The Environmental Data Initiative: connecting the past to the future through data reuse

Corinna Gries[1], Paul Hanson[1], Margaret O'Brien[2], Mark Servilla[3], Kristin Vanderbilt[4], and Robert Waide[3]

[1]University of Wisconsin-Madison
[2]University of California Santa Barbara
[3]University of New Mexico College of Arts and Sciences
[4]National Park Service

October 8, 2022

Abstract

1. The Environmental Data Initiative (EDI) is a trustworthy, stable data repository and data management support organization for the environmental scientist. In a bottom-up community process EDI was built with the premise that freely and easily available data are necessary to advance the understanding of complex environmental processes and change, to improve transparency of research results, and to democratize ecological research. 2. EDI provides tools and support that allow the environmental researcher to easily integrate data publishing into the research workflow. 3. Almost ten years since going into production, we analyze metadata to provide a general description of EDI's collection of data and its data management philosophy and placement in the repository landscape. We discuss how comprehensive metadata and the repository infrastructure lead to highly findable, accessible, interoperable, and reusable (FAIR) data by evaluating compliance with specific community proposed FAIR criteria. 4. Finally, we review measures and patterns of data (re)use, assuring that EDI is fulfilling its stated premise.

The Environmental Data Initiative: connecting the past to the future through data reuse

Authors and affiliation:

Corinna Gries, Center for Limnology, University of Wisconsin, Madison, Wisconsin 53706, USA, cgries@wisc.edu (corresponding author)

Paul C Hanson, Center for Limnology, University of Wisconsin, Madison, Wisconsin 53706, USA, pchanson@wisc.edu

Margaret O'Brien, Marine Science Institute, University of California, Santa Barbara, CA 93106, USA, margaret.obrien@ucsb.edu

Mark Servilla, Department of Biology, University of New Mexico, Albuquerque, New Mexico 87131 USA, mark.servilla@gmail.com

Kristin Vanderbilt, Department of Biology, University of New Mexico, Albuquerque, New Mexico 87131 USA, krvander@fiu.edu

Robert Waide, Department of Biology, University of New Mexico, Albuquerque, New Mexico 87131 USA. rbwaide@unm.edu

Abstract

1. The Environmental Data Initiative (EDI) is a trustworthy, stable data repository and data management support organization for the environmental scientist. In a bottom-up community process EDI was built with the premise that freely and easily available data are necessary to advance the understanding of complex environmental processes and change, to improve transparency of research results, and to democratize ecological research.
2. EDI provides tools and support that allow the environmental researcher to easily integrate data publishing into the research workflow.
3. Almost ten years since going into production, we analyze metadata to provide a general description of EDI's collection of data and its data management philosophy and placement in the repository landscape. We discuss how comprehensive metadata and the repository infrastructure lead to highly findable, accessible, interoperable, and reusable (FAIR) data by evaluating compliance with specific community proposed FAIR criteria.
4. Finally, we review measures and patterns of data (re)use, assuring that EDI is fulfilling its stated premise.

Keywords: data reuse, environmental data repository, FAIR data, metadata, open science

Introduction

Domain-specific data repositories provide services that directly support certain communities of practice or disciplines. They often cater to the needs of that community by archiving and making available data that are of interest, in formats that are usable, and through interfaces that are accessible to the community. A National Science Board refers to these services as “essential, community-proxy functions” (National Science Board, 2005). In turn, the community supports and builds trust in the repository and its content and relies upon it to publish primary data and as a source of data repurposed to answer new scientific questions, either in its original form or combined into a synthetic product or meta-analysis. Data published in a trustworthy and accessible repository provide significant benefits to scientific progress (Hampton et al., 2013), society in general, and the careers and research of individual scientists (Eisenstein, 2022). Evaluating the connection between metadata quality and data reuse will help inform the role of data repositories in the future of ecological science.

The Environmental Data Initiative (EDI) is a domain-specific data repository that was designed for and with input from the environmental and ecological research communities. It was founded in 2016 as a successor to the Long-Term Ecological Research (LTER) Network Information System (NIS) (Servilla et al., 2016) now serving the environmental research community worldwide. The unit of publication in EDI is a “data package”, which consists of data, the metadata, and a quality report. The data may consist of one or more digital files (e.g., tables, spatial raster images and vectors, binary objects, documents, or software code). We distinguish a data package from a dataset by formally including the metadata and quality report as part of the aggregate package in addition to the data. A dataset (Chapman et al., 2020), on the other hand, is often an abstract collection of data files that may or may not include metadata or any other ancillary products relevant to the collection. A data package may undergo an ordered set of revisions, where each revision is an immutable digital snapshot of the data package at the time it was published. The set of revised data packages is called a series. Each data package revision is issued a Digital Object Identifier (DOI), which is registered with DataCite (Brase, 2010), along with a subset of the metadata. Revision-based DOIs not only improve the reuse of data (Groth et al., 2020) but also facilitate the reproducibility of research results that are based on data created at a specific date and time.

EDI has an established data archive of 45,000 unique series (composed of 80,500 individual data packages) containing about 405,000 digital data files and continues to grow in volume. Many data are from early, one-time efforts of the NSF LTER program (EcoTrends synthesis project (Peters et al., 2013) and Landsat imagery), collectively known as the "early collections." The "main collection" is composed of 9,000 unique series (about 30,000 data packages), with new and revised packages added regularly. Contributions to the main collection are from roughly 4,000 scientists and are curated primarily with support from professional information managers at EDI, LTER and other research sites. Data contributions to the EDI data repository have achieved a steady-state growth of roughly 3,000 contributions per year since 2016 with the greatest number being added in the last two years.

Data are described by detailed metadata encoded in the Ecological Metadata Language (EML) standard (Jones et al., 2019a) and must pass a rigorous quality assessment before being published to the repository following community recommendations for best practices (Whitlock, 2011, Goodman et al., 2014, Roche et al., 2015, Briney et al., 2020, Contaxis et al., 2022, Hanisch et al., 2022). Although requirements to fulfill a basic EML document are minimal, EDI's user community agreed on requiring much broader and in-depth metadata for any data to be archived and published as part of the main collection. For example, EDI metadata must include discovery-level information (e.g., title, abstract, creators, and organizations) as well as physical information about the data (e.g., file name, format, size, access location) and attribute-level information about data tables (e.g., column name, data type, data range, units of measurement). Data packages that lack required metadata or whose metadata is not on parity with the data are prevented from submission to the repository. Rules encoded in software that evaluate the metadata and data for quality and consistency enforce this mandate. This evaluation generates a “quality report” that is included as part of the final data package for a successful evaluation but is also available for review if the evaluation fails (O’Brien et al., 2016).

Because requirements for metadata vary across data repositories (Wilkinson et al., 2016), it is valuable to see where EDI falls within a spectrum of other repositories when ease of discovery and reusability of data are plotted against repository requirements for metadata richness, data formatting or specialization of submitted data (Fig. 1). Typically, when metadata and data requirements are stringent, data are easier to find and use. EDI is positioned near the center of this correlation. By requiring more metadata than generalist repositories (but without stringent formats), EDI still provides sufficient information for consumers to determine fitness-of-use and reuse of archived data.

