JLSC

ISSN 2162-3309 | JLSC is published by the Iowa State University Digital Press | http://jlsc-pub.org

Volume 14, 1 (2026)

Assessing Researcher Data Sharing Practices Using Publicly Available Dataset Metadata: A Reproducible Workflow for Institutional Stakeholders

Danielle R. Kirsch & Isaac Wink

Kirsch, D.R. & Wink, I. (2026). Assessing Researcher Data Sharing Practices Using Publicly Available Dataset Metadata: A Reproducible Workflow for Institutional Stakeholders. *Journal of Librarianship and Scholarly Communication*, 14(1), eP22913. https://doi.org/10.31274/jlsc.22913

This article underwent semi-anonymous peer review in accordance with JLSC's peer review policy.

ISSN 2162-3309 10.31274/jlsc.22913

RESEARCH ARTICLE

Assessing Researcher Data Sharing Practices Using Publicly Available Dataset Metadata: A Reproducible Workflow for Institutional Stakeholders

Danielle R. Kirsch
Oklahoma State University
Isaac Wink
University of Kentucky

ABSTRACT

Introduction: Scholarly communities are experiencing increased emphasis on research data sharing and reuse. Given the range of available data repositories and variability in the use of persistent identifiers (PIDs) for individuals and institutions, aggregating datasets by researchers at a specific institution is a significant challenge. We developed a reproducible workflow to evaluate 1) data repository use by researchers at our institutions, 2) metrics of reuse, and 3) quality of metadata records.
Methods: We used the DataCite REST API to locate metadata records for datasets with creator affiliation names that matched our institutions. We performed substantial cleaning and deduplication before comparing repository use, citations, usage metrics, and metadata completeness.
Results: The most common data repositories were Dryad, Harvard Dataverse, figshare, Zenodo, ICPSR, and one institution's institutional repository. View, download, and citation counts were available from a limited number of repositories, with some discrepancies between different citation reporting methods. PIDs were more frequently used for authors and affiliations than funders, and the inclusion of PIDs varied across repositories, with Dryad being the most consistent.
Discussion & Conclusion: Although broad trends in repository and PID use were similar between institutions, our analysis also surfaced examples that illustrate inconsistencies in metadata across repositories. Variable implementation of the DataCite metadata schema requires a significant amount of data cleaning to obtain meaningful results. Even then, incomplete metadata makes some datasets impossible to locate. Repositories, funders, researchers, and institutional open data advocates must coordinate to create complete and usable metadata that integrates datasets into the scholarship ecosystem.

Keywords: datasets, persistent identifiers, metadata, public access, data repository

Received: 01/22/2026 Accepted: 05/01/2026

IMPLICATIONS FOR PRACTICE

1. We provide a reproducible workflow in both R and Python that others can use to obtain metadata from the DataCite database for datasets produced by researchers at a given institution.
2. Among commonly used repositories, the Dryad data repository had the most consistent use of persistent identifiers for authors, affiliations, and funders.
3. Currently, a limited number of repositories report dataset use metrics such as views, downloads, and citations. Citation linking in particular is important to researchers, but those links are often missing.
4. Repositories, researchers, and institutional stakeholders such as librarians should work together to improve the implementation of PIDs (e.g., ORCID, ROR), as those metadata elements are essential for capturing individual and institutional research outputs.

INTRODUCTION

Data sharing, an essential component of open research practices, is becoming increasingly common across disciplines as many funders and journals now require that researchers make their data openly available. For example, since 2023, the U.S. National Institutes of Health has required grant awardees to "maximize the appropriate sharing of scientific data," typically via established data repositories (National Institutes of Health, 2023). While datasets published in repositories are part of the scholarly communication ecosystem, they are not yet as reliably integrated as other outputs (notably journal articles). By understanding both where researchers at their institutions share data and the quality of the metadata records for those datasets, librarians and data curators play a key role in strengthening the impact and connectivity of data within the research ecosystem. Thus far, the results of their engagement with datasets and their metadata include working with researchers to make their data FAIR (findable, accessible, interoperable, and reusable; Wilkinson et al., 2016); better evaluating the metadata support offered by a data repository; and aggregating datasets for tasks such as relocating ephemeral data to institutional repositories and creating public data catalogs to enhance discovery (Briney, 2024; Mannheimer et al., 2021; markuslibrary, 2025).

However, simply determining where researchers share their data is a challenge. Generalist, domain-specific, and consortial or institutional data repositories offer a plethora of options to meet researchers' specific data needs, meaning researchers from a single institution may share datasets across dozens of repositories. Fortunately, comparisons across major repositories are possible because many utilize the DataCite Metadata Schema, which consists of both

mandatory properties (e.g., creator, title, publisher, resource type) as well as recommended and optional properties (e.g., persistent identifiers for creators and their affiliations, funding information, and related resources such as citing works).

Using the DataCite application programming interface (API), we accessed records for thousands of datasets created by researchers at our institutions, Oklahoma State University (OSU) and University of Kentucky (UK). Our analysis was limited to scholarly outputs 1) minted with DataCite DOIs and 2) reported as a "dataset" resource. By leveraging the metadata provided with these records, we worked to understand patterns and practices in data sharing specific to our universities' researchers and better inform our research support services. Specifically, we asked:

1. Which data repositories are researchers using?
2. How often are datasets from our institutions being cited and reused?
3. How is usage measured through alternative metrics of reuse, such as download and view counts?
4. What quality of metadata exists for these datasets?

We provide a reproducible workflow that can serve as a starting point for stakeholders at other institutions to ask these same questions regarding their own researchers' data. We have published scripts written in both R and Python (with similar functionality) for querying the DataCite API, converting metadata records to tabular form, performing data cleaning, and creating visualizations of results.

Toward an institutional understanding of research data sharing

Our research comes at a moment of transition in which datasets are now treated as significant research outputs, and institutions, journals, and repositories are more carefully considering their data sharing and description practices with a focus on long-term preservation and access. Journals such as *Scientific Data* in the Nature portfolio now highlight datasets as standalone outputs via data papers, and data sharing policies in journals across disciplines frequently require datasets accompanying articles to be hosted in a persistent location, such as a data repository. Additionally, there have been major improvements in persistent identifiers (PIDs) to create more reliable linkages in the research ecosystem. While digital object identifiers (DOIs) and Archival Resource Keys (ARKs) as PIDs for publications and ORCID iDs for researchers are well established, and Research Organization Registry (ROR) IDs for institutions are increasing in popularity, there is growing interest in PIDs for other elements as well, such as research instruments and physical samples (Mayernik et al., 2024).

University administrators have also taken an increased interest in where researchers are depositing their data, driven in large part by funder data sharing requirements. Elsevier briefly offered the Data Monitor service to give institutions insight into where their researchers' data could be found (Pure Help Center, n.d.). Although Data Monitor sunsetted in 2025, it is an indication of growing demand for products that offer a view of institutional data collections as a whole.

Publisher and administrator concerns regarding research data intersect with those of librarians and data curators, who must understand data sharing practices at their institutions to inform decision-making. As one example, an academic library looking to purchase an institutional membership with a data repository such as Dryad or figshare would benefit from knowing how frequently researchers are already using a repository; they can also use this information to advocate internally for funds to support such a membership. This investigation seeks to aid all institutional stakeholders who look to understand research data produced at their institutions as collections, improve the services and training they offer, and articulate their technical needs for metadata to data repositories.

LITERATURE REVIEW

Persistent identifiers in the scholarly communications ecosystem

PIDs are critical to linking research outputs to individuals and institutions. ORCID iDs are the most widely used PID for individual researchers, and Research Organization Registry (ROR) IDs are commonly used for organizations, including research funders. DataCite DOIs are themselves persistent identifiers for research outputs (including datasets), and other objects, such as scientific instruments, and the DataCite Metadata Schema enables the inclusion of other PIDs—for example, an ORCID iD included in a dataset's metadata record links the dataset to its creator.

Working individually, PIDs improve the availability of scholarly outputs and aid in disambiguation. Lisa Federer has demonstrated that assigning a DOI to data, code, and other outputs improves their long-term accessibility compared to simply sharing a URL for an output, which may become unavailable over time (Federer, 2022). ORCID iDs can help disambiguate between authors with identical names and link works by an author who may have published under different names over time, though it has been noted that disambiguation is more difficult for "empty" ORCID profiles with no information beyond a name (Heusse & Cabanac, 2022; Fernández-Marcial et al., 2023).

Beyond maintaining access to an individual resource, PIDs provide particular value in the context of scholarly communication by linking together people, organizations, and research

objects (Gould, 2022; Adamich, 2016). PIDs underpin the scholarly linkages tracked in databases such as Web of Science and OpenAlex and enable bibliometric analysis of researcher publishing patterns (Hazarika & Sudhier, 2025). Given the gradual adoption of PIDs for authors, institutions, and research funders, some of these potential benefits of a robustly linked scholarly ecosystem remain speculative, though individual projects present exciting use cases. For instance, the Science and Technology Facilities Council (STFC) in the United Kingdom leveraged PIDs to create a network graph linking publications, authors, and funders, offering "greater oversight of the effectiveness of the research funding [STFC] provides and a better means to track the outcome of the research" (Madden & Bunakov, 2019). DOI metadata that records an item's linkages to other resource DOIs has been noted to have particular relevance for researchers practicing publicly-engaged scholarship and producing a diverse range of scholarly products (Burton et al., 2023).

However, linking nodes in the scholarly ecosystem depends on complete records, especially for research outputs minted with DOIs. Implementation of PIDs and other nonmandatory properties of the DataCite schema is not consistent across repositories or individual datasets, leading to high variance in quality and completeness of metadata. Previous evaluations (Wettere, 2021; Malik & Mandal, 2024) have identified a variety of gaps in dataset metadata records, with one recent article concluding that "the metadata quality landscape for published research datasets is uneven; key information, such as author affiliation, is often incomplete or missing from source data repositories and aggregators" (Johnston et al., 2024).

Research datasets as collections

While our aim here is a quantitative analysis of the dataset output of institutions as a whole, rather than the creation of a public discovery point for institutional datasets, this investigation shares challenges with past work developing data catalogs. Serving as a discovery platform for researchers and institutional stakeholders, data catalogs must determine how to systematically identify datasets produced by researchers at a given institution. The New York University Health Sciences Library Data Catalog only includes datasets that researchers have self-submitted, ensuring all depositors have a genuine institutional affiliation (Lamb & Larson, 2016). A recently-launched data catalog from the University of Alabama at Birmingham (UAB) harvests records from DataCite by querying for the UAB ROR ID, but it separately queries the Zenodo repository for a text string matching the institution name (markuslibrary, 2025). Similar workflows are being developed at the University of Texas at Austin to retrieve dataset metadata records from DataCite, Dataverse, Dryad, and Zenodo APIs (Gee, 2025). The Montana State University Dataset Search team queried data repository APIs for individual faculty names and had human curators disambiguate researchers and enhance metadata records (Lamb & Larson, 2016). These projects showcase opportunities

to leverage APIs to streamline dataset collection, but also highlight the continued importance of human review and manual enhancement of metadata records.

METHODS

The DataCite metadata schema

Since 2014, the DataCite consortium, which specializes in minting digital object identifiers (DOIs) for datasets, has maintained the DataCite Metadata Schema as "a list of core metadata properties chosen for the accurate and consistent identification of a resource for citation and retrieval purposes" (DataCite Metadata Working Group, 2014). The most recent version of the schema (4.6) was released in December 2024 and contains 20 top-level properties, many with additional subproperties (DataCite Metadata Working Group, 2024). Since many data repositories mint DOIs using DataCite, the metadata included for each dataset offers a rich opportunity to not only describe individual datasets but also identify trends across repositories or within a defined population, such as datasets produced by researchers affiliated with a single institution.

Accessing dataset metadata from DataCite

The most user-friendly option for accessing dataset metadata is through DataCite Commons, which allows users to search for individual datasets or datasets associated with a given researcher, institution, or data repository. However, DataCite Commons relies on PIDs to link datasets to related creators. For example, when a user searches for datasets affiliated with a given institution, DataCite displays all datasets containing that institution's ROR ID (DataCite, n.d.c). Due to missing affiliation metadata, many datasets are left out of searches via DataCite Commons. Additionally, while DataCite Commons offers downloadable reports of research funders and related works (DataCite, n.d.e), it does not provide bulk access to the complete metadata records for each dataset, making access to and comparison of specific properties difficult (DataCite, n.d.c).[1]

The DataCite REST API offers a more granular means of accessing metadata. Rather than relying exclusively on PIDs, API queries can more flexibly return all results that match one or multiple pieces of metadata (DataCite, n.d.d). We tested two queries for the DataCite API that returned differing results, an early indicator of discrepancies in how repositories record metadata.

