Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00252

Research paper

Sustainable Connectivity in a Community Repository

Ted Habermann

Metadata Game Changers, 4524 14th St. Unit 7H Boulder, Boulder, Colorado 80304, USA)

Corresponding author: Ted Habermann (E-mail: ted@metadatagamechangers.com; ORCID: 0000-0003-3585-6733)

Citation: Habermann, T. Sustainable connectivity in a community repository. Data Intelligence. 2024. DOI: TBD

Keywords: Metadata Curation, Re-Curation, Connectivity, FAIR Data, Persistent Identifiers

Abstract

Persistent identifiers for research objects, researchers, organizations, and funders are the key to creating unambiguous and persistent connections across the global research infrastructure (GRI).

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00252

Many repositories are implementing mechanisms to collect and integrate these identifiers into their submission and record curation processes. This bodes well for a well-connected future, but metadata for existing resources submitted in the past are missing these identifiers, thus missing the connections required for inclusion in the connected infrastructure. Re-curation of these metadata is required to make these connections. This paper introduces the global research infrastructure and demonstrates how repositories, and their user communities, can contribute to and benefit from connections to the global research infrastructure.

The Dryad Data Repository has existed since 2008 and has successfully re-curated the repository metadata several times, adding identifiers for research organizations, funders, and researchers. Understanding and quantifying these successes depends on measuring repository and identifier connectivity. Metrics are described and applied to the entire repository here.

Identifiers (Digital Object Identifiers, DOIs) for papers connected to datasets in Dryad have long been a critical part of the Dryad metadata creation and curation processes. Since 2019, the portion of datasets with connected papers has decreased from 100% to less than 40%. This decrease has significant ramifications for the re-curation efforts described above as connected papers have been an important source of metadata. In addition, missing connections to papers make understanding and re-using datasets more difficult.

Connections between datasets and papers can be difficult to make because of time lags between submission and publication, lack of clear mechanisms for citing datasets and other research objects from papers, changing focus of researchers, and other obstacles. The Dryad community of members, i.e. users, research institutions, publishers, and funders have vested interests in identifying these connections and critical roles in the curation and re-curation efforts.

Their engagement will be critical in building on the successes Dryad has already achieved and ensuring sustainable connectivity in the future.

1. Introduction

Dryad [1] is a community of academic and research institutions, research funders, scholarly societies and publishers that are committed to leading in best practices for open data sharing and reuse and to the open availability and routine re-use of all research data. Connections across the Dryad community and between Dryad and the broader global research community are critical for supporting these goals. Managing connections across these communities requires consistent monitoring and on-going activity. The repository team and all community members have roles in creating and sustaining those connections through the entire data life cycle.

Persistent identifiers of many kinds are included in research object metadata as related identifiers to realize unambiguous and persistent connections. These include DOI's for articles, datasets, software and other research objects [2], Open Researcher and Contributor IDs (ORCIDs) for researchers, Research Organization Registry identifiers (RORs) for organizations, Funder Ids (either Crossref Funder Ids or RORs) for funders, and (funder) award numbers or DOIs for funded projects. In addition to making connections, these identifiers are critical for ensuring that appropriate credit for a wide variety of contributions is given to community members. These identifiers also serve as persistent "primary keys" in repository systems. Together with metrics like those described below, these primary keys can be used for tracking evolution of repositories through time. Creating data-driven, quantitative baselines and measuring through time are key to on-going tracking processes.

Together these identifiers and the research objects they identify are referred to here as the *global research infrastructure*. This infrastructure is global [3] and is made up of organizations that provide identifiers with repositories of related metadata and on-going identification, connection, and discovery services on top of those repositories. While many organizations from all over the world makeup this infrastructure, here I focus on Crossref, DataCite, ORCID, and ROR, which together form a coherent network with broadly available and well-documented services.

1.2 Dryad History

Understanding repository context and how it evolves over time provides important background for long-term tracking. The context of Dryad has changed significantly over the last several years. It was conceived during 2007 and went live during 2008 [1]. The first data submission instructions read: "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" [4]. This emphasis on connections between datasets and papers has persisted since the beginning of Dryad and is a critical link in re-curation efforts described here.

Several significant changes have occurred during Dryad's history, most important the development of a partnership with California Digital Library during 2018 [5] and the subsequent launch of the "New Dryad" during 2019 [6]. Associated changes included migration to a new metadata model based on the DataCite Schema [7], strengthening the links to the global research infrastructure (GRI) and the pioneering introduction of identifiers for organizations (RORs, [8])

and people (ORCIDs). Finally, Dryad began migration to a membership-based business model with direct financial support from publishers and research institutions in the community.

1.3 Dryad Connections

The original Dryad metadata model [9] focused on connecting multiple data files into packages and administering the preservation of those data packages. It relied on connected articles as critical contributors to the documentation required to discover, understand, and re-use datasets. Even typical discovery metadata such as author names and affiliations were not included in the Dryad metadata as they were available in the related papers.

During 2019 Dryad adopted the DataCite Metadata schema which brought important changes to the metadata model. Part of this evolution included addition of DOIs for the articles related to Dryad datasets, which enabled a richer set of connections to other types of resources (articles, software, preprints, etc.). This evolution is illustrated by the addition of Crossref (C) and DataCite (D) to the Dryad infrastructure shown in Figure *1*.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00252

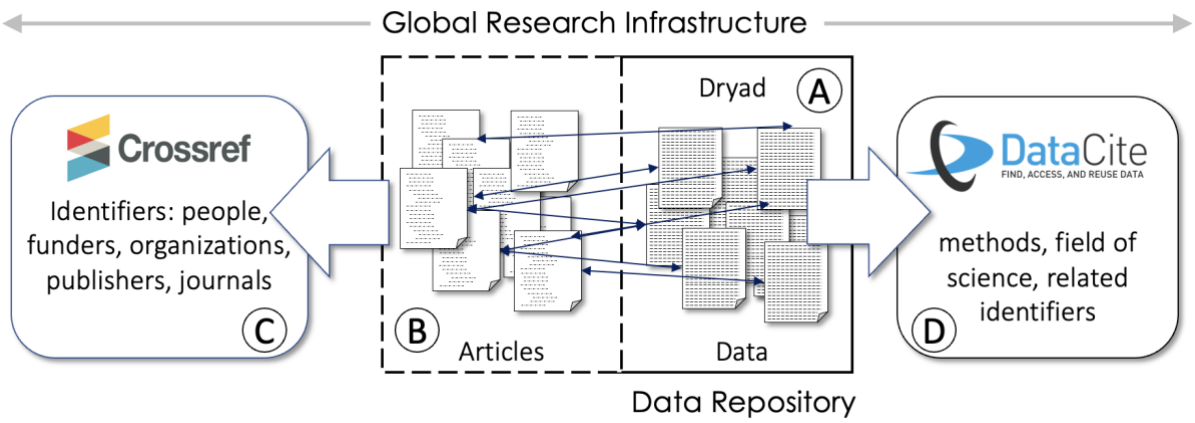

Figure 1. Evolution of Dryad from an isolated data repository (A) to a connected virtual repository with data and related articles (B) and then to a connected element of the global research infrastructure with article metadata in Crossref (and other repositories) (C) and dataset metadata in DataCite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

The adoption of the DataCite metadata model had an important effect on the relationship between Dryad and the GRI. It means that all Dryad metadata are shared with the GRI through DataCite, not just the six mandatory DataCite fields required to get a DOI.

