Assessment Information

CoreTrustSeal Requirements 2017–2019

Repository:	Arctic Data Center
Website:	https://arcticdata.io
Certification Date:	26 March 2020
This repository is owned by:	University of California Santa Barbara, National Center for Ecological Analysis and Synthesis

Arctic Data Center

Notes Before Completing the Application

We have read and understood the notes concerning our application submission.

True

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background & General Guidance

Glossary of Terms

BACKGROUND INFORMATION

Context

R0. Please provide context for your repository.

Repository Type. Select all relevant types from:

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

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of Repository

The National Science Foundation (NSF) Arctic Data Center currently preserves over 34 Terabytes of data and metadata representing over 5000 data sets for the Arctic research community. As of September 2019, visitors have viewed metadata descriptions over 430,000 times, and have downloaded data files over 12.5 million times. The holdings have been used and cited in many publications (including special issues of journals, theses, conference proceedings, reports, journal articles, and books), some of which are listed in the NSF Arctic Data Center publications list (https://arcticdata.io/publications).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of the Repository's Designated Community.

Arctic researchers (Arctic biological, geophysical, chemical, and sociocultural processes, and the interactions of ocean, land, atmosphere, biological, and human systems. Natural sciences, social sciences, system science, observing network, polar cyberinfrastructure, and research and policy support public relations.)

National repository system, including governmental: United States National Science Foundation (NSF) Office of Polar Programs (OPP) Section for Arctic Sciences (ARC)-funded research https://www.nsf.gov/awardsearch/showAward?AWD_ID=1546024&HistoricalAwards=false. From the submissions guideline page (https://arcticdata.io/submit/): The NSF Office of Polar Programs (OPP) requires that metadata, full data sets, and derived data products be deposited in a long-lived and publicly accessible archive. Specific requirements for various OPP programs can be found in the NSF Dear Colleague Letter #16055 (https://www.nsf.gov/pubs/2016/nsf16055/nsf16055.jsp). The Arctic Data Center was created with funding from NSF to assist with compliance of these requirements by providing a long-lived and publicly accessible archive for Arctic Sciences Section (ARC) data and metadata. At a minimum, metadata describing ARC-supported data packages must be submitted

to the Arctic Data Center.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Level of Curation Performed. Select all relevant types from:

C. Enhanced curation – e.g. conversion to new formats; enhancement of documentation, D. Data-level curation – as in C above; but with additional editing of deposited data for accuracy

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Comments

We rely on the expertise of researchers with regard to the integrity of data observations. As such, all data and metadata are preserved as uploaded, and curated to produce new versions that enhance documentation and data organization to improve reusability. This includes conversion to open formats and congruency checks between the data and metadata.. Version control is supported for all uploaded metadata and data content. Datasets must meet certain requirements (sufficient metadata, open file formats) in order to be published with a DOI, and the curation process ensures that these requirements are met. Data files are published as-deposited, occasionally alongside an open-data format version of the data that is generated either by the curation team, or by the submitter upon request. Due to the heterogeneity of measurements, no effort is currently made to harmonize files across datasets except where initiated by a submitter.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Outsource Partners. If applicable, please list them.

National Center for Ecological Analysis and Synthesis (NCEAS) at the University of California, Santa Barbara (UCSB) - Administrative and organizational relationship - Operates the Arctic Data Center. All data are stored in servers on campus in UCSB's North Hall Data Center (NHDC) [0.1], which complies with a subset of the Tier 1 ANSI/TIA Data Center Standards.

Data Observation Network for Earth (DataONE) - Institutional relationship - infrastructure support; the NSF Arctic Data Center is a DataONE member node (MN) [0.2], which provides global, federated data discovery across distributed data centers.

US National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Information (NCEI) - Institutional relationship - offsite replication of data, metadata, software, documents, and provenance relationships for long-term preservation. NCEI is certified to operate the World Data Center (WDC) for Meteorology, the World Data Service (WDS) for Oceanography, and the World Data Service for Paleoclimatology. The WDS data centers are Core Trustworthy certified [0.3]

Amazon Web Services (AWS) - Contractual relationship - periodic, automated archival replication assures that data and metadata remain available even in the case of unplanned local system outages and provides for higher-performance access to data from multiple replica sites.

[0.1] https://www.ets.ucsb.edu/services/north-hall-data-center/service-description
[0.2] https://www.dataone.org/benefits-becoming-member-node
[0.3] https://www.ncdc.noaa.gov/customer-support/world-data-centers

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Other Relevant Information.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

ORGANIZATIONAL INFRASTRUCTURE

I. Mission/Scope

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Data preservation for the Arctic research community is critically important to the NSF Arctic Data Center. We recognize that data preservation is challenging for both for technical and non-technical reasons, and have developed an explicit data preservation plan describing how the NSF Arctic Data Center ensures the long-term preservation of the data entrusted to the repository. Key to this plan is our belief that no single organization can possibly provide sufficient institutional stability to guarantee multi-decadal preservation, and that partnerships among committed archives for replication are necessary for successful data longevity.

