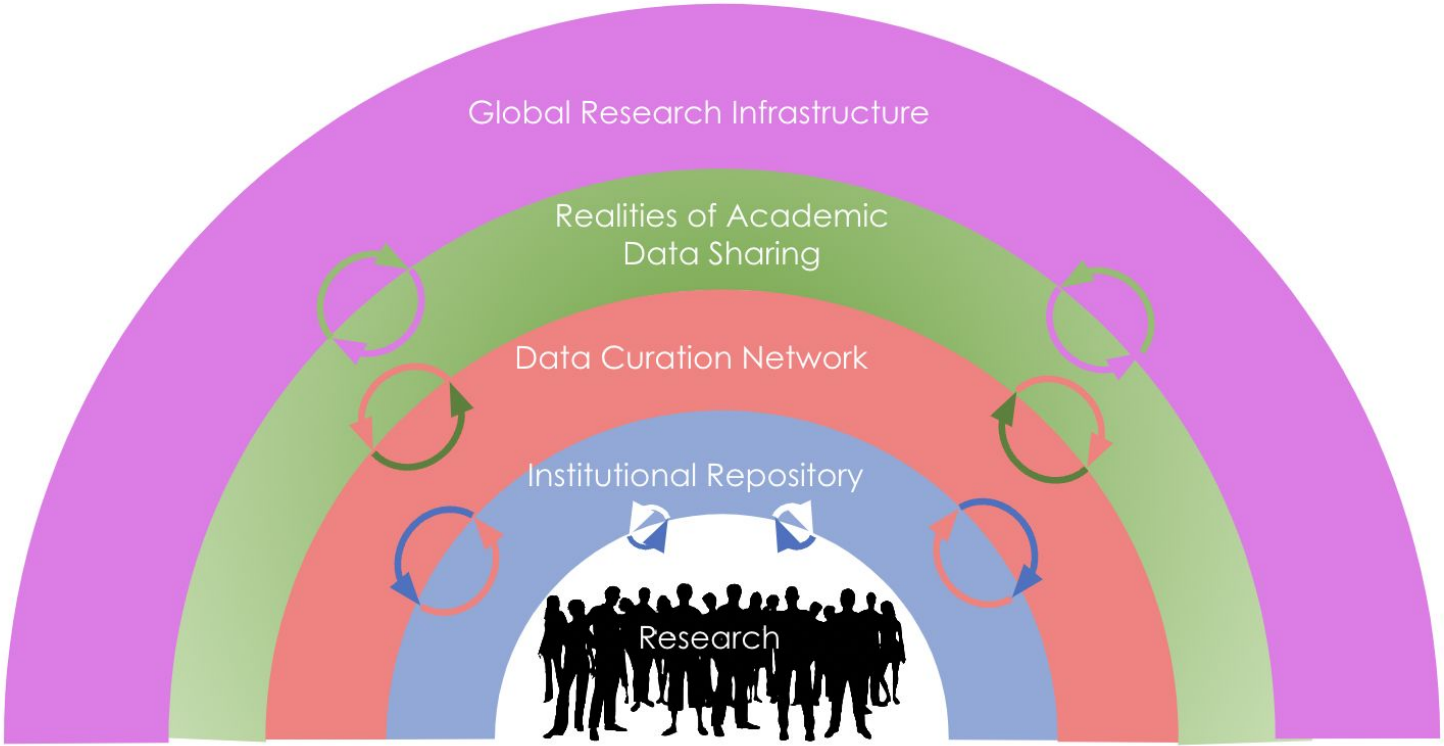
ASSOCIATION OF RESEARCH LIBRARIES

RADS: Serendipitous Improvements

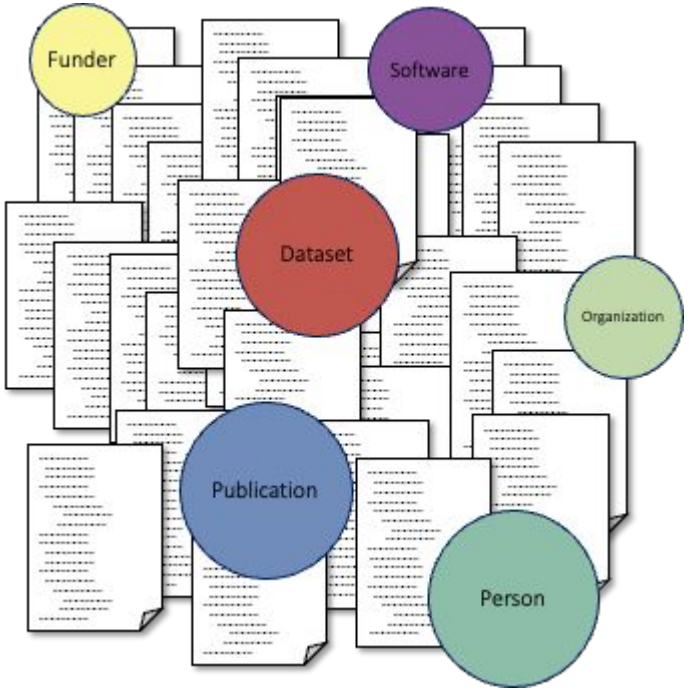
- ✓ Used DataCite's API to transfer metadata from local repositories to DataCite
- ✓ No new metadata content created
- ✓ Can we make further improvements in the local to global transfer using PIDs or documentation such as README files?