1.4 The Dryad Community

Figure *2* shows the number of unique datasets, organizations, and authors for Dryad datasets associated with journals. The size of the community has increased over time with an average of over 5500 unique datasets, 4000 unique organizations, and over 25,000 unique authors per year since the introduction of the new Dryad during 2019. These numbers do not include Dryad datasets that are not associated with journals which add ~4% to the total.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00252

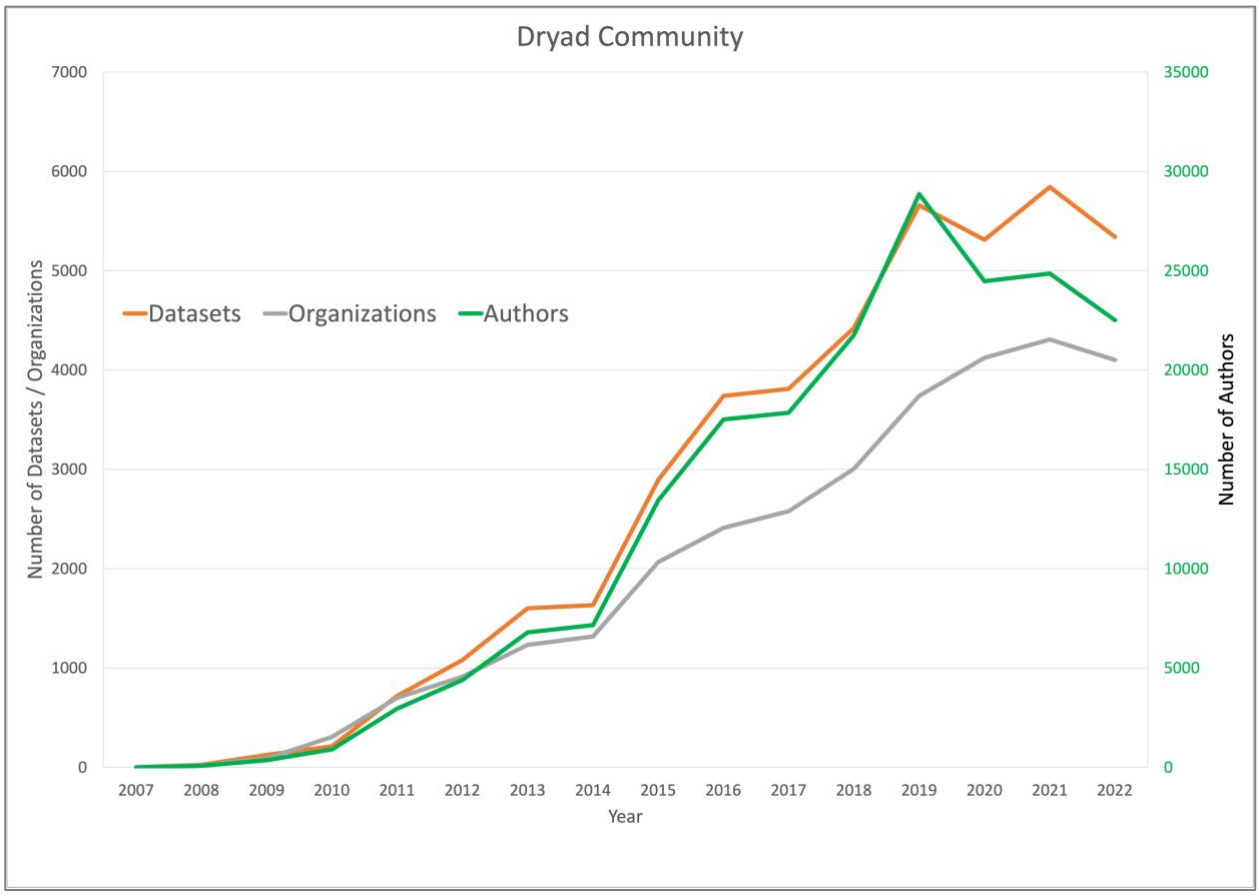

Figure 2. Number of datasets (orange), authors (green), and organizations (grey) associated with journals in the Dryad repository between 2007 and 2022.

2. Repository Guidelines and Identifiers

Many organizations and initiatives have developed and espoused sets of guidelines and practices for repositories of many kinds. These guidelines are generally high-level, can cover many aspects of repository practice, and can be addressed in many ways. In this work we are focused on identifiers, so identifier guidelines and identifier practices are most relevant.

Perhaps the most referenced set of data management principles is the FAIR Data Principles [10] which provide high level guidance for findability, access, interoperability, and re-use of

data. These principles include identifiers for data and metadata and recommend including identifiers for datasets in the metadata that describes them. They do not include guidelines for other kinds of identifiers.

The Generalist Repository Ecosystem Initiative [11], supported by the National Institutes of Health (NIH), was created to support data sharing and reuse by NIH-funded researchers. Dryad is one of six repositories supported by this initiative. Best practice recommendations proposed [12] for sharing data in generalist repositories included leveraging PIDs (RORs, ORCIDs, DataCite DOIs) across the repositories to avoid broken links and create interoperability between infrastructures that include these identifiers. Using the DataCite metadata schema which supports these identifiers was also recommended, along with providing annual updates on data management and sharing activities.

The Confederation of Open Access Repositories (COAR) is an international association with 156 members and partners from 50 countries, representing libraries, universities, research institutions, government funders and others. The COAR Community Framework for Good Practices in Repositories [13] describes essential and desired repository characteristics, including a recommendation to use DOIs that point to landing pages, but nothing about identifiers other than DOIs, or about measurement/reporting.

The U.S. Federal government released several important sets of guidelines during 2022. First, the Subcommittee on Open Science of the National Science and Technology Council released high-level guidance for repositories for federally funded research [14]. Second, the U.S. Office of Science, Technology and Policy (OSTP) released a memorandum during August 2022 [15] recommending that repositories include identifiers for authors, organizations, funders, and

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00252

research objects in publicly available metadata. This memo thus provided explicit guidance related to the interconnected global research infrastructure (GRI) envisioned in this work, at least in the context of distributed repositories.

There are several important practices that are not discussed in any of these recommendations. First, the concept of sharing complete repository metadata with the global research infrastructure. Dryad demonstrates benefits of this recommendation by using the DataCite metadata schema, which includes all relevant identifiers and, sharing all their metadata in DataCite. In addition, Dryad adds improved metadata to DataCite on a regular basis, facilitating an improved and more useful GRI. Second, the concept of measuring compliance with any set of recommendations is also missing. The importance of measurement is well known in the federal [16] and private [17] sectors.

This paper presents some ideas and examples of measurements of connectivity with the goal of helping communities understand, improve, and sustain repository connectivity.

3. Connectivity

Whether research objects get discovered depends on their *connectivity, i.e., the state or extent of being connected or interconnected*. Can connectivity in a repository be measured? A connectivity metric has been defined [18] as the number of existing identifiers divided by the total number of possible identifiers, expressed as a %. This metric can be measured and applied across any interesting collection of research objects. For example, a typical dataset in Dryad has several funders and authors, each of which can have an identifier or an affiliation. Each dataset therefore has connectivity, i.e. the number of identifiers / the number of possible identifiers. The

connectivity can also be calculated for the entire repository or for any subset of the repository, e.g. for all datasets associated with an author, a journal, or a research organization. This finer granularity is important, as these are the organizational units that can take action to improve connectivity for resources they create and manage.

Connectivity can also be calculated for different types of identifiers. For example, dataset connectivity can be calculated for funder identifiers, for ORCIDs, or for RORs, and any kind of connectivity can be calculated over time to track changes at any granularity.

4. Curation and Re-Curation

The definition of curation varies significantly across the spectrum of repositories in the U.S. and around the world. The Data Curation Network [19] 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 the metadata going forward, but existing records, i.e., those for datasets curated in the past, 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.

Repository 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, Scholix, etc.) as well as tools for making those connections and retrieving relevant metadata. 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), communities using identifiers in new ways, and identifiers migrating between types (e.g. IGSNs becoming DOIs). In many cases, these differences mean that new tools are required for facilitating this work.

