

**FAIR Motivated Metadata:  
Metadata Requirement Variation between Environmental Science Data Repositories**

Emily Ann Robles

**ABSTRACT**

Proper management of environmental science data is crucial to the progression of research into areas of large scale analyses and model creation. Data longevity and quality can be improved through metadata requirements implemented by data repositories, however there are no standards currently in place that have identified a single set of metadata fields that should be required by all environmental science data repositories. Each repository is therefore left to independently define their own requirements, and most do so under the guidance of the FAIR principles, a globally accepted set of high quality data characteristics: Findable, Accessible, Interoperable, and Reusable. To identify the variation in metadata requirements implemented by data repositories, I selected a group of 15 common metadata requirements and surveyed their presence in 15 environmental science data repositories. Only one data repository required all 15 of the metadata requirements tested, and many repositories were lacking key fields that have been identified as crucial to the long term usability of data. Additionally, common requirements related to the findability and accessibility characteristics of FAIR data were present more consistently than those that promote interoperability and reusability, suggesting that these FAIR principles may be better addressed by the implementation of strategies that fall outside of metadata requirements alone.

**KEYWORDS**

data longevity, data stewardship, FAIR principles, open data, standards, data management

## INTRODUCTION

Data and metadata quality control within the environmental sciences is crucial not only for the current generation of scientists, but to ensure usability of research data for generations to come (Michener et al. 1997). Because publications are the prevailing method of presenting scientific findings, data and its associated metadata have seldom been properly managed or stored from research by past generations, leading to massive amounts of data “death” (Pepe et al. 2014; Mayernik et al. 2020). This data entropy often results from a lack of accuracy and completeness of the data record and accompanying metadata. Without the prioritization of proper data management during collection, there is no guarantee that even the original researcher will recall enough information needed to reproduce their findings within just months of publication. (Michener et al. 1997). Additionally, advancements in the technological strategies used to store data have accelerated this process of data entropy as proprietary software become obsolete (Michener et al. 1997; Borer et al. 2009). Therefore, complete metadata are necessary for the reusability of data both by outside researchers *and* the original collectors.

Modern efforts in environmental science research, which have grown over the recent decades to become focused on global analysis within extended temporal ranges, have led scientists to routinely use the data of others in conjunction with their own (Michener et al. 1997). Quality metadata and data, which increase the usability and longevity of research, are therefore critical to the advancement of environmental science research and the ability to track long-term and wide spatial changes (Halbritter et al. 2020).

Also critical to this advancement are data repositories, which offer researchers a platform on which to host their data and remove the burden of hosting data long term from the researcher. Long term data repositories also provide users with easily accessible and often publicly available data, accompanied by the associated metadata in the form of a data package. However, repositories can only meet the needs of the scientific community successfully when the data packages they host are of high quality (Palmer et al. 2005; Michener et al. 1997). Although the need for standardized metadata quality requirements for published data packages is not a recent realization, nor is its importance disputable, the difficulties of implementing such a system are abundant. Without proper requirements for metadata completeness before publication, data repositories cannot guarantee the longevity of the data packages they store, and researchers risk

losing years of work to degradation. Individual repositories have implemented their own metadata requirements in response to the concern of data longevity, but strategies across repositories are not uniform (Marcial and Hemminger 2010).

The FAIR principles are organized into four categories, each representing an important element of a quality data package: Findability, Accessibility, Interoperability, and Reproducibility. The guidelines were created with the intent of providing a backbone for quality standardization (Wilkinson et al. 2016). The FAIR principles are being used by data repositories for the creation of data and metadata requirements to improve the quality of the data that they store (Varadharajan et al. 2019). However, in their current stage, the FAIR principles cannot be directly translated into metadata requirements and are largely open to interpretation (Mons et al. 2017). As a result, repositories and archives are implementing reporting requirements loosely based on FAIR. Consequentially, there is no standardized set of metadata requirements for the publication of environmental science data, limiting the overall longevity of archived research.

In this study, I explore the implementation of metadata reporting requirements in a sample of environmental science data repositories. When documented, similarities in metadata requirement implementation can help determine whether a core group of requirements already occurs in practice. Furthermore, recording the characteristics of each repository, such as funding sources, age, and size, may help to identify how these characteristics affect the variability of metadata requirements.

## **BACKGROUND**

### **The FAIR Principles**

The FAIR Guiding Principles for scientific data management and stewardship, published in 2016, defined findability (F), accessibility (A), interoperability (I), and reusability (R) as core components of high quality data (Wilkinson et al. 2016). Since publication, the FAIR principles have been widely adopted by researchers and governments worldwide and across a wide range of scientific domains as a strategy to maximize the value and longevity of data. During the G20 summit in 2016, the principles were referred to as an important element of efforts to promote innovation in scientific and technology (Mons et al. 2017). It is important to recognize FAIR

principles are idealized characteristics of data, not a defined standard that can be directly applied (Mons et al. 2017). Therefore, the task of managing data in a manner consistent with their interpretation of the FAIR principles falls upon researchers and repositories. One key strategy for making data more FAIR compliant is to improve metadata, the set of data which accompanies and describes characteristics of data that are necessary for future reuse and understanding.

## **Data repositories**

Increasingly, journals and funding agencies across scientific disciplines are implementing requirements for scientists to publicly archive data associated with their research studies and findings (Roche et al. 2015; Hillebrand and Gurevitch 2013). These public data archival (PDA) policies drive researchers to publish their data through data repositories. These repositories accept and publish data from researchers and have long term data preservation plans. Data repositories now carry the burden of data management that was once placed solely on the scientist or project team and play a large role in the progression of environmental science research, as they often promote the sharing of data through open access policies (Wolkovich et al. 2012; Baker 2009).

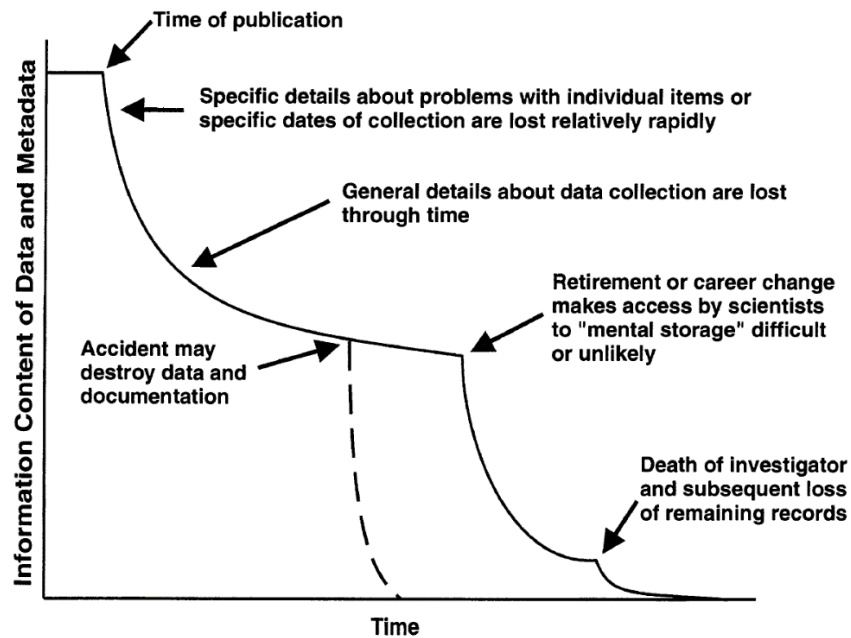
Environmental and earth science repositories differ widely on size, funding, and purpose. Some are built for specific projects and only accept data generated by these projects. For example, they may accept data from projects funded by a specific university or government agency, while others are open to any researchers who would like to contribute their data. Of the repositories I surveyed, three illustrate a few of the main differences between repositories. First, the Cornell University Geospatial Information Repository (CUGIR) is a small repository run by students at Cornell University (Table 1). It focuses on specific data types generated by researchers affiliated with the university and provides open access to data. Second, the Oak Ridge National Laboratory's Distributed Active Archive Center (ORNL DAAC) is a larger repository with over 1,500 datasets in its collection (Table 1). It is funded by the United States Department of Energy and requires that data contributors be approved based on the types of data that they store. Finally, PANGAEA is the largest repository surveyed in this research with over 400,000 datasets in its collection (Table 1). It accepts data from any projects or researchers within its field of interest, regardless of funding background.

