ISSN 2161-3974
Full-Length Paper

*Journal of eScience Librarianship* 12(3): e739
DOI: https://doi.org/10.7191/jeslib.739

Journal of eScience Librarianship
putting the pieces together: theory and practice

# Connecting Repositories to the Global Research Community: A Re-Curation Process

**Ted Habermann**, Metadata Game Changers, Boulder, CO, USA, ted@metadatagamechangers.com

## Abstract

Over the last decade, significant changes have affected the work that data repositories of all kinds do. First, the emergence of globally unique and persistent identifiers (PIDs) has created new opportunities for repositories to engage with the global research community by connecting existing repository resources to the global research infrastructure. Second, repository use cases have evolved from data discovery to data discovery and reuse, significantly increasing metadata requirements.

To respond to these evolving requirements, we need retrospective and on-going curation, i.e. re-curation, processes that 1) find identifiers and add them to existing metadata to connect datasets to a wider range of communities, and 2) add elements that support reuse to globally connected metadata.

The goal of this work is to introduce the concept of re-curation with representative examples that are generally applicable to many repositories: 1) increasing completeness of affiliations and identifiers for organizations and funders in the Dryad Repository and 2) measuring and increasing FAIRness of DataCite metadata beyond required fields for institutional repositories.

These re-curation efforts are a critical part of reshaping existing metadata and repository processes so they can take advantage of new connections, engage with global research communities, and facilitate data reuse.

**Received**: June 6, 2023 **Accepted**: November 29, 2023 **Published**: December 20, 2023

**Data Availability**: The data used in this work were provided by Institutional Repositories and retrieved from DataCite using the public DataCite API during late 2021 and early 2022. These metadata are constantly being maintained and change over time so the now out-of-date metadata are not available. The software used in the analysis was customized to accommodate translation of metadata models used in each repository to the DataCite model and uploading that metadata to DataCite. It is not easily generalizable.

## Introduction

Curation is an important role that data repositories implement in a variety of ways. What does curation include? Where does curation occur in the data life cycle? How is the impact of curation measured? These questions are answered in many ways across the broad spectrum of repositories.

Many conceive of curation as a process that happens early in the preservation part of the data life cycle. Typically, a dataset is submitted to a repository and curation is the process of working with the researcher to enhance the dataset and related metadata to increase suitability for sharing and long-term preservation while increasing the potential for reuse. The result of this curation is an improved data package being accepted into the repository.

The recent emergence of a family of globally unique and persistent identifiers is making it possible to connect research objects from around the world in new and novel ways. However, making these connections requires that unique identifiers for the research objects exist in metadata. While recently created metadata may include them, they are generally absent or rare in metadata created before the identifiers existed, leaving many existing research objects outside of the connected global research community.

The Data Curation Network (Johnston et al. 2018) is made up of curation and digital curation experts from many research institutions. Together, they have proposed and promulgated a model of digital curation which includes seven steps (CURATED): Check files and code, Understand the data, Request missing information, Augment metadata, Transform formats, Evaluate for FAIRness, and Document all activities that are designed to be carried out as a dataset is submitted to and accepted into a repository. This curation process, referred to here as *Record Curation*, clearly results in improved quality of data in many institutional repositories.

The introduction of identifiers as critical metadata elements changes the landscape considerably, adding work to the "Augment metadata" step in record curation processes. Identifiers can be found or created and added to new metadata going forward, but existing records remain without these identifiers. Bringing these existing records up to current standards requires *repository re-curation*, in this case, curating existing records again by augmenting their metadata to include new identifiers.

Re-curation is different from record curation in several ways. First, it involves connections to a wide variety of metadata sources in a variety of metadata dialects (DataCite, Crossref, ORCID, ROR, OpenAlex, ScholeXplorer, etc.). Second, re-curation is an on-going process as the landscape continues to evolve with new kinds of objects getting identifiers (e.g. samples, instruments, projects), new communities using identifiers in new ways, and identifiers migrating between types (e.g. International Generic Sample Number, IGSNs, becoming Digital Object Identifiers, DOIs). In many cases, these differences mean that new tools are required for facilitating the re-curation work.

Examples from Dryad and DataCite will demonstrate approaches for discovering identifiers for papers (DOIs), people (ORCIDs), organizations and funders (RORs) and re-curating existing metadata to include those identifiers as well as additional content supporting reuse. This will include demonstration of simple metrics for measuring the impact of this re-curation and demonstrating the benefits to repository communities and managers.