In addition, re-curation can account for important connections that develop over time, i.e., the article publication process is slower than dataset curation and datasets are contributed before articles are reviewed, revised, and published. Re-curation is needed to find these connections when they occur and add them to the dataset metadata. This is an area where community members, i.e. researchers, funders, and organizations play critical roles.

5. Dryad Re-Curation

As the Dryad community and repository has grown, identifiers have emerged, and metadata dialects have evolved. Dryad has taken an active role in evolving their metadata model and adding new content. As these additions have taken place after the resources are in the repository, they are re-curation projects. Dryad re-curation projects for organizations, individuals, funders, and research objects are described in this section.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00252

5.1 Affiliations and RORs

During 2019 a new community-driven identifier for organizations [20] was being developed and Dryad decided to add this new identifier for nearly 100,000 organizations in over 20,000 dataset metadata records [8].

Given the pre-2019 Dryad metadata model, re-curating the metadata to add identifiers for organizations required two steps: 1) finding author affiliations and 2) using those affiliations to find RORs. Fortunately, the Dryad metadata included connections to Crossref, a source for author affiliations in a standard form that could be retrieved using DOIs included in Dryad metadata (A in Figure 3). This resulted in a long list of "noisy" affiliations with considerable ambiguity and complexity.

This was early in the days of ROR, so approaches to searching these affiliations to convert them to RORs (B in Figure 3) were developed and implemented. This search resulted in nearly 90% of the Dryad datasets having RORs for at least one organization. The New Dryad was using DataCite to mint DOIs and using the DataCite metadata model which includes authors, affiliations, and affiliation identifiers, so the new metadata content could be added to DataCite to become available to the global research infrastructure through the standard DataCite API (C in Figure 3).

This process illustrates using automated tools to augment human curators in re-curation workflows. Affiliation strings were retrieved automatically from Crossref for thousands of DOIs and authors, and algorithms [21, 22] were used to search those strings for organization names and search the ROR registry for the actual RORs. The algorithms work well and save considerable time, but noise in the affiliation strings and other realities such as authors with

multiple affiliations or acronyms [23], requires that the results be manually curated to identify problems and validate final selections.

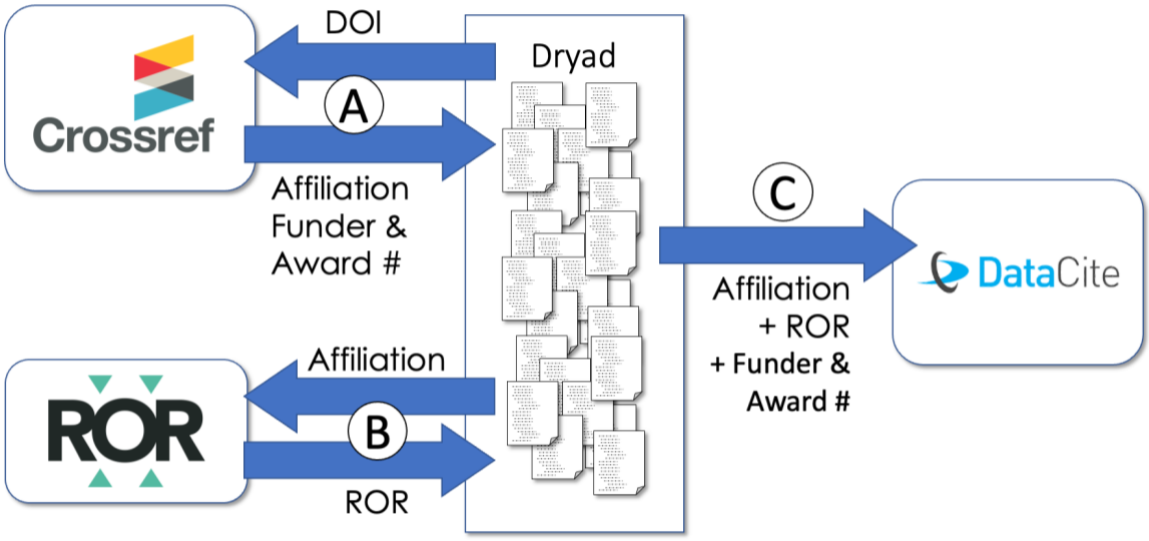

Figure 3. Two Dryad re-curations projects to increase completeness of connected papers and funder information using Crossref as a data source.

Figure 4 shows the % of authors in Dryad journal-related[1] datasets that have affiliations as a function of time (blue) which has been above 80% since 2010 except for a small dip during the transition to the New Dryad during 2019. Since then, affiliation information has been entered by authors during the submission process (indicated by "Curation" in Figure 4).

[1] Dryad "journal-related" datasets are datasets 1) already related to specific articles in a journal or 2) where authors identify the journal they expect the related paper to be published in when they submit the dataset. These data sets can be retrieved by searching Dryad for the International Standard Serial Number (ISSN) associated with the journal. See section 5.5 for discussion of datasets submitted without journals.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00252

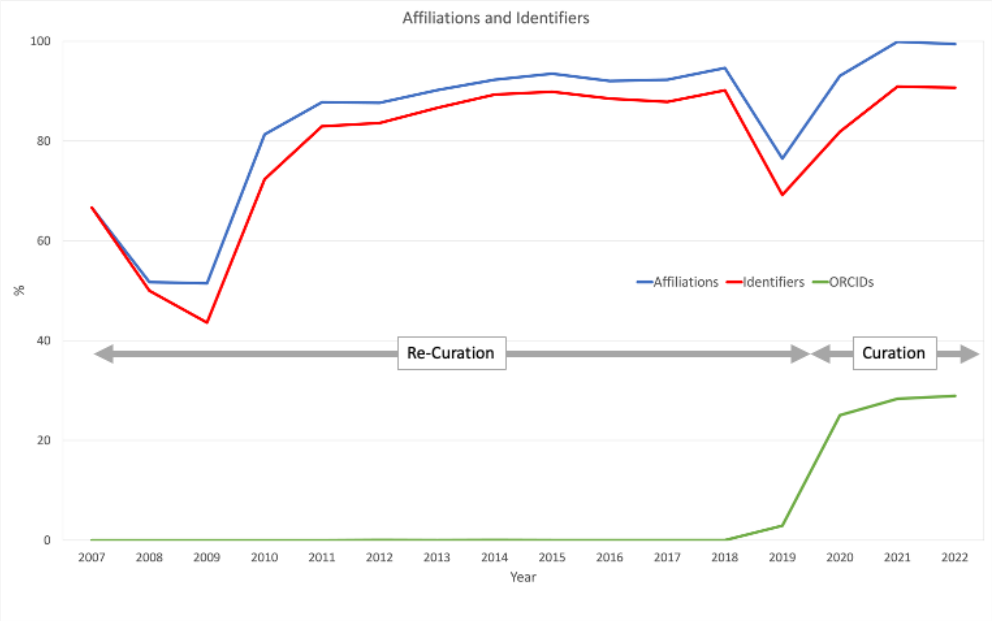

Figure 4. The % of journal related datasets in Dryad that include author affiliations (blue), affiliation identifiers (red), and author identifiers (green) over time. Periods of re-curation and curation are shown.

The red line in Figure 4 shows the % of authors with RORs, which is generally within 5% of the % with affiliations. This gap reflects affiliations for organizations that do not yet have RORs or RORs missed during the curation and re-curation. Even with these gaps, comparing the results of the curation and re-curation periods in Figure 4 shows that the success of the re-curation process is very close to the ongoing curation process. The average % for the re-curation between 2010 and 2018 is 86% compared with 89% during the curation period between 2020 and 2022.

5.2 People

Figure 4 shows the history of occurrences of identifiers for people (ORCIDs) in Dryad metadata between 2007 and 2022 (green). The % in this case is the percentage of authors that

have identifiers rather than the % of DOIs. These identifiers began being included during 2019, when they started being used for users logging into the New Dryad, and that completeness has grown to between 25 and 30% of the authors having ORCIDs.