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

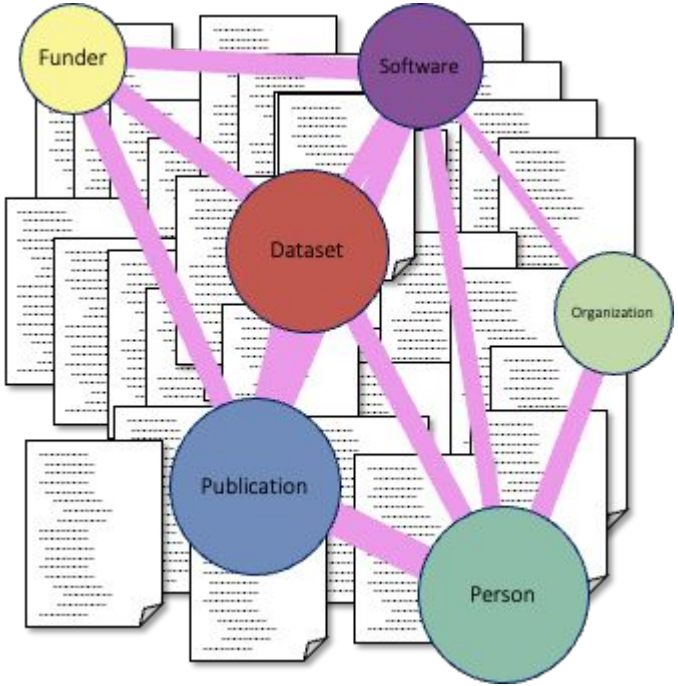
ASSOCIATION OF RESEARCH LIBRARIES


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

DataCite Mandatory

- Resource URL
- Resource Title
- Resource Author
- Resource Identifier
- Resource Type General
- Resource Publication Date
- Resource Publisher

FAIR Findable Essential

- Abstract
- Date Created
- Keyword
- Keyword Vocabulary
- Resource Author Affiliation
- Project Funder
- Funder Project Identifier
- Temporal Extent
- Spatial Extent