Although PDA policies have increased the quantity of environmental science data being contributed to long term data repositories, they do not guarantee that this data will be of high enough quality to be reproduced or used. In 2015, a study of 100 ecological and evolutionary research datasets that had been archived in compliance with PDA requirements found 56% of datasets to be incomplete and 64% were archived in a manner that prevented reuse, concluding that data published in compliance with PDA policies are often lacking in quality (Roche et al. 2015).

Repository certifications have been created as a way to certify the quality of stored data by increasing a repository's trustworthiness. Member repositories must meet requirements in a variety of areas, such as a commitment to data longevity and trustworthiness, proper staffing and financial stability to enable long-term preservation planning, and the implementation of descriptive metadata requirements (CoreTrustSeal Standards And Certification Board 2019).


## **Metadata**

Metadata, which are additional data that provides sufficient information about published data to enable reuse, are essential to the longevity and usability of data. Metadata can be included as separate files within data packages (such as read.me files), within the datasets themselves through descriptive and defined column names, or even as images. Metadata clarifies how to use data and benefits not only future researchers but also the original investigator, by ensuring that no details about the research are forgotten or lost (Figure 1).



**Figure 1. Data death over time.** Degradation of data from time of publication until the death of original investigator (Michener et al. 1997)

A data package's first impression on a user is the metadata included on the public facing landing page of the data package (Figure 2). The metadata on this landing page is often referred to as "package level metadata." Metadata included at this level can vary from simple titles and abstracts to precise geographic information and data collection methods.



ESS-DIVE  
Data Infrastructure for a Virtual Ecosystem

DATA SUPPORT ABOUT [Submit Data](#) [Sign in with ORCID](#)

Damerow J ; Varadharajan C ; Boye K ; Brodie E ; Burnus M ; Chadwick D ; Cholla S ; Crystal-Omelas R ; Elbaashandy H ; Eloy Alves R ; Ely K ; Goldman A ; Hendrix V ; Jones C ; Jones M ; Kakalia Z ; Kemmer K ; Kensing A ; Maher K ; Merino N ; O'Brien F ; Perzan Z ; Robles E ; Snavely C ; Sorensen P ; Stegen J ; Weisenhorn P ; Whitenack K ; Zavarin M ; Agarwal D  
(2020): Sample Identifiers and Metadata Reporting Format for Environmental Systems Science. Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE). doi:10.15485/1660470

Citations 0 Downloads 0 Views 1 [Copy Citation](#) [Assessment report](#)

Files in this dataset: Package: ess-dive-4b702e966850e-20210430T040402157330

Name	File type	Size	Login to Download
Metadata: Sample_Identifiers_and_Metadata_Reporting_Format_for_Environmental_Systems_Science.xml	EML v2.1.1	25 KB	<a href="#">Download</a>
README.md	<a href="#">More info</a> text/markdown	8 KB	<a href="#">Download</a>
material.md	<a href="#">More info</a> text/markdown	6 KB	<a href="#">Download</a>
sampleMetadataTranslationTable.csv	<a href="#">More info</a> text/csv	13 KB	<a href="#">Download</a>

[Show 12 more items in this data set](#)

**General**

Identifier: ess-dive-4b702e966850e-20210430T040402157330

Abstract: The ESS-DIVE sample identifiers and metadata reporting format primarily follows the System for Earth Sample Registration (SESAR) Global Sample Number (IGSN) guide and template, with modifications to address Environmental Systems Science (ESS) sample needs and practicalities (IGSN-ESS). IGSNs are associated with standardized metadata to characterize a variety of different sample types (e.g. object type, material) and describe sample collection details (e.g. latitude, longitude, environmental context, date, collection method). Globally unique sample identifiers, particularly IGSNs, facilitate sample discovery, tracking, and reuse; they are especially useful when sample data is shared with collaborators, sent to different laboratories or user facilities for analyses, or distributed in different data files, datasets, and/or publications. To develop recommendations for multidisciplinary ecosystem and environmental sciences, we first conducted research on related sample standards and templates. We provide a comparison of existing sample reporting conventions, which includes mapping metadata elements across existing standards and Environment Ontology (ENVO) terms for sample object types and environmental materials.

We worked with eight U.S. Department of Energy (DOE) funded projects, including those from Terrestrial Ecosystem Science and Subsurface Biogeochemical Research Scientific Focus Areas. Project scientists tested the process of registering samples for IGSNs and associated metadata in workflows for multidisciplinary ecosystem sciences. We provide modified IGSN metadata guidelines to account for needs of a variety of related biological and environmental samples. While generally following the IGSN core descriptive metadata schema, we provide recommendations for extending sample type terms, and connecting to related templates geared towards biodiversity (Darwin Core) and genomic (Minimum Information about any Sequence, MiX3) samples and specimens. ESS-DIVE recommends registering samples for IGSNs through SESAR, and we include instructions for registration using the IGSN-ESS guidelines. Our resulting sample reporting guidelines, template (IGSN-ESS), and identifier approach can be used by any researcher with sample data for ecosystem sciences.

Keywords: CATEGORICAL\_NONE

Keyword	Type
sample identifier	
Global Sample Number (IGSN)	
ESS-DIVE reporting format	
multidisciplinary	
data linking	
FAIR data	
persistent identifier	
Environment Ontology (ENVO)	
environmental context	

**Figure 2. Data package landing page example.** First portion of a data package landing page from the Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE), including title and citation, identifier, abstract, and keywords fields (Damerow et al. 2020).

For this study, I surveyed metadata at the package level because it is crucial to the initial findability and understanding of archived data, and because requirements are often set by repositories themselves. Increasingly, data repositories are using the FAIR principles to guide their metadata requirements. However, there is no set standard for what should be included in package level metadata across environmental science data repositories.

## METHODS

### Repository selection and characterization

To begin, I selected 15 repositories from the FAIRsharing list of databases filtered under the “environmental science” key phrase (Figure 3).

The screenshot shows the FAIRsharing.org website interface. At the top, there is a search bar and navigation tabs for Standards, Databases, Policies, Collections, Add/Claim Content, Stats, and Log in or Register. A red tag 'Environmental Science' is selected. Below the search bar, there are filters for 'View as Table' and 'View as Grid', a 'Sort by' dropdown set to 'Best Match', and sections for 'Recommended Records', 'Associated Publication?', 'Claimed?', 'Record Status', and 'Domains'. The main table displays search results for 'Environmental Science', showing records 1-50 of 134. The table has columns: Registry, Name, Abbreviation, Type, Subject, Domain, Taxonomy, Related Database, Related Standard, Related Policy, and In Collection/Rec. Three records are visible: Comparative Toxicogenomics Database (CTD), PubMed Central (PMC), and PubMed (PubMed).