The primary mission [1.1] of the NSF Arctic Data Center is data preservation and data access, and high-quality data management is essential to data preservation. All submitted data and metadata are reviewed and edited before acceptance to ensure high-quality data products are available to the research community. Data are physically managed following best practice for systems administration at UCSB's NHDC, which complies with a subset of Tier 1 ANSI/TIA Data Center Standards.

Wherever possible, we utilize and encourage the use of open standards for representation of data and metadata, and for provisioning of services. Metadata are managed in the open Ecological Metadata Language (EML), and we encourage researchers to provide data using open data formats such as text-based CSV for tabular data and open formats for imagery. Open formats support accessibility of the data in the future even in the face of large software changes. In addition, the repository supports open programmatic access via the DataONE REST API [1.2] for automated access to the

holdings.

The mission and role of the NSF Arctic Data Center to preserve and provide access to data (as stated above) is mandated by its funder, the US National Science Foundation, and is stated in the award [1.3], “The archive will provide the capability to preserve and enable discovery of all products of NSF Arctic Science Section funded research, including data, metadata, software, documents, and provenance that link these in a coherent knowledge model, using infrastructure from the DataONE federation of data repositories.”

[1.1] https://arcticdata.io/preservation/
[1.2] https://arcticdata.io/catalog/#api
[1.3] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1546024&HistoricalAwards=false

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

II. Licenses

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

All data and metadata are released [2.1] under either the Creative Commons Public Domain Dedication (CC-0) [2.2] or the Creative Commons Attribution 4.0 International License (CC-BY) [2.3].

As a repository dedicated to helping researchers increase collaboration and the pace of science, this repository needs certain rights to copy, store, and redistribute data and metadata. By uploading data, metadata, and any other content to the NSF Arctic Data Center, users warrant that they own any rights to the content and are authorized to do so under copyright or any other right that might pertain to the content. Data and facts themselves are not covered under copyright law in the US and most countries [2.4], since facts in and of themselves are not eligible for copyright. That said, some associated metadata and some particular compilations of data could potentially be covered by copyright in some jurisdictions.

By uploading content, users grant the NSF Arctic Data Center repository and UCSB all rights needed to copy, store, redistribute, and share data, metadata, and any other content. By marking content as publicly available, users grant the NSF Arctic Data Center repository, UCSB, and any other 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 [2.2] or the Creative Commons Attribution 4.0 International License [2.3], depending on which license users choose at the time of upload. The choice of license is recorded in the metadata record for each submitted data set at the time of submission, and is preserved with the data through all subsequent preservation operations.

CC-0 means that there is no copyright; all data holdings are completely public and without conditions imposed on users. CC-BY means that users are free to share and adapt, so long as they “give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.”

“Appropriate credit” [2.5] means “if supplied, you must provide the names of the creator and attribution parties, a copyright notice, a license notice, a disclaimer notice, and a link to the material.

“Indicate if changes were made” [2.6] means that “you must indicate if you modified the material and retain an indication of previous modifications.”

“Technological measures” [2.7] are “defined with reference to Article 11 of the World Intellectual Property Organization [WIPO] Copyright Treaty.” [2.8]

Noncompliance with conditions of access and use is punishable under the US Copyright Act - 17. U.S.C. § 101. [2.9]. The NSF Arctic Data Center, distributes and makes data accessible under the terms of these licenses, and educates the community on the appropriate reuse and mechanisms for attribution, but does not become involved in enforcing the copyrights held by data submitters. It is incumbent upon submitters to enforce the copyrights that they own.

Data files which legally cannot be released under one of these two licenses (such as sensitive social science data

governed by an institutional review board) are not published on the Arctic Data Center. NSF Arctic Data Center staff are trained to recognize projects that may potentially have collected restricted data and will confirm with NSF program officers whether data should be archived from these projects as necessary. We rely on the expertise of the depositor to uphold ethical norms and to work with NSF program officers to safeguard sensitive information. As such, procedures are in place to limit the deposit of data with disclosure risks, in particular human subjects data.