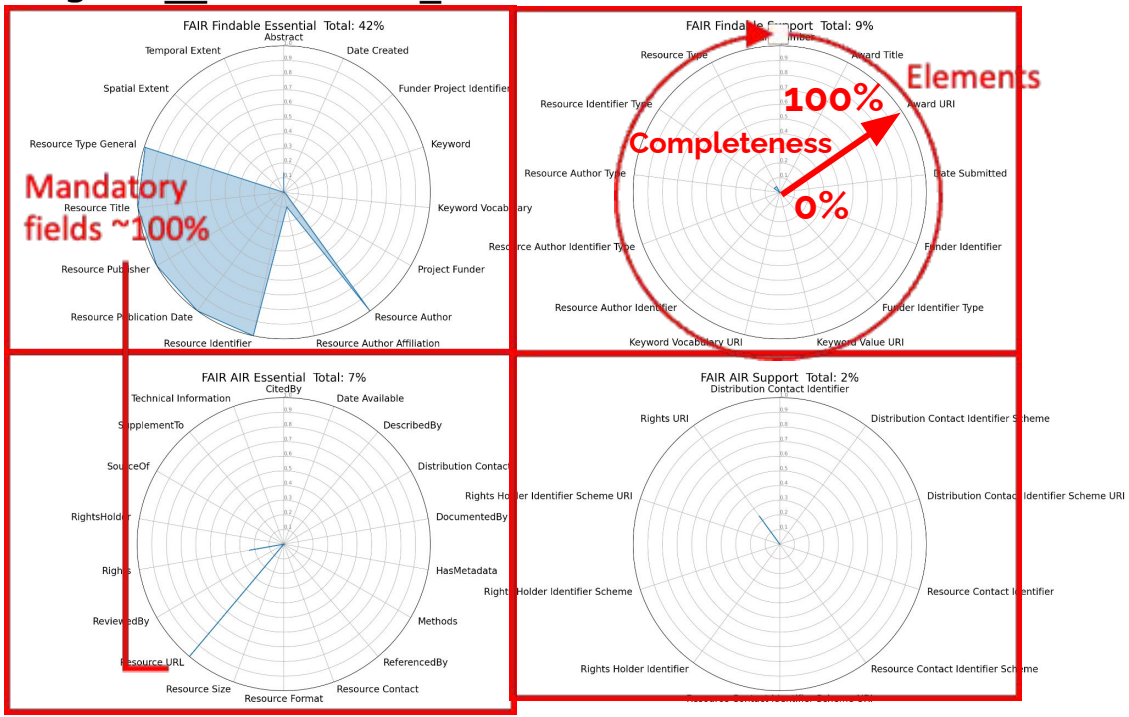
Connections

- Cites
- IsCitedBy
- IsSupplementTo
- IsSupplementedBy
- IsContinuedBy
- Continues
- IsNewVersionOf
- IsPreviousVersionOf
- IsPartOf
- IsSourceOf
- Describes
- IsDescribedBy
- HasVersion
- IsVersionOf
- Requires
- IsRequiredBy
- Obsoletes
- IsObsoletedBy
- IsVariantFormOf
- IsOriginalFormOf
- IsIdenticalTo
- HasMetadata
- IsMetadataFor
- Reviews
- IsReviewedBy
- IsDerivedFrom
- HasPart
- IsPublishedIn
- IsReferencedBy
- References
- IsDocumentedBy
- Documents
- IsCompiledBy
- Compiles

METADATA GAME CHANGERS

Metadata Starting Point: Mandatory Fields

original__20220221_15 609 Records Total: 15%



*DRUM Repository University of Minnesota

Completeness of DataCite metadata* in four categories:

Findable Essential

Findable Supporting

AIR** Essential

AIR** Supporting

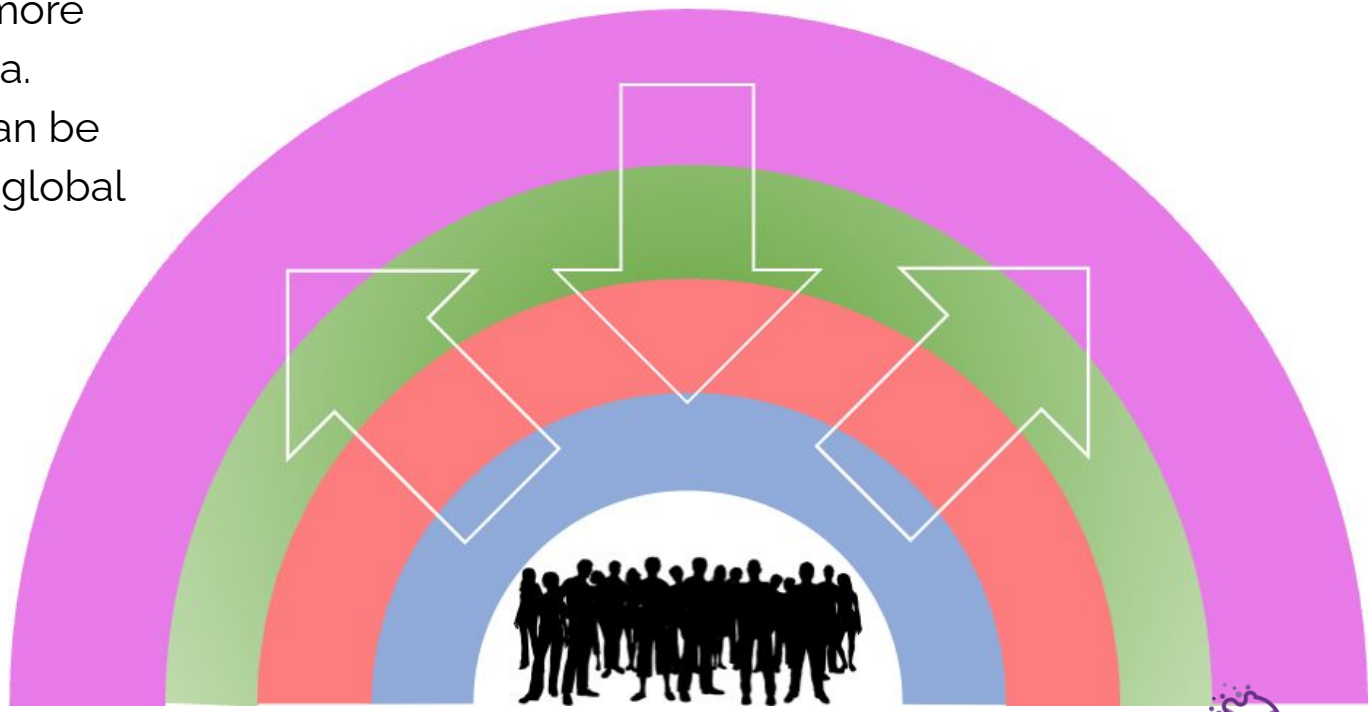
The observations clearly indicate that DataCite metadata is currently dominated by fields required for identification and citation, the mandatory fields.