Registry	Name	Abbreviation	Type	Subject	Domain	Taxonomy	Related Database	Related Standard	Related Policy	In Collection/Rec
	Comparative Toxicogenomics Database	CTD	Database	Anatomy, Biomedical Science, Comparative Genomics, Environmental Science, Systems Biology, Plus 2 more...	Adverse Reactions, Annotation, Drug Interaction, Gene Ontology Enrichment, Gene Expression Data, Plus 11 more...	3D4, 3D4, 3D4, 3D4, 3D4, Plus 5 more...	NCI Gene Wiki, DrugBank, Reactome, PC, Plus 12 more...	CL, NCBITAXON, DOID, GO, MESH, Plus 4 more...	None	Technology, Chemistry, COVID-19 Resources, RDA Covid-19 WG Re
	PubMed Central	PMC	Database	Biomedical Science, Earth Science, Environmental Science, Epidemiology, Life Science, Plus 1 more...	Bibliography, Publication	AI	Europe PMC, ADP/bioDB	CIDO	None	Open Science Prize, COVID-19 Resources, RDA Covid-19 WG Re
	PubMed	PubMed	Database	Biomedical Science, Earth Science, Environmental Science, Life Science, Traditional Medicine	Bibliography, Behavior	AI	NONCODE, Europe PMC, STRING, Guide to Pharmacology, CiteAb, Plus 19 more...	M2CAST, MESH, NN, Common Metadata Elements for Cataloging Biomedical Datasets, CIDO	None	None

**Figure 3. FAIRsharing.org database.** A portion of the FAIRsharing database search, filtered using the Environmental Science tag.

I then assigned each repository a shortened ID built a profile with five characteristics: size (measured by the number of datasets available); funding type (U.S. Government, International government, University, or non-government organization); whether public facing content mentions the FAIR principles; and whether the repository is CoreTrustSeal certified (Table 1).

**Table 1. All repository characteristics surveyed.** Characteristics were found through FAIRsharing.org and/or within the repository's public facing content on websites.

Repository ID	Repository Name	CoreTrust Seal Member	Mentions FAIR	Funding	Size (Datasets Available)	Repository Age
KNB	Knowledge Network for Biocomplexity (KNB)	No	Yes	US Government	27,963.00	>5 years
EDI	Environmental Data Initiative	No	Yes	US Government	7,996.00	>5 years
NCEI	NOAA National Centers for Environmental Information	No	No	US Government	Not available	>5 years
DRYAD	DRYAD	No	Yes	Non-profit NGO	40,539.00	>5 years
ENVD	Environmental Data Portal	No	Yes	US Government	373.00	< 5 years
PANGAEA	PANGAEA - Data Publisher for Earth and Environmental Science	Yes	Yes	Multiple Governments	402,229.00	>5 years
TERN	Advanced Ecological Knowledge and Observation System from the Terrestrial Ecosystem Research Network (TERN AEKOS)	No	Yes	International Government	1,278.00	>5 years
ORNLDAAC	Oak Ridge National Laboratory Distributed Active Archive Center	Yes	No	US Government	1,521.00	Not available
NCAR	Research Data Archive at the National Center for Atmospheric Research (NCAR)	Yes	No	US Government	68,861.00	>5 years
AODN	Australian Ocean Data Network Portal	No	No	International Government	275.00	Not available
CUGIR	Cornell University Geospatial Information Repository	No	No	University	458.00	< 5 years
USGS	USGS Science Data Catalog	No	Yes	US Government	18,487.00	>5 years
ADC	Arctic Data Center	Yes	Yes	US Government	6,406.00	>5 years
ESS	Environmental Systems Science Data Infrastructure for a Virtual Ecosystem	No	Yes	US Government	469.00	< 5 years
BCO	Biological and Chemical Oceanography Data Management Office	Yes	No	US Government	9,883.00	>5 years

## Metadata requirements

To identify variation in metadata requirements, I first chose a set of 15 common metadata fields to test for each repository (Table 2).

**Table 2. Metadata requirements surveyed.** Metadata field descriptions and examples were adapted from the Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) Package Level Metadata Guide at <https://docs.ess-dive.lbl.gov/data-and-metadata-upload/package-level-metadata>.

Requirement	Description	Example
Title	A title will generally include information such as the topic, geographic location, dates, and scale of data. If data is associated with a journal publication, the data package title may include the journal name.	<i>Raw sapflow and soil moisture data from January 2016-April 2016 in Manaus, Brazil</i>
Abstract	A description of the content included which should provide all necessary scientific context needed to promote the reproducibility of data. A clear description of the research question or statement of purpose should also be included.	<i>This data package contains raw output from a data logger connected to 9 sapflow and 5 soil moisture sensors in Manaus, Brazil. The file xxx.dat contains raw data and the metadata file (BR-Ma2_E-fieldlog_20160501.xls) has information on locations where the sensors were installed and other sensor maintenance details. No data processing or QA/QC was done on the raw data packages. Processed data will be uploaded as separate data packages on ESS-DIVE. This research was performed as a part of the NGEE Tropics project.</i>
Keywords	Terms included to increase the findability of data packages, often through repository search capabilities, and identify the themes of the data.	<i>Earth science; land surface; soils</i>
Data variables	A list of data variables included in datasets.	<i>Soil moisture</i>
Usage rights	Determine how data can be shared and reused.	<i>Creative Commons Attribution (CC BY 4.0) requires that the data package be cited by anyone using the data. Creative Commons Public Domain (CC BY 1.0) dedicates the data to the public domain without restriction.</i>
Related references	Full citations of data packages or publications associated with the data package.	<i>Somebody J. (2018), Sapflow and soil moisture coupling in the Amazon, Journal. doi: xx.xxxx</i>
Funding source or project	A list of organizations that funded the research, or the project associated with data production.	<i>U.S. DOE &gt; Office of Science &gt; Biological and Environmental Research (BER) or Next-Generation Ecosystem Experiments (NGEE) Tropics</i>
Data package contact	A single individual who should be contacted by users seeking further information for the data. Including emails or ORCIDs is often recommended.	<i>First name, Last name, Organization, Email, ORCID</i>
Data package authors	The main researchers involved in producing the data. These authors will appear in the data package citations, and full contact information is often recommended.	<i>First name, Last name, Organization, Email, ORCID</i>
Temporal coverage	Start and end dates of data collection.	<i>2017-04-16, 2019-07-13</i>

Geographic description	A description of the location(s) where data were collected. This field may not be relevant for specific data types.	<i>Br-Ma2, Manaus, Brazil: ZF2 K34 Tower. Eddy covariance site established in 1999 on kilometer 34 of the ZF2 highway. It was later expanded into atmospheric and soil sampling hub. It is a 1.5m x 2.5 m- section aluminum tower.</i>
Coordinates	Latitude and longitude of location(s) the data represent. This field may not be relevant for specific data types.	<i>Northwest Coordinates [Lat Long]/Southeast Coordinates [Lat Long]</i>
Methods	A thorough description of all aspects of data production necessary for the reproduction of data. Some repositories may allow for the citation of existing methods that have already been published.	<i>“Step 1: ... Step 2: ... “ <b>OR</b> “See &lt;related reference&gt; for field sample collection and handling methodology”</i>
Persistent identifier	Identifiers that allow for the long-term location of digital objects, such as data packages.	<i>Digital Object Identifier (DOI): <a href="http://dx.doi.org/XXXX">http://dx.doi.org/XXXX</a></i>
Citation available	A full citation of the data package, which may also include relevant persistent identifiers.	<i>Jardine K ; Zorzanelli R ; Gimenez B ; Robles E ; Rosa L (2020): <u>Raw leaf isoprene and monoterpene emission GC-MS chromatograms/calibrations for MassHunter software, Brazil, 2014-2016. Next-Generation Ecosystem Experiments (NGEE) Tropics. doi:10.15486/NGT/1602144</u></i>

Next, I searched for repository requirements using one of 3 pathways: preferably, by following the steps of a data contributor and noting requirements throughout the process. This approach offers the most insight into the data contributor’s experience. Even if there is documentation of required metadata for publication elsewhere on a repository’s website, the instructions and guidelines throughout the submission process will be the final determination of what the researcher provides. However, collecting requirements using this pathway was in some cases inaccessible due to barriers such as requirements for registration as a contributor. Due to those barriers, I also collected data by reviewing available documentation and instructions for data contributors. If the contributor workflow was not accessible, and documentation of metadata requirements for submission were not available, I contacted the repository team to request more information. If the repository team did not respond within one week, I removed the repository from my list. Once located, I used a Google Form to record the presence of the 15 metadata requirements (Table 2).