[2.1] https://arcticdata.io/submit/#licensing-and-data-distribution
[2.2] https://creativecommons.org/publicdomain/zero/1.0/legalcode
[2.3] https://creativecommons.org/licenses/by/4.0/legalcode
[2.4] http://www.bitlaw.com/copyright/database.html
[2.5] https://wiki.creativecommons.org/wiki/License_Versions#Detailed_attribution_comparison_chart
[2.6] https://wiki.creativecommons.org/wiki/License_Versions#Modifications_and_adaptations_must_be_marked_as_such
[2.7] https://wiki.creativecommons.org/wiki/License_Versions#Application_of_effective_technological_measures_by_users_of_CC-licensed_works_prohibited
[2.8] https://www.wipo.int/pct/en/texts/articles/a11.html
[2.9] https://www.govinfo.gov/content/pkg/USCODE-2011-title17/pdf/USCODE-2011-title17.pdf

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

III. Continuity of access

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:

4 – The guideline has been fully implemented in the repository

Response:

The NSF Arctic Data Center is hosted at the University of California, Santa Barbara, at the National Center for Ecological Analysis and Synthesis (NCEAS) and funded under National Science Foundation (NSF) award number 1546024 [3.1], which ends in January 2021. NSF has stated their programmatic goal to continue to fund ongoing operations of the Arctic Data Center in 5 year award increments, with the operations governed by a Cooperative Agreement that can either be renewed or recompeted. NSF started funding this facility in 2007 through the CADIS program, renewed it from 2011-2016 through the ACADIS program, and then recompeted the program and awarded the current cooperative agreement to UCSB in 2016. As part of the NSF cooperative agreement, NSF conducts periodic reviews and site visits to determine compliance and program operating effectiveness, and to determine if a non-competitive renewal or a recompetition for operating the facility will occur. These reviews were successfully completed in 2017 and 2019, and in 2020 UCSB will be submitting a non-competitive renewal proposal to continue to operate the Arctic Data Center for the next 5 year block.

We recognize that over long time periods spanning many decades, it is extremely difficult to predict and sustain funding for single institutions. Our replication policy ensures high-availability during normal operations, but also provides security should NSF's investment in data archival wane. Should the main Arctic Data Center fail to be sustained, then the management of the Arctic Data Center has worked with our partnering institutions to ensure that the archival replicas that they hold continue to be preserved and available to the scientific community. This will mean that the National Centers for Environmental Information (NCEI) would become the authoritative holder of the data until a time when continued support from NSF can be obtained to re-establish operations. The Arctic Data Center has established its partnership with NCEI [3.2] specifically because they are funded independently of the National Science Foundation, and they are fundamentally focused on long-term archival quality preservation as their core mission, and are certified as a World Data System operator. By automating the replication of content from the Arctic Data Center to NCEI, we have provided a strong guarantee of continuity of access.

[3.1] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1546024
[3.2] https://arcticdata.io/about/

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

IV. Confidentiality/Ethics

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

NSF Arctic Data Center staff are trained to recognize projects that may potentially have collected restricted data and will confirm with NSF program officers whether data should be archived from these projects as necessary. We rely on the expertise of the depositor to uphold ethical norms and to work with NSF program officers to safeguard sensitive information. As such, procedures are in place to limit the deposit of data with disclosure risks, in particular human subjects data. Nevertheless, all data are stored with access control rules in place, and are transmitted over encrypted channels to maintain privacy. Licensing and data distribution considerations for all data are outlined on the NSF Arctic Data Center website [4.1], and NSF ultimately ensures compliance with all federal, university, and Institutional Review Board policies on the use of restricted data.

In cases where data cannot be published because it is restricted, submitters create a metadata record to document non-sensitive aspects of the project and data, including the title, contact information for the data set creators and contacts, and an abstract and methods description summarizing the data collection methodologies that does not include any sensitive information or data. [4.2]

[4.1] https://arcticdata.io/submit/#licensing-and-data-distribution
[4.2] https://arcticdata.io/submit/#who-must-submit-

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

V. Organizational infrastructure

R5. 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:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The NSF Arctic Data Center is hosted at the University of California, Santa Barbara, at the National Center for Ecological Analysis and Synthesis (NCEAS) and funded under NSF award number 1546024, which ends in January 2021. Basic operations, including technical operation, of the data center are conducted by the host institute (NCEAS). As part of the NSF cooperative agreement, NSF conducts periodic reviews and site visits to determine compliance and program operating effectiveness, and to determine if a non-competitive renewal or a recompetition for operating the facility will occur. These reviews were successfully completed in 2017 and 2019, and in 2020 UCSB will be submitting a non-competitive renewal proposal to continue to operate the Arctic Data Center for the next 5 year block.

This award funds software developers, data coordinators, student interns, a community engagement officer [5.1], training events for submitters [5.2], and travel for team members to attend conferences.