** Accessible, Interoperable, Reusable

METADATA GAME CHANGERS

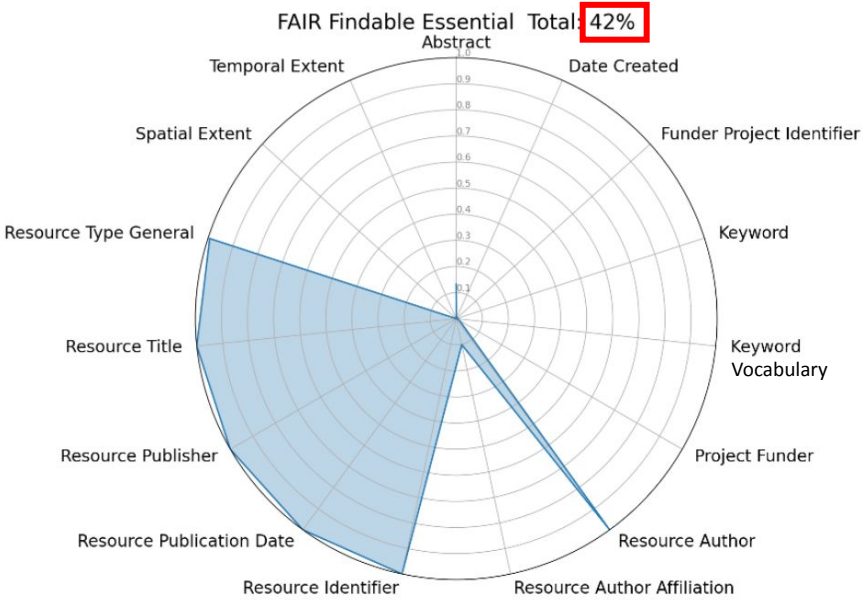
Complete Metadata Flows Out

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