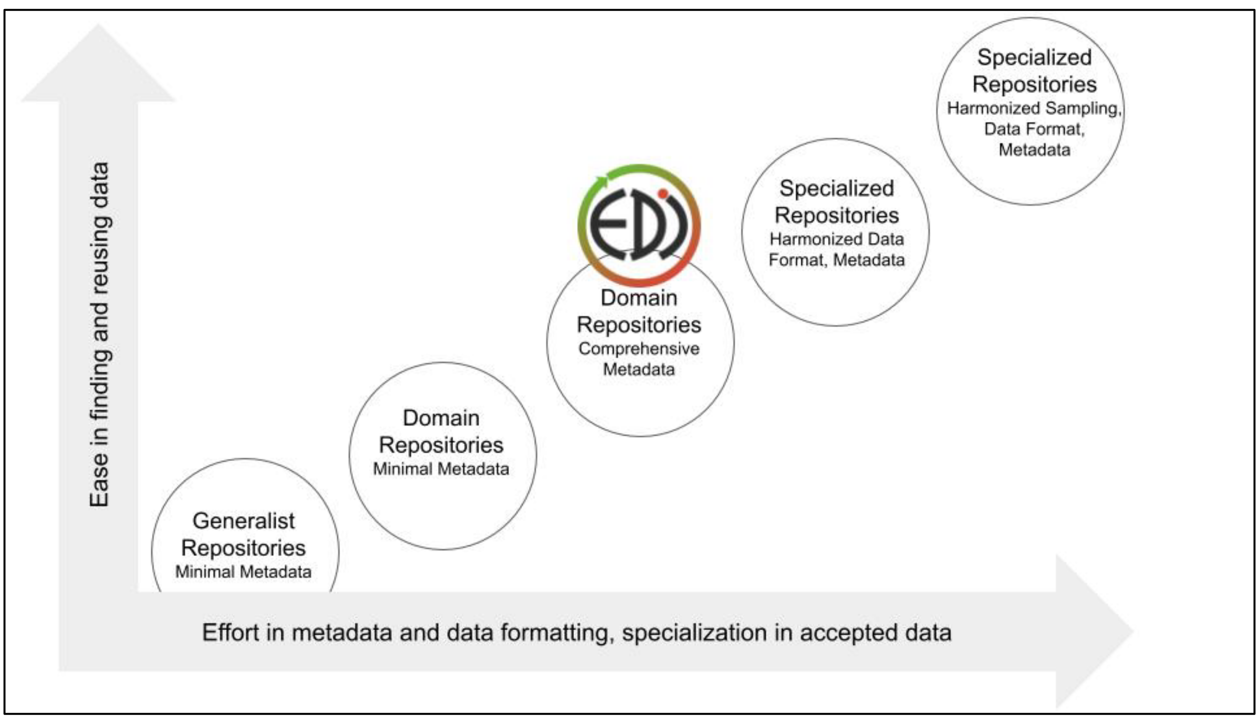

Figure 1: Characteristics of data repositories are plotted qualitatively along axes representing ease of data discovery and reuse versus the perceived effort to create semantically rich metadata or formatted data of a specific type.

EDI simplifies the creation of rich metadata by providing a simple, highly automated, online metadata editor, ezEML (Vanderbilt et al., 2022) and professional curation services. EDI data curators are available to counsel users on best practices in data organization, documentation, and ethical publication practices (Puebla et al., 2021), including procedures to help identify and anonymize sensitive data (e.g., human subject or endangered species data) prior to publishing.

Data submitted to EDI are	Findable	Accessible	Interoperable	Reusable
How it's done in EDI	Highly automated metadata editor Metadata standardization Superior search capability based on extensive science metadata EDI portal DataOne portal Search engine optimized (SEO)	Repository infrastructure optimized for environmental data Custom portals (specific subset of data packages)	Non-binary or community standard file types recommended Community developed best practice recommendations	Quality checks during data submission assure sufficient metadata to determine fitness for use and documentation of data collection context
	EDI staff is available to help planning, data cleaning and formatting, metadata content			
	EDI applications can automate submission, search, download, data use analysis (REST API, R package EDIutils, access code generation)			

Figure 2:Services and approaches provided by EDI to provide optimal reusability of published data packages.

After a decade of repository operations and four decades of organized Information Management experience in the community served by EDI, we are taking stock of the data collection managed by EDI (specifically, the "main collection"). We explore the variability of data within the repository by classifying descriptive attributes found in associated metadata and by analyzing how these attributes stack up against FAIR (Findable, Accessible, Interoperable, and Reusable) criteria (Wilkinson et al., 2016). We then review indications of data reuse by analyzing download statistics and formal data citations found in scientific publications as reported by Google Scholar (and other means). Finally, we discuss how openly available and well documented data have enabled the ecological community to ask and answer important new questions.

Methods

Three primary sets of data were analyzed: the first consists of the EML metadata that accompanies each data package in EDI's main collection; the second is a summary of download events for individual data files; and the third consists of citations of data archived in the EDI repository obtained by a Google Scholar search.

EDI's data collection and FAIR analysis

There is no universal definition of a data package (Lowenberg et al., 2019), nor even within a community does complete agreement exist (Gries et al., 2021) which has ramifications for the following analyses. In environmental sciences, it is important that data packages are designed to document the context of a specific research project and data collection with metadata, data, and code. Hence, in some cases, a data package encompasses a combination of thematically different observations that are needed to fully comprehend the context of a particular research study (e.g., the abiotic conditions during sampling and concurrent observations of the biota). Alternatively, data may be separated into several data packages according to different aspects of a study.

Following the above example, one package may contain meteorological data while a different package contains observations of the biota. In other cases, observations taken over time may be published as a single data series that is regularly updated and versioned (i.e., a series), or as separate packages for each observation period (e.g., annually). Similarly, observations spanning more than one location may be split into different data packages along spatial criteria. High-volume data may also be separated into individual packages to simplify management, download and processing. This heterogeneity should be considered when interpreting the following analyses, which are based on numbers of data series.

Metadata for the approximately 9,000 data series in EDI's main collection (data package of the newest revision were used) were analyzed for specific attributes, including keywords, start and end dates of the data collection period, and the sampling locations. Analysis was performed by using the R statistical programming language to parse and record attribute information from the metadata. This information was then recorded into a corresponding table of key-value pairs for keyword analysis or into time-period bins for temporal analysis or into latitude/longitude pairs for spatial analysis. These data and the R source code are published in the EDI data repository (Gries and Servilla, 2022).

The set of metadata was then processed to determine compliance with criteria identified as being representative of FAIR data. The two sources of FAIR criteria used in this analysis are the FAIR Data Maturity Model proposed by Bahim et al. (2020) and the MetaDIG criteria (Jones and Slaughter, 2019) adopted by DataONE. A detailed discussion of how FAIR criteria were mapped to EML attributes may be found in Gries (2022). In total, 46 criteria combined from each approach were analyzed to determine their presence in EDI's metadata. Again, this analysis was performed by using R, with results being recorded into criteria-based bins.

Download Events

Download “request” events for data files were obtained from the repository audit system database. These events are annotated with the downloaded data file identifier, an event date-timestamp, and the requesting HTTP User-Agent record. To analyze only user-initiated requests for data files, download events that did not contain a valid User-Agent record (i.e., the record was null or contained non-identifiable content) were excluded. The User-Agent record was used to categorize the originating actor of the request as either a “robot”, “human”, or “program”. Download events identified as a “robot” (i.e., initiated by a search engine or other web crawler) were filtered out by matching the string content found in the HTTP User-Agent record with known robot string patterns that are published by the Make Data Count project (Cousijn et al., 2019). The remaining download events were further labeled, also based on the User-Agent strings, as either “human” (i.e., initiated through a web browser) or “program” (i.e., initiated by a computer program). Human requests for data were identified by matching the User-Agent string to known web browser labels, while program requests were identified by User-Agent strings that are associated with the programming environment being used to access the repository web-service API. The approach used to identify robots in this research is not foolproof but does serve the needs of this analysis.

Using the above approach, download events for 2021 were filtered and categorized. Of nearly 3 million download events, 180,000 were identified as either human or program-initiated requests for data. Each download event record lists the data entity which was used to identify the corresponding data package from which data were downloaded. Once the data package is known, its metadata were analyzed to determine the thematic classification of the data and temporal ranges of data-collection time spans.

Data Citations

Journal citations for data series were collected by using Google Scholar to search for the "shoulder" of the data package DOI, which is a unique substring found at the start of all DOIs registered to EDI. A small number of "citations" not found by Google Scholar were added based on author assurance of data package use. The set of citations was restricted to the years 2013 through 2021. Although a formal data citation includes a DOI which points to a specific version within a data series, citations were combined for each series in the main collection. The validity of data package citations was confirmed by accessing the publication through the University of Wisconsin library system. A total of 2,595 data package citations were found. Similar to download events, the data package citations were summed into bins based on the data package identifier and again used as proxies for the reuse of thematic and time-span data.

Results

EDI's Main Collection of Data

EDI houses valuable long-term ecological observations with almost 30% of data series having observations covering 10 or more years (Fig. 3). Some short-duration data packages (e.g., classified as "1 year") are part of longer-term observation, but were published in smaller increments (see Methods). Data packages with tree-ring analyses, modeling results, and records of duration of ice cover provide data records for well over 500 years.

