NISTIR 8341

A Self-Audit of the NIST Public Data Repository Using the CoreTrustSeal Trustworthy Data Repositories Requirements

Andrea Medina-Smith

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8341

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 8341

A Self-Audit of the NIST Public Data Repository Using the CoreTrustSeal Trustworthy Data Repositories Requirements

Andrea Medina-Smith
Information Services Office

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8341

April 2021

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
James K. Olthoff, Performing the Non-Exclusive Functions and Duties of the Under Secretary of Commerce for Standards and Technology & Director, National Institute of Standards and Technology

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

National Institute of Standards and Technology Interagency or Internal Report 8341
Natl. Inst. Stand. Technol. Interag. Intern. Rep. 8341, 37 pages (April 2021)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8341

Abstract

In February 2020, Office of Data and Informatics Group Leader Gretchen Greene asked Metadata Librarian Andrea Medina-Smith to perform a CoreTrustSeal (CTS) self-audit of the Public Data Repository (PDR)[1] in preparation for submitting an application for CoreTrustSeal Certification. This certification results from a lightweight but thorough assessment of a given data repository's policies, documentation, and technical infrastructure supporting 16 required elements. The requirements are broadly split into the following themes: background information, organizational infrastructure, digital object management, and technology. These elements together form a picture of the trustworthiness of the data repository. The following self-audit report is a "snapshot" of where the PDR stands as of August 2020.

Key words

CoreTrustSeal; data; data repository; digital repository; research data; trustworthy digital repository.

[1] The NIST Public Data Repository is the back-end processes and storage behind the NIST Science Data Portal. https://data.nist.gov/

Table of Contents

List of Tables

Introduction

In February 2020, Office of Data and Informatics Group Leader Gretchen Greene asked Metadata Librarian Andrea Medina-Smith to perform a CoreTrustSeal (CTS) self-audit of the Public Data Repository (PDR)[2] in preparation for submitting an application for CoreTrustSeal Certification. This certification results from a lightweight but thorough assessment of a given data repository's policies, documentation, and technical infrastructure supporting 16 required elements. The requirements are broadly split into the following themes: background information, organizational infrastructure, digital object management, and technology. These elements together form a picture of the trustworthiness of the data repository. The following self-audit report is a "snapshot" of where the PDR stands as of August 2020.

A successful CTS Certification application depends on fulfilling all 16 requirements with a rating of 4 (fully implemented) or high 3 (implementation phase). Ratings for each requirement are based on assessment specifications spelled out in the guidance provided by CTS in the CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022.[3] After assessment, the requirement is assigned a compliance level from 0 to 4. Recommendations for improvement of the requirement are also provided.

0 – Not applicable
1 – The repository has not considered this yet
2 – The repository has a theoretical concept
3 – The repository is in the implementation phase
4 – The guideline has been fully implemented in the repository.

The Public Data Repository (referred to as either the PDR or "the Repository" henceforth) is a public-facing repository built by the National Institute of Standards and Technology (NIST) to facilitate discovery, access, and preservation of research data created by NIST researchers. The team responsible for management and development of the PDR includes members from the NIST Office of Data and Informatics (ODI), the Office of Information Systems Management (OISM), the Special Programs Office (SPO), and the Information Services Office (the NIST Research Library). ODI develops and manages the preservation, landing page, and dissemination services used by the PDR; OISM manages the storage, integrity, and authentication services; SPO is the system owner and creates the policies that guide the PDR's use and development; and ISO is involved in curation duties for the PDR and its ingest service known as MIDAS. The self-audit was conducted through interviews of these responsible team members and a review of available documentation between March and May 2020.

The results from the self-audit were encouraging. Out of 16 requirements, the PDR had compliance levels of 4 for 11 of the requirements; the remaining requirements 5 were rated at a level 3. The majority of the work to be done to raise those compliance levels, and to improve even in the case of a fully implemented score of 4, is to produce robust documentation for use by current and future staffers of the Repository.