## Re-Curating Identifiers in the Dryad Data Repository

### *Connections*

The Dryad Data Repository formed during 2008 with the goal of providing curation and preservation for datasets associated with published scientific articles. The original metadata model was simple. The dataset submission guidelines were: "*To deposit data, simply mail it to submit@datadryad.org. Please include a title and short description for each file, as well as a reference to the relevant publication.*"

The decision to require references to "the relevant publication" was a critical one that fundamentally changed the nature of Dryad from an isolated data repository (Figure 1A) into a connected virtual repository of data and articles (Figure 1B). Note that these connections were made in the form of references, like connections between articles in the literature had always been made.

During 2018 Dryad formed a strategic partnership with the California Digital Library (Dryad, 2018) and the metadata model and management processes began to evolve, emerging as the "New Dryad" during late 2019 (Dryad, 2019). Part of this evolution included addition of DOIs for the articles related to Dryad datasets which also enabled a richer set of connections to many types of resources (articles, software, preprints, etc.) for Dryad datasets. This evolution is illustrated by the addition of Crossref (C) and DataCite (D) to Figure 1. Identifiers for papers and other research objects were migrated to DataCite as part of the Dryad DataCite repository.

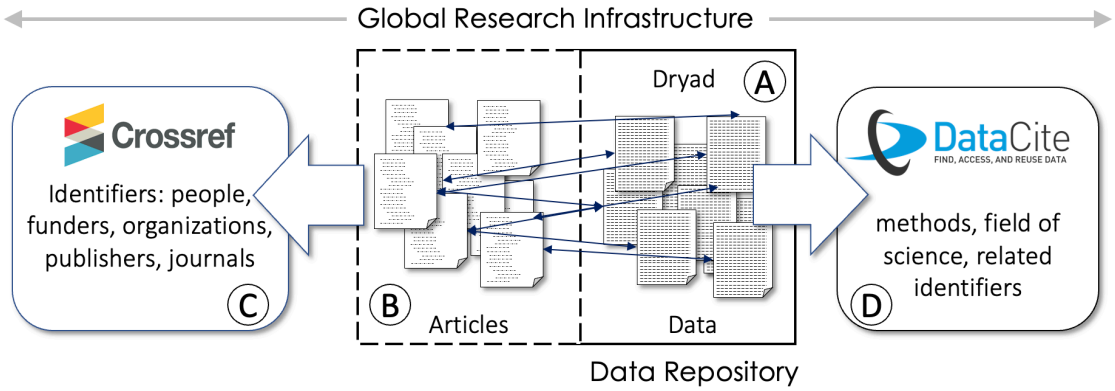


**Figure 1**: Evolution of Dryad from an isolated data repository(A) to a connected virtual repository with data and related papers (B) and then to a connected element of the global research infrastructure with article metadata in Crossref (and other repositories) (C) and dataset metadata in DataCite (D).

It is important to note that connecting Dryad datasets to papers is not the primary purpose of either Crossref or DataCite. Instead, it is a part of the services these identifier infrastructures and their metadata schema provide to their users. Many repositories, journals, and users all over the world use these identifiers for making the same kind of connections building a global PID Graph (Fenner and Aryani 2019).

### *Affiliations and Organizational Identifiers (RORs)*

The original Dryad metadata model (Habermann 2019) focused on connecting multiple data files into packages and administering the preservation of those data packages. It relied on connected papers as critical contributors to the documentation required to discover, understand, and re-use datasets. Even author names and affiliations were not included in the metadata as they were available in the papers.

During 2019 a new community-driven identifier for organizations (ROR) was being developed and Dryad decided to add this new identifier for nearly 100,000 organizations in over 20,000 dataset metadata records (Gould and Lowenberg 2019). Given the Dryad metadata model, re-curating the metadata to add identifiers for organizations required two steps: 1) finding affiliations and 2) using those affiliations to find RORs. Fortunately, the Dryad metadata included connections to Crossref, a standard source for author affiliation strings that could be retrieved using DOIs included in Dryad metadata (Figure 2A). This resulted in a long list of affiliation strings with the well-known ambiguity and complexity of different spellings, abbreviations, and extraneous text.