NSF Arctic Data Center staff work within the NCEAS and UCSB community, which provides opportunities for professional development via both in-house and external training events and seminars. The NSF Arctic Data Center draws on a wide range of staff expertise, and members participate in a variety of professional societies, including the Ecological Society of America (ESA), American Geophysical Union (AGU), Federation of Earth Science Information Partners (ESIP), in addition to participating in Arctic-specific communities such as IARPC [5.3] and the Arctic Data Committee [5.4]

[5.1] https://arcticdata.io/team/

[5.2] https://training.arcticdata.io

[5.3] https://www.iarpccollaborations.org/

[5.4] https://arcticdc.org/

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

VI. Expert guidance

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse or external, including scientific guidance, if relevant).

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The NSF Arctic Data Center is guided by a science advisory board [6.1] that collectively provides vast expertise in the Arctic, including researchers focused on terrestrial, aquatic, oceanographic, atmospheric, and social science in the Arctic. NSF Arctic Data Center team members meet with this board annually to garner feedback on the center.

The center also regularly solicits feedback from users and other members of the community via a variety of mechanisms including: email correspondence, conferences, and user trainings [6.2].

The NSF Arctic Data Center is also an active participant in both Arctic and cyberinfrastructure communities including IARPC, ARCUS, ESIP, and DataONE.

[6.1] https://arcticdata.io/team/

[6.2] https://training.arcticdata.io

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

DIGITAL OBJECT MANAGEMENT

VII. Data integrity and authenticity

R7. The repository guarantees the integrity and authenticity of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The NSF Arctic Data Center uses stable software technology that ensures data integrity and authenticity throughout the lifecycle of the data, including ingest, archival storage, and access. The NSF Arctic Data Center is powered by the Metacat data management system [7.1]. Metacat is a repository application that runs on Linux, Mac OS, and Windows platforms in conjunction with a database, such as PostgreSQL (or Oracle), and a Web server. The Metacat application

primarily stores metadata in an XML format and supports several internationally recognized metadata standards. Metacat implements the DataONE API [7.2] (https://purl.dataone.org/architecture) to upload and access content, and to replicate content over a secure connection to partner repositories so that all records are stored on multiple machines and data are preserved in the event of technical failures. Metadata are primarily stored using the Ecological Metadata Language [7.3] (https://eml.ecoinformatics.org), as well as other ubiquitous metadata standards. The data archive is available to users using the NSF Arctic Data Center's implementation of MetacatUI [7.4] (https://github.com/NCEAS/metacatui), a browser-based web application for DataONE repositories.

Each object inserted in the Metacat system has a unique, immutable identifier which is associated with the object's on-disk bytes in the Metacat repository. Upon dataset approval in the curation process, the most recent (approved) version of a metadata object is assigned a Digital Object Identifer (DOI), and is resolvable through the DOI resolution service. Data objects, resource maps, and intermediate (not approved, in curation process) versions of metadata objects are assigned a UUID identifier. All identifiers are resolvable through the DataONE resolution service. Once a document is inserted into the repository, the object itself cannot be changed or deleted. A digest checksum of the object is calculated upon ingest and is validated against the checksum stated in the DataONE system metadata that accompanies the object during upload. This checksum value is periodically used to determine the authenticity of the stored origin and replica objects using the DataONE fixity auditing service every ninety days. Subsequent uploads of an object to the repository are versioned through an update API call. The Metacat repository tracks versions of uploads by maintaining a linked obsolescence list of identifiers that can be used to trace the lineage of every object.

Authentication for inserting objects (including new versions of objects) is also controlled using the DataONE API. Client applications interacting with the repository provide credentials using a Javascript Web Token (JWT) or an X509 certificate, and all transactions are performed using Transport Layer Security (TLS)-based encryption. Permissions of "read", "write", and "changePermission" can be granted on any object in the Metacat system for individual or group subjects, allowing for private (non-public) storage. The NSF Arctic Data Center uses ORCID [7.5](https://orcid.org/) identifiers to uniquely identify researchers, and the DataONE Identity Service to maintain group subjects and permissions.

Metadata documents submitted to the repository are validated against the associated schema prior to insertion into the Metacat database to ensure documents are valid and complete. Data producers can check the completeness of their metadata by generating a metadata quality assessment using the metadig-engine quality suite [7.6] (https://github.com/NCEAS/metadig-engine) which is built into the NSF Arctic Data Center's web application. Data are ensured to be complete and uncorrupted using the fixity service described above, however it is not in the purview of the NSF Arctic Data Center to determine the scientific merit of any dataset.

Metacat keeps an internal log of events (such as insertions, updates, deletes, and reads) for auditing purposes, and administratively exposes the events through the DataONE API to produce aggregated metrics of objects and their replicas across the network.

The NSF Arctic Data Center describes provenance between data objects in the repository using the ProvONE provenance model. ProvONE is an extension of the W3C recommended standard PROV, aiming to capture the most relevant

information concerning scientific workflow computational processes, and providing extension points to accommodate the specificities of particular scientific workflow systems. Provenance is used to explicitly define the relationships between objects that are not of the same version chain, but which do have a computational relationship (such as a script that produces a data file). These provenance relationships are defined by data producers using the MetacatUI provenance editor or using programmatic libraries in Java, Python, R and Matlab.