[1] To find a ROR ID for an institution, a user may search for that institution's name on DataCite Commons, and the ROR ID will be displayed on the institution's information page. Alternatively, a user may search directly from the ROR website at ror.org/search for additional information on an institution, including its relationship to other organizations.

Query 1:

https://api.datacite.org/dois?affiliation-id=[ROR%20ID]&resource-type-id=dataset&affiliation=true&publisher=true&page[size]=1000

This query was built using the get_dois reference tool provided by DataCite (DataCite, n.d.g). The components of this query are as follows:

- **affiliation-id=[ROR%20ID]** provides the ROR ID for the institution of interest and searches for that ROR ID in both creator and contributor affiliations. For OSU, for example, this was encoded as: affiliation-id=2F01g9vbr38.
- **resource-type-id=dataset** limits results to resources classified as datasets, as opposed to other DataCite resources, such as an image or software.
- **affiliation=true** provides all affiliation metadata, if available.
- **publisher=true** provides publisher identifier details, if available.
- **page[size]=1000** establishes that the API exports 1000 records per page.

This query is limited to datasets where creator and/or contributor metadata includes the specified ROR ID (indicating that the creator/contributor is affiliated with the specified institution). The results from this query nearly identically match the results yielded by searching an institution in DataCite Commons and filtering for only datasets. While ROR IDs accurately identify an affiliated institution and are accompanied in the metadata by the institution's name, there are many instances in which a ROR ID is not provided by a dataset's authors or its publisher. As a result, identifying datasets affiliated with an institution by ROR ID alone risks leaving out a large number of datasets.

Query 2:

https://api.datacite.org/dois?query=creators.affiliation.name:*[Institution\%20Name]*&resource-type-id=dataset&affiliation=true&publisher=true&page[size]=1000

In place of the **affiliation-id=[ROR%20ID]** parameter used in Query 1, this query utilizes the **creators.affiliation.name:*[Institution\%20Name]*** parameter (with "\%20" encoding a space between words), which indicates that the "name" attribute listed under a creator's institutional affiliation must match the name of the institution entered.[2] Importantly, the "name"

[2] More specifically, this parameter means that an affiliation must contain the institution name exactly as entered, but may contain additional text before and after the name. This means that the query still matches affiliations

attribute is just a text string, meaning that while the match is sensitive to typos in how an institution name was entered, it does not require that the record contain a ROR ID for the institution.[3]

Because this query checks only for a text string match for the institution name rather than a ROR ID, it provides a more expansive list of results than searching for a ROR ID alone. However, this call is more likely to generate false matches, especially for institutions that share parts of their names with other entities in the same system or with similar name elements.[4] Our workflow provides a method for flagging potentially incorrect institutional matches and only keeping relevant affiliations.

API calls conducted on July 25, 2025, revealed that limiting searching to ROR IDs missed more than 800 datasets potentially affiliated with UK and more than 250 potentially affiliated with OSU (Table 1).

Institution	Query 1 (ROR ID)	Query 2 (Text String)
Oklahoma State University (OSU)	1505	1780
University of Kentucky (UK)	574	1385

Table 1. A comparison of the number of dataset DOIs returned by the DataCite API when querying by ROR ID vs. by text string of an institution name. Note that at this stage, no data cleaning or deduplication had been performed.

Data cleaning

The DataCite API returns results in a nested JSON format, so these results first need to be parsed to extract information relevant to the user. Certain metadata, such as the dataset DOI, usage metrics, publisher, and publication year, are fairly straightforward to extract as they are

such as "University of Kentucky Research Foundation, Inc.," and not just "University of Kentucky." Depending on a user's institutional context and need, this parameter may need to be adjusted. Additionally, due to changes in the DataCite API since July 2025, the original query described here no longer functions, but comparable results can be achieved by encoding quotation marks using %22. For example, in place of "*University%20of%20Kentucky*", the search string should now be "%22University%20of%20Kentucky%22".

[3] This query only searches affiliations in the "creator" field, not "contributor." See "Metadata Schema and API Query Limitations" for more details.

[4] For instance, querying for "*University of California*" will return results for affiliations such as "University of California, Riverside" and "Dominican University of California," but would miss affiliations written with abbreviations, such as "UC Santa Barbara." DataCite queries support the use of regex, so while it is possible to build more sophisticated queries, there remains a tradeoff between false positives and false negatives.

single-entry items (e.g., a dataset only has a single DOI per record). Other metadata–specifically for authors, funders, and related works–consist of multiple nested elements (e.g., ORCID iDs nested under individual author entries) and/or multiple distinct items (e.g., more than one funder). The following metadata were extracted:

- **Datasets:**
  - DOI, publication year, publisher (repository) name, publisher identifier, title, view counts, download counts, and citation counts
  - Delimited lists of author names, funders, and related work DOIs
- **Authors:**
  - Full name, given name, family name, author identifier, affiliation, and affiliation identifier
- **Funders:**
  - Name, funder identifier, award number, and award title
- **Related works:**
  - Relation type, related work DOI, and resource type

Before extracting any data, we first manually reviewed all institution names that were "hits" for our API query to check for false matches. We discovered our query returned datasets affiliated with Southeast Oklahoma State University and Southwest Oklahoma State University (both distinct from OSU) that were not affiliated with institutions in the OSU system. We manually flagged these as false matches and removed them.

To keep the data manageable, we created six separate tabular data files (dataset metadata, author metadata, funder metadata, related works metadata, a list of all funders returned by the query, and a list of all related works returned) and joined them as needed using index numbers. A visualization of our cleaning workflow and intermediate tables is found in Figure 1.

Name standardization

Given the variability of certain metadata items, we conducted additional data cleaning to standardize and consolidate names of authors, funders, and repositories, as well as the formatting of ORCID iDs.

Authors. We restricted our scope to include only authors affiliated with our respective institutions. To ensure an accurate count of distinct, institutionally-affiliated authors in our data, we identified name variations for individuals and standardized them to a single value. We attempted

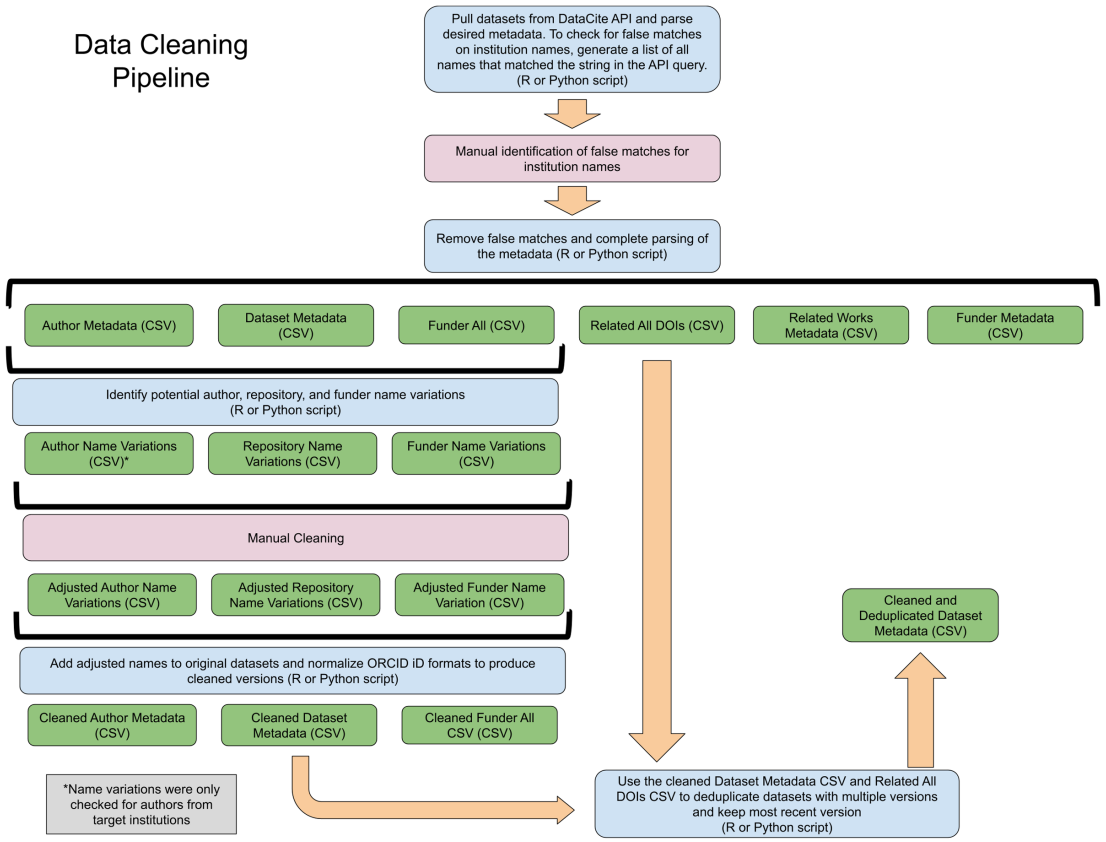

Figure 1. Data cleaning pipeline from the API query through the cleaning, standardization, and deduplication process.

to automate this process where possible by using "fuzzy string matching," which finds near matches between values based on the number of characters—including spaces—that differ between them. For example, "Davis, Craig" and "Davis, Craig A." would show up as a near match, differing by 3 characters: space, A, and period. We set our limit at a difference of no more than 5 characters, which resulted in many false positives (e.g., "Polo, John" and "Hodge, John" differ by only 4 characters but refer to different people) and some false negatives (e.g., "Elmore, Robert" does not match "Elmore, R. Dwayne" but refers to the same person).

We exported this list of matched author names into a CSV file[5] and continued with manual verification of those matches, indicating which name we wanted to keep and which needed to be discarded. We then worked through the full institutional author list to find any

[5] It is important to note that if name adjustments are performed in Excel, there may be errors in character encoding for special characters that appear upon opening the file in Excel but were not originally present in the

unidentified name variations and added those to the spreadsheet as well. Most name discrepancies involved inconsistent use of initials and abbreviated first names—such as Tim instead of Timothy—although there were some rare instances where first names or middle initials preceded the last name. Where possible, we used the most complete name as the individual's name standard (Last, First MI.). This list of original and adjusted author names was then merged with the DataCite author data to complete the author name standardization.

ORCID iDs. Although most ORCID iDs were provided as strictly alphanumeric codes, some were listed as a full URL. To standardize ORCID iD formatting, we removed the URL portion of the ORCID iD (https://orcid.org/) and kept only the alphanumeric code. We identified instances—though rare—of individuals who had multiple ORCID iDs affiliated with them, discrepancies that can be manually addressed as desired. Finally, these ORCID iDs were merged with the DataCite author data so that all institutional authors would have their ORCID iD (assuming one was provided) affiliated with each instance of their name.[6]

Funders. We extracted a de-duplicated list of funder names from the DataCite funder data. Some funder metadata included ROR IDs or DOIs for the specific funder, which facilitated the process of name standardization and consolidation. We opted to consolidate institutes and divisions under their broader funding agency (e.g., the Division of Integrative Organismal Systems was recategorized as NSF).[7] Individuals using the provided code can choose the extent to which they want to consolidate funders, depending on their exact interests and intended uses for the data.

There were also instances of a funder being listed more than once for the same dataset. In some cases, the metadata indicated that the research was funded by multiple awards from the same funder. However, the metadata often lacked sufficient detail to determine whether duplicated funders were true duplicates or the result of multiple awards. We merged the list of original and adjusted funder names with the DataCite funder data to complete this name standardization and consolidation. As a matter of convenience, we also abbreviated lengthy funder names to acronyms.

---

metadata. Names with special characters require additional checks, and the adjusted name CSV should be saved with UTF-8 encoding to ensure it can be properly read in Python or R.

[6] We stored these added ORCID iDs in a separate column to distinguish them from the ORCID iDs provided as part of the original author metadata. Later in the process, when we evaluated completeness of author metadata, we could still check whether an ORCID iD had originally been provided in the dataset's metadata.

[7] Some datasets published by Dryad listed funder names followed by an asterisk (such as "Max Planck Research School for Organismal Biology*"), a relic of a former system of controlled vocabulary that Dryad implemented. Our workflow offers an opportunity to remove these asterisks from funder names in which they appear.

Publishers/Repositories. Finally, we standardized and shortened publisher (repository) names in a similar way. This primarily involved using abbreviations for long repository names (e.g., ICPSR) and consolidating name variations, most commonly needed for federal data repositories (e.g., ScienceBase) and institutional data repositories whose names had shifted over time and required manual reconciliation. For example, some older datasets deposited with UK's institutional repository, UKnowledge, were recorded as being stored at UK or UK Libraries. We merged the list of original and adjusted repository names with the DataCite DOI data to complete this step.