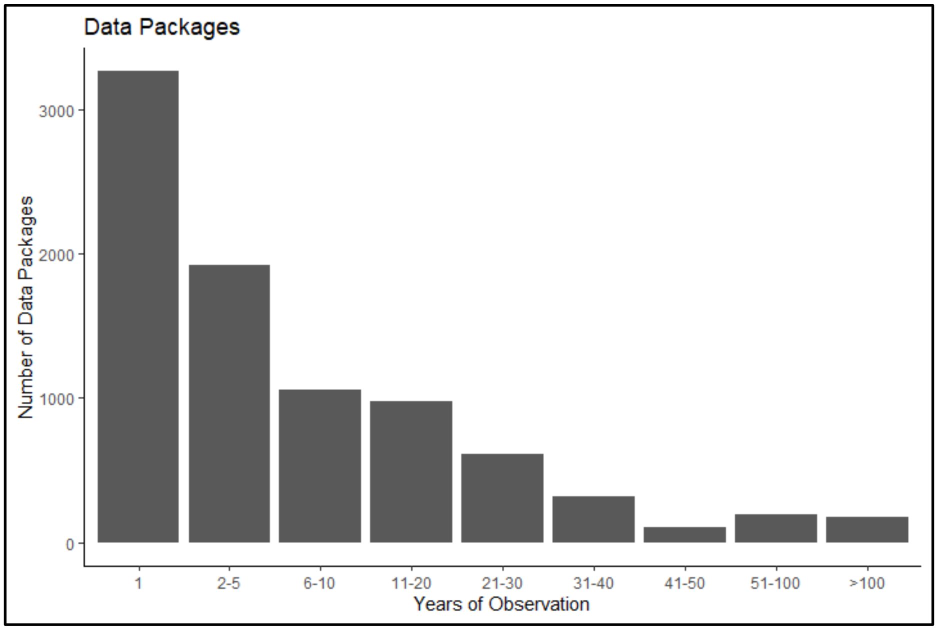

Figure 3:Number of data packages (newest revision within each series) per length of observation in years.

EML metadata include sampling locations as a bounding box or as a list of discrete point locations. Fig. 4 shows sampling locations (or bounding box centroids) for 8500 (97%) data series that provide geographic coverage. Centroids for bounding boxes that span northern Europe and North America appear in the North Atlantic. The EDI repository contains data from all over the world but with a strong emphasis on the US research community. In addition to data packages submitted by international contributors, a wide range of sampling locations can be found in large data products that synthesize many local data packages.

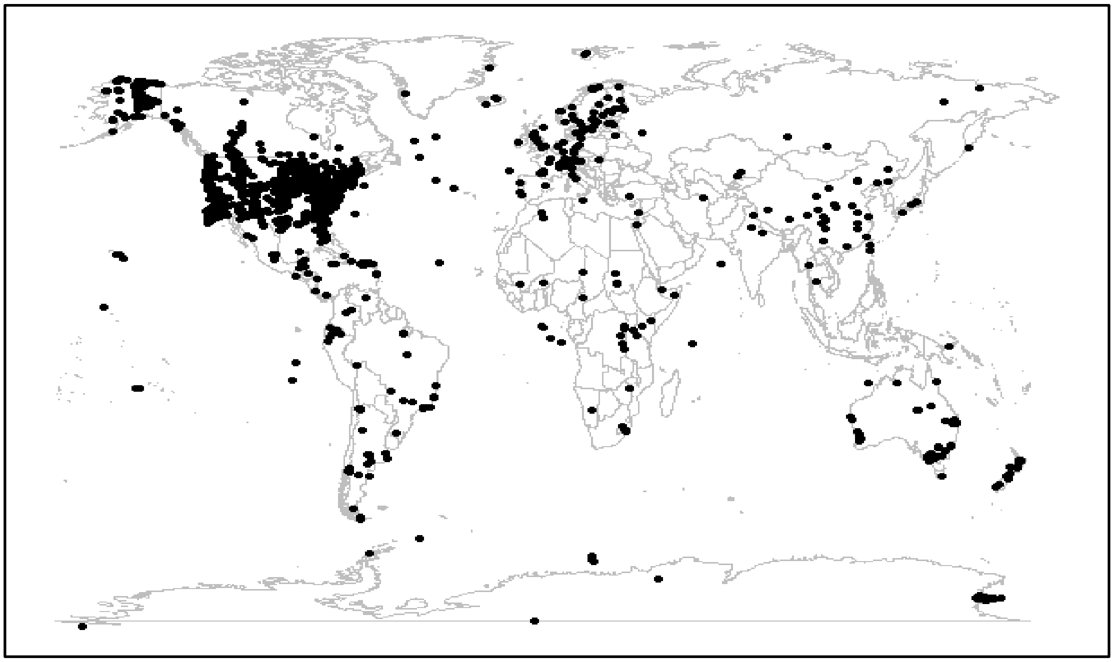

Figure 4: Sampling locations as detailed in metadata, for bounding boxes a centroid was calculated.

The broad subject areas of data in EDI's main collection reflect the complexities of environmental research and are best depicted in an analysis of keywords used by authors in describing their data packages. The 200 most frequently applied keywords are displayed in a word cloud in Fig. 5. Members of the LTER network (EDI's largest contributor) are required to collect data in five core areas: "disturbance", "primary productivity", "populations", "inorganic nutrients", "organic matter". As such, these keywords dominate the word cloud, along with common environmental drivers, like "temperature."

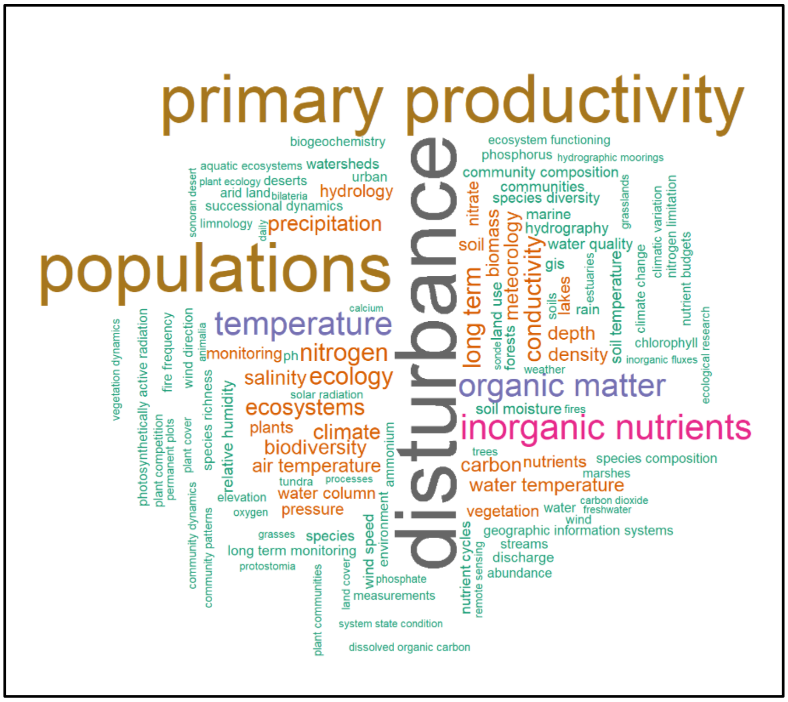

Figure 5: Word cloud of 200 most frequently used keywords to describe research subject of data packages.

A network analysis of the most commonly used keywords further shows how frequently they are used together to describe a single data package (e.g., “primary productivity” and “disturbance” are used together in 11%, “populations” and “disturbance” in 9% of data packages). This overlap in research themes within single data packages denotes the practice of collecting and publishing data of different topics. As described in the discussion of data package design in the methods, observations of organisms and measurements of their abiotic environmental conditions are frequently used to explain organismal behavior. However, each of those observations may very well be reused separately in a meta-analysis. This analysis of keyword grouping further highlights that keywords are often assigned by the data provider without any further requirements for harmonization between projects, therefore the practice of assigning different words for similar concepts is very common. These practices and possible improvements have significant impact on the discoverability of data (Porter, 2019).

Combining the basic count of keyword use, the analysis of keywords used most frequently together, and expert knowledge, we identified groups of keywords that appeared to be describing environmental research areas in their broadest scopes for which data package series are published in EDI. For instance, we expanded the concept of ‘populations’ to ‘biodiversity’ and included data packages with keywords: diversity, community, population, species, density, abundance, competition, cover, organism, habitat, restoration, distribution, plot, inventory, vegetation, fauna, microbe, survey, succession, biota, predation. We also added the concept of ‘abiotic conditions’ which includes the frequently used terms: temperature, precipitation, snow, irradiance, ice, climate, meteorology, waves, radiation, rain, weather, PAR, hydrology, moisture, physical, discharge, elevation. Any single data package may be classified as belonging to more than one thematic area. The group of ‘Not Themed’ data packages is either lacking keywords or cannot be assigned to any of the other environmental themes (e.g., a very few are solely human subject related data). The number of data packages in EDI’s main collection is fairly evenly distributed across these large themes (Fig. 6) with abiotic conditions and biodiversity leading in number of data packages.

