

ESSAY

Facilitating and Improving Environmental Research Data Repository Interoperability

Corinna Gries¹, Amber Budden², Christine Laney³, Margaret O'Brien⁴, Mark Servilla⁵, Wade Sheldon⁶, Kristin Vanderbilt⁵ and Dave Vieglais²

¹ Environmental Data Initiative, University of Wisconsin, Madison, Wisconsin, US

² DataONE, University of New Mexico, Albuquerque, New Mexico, US

³ National Ecological Observatory Network, Boulder, Colorado, US

⁴ Environmental Data Initiative, University of California, Santa Barbara, California, US

⁵ Environmental Data Initiative, University of New Mexico, Albuquerque, New Mexico, US

⁶ Georgia Coastal LTER, University of Georgia, Athens, Georgia, US

Corresponding author: Corinna Gries (cgries@wisc.edu)

Environmental research data repositories provide much needed services for data preservation and data dissemination to diverse communities with domain specific or programmatic data needs and standards. Due to independent development these repositories serve their communities well, but were developed with different technologies, data models and using different ontologies. Hence, the effectiveness and efficiency of these services can be vastly improved if repositories work together adhering to a shared community platform that focuses on the implementation of agreed upon standards and best practices for curation and dissemination of data. Such a community platform drives forward the convergence of technologies and practices that will advance cross-domain interoperability. It will also facilitate contributions from investigators through standardized and streamlined workflows and provide increased visibility for the role of data managers and the curation services provided by data repositories, beyond preservation infrastructure. Ten specific suggestions for such standardizations are outlined without any suggestions for priority or technical implementation. Although the recommendations are for repositories to implement, they have been chosen specifically with the data provider/data curator and synthesis scientist in mind.

Keywords: data repository; standards; data archiving; data publication; interoperability

1. Introduction

Earth observation and environmental research data are valuable for documenting, interpreting, and evaluating the state of natural systems beyond their original purpose and a number of niche data repositories are developing within communities of practice (Waide, Brunt and Servilla, 2017). Their role is to mobilize these data, hold a digital copy of them and make them Findable, Accessible, Interoperable, and Reusable (FAIR, (Wilkinson *et al.*, 2016)). A community of practice, research community, or community and, hence, the data held by their specific repository in this discussion may be delineated by a common research subject (e.g. geochemical, petrological, and geochronological research),¹ geographic research area (e.g. the Arctic)² or funding source (e.g. NSF LTER program).³ They are joined by a corpus of domain agnostic and institutional repositories. Although much data from these repositories are available today through online means, there are major challenges that still exist in discovering and reusing these data for synthesis research and meta-analyses, one of which being the

¹ <http://www.earthchem.org>

² <https://arcticdata.io/>

³ <https://portal.edirepository.org>

Table 1: Overview of recommendations and technical vs. social nature.

Recommendation	technical	social
2.1. Implement one or more metadata standards	X	X
2.2. Automate metadata capture	X	
2.3. Use community vetted vocabularies		X
2.4. Keep digital objects in coherent packages		X
2.5. Assign globally unique resolvable identifiers	X	
2.6. Version immutable digital objects	X	
2.7. Track data provenance	X	X
2.8. Use third party service for user identity functions	X	
2.9. Provide several means for data access	X	
2.10. Facilitate standards in communities of practice		X

non-interoperable state of repositories (Warner *et al.*, 2007). Open access data repositories that work together and provide a common scaffolding for data deposition, discovery, and reuse can help mitigate this dilemma by providing a similar experience for both data producers and data consumers across repositories.

Several organizations, including the EarthCube Council of Data Facilities,⁴ DataONE,⁵ the Research Data Alliance,⁶ and the Earth Science Information Partners,⁷ acknowledge data repositories as well as the scientific data provider and data re-user as their stakeholder community. Each of these organizations has the potential to be leveraged by the community of scientists and repository managers as a hub of discussion, providing guidance for the adoption of repository standards, and in the case of DataONE (Michener *et al.*, 2012) seeking input for developing infrastructure necessary to promote interoperability.