The NSF Arctic Data Center manages datasets, collections of objects as known as “data packages”, which are aggregations of science metadata and data objects, or other nested collections. A data package is typically composed of one science metadata document describing at least one data object with the relationships between them documented in a resource map document. Resource maps are RDF documents that conform to the Open Archives Initiative’s Object Reuse and Exchange (OAI-ORE) specification. The MetacatUI web application allows users to search, link to, and download all datasets defined in the NSF Arctic Data Center. Obsoleted versions of datasets always resolve to the dataset version’s canonical URI, and a banner linking to the most recent version will appear at the top of the landing page. All data objects can also be queried and downloaded programmatically using the DataONE API.

[7.1] https://knb.ecoinformatics.org/knb/docs/
[7.2] https://purl.dataone.org/architecture
[7.3] https://eml.ecoinformatics.org
[7.4] https://github.com/NCEAS/metacatui
[7.5] https://orcid.org

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

VIII. Appraisal

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The NSF Arctic Data Center accepts data primarily from NSF funded projects for Arctic research. Occasionally, non NSF funded projects are accepted, although some portion of the data must have been collected in the Arctic to be included in the archive.

The center does not perform formal checks on the data with regard to its scientific merit or the accuracy of measurements. However, fixity checks are performed on each data object (see section VII), data files are converted to preferred file formats if possible, and data are occasionally reformatted for more efficient usage (converting from wide to long format, for example). Throughout this data curation process, all versions of data objects are preserved through our versioning system (see section VII) and relationships between them are described with explicit provenance, if applicable.

There are robust metadata standards that must be met for a submission to be published, beyond metadata being valid against the EML schema. These include passing all checks in the metadata quality report powered by an assessment framework called metadig [8.1] . The curation team assesses whether datasets contain: a descriptive title, comprehensive abstract, accurate temporal and geospatial coverages and comprehensive methods. Entity (file) level and attribute (variable) level metadata are also required. Required entity level metadata includes: file name, description, access URL, checksum, size, and file type. Required attribute information includes: name, description, units formally linked to base SI units for numeric data, definition of controlled vocabularies for categorical data, and missing value codes. Data congruency checks between data and metadata are performed programmatically by trained curation staff to ensure high-quality, accurate metadata.

Data or metadata that do not meet the requirements above are either returned to the submitter to make corrections, or corrected by the curation staff prior to issuing a DOI. Once a dataset is deemed complete through collaborative improvements by the submitter and the curation staff, a DOI will be issued.

The repository includes a list of preferred formats in the data submission guidelines on the website [8.2]. Data that are not in the preferred formats (not open access) are requested to be converted to a preferred format and resubmitted. The data curation staff ensures that data are in the preferred formats in the dataset review process. In some cases, the curation staff will do the conversion for the researchers (from Excel to CSV, for example), but researchers must approve the changes before the data are published.

[8.1] https://github.com/NCEAS/metadig
[8.2] https://arcticdata.io/submit/#file-format-guidelines

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

IX. Documented storage procedures

R9. The repository applies documented processes and procedures in managing archival storage of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Depositors use client applications (both graphical and scripted) that implement the DataONE Service Application Programming Interface (API), and the repository itself implements this interface to form a set of procedures with regard to upload, storage, and access of the data. The details of this architecture are documented [9.1]. We use the DataONE Data Model [9.2]as the Archive Information Package (AIP) stored on our servers. The model components include a documented specification for packaging a collection, specifications for science metadata that document science data, and a specification for system-level metadata required for interoperability across system nodes in the DataONE network. The DataONE API also includes an authentication architecture, and each depositor is required to log into the system using their credentials (often a username and password) that is provided to them by their home institution (such as universities that participate in the InCommon [9.3] network). Once they are authenticated, they may deposit data packages. Likewise, all updates to data and metadata require authentication, which preserves the integrity of the datasets. The repository uses the ORCiD researcher identifier service to allow researchers to authenticate with their chosen institution.

The repository preservation policy [9.4] outlines high-level principles for storing data and metadata, which includes a section on “preserving the bits”. Data are managed following best practice for systems administration at UCSB’s North Hall Data Center (NHDC), which complies with a subset of Tier 1 ANSI/TIA Data Center Standards. Only authorized staff are allowed to enter the facility, must sign the Access and Visitor Log, and must use their Access Control/Identification KeyCard to enter, and can only access their own hardware. The center uses video surveillance to record all activity.