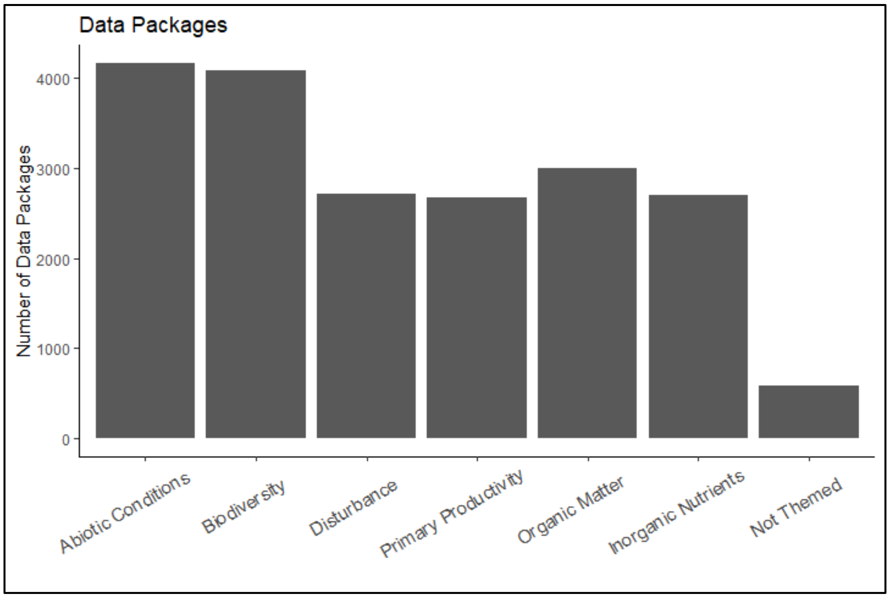

Figure 6: Number of data packages (newest revision within each series) within each major research subject area, as determined by keyword analysis.

FAIR Ranking of Data Packages

Analyzing metadata quality using the newly developed and more specific criteria for evaluating a data package's degree of FAIR implementation clearly shows that the majority of data packages in EDI's repository score high on many of the FAIR criteria (Fig. 7). Most criteria (over 70%) under Findable and Accessible are either checked for upon data submission or the metadata are increasingly inserted automatically by EDI. The most obvious exceptions (fewer than 50% of data packages pass) are criteria that do not apply to all data packages (e.g., taxonomic coverage), plus the adoption, acquisition and use of IDs in metadata (e.g., ORCID for data package authors, Research Organization Registry, ROR ID for institutions and projects). These identifiers are relatively new (e.g., ROR IDs have only recently been assigned for LTER projects) and the practice of obtaining and integrating them into metadata will slowly improve.

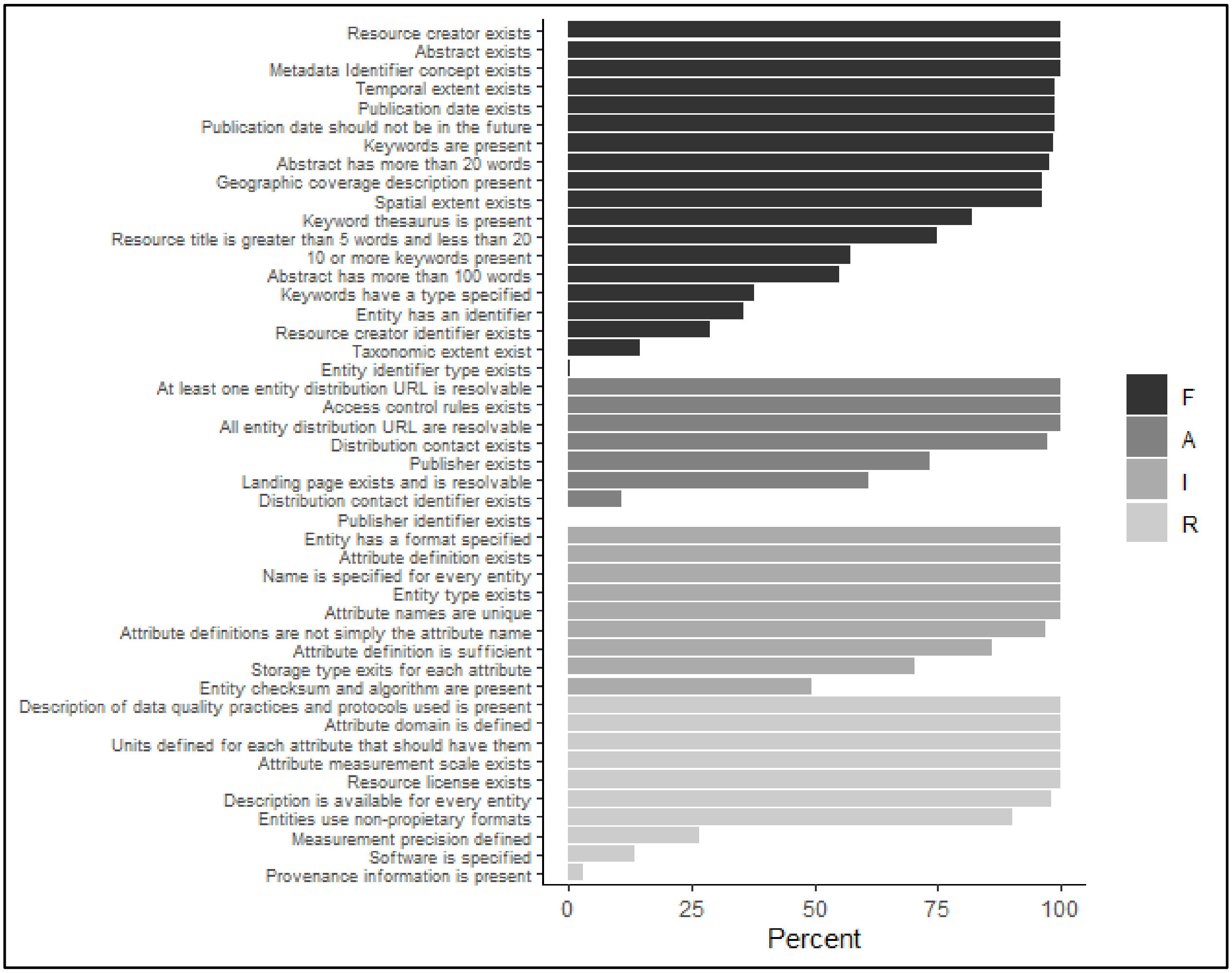

Figure 7: Compliance with a given quality measure in percent of all measured units in EDI's main collection, i.e., measures for data package quality is given as percent of all data packages in EDI's main collection, measure for data entities as percent of data entities, and measures for table attributes as percent of attributes.

In the areas of Interoperability and Reusability, EDI's metadata comply well with criteria suggested by Jones and Slaughter (2019) with the exception of specific data provenance information, measures of data quality and precision. The two lowest categories under 'Reusable' 'provenance information present' and 'software is specified' in Fig. 7 are mainly needed for documenting the generation of synthesis data products (see discussion). The majority of data in EDI are original observations where this does not apply. General provenance information may be found in several places in the metadata. Foremost, provenance information is detailed in the method description that is present in most data packages. Documenting data precision and quality, however, is a concern to data users that is currently not addressed by data contributors.

Data Downloads and Data Citations

By subject (Fig. 8) or time (Fig. 9), the majority of data downloads occurred manually via browser. It should be noted that because a script automates data access, it is likely to execute and record data access many times before the final data analysis is actually happening, which would inflate the importance of that download fraction.

A total of 2,595 citations of 1,563 unique data packages were recorded from 1,382 unique publications. Citations per publication ranged from 1-33 data series, and single data series were cited in 1-25 publications. While it can be assumed that most data series in EDI have been used in at least one publication or thesis, formal documentation of such use accounts only for about 18% of data series in EDI's main collection. The practice of formally citing data packages in publications is rapidly gaining popularity, though, with journals starting to require that data are available in a public repository and a data availability statement be included in the publication. Accordingly, the number of publications containing formal citations of data published in EDI have increased from 13 to over 400 annually between 2013 and 2021.

Given all caveats, the following data analysis does show very important patterns of data use. First, it does not appear that any particular research theme dominates data usage for either measure, download and citation (Fig. 8).

