Upstream

INFORMATE: Where Are the Data?

Ted Habermann, **Jamaica Jones**, **Howard Ratner** and **Tara Packer**

Published December 5, 2023

Citation

Habermann, T., Jones, J., Ratner, H., & Packer, T. (2023). INFORMATE: Where Are the Data?. *Upstream*. https://doi.org/10.54900/vnevh-vaw22

Keywords

Original Research
Feature Image

Acknowledgments

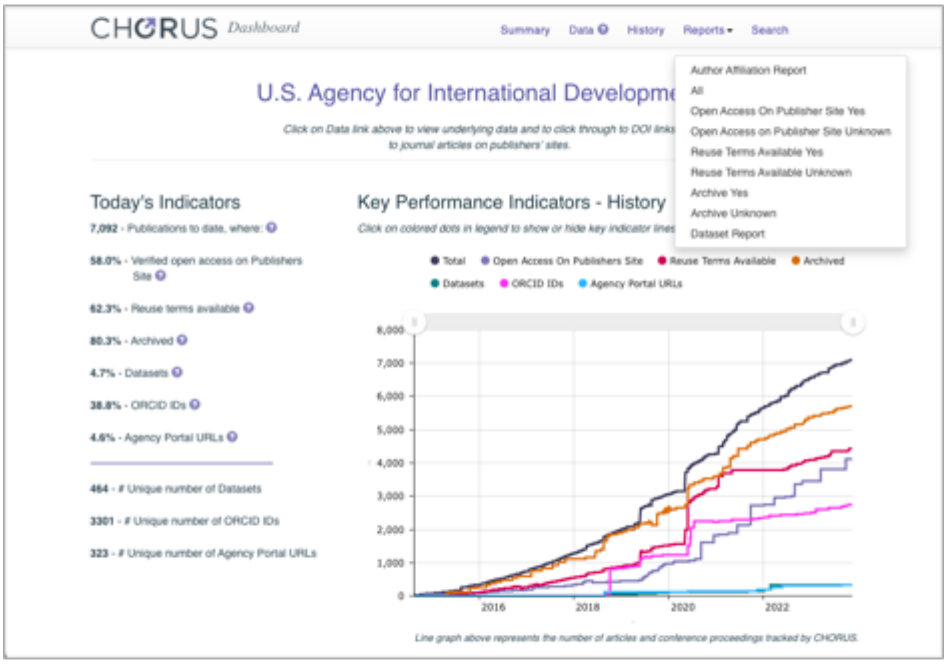
Copyright

Introduction

A recent blog post described a new partnership between Metadata Game Changers and CHORUS aimed at understanding how CHORUS data can help federal agencies, other funders, and other users access and use information from the global research infrastructure to measure this infrastructure and understand connections between research objects. These data are openly available but can be difficult to find or use because they are invisible behind a plethora of sources and APIs. Making the invisible visible is a goal of CHORUS and of this project.

CHORUS helps address this difficulty by retrieving data from across the open research ecosystem, primarily Crossref, ScholeXplorer, and DataCite, and provides several open user services on top of that data: a search (CHORUS Search Service), a dashboard (CHORUS Dashboard, Figure 1), a public API (CHORUS API), and a series of reports in tabular formats accessible to common analysis tools.

Each report is engineered and organized around funding agencies. As will be explained in detail in the following section, CHORUS data begins the "journey" with funder-specific data gathering. All reports and data generated by CHORUS carry over this central funder identity in their organization. In our project, we focus on three of the CHORUS Reports: All, Authors Affiliation, and Dataset. In this blog post we focus on the Dataset reports related to three Federal agencies: The National Science Foundation, the U.S. Geological Survey (USGS), and the U.S. Agency for International Development (USAID). These reports provide views into the work supported by these agencies that augment those provided in agency-specific repositories (NSF PAR, USAID DDL, and USGS repositories).



U.S. Agency for International Developme

Click on Data link above to view underlying data and to click through to DOI links to journal articles on publishers' sites.

Today's Indicators

7,092 - Publications to date, where:

58.0% - Verified open access on Publishers Site

62.3% - Reuse terms available

80.3% - Archived

4.7% - Datasets

38.8% - ORCID IDs

4.6% - Agency Portal URLs

464 - # Unique number of Datasets

3301 - # Unique number of ORCID IDs

323 - # Unique number of Agency Portal URLs

Key Performance Indicators - History

Click on colored dots in legend to show or hide key indicator lines

Line graph above represents the number of articles and conference proceedings tracked by CHORUS.

Figure 1. CHORUS Dashboard visualization of the history of open science parameters for USAID in CHORUS Dashboard. CHORUS data are available in reports that provide analysis-ready data for answering many questions about federally funded research. We will focus on the All, Author Affiliation and Dataset reports.

Upstream

One of our goals is to help increase visibility by looking more deeply into the data and exposing it to users, a process termed informating. In this blog post, we focus on several elements of the Dataset report.

Funder Metadata

The CHORUS data journey for any agency starts with a Crossref query for research objects funded by that agency. The DOIs returned by that query form the basis for the CHORUS All Report. Those DOIs are then searched for related datasets in ScholeXplorer, a data-linking service that returns metadata about links for research objects. For example, a ScholeXplorer search for the journal article DOI https://doi.org/10.1186/s12864-017-3751-1 searches over three million links and gives seven datasets that have been linked to this article. Finally, DataCite metadata are searched for datasets discovered in this way and DataCite metadata are included in the CHORUS Dataset report.

Comparing Crossref DOIs in the Dataset reports to those in the All reports provides information about how many of the funder journal articles cited datasets / software with links that are matched by ScholeXplorer. These numbers are rather low: NSF: 77,116/405,256 (19%), USAID: 217/7043 (3%), and USGS: 1275/6026 (21%)(Figure 2), reflecting a wide variety of challenges with identifying datasets and citing datasets/software from articles. Fortunately, many people and organizations are working to address these challenges, so we will not discuss them here.

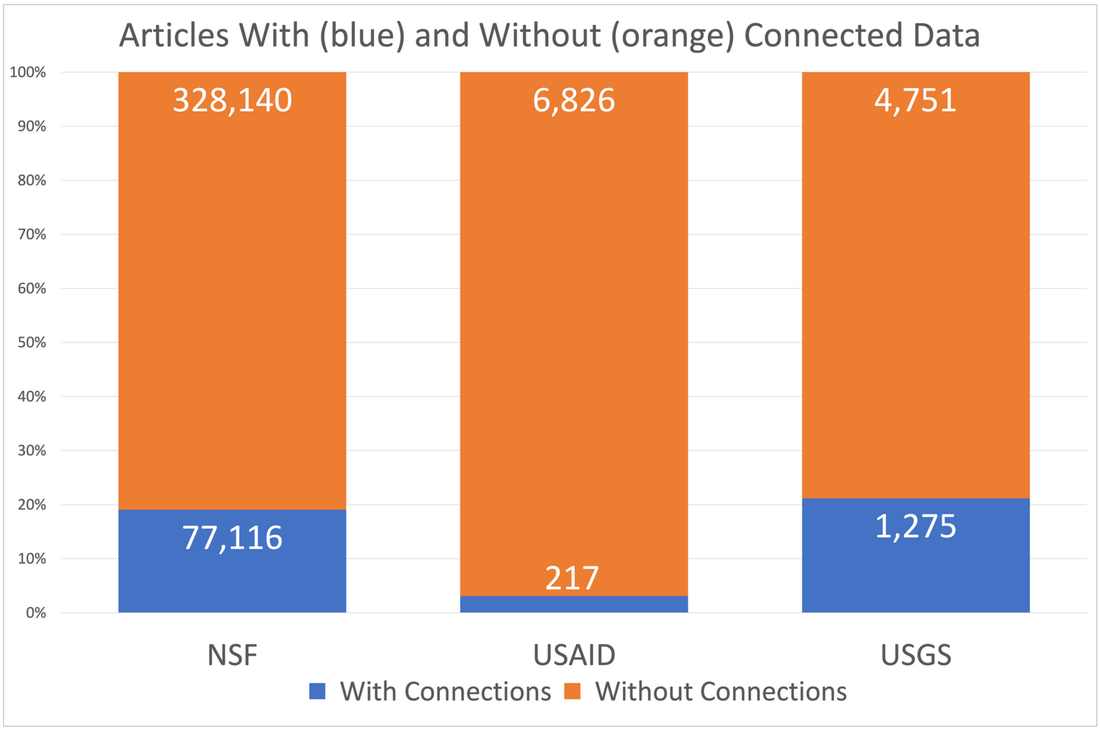


Figure 2. The percentage of article DOIs with connected datasets discovered by CHORUS using ScholeXplorer for three agencies. The withe numbers

The Dataset Report includes article DOIs and Funder Identifiers (Funder Id) from Crossref, and dataset DOIs and Funder Names from DataCite (FunderName). Table 1 summarizes these funder properties from the dataset reports for three agencies. The count % column shows the completeness of each property in the Dataset Report.

Note that all datasets have Funder Ids that match the agency Funder Ids because the articles are selected from Crossref using those Funder Ids. Funder names from DataCite are much less common in these data, i.e. between 4 and 10% of the datasets have metadata about the funders of the dataset and the most common value is always National Science Foundation.

Agency	property	count	count %	unique	most common value	common count
USGS	Funder Id	1272	100	295	10.13039/100000203 (USGS)	592
USAID	Funder Id	217	100	73	10.13039/100000200 (USAID)	90
NSF	Funder Id	76957	100	10539	10.13039/100000001 (NSF)	13594
USGS	Funder Name	49	4	26	National Science Foundation	13
USAID	Funder Name	21	10	14	National Science Foundation	6
NSF	Funder Name	5606	7	1149	National Science Foundation	2095

Table 1. Funder metadata from the CHORUS Dataset Report. Count = number of occurrences, count % = percentage of occurrences, unique = unique values, most common value = most common value, common count = number of occurrences of most common value.

Together these two observations reflect 1) the common practice of providing funder information for articles (whether structured or free-text) and 2) the focus of mandatory DataCite fields on identification and citation use cases, i.e., funder metadata is not mandatory, so it is rare in DataCite metadata.

Keep in mind that the datasets included in CHORUS are those connected to journal articles reporting on work funded by the specific agency. In other words, these are datasets used by agency fundees. They are not necessarily created with direct funding from the agencies or available in agency repositories.

Where Are the Data?

The CHORUS Dataset Report includes a property named “Dataset Repository Name” which is present for all datasets and gives the name of the DataCite repository which registered the dataset DOI and, in many cases, holds the datasets. Given the caveats described above, this property can be used to explore the question “What repositories hold the datasets that researchers funded by a specific agency are citing and, therefore, using in their work." In other words, *where are the data*?

USAID Repositories

Table 2 shows the dataset repositories for the 217 USAID datasets. It is interesting to note that the list includes well-known generalist repositories (Figshare, Dryad, or Mendeley) along with several publishers such as Taylor & Francis and the The Royal Society. Closer examination of the data shows that, while publishers are identified as repositories, the DOIs of these resources indicate that they are held in Figshare, e.g. https://doi.org/10.6084/m9.figshare.8325443 for one of the Taylor & Francis datasets. This is true for Taylor & Francis, The Royal Society, Wiley, and Optica Publishing Group, reflecting a growing number of partnerships (e.g. Taylor and Francis, PLOS) between publishers and commercial data repositories like Figshare for data stewardship and open data access. Considering these partnerships increases the Figshare count from 100 to 118, i.e. 54% of the datasets.

Dataset Repository Name	**#**	**Dataset Repository Name**	**#**
Figshare	100	UK Data Service	1
Dryad	37	James Cook University	1
Mendeley	21	ICPSR - Interuniversity Consortium for Political and Social Research	1
Taylor & Francis	13	Optica Publishing Group	1
Zenodo	9	Interdisciplinary Earth Data Alliance (IEDA)	1
Ag Data Commons	4	Palisades, NY: NASA Socioeconomic Data and Applications	1

		Center (SEDAC)	
F1000Research	4	U.S. Geological Survey	1
Harvard Dataverse	4	PANGAEA - Data Publisher for Earth & Environmental Science	1
The Royal Society	3	Wiley	1
The Global Biodiversity Information Facility	3	Rothamsted Research	1
NASA EOSDIS Land Processes DAAC	3	GFZ Data Services	1
Cambridge Crystallographic Data Centre	2	SAGE Journals	1
London School of Hygiene & Tropical Medicine	2		

Table 2. Repositories and dataset count for datasets in the CHORUS Dataset report for USAID.

Figure 3 depicts these data graphically and the dominance of the Figshare repository is clear even without the addition of the publisher counts.

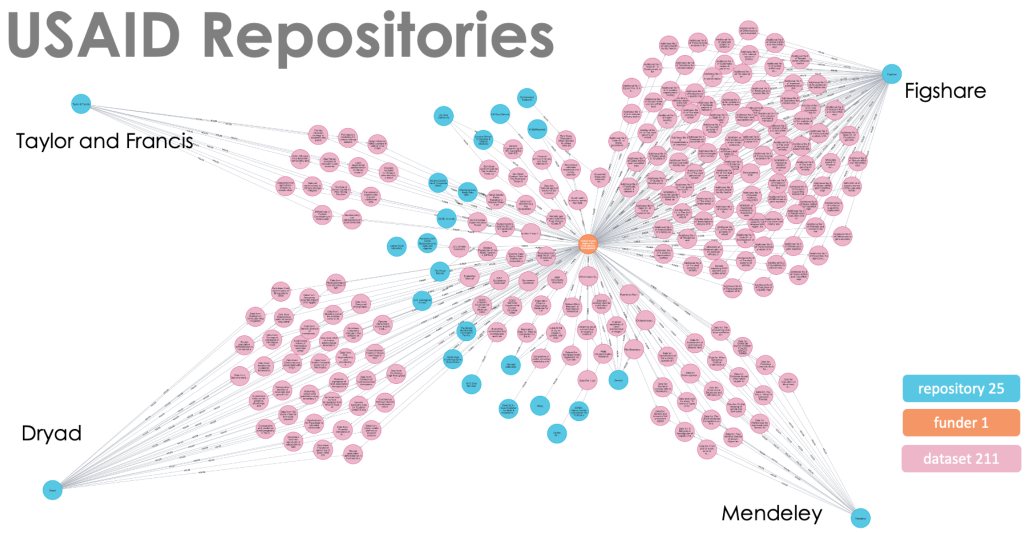


Figure 3. Connections between datasets (pink) and repositories (blue) for datasets used in research funded by USAID.

USGS Repositories

The repository data for the U.S. Geological Survey are shown in Table 3. In this case, the USGS provides a repository for nearly 60% of the datasets used by researchers funded by the USGS. The other common repositories are mostly large, generalist repositories sustained with government or member support. Figshare and other commercial repositories contain only a small number of datasets.

Dataset Repository Name	**#**	**Dataset Repository Name**	**#**
U.S. Geological Survey	741	University of Arizona Research Data Repository	2
PANGAEA - Data Publisher for Earth & Environmental Science	139	Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC)	2
The Global Biodiversity Information Facility	121	BindingDB	2

Upstream

Dryad	99	NCCWSC/CSC	1
Environmental Data Initiative	38	National Geophysical Data Center, NOAA	1
Interdisciplinary Earth Data Alliance (IEDA)	17	UK Polar Data Centre, Natural Environment Research Council, UK Research & Innovation	1
U.S. EPA Office of Research and Development (ORD)	16	Biological and Chemical Oceanography Data Management Office (BCO-DMO)	1
The Royal Society	14	U.S. Environmental Protection Agency	1
Mendeley	11	Northern California Earthquake Data Center	1
Cambridge Crystallographic Data Centre	10	Palisades, NY: Socioeconomic Data and Applications Center (SEDAC)	1
figshare	9	University of Illinois at Urbana-Champaign	1
Zenodo	9	NOAA National Centers for Environmental Information	1
Movebank Data Repository	7	Wiley	1

Upstream

Taylor & Francis	6	University of Kentucky Libraries	1
Harte Research Institute	5	NASA EOSDIS Land Processes DAAC	1
SAGE Journals	4	Arctic Data Center	1
GFZ Data Services	3	NASA Global Hydrometeorology Resource Center DAAC	1
Neotoma Paleoecological Database	3	NERC Environmental Information Data Centre	1
International Federation of Digital Seismograph Networks	2		

Table 3. Repositories and dataset count for datasets in the CHORUS Dataset report for USGS.

Figure 4 depicts these data graphically and the dominant role of the USGS repository, i.e., the similarity between the USGS repository in this case and the Figshare repository in the USAID case (Figure 3) is clear. The secondary repositories in this case (shown on the left) have similar numbers of datasets, however, most of the datasets in PANGEA (68%) and GBIF (83%) are associated with two studies related to large-scale dam removal on the Elwha River, Washington, USA, and open access biodiversity data.