## **Presence of reporting requirements**

To divide repository characteristics from metadata requirement data, I compiled results from the Google Form into a spreadsheet. The first objective of this study focused primarily on identifying the presence of metadata reporting requirements. Each requirement was scored as present (1) or not present (0) and totals were calculated for each repository (Appendix A).

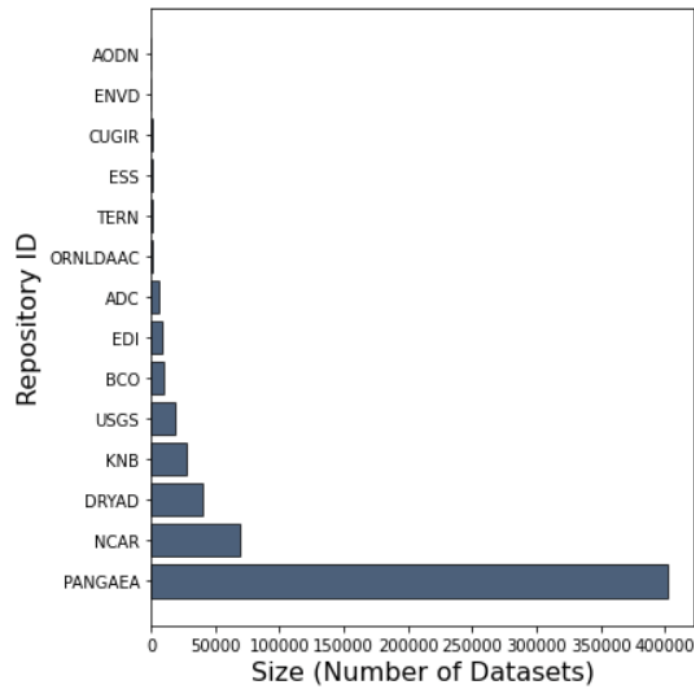
## **FAIR category principles**

The second objective of this study was to determine which categories of the FAIR principles are most often addressed within metadata requirements implemented by repositories. I created four categories based on the definitions of findable, accessible, interoperable, and reproducible provided in 'FAIR Guiding Principles for scientific data management and stewardship' Scientific Data' (Wilkerson et al. 2016). I then divided the original 15 metadata requirements into these categories. Depending on their definitions, some requirements were present in multiple groups. The completeness of each category was calculated based on the number requirements and the percentage of present (1) scores for each requirement.

# **RESULTS**

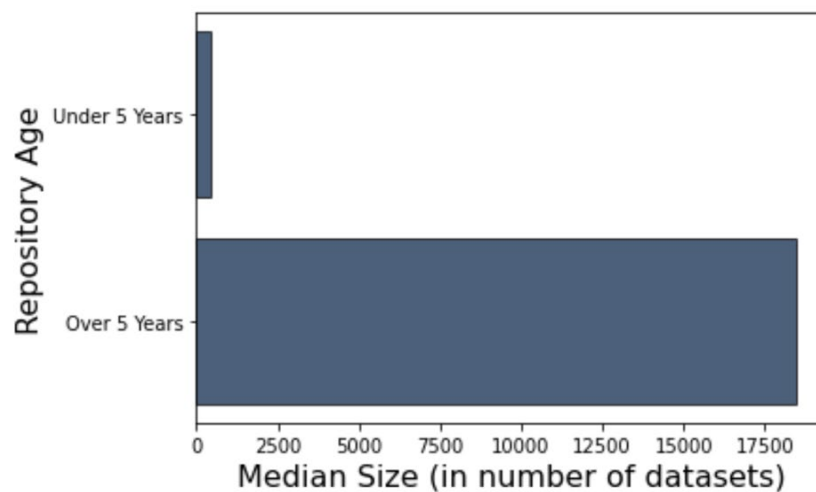
## **Repository characteristics**

The 15 repositories selected included 12 government funded (5 US National Science Foundation, 5 other US government, 2 international government), one non-profit organization, one University repository, and one repository funded by multiple international government organizations (Table 1). The following figures use the shortened repository IDs – full repository names can be found in the table included in Table 1. Ten of the selected repositories have been accepting data for over five years, two have been accepting data for less than five years. Five repositories were CoreTrustSeal Certified repositories (Table 1). The smallest data repository had a collection of 275 datasets, while the largest had a collection of 402,229 datasets (Figure 5). The median dataset collection size was 41,909.86.



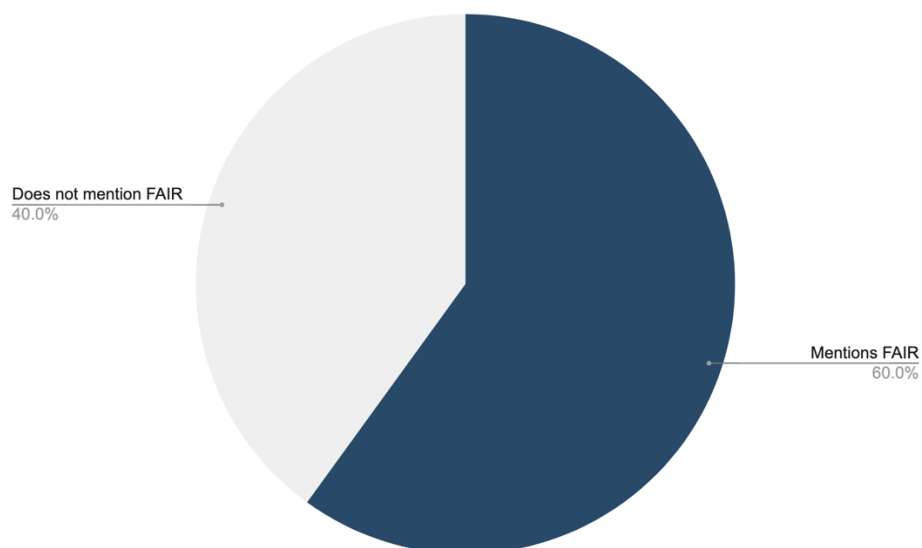
**Figure 4. Size of repositories.** The size of each repository was measured by surveying the number of datasets within the repository's collection.

Repositories that were under 5 years old ( $n=3$ ) had a median size of 458 data packages, while repositories over 5 years old ( $n=9$ ) had a median size of 18487 data packages (Figure 5). Since data collection size and age were not available for all 15 data repositories, three (ORNL DAAC, NCEI, AODN) removed during calculation of medians.