For backups, hourly snapshots of the entire repository are stored at the data center and are transferred off site in Santa Barbara, CA. Likewise, we copy the entire database contents to an Amazon Web Services (AWS) cloud backup on a nightly basis. Individual data files are also replicated to participating Tier 4 replica target repositories in the DataONE network to further increase geographic redundancy and high availability.

In the event that a recovery is needed, the repository system is replaced according to internal system administration procedures to restore the latest hourly snapshot. These snapshots are containerized images managed with operating system-level virtualization, and are portable across three separate hardware installations in case of hardware failures. In the event of a regional emergency (such as a fire), a procedure is followed to migrate the entire system to an Amazon AWS compute instance that is operated out of a different region.

The repository team mitigates risk of data loss at multiple levels. At the hardware level, we incorporate redundancy into the systems such as multiple RAID controllers and using striping, mirroring, and other RAID technologies for high availability. We also mitigate risk at the application level by leveraging the DataONE Fixity Service which audits byte-for-byte content on a 90-day schedule of all files and replicas by comparing a known checksum of the file with a freshly calculated checksum. Any issues are reported via application-level logging. All aspects of server resources (disk, memory, network, OS, and applications) are monitored using an external suite of system and application checks that typically execute every minute or every five minutes. Our system administrators and operations team are notified via SMS messaging, email, and instant messaging channels in the event any component has degraded in status.

[9.1] https://purl.dataone.org/architecture
[9.2] https://purl.dataone.org/architecture/design/DataPackage.html
[9.3] https://incommon.org/
[9.4] https://arcticdata.io/preservation/

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

X. Preservation plan

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The NSF Arctic Data Center preservation plan is available on the website [10.1]. Stated in NSF award [10.2]: “Data will be replicated to administratively diverse institutions at the KNB, the Centers for Environmental Information (NCEI) and the Amazon cloud, as this is critical to long-term preservation.”

All metadata and data are replicated at geographically distinct locations including 1) DataONE replication nodes and 2) at the NOAA National Centers for Environmental Information. A full archival copy is also made nightly on the Amazon AWS cloud service. Replication is automated, and occurs any time that a change to any file in the system is made. Replication ensures that data and metadata remain available even in the case of unplanned local system outages (such as a regional-scale fire or earthquake event), and provides for higher-performance access to data from multiple replica sites.

By uploading content, users grant the NSF Arctic Data Center repository and UCSB all rights needed to copy, store, redistribute, and share data, metadata, and any other content [10.3]. By marking content as publicly available, users grant the NSF Arctic Data Center repository, UCSB, and any other 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 [10.4] or the Creative Commons Attribution 4.0 International License [10.5], depending on which license users choose at the time of upload. The choice of license is recorded in the metadata record for each submitted data set at the time of submission, and is preserved with the data through all subsequent preservation operations.

Finally, the rigorous appraisal process and file format requirements described in section VIII helps ensure that metadata and data can be usable and understandable in the long term (>20 years).

[10.1] https://arcticdata.io/preservation/

[10.2] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1546024&HistoricalAwards=false

[10.3] https://arcticdata.io/submit/#licensing-and-data-distribution

[10.4] http://creativecommons.org/publicdomain/zero/1.0/

[10.5] http://creativecommons.org/licenses/by/4.0/

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XI. Data quality

R11. 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:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The NSF Arctic Data Center facilitates high quality, long-lived datasets. Each dataset must have a complete and valid metadata record, which passes checks performed by both the metadig-engine [11.1] quality suite, and programmatic checks using functions written in R which ensure congruency between entity and attribute level metadata and data file contents. Researchers are encouraged to name measurement variables in accordance with the standard practices in their scientific discipline,and we default to the scientist’s expertise in this area.

Data providers can provide citations to additional datasets or publications which are related to the datasets.

Other data and metadata quality considerations, such as using preferred file formats and ensuring data integrity are described in detail in sections 7 and 8.

Dataset citation, view, and download metrics are visible on each dataset landing page.

[11.1] https://github.com/NCEAS/metadig-engine

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XII. Workflows

R12. Archiving takes place according to defined workflows from ingest to dissemination.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The NSF Arctic Data Center submission workflow is described in the “Publication” section of the submission guidelines [12.1].

In the typical submission process, researchers will submit data to the NSF Arctic Data Center through the NSF Arctic Data Center website [12.2] (see Submission Support [12.3]). Then, the Center's support team will review the data package. If the support team discovers any issues with the initial submission, the data team will work with the submitter to resolve the issues as quickly as possible (we primarily correspond via emails from support@arcticdata.io to the email address registered with the submitter's ORCID iD).

The NSF Arctic Data Center publicly releases data packages once complete data packages have been submitted and compiled. Depending on the complexity of the data package and the quality of the initial submission, the review process can take anywhere from a few hours to several weeks. Long processing times generally occur when initial submissions have incomplete metadata and/or poorly organized files and/or the submitter is not responsive to follow-up emails. After the review process, each data package is given a unique Digital Object Identifier (DOI) that will assist with attribution and discovery. The DOI is registered with DataCite using the EZID service, and will be discoverable through multiple data citation networks, including DataONE and others.