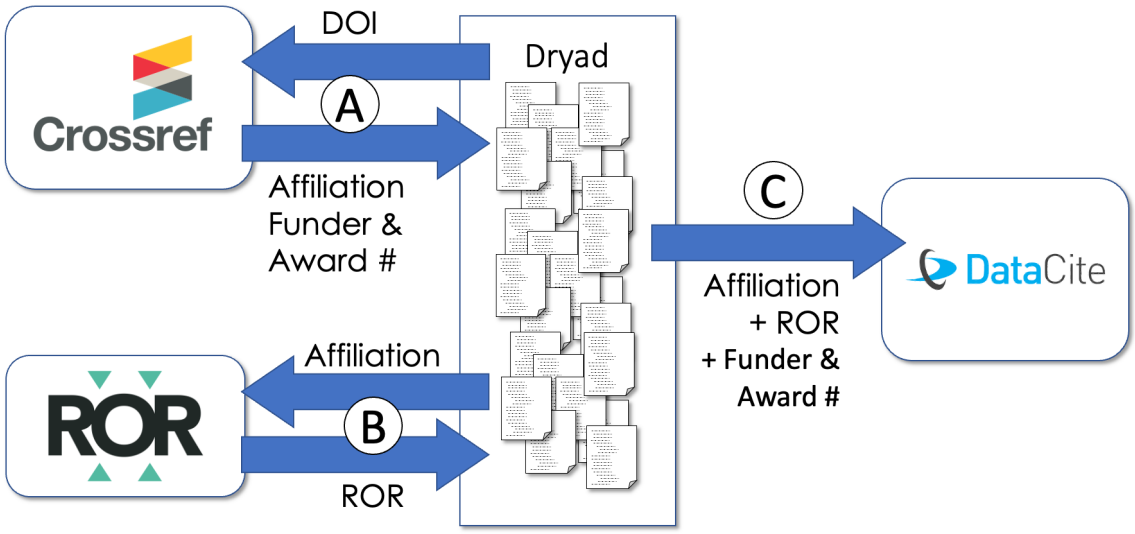


**Figure 2**: Re-curation of Dryad with consisted of three steps. A) DOIs from Dryad were used to search Crossref for affiliations, funder names, and award numbers, B) those affiliations were used to search ROR for organizational identifiers (RORs), C) the affiliations, RORs, funder names and identifiers were added to DataCite metadata.

This was early in the days of ROR, so approaches were developed to searching these affiliations to convert them to RORs (Figure 2B). This search resulted in 91% of the Dryad datasets having RORs for at least one organization (Lowenberg and Habermann 2019). The New Dryad was using the DataCite metadata model which includes authors, affiliations, and affiliation identifiers, so the new content could be easily added to DataCite to become available to the global research infrastructure (Figure 2C).

Since that initial effort, Dryad has incorporated RORs into their standard dataset submission interface (Gould and Lowenberg 2019), collecting RORs for most incoming dataset affiliations. At the same time, the number of organizations with RORs is increasing (e.g. 2,010 new organization records during 2022) and methods for finding RORs from affiliations are improving, so on-going re-curation is needed to keep the RORs current.

### *Identifiers for Funders and Awards*

NIdentifiers for funders and awards are metadata elements critical for bringing together resources supported by a particular funder or created in any funded project. Like affiliations, funder metadata (names and award numbers) and how they are used in research objects and metadata vary significantly. Thus, funders can benefit from creating and using unique identifiers just like research organizations are benefitting from RORs.

This problem was addressed by Elsevier and Crossref with the implementation of the Crossref Funder Registry (Crossref 2020) as a shared resource for funder identifiers that disambiguate the many names that are used for organizations that fund research projects. Typical funder metadata combines these identifiers with the funder name and the award number to create unique and permanent connections between funders and research objects.

During 2021 Dryad took on a second major re-curation effort, in this case focused on funder identifiers rather than RORs. The archive was searched for funder names that were normalized when possible and Crossref was searched for funder metadata provided for related articles. When this information was found, it was added to the Dryad dataset metadata (Figure 2A, C).

Figure 3 shows the results of this funder metadata re-curation project, comparing the number of award numbers, funder names, and funder identifiers in all Dryad metadata (~10,000 datasets) during 2020 and 2021 before and after the update. The focus on funder identifiers resulted in addition of identifiers for nearly all the datasets that had funder names and award numbers.