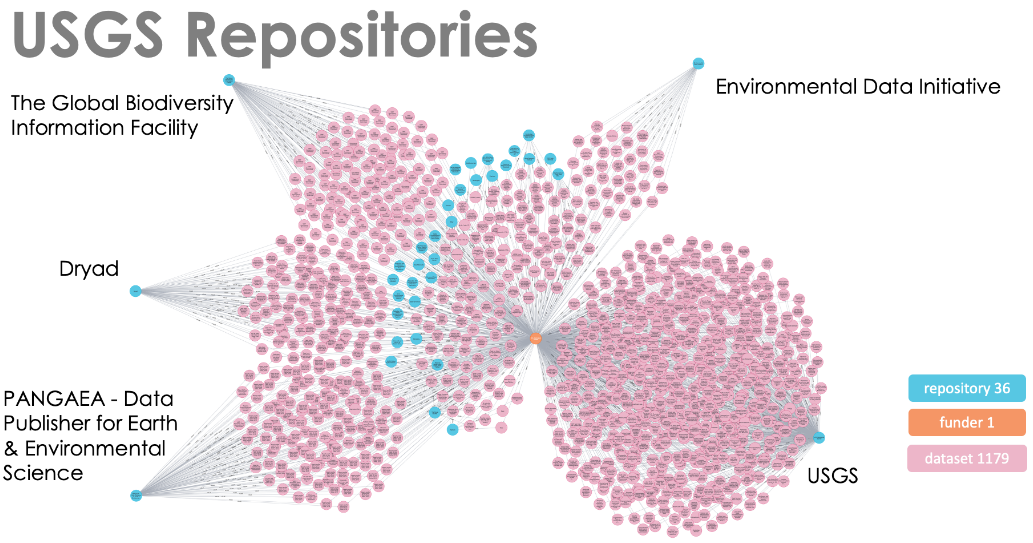

USGS Repositories
The Global Biodiversity Information Facility
Environmental Data Initiative
Dryad
PANGAEA - Data Publisher for Earth & Environmental Science
USGS
repository 36
funder 1
dataset 1179

Figure 4. Connections between datasets (pink) and repositories (blue) for datasets used in research funded by USGS.

NSF Repositories

The NSF CHORUS dataset is much larger than either the USAID or USGS collections with over 75,000 datasets and the corresponding list of repositories holding datasets referenced by NSF-funded research is also large and diverse. Table 4 shows the distribution of these datasets across repositories (those with more than 100 datasets are listed, 20 / 244). Large collections generated by the crystallographic structures and high-energy physics communities account for over 60% of the datasets in CHORUS and domain repositories associated with those communities dominate the list. Figshare shows up again and considering the datasets listed for Taylor & Francis, The Royal Society, and other publishers, accounts for just over 3000 datasets.

Dataset Repository Name	**#**
Cambridge Crystallographic Data Centre	21,403
HEPData	10,743
Dryad	3,412
The Global Biodiversity Information Facility	2,633
Figshare	2,348
PANGAEA - Data Publisher for Earth & Environmental Science	1,503
Zenodo	1,292
FIZ Karlsruhe - Leibniz Institute for Information Infrastructure	983

Upstream

Environmental Data Initiative	968
Mendeley	673
PANGAEA	588
Biological and Chemical Oceanography Data Management Office (BCO-DMO)	550
The Royal Society	368
Taylor & Francis	340
SEANOE	313
Interdisciplinary Earth Data Alliance (IEDA)	249
U.S. Geological Survey	207
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure	126
Movebank Data Repository	117
Harte Research Institute	108

Table 4. Repositories and dataset count for datasets in the CHORUS Dataset report for NSF.

Figure 5 depicts the repository distribution graphically and, even though a small portion of the total collection is shown, it shows the main features of this dataset.