[2] The NIST Public Data Repository is the back-end processes and storage behind the NIST Science Data Portal. https://data.nist.gov/.
[3] https://www.coretrustseal.org/why-certification/requirements/

Repository team members have worked hard for the past five years. The collaboration has proven to be fruitful in a number of tangible ways. Since its release in 2017 more than 500 datasets have been published, and a major update was released in August 2020 that allows for direct editing of the metadata displayed on the dataset landing pages, thus giving researchers another tool to improve the FAIRness of their data.[4]

Fulfillment of the recommendations in this report will put the PDR on a good footing to apply for CTS Certification before the next large development cycle begins.

Table 1. Summary Table of CTS Certification Ratings and Recommendations

Requirement	Rating	Recommendations
R1. Mission	4	• A review process for updating the mission both in terms of policies and language should be considered after five or more years. • It would also be useful to have the mission reaffirmed by NIST leadership as it evolves.
R2. Licenses	4	• Licenses should be reviewed by NIST Office of Chief Counsel periodically to maintain accuracy and alignment with current policies.
R3. Continuity of Access	3	• A plan should be developed for relocation or transition to another organization or return of the data to data creators in the unexpected event of a loss of funding or a shift in organizational priorities.
R4. Confidentiality/Ethics	3(/2?)	• The Repository and the related MIDAS system should be clearer about what data can and cannot be stored by the system. This could be achieved by explicitly calling out in the FAQ and About pages that data with PII and BII are prohibited. • Further, the NIST Directive on risk management in data management plans (DMPs) will be a welcome addition to the guidance on this subject. • The ability for the Repository to mediate access to specific datasets, under development at the time of this writing, will allow for inclusion of more sensitive data.
R5. Organizational Infrastructure	4	• An assessment of the team's skills and identifying any skill gaps will allow for succession planning and future additions to the team.
R6. Expert Guidance	3	• Create an outside user or partner group to solicit opinions and feedback. These users are the researchers who reuse the data, industry partners, academics, and the public. These stakeholders will add a needed point of view because currently only "in-house" users (NIST researchers and management who interact with the system from the submission side) are part of the feedback loop.
R7. Data integrity and authenticity	4	• The Repository should improve the amount of documentation for all integrity and authenticity workflows and checks.

[4] FAIR Data Principles, https://www.go-fair.org/fair-principles/

Requirement	Rating	Recommendations
R8. Appraisal	4	• The Repository team should work with OU leadership to spell out collection development policies at the OU level. This guidance would complement guidance at the NIST level, including the extant directives and the data taxonomy.
R9. Documented Storage Procedures	3	• While the work of the team to implement archival workflows is nearly complete, there is little documentation. • A comprehensive, end-to-end documentation file would move this rating from a 3 to a 4.
R10. Preservation Plan	4	• The Repository should complete documentation of the Preservation Service and associated workflows. • An explicit notification of the transfer of custody and responsibility should be included in a "click-through" notification when datasets are submitted for review.
R11. Data Quality	4	• The Repository should assess if and how the Project Open Data (POD) field dataQuality could be implemented.
R12. Workflows	3	• Technical documentation should be completed for each step in this workflow.
R13. Data Discovery and Identification	4	• The Repository should move to the open OAI-PMH standard for exposing data.
R14. Data Reuse	4	• Datafiles should automatically download with a copy of the metadata record OR any creator supplied documentation. • A method to require metadata at the distribution level for all datafiles (even those submitted in bulk) should be explored.
R15. Technical Infrastructure	4	• Documentation for all technical infrastructure should be completed. This could be structured around the repository functions as spelled out in the OAIS reference model.
R16. Security	4	• Regularly review and update the security plan currently in place to ensure standard practices are followed over time. • Develop user-level auditing to detect malicious or erroneous use of the system before an embarrassing error occurs.