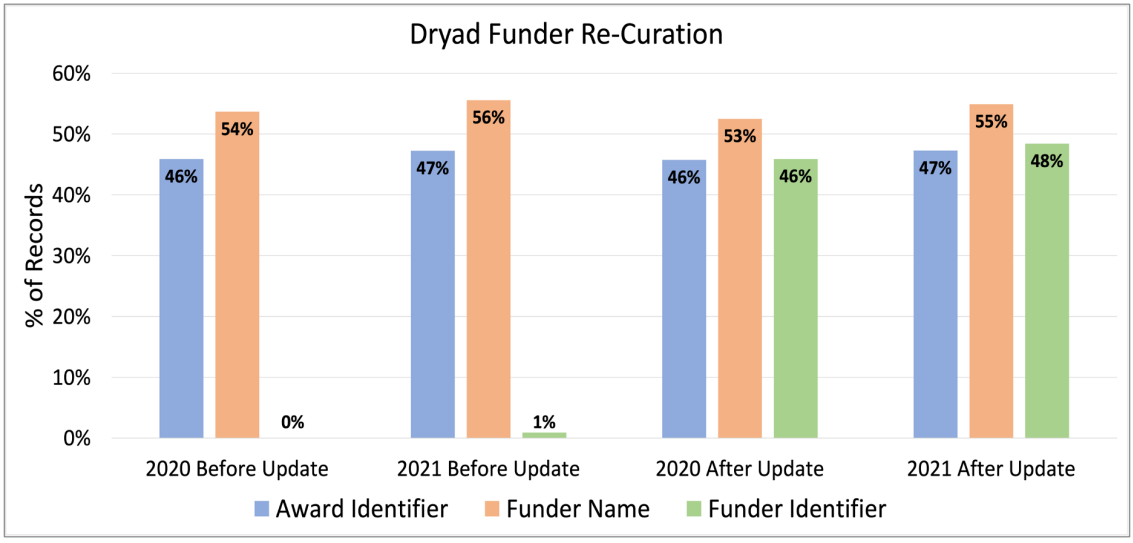


**Figure 3**: % of records with funder names, award numbers, and funder identifiers in all Dryad dataset metadata (~10,000 records).

### *Re-Curation*

Re-curation workflows were introduced by Hoyt et al. 2019 as part of the process of keeping metadata for biological research up to date. Addition of new metadata elements into repositories to enable new capabilities and connections is a critical part of these processes. The Dryad repository evolution described above included several phases of re-curation: the introduction of Crossref DOIs for connecting Dryad datasets to papers, the adoption of the DataCite metadata model and the migration of the metadata to DataCite, the addition of author affiliations, the addition of RORs for affiliated organizations, and the addition of identifiers for funders and awards. All these phases have resulted in more complete metadata and greatly improved connections between Dryad research objects, related papers, research organizations, and funders.

## Improving FAIRness of Metadata in Institutional Repositories

Institutional repositories serve academic researchers by providing detailed data management guidance and resource preservation. The guidance typically includes specific recommendations for metadata required to support many FAIR Principles. This guidance and interaction between researchers and data management experts results in resources with high-quality metadata addressing many of the FAIR requirements. Digital Object Identifiers (DOIs) for identifying resources are important elements included in this guidance.

Many institutional repositories use DataCite as a source for DOIs for datasets and other research resources. DataCite metadata focus on resource identification and citation and require only a small number of metadata elements. In most cases, only these minimum metadata are provided to DataCite by the Institutional Repository to minimize the effort required to satisfy the requirement to get a DOI for the institutional

repository resources (Figure 4A). In many cases, this is related to tools the repository uses to share metadata with DataCite and it can be difficult or expensive for the repository to evolve. These metadata can be retrieved from the institutional DataCite repositories by searching the DataCite API for repository-id = Institutional repository id.


**Figure 4**: Academic Pathways to DataCite. A is the minimum metadata pathway used by the institutional repository to get a DOI with minimum input. B is the pathway used directly by researchers through other repositories. In many cases path B results in more complete metadata in DataCite.

Academic researchers also submit datasets and other resources to DataCite through other repositories like Zenodo, Harvard DataVerse, Dryad, ICSPR, and many others (Figure 4B). Datasets submitted this way can sometimes be found and associated with institutions by searching creator affiliations or RORs in the DataCite metadata, i.e. searching for creator affiliation = '*University*of*X*'.