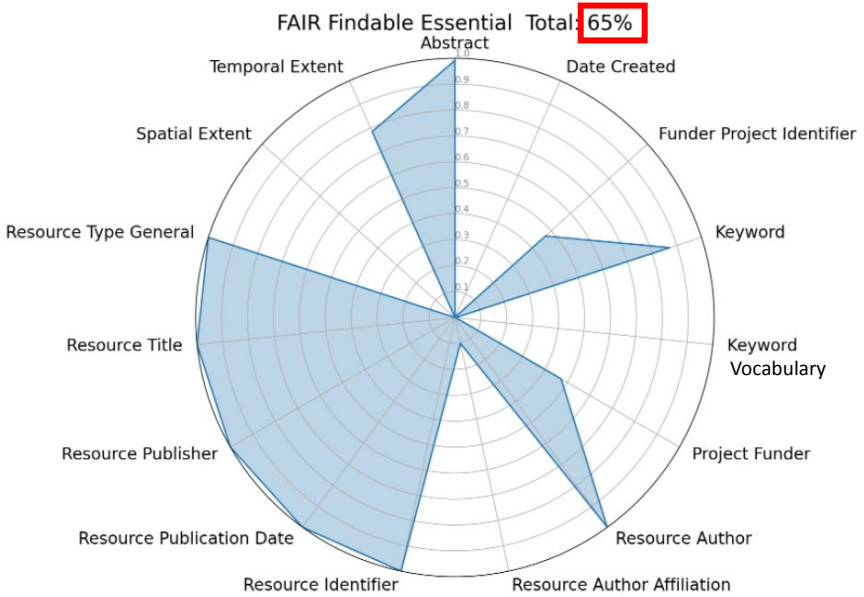


METADATA
GAME CHANGERS

DRUM Metadata: Findable Essential

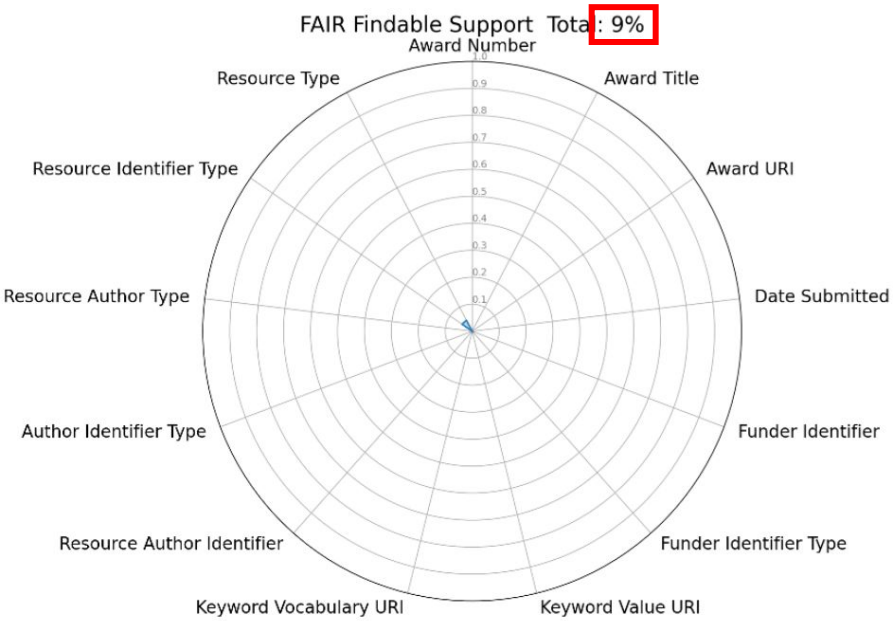


Before

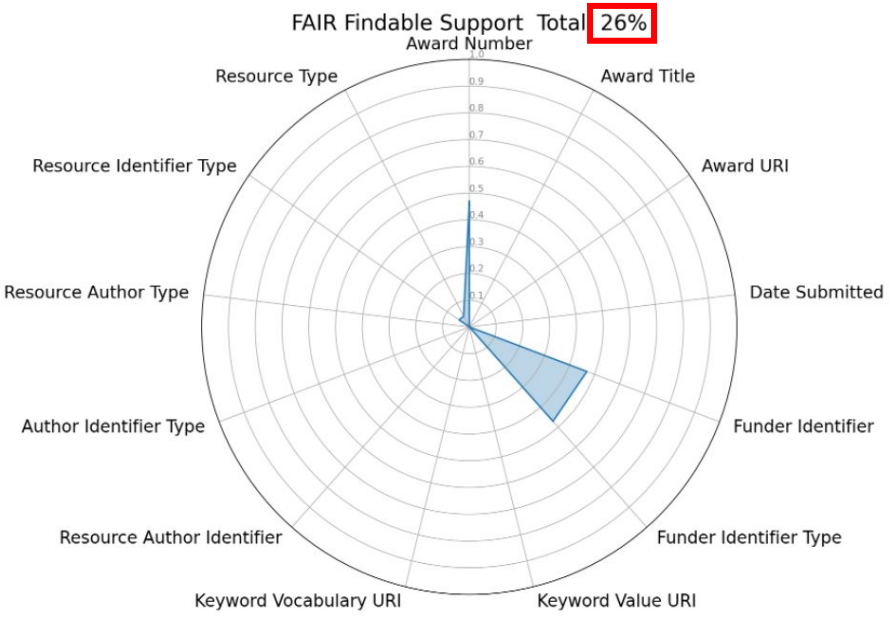


After