Dataset versioning and deduplication

Finally, we addressed cases where multiple versions of the same dataset were present and minted with separate DOIs. To prevent datasets with multiple versions from being over-weighted in our analysis, we sought to remove any previously published versions of a dataset and keep only the most recent version.[8]

Publishers are not consistent in how they handle a new version of an existing dataset, necessitating different methods for identifying and removing them. figshare, Taylor & Francis, and ICPSR indicate a new version of a dataset by minting a new DOI with a version number added to it. For example, 10.3886/e105000 and 10.3886/e105000v1 refer to two versions of the same dataset in ICPSR. Our deduplication procedure used regex to find DOIs ending in a version number, identify all other DOIs with the same base, and drop everything but the most recent version.

Zenodo, meanwhile, creates two DOIs when a dataset or code package is first published, with one reserved for the "concept" of the material and the other applied to the specific version. When a new version of the material is published, it receives a new DOI, but the concept DOI remains the same, though it will now resolve to the most recent version of the material (Zenodo, n.d.). This relationship is expressed using the "relationType" field in the DataCite schema. Specifically, the "concept" DOI will include all versions of the material in the "HasVersion" field. Mendeley Data uses a similar process for versioning. To deduplicate Zenodo and Mendeley Data datasets, we dropped all Zenodo/Mendeley Data DOIs that appeared in the "HasVersion" field for another Zenodo/Mendeley Data DOI.

Similarly, Harvard Dataverse previously minted separate DOIs for every file in a project, linked to the main DOI via the "HasPart" relation type. We dropped every DOI that appeared under

[8] While this section focuses on the removal of multiple DOIs that correspond to different versions of the same dataset, not all repositories update DOIs when a new version is published. Dryad, for example, primarily does versioning locally, keeping the original DOI for a dataset and updating its metadata record. We did not seek to identify datasets that had been updated in this way, as they would not impact the total number of unique datasets from our institutions.

"HasPart" for another DOI in our table, keeping only the record for the main DOI. Other Dataverse instances, such as UNC Dataverse, appear to have followed the same practice, although we did not deduplicate them in our workflow as they were not among our most used repositories.

Finally, we encountered one case each for UK and OSU, where what was clearly a conceptually singular dataset had been spread across many different DOIs. One UK researcher had a project with 184 different tables, all minted with separate DOIs. Additionally, an OSU researcher was an author on 1,250 datasets published in the PANGAEA repository, which alone account for about 70% of datasets returned by the DataCite API for that institution. Although no metadata field indicated a relationship among the DOIs in these two instances, we added portions to our code to collapse them each down to a single DOI. Table 2 tabulates the total number of datasets that were removed through this process.

Repository	UK	OSU
ICPSR	79	15
Zenodo/Mendeley Data	75	61
Harvard Dataverse	355	51
figshare/Taylor & Francis	65	26
UKnowledge	183	0
PANGAEA	0	1249
Total	**757**	**1402**

Table 2. Number of datasets removed in our deduplication workflow by publisher/repository.

Analysis of datasets

A total of 43 repositories were represented in the OSU and UK data, but we primarily focused on the most commonly used repositories. All other repositories were categorized as "Other" except for instances where we were interested in exploring results across all applicable repositories.

Dataset citation and reuse. Dataset downloads, views, and citations are optional parts of DataCite metadata reporting but are not yet in widespread use, though the Make Data Count Initiative is encouraging their adoption by both repositories and researchers (Puebla & Chodacki, 2024). There are two ways in which citation information can be recorded and reported in DataCite. One method is through the citationCount property, which reports the number of citations for a particular DOI. The other is through the related works metadata, where related works with relationType "IsCitedBy" indicate a work that cites the DOI in question. We tabulated the total number of DOIs with the "IsCitedBy" relationType for each dataset to calculate the number of "IsCitedBy" citations for that dataset.

Dataset views and downloads. Although not part of the DataCite schema proper, DataCite enables repositories to report the number of views and downloads that a dataset has received using the DataCite Usage Reports API (DataCite, n.d.b). Specifically, a view is "a retrieval of the item in any way, whether the content was accessed or downloaded in full or in part," and a download is "the retrieval of all or part of the files within the item," defined using the COUNTER Code of Practice terminology (DataCite, n.d.h). Dataset views and downloads are useful indicators of the volume of interest in the dataset and its contents. Although not as impactful as explicit reuse of the dataset—primarily indicated by citations of the dataset by other publications—views and downloads are still important usage metrics. Given the issues in reliable dataset citation (discussed elsewhere in this article), views and downloads provide researchers with a more consistent measure of the impact of their shared data.

Metadata quality and completeness. The presence of persistent identifiers is crucial for the long-term findability and availability of resources, making it possible to aggregate datasets created by a particular researcher, affiliated with a given institution, or supported by a specific funder. Building on the work conducted by the Realities of Academic Data Sharing Initiative (Johnston et al., 2024), we examined the completeness of certain metadata fields identified in the 2022 Nelson Memo (White House Office of Science and Technology Policy, 2022), given that the inclusion of these fields is now of increased importance to U.S. funders and researchers. We focused on creator, affiliation, and funder information, investigating how many datasets had a complete record (meaning both a text name and a persistent identifier) for at least one author, institution, and funder.

RESULTS

We queried the DataCite API for datasets from both UK and OSU on July 25, 2025. All references to the number of datasets reflect those that were returned by the API on that date. Following data cleaning and deduplication, we conducted our analysis of 376 datasets for OSU and 628 for UK, with records that had been standardized where possible. By addressing our four research questions with results for two institutions, we offer two distinct case studies of data sharing practices and utilization of metadata infrastructure. While these results will surely be different for other institutions, we have endeavored to provide a reproducible workflow that can easily be adapted by stakeholders in other settings.

Question 1: Which data repositories are our researchers using?

The top 6 repositories were Dryad, Harvard Dataverse, Zenodo, figshare, ICPSR, and UKnowledge (UK's institutional repository). These repositories collectively housed more than 80% of

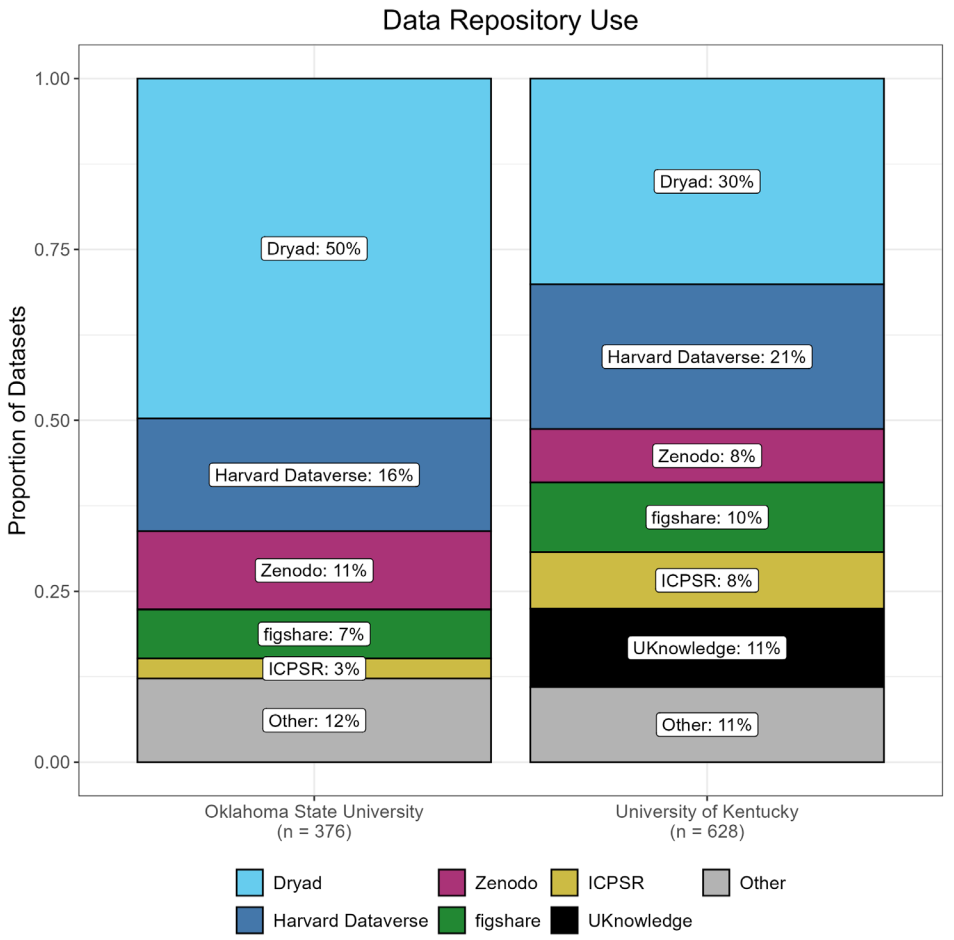

Figure 2. Comparison of data repository use between OSU and UK. Bars represent the proportion of all datasets from that institution that are in each data repository, with infrequently used repositories captured in the "Other" category.

datasets affiliated with each of our respective institutions. Dryad alone contained nearly 50% of OSU datasets and 30% of UK datasets. The distribution of datasets is shown in Figure 2.

Question 2: How often are datasets from our institutions being cited and reused?

citationCount

At least 1 citation via the citationCount metadata property was reported in 37% ($N = 140$) of OSU datasets and 38% ($N = 238$) of UK datasets, with a maximum citation count of 3 and 16, respectively. For OSU, Dryad accounted for about 84% ($N = 117$) of these reported citations, with 9 other repositories making up the remaining 16%. For UK, Dryad accounted for about

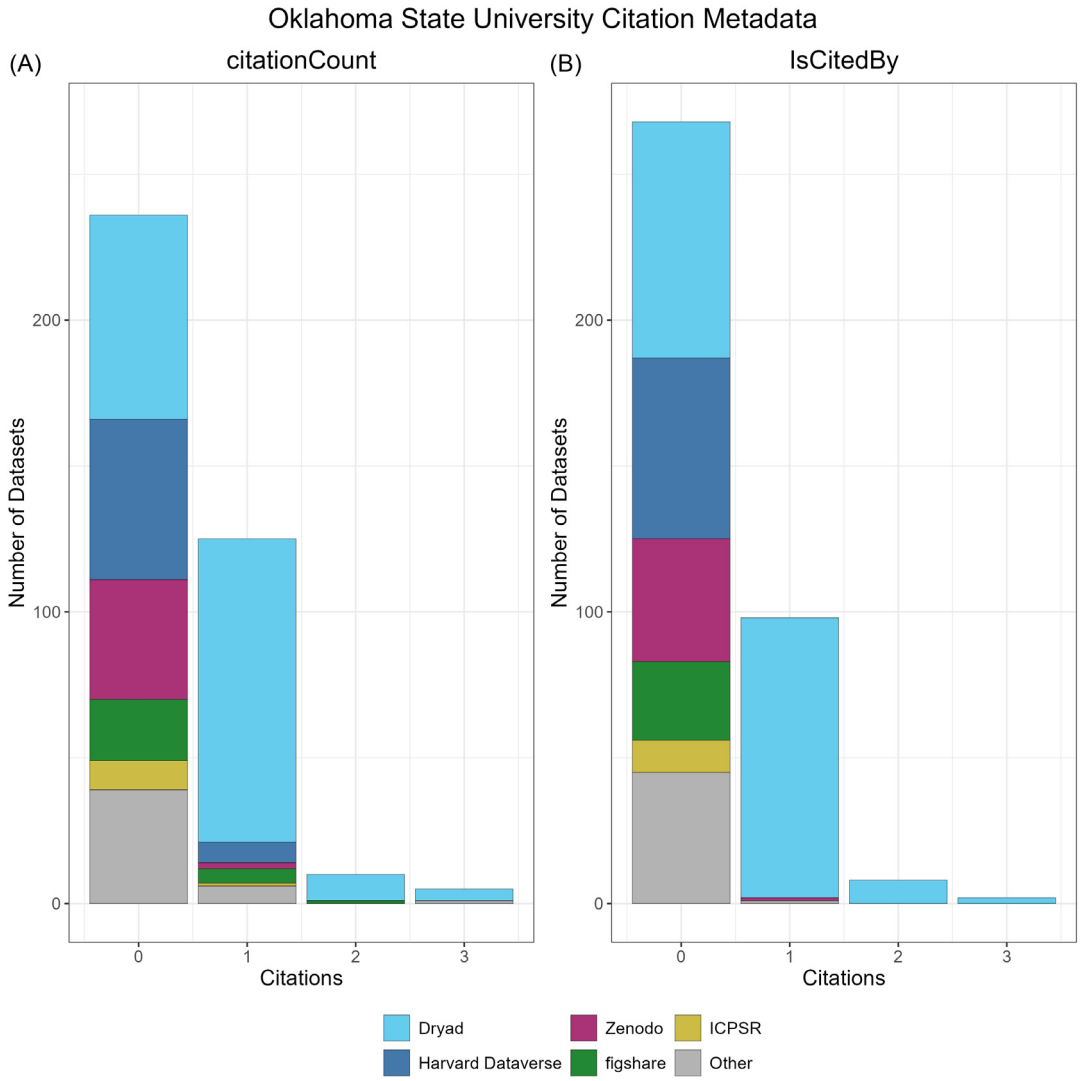

Figure 3. OSU dataset citations as reported by A) citationCount and B) IsCitedBy. Numerical summaries of graph contents are available in Supplemental Tables 1 and 2, respectively.