As such, we present ten specific suggestions for facilitating and improving environmental research data repository interoperability. These suggestions align with (and reinforce) the criteria already put forth by certification bodies (e.g., CoreTrustSeal from the ICSU World Data System),⁸ journals (e.g. PLOS ONE Editors, n.d.), and the principles of the “Enabling FAIR Data” project (Wilkinson *et al.*, 2016; Husen *et al.*, 2017). Here we emphasize an overarching goal of coordinated interoperability among repositories to improve the user experience for both, the data submitter or curator as well as the data re-user or synthesis researcher. Included are recommendations for planning, but not necessarily implementation details. They are loosely categorized into two areas: first, technical interoperability (i.e., interoperability s.str., enabling information exchange) and second expanding to recommendations for the repository’s reinforcement of the curation effort in its community (including the semantic, social and organizational aspects of interoperability s.l.) (Table 1). They cover a wide swath of recommendations taking the discussion beyond machine operable interfaces and data exchange protocols (Members of the RDA Research Data Repository Interoperability Group, 2017) and best practices for data providers (Goodman *et al.*, 2014; Hart *et al.*, 2016), into common patterns of community engagement which will allow separation of data curation from storage and delivery and improve efficiency, while simultaneously endorsing and building the expertise of domain data curators. We intend for these recommendations to be general enough so that they may persist through time, but they should be revisited and evolve as repositories and their communities of data providers and data re-users mature. Lastly, we make no determination about prioritization, as the usefulness will vary depending on the community, but some recommendations build on each other.

2.1 Community-supported metadata standards

Most basic and generally agreed upon is that implementing a community-recognized and supported metadata standard (e.g., (McQuilton *et al.*, 2016)⁹ will make data easily understood, exchanged and compatible. Furthermore, implementation of a recognized standard enables the burden of software development for

⁴ <https://www.earthcube.org/group/council-data-facilities>

⁵ <https://www.dataone.org/>

⁶ <https://www.rd-alliance.org/>

⁷ <http://www.esipfed.org/>

⁸ <https://www.coretrustseal.org/>

⁹ <https://fairsharing.org/>

both data submission and consumption to be shared across the community. However, repositories may need to consider adopting more than one metadata standard, or to export metadata in more than one specification. A smaller number of standards or profiles for Earth science data (e.g., Ecological Metadata Language,¹⁰ ISO/TS 19115-3:2016,¹¹ Dublin Core)¹² means that crosswalks between them exist and metadata can be shared, while still allowing the repository to use the specification which is best-suited to their data entities and curation practices. We recommend that a repository choose the metadata standard that allows for the richest description of the data they manage to maximize potential understanding and integration. Starting with the richest metadata possible will also enable better support for transformation to other specifications, especially with the likelihood that most crosswalks will lose some information during translation.

2.2 Prioritizing automated metadata capture

Usability and user-friendliness of data submission tools are critical to the success and reputation of a repository. A common barrier to the adoption of a repository by a community is the over-reliance on manual metadata entry tools (e.g., form-based entry methods); these tools have considerable differences in their interface workflows, thereby frustrating users that rely on multiple repositories and, because they often accept natural-language input into their form fields, their content can be wildly inconsistent between users. An alternative solution that reduces the level of user interaction (and frustration) and provides greater consistency in content across repositories is to automate metadata creation wherever possible. Automation also has the potential to improve metadata quality and richness since data minutiae that may be overlooked or ignored during manual generation can be codified and easily included time-and-time again. Possibilities include: sensor- and data logger-level metadata capture while in the field or when importing data from loggers (e.g. mining configuration information from sensor logs, deriving data types and attribute names from files or scripts); use of metadata augmentation services (e.g. keyword expansion, geographic gazetteers, directory services for personnel contact records); and providing support for metadata content re-use (e.g. templating or import services). There are also end-user tools or code libraries (e.g. DataONE-R,¹³ EMLAssemblyLine¹⁴) that can streamline both metadata creation and data submission to repositories supporting web services.

2.3 Open, standardized, community vetted vocabularies

A key component of data repositories are controlled vocabularies that unambiguously describe data concepts. Controlled vocabulary terms should be associated with unique identifiers, resolvable URIs, and authoritative definitions. It is important that the terms described by the controlled vocabulary have persistent identifiers to enable linked data (Bizer, Heath and Berners-Lee, 2009) or integration with formal modeling frameworks. Ideally, a controlled vocabulary is not simply a flat glossary of terms but should encompass a formal model that defines concepts and their relationships within a scientific domain, e.g., in an ontology. When repositories use the same controlled vocabulary to annotate datasets, computers can more easily find, extract, aggregate and integrate related information broadly. Controlled vocabularies will change as knowledge increases, and repositories should ensure that their sources are reputable and include mechanisms for community input for the creation of new terms or to clarify the definition of existing terms. Typically, terms in controlled vocabularies may be deprecated but not deleted, to promote backward compatibility; guidelines for vocabulary maintenance are beyond the scope of this document (e.g., Smith et al., 2007). Some controlled vocabularies include Library of Congress Subject Headings,¹⁵ Dublin Core Metadata Type Vocabulary,¹⁶ EnvThes (Schentz, Peterseil and Bertrand, 2013), LTER Controlled Vocabulary (Porter, 2010) SWEET earthrealm ontologies (Raskin and Pan, 2005), and NASA's Global Change Master Directory (Blumenfeld, no date).