Once the data package is published with the NSF Arctic Data Center, it can still be edited and updated with new data or metadata. The original data and metadata will remain archived and available to anyone who might have cited it. Each data package DOI represents a unique, immutable version, just like for a journal article. Therefore, any update to a data package qualifies as a new version and therefore requires a new DOI. DOIs and URLs for previous versions of data packages remain active on the NSF Arctic Data Center (i.e., they will continue to resolve to the data set landing page for the specific version they are associated with), but a clear message will appear at the top of the page stating that "A newer version of this data package exists" with a hyperlink to the latest version. With this approach, any past uses of a DOI (such as in a publication) will remain functional and will reference the specific version of the data package that was cited, while pointing researchers to the newest version if one exists. All metadata records receive DOIs once they have been curated. Metadata records which are mid-curation process receive UUIDs. Datasets that are frequently updated are occasionally assigned a UUID with a DOI "series identifier." The Arctic Data Center also will sometimes generate a "parent" dataset to group together related datasets. This dataset is typically not issued a DOI, though the child metadata records are. The Arctic Data Center has recently begun moving away from this nesting practice and towards a data portal approach with the release of new software features.

[12.1] https://arcticdata.io/submit/
[12.2] https://arcticdata.io/catalog/#share
[12.3] https://arcticdata.io/submit/#submission-support

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:

Accept

XIII. Data discovery and identification

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Data and metadata in the NSF Arctic Data Center repository are searchable using the MetacatUI web application. All metadata are stored in the Ecological Metadata Language (EML), which is then stored using Metacat, and indexed using Solr. MetacatUI leverages this Solr index to allow for a user to generate queries of all repository metadata. Section VII contains more details on the interactions between these software systems. Indexed metadata fields [13.1] allow for a wide range of queries - including by entity/attribute information, taxonomic, geographic, and temporal coverage, dataset creator, identifier, etc.

The NSF Arctic Data Center datasets are also searchable via the DataONE network. The repository itself is also registered with other resource registries such as re3data. The NSF Arctic Data Center is also the official repository for Arctic research projects funded by NSF, and participates in a number of collaborations and events (see section V) to facilitate data sharing.

The NSF Arctic Data Center includes a recommended data citation format at the top of each dataset landing page, and issues persistent identifiers including DOIs, and tracks the citations of these datasets as part of the Making Data Count project [13.2]

[13.1] https://cn.dataone.org/cn/v2/query/solr
[13.2] http://mdc.lagotto.io/

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XIV. Data reuse

R14. 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:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The metadata standard used by the NSF Arctic Data Center is formatted using the Ecological Metadata Language. Although the NSF Arctic Data Center will accept data that is submitted in any format, where possible the data is converted to an open data format (e.g. Excel to CSV) to ensure long-term reuse of the data. As proprietary and vendor-specific file formats become obsolete over time, we require that data files be converted to open and ubiquitous data formats.These conversions are tracked using the explicit versioning and provenance systems implemented in the NSF Arctic Data Center. We strongly recommend the use of text-based formats, or when appropriate, open binary formats like NetCDF for large-volume data. The repository ensures understandability of the data using thorough metadata checks, implemented in both the Metadig quality engine in the web submission application, and via the trained data curation team. The curation team has a set of standards that each metadata record must reach prior to publication with a DOI, including but not limited to: a complete methods section, detailed abstract, descriptive title, and complete entity and attribute information including units. The teams metadata curation process is updated as new metadata formats and fields become relevant, and we regularly migrate from older metadata schemas to improved schemas during the operation of the repository.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

TECHNOLOGY

XV. Technical infrastructure

R15. 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:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The NSF Arctic Data Center's repository software infrastructure is based on top of the Linux operating system, and uses a client-server model according to international standards. The repository software is a stable, open source platform developed and maintained by NCEAS software engineers over the last two decades. As described in Section 7, the repository uses Metacat [15.1], the DataONE API [15.2], and MetacatUI [15.3]. The repository stores science metadata in the W3C XML [15.4] format, encoded in either ISO 19139 documents or the Ecological Metadata Language (EML) [15.5] documents. Packaging metadata are also stored in XML documents encoded in the Resource Description Framework (RDF) [15.6] syntax that conform to the Open Archives Initiative's Object Reuse and Exchange (OAI-ORE) [15.7]

specification. The DataONE Data Model builds on this specification, and requires that objects registered across the DataONE system uses HTTP IRIs that are resolvable using the DataONE Resolution Service. Provenance metadata are also encoded into these RDF documents, according to the W3C PROV [15.8], Ontology, an international standard expressed in the Web Ontology Language (OWL) [15.9]. The provenance assertions are further refined with concepts derived from the ProvONE [15.10] provenance ontology, which is an extension to the W3C PROV model designed specifically for scientific workflows.