64% ($N = 152$) of these reported citations, with 11 other repositories making up the remaining 36%. The distribution of dataset citationCount values is shown in Figures 3A and 4A.

IsCitedBy

About 29% ($N = 108$) of OSU datasets and about 25% ($N = 155$) of UK datasets reported at least 1 citation via the IsCitedBy relationship, with a maximum citation count of 3 and 7, respectively. For both institutions, Dryad accounted for more than 95% ($N = 106$ for OSU;

$N = 150$ for UK) of these reported citations. Overall, only Dryad, ScienceBase (USGS), Zenodo, ICPSR, and SBGrid Data Bank provided "IsCitedBy" DOI relation types in spite of the fact that 15 total data repositories reported aggregate citation counts via citationCount. Repository representation and numbers of "IsCitedBy" citations are illustrated in Figures 3B and 4B.

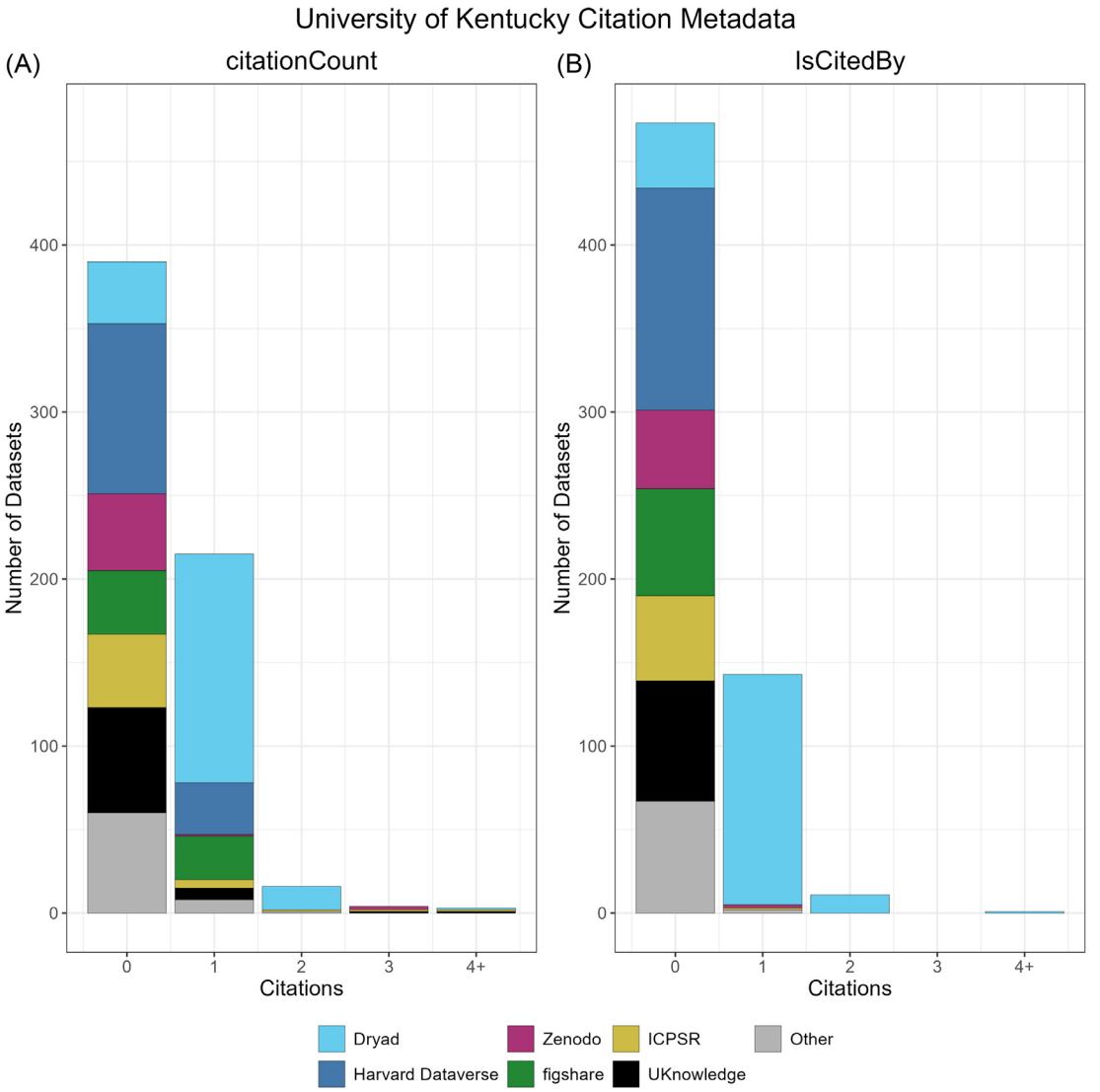

Figure 4. UK dataset citations as reported by A) citationCount and B) IsCitedBy. Top citation counts in panel A come from UKnowledge (5 citations), Dryad (10 citations), and ICPSR (16 citations), while the top IsCitedBy count in panel B comes from Dryad (7 citations). Numerical summaries of graph contents are available in Supplemental Tables 1 and 2, respectively.

citationCount vs. IsCitedBy

There were some discrepancies between the reported citation counts and the number of DOIs listed as citing the datasets (relationType = "IsCitedBy"). In theory, a dataset's total reported citation count should equal the sum of records with the "IsCitedBy" relationship. In many cases, however, datasets with reported citationCount values did not provide any metadata about which DOIs cited the dataset. In other cases, datasets reported both citationCount values and IsCitedBy relationships, but those values did not match. These comparisons between the metadata fields are reported in Table 3.

Repositories Not Reporting Citation Information

There were 11 repositories for OSU and 22 for UK that did not report any citation information for our institutions' datasets, from either the citationCount property or the IsCitedBy relationship.

Question 3: How is usage measured through alternative metrics of reuse, such as download and view counts?

The repositories that reported views and downloads include Dryad, figshare, Mendeley Data, Harvard Dataverse, and KNB Data Repository, with Dryad reporting views and downloads

Repository	Institution	−2	−1	0	1	2	3	5	16	Total
Dryad	OSU	1	4	161	20	1	–	–	–	187
Dryad	UK	–	2	179	7	–	1	–	–	189
Harvard Dataverse	OSU	–	–	55	7	–	–	–	–	62
Harvard Dataverse	UK	–	–	102	31	–	–	–	–	133
Zenodo	OSU	–	–	42	1	–	–	–	–	43
Zenodo	UK	–	1	46	–	–	2	–	–	49
figshare	OSU	–	–	21	5	1	–	–	–	27
figshare	UK	–	–	38	26	–	–	–	–	64
ICPSR	OSU	–	–	10	1	–	–	–	–	11
ICPSR	UK	–	–	45	4	1	1	–	1	52
UKnowledge	OSU	–	–	–	–	–	–	–	–	0
UKnowledge	UK	–	–	63	7	–	1	1	–	72

Table 3. Comparison between citation data retrieved from the "citationCount" property versus the "IsCitedBy" relationType property. A value of 0 indicates equal values for citationCount and IsCitedBy. Negative values indicate more IsCitedBy records and positive values indicate more citationCount records. Only the top 6 most common repositories were included in this table.

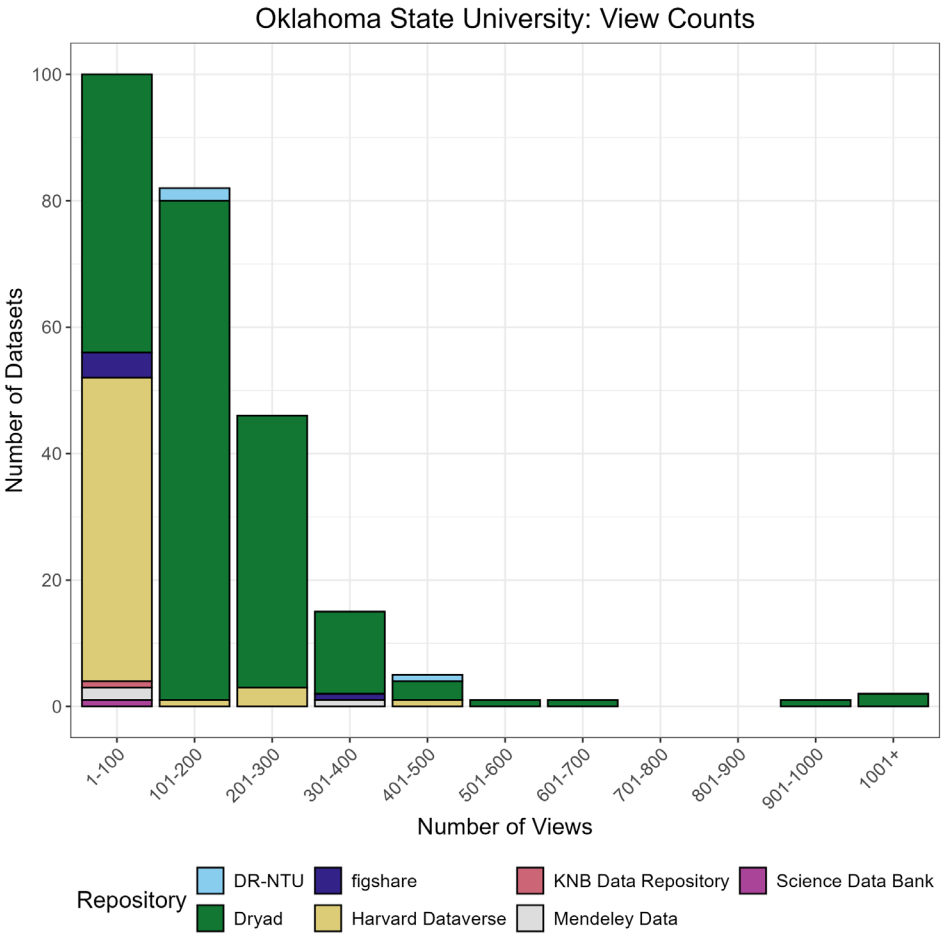

Figure 5. OSU datasets reporting at least one view. OSU had two extreme outliers with view counts of 3025 and 3261. Numerical summaries of graph contents are available in Supplemental Table 3.

most reliably. View and download counts for OSU (Figures 5 and 7) and UK (Figures 6 and 8) are visualized for the subset of datasets where these metrics are available, with the vast majority coming from Dryad for both institutions. Here, OSU and UK had contrasting distributions. While a plurality of datasets had downloads of less than 100 and views of less than 400, the distribution of UK datasets had a longer rightward tail, while OSU had two extreme rightward outliers with view and download counts several times greater than the average.

Question 4: What quality of metadata exists for these datasets?