**Figure 5. Median size of repository by age group.** Median size was measured by the number of datasets available in a repository's collection. Two groups were created: over 5 years old (n=9) and under 5 years old (n=3)

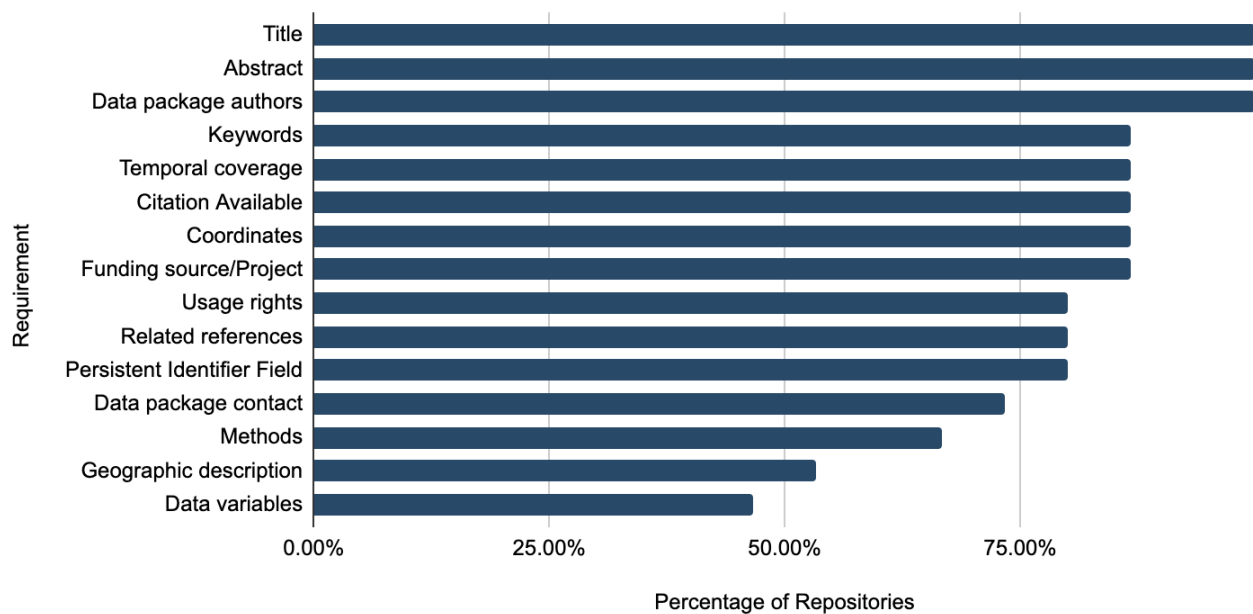
Over half (60%) of the repositories surveyed included statements about promoting FAIR data on their public facing content, including in website or documentation material (Figure 6).



**Figure 6. FAIR motivation in repositories.** Percentage of the 15 surveyed data repositories that did or did not clearly mention the FAIR principles within their public facing website or documentation.

## Repository metadata requirements

Only three metadata requirement fields, abstract, title, and authors, were present in every repository surveyed (Figure 7). The least common requirements were data variables (47%) and geographic descriptions (53%). Data package contacts were available in 11 of the surveyed repositories (73%). More repositories included data package citations (87%) than persistent identifiers (80%) (Figure 7).



**Figure 7. Metadata requirement presence in percentage of data repositories.** Each metadata field was ranked based on the percentage of the 15 surveyed data repositories that required the field.

## FAIR Categories

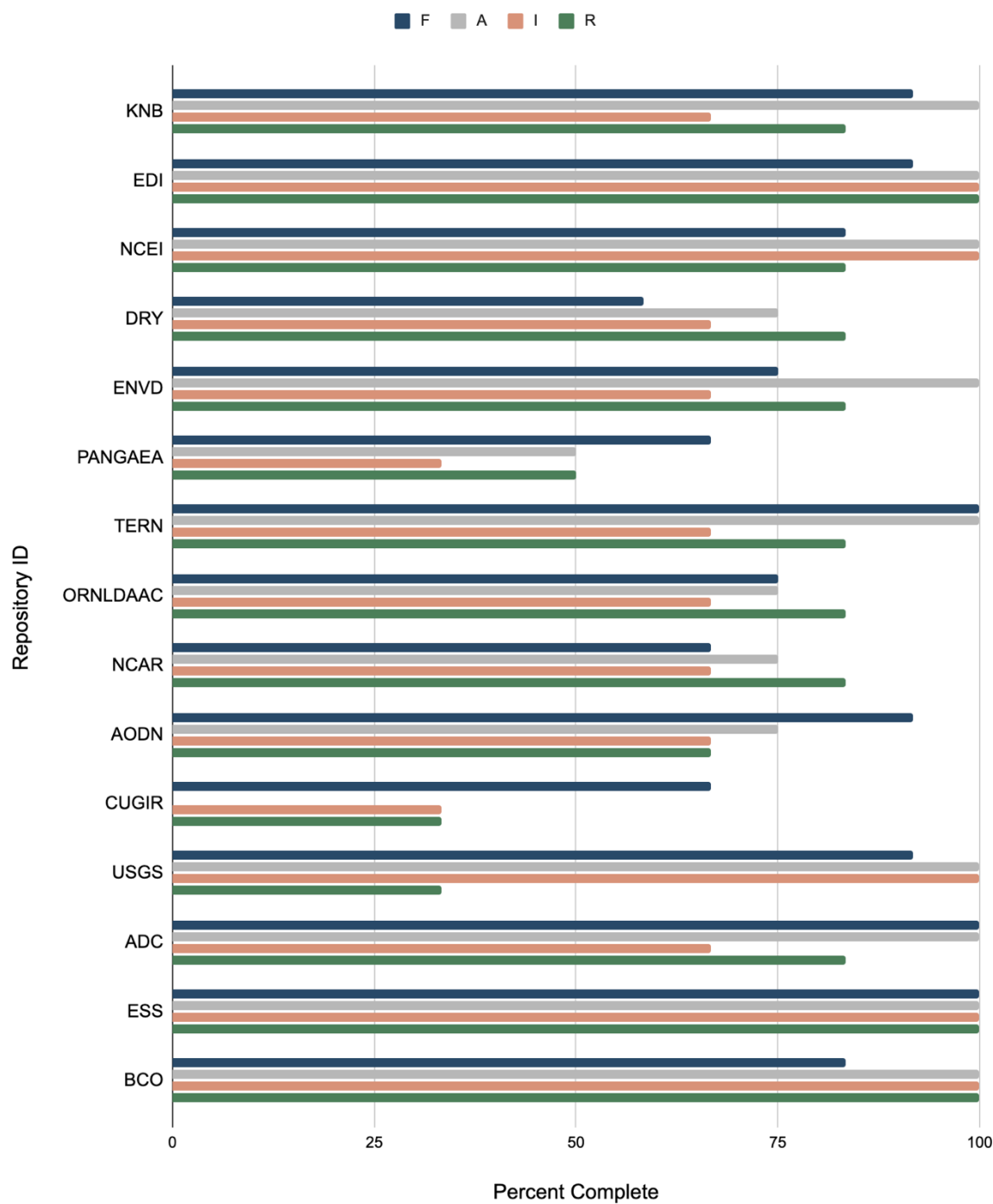
After categorizing each of the 15 metadata requirement by its relation to the FAIR principles, the findability category contained 12 requirements, the accessibility category contained 4 requirements, the interoperability category contained 3 requirements, and the reusability category contained 6 requirements (Table 3). Of the requirements, 7 were categorized into more than one category (Table 3)

**Table 3. Requirement fields and FAIR categories.** Each requirement was assigned to one or more FAIR category, dependent on which principle the associated field promotes.