DRUM Metadata: Findable Supporting



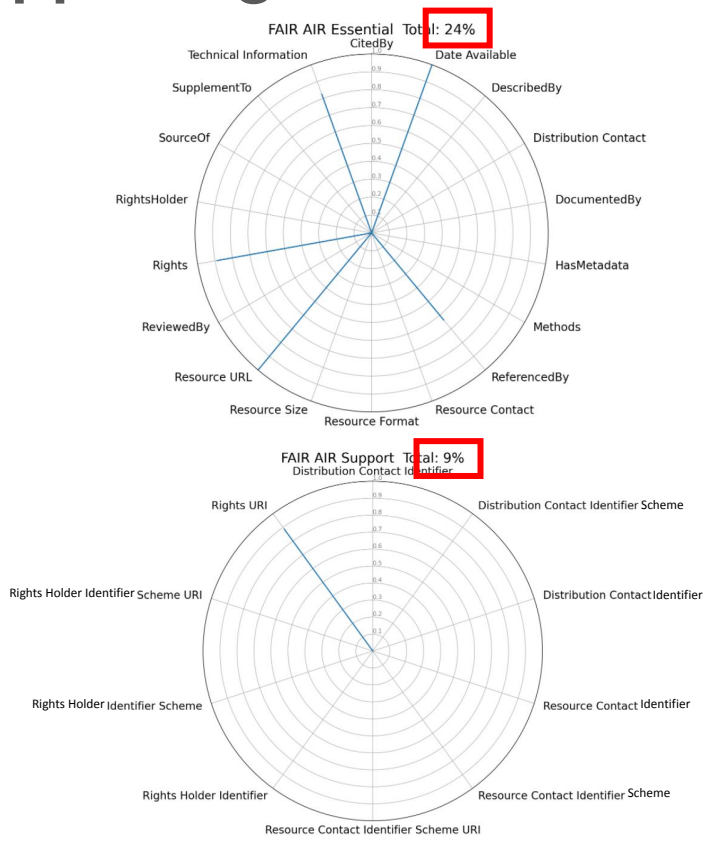
Before



After

DRUM Metadata: Findable Supporting

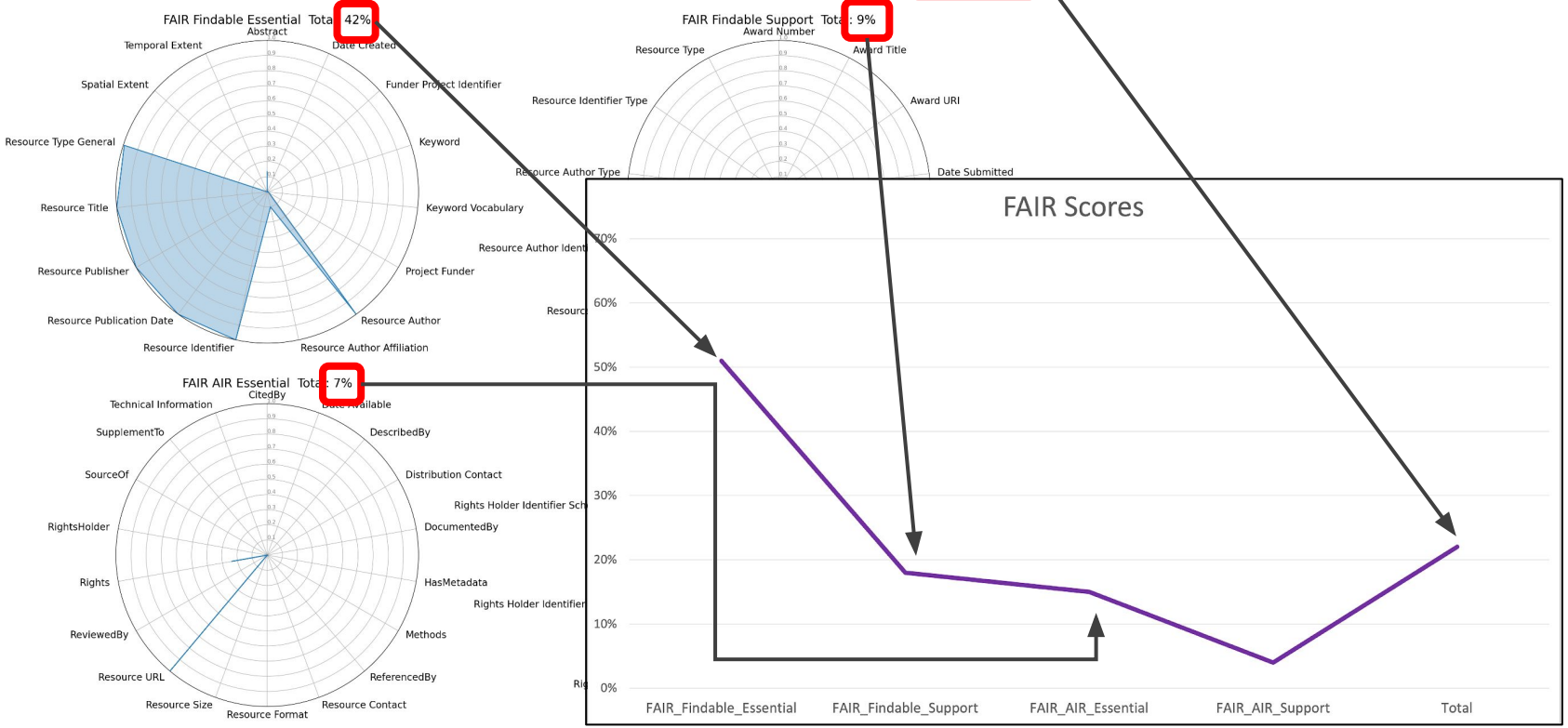
Before



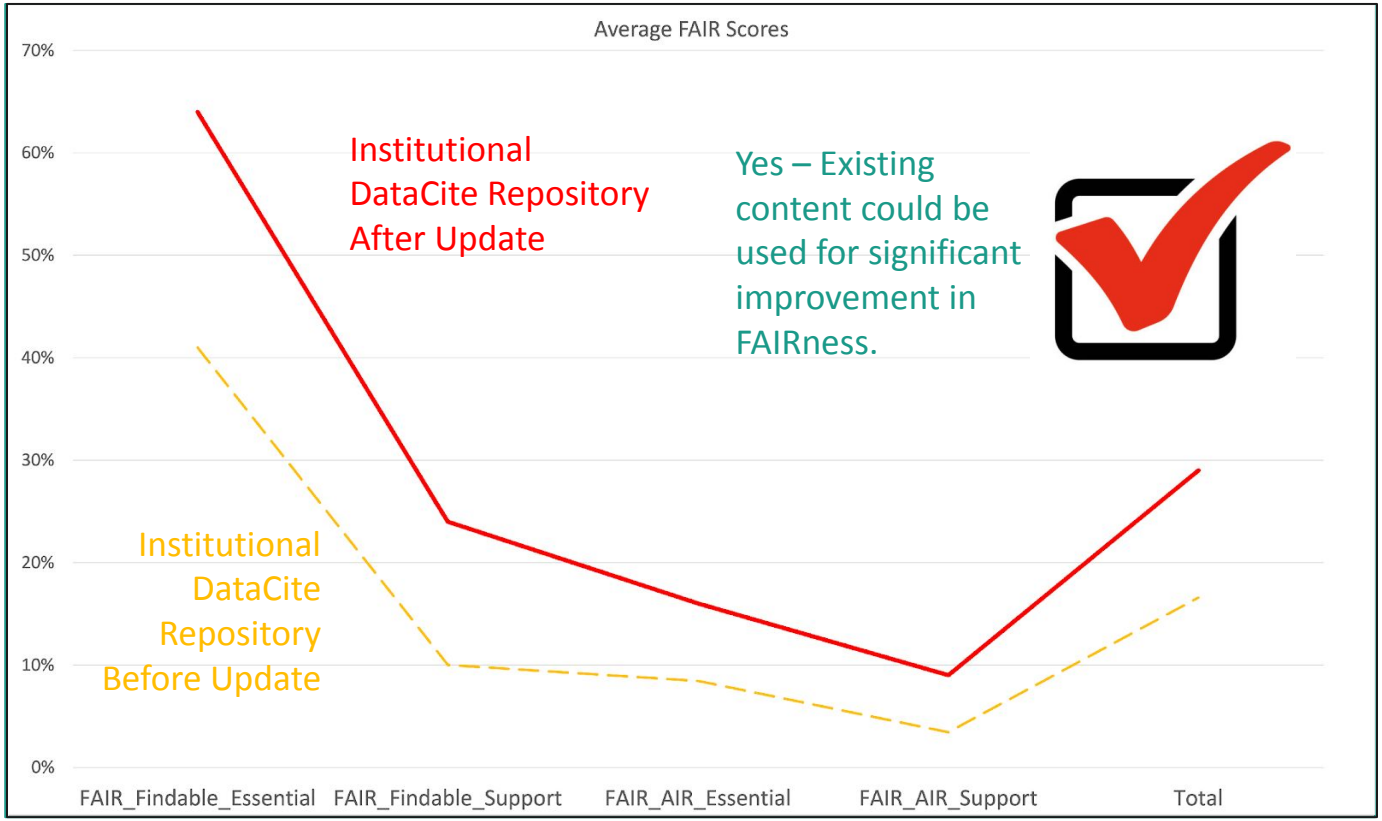