We found it was rare that a dataset with multiple creators included a PID for each creator and their affiliated institution(s). Therefore, to determine whether a dataset had "complete"

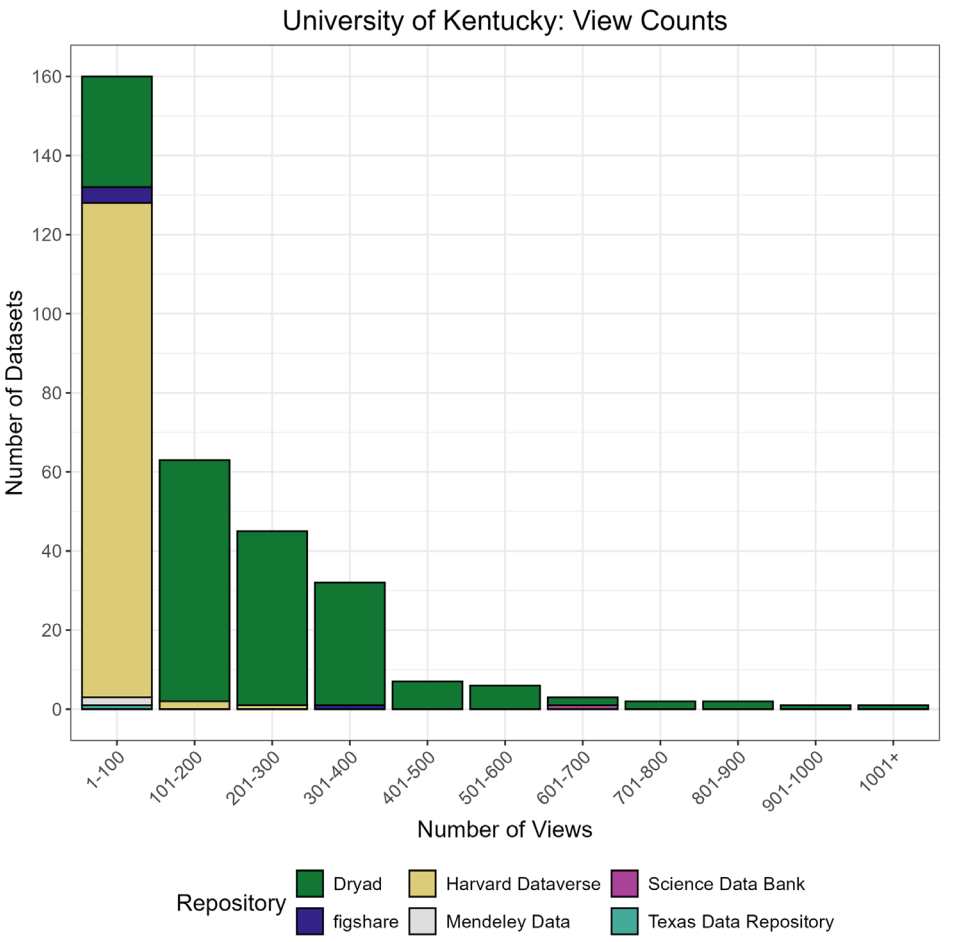

Figure 6. UK datasets reporting at least one view. Numerical summaries of graph contents are available in Supplemental Table 3.

metadata in a particular category, we checked if it had at least one named entity (creator, affiliation, or funder) with a PID attached.[9]

For both OSU and UK, author and affiliation PIDs were much more commonly included than funder PIDs (Table 4). Although alternate PIDs for individuals exist, in our case, the only one used for creators was an ORCID iD, which first launched in 2012 and is now the most popular PID for researchers. There was slightly more diversity in PIDs for institutions, which primarily used ROR IDs but occasionally used GRIDs

[9] Note that the complete entity did not need to belong to our own institutions. Indeed, because our API query looked only for the text string for our institution names, we had many datasets where our affiliated researcher had a name and affiliation listed, but no PIDs attached to them.

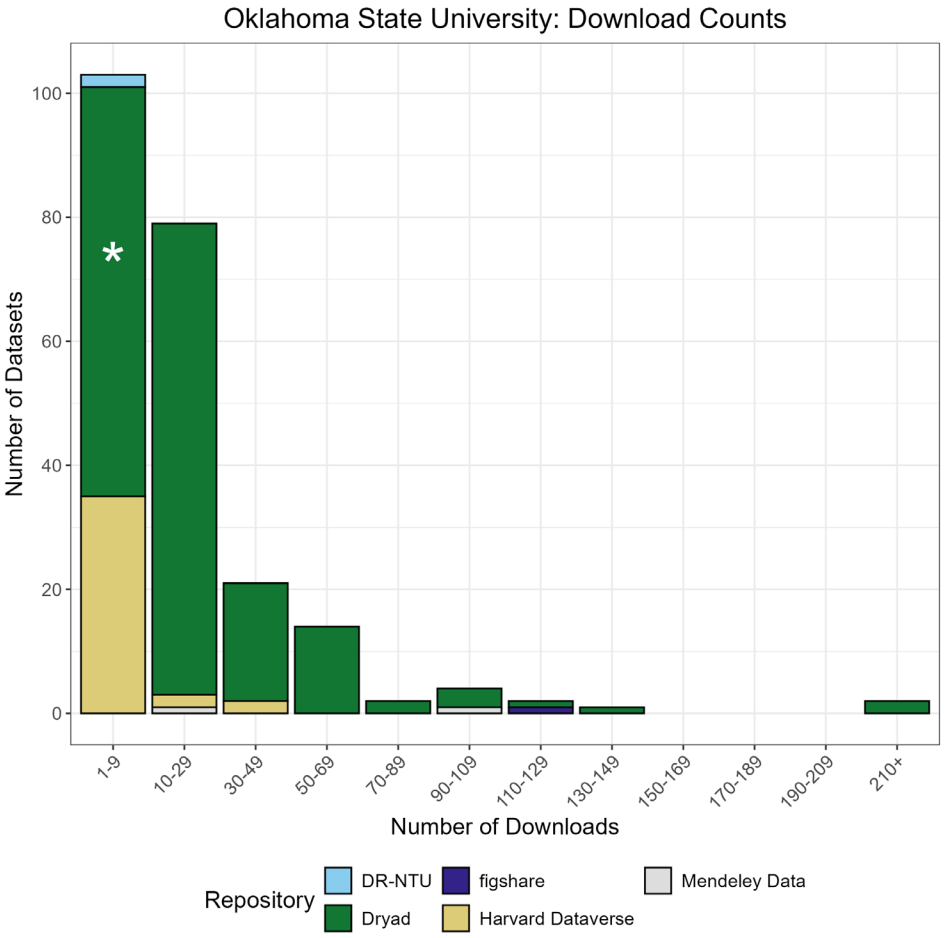

Figure 7. OSU datasets reporting at least one download. Asterisk (*) denotes a bin containing download counts from 1–9. All other bins have a range of 20, except for the final bin, which captures all download counts of 210 or greater. OSU had two extreme outliers with download counts of 753 and 1268. Numerical summaries of graph contents are available in Supplemental Table 4.

instead.[10] Funder PIDs were the least consistent, with a plurality that utilized ROR IDs, others with DOIs, and a small number with IDs from an unidentified schema.

Metadata completeness over time

Figures 9 and 10 show the number of datasets deposited by year in each repository and their metadata completeness in the three examined categories for OSU and UK, respectively. These charts demonstrate the individual metadata practices of specific repositories.

[10] UK-affiliated datasets also featured one institutional affiliation indicated with a record ID from the European Directory of Marine Organisations.

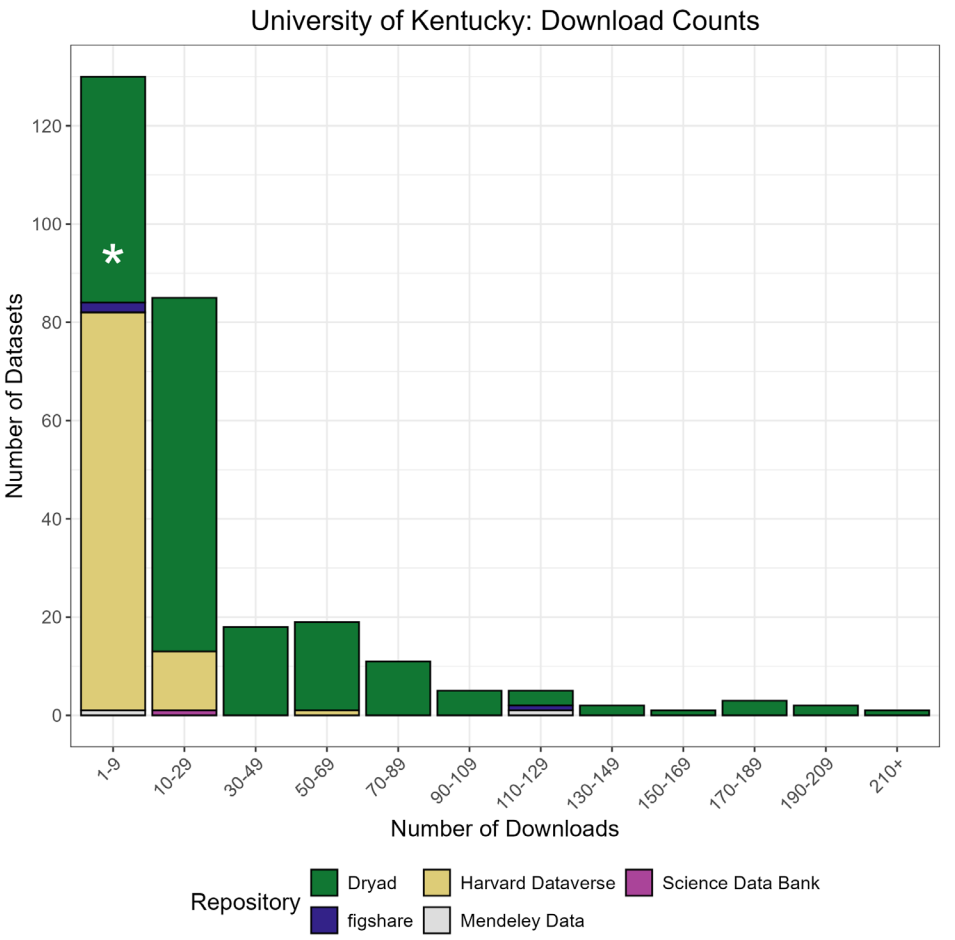

Figure 8. UK datasets reporting at least one download. Asterisk (*) denotes a bin containing download counts from 1–9. All other bins have a range of 20, except for the final bin, which captures all download counts of 210 or greater. Numerical summaries of graph contents are available in Supplemental Table 4.

Institution	Author	Affiliation	Funder	Author, Affiliation, & Funder
OSU ($N = 376$)	58.2%	65.4%	29.8%	26.0%
UK ($N = 628$)	47.6%	51.6%	19.1%	14.8%

Table 4. Percent of datasets that included at least one PID for different metadata categories.

Funder metadata completeness

Figures 11 and 12 provide additional detail for the levels of completeness of funder metadata by repository. In addition to a name and PID for a funder, the DataCite schema also allows for an award number.

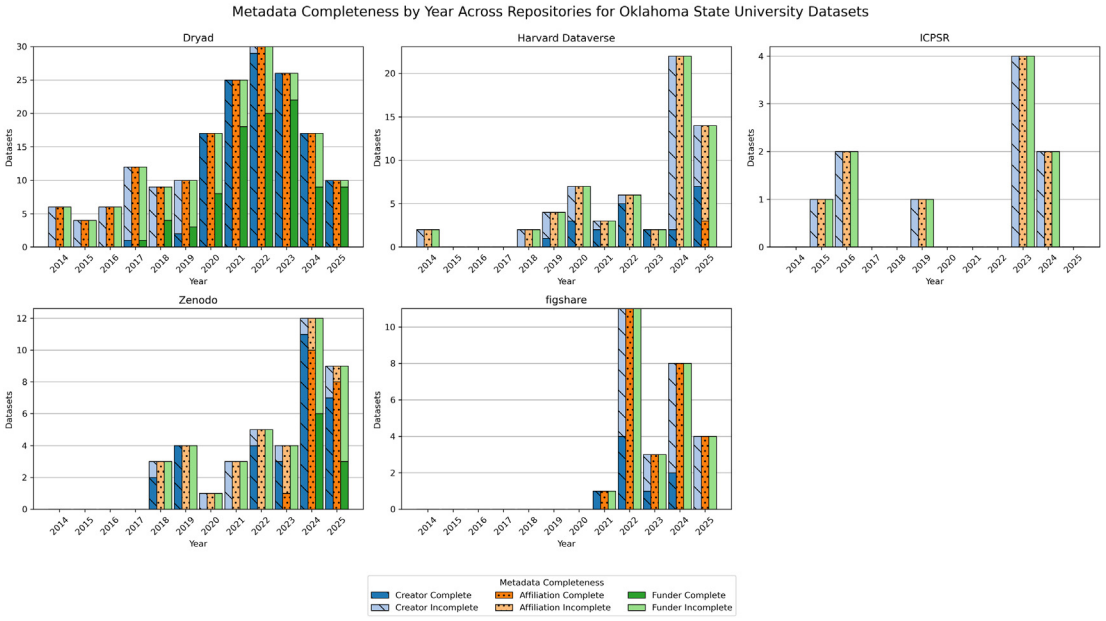

Figure 9. Metadata completeness for OSU datasets in the 5 most used repositories (excluding UKnowledge, which contained no OSU-affiliated datasets).