## *Measuring Metadata FAIRness*

The FAIR Principles (Wilkinson et al. 2016) provide high level guidance for improving findability, access, interoperability, and re-use of data. Evaluating compliance with these principles is typically done at the level of repository practices (see Devaraju and Huber 2021 for an overview).

Jones et al. 2016 describe a generalized approach to evaluating compliance with community standards that focuses more attention on completeness of specific metadata elements. This approach was extended to facilitate evaluation of DataCite metadata FAIRness (Habermann 2019B) and applied to over 100 DataCite member repositories managed by the German Technical Information Library (Burger et al. 2021 and Habermann 2021).

The FAIR evaluation determines completeness (% of records that include the element) of over fifty DataCite metadata elements in four categories (Findable Essential, Findable Supporting, AIR Essential, AIR Supporting). Results are shown in a collection of four rose diagrams, one for each category (Figure 5). The

diagrams show completeness (0 in the center, 100% on the edge) of ~fifty documentation concepts that map to appropriate metadata elements in various dialects.

The pattern seen in Figure 5 is common across many DataCite repositories and reflects the repository practices denoted in Figure 4A, i.e. only the required fields (in the Findable Essential and AIR Essential categories) are complete. The totals show the completeness for each category. These four completeness measures, along with the completeness over the total set of elements, are compared in Figure 6 and Figure 7 for the two metadata sets described above and illustrated in Figure 4.

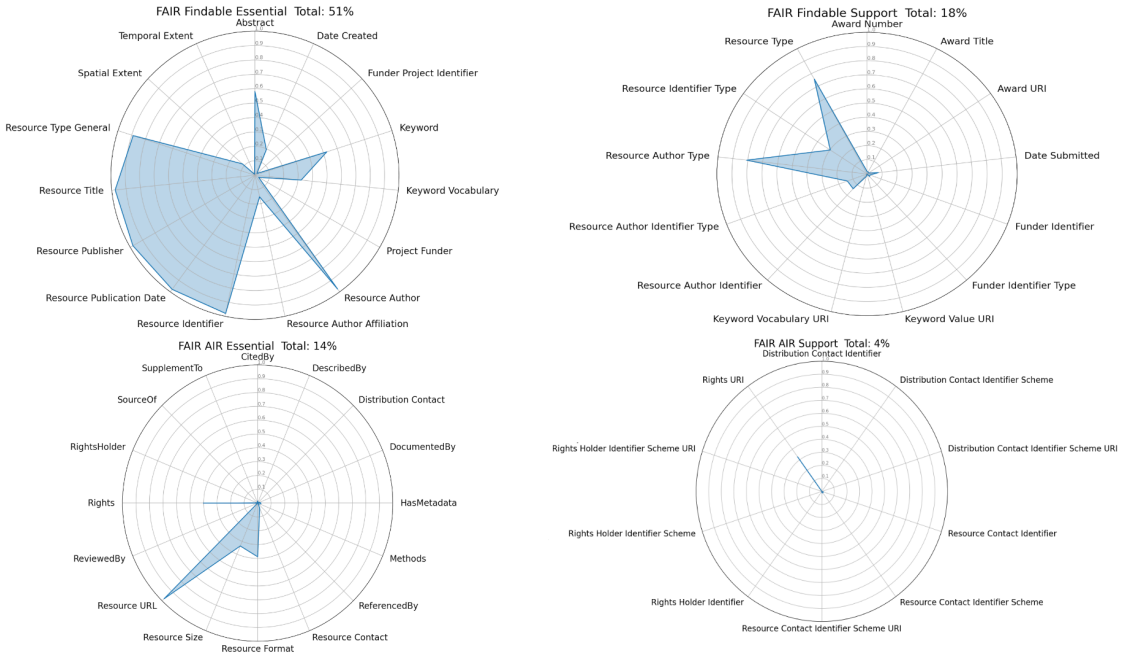


**Figure 5**: Schematic diagram showing completeness of a representative selection of 190 DataCite metadata records in four categories given above each rose diagram. The names of documentation concepts in each category are given around the edges of each diagram and completeness is shown from 0% in the center to 100% at the edge. Total completeness (%) is shown for each category is shown in Figures 6 and 7 for several sets of repositories.