2.4 Coherent packages of digital objects

Datasets, unlike traditional resources such as journal articles or books in PDF files, are complex, composite digital structures containing multiple objects of varying formats and may include cross references to related information such as provenance. Environmental science studies often produce multiple files of raw

¹⁰ <https://knb.ecoinformatics.org/#external//emlparser/docs/index.html>

¹¹ <https://www.iso.org/standard/32579.html>

¹² <http://dublincore.org/>

¹³ <https://cran.r-project.org/web/packages/dataone/index.html>

¹⁴ <https://github.com/EDIorg/EMLassemblyline>

¹⁵ <http://id.loc.gov/authorities/subjects.html>

¹⁶ <http://dublincore.org/documents/2003/02/12/dcmi-type-vocabulary/>

observational data in CSV, netCDF, JPG, and other formats, plus XML metadata files describing the context and structure of those data. Ancillary data may include calibration information, files derived by processing the data through quality control, calibration and transformation algorithms, and products such as visualizations (graphs, animations) with code and provenance traces to explain or reproduce these (see further discussion,¹⁷ (Bechhofer *et al.*, 2010)). Publishing data, or collections of digital research objects as a coherent package enables researchers to properly understand, use and cite complex datasets (The Consultative Committee for Space Data Systems (CCSDS), 2012). While there are no hard rules of what constitutes a dataset, repositories should ensure that each published package represents a coherent collection of data or research output so that each package represents a citable unit. Packages can also reference other child packages, allowing repositories to build hierarchies of data for convenient navigation of complex data spaces.

Although different approaches to creating such data package units exist, all data packages should include a package manifest such as an OAI-ORE¹⁸ document that aggregates its members by referencing their globally unique identifiers. The entire data package should be assigned its own globally unique and resolvable identifier, which is typically a DOI or other recognizable identifier that facilitates citation (see 2.5). Relationships among package members are represented as explicit statements linking two objects through their IDs.

2.5 Persistent, globally unique, resolvable identifiers

One of the key means for enabling interoperability among data repositories involves the use of globally unique, persistent, web-resolvable identifiers. Identifiers should be globally unique, meaning that they consistently “point” to a digital description of a single resource and only to that resource. However, the exact specification of what constitutes a resource, and what makes it “consistent” is subject to interpretation. A resource may be a dataset, a team, or a set of events, as well as more solitary entities such as an individual person or ontology class. For digital resources (e.g., datasets) “consistent” is interpreted as fully “immutable”, meaning that they do not change at all and retain an exact digital signature (see also 2.6). Two such identifier systems have emerged for use with data which are centrally managed to be resolvable. These are the Digital Object Identifier (DOI, managed by DataCite)¹⁹ and the Archival Resource Key (ARK, managed by the California Digital Library).²⁰ It is recommended that these identifiers be used to point to the highest-level description of the resource giving access to all parts (e.g., metadata, resource map). Other identifiers are available, e.g., PURL,²¹ UUID, GUID²² (Duerr *et al.*, 2011), which usually follow a standard pattern and may be generated by the repository and used internally for higher granularity (see 2.4) to ensure all discrete entities in a dataset are uniquely identified.

2.6 Versioning of digital objects and packages

This recommendation addresses the immutability of a data package in more detail. Data and research objects evolve over time, but it is important for users to be able to consistently refer to specific data collected or used in an analysis or paper (Force 11, no date). Examples of data that change over time are long-term time-series with regular additions (e.g., monthly, annually), or advances in methodology or processing methods applied to extant data, to generate a dataset revision. The ability to specifically identify and access an exact revision of data is critical for ensuring reproducibility of analyses utilizing those data. However, when an older version is accessed, the repository should highlight the fact that a newer version is available. In most cases this functionality is achieved by assigning a new identifier (e.g., DOI) to every version and keeping all versions accessible while general searches would present only the most recent version. Relationships between revisions of data should be preserved by a repository and when making such information available to users, the relationships should utilize already defined concepts intended for such purpose (e.g., a versioning system like software versioning).