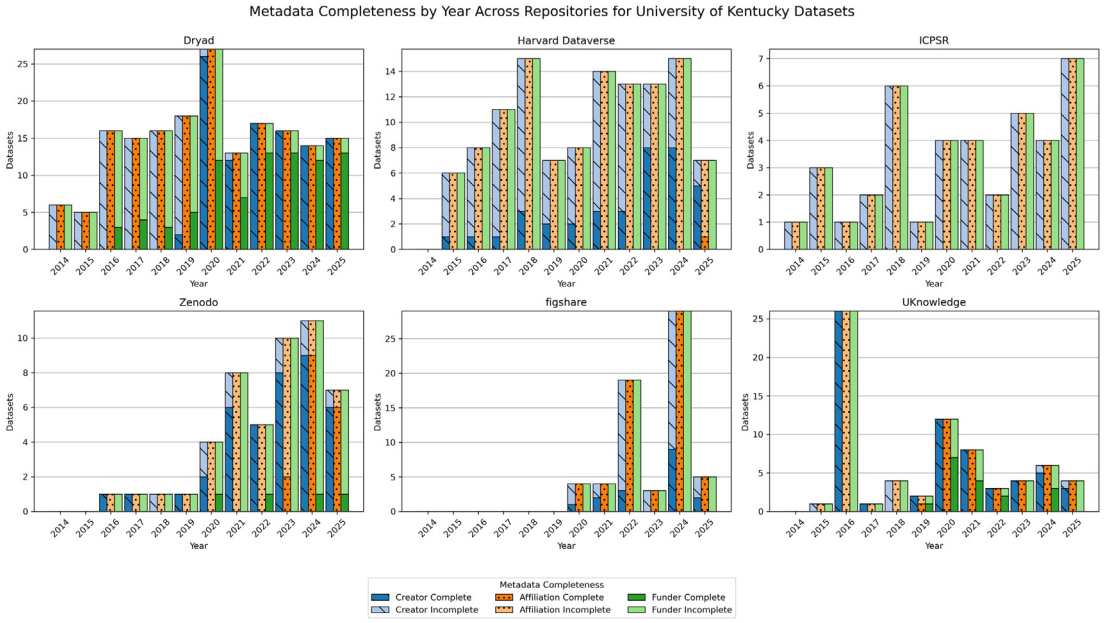

Figure 10. Metadata completeness for UK datasets in the 6 most used repositories.

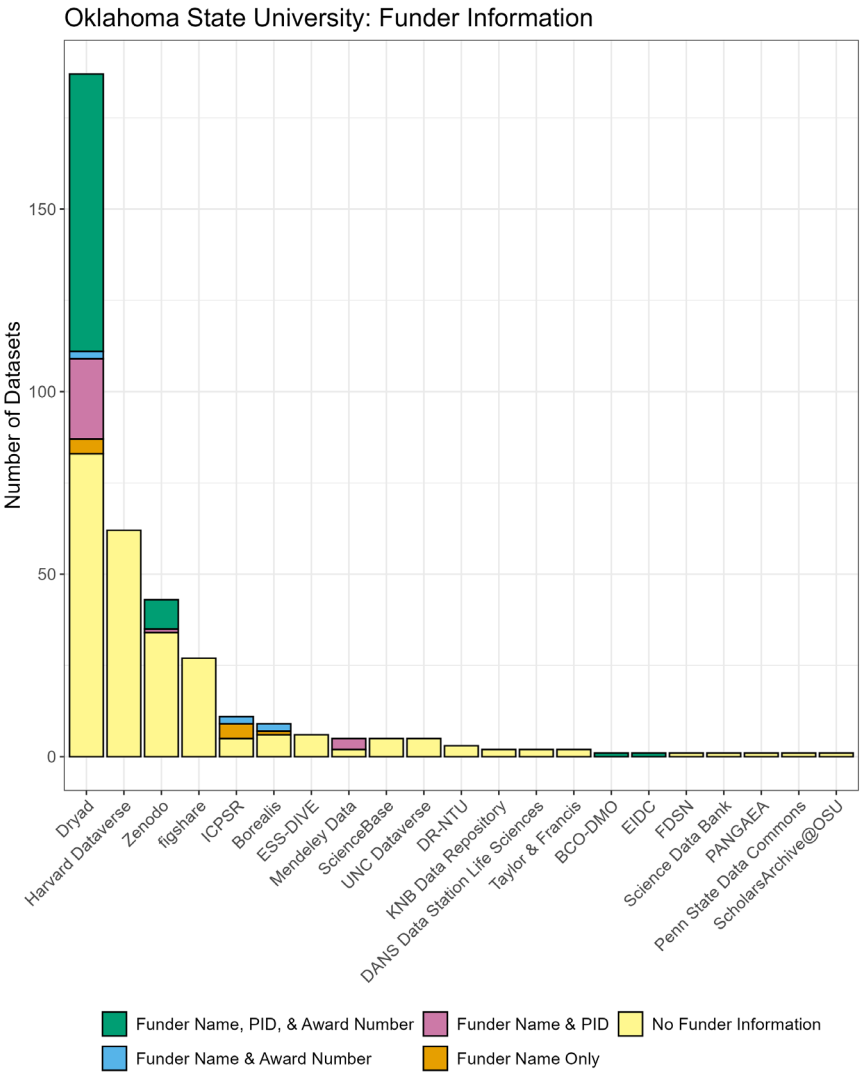

Figure 11. OSU funder metadata completeness. Among the repositories with only 1 institutional dataset, BCO-DMO and EIDC provided complete funder information, and all other repositories provided no funder data. Numerical summaries of graph contents are available in Supplemental Table 5.

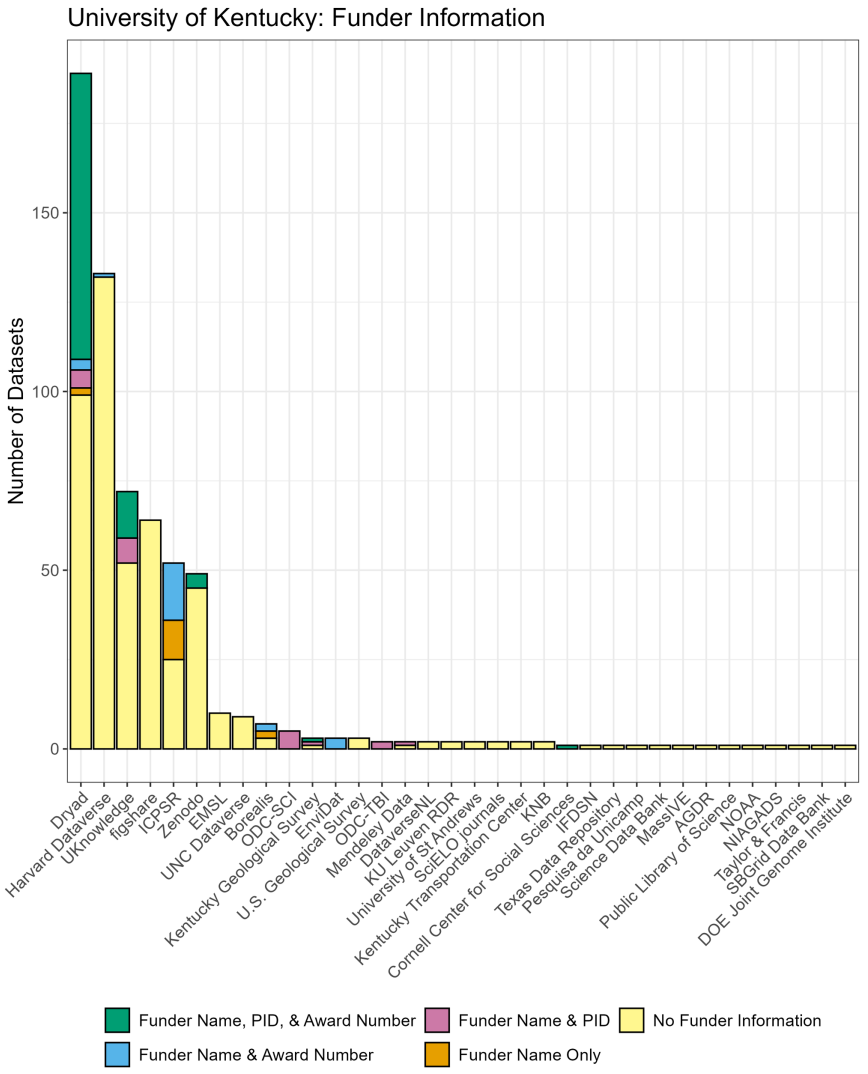

Figure 12. UK funder metadata completeness. Among the repositories with only 1 institutional dataset, the Cornell Center for Social Sciences provided complete funder information, and all other repositories provided no funder data. Numerical summaries of graph contents are available in Supplemental Table 5.

DISCUSSION

Institutional comparison

Overall, OSU and UK datasets followed roughly similar trends on the questions investigated here. Both were found primarily in the same set of commonly used repositories—Dryad, Dataverse, Zenodo, figshare, and ICPSR—with UK's institutional repository containing a notable proportion of UK datasets (Figure 2). Less than half of all datasets from both institutions reported citation information, with the citationCount property used more frequently than a discrete tracking of citations with the IsCitedBy relation (Table 3). The most notable difference appeared in counts of views and downloads (Figures 5–8), where the distributions were roughly similar for both institutions but featured two outliers for OSU.[11] Both datasets, which were supplemental to articles published in *PLOS Genetics* and *Scientific Data,* had view and download counts far greater than the normal range. While the reason for the relative popularity of these two datasets cannot be confirmed from metrics alone, they indicate the potential for datasets to function as first-class research outputs.

Metadata completeness

For both institutions, Dryad and figshare both had at least one complete creator affiliation for every deposited dataset (Figures 9 and 10). Meanwhile, datasets in ICPSR, a repository first founded in 1962 that predates all the PID schemas included here, frequently lacked persistent identifiers of any kind. The fact that the general trends for both UK and OSU match up across the five overlapping repositories suggests that these rates of completeness are in part due to the implementation of the DataCite schema by the individual repositories. For example, Dryad requires that users link their ORCID iD when creating an account before they can make a submission, which therefore ensures that every dataset has an ORCID iD listed for at least one author (Dryad, n.d.). Figure 10 also provided a useful perspective on UKnowledge, UK's institutional repository (IR). IRs represent a unique space for institutional stakeholders because they frequently play roles both in educating researchers on metadata and maintaining the technical infrastructure for creating robust records. For UKnowledge, the breakdown of repository completeness by year demonstrated strong implementation of ORCID iDs for researchers but also highlighted the potential to update records to include ROR IDs.

Additionally, Figures 9 and 10 also reveal changes over time in the use of PIDs across the repository ecosystem as a whole. Most notable is the gradual increase in inclusion of PIDs

[11] 10.5061/dryad.th092 and 10.5061/dryad.723m1

for individuals (ORCID iDs), and a more uneven increase in funder PIDs. It is also noteworthy that repositories have updated records to supplement them with persistent identifiers (for example, adding ROR IDs to datasets that predated the ROR schema), representing an active stewardship of datasets past their initial creation.

While it was most common for no funder information to be included at all (59% of OSU datasets and 61% of UK datasets), Dryad—which was most likely to provide funder metadata—often recorded a funder name, PID, and award number. Zenodo and UKnowledge also included all three of these components in a proportion of their datasets (Figures 11 and 12). Robust funder metadata is a particularly complex challenge, given that it requires coordination between research teams that must provide the award number, funders that must communicate a preferred PID, and repositories that must implement the technical infrastructure to include this field in their records.

Metadata schema and API query limitations

As noted in the Methods, the text string API query only searched for a given institution name in the creator affiliation and not the contributor affiliation, while the ROR ID API query searched in both the creator and contributor fields. We tested a separate API query based on contributors' affiliation, but did not surface many additional datasets from either of our institutions.[12] Furthermore, we found that in cases where a contributor was not also an author, their role on the project did not firmly suggest authorship (e.g., data collector, contact person). We believe these discrepancies are more of a reflection of the lack of a commonly agreed-upon implementation of the creator (author) and contributor labels among individual researchers and repositories, rather than a problem with the metadata schema itself.

Regarding data cleaning and identification of name variations for authors, repositories, and funders, we acknowledge that our manual identification of variant names for a single entity introduces room for error, especially for author names. We erred on the side of caution, marking author names as variations of one another only in clear cases (such as alternate versions of a name with or without a middle initial). The difficulty in disambiguating authors, especially those with common names, underscores the importance of ORCID iDs as a unique identifier for grouping outputs by the same creator.

[12] Because the API query using an institution's ROR ID searches for the ROR in all fields, our ROR results included a small number (under 10 for both institutions) of datasets that were not included in the response to the institution name text query.

A further point of complication is the potential modifications and loss of metadata content as records are transferred from the host repository to DataCite. Although many of these repositories utilize the DataCite schema, repositories may also customize their metadata schema, and these customizations do not always map cleanly onto the DataCite schema. This is especially relevant for usage metrics, such as downloads and views. Some repositories, such as Zenodo and ICPSR, measure usage but do not report these metrics to DataCite, which attempts to standardize usage metrics by following the COUNTER Code of Practice (DataCite, n.d.h). Since 2024, DataCite has participated in the Make Data Count Initiative, an effort to bring standardization to disparate methods of measuring dataset usage (Puebla & Chodacki, 2024).