## *Improving Global Metadata Completeness*

Figure 4 shows two pathways into the global research infrastructure for metadata from academic institutions: one through Institutional Repositories (A) and one through other repositories (B). The FAIR analysis approach described here provides a way to compare DataCite metadata for these two pathways. Differences in completeness between these two types of metadata are shown by category averages for six institutional repositories in Figure 6. Metadata from the Other Repositories (blue, solid) is more complete that metadata from the Institutional Repositories (red, dashed) and the differences are clear and consistent across all the institutions, suggesting a general characteristic.

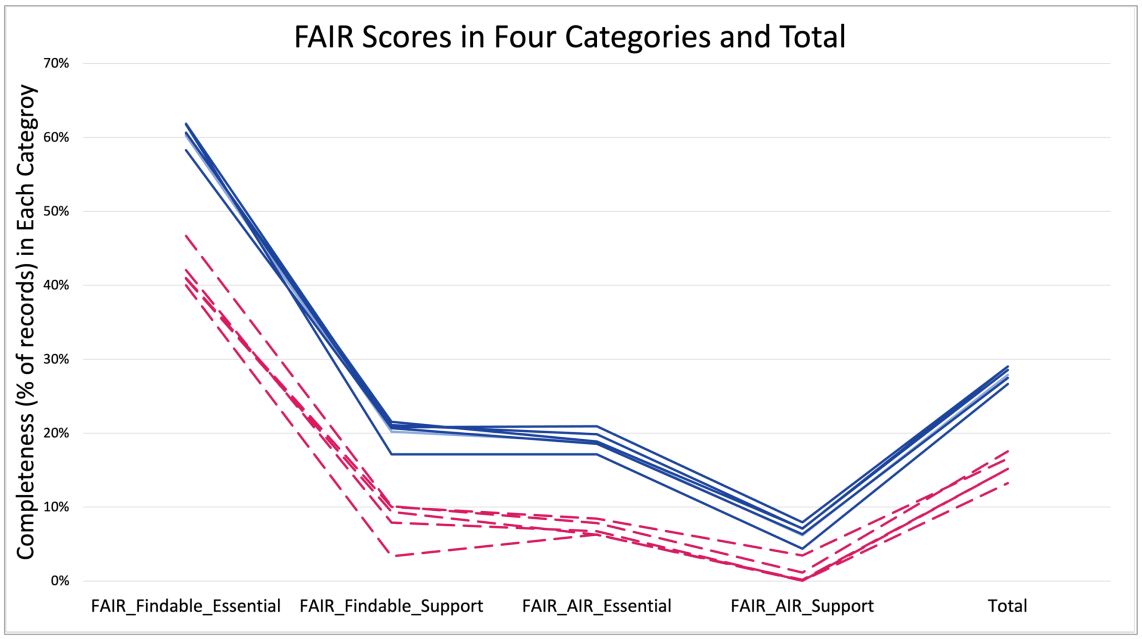


**Figure 6**: Completeness scores in four FAIR categories and total for DataCite metadata from six Institutional Repositories (red dashed, Figure 4A) and Other Repositories (blue solid, Figure 4B). The metadata sets were retrieved using the DataCite API searching for Institutional Repository IDs and for Institutional Affiliations. The average number of records per set was 520.

As described above, the Institutional Repositories partner with the researchers to create complete metadata. The observations in Figure 6 suggest that many metadata elements for which content is known do not make it from the Institutional Repository into DataCite. Examining metadata in the Institutional Repositories shows that this is, in fact, the case. Many of the metadata elements included in the FAIR evaluation exist in that metadata. For example, the Abstract element is available in nearly all the Institutional Repository metadata but not in DataCite. Abstract is not required in DataCite so it does not make it through the Institutional Repository metadata submission process.

This example is different from the Dryad case shown above in that the need for re-curation is related to technical repository processes instead of new identifiers. To address it, therefore, processes must be evolved. We developed custom tools to transfer missing metadata elements to DataCite and ran the evaluations on the re-curated metadata. The results in Figure 7 show that increases in completeness across the board can be accomplished using new processes to transfer existing metadata. See Habermann 2022 for another example.