Requirement	FAIR Category
Title	F
Abstract	F, R
Data package authors	F
Keywords	F
Temporal coverage	F
Citation available	F, A, R
Coordinates	F
Funding source/project	F, A
Usage rights	R
Related references	I, R
Persistent identifier	F, A
Data package contact	F, A, I, R
Methods	I, R
Geographic description	F
Data variables	F

Repositories included the requirements that fell under the findable and accessible categories (82.8% and 83.3% average completeness respectively) more often than those in the interoperable and reusable categories (73.3% and 76.6% respectively). The category for requirements that promoted findability was the largest, meaning that more of the 15 core requirements promoted findability than any other FAIR principle, and also the most fulfilled by the repositories, meaning that on average more repositories required the fields in the findable category than they required fields in the accessible, interoperable, and reproducible categories.

Only one data repository, the Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE), required every requirement surveyed. Seven repositories included at least 75% of the findable (F) requirements. Each of the 15 surveyed repositories included at least 25% of the requirements in each of the FAIR categories, except for the Cornell University Geospatial Information Repository (CUGIR), which did not have any requirements that fell within the reusable (R) category.



**Figure 8. FAIR category completeness by repository.** These percentages illustrate the percentage of possible fields in each FAIR category that were required by the data repository.

## DISCUSSION

Metadata has the ability to ensure that environmental science data remains of high enough quality to support meta-analyses and encourage reuse well past the original publication of research findings. As well accepted indications of high quality data, the FAIR principles, while not a standard that can be directly translated into metadata requirements, have driven the creation of metadata reporting requirements within many data repositories. However, even while the majority of the 15 repositories surveyed referenced the FAIR principles within their public facing content, only three of the 15 common metadata fields were present across every repository. Only one repository included all 15 fields in their requirements. Additionally, the distribution of the 15 metadata fields into FAIR categories identified strengths in the promotion of findability and accessibility and weaknesses in the support of interoperability and reusability.

### Repository characteristics

The funding background of a data repository may influence the ability to implement thorough metadata requirements. Metadata curation, which often occurs in repositories after a data package is submitted for publication, requires the dedication of resources to a publication review process. The length of the curation process for each data package can be impacted by the number of requirements that must be verified before approval for publication. As repositories grow, the curation process can become a barrier to scaling up (O'Brien 2019). This research showed that data repositories below 5 years of age had a much smaller data collection sizes when compared to the repositories older than 5 years (Figure5).

To mitigate the increase in resources demanded by high volumes of publication requests, repositories have implemented automated components to confirm the presence of certain fields before a manual review of content is performed to shorten review time (O'Brien et al. 2019). If a repository does not have the resources to create such a system, their metadata requirements may have to be made broader to compromise, or the review time may be quite long. Lengthy review times is prevalent for PANGAEA, the largest repository surveyed, which states in their data contributor documentation that the publication process can take up to several months due to high volumes of data submissions.

Five of the surveyed repositories were CoreTrustSeal certified repositories, and I expect that number will grow in the coming years. As funders and journals are increasingly requiring researchers to publish their data, they are likely to also ask that data be archived with repositories that meet certain requirements. Through certification, repositories make themselves appear more trustworthy to both the data contributor and data user (CoreTrustSeal Standards And Certification Board 2019). The certification itself requires a certain amount of resources to achieve, though, as the process to be accepted as a member repository requires the organization to perform a detailed self-evaluation and meet a list of 16 requirements specific to areas such as data curation levels, organizational infrastructure and well documented storage procedures, and thorough plans for long-term preservation (CoreTrustSeal Standards And Certification Board 2019).

### **Metadata requirements**

The three fields that were most commonly required across all sampled repositories are also fields commonly required by research journals. When submitting to a journal, it is highly likely that a title, an abstract, and a list of authors will be required by the publisher. Therefore, the frequency of these requirements in data repositories could be associated with the fact that researchers may be more willing to provide this information due to ease and familiarity. Other fields that were not as common, such as a geographic description (53%) and data variables (47%), may have variability related to the purpose of the repository and the data type included. For example, a full geographic description may not be necessary for a global dataset. However, two of the least common fields, a data package contact and a clear description of data collection methods, are vital to the reproducibility of data. Relying on the contact information or names of authors alone does not guarantee that a researcher will be accessible over the lifetime of the data. A study of 516 research articles from 2-22 years of age found that the odds of finding working email addresses for authors fell by 7% each year after publication (Vines et al. 2014). Whether a project PI, a data manager, or even one of the authors, designating a single researcher as the long term point of contact for any questions about the data is essential to its longevity.

The expectation of a statement of usage rights in each data repository also was not consistently present, which was concerning. The clear documentation of user and ownership rights is essential to mitigating possible risks related to restrictive legal limits on data reuse (Mayernik et al. 2020).

Although some repositories that did not have this requirement visible may have had overarching usage rights that apply to every data package in their collection, it is still vital that this information be provided clearly to the data user.

Although it was not a part of my original data collection plan, I noticed that no repository offered a field for data contributors to disclose uncertainties. Uncertainty is a field that could become increasingly useful for data users who plan to integrate data or use the data to train models. In some cases, providing details of uncertainty in data collection, measurement, or statistical measures can be more influential on the success of accurate integration than the quality of the data itself (Raupach et al. 2005; Merchant et al. 2017). Uncertainty information, as with many metadata fields, can be difficult for researchers to provide retroactively, especially as the period after publication increases. This highlights the importance of establishing a concrete metadata life cycle that runs parallel to the data production process (Habermann 2019).

The metadata reporting process can become intimidating if a scientist or project waits until data must be published in compliance with PDA policies implemented by journals. Metadata creation in this case becomes a barrier, and the rush to produce metadata sufficient for publication may result in poor quality or incomplete metadata records. By breaking down the metadata reporting process into phases that align with data collection, such as creating a profile of the geographic location or defining the purpose of planned research before data collection begins, could result in more complete metadata records. Therefore, the common definition of metadata as simply “data about data” may be limiting and outdated, as metadata can also describe multiple steps in the data collection workflow, from site selection to data analysis. Additionally, releasing information about the planned scope of data collection before data collection begins could help prevent the recreation of identical research by multiple projects (Habermann 2019).

## **FAIR Variability**

The majority of surveyed repositories included language on their website that connected their mission with the FAIR principles. Although the findability and accessibility characteristics were well represented, this survey of 15 environmental science repositories indicated a lower presence of requirements that fell into the interoperable reusable categories. However, these fields may be more appropriately fulfilled by other repository features outside of package level metadata

requirements. Interoperability and reusability are centered around machine readability and the promotion of integration, and while in package level metadata this may look like providing thorough methods descriptions and detailing any software used during data production, other requirements created by data repositories may be better fit for promoting these principles (Wilkerson et al. 2016; Mons et al. 2017).

File level metadata specific to certain data types could be more useful for the enablement of machine readability than package level metadata (Christianson et al. 2017). For example, workflows for the reporting of metadata for sample data have been implemented to supplement existing package level metadata and increase the FAIRness of these data types (Damerow et al. 2021). Requirements related to data file types have also become common, as data stored in proprietary file formats that rely on the use of specific software, such as Excel files, are at a higher risk of degradation if the software becomes obsolete (Michener et al. 1997; Mayernik et al. 2020).

Repository characteristics such as search capabilities and open access policies also play a role in the findability and accessibility of data. An efficient user interface allows data users to easily identify data relevant to their research. The findable and accessible qualities of data can also be impacted by a repository's engagement with the relevant scientific community, whether through workshops or 1:1 discussion with research projects. Additionally, providing thorough but clear information about the repository's purpose through "About" pages on websites and writing content in accessible language can further promote these FAIR principles.