After

Visualizing FAIRness

original__20220221_15 609 Records Total: 15%

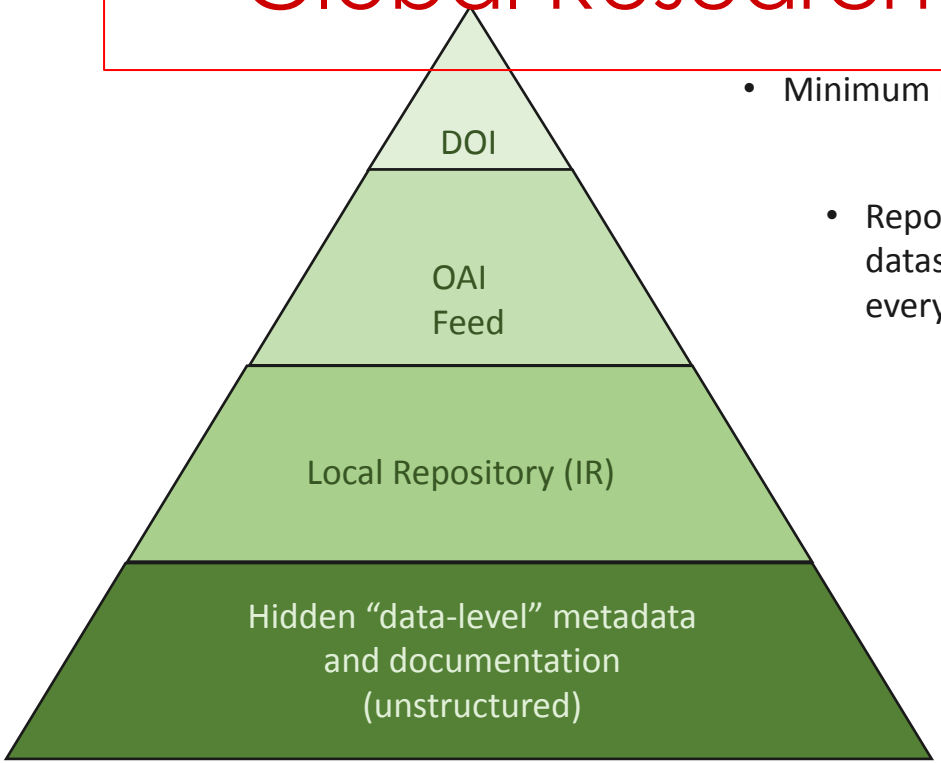


Real-World Improvements in the Global Infrastructure!

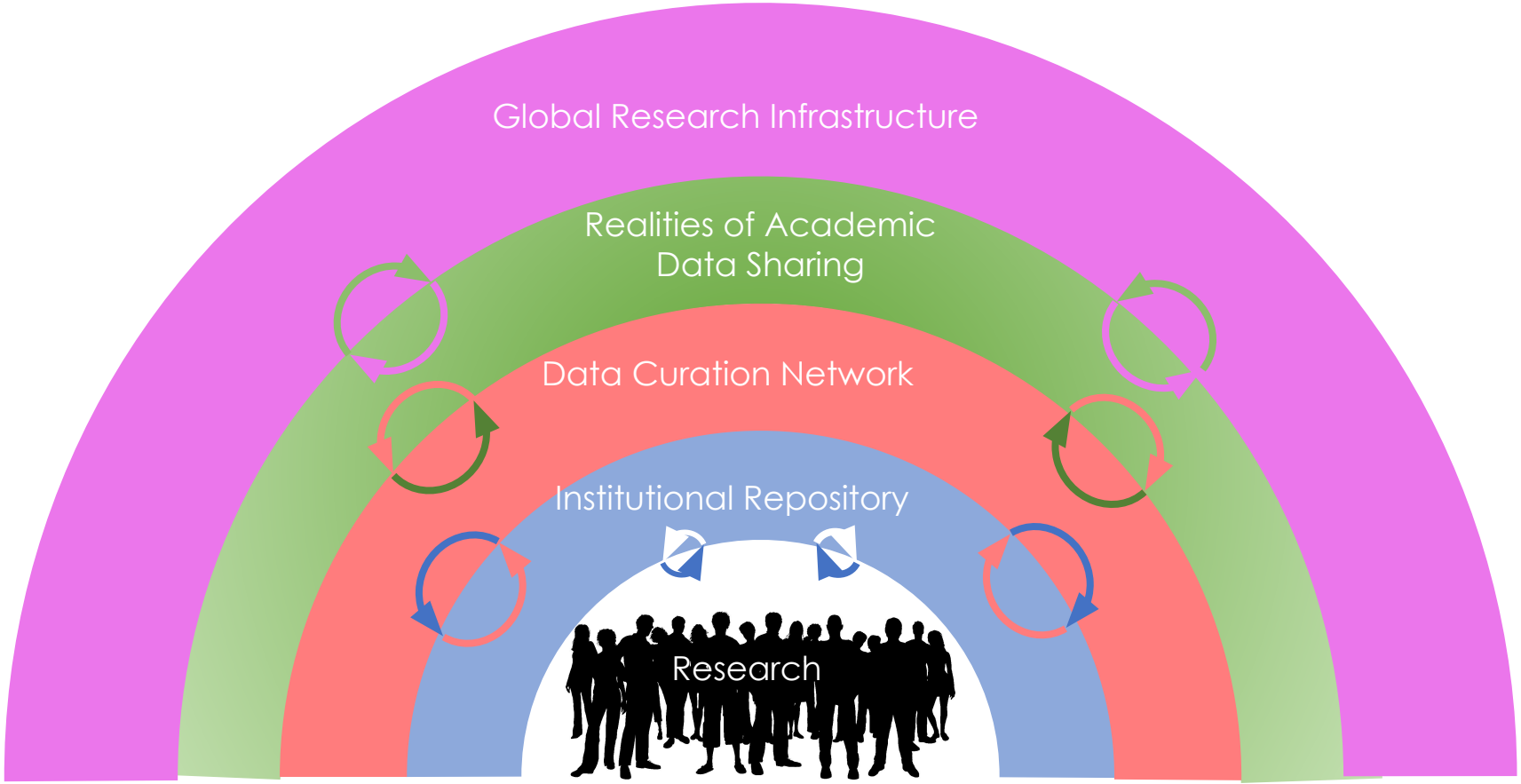


Metadata Mountain

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)


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

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