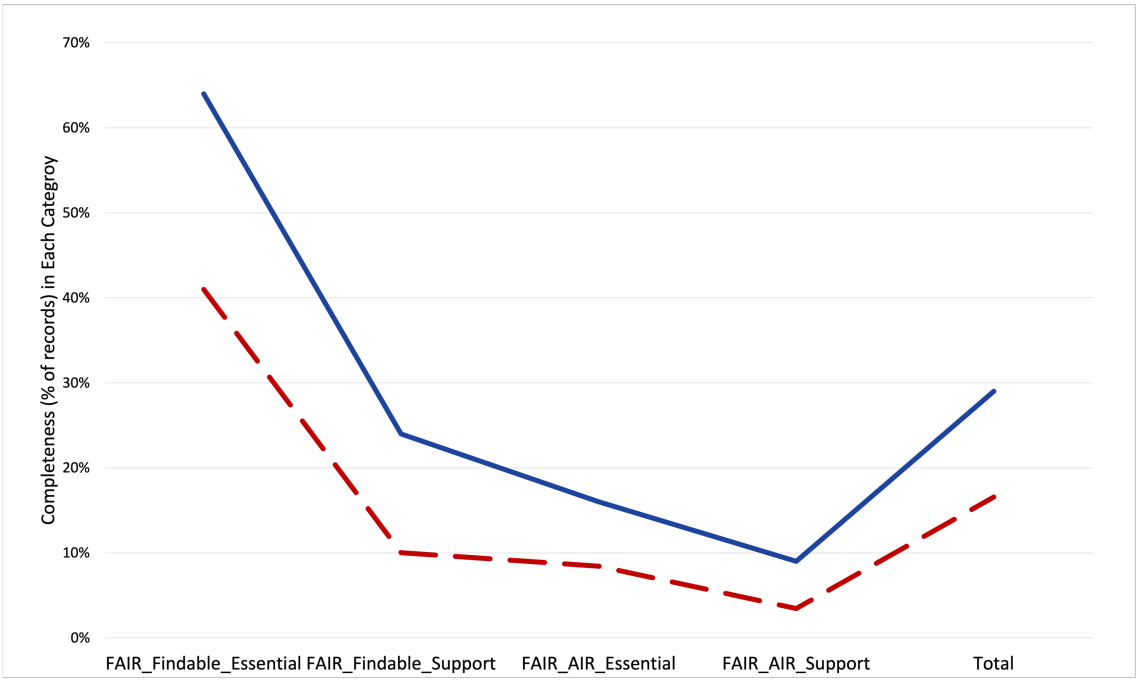


**Figure 7**: Representative Institutional DataCite metadata FAIRness before (red) and after (blue) transferring all possible metadata to DataCite.

## Conclusion

The global research infrastructure that has emerged over the last decade continues to provide an expanding set of connections between journal articles, datasets, researchers, research organizations and many other kinds of research objects. Participating in this network requires unique and persistent identifiers for these objects.

Research repositories around the world have existed for many years and metadata in these repositories created prior to the emergence of these identifiers needs to be augmented to facilitate connections. This metadata augmentation is an on-going curation process, termed *re-curation*. Two re-curation efforts are described here.

The first described a process of taking advantage of existing DOIs in Dryad metadata to retrieve affiliations and identifiers for organizations and funders from the global research infrastructure (Crossref and ROR). Once found, these identifiers were added to Dryad metadata and then submitting to DataCite, enriching the network of connections to these datasets.

The second example demonstrated augmenting the global infrastructure with existing metadata from institutional repositories that goes beyond the minimum metadata required by DataCite for identification and citation. This included abstracts, keywords, temporal extents, and funder information. These metadata

extend the possible use cases of the global infrastructure beyond the "get a DOI" use case and provide some support for the FAIR principles beyond findability.

Participating in the global research community requires on-going metadata evaluation and improvement efforts that augment the current practice of curating metadata and datasets only during the submission process. These two examples demonstrate that information required to connect repositories to the growing research infrastructure exists and can be harvested and utilized to add connections and re-use metadata to existing repositories. This work benefits the entire research community by integrating existing resources into on-going research with rich connections between people, institutions, funders, and results.

## Data Availability

The data used in this work were provided by Institutional Repositories and retrieved from DataCite using the public DataCite API during late 2021 and early 2022. These metadata are constantly being maintained and change over time so the now out-of-date metadata are not available. The software used in the analysis was customized to accommodate translation of metadata models used in each repository to the DataCite model and uploading that metadata to DataCite. It is not easily generalizable.

## Acknowledgements

This work was funded by the U.S. National Science Foundation awards 2134956 and 2135874.