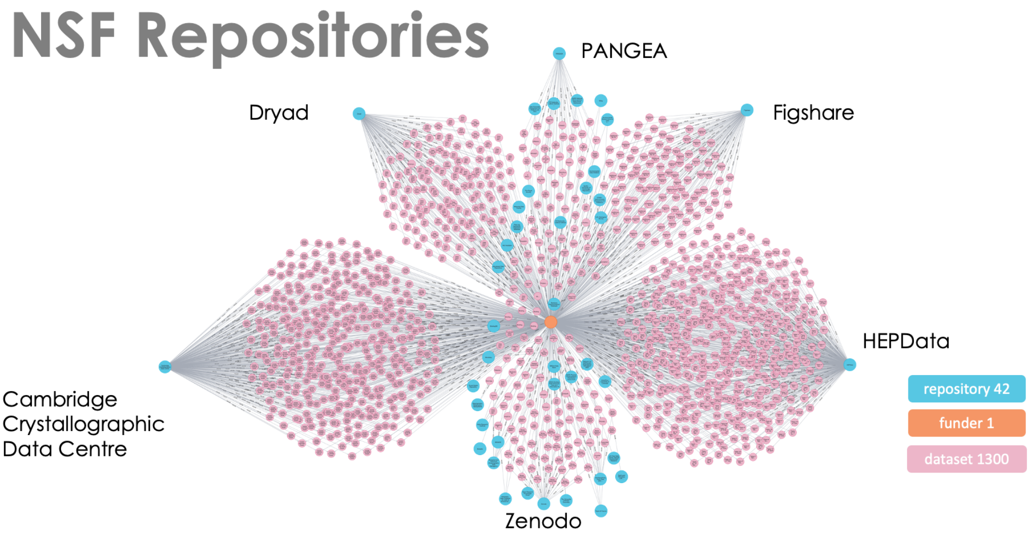


Figure 5. Connections between datasets (pink) and repositories (blue) for datasets used in research funded by NSF.

Conclusion

One goal of the INFORMATE project is to bring data harvested from the global research infrastructure by CHORUS "into the light" and share interesting observations derived from the data and the process of exploring it. This is our first blog post serving this goal. Hope you enjoy it.

The CHORUS data examined here indicate that USAID, USGS, and NSF researchers they fund have different data discovery and storage practices. The CHORUS dataset for USAID includes 217 items. The USAID Development Data Library includes metadata for over 13,000 items and over 9,000 of these include citation instructions. These instructions rely on URLs as unique identifiers for these datasets rather than DOIs.

USGS-funded researchers in CHORUS rely primarily on USGS-owned repositories, consistent with USGS Office of Science Quality and Integrity guidance. The USGS owns and operates a number of trusted repositories, with ScienceBase being one of the more widely utilized and approved repositories for USGS data releases. ScienceBase includes over 10,000 data releases with the majority of these being available in various data services (OGC Web Map or Feature) and just over 1,000 being downloadable. All of the downloadable datasets include DOIs.

The NSF CHORUS dataset is the largest of the three, with over 77,000 items. Most of these are held in two large domain repositories with the others being held in a mixture of domain and generalist repositories.

One observation is that finding datasets funded by specific agencies can be difficult because, while funder metadata are common in journal articles and Crossref, those metadata are rare for datasets in DataCite. A recent examination of funder metadata by ROR found more than 13 million funder identifiers in Crossref and less than 1 million in DataCite. This paucity of funder metadata leads CHORUS to depend on connections between articles and datasets in ScholeXplorer for discovery.

There are two ways to improve this situation: Repositories creating identifier metadata in DataCite can facilitate improved discovery by providing funder names and identifiers in addition to the required fields and 2) researchers, data curators, and journal publishers can improve their tools for finding connections between articles and papers and adding those connections into the metadata at Crossref or DataCite to be harvested into ScholeXplorer.

Another important avenue for improvement is in the development of guidance for repositories of federally funded research. The recently published White House Report on Desirable Characteristics of Data Repositories recommends identifiers for datasets but does not mention the importance of sharing those identifiers in the global research infrastructure. The more recent Public Access Memo recommends using identifiers for datasets, people, funders, and organizations and making those publicly available in public-access repositories. Implementation of these guidelines in data management design and tools (e.g. DataCite

Commons, and Plankytė et al., 2023) will certainly increase the capabilities of the global research infrastructure for all users.

Acknowledgments

This work was funded by the U.S. National Science Foundation, award 2134956.

References

Plankytė, V., Macneil, R., & Chen, X. (2023). *Guiding principles for implementing persistent identification and metadata features on research tools to boost interoperability of research data and support sample management workflows.* https://doi.org/10.5281/ZENODO.8284206