The increased ORCID occurrence since 2019 reflects the Dryad practice of using ORCIDs as logins. This ensures that each dataset submitted to Dryad includes an ORCID for at least the author that submits the dataset to ORCID. The % between 25 and 30% reflects the fact that many times there is only one ORCID associated with a dataset even if there is more than one author.

Three approaches can be used to increase the completeness of ORCIDs in the repository. The first is the same as that used in the ROR case – searching Crossref or other sources for author ORCIDs. This approach is limited by incompleteness of ORCIDs in Crossref and other journal article metadata which is related to the common practice of requiring ORCIDs only for corresponding authors. This practice is becoming less common with growing acceptance and understanding of the benefits of ORCIDs, but ORCIDs remain much less common in journal metadata than affiliations.

The second approach to increasing ORCID completeness, termed 'spreading' [18], works in situations where authors make multiple contributions to a repository, but only include their ORCID for some of them. This situation is demonstrated in Table 1 which shows twelve Dryad datasets for Dr. Todd Vision, a co-founder and long-time user of Dryad. These datasets illustrate the need for and some of the problems with spreading.

Publication Date	DOI	Name	Identifier
2008-06-18	doi:10.5061/dryad.162	Todd J. Vision	
2010-10-18	doi:10.5061/dryad.7881	Todd J. Vision	

2011-04-28	doi:10.5061/dryad.j1fd7	Todd J. Vision	
2013-10-01	doi:10.5061/dryad.781pv	Todd J. Vision	
2014-12-12	doi:10.5061/dryad.41dq8	Todd J. Vision	
2015-12-15	doi:10.5061/dryad.51vs3	Todd J. Vision	
2016-07-15	doi:10.5061/dryad.239sm	Todd J. Vision	
2016-10-31	doi:10.5061/dryad.8q931	Todd J. Vision	
2019-10-11	doi:10.5061/dryad.0373j7r	Todd Vision	
2020-04-08	doi:10.5061/dryad.3xsj3txbz	Todd Vision	0000-0002-6133-2581
2022	doi:10.5061/dryad.59zw3r27c	Todd Vision	
2022	doi:10.5061/dryad.vdncjsxwt	Todd Vision	

Table 1. Dryad datasets for Dr. Todd Vision

First, these twelve datasets have two different versions of the author's name: Todd J. Vision and Todd Vision. Small differences like this are easy to identify manually, but, with over 166,000 unique author names in the Dryad repository, they introduce disambiguation complexities. In this case, the ORCID record (https://orcid.org/0000-0002-6133-2581) confirms the middle initial J., but similar checks for all cases inevitably introduce manual work and related challenges.

Once a decision is made that all authors are the same person, the ORCIDs can be focused on. Only one of the twelve datasets include Dr. Vision's ORCID, so spreading in this case can gain ORCIDs for eleven new datasets. This is a very common situation in the Dryad repository. **Error! Reference source not found.** shows nine community members with 50 or more datasets in Dryad. Together these nine contributors with known ORCIDs add up to over 450 missing ORCIDs in the repository.

Table 2. Common contributors to Dryad with number of datasets and number of ORCIDs. The difference is an opportunity for spreading ORCIDs to records that are currently missing them.

Name	Dataset Count	ORCIDs
Louis Bernatchez	91	2
Richard Shine	63	18
Bart Kempenaers	58	8
Leigh W. Simmons	54	3
Ole Seehausen	52	6
Juha Merilä	52	1
Yang Liu	51	13
Pierre Taberlet	50	1
Axel Meyer	50	2

A second example that includes searching and spreading is provided by one of the recent DOIs in Table 1 (doi:10.5061/dryad.vdncjsxwt). In Dryad this dataset includes the ORCID for one of seven authors (Diego Porto, without * in Table 3) and affiliations for all authors. The dataset does not include a related article in Dryad, but searching for the name of the dataset using Google finds the related article in the journal Systematic Biology with the DOI: https://doi.org/10.1093/sysbio/syac022 [24]. Retrieving metadata for the article DOI from Crossref yields two more ORCIDs indicated by * in Table 3 and spreading ORCIDs from other Dryad datasets finds two more ORCIDs indicated by ** in Table 3. Combining these two

techniques (searching and spreading) increases completeness of ORCIDs for this dataset from 14% to 86%.

Name	ORCID	Affiliation
Diego Porto	0000-0002-1657-9606	Virginia Tech
Wasila Dahdul	0000-0003-3162-7490**	University of California, Irvine
Hilmar Lapp	0000-0001-9107-0714*	Duke University
James Balhoff	0000-0002-8688-6599*	Renaissance Computing Institute
Todd Vision	0000-0002-6133-2581**	University of North Carolina at Chapel Hill
Paula Mabee	0000-0002-8455-3213***	National Ecological Observatory Network
Josef Uyeda	0000-0003-4624-9680**	Virginia Tech

*Table 3. Authors, Identifiers, and Affiliations for https://doi.org/10.1093/sysbio/syac022. * show ORCIDs found by searching Crossref for this DOI, ** show ORCIDs found by spreading from other Dryad datasets, *** orcid.com lookup.*

Finally, names can be searched for ORCIDs directly on the orcid.org website. In cases like the one remaining name here, Paula Mabee, only one occurrence of the name is found and Dr. Mabee has chosen to make her ORCID profile public, so we can add the last ORCID for this dataset manually.

This example demonstrates the sometimes-circuitous path to re-curating ORCIDs in Dryad and other repositories. It is more difficult than re-curating affiliations because of the relative paucity of ORCIDs in the literature, identical or similar names for multiple people, ORCID profiles that are not open to the public, and inconsistency in the names that individuals use in dataset and journal article submission processes. Considerable work has been done in name disambiguation [25, 26] that can help further improve accuracy of these approaches.

Community members can be important contributors to increasing the completeness of ORCIDs in repositories of journal articles and datasets but individual vigilance and monitoring is

required for existing resources. Using ORCIDs in the login process can facilitate on-going collection of ORCIDs for community members.

5.3 Funder Identifiers

Organizations that provide funding for scientific research face the same identification problems described above for research organizations and authors and similar re-curation approaches can be used to add funder metadata into repositories. In this case the most common identifiers are Crossref Funder Identifiers [27] although use of RORs for funders is increasing [28].

During late 2021 Dryad undertook a multi-faceted re-curation project aimed at improving completeness of funder identifiers. It included normalization of funder names in the repository and searches for funder identifiers in Crossref (A in Figure 3).

The results of this effort are shown in Figure 5. The two histograms on the left show the % of funder names (orange), award numbers (blue), and funder identifiers (green) in all Dryad metadata during 2020 and 2021 before the re-curation. Note that funder identifiers were essentially absent from the repository prior to the re-curation. The histograms on the right show the same data after the re-curation project. The green bars show that funder identifiers were found for ~47% of the Dryad datasets and for ~88% of the funder names.