Thanks to partners in the EAGER: Completing the Lifecycle: Developing Evidence Based Models of Research Data Sharing Project for access to and help with metadata for their institutional and DataCite repositories and insights into their repository practice and to the Dryad Team and the California Digital Library digital curation group (UC3) for many discussions along the way.

The content of this article is based on the presentation entitled "Re-curation Processes-Connecting Repositories to the Global Research Community" originally presented at RDAP Summit 2023, available from Open Science Framework: https://osf.io/cgqvk.

## Competing Interests

The author declares that they have no competing interests.

## References

Burger, Marleen, Anette Cordts, and Ted Habermann. 2021. "Wie FAIR sind unsere Metadaten? Eine Analyse der Metadaten in den Repositorien des TIB-DOI-Services." *Bausteine Forschungsdatenmanagement* (3): 1–13. https://doi.org/10.17192/bfdm.2021.3.8351.

Crossref. 2020. Funder Registry. Retrieved March 22, 2023. https://www.crossref.org/services/funder-registry.

Devaraju, Anusuriya and Robert Huber. 2021. "An automated solution for measuring the progress toward FAIR research data" *Patterns* 2(11): 100370. https://doi.org/10.1016/j.patter.2021.100370.

Dryad. 2018. Dryad partnering with CDL to accelerate data publishing. *Dryad news* (blog). https://blog.datadryad.org/2018/05/30/dryad-partnering-with-cdl-to-accelerate-data-publishing.

Dryad. 2018. New Dryad is Here. *Dryad news* (blog). https://blog.datadryad.org/2019/09/24/new-dryad-is-here.

Fenner, Martin and Amir Aryani. 2019. Introducing the PID Graph, Introducing the PID Graph. *FREYA Blog* (blog). https://www.project-freya.eu/en/blogs/blogs/the-pid-graph.

Gould, Maria and Daniella Lowenberg. 2019. ROR-ing Together: Implementing Organization IDs in Dryad. *ROR Blog* (blog). https://ror.org/blog/2019-07-10-ror-ing-together-with-dryad.

Habermann, Ted. 2019A. Dryad Data Packages and Files. *Metadata Game Changers Blog* (blog). https://metadatagamechangers.com/blog/2019/2/11/dryad-data-packages-and-files-1?rq=Package.

Habermann, Ted. 2019B. MetaDIG recommendations for FAIR DataCite metadata. *DataCite Blog* (blog). https://doi.org/10.5438/2CHG-B074.

Habermann, Ted. 2021. A PID Feast for Research – PIDapalooza 2021. *Metadata Game Changers Blog* (blog). https://metadatagamechangers.com/blog/2021/2/2/a-pid-feast-for-research-pidapalooza-2021.

Habermann, Ted. 2022. Metadata Life Cycle: Mountain or Superhighway? *Metadata Game Changers Blog* (blog). https://metadatagamechangers.com/blog/2022/3/7/ivfrlw6naf7am3bvord8pldtuyqn4r.

Hoyt, Charles Tapley, Daniel Domingo-Fernández, Rana Aldisi, Lingling Xu, Kristian Kolpeja, Sandra Spalek, Esther Wollert, John Bachman, Benjamin M. Gyori, Patrick Greene, and Martin Hofmann-Apitius. 2019. "Re-curation and rational enrichment of knowledge graphs in Biological Expression Language." *Database* 2019(2019): baz068. https://doi.org/10.1093/database/baz068.

Johnston, Lisa R., Jake Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, Claire Stewart, Mara Blake, Joel Herndon, Timothy M. McGeary, and Elizabeth Hull. 2018. "Data Curation Network: A Cross-Institutional Staffing Model for Curating Research Data. In International Journal of Digital Curation." *International Journal of Digital Curation* 13(1): 125–140. https://doi.org/10.2218/ijdc.v13i1.616.

Jones, Matthew, Peter Slaughter, Ben Leinfelder, Bryce Mecum, Ted Habermann, Lindsay Powers, Sean Gordon. 2016. "MetaDIG?: Engaging Scientists in the Improvement of Metadata and Data." figshare. Journal contribution. https://doi.org/10.6084/m9.figshare.4055808.v1.

Lowenberg, Daniella and Ted Habermann. 2019. "Metadata and Dialect Evolution - Affiliations in the Dryad Data Repository." ESIP. Presentation. https://doi.org/10.6084/m9.figshare.9252824.v1.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* 3(1): 1–9. https://doi.org/10.1038/sdata.2016.18.