## **Limitations and Future directions**

Although background information from each repository was collected for this research, a larger study with more repositories included is needed to make definitive relationships between metadata requirements and repository characteristics, such as size and funding background. For this research, although the funding categories of US government, international government, university, and NGO were established, the exact amount of funding for each repository was not collected. The majority of data repositories were funded by the US government through areas such as the Department of Energy or the National Science Foundation, however, two repositories with the same funding agencies could still have vastly different funding amounts.

Limitations to this research are also related to the increasing number of data repositories available to researchers, even within the environmental science area. Further research with larger sample sizes of data repositories could be better suited for identifying trends in metadata requirement variability.

Documenting metadata requirements outside of the 15 surveyed fell outside the scope of this research, so it is possible that some repositories may have had other requirements related to the FAIR principles that were not recorded. The results of this research therefore do not suggest that a repository that scored low on the presence of the 15 requirements has an underdeveloped metadata review process. Rather, the results indicate which repositories are missing *key* metadata fields that build the necessary context for data users for the reproduction and integration of data (Hillebrand and Gurevitch 2013). Additionally, this study did not dive into the specific content of each requirement. For example, two repositories may have the requirement of a data package abstract, however one may have the added *content* requirement of at least 100 words. Even if repositories move towards a core set of metadata fields present in all data packages, the exact wording of the content requirements may still differ because the inclusion of certain metadata ultimately requires judgement by those working in the subject area field and the maintainers of the repository.

## **Broader Implications**

Although further research efforts are necessary to illustrate clear trends in the types of metadata requirements implemented by environmental science data repositories, this study has identified key components that are missing from data repository landing pages, especially those that support interoperability and reusability. Establishing thorough metadata requirements is not the only strategy for increasing the findability, accessibility, interoperability and reusability of data, but the absence of this information could be detrimental to the ability of data users to perform the large scale meta-analyses that are becoming more common in the environmental science research area (Michener et al. 1997). As journals and funding organizations begin to implement policies that require the archival of data related to publications, it is vital that repositories help ensure these data packages are of high enough quality for reuse (Roche et al. 2015).

Furthermore, the lack of proper metadata describing the most basic steps of data collection impacts the ability of scientists to test the validity of published research and hinders the sharing and reuse of data by future generations in addition to the original investigators (Michener et al. 1997; Tenopir et al. 2011). Although the implementation of a thorough metadata requirements may demand the allocation of increased repository resources, such as focusing team efforts on the publication process, these quality standards benefit the original users, future data users, and the repositories themselves by increasing their trustworthiness and establishing a reputation of providing high quality data that will continue to contribute to research efforts far beyond their time of publication.

## ACKNOWLEDGEMENTS

Thank you to my team at Lawrence Berkeley National Laboratory for their guidance in throughout writing of this thesis and for introducing me to this research, especially my mentors Charuleka Varadharajan and Robert Crystal-Ornelas. To the ESPM175 teaching team, Patina Mendez, Kyle Leathers, and Leslie McGinnis: I cannot imagine a better group of individuals to cheer us on through two full semesters of online classes. Your motivation and passion never went unnoticed, even as you dragged us all across the finish line. Thank you for all of your trust, confidence, and understanding. To the “Big Pictures,” Isa Avila Breach, Sarah Bui, Phoebe Goulden, Sasha Mizenin, and Katie Wimsatt, thank you for the feedback, reassurance, and support you have provided over the past two semesters.

To my family, who have been my constant cheerleaders through my undergraduate career – being the first wasn’t easy, but your unwavering belief that I would see the other side of the tunnel kept me sane. Eva and Estella, you always have and you forever will be my greatest motivation in life. Alexander, thank you for inspiring me each day with your drive and ambition, for your optimism that I would always find a way to succeed, and for letting me ruin every attempt at healthy eating with my near constant stress baking. Lastly, to the goofiest and most entertaining pet, Normi, thank you for being a constant source of laughter and cuddles, and for nearly doubling your life expectancy to see me through my time at Cal.

## REFERENCES

- Baker, K. S. 2009. Data Stewardship: Environmental Data Curation and a Web of Repositories. *Digital Discourse: The e-evolution of Scholarly Communication* 1.
- Christianson, D. S., C. Varadharajan, B. Christoffersen, M. Detto, B. Faybishenko, B. O. Gimenez, V. Hendrix, K. J. Jardine, R. Negron-Juarez, G. Z. Pastorello, T. L. Powell, M. Sandesh, J. M. Warren, B. T. Wolfe, J. Q. Chambers, L. M. Kueppers, N. G. McDowell, and D. A. Agarwal. 2017. A metadata reporting framework (FRAMES) for synthesis of ecohydrological observations. *Ecological Informatics* 42:148–158.
- CoreTrustSeal Standards And Certification Board. 2019. CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022.
- Damerow, J. E., C. Varadharajan, K. Boye, E. L. Brodie, M. Burrus, K. D. Chadwick, R. Crystal-Ornelas, H. Elbashandy, R. J. E. Alves, K. S. Ely, A. E. Goldman, T. Haberman, V. Hendrix, Z. Kakalia, K. M. Kemner, A. B. Kersting, N. Merino, F. O'Brien, Z. Perzan, E. Robles, P. Sorensen, J. C. Stegen, R. L. Walls, P. Weisenhorn, M. Zavarin, and D. Agarwal. 2021. Sample Identifiers and Metadata to Support Data Management and Reuse in Multidisciplinary Ecosystem Sciences. *Data Science Journal* 20:11.
- Damerow, J., C. Varadharajan, K. Boye, E. Brodie, M. Burrus, D. Chadwick, S. Cholia, R. Crystal-Ornelas, H. Elbashandy, R. Eloy Alves, K. Ely, A. Goldman, V. Hendrix, C. Jones, M. Jones, Z. Kakalia, K. Kemner, A. Kersting, K. Maher, N. Merino, F. O'Brien, Z. Perzan, E. Robles, C. Snively, P. Sorensen, J. Stegen, P. Weisenhorn, K. Whitenack, M. Zavarin, and D. Agarwal. 2020. ESS-DIVE Global Sample Numbers and and Metadata Reporting Format for Environmental Systems Science (IGSN-ESS). *Environmental System Science Data Infrastructure for a Virtual Ecosystem; Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE)*.
- Edwards, P. N., M. S. Mayernik, A. L. Batcheller, G. C. Bowker, and C. L. Borgman. 2011. Science friction: Data, metadata, and collaboration. *Social Studies of Science* 41:667–690.
- Gordon, S., and T. Habermann. 2018. The influence of community recommendations on metadata completeness. *Ecological Informatics* 43:38–51.
- Habermann, T. 2018. Metadata Life Cycles, Use Cases and Hierarchies. *Geosciences* 8:179.
- Halbritter, A. H., H. J. De Boeck, A. E. Eycott, S. Reinsch, D. A. Robinson, S. Vicca, B. Berauer, C. T. Christiansen, M. Estiarte, J. M. Grünzweig, R. Gya, K. Hansen, A. Jentsch, H. Lee, S. Linder, J. Marshall, J. Peñuelas, I. Kappel Schmidt, E. Stuart-Haëntjens, P. Wilfahrt, the ClimMani Working Group, V. Vandvik, N. Abrantes, M. Almagro, I. H. J. Althuizen, I. C. Barrio, M. te Beest, C. Beier, I. Beil, Z. C. Berry, T. Birkemoe, J. W. Bjerke, B. Blonder, G. Blume-Werry, G. Bohrer, I. Campos, L. A. Cernusak, B. H. Chojnicki, B. J. Cosby, L. T. Dickman, I. Djukic, I. Filella, L. Fuchslueger, A. Gargallo-Garriga, M. A. K. Gillespie, G. R. Goldsmith, C. Gough, F. W. Halliday, S. Joar Hegland,

- G. Hoch, P. Holub, F. Jaroszynska, D. M. Johnson, S. B. Jones, P. Kardol, J. J. Keizer, K. Klem, H. S. Konestabo, J. Kreyling, G. Kröel-Dulay, S. M. Landhäusser, K. S. Larsen, N. Leblans, I. Lebron, M. M. Lehmann, J. J. Lembrechts, A. Lenz, A. Linstädter, J. Llusià, M. Macias-Fauria, A. V. Malyshev, P. Mänd, M. Marshall, A. M. Matheny, N. McDowell, I. C. Meier, F. C. Meinzer, S. T. Michaletz, M. L. Miller, L. Muffler, M. Oravec, I. Ostonen, A. Porcar-Castell, C. Preece, I. C. Prentice, D. Radujković, V. Ravolainen, R. Ribbons, J. C. Ruppert, L. Sack, J. Sardans, A. Schindlbacher, C. Scoffoni, B. D. Sigurdsson, S. Smart, S. W. Smith, F. Soper, J. D. M. Speed, A. Sverdrup-Thygeson, M. A. K. Sydenham, A. Taghizadeh-Toosi, R. J. Telford, K. Tielbörger, J. P. Töpfer, O. Urban, M. Ploeg, L. Van Langenhove, K. Večeřová, A. Ven, E. Verbruggen, U. Vik, R. Weigel, T. Wohlgemuth, L. K. Wood, J. Zinnert, and K. Zurba. 2020. The handbook for standardized field and laboratory measurements in terrestrial climate change experiments and observational studies (ClimEx). *Methods in Ecology and Evolution* 11:22–37.
- Hankin, S., L. Bermudez, J. D. Blower, B. Blumenthal, K. S. Casey, M. Fornwall, J. Graybeal, R. P. Guralnick, T. Habermann, E. Howlett, B. Keeley, R. Mendelssohn, R. Schlitzer, R. Signell, D. Snowden, and A. Woolf. 2010. Data Management for the Ocean Sciences - Perspectives for the Next Decade. Pages 570–579 *Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society*. European Space Agency.
- Hart, E. M., P. Barmby, D. LeBauer, F. Michonneau, S. Mount, P. Mulrooney, T. Poisot, K. H. Woo, N. B. Zimmerman, and J. W. Hollister. 2016. Ten Simple Rules for Digital Data Storage. *PLOS Computational Biology* 12:e1005097.
- Hillebrand, H., and J. Gurevitch. 2013. Reporting standards in experimental studies. *Ecology Letters* 16:1419–1420.
- Marcial, L. H., and B. M. Hemminger. 2010. Scientific data repositories on the Web: An initial survey. *Journal of the American Society for Information Science and Technology* 61:2029–2048.
- Mayernik, M. S., K. Breseman, R. R. Downs, R. Duerr, A. Garretson, and C.-Y. (Sophie) Hou. 2020. Risk Assessment for Scientific Data. *Data Science Journal* 19:10.
- Merchant, C. J., F. Paul, T. Popp, M. Ablain, S. Bontemps, P. Defourny, R. Hollmann, T. Lavergne, A. Laeng, G. de Leeuw, J. Mittaz, C. Poulsen, A. C. Povey, M. Reuter, S. Sathyendranath, S. Sandven, V. F. Sofieva, and W. Wagner. 2017. Uncertainty information in climate data records from Earth observation. *Earth System Science Data* 9:511–527.
- Michener, W. K., J. W. Brunt, J. J. Helly, T. B. Kirchner, and S. G. Stafford. 1997. Nongeospatial Metadata for the Ecological Sciences. *Ecological Applications* 7:330–342.
- Mons, B., C. Neylon, J. Velterop, M. Dumontier, L. O. B. da Silva Santos, and M. D. Wilkinson. 2017. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use* 37:49–56.

- O'Brien, F. 2019, December. Increasing Efficiency in Data Publication using Semi-Automated Workflow. AGU Fall Conference, San Francisco.
- Palmer, C., M. Cragin, P. Heidorn, and L. Smith. 2007. Data curation for the long tail of science: The case of environmental sciences.
- Pepe, A., A. Goodman, A. Muench, M. Crosas, and C. Erdmann. 2014. How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. *PLoS ONE* 9:e104798.
- Raupach, M. R., P. J. Rayner, D. J. Barrett, R. S. DeFries, M. Heimann, D. S. Ojima, S. Quegan, and C. C. Schmullius. 2005. Model–data synthesis in terrestrial carbon observation: methods, data requirements and data uncertainty specifications. *Global Change Biology* 11:378–397.
- Roche, D. G., L. E. B. Kruuk, R. Lanfear, and S. A. Binning. 2015. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLOS Biology* 13:e1002295.
- Tenopir, C., S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame. 2011. Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE* 6:e21101.
- Vines, T. H., A. Y. K. Albert, R. L. Andrew, F. Débarre, D. G. Bock, M. T. Franklin, K. J. Gilbert, J.-S. Moore, S. Renaut, and D. J. Rennison. 2014. The Availability of Research Data Declines Rapidly with Article Age. *Current Biology* 24:94–97.
- Wilkinson, M. D., M. Dumontier, Ij. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018.
- Wolkovich, E. M., J. Regetz, and M. I. O'Connor. 2012. Advances in global change research require open science by individual researchers. *Global Change Biology* 18:2102–2110.

**APPENDIX A: Repository Metadata Requirements****Table 4. Presence of metadata requirements in each repository.**

Repository ID	Title	Abstract	Keywords	Data variables	Usage rights	Related references	Funding source/Project	Data package contact	Temporal coverage	Geographic description
KNB	1	1	1	0	1	0	1	1	1	1
EDI	1	1	1	0	1	1	1	1	1	1
NCEI	1	1	1	0	0	1	0	1	1	0
DRY	1	1	1	0	1	1	1	0	0	0
ENVD	1	1	1	0	1	1	1	1	0	0
PANGAEA	1	1	1	0	0	1	1	0	1	0
TERN	1	1	1	1	1	0	1	1	1	1
ORNLDAAC	1	1	1	0	1	1	1	0	1	0
NCAR	1	1	0	1	1	1	1	1	1	0
AODN	1	1	1	1	1	1	1	1	1	1
CUGIR	1	1	1	1	0	1	0	0	1	1
USGS	1	1	1	0	1	1	1	1	1	1
ADC	1	1	1	1	1	0	1	1	1	1
ESS	1	1	1	1	1	1	1	1	1	1
BCO	1	1	0	1	1	1	1	1	1	0

Coordinates	Data package authors	Methods	Persistent Identifier Field	Citation Available
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
0	1	1	1	1
1	1	0	1	1
1	1	0	0	1
1	1	1	1	1
1	1	1	1	1
0	1	0	0	1
1	1	0	1	0
1	1	0	0	0
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1

**APPENDIX B: Metadata Requirement Totals by Repository****Table 5.** Total scores indicate the number of core requirements present in each repository.

<b>Repository ID</b>	<b>Total Score (Number of requirements present)</b>
KNB	13
EDI	14
NCEI	11
DRY	10
ENVD	11
PANGAEA	9
TERN	14
ORNLDAAC	12
NCAR	10
AODN	13
CUGIR	9
USGS	14
ADC	14
ESS	15
BCO	13