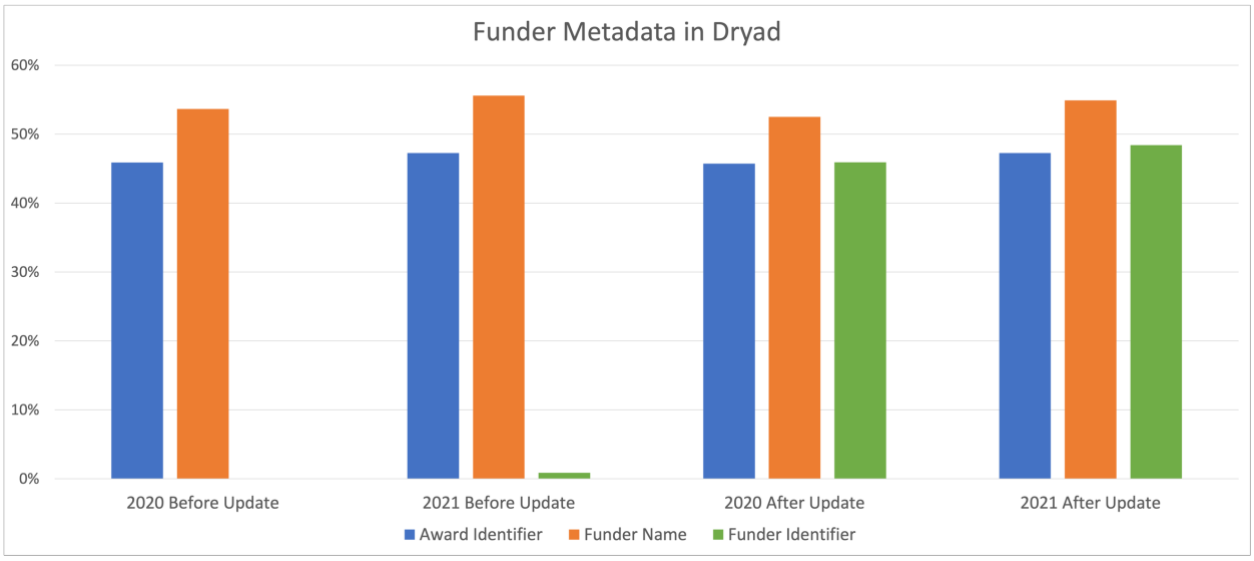

Figure 5. Results of a pilot project to increase the completeness of funder identifiers in Dryad. The % of records with award identifiers (blue), funder names (orange), and funder identifiers (green) during 2020 and 2021 are shown before and after the re-curation project.

Figure 6 shows the time history of the % of authors with funder metadata between 2008 and 2022. The increase in these numbers after 2019 reflects increased attention to identifying funders and awards during this time as well as the focused effort described above.

The shape of the curves in Figure 6 are like the ORCID curve in Figure 4 (green) and we showed above how spreading could be used to increase ORCID completeness earlier in the history of the repository. Spreading can also be used with funder identifiers but only after funder name disambiguation and grouping is done on data prior to 2019.

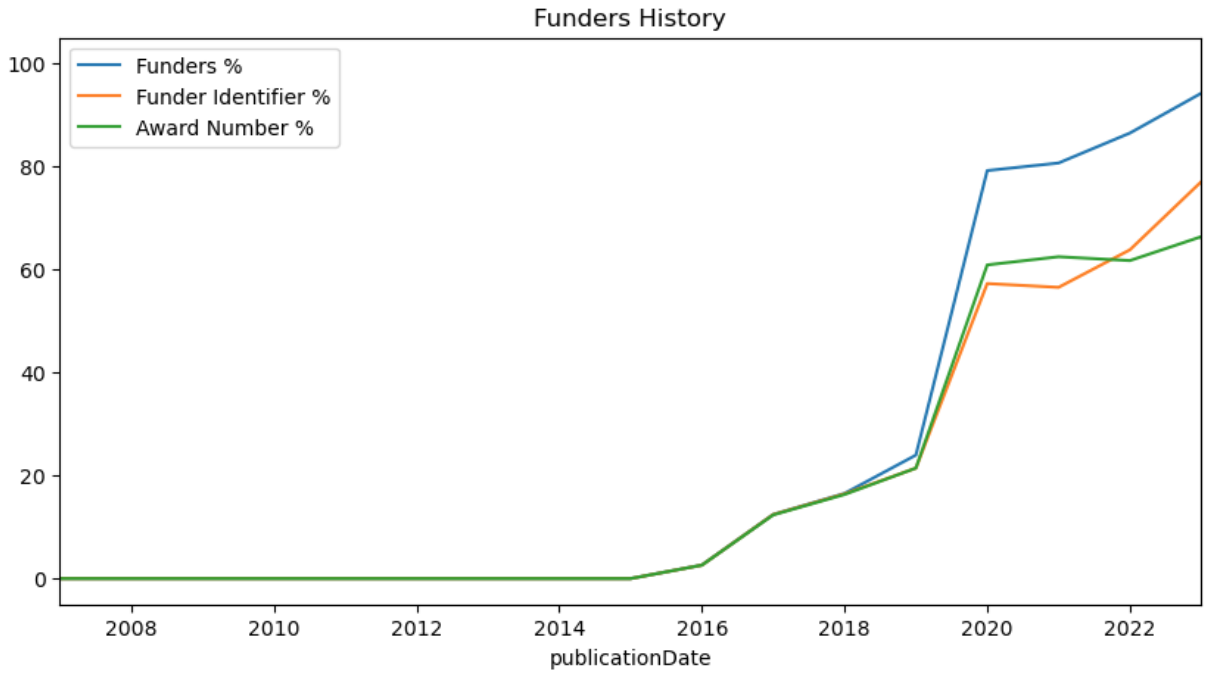

Figure 6. The % of authors with funder names (blue), funder identifiers (orange), and award numbers (green) in Dryad journal related records.

The identification of funder identifiers from name strings brings many of the same challenges as identification of organization names in affiliation strings. In particular, the use of acronyms in funder names can make reliable recognition of identifiers difficult or impossible [23]. As an example, a set of over 45,000 funder names and identifiers from Dryad was checked for consistency. Table 4 shows the identifiers associated with the funder name "NSF", typically an acronym used for the U.S. National Science Foundation. The last three, which occur 60/89 times are apparently incorrect interpretations of the acronym and emphasize the need for a combination of automated and manual tools in all re-curation processes.

Funder Identifier	**Funder Name**	**Count**
http://dx.doi.org/10.13039/100000001	National Science Foundation	28

http://dx.doi.org/10.13039/100000155	Division of Environmental Biology	1
http://dx.doi.org/10.13039/100016620	Nick Simons Foundation	31
http://dx.doi.org/10.13039/501100008982	National Science Foundation of Sri Lanka	21
http://dx.doi.org/10.13039/501100020414	Neurosciences Foundation	8

Table 4. Funder identifiers associated with the acronym NSF in Dryad.

Increasing the accuracy and completeness of funder metadata in repositories also depends critically on community members. Many repository metadata schemas, including the DataCite schema used by Dryad, now include specific elements for funder metadata. Using these elements, in addition to providing funder acknowledgements in free text, can ensure funders are identified and acknowledged correctly and that connections between researchers, funders, and specific awards can be made automatically and unambiguously.

5.3 Connecting Datasets to Papers

The examples given above, and the workflow shown in Figure 3, emphasize the importance of the global research infrastructure as a source for identifiers that can be re-curated into the Dryad repository to improve identifier completeness and dataset connectivity. This is particularly true prior to 2019, before the Dryad submission process focused more attention on collecting identifiers for RORs during initial curation and using ORCIDs for logins.

Connecting datasets and papers has been at the core of Dryad since its inception during 2008 [4]. Connections between datasets and papers in Dryad are made using related identifiers [2] with the “primary_article” relation type. Figure 7 shows the % of Dryad journal-related datasets that have these connections. The steep drop in the % of connections that occurs after 2019 coincides with the number of datasets submitted to Dryad increasing above 5000 / year (Figure

2). This decrease reflects the difficulty of finding these connections in a rapidly growing repository and challenges in record curation processes at Dryad.

A principal component of the challenge is the period between submission of a dataset and publication of a related article with the DOI for making the link. This delay automatically puts finding links and adding them into the Dryad repository outside of the typical curation timeframe and into the re-curation timeframe. The general approach described above, i.e. searching Crossref for metadata and adding that metadata to the record cannot be used because the connection to the article does not exist. Other possibilities include ScholeXplorer [29] and several title search strategies like the Google search used above to find the article associated with an existing Dryad dataset.