2.7 Dataset provenance information

Provenance is a chain, or network of stages of processing that ultimately conclude in the data or output of interest. Provenance relationships include relationships between raw data, re-processed data, derived data, derived outputs such as figures and graphs, and detailed information on computational steps that gener-

¹⁷ <https://environmentaldatainitiative.org/resources/five-phases-of-data-publishing/phase-1/>

¹⁸ <http://www.openarchives.org/ore/1.0/toc>

¹⁹ <https://www.datacite.org/>

²⁰ <https://www.cdlib.org/>

²¹ https://en.wikipedia.org/wiki/Persistent_uniform_resource_locator

²² https://en.wikipedia.org/wiki/Universally_unique_identifier

ated these. Provenance may be documented in different ways. Some metadata standards allow inclusion of this information. Preferred, however, is the use of the ProvONE²³ or PROV-O²⁴ standard to represent data and computing provenance relationships. In addition, repositories should show a human readable representation of provenance, ideally, with identifiers that resolve to associated data.

Especially in an ecosystem of several data repositories provenance traces may document data aggregation from across different repositories. This approach allows a future data re-user to understand possible data duplication. Furthermore, reproducibility of scientific results depends on transparent workflows where the actors and elements involved are named, identified, and documented. Documenting the whole analytical workflow inspires confidence that the output was produced properly. Data preparation and analysis should be scripted in an open source scripting language which can be archived along with the data input and output. If that is not possible, software and its version, parameters, decisions and approaches should be described in the methods of a data package.

2.8 Third-party services for user identity functions

Unique identities and user authentication is critical for repositories to interoperate in a safe and efficient manner. Individuals must be disambiguated from others with similar identities and their identities must be verified to determine authenticity (i.e., is the user really who they claim to be?). Although most repositories provide open access to their data holdings, this is important when a user requests read access to protected data, write access to infrastructure resources, or for assigning attribution to a data contributor. Identities must be unique in the context of a local repository but should also map to identities in use at other repositories to facilitate interoperability. Even better, a single community-wide identity eliminates the need for identity mapping at all. Ensuring that an identity is unique and verifiable is often provided as a bundled service of an identity provider. A common option for user identity management is to utilize a trusted third-party service like those offered by Google, Facebook, or GitHub, instead of deploying a locally managed registry of user profile and credential information. Such providers often utilize a combination of the OpenID Connect²⁵ and OAuth²⁶ protocols to deliver easy-to-use and secure authentication and authorization services for clients relying on the disambiguated and verified identification of their user base. It is also important to choose an identity provider that is recognized by your community for both visibility and scalability. A familiar identity provider for the Earth and environmental science communities is the Open Researcher and Contributor ID (ORCID)²⁷ service. ORCID identifiers provide both a unique user identity that is guaranteed to disambiguate individuals and the ORCID service is a recognized identity provider that uses the OpenID Connect and OAuth protocols.

2.9 Data access via a variety of mechanisms

Different access options will appeal to different users. Although not directly supporting repository interoperability, a web browser landing page including human readable metadata, citation information, links for accessing the data, and cross references to related resources such as provenance information (see 2.7) will be appreciated by less technical users. Embedding the same information in machine readable form within the landing page, such as with schema.org²⁸ constructs, will help ensure that technical users are equally satisfied with the representation since access to the dataset and related components using scripts or other software tools is equally well supported. Such a landing page provides an appropriate target for identifier resolution since regardless of whether the identifier is resolved by human or machine, complete information about the dataset can be discerned at the resolved location.

Providing access to dataset components through alternative protocol choices ensures that a user can access the resources using a mechanism appropriate for the data and the tools they have at their disposal. For example, whilst most components may be accessible via HTTPS or SFTP/FTPS, other protocols may be more appropriate for very large data resources where users may choose to access portions of, rather than the entire resource. It is important that where multiple options are available for accessing a resource, it is clear that the choices offer alternative access to the same resource so that users may select a protocol of preference without unnecessarily accessing the same resource multiple times.

²³ <http://vcvcomputing.com/provone/provone.html>

²⁴ <https://www.w3.org/TR/prov-o/>

²⁵ <http://openid.net/connect/>

²⁶ <https://oauth.net/>

²⁷ <https://orcid.org/>

²⁸ <http://schema.org/>

It is also important that the landing page reflects the current state of mechanisms available for accessing dataset components. If new services are added, or changes are made, then these should be clearly communicated through the landing pages and via other distribution mechanisms such as a user forum, a subscription-based alert service or other mechanism. Replacement services should be backwards compatible for at least a reasonable period of time to ensure resource access is not inadvertently disrupted. Web services APIs²⁹ should provide options for subsetting by common spatio-temporal and thematic factors (e.g. date range, latitude longitude bounding box, sampling event, platform, etc.) and advertise availability of such options. As much as possible, responses from queries should not be limited, or in the alternative, a simple means of retrieving remaining pages of data from the original query should be advertised and supported.

2.10 Facilitating communities of practice and standards for research data

Repositories contain thousands of datasets, many of them suitable for reuse in synthesis projects, meta-analyses or to contextualize new studies. Synthesis and integration would be much easier, however, if similar data were presented similarly. However, “similar” can have a variety of meanings. For example, that datasets a) can be found together because they use the same key terms in catalogs; b) use standardized variable descriptions to minimize heterogeneity and provide explicit context; c) have similar table layout and content; or that d) data that are likely to be used together (e.g., repeated surveys, time series) are aggregated, linked or otherwise packaged together. Some research communities have coalesced around certain practices. In the environmental sciences, this is particularly true when collections are mission-oriented or instruments are involved, as the formats and processing scenarios are often determined by the manufacturers. However, for observational research data (e.g., organism abundance, or concentrations of chemical constituents), datasets tend to be designed ad hoc and specific to the study.

Two processes improve data reusability and repository interoperability: peer review of data papers and professional data curation. Repositories should collaborate to identify performance standards that already exist among their submitters, plus additional community practices that could become standardized, and build coalescence around these. Registries of recommendations for classes of data where specific domain knowledge is required (e.g. “stable isotope ratios” or “populations and organism abundance”) would have the added benefit of creating a curriculum area of expertise for data curators, emphasizing their relevance.

Conclusions

In the current landscape of small Earth and environmental data repositories tension exists between providing community specific services and implementing interoperability to serve the larger needs of open and accessible data. As FAIR data services mature implementing the technical suggestions will hasten repository interoperability, while data interoperability will mostly rely on implementing the social recommendations. The latter will also require further developments in semantic as well as structural interoperability automation.

Acknowledgements

A workshop was held at the National Center for Ecological Analysis and Synthesis (NCEAS), we are grateful for the use of their facilities. We would like to thank the following workshop participants for their contributions to the initial draft of this manuscript, R. Arko, M. Jones, K. Lehnert, M. Leon, M. Schildhauer, A. Shepherd, R. Simons, and B. Vandenberg. A big thank you goes to the reviewers of this manuscript for their scrutiny and extensive suggestions for improvements.

Funding Information

This work was supported by the National Science Foundation grant DBI-EAGER 1500306.

Competing Interests

All authors and workshop participants are representatives of community data repositories in the domain of Earth and environmental sciences. United by their work in data curation they have backgrounds in computer science, information science or domain science. The represented repositories, although maintained in the US hold data from a wide variety of subdomains, research themes within subdomains, geographic regions, and funding sources.

²⁹ https://en.wikipedia.org/wiki/Application_programming_interface

Author Information

This manuscript was conceived during a workshop titled “Cross-institutional repository collaborations”. The authors participated in the workshop and were all involved in developing this manuscript. Gries did the initial editing of the workshop material, final editing of the manuscript and editing in response to reviewer comments. Servilla substantially edited the introduction while contributing to chapters 2.2, 2.8. All other authors’ contributions are listed by chapter numbers: Budden 2.1, 2.5, 2.9, 2.10; Laney, O’Brien 2.1, 2.3, 2.10; Sheldon 2.2; Vanderbilt 2.3; Vieglais 2.4, 2.5, 2.9.

References


- Bechhofer, S**, et al. 2010. Research Objects: Towards Exchange and Reuse of Digital Knowledge. *Nature Precedings*. Nature Publishing Group. DOI: <https://doi.org/10.1038/npre.2010.4626.1>
- Bizer, C, Heath, T and Berners-Lee, T**. 2009. Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3): 1–22. DOI: <https://doi.org/10.4018/jswis.2009081901>
- Blumenfeld, J**. no date. *The Global Change Master Directory: Data, Services, and Tools Serving the International Science Community*. Available at: <https://earthdata.nasa.gov/gcmd-retrospective-and-future> (Accessed: 11 May 2018).
- Duerr, RE**, et al. 2011. On the utility of identification schemes for digital earth science data: An assessment and recommendations. *Earth Science Informatics*, 4(3): 139–160. Springer-Verlag. DOI: <https://doi.org/10.1007/s12145-011-0083-6>
- Force 11**. no date. *Joint Declaration of Data Citation Principles*. DOI: <https://doi.org/10.25490/a97f-egykh>
- Goodman, A**, et al. 2014. Ten Simple Rules for the Care and Feeding of Scientific Data. Bourne, PE (ed.), *PLoS Computational Biology*, 10(4): e1003542. Public Library of Science. DOI: <https://doi.org/10.1371/journal.pcbi.1003542>
- Hart, EM**, et al. 2016. Ten Simple Rules for Digital Data Storage. Markel, S (ed.), *PLoS Computational Biology*, 12(10): e1005097. Public Library of Science. DOI: <https://doi.org/10.1371/journal.pcbi.1005097>
- Husen, SE**, et al. 2017. Recommended versus Certified Repositories: Mind the Gap. *Data Science Journal*, 16: 42. Ubiquity Press. DOI: <https://doi.org/10.5334/dsj-2017-042>
- McQuilton, P**, et al. 2016. BioSharing: Curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database*, 2016: baw075. Oxford University Press. DOI: <https://doi.org/10.1093/database/baw075>
- Members of the RDA Research Data Repository Interoperability and Group**. 2017. *Research Data Repository Interoperability Primer*. DOI: <https://doi.org/10.15497/RDA00020>
- Michener, WK**, et al. 2012. Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences. *Ecological Informatics*, 11: 5–15. Elsevier B.V. DOI: <https://doi.org/10.1016/j.ecoinf.2011.08.007>
- PLOS ONE Editors**. no date. *PLOS Criteria for Recommended Data Repositories|EveryONE: The PLOS ONE blog*. Available at: <http://blogs.plos.org/everyone/2018/03/01/criteria-for-recommended-data-repositories/> (Accessed: 11 May 2018).
- Porter, J**. 2010. A Controlled Vocabulary for LTER Datasets. *LTER DataBits*. Spring. Available at: <http://data-bits.lternet.edu/issues/115> (Accessed: 11 May 2018).
- Raskin, RG and Pan, MJ**. 2005. Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). *Computers & Geosciences*, 31(9): 1119–1125. Pergamon. DOI: <https://doi.org/10.1016/J.CAGEO.2004.12.004>
- Schentz, H, Peterseil, J and Bertrand, N**. 2013. EnvThes – interlinked thesaurus for long term ecological research, monitoring, and experiments. In: *EnviroInfo 2013: Environmental Informatics and Renewable Energies*, 824–832. Aachen: Shaker Verlag. Available at: <https://pdfs.semanticscholar.org/2606/44046e0a0cb1583a002a56449e512183b913.pdf> (Accessed: 11 May 2018).
- Smith, B**, et al. 2007. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11): 1251–1255. Nature Publishing Group. DOI: <https://doi.org/10.1038/nbt1346>
- The Consultative Committee for Space Data Systems (CCSDS)**. 2012. Reference Model for an open Archival Information System (OAIS). *CCSDS 650.0-M-2*. Available at: <https://public.ccsds.org/pubs/650x0m2.pdf> (Accessed: 19 March 2018).
- Waide, RB, Brunt, JW and Servilla, MS**. 2017. Demystifying the Landscape of Ecological Data Repositories in the United States. *BioScience*, 67(12): 1044–1051. Oxford University Press. DOI: <https://doi.org/10.1093/biosci/bix117>

- Warner, S**, et al. 2007. Pathways: Augmenting interoperability across scholarly repositories. *International Journal on Digital Libraries*, 7(1–2): 35–52. Springer-Verlag. DOI: <https://doi.org/10.1007/s00799-007-0016-7>
- Wilkinson, MD**, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3: 160018. DOI: <https://doi.org/10.1038/sdata.2016.18>

How to cite this article: Gries, C, Budden, A, Laney, C, O'Brien, M, Servilla, M, Sheldon, W, Vanderbilt, K and Vieglaiss, D. 2018. Facilitating and Improving Environmental Research Data Repository Interoperability. *Data Science Journal*, 17: 22, pp. 1–8. DOI: <https://doi.org/10.5334/dsj-2018-022>

Submitted: 26 June 2018 **Accepted:** 07 September 2018 **Published:** 20 September 2018

Copyright: © 2018 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 