Additionally, not all citation-related information can be perfectly converted to the DataCite metadata properties. For example, Dryad metadata includes the relatedWorkIdentifier and relatedWorkRelationship properties, which are similar to the relatedIdentifier, relationType, and resourceTypeGeneral properties from the DataCite metadata schema (Dryad, 2025; DataCite, n.d.f). However, Dryad includes relationships such as "primary_article" in addition to the more generic "article," which serves to distinguish the original citing article for a dataset from subsequent reuse cases. This distinction is lost upon incorporation of metadata from Dryad DOIs into the DataCite database. Although not part of the workflow presented here, individuals who want to maximize the utility and accuracy of their dataset metadata may consider querying repository APIs as well to retrieve unmodified metadata, following existing workflows that incorporate APIs from Zenodo, Dataverse, and Dryad in addition to DataCite (Gee, 2025; markuslibrary, 2025).

Missing, repeated, and incorrectly included datasets

Notably absent from our list of research questions is the deceptively simple question, "How many datasets have researchers at our institutions produced?" While this paper has presented a process for collecting metadata records for datasets created by researchers affiliated with a given institution, the resulting records do not represent a complete catalog of all datasets produced by an institution's researchers. Most crucially, the pipeline only queries DataCite and therefore does not identify datasets whose DOI records were created by other DOI registration agencies (DOI Foundation, n.d.), or datasets registered with a different persistent identifier, such as an Archival Resource Key (ARK). A very rough estimate of our coverage can be gleaned from Elsevier's former dataset aggregation service Data Monitor, which reportedly drew 70% of all its dataset records from DataCite (Pure Help Center, n.d.). As members of U.S.-based institutions with heavy use of generalist repositories that utilize DataCite, we found the DataCite API an adequate focus for our research, but this may not be the case for institutions in other contexts.

Additionally, the initial query of the DataCite API is limited to records with a "resource-type-id" of "dataset." Research data can also include other outputs such as images, videos, audio recordings, and other file types that may not be classified as the "dataset" resource type in repositories, and those outputs would be overlooked by our query. Most notably, this meant our query did not return any results from the Open Science Framework (OSF) repository, which classified all resources as "text" by default, unless researchers manually change the resource metadata. Furthermore, institutional affiliations are currently only available to OSF institutional members (Center for Open Science, 2023), so our query would also be incapable of uncovering OSF datasets from nonmember institutions, even if the resource were properly classified. A manual search for our institution names on OSF, conducted at the same time as the DataCite API queries, did not yield any research datasets.

Regarding deduplication, we are aware that our automated methods of deduplication do not catch every alternate version of datasets. Most notably, some researchers appear to have published updated versions of their datasets as though they were entirely new entities, but the metadata is nearly identical, other than the DOI. For example, the figshare DOIs 10.6084/m9.figshare.21370941.v1, 10.6084/m9.figshare.21408750.v1, and 10.6084/m9.figshare.21411399.v1 all appear to reference different versions of the same dataset, yet their metadata records contain no indication that they are linked. These duplications are best detected by manually examining the dataset, which is not feasible at scale. For consistency and to reflect the real-world data sharing practices of researchers, we chose not to attempt to remove this category of duplicated datasets.

Finally, our workflow does not address cases where an institutional affiliation was listed for a researcher by mistake. We identified several cases for datasets published via figshare where some researchers had eight or more affiliations listed, including OSU or UK. Manual investigations confirmed that those researchers were not and never had been affiliated with our institutions, although researchers with the same name were, suggesting that these affiliations may have been automatically included based on name matching. Unfortunately, when these incorrect affiliations are included by either a researcher or a repository, there is no easy method for automatically identifying and removing these datasets.

Implications for scholarly communication services

This investigation demonstrates the importance of PIDs in accurately linking researchers, institutions, and scholarly resources. As the initial API queries for ROR IDs and text strings show, workflows that rely exclusively on ROR IDs have a higher degree of certainty that they are only identifying works with a genuine link to an institution, but they may be missing hundreds of works where a ROR ID is absent. Librarians can leverage their

connections to researchers and data repository providers to address this challenge from two angles.

First, they can provide researchers and administrators with education on the importance of PIDs as part of research output metadata. This engagement could take many forms—for example, the initiation of campus policies or priorities on the use of ORCID iDs; communication about reporting PIDs (including institutional and funder ROR IDs) in article and data submissions; or the creation of research data impact reporting through similar summaries as those provided in this article. By intervening with researchers before data submission, they can improve the quality of metadata researchers provide to data repositories.

Additionally, by working with service providers, librarians can advocate for the inclusion of PIDs in the dataset deposit process to make sure that necessary information is included by default. Ideally, this will simplify and standardize the metadata submission process for researchers and improve metadata completeness overall. Those who steward research data, such as institutional repository managers, may also have options for updating metadata for existing records in bulk (for example, applying ROR IDs).

While the specific distribution of where datasets are shared will certainly vary for other institutions, stakeholders may apply the workflow described here using the provided R or Python scripts to understand where researchers at their own institutions share data. That information may drive spending decisions; for example, an institution may consider a partnership with a generalist repository frequently used by its researchers, or researchers who have published in a specific repository may be contacted to provide feedback to help determine whether the institution should initiate a repository membership. Alternatively, it may help librarians better understand where researchers are publishing their data so they can familiarize themselves with those platforms and provide targeted support. Some institutions may be interested in combining the dataset metadata—which includes author names—with internal data such as departmental affiliations to assess disciplinary and departmental trends in data sharing and repository use.

Finally, librarians involved with institutional research impact evaluation should bring a note of caution to discussions that claim to evaluate an institution's full research data output. As shown here, the variety of infrastructures in place makes it functionally impossible to identify every dataset produced by an institution, and results will be highly sensitive to seemingly small choices. Creating a robust report on shared data is possible, but human labor is necessary to bring datasets together from disparate sources, identify duplicate or erroneously included datasets, and enhance records.

Future directions

Our focus was on metadata related to authors, affiliation, and funding, which are likely the most important to individuals working in libraries, research administration, and other positions that evaluate research impact and outputs. Our assessment does not evaluate other metadata components, which might be more relevant to potential data reusers. For example, other work has explored limitations in metadata beyond the properties represented in the DataCite schema and database, such as metadata related to demographics of a study population or data processing methods (Huang et al., 2025). These types of metadata elements are likely better implemented at the repository level, since metadata standards and needs can vary greatly between disciplines and projects. We encourage further work in this area, as metadata quality and completeness can benefit many stakeholders in the research ecosystem.

Finding ways to convey the impact of PIDs for individuals, affiliations, and funders may help motivate more researchers to include this information in their research outputs. The DataCite GraphQL API also provides a means for acquiring data to visualize connections between research outputs for individual researchers, organizations, and funders (Cousijn et al., 2021; DataCite, n.d.a). Other recent work has explored the use of PIDs to visualize publication collaborations within and between institutions (Aghassibake et al., 2023). It is important to emphasize that this approach to visualizing research outputs may be ineffective or misleading for disciplines or individuals whose research outputs are not currently represented in the PID landscape.

Another notable area for improvement is uptake in the tracking and reporting of citations, which will require increased participation by data repositories—who must push this information to DataCite—as well as researchers, who need to be sure to include DOIs when citing datasets (including their own) and reporting their own related work(s) as part of dataset metadata.[13] But repositories automatically and reliably linking datasets to articles that cite them is a necessary infrastructure piece to encourage data sharing. Surveys of researchers on their willingness to share their data have consistently shown majorities express concern that they be properly cited when others reuse their data (Schmidt et al., 2016; Tenopir et al., 2015, 2020). If more repositories can succeed in accurately and smoothly leveraging DataCite

[13] However, we caution against relying too heavily on quantitative metrics such as citation counts alone as a means of assessing the impact of individual researchers' work. There has been ample work already proposing and evaluating various approaches to assessing the impact of research outputs. For example, the Leiden Manifesto proposes ten principles for "metrics-based research assessment" and cautions against relying too heavily or blindly on quantitative assessments (Hicks et al., 2015). A commissioned report from several years later suggested a variety of ways to assess research data impact specifically, including the use of altmetrics (Konkiel, 2020).

relationships, such as IsCitedBy and metrics such as citation counts, they will play a key role in reducing this barrier to data sharing.

CONCLUSION

The metadata collected for both institutions indicates the need for greater coordination and education regarding metadata content. Variability in the use of persistent identifiers such as ORCID iDs for individuals and ROR IDs for institutions hampers proper linking of datasets to their creators. We found Dryad to be a laudable example of optional metadata content, specifically relating to the reporting of usage metrics and the inclusion of creator and affiliation PIDs, which other repositories may wish to adopt. Research stakeholders could greatly benefit from the ability to leverage this optional metadata in research impact reporting at both an individual and institutional level, but there needs to be greater investment, coordination, and education for all stakeholders on providing complete and accurate metadata for research datasets.

DATA AVAILABILITY STATEMENT

The code (available for R and Python), JSON files from the initial API queries, cleaned datasets, and name standardization files are openly available in Zenodo at https://doi.org/10.5281/zenodo.20025492.

REFERENCES

Adamich, T. (2016). Accessing scholarly research datasets using DataCite, ORCID, and CASRAI. *Technicalities*, *36*(3), 18–21.

Aghassibake, N., Castello, O. G., Gujilde, P., & Rabun, S. (2023). Visualizing institutional activity using persistent identifier metadata. *Information Services and Use*, *43*(3–4), 335–342. https://doi.org/10.3233/ISU-230218

Briney, K. A. (2024). Measuring data rot: An analysis of the continued availability of shared data from a Single University. *PLOS ONE*, *19*(6), e0304781. https://doi.org/10.1371/journal.pone.0304781

Burton, K., Cocks, C., & Russell, B. (2023). Recognizing and harnessing the transformational power of persistent identifiers (PIDs) for publicly-engaged scholars. *Information Services and Use*, *43*(3-4). https://doi.org/10.3233/ISU-230212.

Center for Open Science. (2023, May 5). Affiliate an OSF project with an institution. https://help.osf.io/article/192-affiliate-an-osf-project-with-an-institution

Cousijn, H., Braukmann, R., Fenner, M., Ferguson, C., Horik, R. van, Lammey, R., Meadows, A., & Lambert, S. (2021). Connected research: The potential of the PID graph. *Patterns*, *2*(1). https://doi.org/10.1016/j.patter.2020.100180

DataCite. (n.d.a). *DataCite GraphQL API guide*. Retrieved December 3, 2025, from https://support.datacite.org/docs/datacite-graphql-api-guide

DataCite. (n.d.b). *DataCite usage reports API guide*. Retrieved December 3, 2025, from https://support.datacite.org/docs/views-and-downloads

DataCite. (n.d.c). *Introduction to DataCite Commons*. Retrieved January 7, 2026, from https://support.datacite.org/docs/datacite-commons

DataCite. (n.d.d). *Introduction to the DataCite REST API*. Retrieved December 3, 2025, from https://support.datacite.org/docs/api

DataCite. (n.d.e). *Organizations in DataCite Commons*. Retrieved January 7, 2026, from https://support.datacite.org/docs/datacite-commons

DataCite. (n.d.f). RelatedIdentifier. Retrieved December 3, 2025, from https://datacite-metadata-schema.readthedocs.io/en/4.5/properties/relatedidentifier/

DataCite. (n.d.g). *Return a list of DOIs*. Retrieved December 3, 2025, from https://support.datacite.org/reference/get_dois

DataCite. (n.d.h). *Usage stats in DataCite Commons.* Retrieved December 12, 2025, from https://support.datacite.org/docs/usage-stats

DataCite Metadata Working Group. (2014). *DataCite metadata schema documentation for the publication and citation of research data* (No. Version 3.1). http://doi.org/10.5438/0010

DataCite Metadata Working Group. (2024). *DataCite metadata schema for the publication and citation of research data and other research outputs* (No. Version 4.6). https://doi.org/10.14454/mzv1-5b55

DOI Foundation. (n.d.). *Existing registration agencies.* Retrieved December 3, 2025, from https://www.doi.org/the-community/existing-registration-agencies/

Dryad. (n.d.). *Creating a Dryad account*. Retrieved December 3, 2025, from https://datadryad.org/help/account/login

Dryad. (2025). *Dataset Metadata*. GitHub. https://github.com/datadryad/dryad-app/blob/main/documentation/apis/search.md

Federer, L. M. (2022). Long-term availability of data associated with articles in PLOS ONE. *PLOS ONE*, *17*(8), e0272845. https://doi.org/10.1371/journal.pone.0272845