The software infrastructure that incorporate these technologies are are all open source, community-maintained projects that have individual plans for development, largely driven by and described in foundation-based proposals. A technical description of how connectivity is maintained to the repository is described in Section 3.

Primary systems are maintained at the North Hall Data Center, which complies with a subset of the Tier 1 ANSI/TIA Data Center Standards. Networking at 10GbE is via redundant connections to the public Internet and Internet2 through the CalREN2 and CENIC networks. Room UPS power backed by an emergency generator is available up to the 162kW capacity of the data center. Primary cooling capacity is derived from the campus chilled water loop. With the campus chilled water loop subject to regional power outages, secondary emergency cooling is from two locally installed chillers with a total 60 Tons of capacity. When NHDC is on emergency power, the emergency chilled water is used for the UPS room, AHU 5 (campus networking) and chilled water distribution to advanced rack cooling technologies. All racks are mounted on zone 4 ISO-Base platforms for seismic protection. The NHDC is subject to the environmental conditions of the campus and the region. Planned outages involving all equipment within NHDC will be uncommon, but occasionally necessary for certain types of maintenance activity. During such outages, data and metadata from the NSF Arctic Data Center will still be available via our replica holdings, but data submissions will be delayed until normal operations are restored.

[15.1] http://github.com/NCEAS/metacat
[15.2] https://releases.dataone.org/online/api-documentation-v2.0/index.html
[15.3] https://github.com/NCEAS/metacatui
[15.4] https://www.w3.org/XML/
[15.5] https://eml.ecoinformatics.org
[15.6] https://www.w3.org/RDF/
[15.7] https://openarchives.org/ore/
[15.8] https://www.w3.org/TR/prov-overview/
[15.9] https://www.w3.org/OWL/
[15.10] https://purl.dataone.org/provone-v1-dev

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:

Accept

XVI. Security

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

In addition to the physical security measures described in Section 9 (Documented Storage Procedures), we employ cyber-security procedures and processes to maintain system integrity. At the operating system level, we regularly patch software with security updates to reduce or eliminate possible attack vectors. We fast-track fixes to all security vulnerabilities that compromise administrative access to system, and closely monitor the Common Vulnerabilities and Exposures (CVE) [16.1] disclosures. This work is handled largely by system administration staff, but also includes development staff responsible for application security. To enhance our code security, we periodically request security audits of components of our software stack from colleagues at the Cybersecurity and Networking Division at the National Center for Supercomputing Applications (NCSA) [16.2] at the University of Illinois Urbana-Champaign. We mitigate vulnerabilities in the software that are discovered in order to harden our services to cyber attacks.

Similarly, we rely on UCSB's Office of the Chief Information Officer (CIO) and the core Enterprise Technology Services group to maintain a secure and resilient network infrastructure. Within that group, the Chief Information Security Officer (CISO) directs activities that implement the University of California system-wide Information Security Policies [16.3] at UCSB. In particular, the campus Security Operations Center (SOC) Manager operates, maintains, and monitors the campus network, including the data center network. This monitoring includes TCP-based service monitoring, intrusion detection, active scanning of hosts for known vulnerabilities, unusual traffic spikes and patterns that suggest compromises, and other techniques to keep the network secure.

In the event of a compromise, our repository staff execute a procedure of isolating the system, reporting the issue, and

restoring the system to a known secure snapshot in the same manner that we would for hardware or operating system failures. We then analyze the cause of the breach, and take measures to ensure it does not recur.

As part of our preservation plan and the procedures used to securely maintain the systems, we recognize that over long time periods spanning many decades, it is extremely difficult to predict and sustain funding for single institutions. Our replication policy ensures high-availability during normal operations, but also provides security should NSF's investment in data archival wane. Should the main NSF Arctic Data Center fail to be sustained, then the management of the NSF Arctic Data Center will work with our partnering institutions to ensure that the archival replicas that they hold continue to be preserved and available to the scientific community. This will likely mean that the National Centers for Environmental Information would become the authoritative holder of the data until a time when continued support from NSF can be obtained to re-establish operations.

[16.1] https://cve.mitre.org/
[16.2] https://ncsa.illinois.edu
[16.3] https://security.ucop.edu/policies/it-policies.html

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

APPLICANT FEEDBACK

Comments/feedback

These requirements are not seen as final, and we value your input to improve the core certification procedure. To this end, please leave any comments you wish to make on both the quality of the Catalogue and its relevance to your organization, as well as any other related thoughts.

Response:

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments: