Assessment Information

[CoreTrustSeal Requirements 2020–2023]

Repository:	Environmental Data Initiative
Website:	https://edirepository.org/
Certification period:	06 June 2023 - 05 June 2026
Requirements version:	CoreTrustSeal Requirements 2020-2022

This repository is owned by:	**University of New Mexico**

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background Information

Repository Type

Please provide context for your repository. You can select one or multiple options.

Compliance level:

Not Applicable - 0

Response:

- Domain or subject-based repository
- National repository system; including governmental

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

No comment

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Description of Repository

Provide a short overview of the repository.

Compliance level:

Not Applicable - 0

Response:

The Environmental Data Initiative (EDI) is a comprehensive data management organization serving several stages in the life cycles of ecological and environmental data. EDI builds on over 40 years of data management experience in the Long Term Ecological Research (LTER) Network and works in close collaboration with the National Ecological Observatory Network (NEON) and other data-focused organizations. The EDI team of experienced data practitioners, software developers, and research scientists provides support, training, and software to help archive and publish high-quality data and metadata. The repository is a feature-rich extension of the system originally developed for the NSF Long-Term Ecological Research (LTER) Network and refined through nearly a decade of further development, experience, and user feedback. EDI’s expert data curation, detailed metadata and data discovery services set it apart from other data repositories in the ecological and environmental realms. EDI is registered with re3data.org [0.1].

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

No comment

Designated Community

Provide a clear definition of the Designated Community

Compliance level:

Not Applicable - 0

Response:

EDI works with NSF's largest and most prominent programs for ecology and ecosystem science, including the LTER Network [0.2], Long Term Research in Environmental Biology (LTREB, [0.3]), and Macrosystems programs [0.4]. Other site-based organizations are affiliated with the Organization of Biological Field Station [OBFS, 0.5], the National Park Service, California Coastal Commission, California Department of Water Resources Interagency Ecological Program [0.6]. Many of the larger, integrated projects (such as LTER) support their own data managers, and so EDI is the central hub for a large community of practice (over 100 members), with webinars, tutorials, videos, code-sharing, and working groups to address common issues. Additionally, EDI serves many independent researchers and labs in their data publication needs. EDI is the only organization to provide this range of technologies and services to the ecological research communities, which are essential to improving their data discovery, reuse, archive and appropriate attribution.

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

No comment

Level of Curation

Select all relevant types of curation.
- Content distributed as deposited
- Basic curation – e.g., brief checking, addition of basic metadata or documentation
- Enhanced curation – e.g., conversion to new formats, enhancement of documentation
- Data-level curation – as above, but with additional editing of deposited data for accuracy

Compliance level:

Not Applicable - 0

Response:

- A. Content distributed as deposited
- B. Basic curation – e.g. brief checking; addition of basic metadata or documentation
- C. Enhanced curation – e.g. conversion to new formats; enhancement of documentation

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

No comment

Level of Curation - explanation

Please add the description for your Level(s) of Curation.

Compliance level:

Not Applicable - 0

Response:

Once added, data packages in the EDI Repository are immutable, and all copies are available. Value added by EDI includes a Quality Report (added to every package) to assist with evaluation. A Digital Object Identifier added to metadata (DOI). EDI uses original data to perform two types of transformations (described below): reformatted data using common harmonized formats for specific areas of synthesis, and name normalization in the repository search interface to assist with discovery. Original uploads remain unchanged. All work is within the terms of the license accepted by the data producer and within the skill set of EDI personnel, either curators with training in ecosystem science or trained programmers.

Reformatted data: Our work with synthesis scientists has highlighted the need for high-priority primary data to be converted to analysis ready formats, thereby reducing a time-consuming step in data synthesis. EDI led one such project and produced over 70 ecological community observation datasets in the same data model. The original data deposit is the authoritative version (unchanged), and EDI's standard licensing (see Section R2) allows for transformation to produce derived data products. Data scientists within the community are encouraged to create transforms, or they are reformatted by EDI curators (see Section R12). Data are selected for transformation based on their potential for synthesis, e.g., a dataset covering a long time series of organismal abundance is likely to be selected for transformation to our harmonized format, ecocomDP [0.12]. Similarly, a dataset of stream flow would be selected for transformation to the format used by the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI, [0.13]). Transformed data become a derived data package, with a provenance chain linking to the original (see sections R14, R15).

Metadata normalization: The EDI programming team has created improved search capability based on dataset creator names, since names often occur in multiple variations in different datasets [0.14]. A simple search interface would record each variation separately in a drop-down list. However, we perform name normalization in order to display a single, canonical name for an individual, and the search function returns the union of all datasets for which that individual's name appears in one of its variations. As for data that we reformat (described above), the originally deposited dataset is unchanged and normalized metadata is stored independently. All normalization-related interactions are via web services to (a) update a names database with information from newly submitted datasets and (b) highlight cases that appear to require manual inspection. We expect that this type of metadata mediation will act as a model for other kinds of metadata normalization and harmonization in the future.

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

I've had a look through your website - just the "About" page and the data search so far, and it seems to be very good It sounds like you do plenty of good work

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Insource/Outsource Partners

If applicable, please list them.

Compliance level:

Not Applicable - 0

Response:

All code and curation is generated by EDI itself (insourced). EDI does not outsource, except to contract with Amazon’s S3 Glacier online storage facility for backup storage.

For much of EDI’s scientific community, data publication is voluntary, although practices are changing and pressure from funders and journals to publish data products is increasing. Therefore, significant EDI activities also include outreach to scientists about the value of data curation and publication. These activities are also insourced, although we collaborate with several related data organizations for some activities, such as Data Help Desks at national meetings [0.7]. EDI has built its mission and goals in concert with foundational data repository technology, resulting in collaboration in several technical areas, such as development of standard practices and metadata formats and content, some of which are managed by the Earth Science Information Partners (ESIP, 0.8]).

NEON: EDI has a formal collaboration with the NEON network, to provide data publication services for derived NEON data, and to create common tools for data reuse. [0.9].

DataONE DataONE provides access to multiple heterogeneous data collections through a centralized metadata registry. EDI manages two member nodes for DataONE.

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

No comment

Significant Changes

Summary of Significant Changes Since Last Application if applicable.

Compliance level:

Not Applicable - 0

Response:

-

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

Other Relevant Information

You may provide other relevant information that is not covered by the requirements.

Compliance level:

Not Applicable - 0

Response:

Unit of storage, “data package”:

Frequently in our responses, we have referred to the concept of a “data package”, which is the core unit of storage for the EDI Repository. A “data package”, is an aggregation of science metadata, quality report, and one or more data objects. Each data package is described by a simple resource map that acts as a product manifest and a more complex OAI-ORE (Open Archives Initiative Object Reuse and Exchange) document. The search interface returns the most recent revision of a package. Obsolete versions remain available, with a banner linking to the most recent version at the top of the landing page. Once added, data packages in the EDI Repository are immutable, and all copies are available. Value added by EDI includes a Quality Report (added to every package) to assist with evaluation, plus a Digital Object Identifier added to metadata (DOI) and name normalization in the repository search interface to assist with discovery. Original uploads remain unchanged. EDI also reformats selected data objects into new packages using common harmonized formats for specific areas of synthesis. All work is within the terms of the license accepted by the data producer and within the skill set of EDI personnel, as they all have training in ecosystem science.

History:

The National Science Foundation (NSF) initially funded the infrastructure as part of the Long Term Ecological Research (LTER) Network Information System in 2009. In 2016, NSF established EDI as an independent entity to expand the data publication expertise gained from working with the LTER network to provide services to a broader community of ecosystem researchers. EDI’s most recent funding from NSF was by non-competitive renewal during 2019 [0.10, 0.11]. As of late 2021, EDI holds approximately 45,000 unique researcher-contributed datasets, with the bulk coming from LTER. Five years after the expansion, holdings of non-LTER data totaled to approximately 10 % (by number) and are increasing. Holdings represent almost 4000 scientists and over 500 funding awards. Most awards are from NSF but also include USDA and USFS, NASA, NOAA, EPA, and DOE.

[0.1] https://www.re3data.org/repository/r3d100010272

[0.2] https://lternet.edu

[0.3] https://beta.nsf.gov/funding/opportunities/long-term-research-environmental-biology-ltreb

[0.4] https://beta.nsf.gov/funding/opportunities/macrosystems-biology-and-neon-enabled-science-msb-nes

[0.5] https://www.obfs.org/

[0.6] https://water.ca.gov/Programs/Environmental-Services/Interagency-Ecological-Program

[0.7] https://www.idigbio.org/content/data-help-desk-ecological-society-america

[0.8] https://esipfed.org

[0.9] https://www.neonscience.org/impact/observatory-blog/neon-program-enters-collaboration-environmental-data-initiative

[0.10] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1931143

[0.11] https://www.nsf.gov/awardsearch/showAward?AWD_ID= 1931174

[0.12] O'Brien, M., Smith, C. A., Sokol, E. R., Gries, C., Lany, N., Record, S., & Castorani, M. C. (2021). ecocomDP: A flexible data design pattern for ecological community survey data. Ecological Informatics, 101374. DOI:10.1016/j.ecoinf.2021.101374

[0.13] https://www.cuahsi.org/

[0.14] https://edirepository.org/news/news-20211102.01

Reviews

Reviewer 1:

Compliance level:

Not Applicable - 0

Comments:

Reviewer 2:

Compliance level:

Not Applicable - 0

Comments:

No comment

Organizational Infrastructure

R1 Mission/Scope

The repository has an explicit mission to provide access to and preserve data in its domain.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

EDI's mission is to preserve ecological and environmental data for open and reproducible science, for use in further synthesis across space and time, aiding the assessment of environmental change and its consequences [1.1]. This mission is assigned by the National Science Foundation (NSF), and EDI is included in proposal announcements for several NSF programs [e.g., 1.2, 1.3]. EDI's key services to the scientific community include technical expertise to ensure that environmental and ecological data are well curated and accessible for discovery and re-use well into the future. We assist researchers from field stations, individual laboratories, and research projects of all sizes to archive and publish their environmental data.
EDI is built on the premise that data published in a trustworthy and accessible repository provide significant benefits to scientific progress, society in general, and the careers and research of individual scientists. EDI was founded by NSF in 2016 as a collaborative initiative between the University of New Mexico and the University of Wisconsin. The project, which is funded by the National Science Foundation (NSF), grew out of over 30 years of data management experience in the LTER community. EDI now supports and enables data curation and reuse for a broad community of ecological data providers and users, including but not limited to scientists funded by NSF.
The EDI repository currently provides access to almost 45,000 unique data packages. Of the data packages in the repository, 65% are from early, one-time LTER-wide efforts; namely, the EcoTrends synthesis project [1.4] and Landsat imagery acquisitions. The remaining 35% were contributed by almost 4000 scientists and curated by LTER and EDI information managers with metadata increasingly being provided by the researchers themselves. EDI archives (and makes accessible) all revisions of all data packages, and so the total number managed is nearly 80,000 when all revisions are counted.

[1.1] https://edirepository.org/about/about-edi#vision-mission
[1.2] https://www.nsf.gov/pubs/2022/nsf22504/nsf22504.htm
[1.3] https://www.nsf.gov/pubs/2022/nsf22543/nsf22543.htm
[1.4] Peters DC, Fraser WR, Kratz T, Ohman MD, Rassweiler A, Holbrook SJ, Schmitt RJ. 2013. Cross-site comparisons of state-change dynamics. Long-Term Trends in Ecological Systems: A Basis for Understanding Responses to Global Change. :36-41.

1. https://edirepository.org/about/about-edi#vision-mission
2. https://www.nsf.gov/pubs/2022/nsf22504/nsf22504.htm
3. https://www.nsf.gov/pubs/2022/nsf22543/nsf22543.htm
4. https://naldc.nal.usda.gov/catalog/7048718

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Accept at level 4
Compliance Level Comment: Accept at level 4.

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R2 Licenses

The repository maintains all applicable licenses covering data access and use and monitors compliance.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

Because the EDI repository's mission is making data available for reuse and synthesis, it needs rights to copy, store, and redistribute data and metadata. The Data Policy statement [2.1] declares an understanding between EDI and the individual(s) responsible for data submitted. EDI strives to make environmental research data open and accessible to the general public without undue restrictions or barriers. Although EDI advocates for open and unfettered access to data packages, we do not forbid data providers from declaring more restrictive licensing agreements. Data providers should include a statement of Intellectual Rights in the metadata of their submissions. If they do not, EDI adds a default declaration of intellectual rights to the data package metadata which is based on the Creative Commons CC0 "No Rights Reserved" waiver [2.1].

We also recognize that some data may require limited access while it is under review or during manuscript preparation, and in these cases we support access control to data to limit exposure to appropriate users. Such control must be specified in the data package metadata. EDI will also accept data that requires a permanent embargo due to issues of sensitivity (e.g., the location of endangered species or antiquities). If data are to be submitted to the EDI Data Repository with restricted access, we request that an explanation of the data embargo, including if and when the data will be made available to the general public, be provided in the data entity description field of the data package metadata. EDI periodically reviews restricted data to determine if embargos continue to be justified.

EDI does not knowingly accept data that is protected by Federal, State, or local laws (e.g., FERPA, HIPAA, or IRB restrictions on human subject data). We require that the individual or individuals responsible for submitting science data packages to EDI for the purpose of data publication and archive acknowledge that such science data and metadata is not restricted by any governing laws, and curators are trained to recognize such data.

Although EDI will enforce access control of data as specified in the data package metadata, we cannot guarantee the privacy of such information (e.g., account names). In addition, science metadata often contains personal data of individuals involved in scientific research. These personal data may be available to other EDI customers and the general public through an EDI website. Depositors are asked to ensure that personal information within science metadata is included only with the explicit knowledge and permission of the individual or individuals it affects.

[2.1] https://edirepository.org/about/edi-policy#data-policy

1. https://edirepository.org/about/edi-policy#data-policy

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Accept at level 4
Compliance Level Comment: Accept at level 4.

R3 Continuity of access

The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

Compliance level:

The repository is in the implementation phase - 3

Response:

EDI's operations are funded by the National Science Foundation (NSF) [3.1, 3.2]. NSF has committed to preservation and publication of research data [3.3]; they have funded repositories like EDI for decades (see Background) and require all research proposals to include data management plans specifying where data will be published for sharing. NSF support for EDI is evidenced by EDI's inclusion in proposal announcements for many NSF programs [3.4].

EDI's responsibility for its holdings is indefinite, as the data we hold will have value to ecosystem researchers into the unforeseeable future. For the near-term (3-5 years), infrastructure includes redundant storage arrays (deployed in separate zones of the UNM campus), redundant virtual servers, and metadata and data that are replicated to Amazon's S3 Glacier online storage facility as a backup in the event that local storage suffers catastrophic failure.

We also recognize that over decadal time periods, it is extremely difficult to predict funding for sustained operations, and some EDI activities might be sponsored by other sources. Therefore, EDI is engaged in several activities geared toward sustainability, for example the Sustainable Infrastructure program of the Ecological Society of America [3.5] and the Development to Product (D2P) program at the University of Wisconsin [3.6]. EDI uses these programs to quantify the work involved in publishing data to better predict future costs.

At present, EDI depends entirely on support from the NSF, and while we are confident that we can justify continued backing, some EDI activities might be sponsored by other sources, which would help ensure stable, long-term access to data for the communities that already rely on EDI. With this in mind, we are developing a business model that identifies the core services required for EDI's sustained operation and engages additional sources of support to address current and future user needs beyond the identified core services. We have already collected much of the information needed to inform development of this new business model. We also have plans for replication of data holdings, to preserve data in the case of a major change in circumstances, such as complete cessation of funding. Given the high granularity of our EML metadata and standard formats for data, holdings can be easily transformed to other repositories if necessary. EDI is finalizing an agreement with the Dryad Digital Repository [3.7] to replicate all data into Dryad to serve as a permanent preservation copy of the data, should the need arise. Details are forthcoming, but EDI expects that data flow from EDI to Dryad will begin in late 2022.

[3.1] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1931143
[3.2] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1931174
[3.3] https://www.nsf.gov/bfa/dias/policy/dmp.jsp
[3.4] https://www.nsf.gov/pubs/2022/nsf22541/nsf22541.htm
[3.5] https://esa.org/sbi
[3.6] https://d2p.wisc.edu
[3.7] https://datadryad.org

1. https://www.nsf.gov/awardsearch/showAward?AWD_ID=1931174
2. https://www.nsf.gov/bfa/dias/policy/dmp.jsp
3. https://www.nsf.gov/pubs/2022/nsf22541/nsf22541.htm
4. https://esa.org/sbi
5. https://d2p.wisc.edu
6. https://datadryad.org
7. https://www.nsf.gov/awardsearch/showAward?AWD_ID=1931143

Reviews

Reviewer 1:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Reviewer 2:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Response Text: Accept at level 3
Compliance Level Comment: Accept at level 3.

R4 Confidentiality/Ethics

The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

EDI explicitly does not accept data with disclosure risk (e.g., human subject data), and this policy is posted [4.1] and curators recognize projects which potentially include restricted data, and resolve issues with the depositors. EDI operates under the understanding that depositors will have complied with requirements and guidelines of their respective institutions and funders in choosing an appropriate repository to publish their data. Although most submissions are published with public access, EDI's infrastructure supports restrictions to individuals or groups with explicit permissions, and also supports provenance so that the source data can be traced (see section R15, Technology).

Typically, some personal data is found within science metadata in the form of contact or originator information, which is critical for consumers of science data and metadata to better determine its nature and origin when ascertaining fitness for use. EDI does not actively collect such data; it is provided by depositors who wish to publish and archive science data and metadata. EDI requires depositors to acknowledge that the owners of this personal data have agreed to its release as part of the publication process. Science metadata may also contain identifiers that enable the processing of access control. All collected personal data are transmitted using HTTP SSL encryption when on the open Internet and restricted behind EDI system firewalls within EDI repository [4.2]. If confidentiality issues arise (that might impact policies), these are discusses among the entire team during regular meetings (see section R12, Workflows).

Further, EDI personnel have taken a leadership role in developing guidelines for repositories on the application of principles of data governance and sovereignty related to Indigenous Populations [4.3], and will be well placed to implement further guidelines as these arise.

[4.1] https://edirepository.org/about/edi-policy#data-policy
[4.2] https://edirepository.org/about/edi-policy#privacy-policy
[4.3] O'Brien, M.; Stall, S.; Duerr, R.; Downs, R.; Tarrant, P.; Antognoli, E. (2021): ESIP and CARE Principles - ESIP summer 2021. ESIP. https://doi.org/10.6084/m9.figshare.16926739.v1

1. https://doi.org/10.6084/m9.figshare.16926739.v1
2. https://edirepository.org/about/edi-policy#data-policy
3. https://edirepository.org/about/edi-policy#privacy-policy

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Accept at level 4
Compliance Level Comment: Accept at level 4.

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R5 Organizational infrastructure

The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

EDI is supported by NSF awards to two collaborating institutions: the University of New Mexico (UNM) and the University of Wisconsin (UW). Principal investigators at each institution oversee technical operational infrastructure, and curation and outreach at UNM and UW, respectively. The award funds all basic functions and personnel, which are distributed across the United States. EDI holds weekly meetings to review progress, set new tasks, and address issues, a model that has worked well. Staff expertise covers the full range of expertise needed: curation, software development, community outreach [5.0]. Combined, staff members have decades of experience working with ecosystem-level science and the data it produces, through its history with the LTER and other networks. Travel is covered for staff members to attend scientific meetings, to present their work, to represent EDI, and explore collaborations. These societies include the Research Data Alliance [5.1], the American Geophysical Union [5.2] and the Ecological Society of America [5.3]. EDI also supports staff in personal development and training such as Software Carpentries, and two of our staff members are certified Carpentries [5.4] instructors.

Affiliations: EDI is an active member of the DataONE community [5.5], maintaining two member nodes for metadata contributions. EDI staff members are involved in the Council of Data Facilities (on the executive board) [5.6], and lead several work groups of the Earth Systems information Partners (ESIP, [5.7]). EDI is also a signatory of the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS) Statement of Commitment [5.9], and the Enabling FAIR Data Commitment Statement in the Earth, Space, and Environmental Sciences]. EDI is listed in the Registry of Research Data Repositories [5.9]. Actively engaging in these informatics communities of practice allows EDI to access both domain science and information science expertise, while maintaining awareness of novel technologies and best practices.

EDI also has a formal collaboration with the NSF’s NEON program [5.10]. This joint initiative promotes data accessibility and usability in the environmental sciences by creating tools, templates, and standards to simplify data synthesis from the NEON program, the Long Term Ecological Research Network (LTER), and others. EDI also provides a place for individual researchers to publish derived data from the NEON collections.

[5.0] https://edirepository.org/about/about-edi#team
[5.1] https://www.rd-alliance.org/
[5.2] https://www.agu.org/
[5.3] https://esa.org
[5.4] https://carpentries.org/
[5.5] https://dataone.org
[5.6] https://www.earthcube.org/council-of-data-facilities
[5.7] https://esipfed.org
[5.8] https://copdess.org/enabling-fair-data-project/commitment-statement-in-the-earth-space-and-environmental-sciences/
[5.9] https://www.re3data.org/repository/r3d100010272
[5.10] https://www.neonscience.org/

1. https://edirepository.org/about/about-edi#team
2. https://www.rd-alliance.org/
3. https://www.agu.org/
4. https://esa.org
5. https://carpentries.org/
6. https://dataone.org
7. https://www.earthcube.org/council-of-data-facilities
8. https://esipfed.org
9. https://copdess.org/enabling-fair-data-project/commitment-statement-in-the-earth-space-and-environmental-sciences/
10. https://www.re3data.org/repository/r3d100010272
11. https://www.neonscience.org/

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Accept at level 4
Compliance Level Comment: Accept at level 4.

R6 Expert guidance

The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either in-house, or external, including scientific guidance, if relevant).

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

EDI’s management team includes active ecological researchers to provide valuable connections to the research community. All staff have a scientific background in addition to their informatics expertise. EDI receives external feedback on operations through several avenues. Our advisory board ([6.0]) is composed of ecologists and informatics professionals, who are selected mainly via suggestions from the community. We engage with them during regular meetings and reviews, receiving insightful advice and guidance. Since the advisory board encompasses a wide range of domains (ecology, data management, software engineering), meetings are generally held virtually. This allows for a) more flexibility to tailor discussion to specific areas of expertise, and b) shorter meetings overall to respect advisors’ schedules. The advisory board terms are not constrained, although they typically are 3 years. The longest serving advisory board member has served for 6 years.

We also engage data management colleagues (such as information managers of the LTER Network) and customers through regular Town Hall style meetings, who bring to us a working perspective of our operation and suggestions for improvement. Most Town Hall meetings are held online, three times annually. When possible, for example at the LTER All Scientists Meetings, EDI holds Town Halls in person (e.g., [6.2]). At scientific meetings, EDI is a founding member of a coalition of informatics professionals who run “Data Help Desk” events [6.1], which allow us to engage with the broad scientific and data management community.

[6.0] https://edirepository.org/about/edi-advisory-board
[6.1] https://www.idigbio.org/wiki/index.php/ESA_2020_Data_Help_Desk
[6.2] https://2022lterasm.sched.com/event/13573

1. https://edirepository.org/about/edi-advisory-board
2. https://www.idigbio.org/wiki/index.php/ESA_2020_Data_Help_Desk
3. https://2022lterasm.sched.com/event/13573

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Accept at level 4
Compliance Level Comment: Accept at level 4.

Digital Object Management

R7 Data integrity and authenticity

The repository guarantees the integrity and authenticity of the data.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

EDI requires all data that is submitted to the repository for archive be accompanied with an Ecological Metadata Language (EML) metadata document [7.1]. Only vetted users may submit data to the repository and must authenticate using EDI LDAP credentials. Upon upload, both the data and metadata objects have checksum fixity values computed (MD5 and SHA-1) and stored in a repository database for future object validation (See section R9, Storage). Data submitters are strongly encouraged to use EML's metadata structures for documenting data integrity checksum values (these are added by default in EDI's EML creation software, see Section R12, Workflows). Checksums in EML metadata are then used to ensure data objects have not been compromised due to technical issues during upload and to confirm the authenticity of the data as referenced in the EML metadata. EDI performs continuous fixity checks against all data and metadata objects archived in the repository. This information is logged and reported to system administrators on a daily basis (see Section R9, Storage).

Data and metadata are processed through a series of quality verification checks prior to acceptance into the repository [7.2]. This automated verification ensures that data are accurately described by the metadata, including attribute-level verification of tabular data for structure and data type. Gross errors will result in the rejection of a data package for permanent archiving. A quality verification report is created and archived as part of the final data package. This report is available to all users. Some metadata (e.g., text description) cannot easily be checked by automated processes, and these are evaluated by EDI curators (see section R12 Workflows).

The EDI Data Repository enforces strict versioning of both data and metadata. That is, once a data package is published and archived, all data and metadata within the package become immutable. As such, updates require a new version to be added by an authorized submitter, which goes through the same quality verification as previous versions. New versions are assigned an integer value greater than the previous version, which becomes an indelible part of the data package identifier. All versions of a data package series remain accessible through the repository REST API, which allows access to individual package components using repository URIs or to the data package landing page using Digital Object Identifiers (DOI). The REST API also makes information about each version, including listing all versions, accessible to all users.

Provenance metadata is encoded within the EML metadata and captures creator information (e.g., role, contact, location), creation date-time information, and may be extended to document progenitor information of data and processes used in the creation of the current data package. Submitter information is logged separately since this party may not be attributed in metadata.

[7.1] Jones,M.B, M. O'Brien, B. Mecum, C. Boettiger, M. Schildhauer, M. Maier, T. Whiteaker, S. Earl, S. Chong. 2019. Ecological Metadata Language version 2.2.0. KNB Data Repository. doi:10.5063/F11834T2
[7.2] O'Brien, M. D. Costa, and M. Servilla. 2016. Ensuring the quality of data packages in the LTER network data management system. Ecological Informatics, 36: 237-246. DOI: 10.1016/j.ecoinf.2016.08.001

1. https://doi.org/10.1016/j.ecoinf.2016.08.001
2. https://doi.org/10.5063/F11834T2

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Accept at level 4
Compliance Level Comment: Accept at level 4.

R8 Appraisal

The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The core unit of EDI's holdings is a "data package", which is an aggregation of science metadata, quality report, and one or more data objects (tables, rasters, images, or documents), an EML metadata document, a quality report, other internal management documents, and possibly processing code. Coherence is an important concept in data package design [8.1]; data entities that share high-level metadata such as methods, sampling sites and people can be efficiently grouped together in one data package. A data package is described by a simple resource map that acts as a product manifest and a more complex OAI-ORE (Open Archives Initiative Object Reuse and Exchange) document. Data packages are versioned (see Section R7, Data Integrity), and searches return the most recent revision of a package (obsolete versions remain available, with a banner linking to the most recent version at the top of the landing page).

Per its mission, EDI accepts data from all fields of environmental science. The repository was designed to accommodate all data produced by the LTER Network, which generates a wide range of data types related to ecosystem-level research, including but not limited to investigations of population dynamics, productivity, disturbance patterns, and organic and inorganic matter accumulation and transport. This same diversity of data types is reflected in EDI's acquisitions since 2016. Therefore, there is no explicit collection development policy that guides data selection, and EDI's infrastructure is capable of accepting data in any digital format. Further, since EDI serves the LTER Network and each LTER site supports a data manager who interacts directly with the repository, selection of data is left to those groups. EDI organizes working groups of LTER data personnel to develop best practices for data selection, arrangement and metadata content [8.2]

However, since our mission is to provide environmental data for reuse, data with tight restrictions on reuse is not accepted. These would include data on human subjects (e.g., biomedical). All data packages are screened by a curator to confirm their applicability to this rule. It is not within EDI's purview to assess a dataset's scientific merit or the accuracy of its measurements.

Each metadata record is carefully screened to ensure the completeness and understandability of the data, both automated and manually. The EML metadata specification is highly granular, allowing observations to be described in great detail. The automated mechanisms to ensure data's congruence with the asserted metadata were developed in 2010-2013 with guidance from contributing data managers [8.3]. Those screening mechanisms have continued to be adapted, with additional feedback [8.4]. Automated checking produces a text report which can be reviewed by the depositor. The report for every data package is also available from its landing page, under the "Resources" link "View Quality Report" (example, [8.5]), as well as with the package download. Metadata are also examined by an EDI curator for features that are difficult to assess by computer, e.g., to confirm that a data package contains metadata sufficient for later interpretation, e.g, a descriptive title, comprehensive abstract, accurate temporal and geospatial coverages and comprehensive methods. Taxonomic metadata may also be requested or added.

Although data in any digital form can be accepted, EDI recommends that data be published in open formats, e.g, ASCII tables (as these will be readable long into the future) and discourages use of any proprietary format such as spreadsheets. The presence of text tables is confirmed by the curators, and as of 2019, about 80% of data entities were text tables. Other formats include GIS, imagery or zipped files (often large tables or NetCDF), and for these, we are considering mechanisms to help to ensure their future readability of non-tables (See R10, Preservation Plan).

For data tables, additional data congruence checks confirm that tables match the asserted metadata (see section R7, Data Integrity). Descriptive metadata for the data files and variables are also highly recommended (although not absolutely required), including a checksum, size (as used for integrity checking, see section R7, Data Integrity). Data variable information required by the schema includes a name, label and description, with units formally linked to base SI units for numeric data. Definition of controlled vocabularies for categorical data, and missing value codes are encouraged. All tools for generating metadata provided by EDI add these elements by design, e.g., EMLAssembly line [8.6] and ezEML [8.7]. Further details about ezEML and guidance during data package creation are covered in section.R12, Workflows.

Issues that may arise due to insufficient metadata or proprietary formats are resolved by communicating with the depositor, with changes suggested by the EDI curator. Once educated about the benefits, most depositors are motivated to provide data and metadata in a form with long-term usability (see R10, Preservation, for more information about format obsolescence).

[8.1] https://edirepository.org/resources/resources-for-data-authors#
[8.2] https://ediorg.github.io/data-package-best-practices/data-package-design-for-special-cases.html
[8.3] O'Brien, M. D. Costa, and M. Servilla. 2016. Ensuring the quality of data packages in the LTER network data management system. Ecological Informatics, 36: 237-246. DOI: 10.1016/j.ecoinf.2016.08.001
[8.4] https://github.com/EDIorg/ECC
[8.5] https://portal.edirepository.org/nis/reportviewer?packageid=knb-lter-jrn.202.1
[8.6] https://ediorg.github.io/EMLassemblyline

[8.7] http://ezeml.edirepository.org

1. https://edirepository.org/resources/resources-for-data-authors
2. https://ediorg.github.io/data-package-best-practices/data-package-design-for-special-cases.html
3. https://doi.org/10.1016/j.ecoinf.2016.08.001
4. https://github.com/EDIorg/ECC
5. https://portal.edirepository.org/nis/reportviewer?packageid=knb-lter-jrn.202.1
6. https://ediorg.github.io/EMLassemblyline
7. https://ezeml.edirepository.org/eml/auth/login

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Accept at level 4
Compliance Level Comment: Accept at level 4.

R9 Documented storage procedures

The repository applies documented processes and procedures in managing archival storage of the data.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

An archived data package begins with the submission of an Ecological Metadata Data (EML) document to the repository. This ingest process may occur either directly through the repository's create or update REST API methods (see Section R16), or indirectly through the EDI Data Portal web browser interface when an EML document is uploaded to the Data Portal by an end user followed by the Data Portal's use of the REST API on behalf of the user. Next, the repository validates the EML document and then proceeds to copy all data objects to a local cache for processing (including fixity checksum generation and verification). After the data and EML metadata are cached on the repository's local filesystem, quality verification of data, metadata, and metadata and data congruence (i.e., does the metadata accurately describe the data) take place. If no gross errors occur during quality verification, the data and EML metadata are moved to a permanent block storage device, a data package DOI is created and registered with DataCite [9.1], and authorization information and resource information (including storage location and fixity) are stored in a system database. This completes the archival process.

Archival data and EML metadata are written directly to a block storage filesystem in a directory hierarchy defined by the repository's internal identification system. This block storage system consists of a 56TB NetApp storage array hosted by the Center for Advanced Research Computing at the University of New Mexico (UNM) [9.2], Albuquerque, New Mexico USA. Simultaneously, these data and metadata are replicated to an identical NetApp storage array at the university's Central Information Technology center, also in Albuquerque, and are available only for fault recovery purposes. In addition to the primary and secondary block storage systems, all data and metadata are replicated to a fixed backup staging system located in the UNM EDI office building, which is used for creating removable backup SSD drives that are stored in an environmentally protected safe. Also from this staging system, a compressed version of the data package is written to Amazon Web Services S3 Glacier online storage for off site redundancy. Consistency between the primary and secondary block storage systems occurs through internal controller software, while other replicated storage undergoes automatic fixity checks on a continuous basis. Consistency of data packages on Glacier occur rarely and are a manual process. The risk management strategy used to inform this process is primarily "Lots of Copies Keeps Stuff Safe" (LOCKSS), in addition to ensuring copies are geographically distributed. Longer term storage is being coordinated with the Dryad Digital Repository (see section R3).

Many of the steps above are automated. Technical documentation is managed using the “Read the Docs” application, for example: the basic PASTA+ design ([9.3]), and specifics of PASTA+ web services, such as the Gatekeeper, or entry point for all external interactions ([9.4]), and Operating Procedures (9.5]). Other parts of the workflow are documented internally, and more information can be made available on request. Changes to PASTA+ infrastructure are handled as for other repository workflows (see Section R12), and use typical productivity tools such as issue trackers, software repositories (GitHub) and task managers (e.g., Trello). Changes to specific PASTA+ components or procedures are discussed during weekly group meetings so that unforeseen consequences or edge cases can be considered, and issues are commonly resolved by consensus. The PASTA+ system design and its use has been the subject of several webinars ([e.g., [9.7], [9.8]).

[9.1] https://datacite.org
[9.2] https://carc.unm.edu/
[9.3] PASTA+ design overview: https://pastaplus-core.readthedocs.io/en/latest/doc_tree/pasta_design/index.html
[9.4] Gatekeeper web service design: https://pastaplus-core.readthedocs.io/en/latest/doc_tree/pasta_design/gatekeeper.html
[9.5] Operating procedures: https://pastaplus-core.readthedocs.io/en/latest/doc_tree/pasta_sop/index.html
[9.7] https://edirepository.org/webinars/webinars-20180410.00
[9.8] https://edirepository.org/webinars/webinars-grid

1. https://datacite.org
2. https://carc.unm.edu/
3. https://pastaplus-core.readthedocs.io/en/latest/doc_tree/pasta_design/index.html
4. https://pastaplus-core.readthedocs.io/en/latest/doc_tree/pasta_design/gatekeeper.html
5. https://pastaplus-core.readthedocs.io/en/latest/doc_tree/pasta_sop/index.html
6. https://edirepository.org/webinars/webinars-20180410.00
7. https://edirepository.org/webinars/webinars-grid

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Accept at level 4
Compliance Level Comment: Accept at level 4.

R10 Preservation plan

The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance level:

The repository is in the implementation phase - 3

Response:

EDI’s preservation policies and processes are posted on the website [10.1]. Our primary mission is data preservation and data access, and so high-quality data management is essential. All submitted data and metadata are reviewed and edited before acceptance to ensure high-quality data products are available to the research community [10.2]. Basic principles include the use of open standards to promote open science, such as EML metadata and data files in text formats such as ASCII comma-separated values [10.3], replication of metadata [10.4, 10.5, 10.6] and backup of both data and metadata (see Section R9, Storage), plus strong versioning and frequent auditing [10.7]. Technical details (including software choices and maintenance to forestall its obsolescence) are covered in Section R15. Our process for curation of data packages and communication with depositors is described in Section R12, Workflows. Curation of the 80% of holdings that are data tables already includes data-metadata congruence checking (See Section R8, Appraisal, Section R11 Data Quality), and we are developing mechanisms for checking other formats, e.g., a format checker based on mime types, to identify data entities which might be proprietary or out of date.

By uploading content, users grant the EDI Repository all rights needed to copy, store, redistribute, and share data, metadata, and any other content. By marking content as publicly available, users grant the EDI Repository and other future users the right to copy the content and redistribute it to the public without restriction under the terms of the CC-0 Public Domain Dedication (our default) [10.8] or the Creative Commons Attribution 4.0 International License [10.9]. The license text is recorded in the metadata for each submitted data set at the time of submission. Our use of CC-BY or CC-0 licensing also allows us to transform or reformat data where necessary to maintain its accessibility in the future.

EDI's success in achieving its goal of preserving data far into the future should also include a sustainable business model that protects data in the repository for the long-term. See section R3, Continuity of Access, for more information.

[10.1] https://edirepository.org/about/about-edi#physical-management-and-curation-of-digital-products
[10.2] https://edirepository.org/resources/resources-for-data-authors#
[10.3] https://edirepository.org/resources/resources-for-data-authors#step-1-preparing-for-submission
[10.4] https://edirepository.org/news/news-20170418.00
[10.5] https://search.dataone.org/portals/EDI
[10.6] https://search.dataone.org/portals/LTER
[10.7] https://edirepository.org/about/about-edi#risk-management
[10.8] https://edirepository.org/about/edi-policy#intellectual-rights-of-the-data-contributor
[10.9] https://creativecommons.org/licenses/by/4.0/

1. https://edirepository.org/about/about-edi#physical-management-and-curation-of-digital-products
2. https://edirepository.org/resources/resources-for-data-authors
3. https://edirepository.org/resources/resources-for-data-authors#step-1-preparing-for-submission
4. https://edirepository.org/news/news-20170418.00
5. https://search.dataone.org/portals/EDI
6. https://search.dataone.org/portals/LTER
7. https://edirepository.org/about/about-edi#risk-management
8. https://edirepository.org/about/edi-policy#intellectual-rights-of-the-data-contributor
9. https://creativecommons.org/licenses/by/4.0/

Reviews

Reviewer 1:

Compliance level:

The repository is in the implementation phase - 3

Comments:

Response Text: Accept at level 3
Compliance Level Comment: Accept at level 3.

Reviewer 2:

Compliance level:

The repository is in the implementation phase - 3

Comments:

R11 Data quality

The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality- related evaluations.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

Per our mission, EDI facilitates reuse of long lived, high quality data packages (described in Section R8, Appraisal). Deposited data packages must have a complete metadata record (i.e., in Ecological Metadata Language, EML), and completeness is ascertained by the data quality engine of the PASTA software ([11.0], and fully described in Section R15, Technical Infrastructure).

EDI screens all packages entering its long-term archive to ensure completeness and quality, and to ascertain that metadata and data are structurally congruent, i.e., that the data typing and formats expressed in metadata agree with that found in data entities [11.1]. The checking system also looks for certain data and metadata features, such as ISO date formats and keywords from known catalogs. The automated checking system generates a Quality Report, which becomes part of the data package (example links available with Section R8, Appraisal). Quality Reports are also reviewed by EDI curators prior to acceptance, because the semantics of some text fields cannot be adequately reviewed by software at this time (see Section R8, Appraisal). The system designed by EDI personnel was one of the first instances of checking used for ecological data tables (see Introduction in [11.1]), and has served as a benchmark for other data table checking systems (e.g., [11.2]). EDI's checking system continues to evolve. As originally implemented in 2013, 32 individual checks were included; currently 44 features of data packages are checked automatically by the PASTA software. EDI honors quality flags when included by the depositor, but our checking system does not apply data quality flags routinely to individual measurements. Researchers are encouraged to describe their measurement variables in accordance with practices in their scientific communities, and descriptions are reviewed manually by the data curation team (with automated checks for these features in the planning stages).

To assist with designing data packages, EDI coordinates a library of recommendations for design which are written by a working group of information managers representing contributing research groups [11.3]. These recommendations cover special cases for archiving research data based on type, format, acquisition method, and recommend practices that ensure optimal re-usability of the data. These recommendations focus on improved documentation to avoid misinterpretation, improve usability in terms of data size/volume, or to connect related data.

11.0] Servilla, M. J. Brunt, D. Costa, J. McGann, R. Waide. 2016. The contribution and reuse of LTER data in the Provenance Aware Synthesis Tracking Architecture (PASTA) data repository. Ecological Informatics, 36: 247-258. DOI: 10.1016/j.ecoinf.2016.07.003
[11.1] O'Brien, M. D. Costa, and M. Servilla. 2016. Ensuring the quality of data packages in the LTER network data management system. Ecological Informatics, 36: 237-246. DOI: 10.1016/j.ecoinf.2016.08.001
[11.2] Gordon, S. and T Habermann 2018. The influence of community recommendations on metadata completeness. Ecological Informatics, 43: 38-51 DOI: doi.org/10.1016/j.ecoinf.2017.09.005
■■[11.3] https://ediorg.github.io/data-package-best-practices/

1. https://doi.org/10.1016/j.ecoinf.2016.07.003
2. https://doi.org/10.1016/j.ecoinf.2016.08.001
3. https://doi.org/10.1016/j.ecoinf.2017.09.005
4. https://ediorg.github.io/data-package-best-practices/

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Accept at level 4
Compliance Level Comment: Accept at level 4.

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R12 Workflows

Archiving takes place according to defined workflows from ingest to dissemination.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

The business processes in place at EDI are related to data curation tasks, infrastructure improvements, and to communication with community members and the public. They use typical productivity tools such as issue trackers, software repositories (GitHub) and task managers like Trello. Management of EDI-generated products for reporting (e.g., papers, presentations, workshops held) utilizes spreadsheets and open-source bibliographic software. Because EDI personnel are distributed across the US, communication within the project is typically electronic (Slack), with weekly video meetings attended by all. The data curation team holds a second weekly meeting, and the technical team holds a weekly meeting and daily “stand up”. Ad hoc video meetings on specific topics are common, and during most years, we hold an annual in-person meeting for all EDI personnel as well. Changes to the management of workflows is a topic commonly resolved by consensus during group meetings. Personnel also report on recent activities during regular meetings, and agendas for all meetings are maintained in electronic documents and issue boards.

Data Curation: EDI personnel interact with data depositors in several different ways. First, we serve a large community of data managers and professionals (>100) from the LTER community and other field stations, and EDI input to them generally takes the form of occasional technical assistance (through an email list and ticketing system), webinars (for information of general interest) ad hoc one-on-one training for new personnel in our community, and organized working groups to develop solutions to common issues (e.g, see [12.0]). All materials are posted on the EDI website, or available in shared electronic documents.

Secondly, we accept data from independent researchers and labs. Data deposits from independent labs generally use the ezEML forms [12.1] to record metadata and assemble a data package themselves. An extensive user manual with videos accompanies ezEML. [12.1]. (Researchers can also submit text metadata via templates, although this is becoming rarer [12.2]). The ezEML system submits the data package directly to the curators. In both cases, a curator reviews metadata content to assure completeness and usability (see sections R8, R11), and stages a draft package for the depositor’s review and approval [12.3]. Data curation protocols are documented in internal electronic documents accessible to all staff via a password-accessible drive (which can be made public if requested). If necessary, the entire team is available via messaging to respond to a curator request for input for an unusual data package, and protocols are amended to record subtle changes that may have arisen.

Protocols for data transform: The data packaged as deposited (after review) is the authoritative version. EDI’s standard licensing (see Section R2) allows for transformation to produce derived products, which is occasionally carried out by EDI personnel (data scientists within the community are encouraged to create transforms). Data are selected for transformation based on their potential for synthesis, e.g., a dataset covering a long time series of organismal abundance is likely to be selected for transformation to our harmonized format, ecocomDP [12.4]. Similarly, a dataset of stream flow would be selected for transformation to the format used by the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI, [12.5]). Transformed data become a derived data package, with a provenance chain linking to the original (see sections R14, R15)

Technical protocols: Data storage procedures are fully described in section R9. Briefly, after a data package is written to physical storage data and metadata are replicated to both a permanent mirrored storage device and a removable storage device using a combination of copy and checksum verification. Reverification of checksums is accomplished by periodic monitoring. These protocols are recorded in the PASTA software documentation [12.6].

Guarding the privacy of depositors and individuals named in data package metadata has not, to date, significantly impacted our workflows. EDI requires depositors to acknowledge that the data owners have agreed to release small amounts of personal data (names, emails, login names) as part of the publication process. Our use of HTTP SSL encryption to transfer data has been stable for several years.

Communication and outreach Outreach and general communication is coordinated by a specific staff member who is charged with organizing webinars, producing a bi-monthly newsletter, maintaining the website and social media feeds, and managing an annual fellowship program to train up to 15 community members in data publication practices. Principles followed for handling of data are posted on the EDI website [12.7]. Town hall style meetings allow the community to comment on any aspect of EDI’s activities, and frequently generate suggestions and new initiatives about handling of data.

[12.0] https://edirepository.org/resources/resources-for-information-managers
[12.1] https://ezeml.edirepository.org, https://edirepository.org/resources/creating-metadata-for-publication#ezeml
[12.2] https://edirepository.org/support/contact-us
[12.3] https://edirepository.org/resources/the-review-process
[12.4] O'Brien, M., Smith, C. A., Sokol, E. R., Gries, C., Lany, N., Record, S., & Castorani, M. C. (2021). ecocomDP: A flexible data design pattern for ecological community survey data. Ecological Informatics, 101374. DOI:10.1016/j.ecoinf.2021.101374
[12.5] https://www.cuahsi.org/
[12.6] https://pastaplus-core.readthedocs.io
[12.7] https://edirepository.org/resources/resources-for-data-authors

1. https://edirepository.org/resources/resources-for-information-managers
2. https://edirepository.org/resources/creating-metadata-for-publication#ezeml
3. https://edirepository.org/support/contact-us
4. https://edirepository.org/resources/the-review-process
5. https://doi.org/10.1016/j.ecoinf.2021.101374
6. https://www.cuahsi.org/
7. https://pastaplus-core.readthedocs.io
8. https://edirepository.org/resources/resources-for-data-authors

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Accept at level 4
Compliance Level Comment: Accept at level 4.

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R13 Data discovery and identification

The repository enables users to discover the data and refer to them in a persistent way through proper citation.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

EDI excels in data discoverability in part because of the comprehensive metadata (i.e., EML) it requires to accompany scientific data. Specific elements of the metadata are indexed by Apache Solr. In addition to simple search across all metadata, users can create structured queries for creator (person, with name reconciliation or organization), geographic area (map or place name), taxon, subject, project or ID [13.1]. Searching by name (creators or project personnel) is aided by EDI's name normalization system, which reduces idiosyncratic differences in spelling and name presentations.

The incoming EML metadata is converted to Dublin Core and schema.org formats for harvest and further indexing. Specifically, EDI's schema.org metadata has been harvested by Google's Dataset Search since that service's inception [13.2]. This use of SEO metadata adds structure and meaning gleaned from science metadata to generally unstructured web pages, thereby elevating science data to the level of other first class objects on the internet.

In addition to harvesting by the generic Google registry, EDI manages two nodes for the DataONE registry of metadata [13.3] using the turn-key "Generic Member Node" (GMN) software stack (LTER since 2014 and EDI since 2016). Analysis of access statistics from each instance indicates approximately 2.75M datasets were viewed over their lifetimes, with nearly 2.5M data downloads through DataONE.

Digital object identifiers (DOI) are assigned to all datasets (see section R7, Data Integrity) in collaboration with DataCite [13.4] to enable data citation and data provenance documentation and both are tracked by EDI for maximum credit to data providers. The landing page for every data package in the EDI repository contains a suggested citation that follows the ESIP guidelines for dataset citation [13.5].

[13.1] https://portal.edirepository.org
[13.2] https://datasetsearch.research.google.com/
]13.3] https://dataone.org
[13.4] https://datacite.org
[13.5] ESIP Data Preservation and Stewardship Committee. 2019. Data Citation Guidelines for Earth Science Data. Ver. 2. Earth Science Information Partners. https://doi.org/10.6084/m9.figshare.844181.v1

1. https://doi.org/10.6084/m9.figshare.8441816.v1
2. https://portal.edirepository.org
3. https://datasetsearch.research.google.com/
4. https://dataone.org
5. https://datacite.org

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Accept at level 4
Compliance Level Comment: Accept at level 4.

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R14 Data reuse

The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

Part of EDI's primary mission is to make data available for reuse. The EDI repository monitors activity of data downloads on a regular basis. Over the second half of 2021, average downloads rates of about 20,000 data files per month were recorded. Once published in the repository, data packages become immutable (see section R7 Data Integrity) thereby providing a reproducible data source for science consumers. Data access events can also be monitored by data depositors, via the Audit Services of the PASTA API (see section R15 Technical). To help ensure reuse in the future, EDI encourages data objects to be stored in text formats, avoiding the likelihood that proprietary formats will become obsolete. Our metadata checking system was designed specifically to allow machine reading of text data during reuse (see Section R11 Data Quality, R12 Workflows). Further, EDI's data explorer, DeX [14.1] provides views into EDI's tabular data: 1) a statistical profiler that displays detailed information for each column; 2) a filter and subsetting application (including a new EML metadata document describing the subset); and 3) a scatter and line plotting application for a visual glimpse into data trends. These tools assist scientists in assessing a dataset's fitness for use.

Perhaps the greatest recognition of data reuse is its attribution within a peer-reviewed article which is most accurately determined by citations with DOIs (EDI data packages have received DOIs since 2013). As the practice of data citation is still taking hold within research communities, citation rates are still difficult to quantify. However, hundreds of journal articles (identified by Google Scholar) now cite data using an EDI-assigned DOI, with the rate doubling approximately annually. EDI also maintains a feature to explicitly annotate the relationship between a data package (in EDI) and a published paper, which are published to DataCite [14.2] as an update to data package metadata. To date (late 2021), over 1500 datasets have been cited by 1300 papers, spanning the time period 2016 - 2021. Another important aspect of reusability (to consumers) is the ability to track provenance of a dataset - to know its derivation, or if data has been used in other workflows. By design, PASTA supports references to data sources, and includes web services to make provenance relationships discoverable to outside data sources (see Section R16, Technical).

Our work with synthesis scientists has highlighted the need for high-priority primary data to be converted to analysis ready formats. These harmonized, analysis-ready datasets will eliminate a time-consuming step in data synthesis. EDI led one such project and produced over 70 ecological community observation datasets in the same data model [14.3]. The model has also been applied by the NEON project to their community surveys. Outputs are also being made available to the large data aggregator, the Global Biodiversity Information Facility (GBIF [14.4]). Another harmonization project will focus on weather station data, which will be formatted in the Observation Data Model (ODM, Tarboten et al., 2008) and made available on the CUAHSI data portal. The systems producing these harmonized products make use of the PASTA event notification system (see section R15 Technical).

[14.1] https://edirepository.org/news/news-20220128.00
[14.2] https://datacite.org
[14.3] O'Brien, M, C.A. Smith, E.R.Sokol, C. Gries, N. Lany, S. Record, M.C.N. Castorani. 2021. ecocomDP: A flexible data design pattern for ecological community survey data. https://doi.org/10.1016/j.ecoinf.2021.101374
[14.4] https://gbif.org

1. https://edirepository.org/news/news-20220128.00
2. https://datacite.org
3. https://doi.org/10.1016/j.ecoinf.2021.101374
4. https://gbif.org

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Accept at level 4
Compliance Level Comment: Accept at level 4.

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Technology

R15 Technical infrastructure

The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

EDI's Data Repository infrastructure, PASTA, follows a Service Oriented Architecture (SOA) design pattern that exposes functionality through a REST web-service Application Programmable Interface (API). Community-based technologies incorporated into the EDI Repository include Ecological Metadata Language (EML), the Open Archives Initiative (OAI), Apache Solr, Open Researcher and Contributor ID (ORCID),and schema.org vocabularies.
It is designed using a metadata-driven workflow that relies on content within an EML document to perform various repository tasks, including data centralization and quality assessment. PASTA was designed in 2009 by the former LTER Network Office and guided by input from information managers and research scientists. Core web services use the Java programming language and Servlet framework to expose API methods. Incoming web service calls go through a "Gatekeeper" service ([15.1]) which forwards the calls to the appropriate PASTA service: "Data Package Manager," "Audit Manager," or "Search" ([15.2]). PASTA utilizes other technologies, including Schema.org, OAI, ORCID identifiers, Apache Solr, and OAuth/OpenIDConnect. The software stack, web services, and the EDI Data Portal website are deployed on independent virtual machines. Collectively they comprise a complete PASTA data repository tier and are replicated three times by EDI as development (developer sandbox), staging (customer sandbox), and production systems. Virtual machines (VM) run the Ubuntu operating system for optimal performance and stability with additional software packages (e.g., Nginx, Apache Tomcat). EDI manages 15 VMs for the EDI Data Repository infrastructure, plus approximately 36 others to support ancillary services.

The multi-tier architecture supports seamless deployment of new features and bug fixes with limited impact on production uptime and usability. EDI's Data Repository has been in continuous operation since January 2013 and has never experienced a significant service outage. However, EDI does perform scheduled system maintenance and software updates on a weekly basis, which may result in a nominal downtime of up to 2 hours from 7 to 9 PM (US Mountain Time Zone). An alert about this event is published 12 hours prior on the EDI Data Portal. Security updates may occur outside of the scheduled maintenance period when deemed critical. System maintenance requiring longer periods of downtime occur rarely (1-2 times per year) and generally last for no more than 24 hours. Notification of these longer events are published to the EDI Data Portal and sent out using customer email at least one week prior.

EDI's hardware infrastructure was provisioned in Spring 2020 and is co-hosted at the Center for Advanced Research Computing (CARC) and Central IT (CIT) at the University of New Mexico, Albuquerque, NM USA. The infrastructure includes two Cisco blade servers configured with 7.2 TB of high-speed storage for PASTA and related service virtualization, along with two 56 TB NetApp storage arrays that host our environmental data archive. Each storage array is deployed within separate zones of campus for reliability and redundancy purposes. VMware ESXi (Version 7) is used on both Cisco Blade Servers, and the infrastructure has a minimum 6 year life horizon. In addition, all metadata and data are replicated to Amazon's S3 Glacier online storage facility as a backup in the event of a local catastrophic failure in data storage.

EDI's technical support staff is available 7 days a week during extended business hours (6 AM to 10 PM MTZ) and can receive electronic notification of any system malfunction. General health of the system can be monitored publicly from the EDI Repository service dashboard [15.3]

Provenance/workflow processes: Provenance is used to explicitly define the relationships between objects that are not of the same version chain, but which have a computational relationship (such as a script that derives a data file). These provenance relationships are defined by the data depositors, and stored within the EML metadata. The PASTA software allows a program or script to be triggered when a data package is updated, providing a mechanism for keeping derived data packages up to date with their originals.

[15.1] https://pastaplus-core.readthedocs.io/en/latest/doc_tree/pasta_design/gatekeeper.html

[15.2] https://pastaplus-core.readthedocs.io/en/latest/doc_tree/pasta_api/index.html

[15.3] https://dashboard.edirepository.org/dashboard/

1. https://pastaplus-core.readthedocs.io/en/latest/doc_tree/pasta_design/gatekeeper.html
2. https://pastaplus-core.readthedocs.io/en/latest/doc_tree/pasta_api/index.html

3. https://dashboard.edirepository.org/dashboard/

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Accept at level 4
Compliance Level Comment: Accept at level 4.

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

R16 Security

The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

Critical infrastructure of the EDI Data Repository, including the primary data archive storage system and mirror, is co-hosted at the Center for Advanced Research Computing (CARC [16.1]) and Central IT (CIT [16.2]) at the University of New Mexico, Albuquerque, NM USA. As such, infrastructure security and safety fall under the governance of the University of New Mexico's Office of Information Technologies per Policy 2520. For this reason, EDI follows the university's policy for risk assessment and mitigation, including cybersecurity emergency response planning.

Direct access to hardware infrastructure and data storage devices is controlled 24 hours per day 7 days per week through the use of managed digital key-card entry systems at both building (CARC and CIT) perimeters and at access points into data center machine rooms to only CIT authorized personnel. Both locations utilize state-of-the-art entry alarm systems and multiple surveillance video feeds of surrounding grounds and machine rooms. Backup data are maintained within a separate key-carded access controlled server room with access limited to only key personnel.

EDI works closely with institutional personnel (CARC and CIT) to ensure that infrastructure is accessible, performant, and secure. Network access to infrastructure is addressed at three levels: 1) all network connections are monitored and analyzed by intrusion detection software at the university's network boundary using Palo Alto technology, 2) all network connections within the CARC environment are limited through hardware firewalls that pre-filter packets between UNM IT and EDI infrastructure, and finally 3) all EDI compute infrastructure limits network connections to specific ports and inbound IP addresses through software firewalls (Linux-based iptables). Host access is limited to key personnel on a "need-to-access" policy and requires SSH key authentication (password-based logins are not permitted), in addition to specific client authorization within the firewall rule-set. Network connections to infrastructure from outside of the university require use of the GlobalProtect VPN software.

EDI monitors ongoing system state-of-health using automated state-change detection and standard real time log analysis (Elastic-Logstash-Kibana). Detected issues are transmitted automatically to administrators via email and SMS for immediate notification. Personnel coverage for general issue mitigation is provided during regular business hours and 5 days per week throughout the year and generally within one to two hour mitigation at other times; all critical security issues are addressed immediately. EDI host operating systems are consistent across all infrastructure and are updated ("patched") on a weekly basis. Security updates are applied more frequently if deemed critical by the university or by the United States Cybersecurity and Infrastructure Security Agency.

The EDI Data Repository software stack, PASTA, has been reviewed by the US National Science Foundation's Center for Trustworthy Scientific Computing [16.3] once in 2014 for general security practice compliance and again in 2019 for counsel on practices to advance our customer authentication system. To date, the EDI Data Repository has not experienced any cybersecurity breaches or infractions.

All archived "product-level" data are replicated and backed-up to multiple onsite and offsite locations. Data are automatically and immediately mirrored between the primary data storage system at CARC to a mirrored replica at UNM CIT; both sites are within the university property but are physically separated by 2-3 KM distance and using independent utilities. Data is also replicated on a daily basis to a secondary storage system located in the offices of EDI (also on campus); this replication delay is intentional and provides a time-gap insulation for backup content from potential corruption on primary storage. These data are also copied to removable storage from the secondary storage. EDI rotates the removable storage to a local offsite location on a weekly basis where the media are stored in a fire and waterproof safe. Finally, data from the secondary storage system are uploaded to the Amazon Web Services S3 Glacier online storage as a “dark-archive”. EDI follows the “Lots of copies keeps stuff safe” (LOCKSS) principle. Backup data are regularly examined to ensure their integrity and accessibility.

[16.1] https://carc.unm.edu/
[16.2] https://it.unm.edu/
[16.3] https://www.trustedci.org/

1. https://carc.unm.edu/
2. https://it.unm.edu/
3. https://www.trustedci.org/

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Response Text: Accept at level 4
Compliance Level Comment: Accept at level 4.

Applicant Feedback

R17 Applicant Feedback

We welcome feedback on the CoreTrustSeal Requirements and the Certification procedure.

Compliance level:

The guideline has been fully implemented in the repository - 4

Response:

-

Reviews

Reviewer 1:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

Thanks for more than adequately addressing my comments!

Board comment:
For recertification in 3 years the board would like to see progress on R3 (Continuity of access) and R10 (Preservation plan) needs to be fully implemented.

Reviewer 2:

Compliance level:

The guideline has been fully implemented in the repository - 4

Comments:

A strong application underpinned by what look to be well-designed services that are professionally supported, enabling the trustworthy preservation and reuse of ecological and environmental data.

Additional information and corrections have been provided and public documentation updated in response to reviewer feedback.

Board comment:
For recertification in 3 years the board would like to see progress on R3 (Continuity of access) and R10 (Preservation plan) needs to be fully implemented.