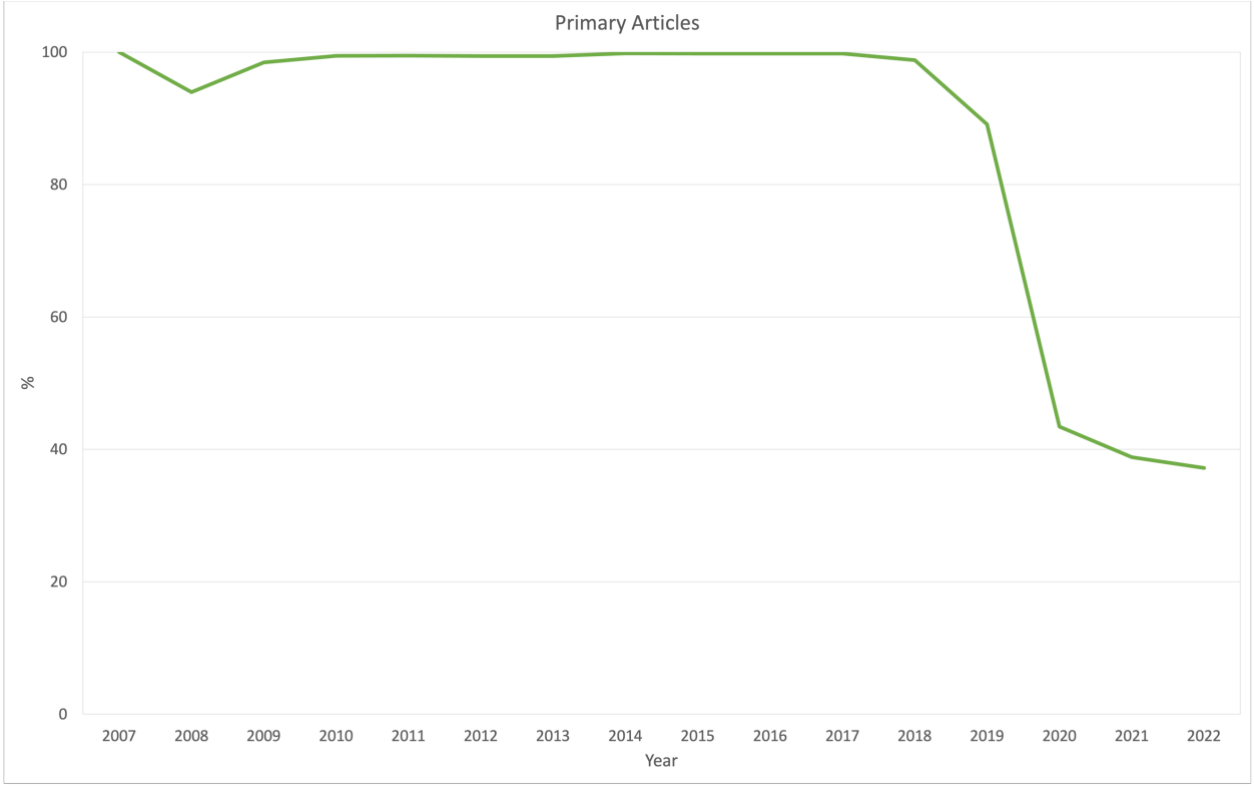

Figure 7. % of journal-related datasets with primary articles identified.

The Framework for Scholarly Link Exchange (Scholix, [30]) is a service aimed at establishing guidelines for exchanging metadata about links between scholarly literature and scientific data and a high-level framework for accessing those metadata. The guidelines have been created by the Research Data Alliance (RDA) and the World Data System (WDS) Scholarly Link Exchange Working group [31] and the framework is operational based on the Scholix Metadata Schema [32] and API. Searching this framework for Dryad DOIs should surface links to those DOIs created by Crossref or by journals when articles referencing the datasets are published.

The second option, searching for related papers using Dryad dataset titles is made easier by the common practice of naming Dryad datasets using the expected name of the published paper. For example, the dataset “Data from: Wildfire catalyzes upward range expansion of trembling aspen in southern Rocky Mountain beetle-killed forests” published in Dryad [33] during January, 2022, is likely data used in a paper titled “Wildfire catalyzes upward range expansion of trembling aspen in southern Rocky Mountain beetle-killed forests” [34]. Searching Google for this title yields two links to the article, one on a journal page and one in the U.S. Forest Service library. The journal page contains two machine-readable meta tags that give the DOI: <meta name="citation_doi" content="10.1111/jbi.14302"/> and <meta name="dc.identifier" content="10.1111/jbi.14302"/> which can then be searched for article metadata. In this case, the Crossref search yields no new affiliations or ORCIDs, but it does include two funders.

This example clearly depicts how these title searches can happen in a perfect world, but automating google searches and matching titles across thousands of datasets in the real-world is a

more complicated task. Dryad is currently exploring this option with the goal of integrating it into the standard processing.

5.4 Preprint Datasets

Preprint datasets are a special category of datasets without primary_articles because preprints typically have DOIs that will be connected to the DOI of the associated peer-reviewed paper when it is published. This time delay is like that discussed above for all Dryad datasets, but, in the preprint case, the preprint repositories and journals are enlisted in the dataset-paper linking process.

Despite this community involvement, considerable problems linking preprints to papers still exist. Cabanac et al. [35] discussed these problems in detail and described a technique for finding links using Crossref metadata and criteria that combined titles, publication dates, and first author names. Eckmann and Bandrowski [36] described a preprint-publication linker that uses broader measures of similarity including the abstracts.

The number of preprints in Dryad is relatively small (~1000) but they do contribute to the datasets without primary articles shown in Figure 7. Most preprints with datasets in Dryad are in the BioRxiv repository [37] which provides community supported links to published papers for some of these preprints. Keeping the caveat of incomplete coverage in mind, the BioRxiv API [38] was used to find published DOIs for these preprints. In a sample of 721 preprints, 389 published articles were found (54%). This approach could also be integrated into standard Dryad processing to improve recognition of peer-reviewed articles related to preprints.

5.5 Datasets submitted without papers

Dryad has recently begun accepting independent datasets without expectations of connected papers. Examining 44,486 Dryad datasets associated with organizations showed that 1,727 of those (4%) do not have a related ISSN identifying an associated journal. This percentage may grow in the future, but these datasets only make a small contribution to the missing connections identified in Figure 7.

6. Funder / Journal / Organization / Connectivity

The results reported above are examples of *repository connectivity* – calculated over entire repositories, Dryad in this case. Connectivity can also be calculated across repository subsets, for example all datasets associated with a funder, a journal, or an organization, to determine whether the available identifiers are in place. High-level summaries of those observations are shown here using connectivity visualizations described by Habermann, 2023 [18].

6.1 Funder Connectivity

Funder connectivity depends on funder and award identifiers, and each has independent connectivity. Funders with complete connectivity (lower band in Figure 8, green) include funder identifiers in the metadata for all the datasets they are associated with. That is, 30% of the funders in the dataset (3538) always have an associated identifier. Those identified as Missing (red) have no identifiers. Funders with some identifiers (yellow) have identifiers in some cases.