Fernández-Marcial, V., González-Solar, L., & Vale, A. (2023). Is ORCID your ID? A case study at the Faculty of Arts and Humanities of the University of Porto. *Learned Publishing*, *36*(4), 564–576. (172875770). https://doi.org/10.1002/leap.1562

Gee, B. (2025, November 25). *Scripted process for retrieving metadata on institutional-affiliated research dataset publications*. GitHub. https://github.com/utlibraries/research-data-discovery

Gould, M. (2022). People, places, and things: Persistent identifiers in the scholarly communication landscape. *College & Research Libraries News*, *83*(9), 398–402. (159535973). https://doi.org/10.5860/crln.83.9.398

Hazarika, R., & Sudhier, K. G. (2025). Mapping open access publishing trends in central universities of North East India: A scientometric approach. *Science & Technology Libraries*, *44*(3), 247–265. https://doi.org/10.1080/0194262X.2024.2416952

Heusse, M., & Cabanac, G. (2022). ORCID growth and field-wise dynamics of adoption: A case study of the Toulouse scientific area. *Learned Publishing*, *35*(4), 454–466. (159630708). https://doi.org/10.1002/leap.1451

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429–431. https://doi.org/10.1038/520429a

Huang, Y.-N., Munteanu, V., Love, M. I., Ronkowski, C. F., Deshpande, D., Wong-Beringer, A., Corbett-Detig, R., Dimian, M., Moore, J. H., Garmire, L. X., Reddy, T. B. K., Butte, A. J., Robinson, M. D., Eskin, E., Abedalthagafi, M. S., & Mangul, S. (2025). Perceptual and technical barriers in sharing and formatting metadata accompanying omics studies. *Cell Genomics*, *5*(5). https://doi.org/10.1016/j.xgen.2025.100845

Johnston, L. R., Mohr, A. H., Herndon, J., Taylor, S., Carlson, J. R., Ge, L., Moore, J., Petters, J., Kozlowski, W., & Vitale, C. H. (2024). Seek and you may (not) find: A multi-institutional analysis of where research data are shared. *PLOS ONE*, *19*(4), e0302426. https://doi.org/10.1371/journal.pone.0302426

Konkiel, S. (2020). *Assessing the Impact and Quality of Research Data Using Altmetrics and Other Indicators | Scholarly Assessment Reports*. https://doi.org/10.29024/sar.13

Lamb, I., & Larson, C. (2016). Shining a light on scientific data: Building a data catalog to foster data sharing and reuse. *The Code4Lib Journal*, *32*. https://journal.code4lib.org/articles/11421

Madden, F., & Bunakov, V. (2019). Using persistent identifiers to track PhD outcomes. *Cadernos de Biblioteconomia, Arquivística e Documentação*, (1), 53–60. (141640894)

Malik, S., & Mandal, S. (2024). Research organization registry and its impact on Indian Institutes of Technology: A data perspective. *World Digital Libraries*, *17*(2), 101–116. https://doi.org/10.18329/09757597/2024/17208

Mannheimer, S., Clark, J. A., Hagerman, K., Schultz, J., & Espeland, J. (2021). Dataset Search: A lightweight, community-built tool to support research data discovery. *Journal of eScience Librarianship*, *10*(1). https://doi.org/10.7191/jeslib.2021.1189

markuslibrary. (2025, November 10). *UAB Dataset Catalog*. GitHub. https://github.com/markuslibrary/uab-dataset-catalog

Mayernik, M., Johnson, A., Julian, R., Murray, M., Mundoma, C., Ranganath, A., & Stossmeister, G. (2024). Persistent identifiers for instruments and facilities: Current state, challenges, and opportunities. *Journal of eScience Librarianship*, *13*(3). https://doi.org/10.7191/jeslib.964

National Institutes of Health. (2023). *Final NIH Policy for Data Management and Sharing*. (NOT-OD-21-013). https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

Puebla, I., & Chodacki, J. (2024). Make data count: Driving metrics for the meaningful evaluation of data. *Zenodo*. https://doi.org/10.5281/zenodo.14261211

Pure Help Center. (n.d.). *Datacite as import source (vision)*. Retrieved December 3, 2025, from https://web.archive.org/web/20251009073945/https://helpcenter.pure.elsevier.com/data-sources-and-integrations/datacite-as-import-source-vision

Schmidt, B., Gemeinholzer, B., & Treloar, A. (2016). Open data in global environmental research: The Belmont Forum's open data survey. *PLOS ONE*, *11*(1), e0146695. https://doi.org/10.1371/journal.pone.0146695

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLOS ONE*, *10*(8), e0134826. https://doi.org/10.1371/journal.pone.0134826

Tenopir, C., Rice, N. M., Allard, S., Baird, L., Borycz, J., Christian, L., Grant, B., Olendorf, R., & Sandusky, R. J. (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLOS ONE*, *15*(3), e0229003. https://doi.org/10.1371/journal.pone.0229003

Wettere, N. V. (2021). Affiliation information in DataCite dataset metadata: A Flemish case study. *Data Science Journal*, *20*(1). https://doi.org/10.5334/dsj-2021-013

White House Office of Science and Technology Policy. (2022, August 25). *Ensuring free immediate & equitable access to federally funded research*. Executive Office of the President. https://bidenwhitehouse.archives.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Zenodo. (n.d.). *DOI versioning*. Retrieved December 3, 2025, from https://zenodo.org/help/versioning

SUPPLEMENTAL TABLES

Repository	Institution	0 Citations	1 Citation	2 Citations	3 Citations	4+ Citations	Total
Dryad	OSU	70	104	9	4	NA	187
Harvard Dataverse	OSU	55	7	NA	NA	NA	62
Zenodo	OSU	41	2	NA	NA	NA	43
figshare	OSU	21	5	1	NA	NA	27
ICPSR	OSU	10	1	NA	NA	NA	11
UKnowledge	OSU	NA	NA	NA	NA	NA	0
Other	OSU	39	6	NA	1	NA	46
Dryad	UK	37	137	14	NA	1	189
Harvard Dataverse	UK	102	31	NA	NA	NA	133
Zenodo	UK	46	1	NA	2	NA	49
figshare	UK	38	26	NA	NA	NA	64
ICPSR	UK	44	5	1	1	1	52
UKnowledge	UK	63	7	NA	1	1	72
Other	UK	60	8	1	NA	NA	69

Supplemental Table 1

Repository	Institution	0 Citations	1 Citation	2 Citations	3 Citations	4+ Citations	Total
Dryad	OSU	81	96	8	2	NA	187
Harvard Dataverse	OSU	62	NA	NA	NA	NA	62
Zenodo	OSU	42	1	NA	NA	NA	43
figshare	OSU	27	NA	NA	NA	NA	27
ICPSR	OSU	11	NA	NA	NA	NA	11
Uknowledge	OSU	NA	NA	NA	NA	NA	0
Other	OSU	45	1	NA	NA	NA	46
Dryad	UK	39	138	11	NA	1	189
Harvard Dataverse	UK	133	NA	NA	NA	NA	133
Zenodo	UK	47	2	NA	NA	NA	49
figshare	UK	64	NA	NA	NA	NA	64
ICPSR	UK	51	1	NA	NA	NA	52
UKnowledge	UK	72	NA	NA	NA	NA	72
Other	UK	67	2	NA	NA	NA	69

Supplemental Table 2

Repository	Institution	1-100	101-200	201-300	301-400	401-500	501-600	601-700	701-800	801-900	901-1000	1001+	Total
DR-NTU	OSU	NA	2	NA	NA	1	NA	NA	NA	NA	NA	NA	3
Dryad	OSU	44	79	43	13	3	1	1	NA	NA	1	2	187
Harvard Dataverse	OSU	48	1	3	NA	1	NA	NA	NA	NA	NA	NA	53
KNB Data Repository	OSU	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1
Mendeley Data	OSU	2	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	3
Science Data Bank	OSU	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1
figshare	OSU	4	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	5
Dryad	UK	28	61	44	31	7	6	2	2	2	1	1	185
Harvard Dataverse	UK	125	2	1	NA	NA	NA	NA	NA	NA	NA	NA	128
Mendeley Data	UK	2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2
Science Data Bank	UK	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	1
Texas Data Repository	UK	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1
figshare	UK	4	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	5

Supplemental Table 3

Repository	Institution	9-Jan	29-Oct	30-49	50-69	70-89	90-109	110-129	130-149	150-169	170-189	190-209	210+	Total
DR-NTU	OSU	2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2
Dryad	OSU	66	76	19	14	2	3	1	1	NA	NA	NA	2	184
Harvard Dataverse	OSU	35	2	2	NA	NA	NA	NA	NA	NA	NA	NA	NA	39
Mendeley Data	OSU	NA	1	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	2
figshare	OSU	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	1
Dryad	UK	46	72	18	18	11	5	3	2	1	3	2	1	182
Harvard Dataverse	UK	81	12	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	94
Mendeley Data	UK	1	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	2
Science Data Bank	UK	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1
figshare	UK	2	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	3

Supplemental Table 4

Repository	Institution	No Funder Information	Funder Name Only	Funder Name & Award Number	Funder Name & PID	Funder Name, PID, & Award Number	Total
BCO-DMO	OSU	NA	NA	NA	NA	1	1
Borealis	OSU	6	1	2	NA	NA	9
DANS Data Station Life Sciences	OSU	2	NA	NA	NA	NA	2
DR-NTU	OSU	3	NA	NA	NA	NA	3
Dryad	OSU	83	4	2	22	76	187
EIDC	OSU	NA	NA	NA	NA	1	1
ESS-DIVE	OSU	6	NA	NA	NA	NA	6
FDSN	OSU	1	NA	NA	NA	NA	1
Harvard Dataverse	OSU	62	NA	NA	NA	NA	62
ICPSR	OSU	5	4	2	NA	NA	11
KNB Data Repository	OSU	2	NA	NA	NA	NA	2
Mendeley Data	OSU	2	NA	NA	3	NA	5
PANGAEA	OSU	1	NA	NA	NA	NA	1
Penn State Data Commons	OSU	1	NA	NA	NA	NA	1
ScholarsArchive@OSU	OSU	1	NA	NA	NA	NA	1
Science Data Bank	OSU	1	NA	NA	NA	NA	1
ScienceBase	OSU	5	NA	NA	NA	NA	5
Taylor & Francis	OSU	2	NA	NA	NA	NA	2
UNC Dataverse	OSU	5	NA	NA	NA	NA	5
Zenodo	OSU	34	NA	NA	1	8	43
figshare	OSU	27	NA	NA	NA	NA	27
AGDR	UK	1	NA	NA	NA	NA	1
Borealis	UK	3	2	2	NA	NA	7
Cornell Center for Social Sciences	UK	NA	NA	NA	NA	1	1
DOE Joint Genome Institute	UK	1	NA	NA	NA	NA	1
DataverseNL	UK	2	NA	NA	NA	NA	2
Dryad	UK	99	2	3	5	80	189
EMSL	UK	10	NA	NA	NA	NA	10
EnviDat	UK	NA	NA	3	NA	NA	3

Supplemental Table 5. (Table continues on following page)

Harvard Dataverse	UK	132	NA	1	NA	NA	133
ICPSR	UK	25	11	16	NA	NA	52
IFDSN	UK	1	NA	NA	NA	NA	1
KNB	UK	2	NA	NA	NA	NA	2
KU Leuven RDR	UK	2	NA	NA	NA	NA	2
Kentucky Geological Survey	UK	1	NA	NA	1	1	3
Kentucky Transportation Center	UK	2	NA	NA	NA	NA	2
MassIVE	UK	1	NA	NA	NA	NA	1
Mendeley Data	UK	1	NA	NA	1	NA	2
NIAGADS	UK	1	NA	NA	NA	NA	1
NOAA	UK	1	NA	NA	NA	NA	1
ODC-SCI	UK	NA	NA	NA	5	NA	5
ODC-TBI	UK	NA	NA	NA	2	NA	2
Pesquisa da Unicamp	UK	1	NA	NA	NA	NA	1
Public Library of Science	UK	1	NA	NA	NA	NA	1
SBGrid Data Bank	UK	1	NA	NA	NA	NA	1
SciELO journals	UK	2	NA	NA	NA	NA	2
Science Data Bank	UK	1	NA	NA	NA	NA	1
Taylor & Francis	UK	1	NA	NA	NA	NA	1
Texas Data Repository	UK	1	NA	NA	NA	NA	1
U.S. Geological Survey	UK	3	NA	NA	NA	NA	3
UKnowledge	UK	52	NA	NA	7	13	72
UNC Dataverse	UK	9	NA	NA	NA	NA	9
University of St Andrews	UK	2	NA	NA	NA	NA	2
Zenodo	UK	45	NA	NA	NA	4	49
figshare	UK	64	NA	NA	NA	NA	64

Supplemental Table 5 (continued)