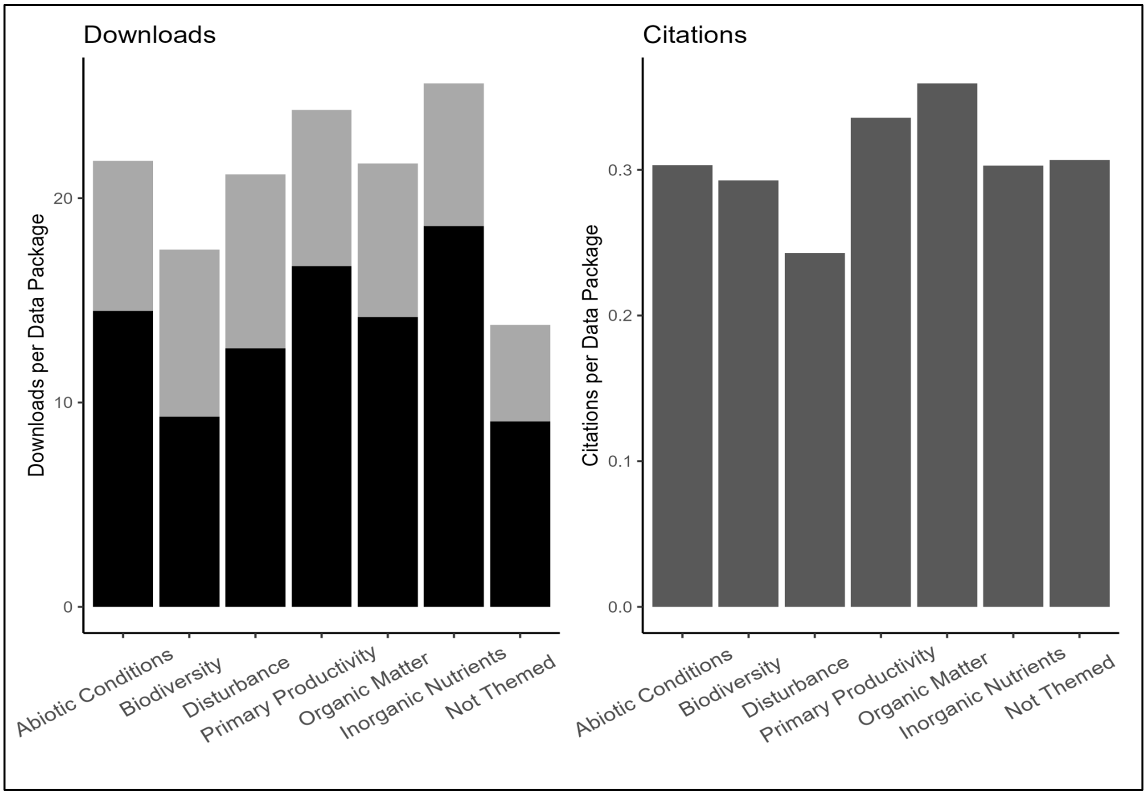

Figure 8: Data downloads (left) and citations (right) data package in category. Categories are major research themes as determined by author assigned keywords. For Downloads, gray = program and black = human.

However, when comparing data use by length of observation, long-term data packages are being used proportionally more frequently than short-term data packages. Another interesting result is that download numbers are particularly low for data packages providing observation for only one year (Fig. 9).

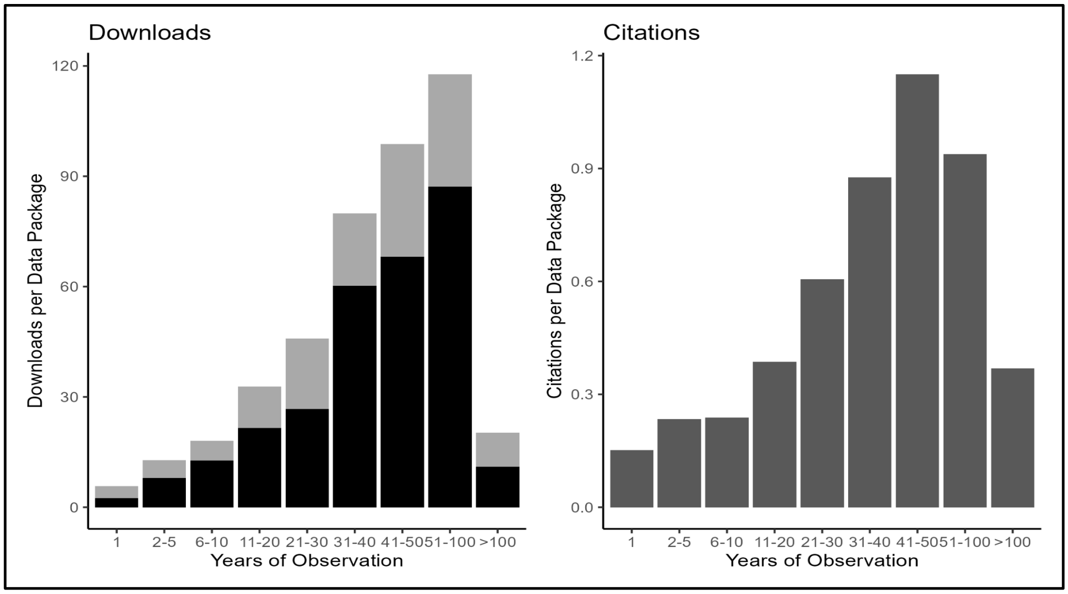

Figure 9: Data downloads (left) and Citations (right) per data package in duration in years bin. For Downloads, black = human and gray = program)

To further explore the impact of publicly available data packages, we retrieved citation indexes for each journal article citing a data package and the impact factors for the journals which range from 0 to 590 and 0.5 to 50 (Web of Science, 2021), respectively.

Discussion

EDI provides access to data from the ‘long-tail’ of environmental research and a large proportion of the data are long-term monitoring efforts in most environmental research areas. The distribution of reported data collections is worldwide with emphasis on North America. Our examination of the subject areas covered by dataset keywording entailed manual analysis that relied on EDI’s expert knowledge of the research fields covered by data packages. This work could have been accelerated had the use of controlled vocabularies supported by ontology and related technologies been embraced earlier. However, EDI and its data management community are gearing up to retrospectively implement more meaningful annotations to the metadata. Developing community endorsed vocabularies and ontologies (e.g., Buttigieg et al., 2016) show great promise for linking data both within and across scientific domains and improving findability and interoperability of the data.

Our FAIR analysis addresses the utmost importance of carefully documenting the context in which data were collected, which has long been recognized in environmental research (Catford et al., 2022) and has important ramifications for metadata and the makeup of data in a data package (Lowenberg et al., 2019, Gries et al., 2021). Some of the RDA and DataONE criteria used for our FAIR evaluation are enforced by constraints in the EML XML schema. Furthermore, metadata content was collaboratively improved by the data providers since the data repository went into production in 2013 resulting in the development of the EML congruence checker (O'Brien et al., 2016), continuous improvements to the repository infrastructure, and its metadata editor, ezEML (Vanderbilt et al., 2022). Upon submission, all metadata and data files are passed through the EDI congruence checker, which compares metadata to data structures. By implementing the EML standard and developing community endorsed best practices, data in the EDI repository are inherently FAIR and were so long before the term was coined (Jones et al., 2019b).

In addition to the FAIR criteria recommendations used here, several data user interviews (Kratz and Strasser, 2015, Schmidt et al., 2016, Gregory et al., 2020) have identified a number of high-priority criteria for evaluating the fitness for use of open data, some of which align well with the reported FAIR criteria and EDI's mission. Free access, ease of access, data coverage, and adequate metadata rank high. Open data users do not expect a data package review process (Kratz and Strasser, 2015), but also consider transparency of collection and processing methods, lack of data errors, or reputation of the data creator important when determining fitness for use of a data package. These criteria are difficult to judge reliably and report without human input. FAIR criteria suggested by Jones and Slaughter (2019) are designed to be machine-actionable and are mostly evaluating metadata completeness and not content. Hence, our FAIR analysis evaluates the existence and length of a method and other descriptive elements in the metadata but cannot judge the completeness or quality of such descriptions provided. Reporting use for data packages (downloads and citations) will be the best proxy indicator for these qualitative criteria.

Not addressed in the FAIR analysis are Bahim et al. (2020) recommendations of using machine-understandable knowledge representation for data, community data models, and FAIR-compliant vocabularies. Given EDI's primary goals, (and hence position in the curation effort vs. usability diagram, Fig. 1), achieving higher ratings for criteria related to machine readability would require a major effort and expense. However, in collaboration with the research community, EDI increasingly hosts data in community-developed standardized formats (Vanderbilt and Gries, 2021, O'Brien et al., 2021).

Standards in reporting and analyzing data use are still a developing area and are strongly influenced by community practices (Lowenberg et al., 2019). EDI serves data communities (Cooper and Springer, 2019) within larger, place-based, cross-institutional environmental research programs (e.g., LTER sites, biological field stations, California Interagency Ecological Program). These data communities are marked by their early recognition of the value of data sharing and comprehensive metadata, expert data management support, and a bottom-up development of data management infrastructure (Gries et al., 2016, Kaplan et al., 2021, Stafford, 2021), leading to the EDI repository of today with a well-defined scope and mission (Servilla et al., 2016). These communities are composed of thousands of researchers, representing both data providers and users, plus research collaborators. These communities are central to EDI, a feature not typically exhibited by generic repositories (Fig. 1, left) or those focused mainly on aggregation and harmonization of specific data (Fig. 1, right).

For example, for more than 40 years, observational data packages now available in EDI were used repeatedly within their respective data communities but without formal acknowledgment. The LTER program reports over 25,000 published products (https://www.zotero.org/groups/2055673/lter_network/library) (~19,000 peer-reviewed journal articles). It can safely be assumed that most of these products are directly using data now available in the EDI repository or are building on the knowledge gained from these data.

It should be noted that throughout this study, we report total data use, and do not distinguish between primary use and reuse. Although there are several definitions for data reuse in the literature (Pasquetto et al., 2017), we are following the guidance of van de Sandt et al. (2019), who after extensive research into definitions plus modeling of data use scenarios, concluded that ‘data use’ is the most accurate way to describe all uses of a research resource in a very complex, nonlinear, and evolving open research environment.

Such nonlinear use of new and existing data is well established in synthesis science, which has been strongly promoted through the establishment of Synthesis Centers (Baron et al., 2017) over the last 25 years. Synthesis research is considered highly important in environmental science (Carpenter et al., 2009) addressing complex questions at broad scales (e.g., Wieder et al., 2021) with long-term observations proving critical to the understanding of drivers of environmental change and its implications (e.g., Patel et al., 2021). Synthesis involves meta-analyses, reviews, new combinations of existing data, and advances in statistical methods (Collins, 2020). In addition to making effective use of existing data, synthesis research leads to novel insights and provides usable information for decision-makers (Hackett et al., 2008). Although data products from several such synthesis efforts have been published in the EDI repository (e.g., Collins et al., 2018, Soranno et al., 2019, Wieder et al., 2020), other synthesis studies have not formally cited data packages that are published in EDI (Batt et al., 2017, Li and Pennings, 2016) but are assuring data use in other ways. In a recent study documenting the importance of such data use in advancing knowledge, Halpern et al., (2020) found a five-fold higher citation rate for synthesis publications compared to the broader ecological literature.

In addition to data downloads and citations, EDI provides the option to document data use in the form of specific provenance information in the metadata along with processing scripts. This formal encoding of data used to develop a synthesis data product can handle many more data ‘citations’ (links) than a regular journal publication would, and documents decisions made during data preparation (AlNoamany and Borghi, 2018, Brinckman et al., 2019). For instance, the above-mentioned data package by Soranno et al. (2019) documents 90 data packages that were used to synthesize it. Furthermore, Soranno et al. (2019) has been used to create the data package by Cheruvelil et al., (2022). One of the articles citing an earlier version of the Soranno et al. data package is what is called a ‘data paper’ (Belter, 2014, Kratz and Strasser, 2014), i.e., a journal article style discussion of the metadata for and content of a data package. This data paper (Soranno et al., 2017) in turn has been cited over 80 times. Hence, we see formal citations of the data package DOI and the data paper DOI both may indicate data use. This short discourse on the complexities of data package use shows that the research community needs more extensive data use reporting and the difference between use and reuse is almost impossible to determine or measure.

Although complex, the above examples of data use are documented and therefore transparent. They may be discovered by citation indexes and machine-readable metadata. Many data uses cannot be traced, however, and evaluating data downloads as a proxy is the only viable approach. EDI provides unfettered access to data (no login or registration is required) and does not ask a user to specify what the intended application of the data will be. Based on survey results by Gregory et al. (2020) other uses include data for teaching and exploring (and discarding) new ideas, and these are not likely to ever have a mechanism for formal documentation and reporting.

Conclusion

Studying the highly complex living environment to understand its connections and drivers and monitor and document its changes requires a multidisciplinary research endeavor. Although data sharing and reuse has become integral to advancing knowledge in environmental science, data stewardship and enabling such reuse are still in the early stages of socio-technical inventions (Michener, 2015). However, it is recognized that data publishing improves the scientific enterprise (McKiernan et al., 2016) by increasing transparency and reproducibility of published results (Roche et al., 2015, 2021, Borghi and Van Gulick, 2021) and encouraging new collaborations (e.g., Boland et al., 2017, Walter et al., 2021).

EDI is a data repository and data management support organization providing the environmental research community with a stable platform of well documented and, hence, reusable data. As the open data landscape is changing toward data publishing requirements to increase transparency and reproducibility of scientific results (Roche et al., 2021) EDI provides tools and support to streamline publication workflows and review processes (e.g., Fox et al., 2021). The current rapid and dramatic environmental changes in particular, increasingly prompt researchers to publish and seek historic observations for comparison and context in EDI.

Acknowledgements

This work was supported by National Science Foundation awards #1931143 and #1931174.

Conflict of Interest Statement

All authors declare no conflict of interest.

Author contributions

Gries, conception and design, acquisition of data, analysis and interpretation of data, drafting the article, revising it critically for important intellectual content

Servilla, acquisition of data, revising it critically for important intellectual content

Hanson, O’Brien, Vanderbilt, Waide, revising it critically for important intellectual content

Data Availability statement

Gries, C. and M. Servilla. 2022. Data and code for EDI overview paper, data collection characteristics, FAIR evaluation, downloads, and citations ver 1. Environmental Data Initiative. https://doi.org/10.6073/pasta/a2aa41040c3e655eeb4406808a442e50 (Accessed 2022-08-02).

References

AlNoamany, Y., Borghi, J.A., 2018. Towards computational reproducibility: researcher perspectives on the use and sharing of software. PeerJ Comput. Sci. 4, e163. https://doi.org/10.7717/peerj-cs.163

Bahim, C., Casorrán-Amilburu, C., Dekkers, M., Herczog, E., Loozen, N., Repanas, K., Russell, K., Stall, S., 2020. The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments. Data Sci. J. 19, 41. https://doi.org/10.5334/dsj-2020-041

Baron, J.S., Specht, A., Garnier, E., Bishop, P., Campbell, C.A., Davis, F.W., Fady, B., Field, D., Gross, L.J., Guru, S.M., Halpern, B.S., Hampton, S.E., Leavitt, P.R., Meagher, T.R., Ometto, J., Parker, J.N., Price, R., Rawson, C.H., Rodrigo, A., Sheble, L.A., Winter, M., 2017. Synthesis Centers as Critical Research Infrastructure. BioScience 67, 750–759. https://doi.org/10.1093/biosci/bix053

Batt, R.D., Carpenter, S.R., Ives, A.R., 2017. Extreme events in lake ecosystem time series. Limnol. Oceanogr. Lett. 2, 63–69. https://doi.org/10.1002/lol2.10037

Belter, C.W., 2014. Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets. PLoS ONE 9, e92590. https://doi.org/10.1371/journal.pone.0092590

Boland, M.R., Karczewski, K.J., Tatonetti, N.P., 2017. Ten Simple Rules to Enable Multi-site Collaborations through Data Sharing. PLOS Comput. Biol. 13, e1005278. https://doi.org/10.1371/journal.pcbi.1005278

Borghi, J.A., Van Gulick, A.E., 2021. Promoting Open Science Through Research Data Management. https://doi.org/10.48550/ARXIV.2110.00888

Brase, J., 2010. DataCite - A global registration agency for research data (No. 149), RatSWD Working Paper Series. Rat für Sozial- und Wirtschaftsdaten, Berlin.

Brinckman, A., Chard, K., Gaffney, N., Hategan, M., Jones, M.B., Kowalik, K., Kulasekaran, S., Ludäscher, B., Mecum, B.D., Nabrzyski, J., Stodden, V., Taylor, I.J., Turk, M.J., Turner, K., 2019. Computing environments for reproducibility: Capturing the “Whole Tale.” Future Gener. Comput. Syst. 94, 854–867. https://doi.org/10.1016/j.future.2017.12.029

Briney, K., Coates, H., Goben, A., 2020. Foundational Practices of Research Data Management. Res. Ideas Outcomes 6, e56508. https://doi.org/10.3897/rio.6.e56508

Buttigieg, P.L., Pafilis, E., Lewis, S.E., Schildhauer, M.P., Walls, R.L., Mungall, C.J., 2016. The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. J. Biomed. Semant. 7, 57. https://doi.org/10.1186/s13326-016-0097-6

Carpenter, S.R., Armbrust, E.V., Arzberger, P.W., Chapin, F.S., Elser, J.J., Hackett, E.J., Ives, A.R., Kareiva, P.M., Leibold, M.A., Lundberg, P., Mangel, M., Merchant, N., Murdoch, W.W., Palmer, M.A., Peters, D.P.C., Pickett, S.T.A., Smith, K.K., Wall, D.H., Zimmerman, A.S., 2009. Accelerate Synthesis in Ecology and Environmental Sciences. BioScience 59, 699–701. https://doi.org/10.1525/bio.2009.59.8.11

Catford, J.A., Wilson, J.R.U., Pyšek, P., Hulme, P.E., Duncan, R.P., 2022. Addressing context dependence in ecology. Trends Ecol. Evol. 37, 158–170. https://doi.org/10.1016/j.tree.2021.09.007

Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.-D., Kacprzak, E., Groth, P., 2020. Dataset search: a survey. VLDB J. 29, 251–272. https://doi.org/10.1007/s00778-019-00564-x

Cheruvelil, K.S., Webster, K.E., King, K.B.S., Poisson, A.C., Wagner, T., 2022. LAGOS-NE Shallow Lakes: a dataset of lake variables and multi-scaled ecological context variables used to predict and compare trophic status and TP:CHLa relationships between shallow and non-shallow lakes in the Upper Midwest and Northeastern United States. https://doi.org/10.6073/PASTA/BE49507B941815D7A6807A273EE02D1E

Collins, S., Avolio, M., Gries, C., Hallett, L., Koerner, S., La Pierre, K., Rypel, A., Sokol, E., Fey, S., Flynn, D., Jones, S., Ladwig, L., Ripplinger, J., Jones, M., 2018. Compiled long-term community composition datasets of primary producers and consumers in both freshwater and terrestrial communities. https://doi.org/10.6073/PASTA/91789C93A8930C3091BC5849060FF672

Collins, S.L., 2020. Synthesis in Ecology. BioScience 70, 1041–1041. https://doi.org/10.1093/biosci/biaa149

Contaxis, N., Clark, J., Dellureficio, A., Gonzales, S., Mannheimer, S., Oxley, P.R., Ratajeski, M.A., Surkis, A., Yarnell, A.M., Yee, M., Holmes, K., 2022. Ten simple rules for improving research data discovery. PLOS Comput. Biol. 18, e1009768. https://doi.org/10.1371/journal.pcbi.1009768

Cooper, D., Springer, R., 2019. Data Communities: A New Model for Supporting STEM Data Sharing. Ithaka S+R. https://doi.org/10.18665/sr.311396

Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., Simons, N., 2019. Bringing Citations and Usage Metrics Together to Make Data Count. Data Sci. J. 18, 9. https://doi.org/10.5334/dsj-2019-009

Eisenstein, M., 2022. In pursuit of data immortality. Nature 604, 207–208. https://doi.org/10.1038/d41586-022-00929-3

Fox, P., Erdmann, C., Stall, S., Griffies, S.M., Beal, L.M., Pinardi, N., Hanson, B., Friedrichs, M.A.M., Feakins, S., Bracco, A., Pirenne, B., Legg, S., 2021. Data and Software Sharing Guidance for Authors Submitting to AGU Journals. https://doi.org/10.5281/zenodo.5124741

Goodman, A., Pepe, A., Blocker, A.W., Borgman, C.L., Cranmer, K., Crosas, M., Di Stefano, R., Gil, Y., Groth, P., Hedstrom, M., Hogg, D.W., Kashyap, V., Mahabal, A., Siemiginowska, A., Slavkovic, A., 2014. Ten Simple Rules for the Care and Feeding of Scientific Data. PLoS Comput. Biol. 10, e1003542. https://doi.org/10.1371/journal.pcbi.1003542

Gregory, K., Groth, P., Scharnhorst, A., Wyatt, S., 2020. Lost or Found? Discovering Data Needed for Research. Harv. Data Sci. Rev. https://doi.org/10.1162/99608f92.e38165eb

Gries, C., 2022. Analyze Metadata Content [WWW Document]. URL https://ediorg.github.io/eml_content_analyses/ (accessed 8.2.22).

Gries, C., Beaulieu, S., Brown, R.F., Elmendorf, S., Garritt, H., Gastil-Buhl, G., Hsieh, H.-Y., Kui, L., Martin, M., Maurer, G., Nguyen, A.T., Porter, J.H., Sapp, A., Servilla, M., Whiteaker, T.L., 2021. Data package design for special cases. https://doi.org/10.6073/PASTA/9D4C803578C3FBCB45FC23F13124D052

Gries, C., Gahler, M.R., Hanson, P.C., Kratz, T.K., Stanley, E.H., 2016. Information management at the North Temperate Lakes Long-term Ecological Research site — Successful support of research in a large, diverse, and long running project. Ecol. Inform. 36, 201–208. https://doi.org/10.1016/j.ecoinf.2016.08.007

Gries, C., Servilla, M., 2022. Data and code for EDI overview paper, data collection characteristics, FAIR evaluation, downloads, and citations. https://doi.org/10.6073/PASTA/A2AA41040C3E655EEB4406808A442E50

Groth, P., Cousijn, H., Clark, T., Goble, C., 2020. FAIR Data Reuse – the Path through Data Citation. Data Intell. 2, 78–86. https://doi.org/10.1162/dint_a_00030

Hackett, E.J., Parker, J.N., Conz, D., Rhoten, D., Parker, A., 2008. Ecology Transformed: The National Center for Ecological Analysis and Synthesis and the Changing Patterns of Ecological Research, in: Olson, G.M., Zimmerman, A., Bos, N. (Eds.), Scientific Collaboration on the Internet. The MIT Press, pp. 277–296. https://doi.org/10.7551/mitpress/9780262151207.003.0016

Halpern, B.S., Berlow, E., Williams, R., Borer, E.T., Davis, F.W., Dobson, A., Enquist, B.J., Froehlich, H.E., Gerber, L.R., Lortie, C.J., O’connor, M.I., Regan, H., Vázquez, D.P., Willard, G., 2020. Ecological Synthesis and Its Role in Advancing Knowledge. BioScience biaa105. https://doi.org/10.1093/biosci/biaa105

Hampton, S.E., Strasser, C.A., Tewksbury, J.J., Gram, W.K., Budden, A.E., Batcheller, A.L., Duke, C.S., Porter, J.H., 2013. Big data and the future of ecology. Front. Ecol. Environ. 11, 156–162. https://doi.org/10.1890/120103

Hanisch, R., Chalk, S., Coulon, R., Cox, S., Emmerson, S., Flamenco Sandoval, F.J., Forbes, A., Frey, J., Hall, B., Hartshorn, R., Heus, P., Hodson, S., Hosaka, K., Hutzschenreuter, D., Kang, C.-S., Picard, S., White, R., 2022. Stop squandering data: make units of measurement machine-readable. Nature 605, 222–224. https://doi.org/10.1038/d41586-022-01233-w

Jones, M., O'Brien, M., Mecum, B., Boettiger, C., Schildhauer, M., Maier, M., Whiteaker, T., Earl, S., Chong, S., 2019a. Ecological Metadata Language version 2.2.0. https://doi.org/10.5063/F11834T2

Jones, M., Slaughter, P., Habermann, T., 2019b. Quantifying FAIR: metadata improvement and guidance in the DataONE repository network. Knowledge Network for Biocomplexity. https://doi.org/10.5063/F14T6GP0

Jones, M.B., Slaughter, P., 2019. Quantifying FAIR: metadata improvement and guidance in the DataONE repository network.

Kaplan, N.E., Baker, K.S., Karasti, H., 2021. Long live the data! Embedded data management at a long-term ecological research site. Ecosphere 12. https://doi.org/10.1002/ecs2.3493

Kratz, J., Strasser, C., 2014. Data publication consensus and controversies. F1000Research 3, 94. https://doi.org/10.12688/f1000research.3979.3

Kratz, J.E., Strasser, C., 2015. Researcher Perspectives on Publication and Peer Review of Data. PLOS ONE 10, e0117619. https://doi.org/10.1371/journal.pone.0117619

Li, S., Pennings, S.C., 2016. Disturbance in Georgia salt marshes: variation across space and time. Ecosphere 7. https://doi.org/10.1002/ecs2.1487

Lowenberg, D., Chodacki, J., Fenner, M., Kemp, J., Jones, M.B., 2019. Open Data Metrics: Lighting the Fire. Zenodo. https://doi.org/10.5281/zenodo.3525349

McKiernan, E.C., Bourne, P.E., Brown, C.T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B.A., Ram, K., Soderberg, C.K., Spies, J.R., Thaney, K., Updegrove, A., Woo, K.H., Yarkoni, T., 2016. How open science helps researchers succeed. eLife 5, e16800. https://doi.org/10.7554/eLife.16800

Michener, W.K., 2015. Ecological data sharing. Ecol. Inform. 29, 33–44. https://doi.org/10.1016/j.ecoinf.2015.06.010

National Science Board, 2005. Long-Lived digitaL data collections: enabling Research and education in the 21st Century. National Science Foundation, Arlington, VA.

O’Brien, M., Costa, D., Servilla, M., 2016. Ensuring the quality of data packages in the LTER network data management system. Ecol. Inform. 36, 237–246. https://doi.org/10.1016/j.ecoinf.2016.08.001

O’Brien, M., Smith, C.A., Sokol, E.R., Gries, C., Lany, N., Record, S., Castorani, M.C.N., 2021. ecocomDP: A flexible data design pattern for ecological community survey data. Ecol. Inform. 64, 101374. https://doi.org/10.1016/j.ecoinf.2021.101374

Pasquetto, I., Randles, B., Borgman, C., 2017. On the Reuse of Scientific Data. Data Sci. J. 16, 8. https://doi.org/10.5334/dsj-2017-008

Patel, K.F., Fernandez, I.J., Nelson, S.J., Norton, S.A., Spencer, C.J., 2021. The Bear Brook Watershed in Maine: Multi-decadal whole-watershed experimental acidification. Hydrol. Process. 35, e14147. https://doi.org/10.1002/hyp.14147

Peters, D.P.C., Christine M. Laney, Ariel E. Lugo, Scott L. Collins, Charles T. Driscoll, Peter M. Groffman, J. Morgan Grove, Alan K. Knapp, Timothy K. Kratz, Mark D. Ohman, Robert B. Waide, Jin Yao, United States. Agricultural Research Service, 2013. Long-term trends in ecological systems : a basis for understanding responses to global change 1 online book (378 pages) : illustrations, color maps.-USDA.

Porter, J.H., 2019. Evaluating a thesaurus for discovery of ecological data. Ecol. Inform. 51, 151–156. https://doi.org/10.1016/j.ecoinf.2019.03.002

Puebla, I., Lowenberg, D., FORCE11 Research Data Publishing Ethics WG, 2021. Joint FORCE11 and COPE Research Data Publishing Ethics Working Group Recommendations. Zenodo. https://doi.org/10.5281/ZENODO.5391293

Roche, D., Berberi, I., Dhane, F., Lauzon, F., Soeharjono, S., Dakin, R., Binning, S., 2021. The quality of open datasets shared by researchers in ecology and evolution is moderately repeatable and slow to change. https://doi.org/10.32942/osf.io/d63js

Roche, D.G., Kruuk, L.E.B., Lanfear, R., Binning, S.A., 2015. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? PLOS Biol. 13, e1002295. https://doi.org/10.1371/journal.pbio.1002295

Schmidt, B., Gemeinholzer, B., Treloar, A., 2016. Open Data in Global Environmental Research: The Belmont Forum’s Open Data Survey. PLOS ONE 11, e0146695. https://doi.org/10.1371/journal.pone.0146695

Servilla, M., Brunt, J., Costa, D., McGann, J., Waide, R., 2016. The contribution and reuse of LTER data in the Provenance Aware Synthesis Tracking Architecture (PASTA) data repository. Ecol. Inform. 36, 247–258. https://doi.org/10.1016/j.ecoinf.2016.07.003

Soranno, P.A., Bacon, L.C., Beauchene, M., Bednar, K.E., Bissell, E.G., Boudreau, C.K., Boyer, M.G., Bremigan, M.T., Carpenter, S.R., Carr, J.W., Cheruvelil, K.S., Christel, S.T., Claucherty, M., Collins, S.M., Conroy, J.D., Downing, J.A., Dukett, J., Fergus, C.E., Filstrup, C.T., Funk, C., Gonzalez, M.J., Green, L.T., Gries, C., Halfman, J.D., Hamilton, S.K., Hanson, P.C., Henry, E.N., Herron, E.M., Hockings, C., Jackson, J.R., Jacobson-Hedin, K., Janus, L.L., Jones, W.W., Jones, J.R., Keson, C.M., King, K.B.S., Kishbaugh, S.A., Lapierre, J.-F., Lathrop, B., Latimore, J.A., Lee, Y., Lottig, N.R., Lynch, J.A., Matthews, L.J., McDowell, W.H., Moore, K.E.B., Neff, B.P., Nelson, S.J., Oliver, S.K., Pace, M.L., Pierson, D.C., Poisson, A.C., Pollard, A.I., Post, D.M., Reyes, P.O., Rosenberry, D.O., Roy, K.M., Rudstam, L.G., Sarnelle, O., Schuldt, N.J., Scott, C.E., Skaff, N.K., Smith, N.J., Spinelli, N.R., Stachelek, J.J., Stanley, E.H., Stoddard, J.L., Stopyak, S.B., Stow, C.A., Tallant, J.M., Tan, P.-N., Thorpe, A.P., Vanni, M.J., Wagner, T., Watkins, G., Weathers, K.C., Webster, K.E., White, J.D., Wilmes, M.K., Yuan, S., 2017. LAGOS-NE: a multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of US lakes. GigaScience 6. https://doi.org/10.1093/gigascience/gix101

Soranno, P.A., Lottig, N.R., Delany, A.D., Cheruvelil, K.S., 2019. LAGOS-NE-LIMNO v1.087.3: A module for LAGOS-NE, a multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of U.S. Lakes: 1925-2013. https://doi.org/10.6073/PASTA/08C6F9311929F4874B01BCC64EB3B2D7

Stafford, S.G., 2021. A Retrospective of Information Management in the Long Term Ecological Research Program, in: Waide, R.B., Kingsland, S.E. (Eds.), The Challenges of Long

Term Ecological Research: A Historical Analysis, Archimedes. Springer International Publishing, Cham, pp. 375–402. https://doi.org/10.1007/978-3-030-66933-1_13

van de Sandt, S., Dallmeier-Tiessen, S., Lavasa, A., Petras, V., 2019. The Definition of Reuse. Data Sci. J. 18, 22. https://doi.org/10.5334/dsj-2019-022

Vanderbilt, K., Gries, C., 2021. Integrating long-tail data: How far are we? Ecol. Inform. 64, 101372. https://doi.org/10.1016/j.ecoinf.2021.101372

Vanderbilt, K., Ide, J., Gries, C., Grossman-Clarke, S., Hanson, P., O'Brien, M., Servilla, M., Smith, C., Waide, R., Zollo-Venecek, K., 2022. Publishing Ecological Data in a Repository: An Easy Workflow for Everyone. Bull. Ecol. Soc. Am. https://doi.org/10.1002/bes2.2018

Walter, J.A., Hallett, L.M., Sheppard, L.W., Anderson, T.L., Zhao, L., Hobbs, R.J., Suding, K.N., Reuman, D.C., 2021. Micro-scale geography of synchrony in a serpentine plant community. J. Ecol. 109, 750–762. https://doi.org/10.1111/1365-2745.13503

Web of Science, 2021. Journal Impact Factor List 2021 – JCR, Web Of Science (PDF, XLS). J. Impact Factor. URL https://impactfactorforjournal.com/jcr-2021/ (accessed 12.7.21).

Whitlock, M.C., 2011. Data archiving in ecology and evolution: best practices. Trends Ecol. Evol. 26, 61–65. https://doi.org/10.1016/j.tree.2010.11.006

Wieder, W.R., Pierson, D., Earl, S., Lajtha, K., Baer, S.G., Ballantyne, F., Berhe, A.A., Billings, S.A., Brigham, L.M., Chacon, S.S., Fraterrigo, J., Frey, S.D., Georgiou, K., de Graaff, M.-A., Grandy, A.S., Hartman, M.D., Hobbie, S.E., Johnson, C., Kaye, J., Kyker-Snowman, E., Litvak, M.E., Mack, M.C., Malhotra, A., Moore, J.A.M., Nadelhoffer, K., Rasmussen, C., Silver, W.L., Sulman, B.N., Walker, X., Weintraub, S., 2021. SoDaH: the SOils DAta Harmonization database, an open-source synthesis of soil data from research networks, version 1.0. Earth Syst. Sci. Data 13, 1843–1854. https://doi.org/10.5194/essd-13-1843-2021

Wieder, W.R., Pierson, D., Earl, S.R., Lajtha, K., Baer, S., Ballantyne, F., Berhe, A.A., Billings, S., Brigham, L.M., Chacon, S.S., Fraterrigo, J., Frey, S.D., Georgiou, K., De Graaff, M.-A., Grandy, A.S., Hartman, M.D., Hobbie, S.E., Johnson, C., Kaye, J., Snowman, E., Litvak, M.E., Mack, M.C., Malhotra, A., Moore, J.A.M., Nadelhoffer, K., Rasmussen, C., Silver, W.L., Sulman, B.N., Walker, X., Weintraub, S., 2020. SOils DAta Harmonization database (SoDaH): an open-source synthesis of soil data from research networks. https://doi.org/10.6073/PASTA/9733F6B6D2FFD12BF126DC36A763E0B4

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., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., ’t Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018. https://doi.org/10.1038/sdata.2016.18