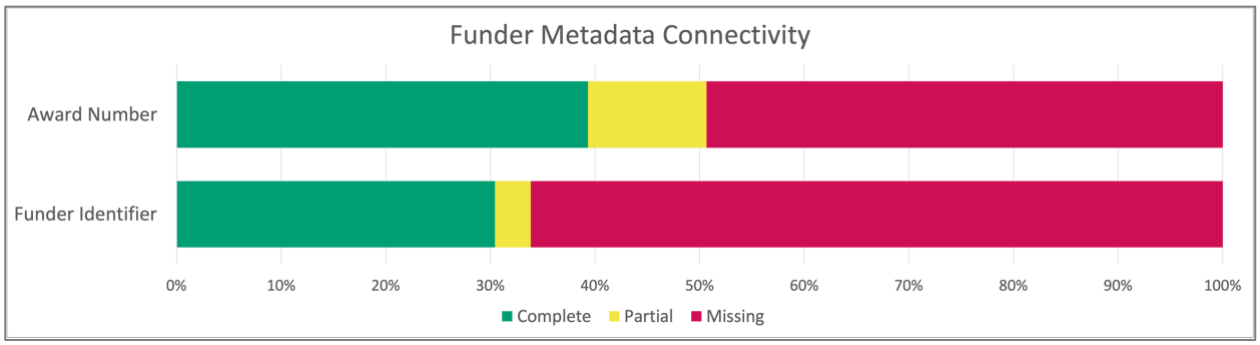

Figure 8. Connectivity for funder and award identifiers.

The data in Figure 6 shows that most of the Dryad funder metadata is for datasets published during the last several years. The funders identified as partial in Figure 8 are, therefore, opportunities for spreading funder identifiers to earlier datasets as described above for ORCIDs.

The upper band in Figure 8 shows the same data for award identifiers. The award identifier data are complete for more funders than the funder identifier (39%) and fewer funders are missing all award information (49%). This suggests that funder identifiers are more difficult for researchers to locate than award numbers for awards they have received.

6.2 Journal and Organization Connectivity

Journal connectivity depends on organizational and individual identifiers. The data in Figure 4 show that the number of organizational identifiers (RORs) in Dryad is much larger than individual identifiers (ORCIDs) and the journal connectivity shown in Figure 9 conforms with the expectations based on that data.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00252

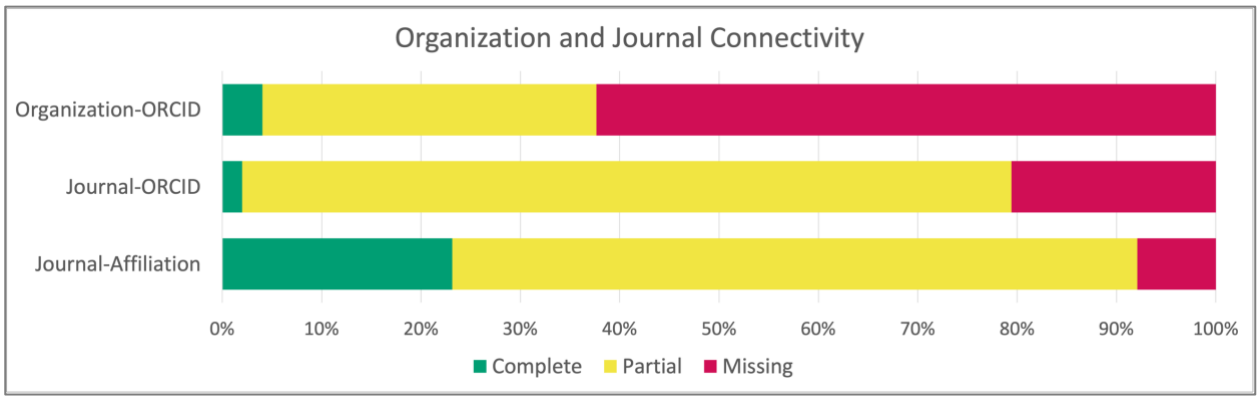

Figure 9. Journal and Organization ORCID and Affiliation Connectivity

The bottom band in Figure 9, labeled Journal-Affiliation, shows that many journals have organizational identifiers for all their organizations (23%, green) and only 8% of the journals are missing all organization identifiers (red). The rest (69%, yellow) have identifiers for some organizations.

The second band (Journal-ORCID) shows that only 2% of the journals have identifiers for all authors (green) while 21% have no individual identifiers (red), and 77% have some identifiers (yellow).

The Dryad repository includes datasets from many research organizations (mostly colleges and universities). These data were retrieved by organization to determine ORCID connectivity for each organization. These data (top band in Figure 9) show a pattern like the journals but with twice as many complete organizations and more missing (62%).

7. Conclusion

Identifiers of many kinds are the key to creating unambiguous and persistent connections between research objects and other items in the global research infrastructure. Many repositories

include research objects that were submitted and curated before these identifiers were created or implemented, making it difficult to connect those research objects into the big picture. Repository re-curation can be used to ameliorate this problem by finding identifiers and augmenting existing metadata. This approach has been used in the Dryad Data Repository to increase identifier completeness for organizations, people, funders, and related papers.

The first re-curation effort was undertaken during 2018-2019 as part of the migration of the Dryad repository to the California Digital Library. This work took advantage of DOIs for papers connected to Dryad datasets, searched metadata for those DOIs to find affiliations and searched the Research Organization Registry (ROR) for identifiers for those affiliations. Figure 4 shows that the results of that effort come very close to the results of collecting RORs during the submission process since 2020.

The second re-curation effort focused on Funder identifiers for datasets in Dryad since 2020. This effort introduced identifiers for ~88% of the funders for datasets in the Dryad repository since that time (Figure 3). Improving these results and extending their temporal coverage depends on consistent funder names and award numbers as datasets are submitted to the repository.

Re-curating identifiers for people into the Dryad repository remains as a significant challenge even though ORCIDs have been used as Dryad logins since 2019. The % of author occurrences with ORCIDs remains close to 30%. The approach used for organizations and funders, i.e., searching DOIs for related papers for identifiers, does not work well because of the paucity of ORCIDs in journal metadata. Spreading known ORCIDs through the repository and

484 searching orcid.org for authors can both help improve individual connectivity, but both
485 approaches have significant challenges.

486 All these re-curation efforts depend critically on connections between Dryad datasets and
487 journal articles produced using those datasets. These connections have been a critical part of the
488 Dryad mission since its formation during 2008. As the Dryad community has grown to include
489 over 5,000 unique datasets from over 20,000 unique authors and over 4,000 unique organizations
490 per year (Figure *2*), the % of datasets with connections to journal articles has dropped
491 significantly (Figure 7) to <40%.

492 This unexpected decrease in Dryad connectivity raises important questions about continuing
493 the long-term Dryad commitment to connecting data with journal articles in the face of the five-
494 fold increase in repository submissions. All members that make up the growing Dryad
495 community shown in Figure *2* have a stake in finding more a sustainable approach to finding and
496 recording these connections. Increased utilization of automated tools for finding these
497 connections may be part of the solution, but current automated efforts [36] have not been
498 successful. Increased engagement of the journals and research organizations that support Dryad
499 is also important and the community needs find mechanisms for working together to sustain
500 these connections. The techniques described here can provide metrics for quantitatively
501 demonstrating future progress.

502 The complete global research infrastructure includes many repositories: institutional,
503 generalist, commercial, and non-profit. Like Dryad, these repositories are faced with challenges
504 related to getting connected and staying connected in an ever-changing landscape. Dryad has
505 taken an active approach to addressing these challenges reflected in the re-curation efforts and

results described here. Measuring connectivity and the results of re-curation work are important for identifying opportunities, defining baselines for measuring future improvements, and for demonstrating successes and impacts and the techniques described here can be useful across many repositories.

8. Acknowledgements

This work was funded by the U.S. National Science Foundation (Crossref Funder ID: 100000001, ROR: https://ror.org/021nxhr62) Award 2134956. Current and past Dryad staff, particularly Daniella Lowenberg and Ryan Sherle, were very helpful in initiating this work and in understanding technical aspects of the Dryad Repository. John Chodacki was very helpful with Dryad history.

9. Data Availability

The data used in this work are available in the Dryad Data Repository (DOI: 10.5061/dryad.nzs7h44xr)

10. References

1. Dryad: Who We Are. Available at: https://datadryad.org/stash/about, Accessed June 7, 2023.
2. DataCite, (2023), Connecting to Works. Available at: https://support.datacite.org/docs/connecting-to-works, Accessed January 12, 2024.

3. Hendricks, G. and Buys, M.: Working for Global Equity through Digital Object Identifiers. Available at: https://upstream.force11.org/working-for-global-equity-through-digital-object-identifiers/ (2023). https://doi.org/10.54900/6sz4q-47185, Accessed October 12, 2023.
4. Dryad: Depositing Data to Dryad. Available at https://web.archive.org/web/20080602032626/http://datadryad.org/depositing.html. Accessed May 29, 2023.
5. Dryad: Dryad partnering with CDL to accelerate data publishing. Available at: https://blog.datadryad.org/2018/05/30/dryad-partnering-with-cdl-to-accelerate-data-publishing/ (2018). Accessed March 22, 2023.
6. Dryad: New Dryad is Here. Available at: https://blog.datadryad.org/2019/09/24/new-dryad-is-here/ (2019). Accessed March 22, 2023.
7. DataCite, (2021). DataCite Metadata Schema. Available at: https://schema.datacite.org/, Accessed May 29, 2023.
8. Gould, M., and Lowenberg, D., (2019). ROR-ing Together: Implementing Organization IDs in Dryad. Available at: https://ror.org/blog/2019-07-10-ror-ing-together-with-dryad/, Accessed March 22, 2023.
9. Habermann, T., (2019). Dryad Data Packages and Files. Available at: https://metadatagamechangers.com/blog/2019/2/11/dryad-data-packages-and-files-1, Accessed March 22, 2023.
10. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

11. GREI, (2022), The Generalist Repository Ecosystem Initiative. Available at: https://datascience.nih.gov/data-ecosystem/generalist-repository-ecosystem-initiative, Accessed March 22, 2023.
12. GREI, (2022), Best practices for sharing data in a generalist repository. Available at: https://osf.io/h59ge, Accessed March 22, 2023.
13. COAR, (2022), COAR Community Framework for Good Practices in Repositories. Available at: https://www.coar-repositories.org/coar-community-framework-for-good-practices-in-repositories/, Accessed May 29, 2023.
14. The National Science and Technology Council, Desirable Characteristics of Data Repositories for Federally Funded Research, 2022. Available at: https://doi.org/10.5479/10088/113528
15. OSTP, (2022), Public Access Memo. Available at: https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf, Accessed March 22, 2023.
16. OPM, (2023), Good Measurement Makes a Difference in Organizational Performance. Available at: https://www.opm.gov/policy-data-oversight/performance-management/measuring/good-measurement-makes-a-difference-in-organizational-performance/, Accessed March 22, 2023.
17. Voehl, F. and Harrington, H.J.: Change Management: Manage the Change or It Will Manage You, CRC Press, Boca Raton, FL. (2016),
18. Habermann, T. (2023). Improving Domain Repository Connectivity. Data Intelligence, 5(1), 6–26 (2023). https://doi.org/10.1162/dint_a_00120.

19. Johnston, L. R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., Stewart, C., Blake, M., Herndon, J., McGeary, T. M., & Hull, E. (2018). Data Curation Network: A Cross-Institutional Staffing Model for Curating Research Data. In International Journal of Digital Curation (Vol. 13, Issue 1, pp. 125–140). Edinburgh University Library. https://doi.org/10.2218/ijdc.v13i1.616
20. ROR, (2023) . Available at: About, https://ror.org/about/, Accessed March 22, 2023.
21. ROR, (2023), Match organization names to ROR IDs. Available at: https://ror.readme.io/docs/match-organization-names-to-ror-ids, Accessed January 12, 2024.
22. Habermann, T., (2022), Need help searching for RORs? Try RORRetriever!. Available at: https://metadatagamechangers.com/blog/2022/6/30/rorretriever, Accessed January 12, 2024.
23. Habermann, T., (2021). Acronyms Are Definitely Not Enough. Available at: https://metadatagamechangers.com/blog/2021/7/16/acronyms-are-definitely-not-enough, Accessed March 22, 2023.
24. Diego S Porto, Wasila M Dahdul, Hilmar Lapp, James P Balhoff, Todd J Vision, Paula M Mabee, Josef Uyeda, Assessing Bayesian Phylogenetic Information Content of Morphological Data Using Knowledge From Anatomy Ontologies, Systematic Biology, Volume 71, Issue 6, November 2022, Pages 1290–1306, https://doi.org/10.1093/sysbio/syac022.
25. Anderson A. Ferreira, Marcos André Gonçalves, and Alberto H.F. Laender. 2012. A brief survey of automatic methods for author name disambiguation. SIGMOD Rec. 41, 2 (June 2012), 15–26. https://doi.org/10.1145/2350036.2350040.

26. Sanyal, D. K., Bhowmick, P. K., & Das, P. P. (2021). A review of author name disambiguation techniques for the PubMed bibliographic database. Journal of Information Science, 47(2), 227–254. https://doi.org/10.1177/0165551519888605

27. Crossref, (2023), Funder Registry. Available at: https://www.crossref.org/services/funder-registry/, Accessed May 29, 2023.

28. French, A. and Buttrick, A., (2023) How ROR and the Open Funder Registry Overlap: A Closer Look at the Data. Available at: https://ror.org/blog/2023-10-12-ror-funder-registry-overlap/, Accessed November 8, 2023.

29. ScholeXplorer, (2023) . Available at: https://scholexplorer.openaire.eu/#/, Accessed January 12, 2024.

30. Cousijn, H., Feeney, P., Lowenberg, D., Presani, E. and Simons, N., 2019. Bringing Citations and Usage Metrics Together to Make Data Count. Data Science Journal, 18(1), p.9. DOI: http://doi.org/10.5334/dsj-2019-009

31. RDA, (2019), RDA/WDS Scholarly Link Exchange (Scholix) WG, Accessed March 22, 2023.

32. La Bruzzo, Sandro and Manghi, Paolo. (2022). The Scholix Metadata JSON Schema (4.0). Zenodo. https://doi.org/10.5281/zenodo.6351557, Accessed March 22, 2023.

33. Nigro, Katherine et al. (2022). Data from: Wildfire catalyzes upward range expansion of trembling aspen in southern Rocky Mountain beetle-killed forests [Dataset]. Dryad. https://doi.org/10.5061/dryad.crjdfn348

34. Nigro, K. M., Rocca, M. E., Battaglia, M. A., Coop, J. D., & Redmond, M. D. (2022). Wildfire catalyzes upward range expansion of trembling aspen in southern Rocky

610 Mountain beetle-killed forests. Journal of
611 Biogeography, 49, 201– 214. https://doi.org/10.1111/jbi.14302
612 35. Cabanac, G., Oikonomidi, T. & Boutron, I., (2021), Day-to-day discovery of preprint–
613 publication links. *Scientometrics* **126**, 5285–5304, https://doi.org/10.1007/s11192-021-
614 03900-7.
615 36. Eckmann P, Bandrowski A (2023) PreprintMatch: A tool for preprint to publication detection
616 shows global inequities in scientific publication. PLoS ONE 18(3): e0281659.
617 https://doi.org/10.1371/journal.pone.0281659.
618 37. BioRxiv, (2023), BioRxiv, The Preprint Server for Biology. Available at:
619 https://www.biorxiv.org/, Accessed January 12, 2024.
620 38. BioRxiv, (2023), Machine access and text/data mining resources. Available at:
621 https://www.biorxiv.org/tdm, Accessed January 12, 2024.
622 39. Dryad, (2017). Improvements in data-article linking. Available at:
623 https://blog.datadryad.org/2017/12/18/improvements-in-data-article-linking/, Accessed May
624 29, 2023.