Background Information

Context

R0. Please provide context for your repository

- Repository Type.
    - Institutional repository
    - Research project repository
- Brief Description of Repository
    - "This National Institute of Standards and Technology (NIST) **Science Data Portal** provides a user-friendly discovery and exploration tool for publicly available datasets at NIST. These data products are generated as part of the NIST mission, spanning multiple disciplines of scientific, engineering and technology research. NIST's publicly available data sets showcase its commitment to providing accurate, well-curated measurements of physical properties, exemplified by the Standard Reference Data program, as well as its commitment to advancing basic research."[5]
- Brief Description of Designated Community
    - The PDR's designated community is NIST researchers and associates and members of the public who use the published datasets.
- Level of Curation Preformed.
    - Basic curation – e.g., brief checking, addition of basic metadata or documentation

[5] From the NIST Scientific Data Portal About NIST Data Page: https://data.nist.gov/sdp/#/about

Organizational Infrastructure

1. Mission/Scope

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

Guidance:

Repositories take responsibility for stewardship of digital objects, and for ensuring that materials are held in the appropriate environment for appropriate periods of time. Depositors and users must be clear that preservation of and continued access to the data is an explicit role of the repository.

Response: Use of the NIST Public Data Repository (PDR) is an option for NIST researchers who would like to make their data public in a NIST-authorized repository. The mission of the repository is laid out on the Science Data Portal (SDP) "About" page and on the SDP Policy[6] page:

> *Data products developed and distributed by the National Institute of Standards and Technology span multiple disciplines of research and are widely used in research and development programs by industry and academia.*
>
> *NIST's publicly available data sets showcase our commitment to providing accurate, well-curated measurements of physical properties, exemplified by the Standard Reference Data program, as well as its commitment to advancing basic research.*
>
> *In accordance with U.S. Government Open Data Policy and the NIST Policy for Managing Public Access to the Results of Federally Funded Research, NIST maintains a publicly accessible listing of available data, the NIST Public Dataset List (json). Additionally, these data are assigned a Digital Object Identifier (DOI) to increase the public's ability to discover and access our research output; these DOIs are registered with DataCite and provide globally unique persistent identifiers.* ***The NIST Science Data Portal provides a user-friendly discovery and exploration tool for publicly available datasets at NIST****. – About Page*
>
> *The NIST Public Data Listing and associated data are stored in NIST's public data repository. – Policy Page*

In addition, *The NIST Plan for Providing Public Access to Results of Federally Funded Research,* signed by the NIST Director, makes the Associate Director for Laboratory Programs responsible for executing the plan and ensuring its effectiveness.[7] This plan

[6] The Science Data Portal is the public facing side of the PDR. The SDP allows for searching the PDR, viewing landing pages and downloading data stored in the PDR. https://data.nist.gov/

[7] https://doi.org/10.6028/NIST.IR.8084

developed a taxonomy of data types and outlines what data should be reviewed, preserved, and discoverable based on the data type.

Recommendations for Repository to meet/improve compliance with this Requirement:

- A review process for updating the mission both in terms of policies and language should be considered after five or more years.
- As the mission and product evolves NIST leadership should reaffirm their support.

Rating: 4 (The guideline has been fully implemented in the repository)

2. Licenses

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

Guidance:

Repositories must have an appropriate rights model covering data access and use, communicate about them with users, and monitor compliance. This Requirement relates to the access regulations and applicable licenses set by the data repository itself, as well as any codes of conduct that are generally accepted in the relevant sector for the exchange and proper use of knowledge and information. Evidence should demonstrate that the repository has sufficient controls in place according to the access criteria of their data holdings, as well as evidence that any relevant licenses or processes are well managed.

Response: All datasets and software stored in the PDR are subject to the NIST copyright, fair use, and licensing information, which is provided through a link to the Fair Use Statement on every repository record page.[8] With the exception of Standard Reference Data, NIST data are not subject to copyright within the United States.

The Standard Reference Data Act of 1968, updated in 2017, allows Standard Reference Data (SRD) to be copyrighted by the U.S. Secretary of Commerce on behalf of the United States of America.[9]

The repository does not provide access to sensitive or controlled data and makes that clear on the Public Access to NIST Research page: *"To the extent feasible and consistent with law, agency mission, resource constraints, and U.S. national, homeland, and economic security, NIST will promote the deposit of scientific data arising from unclassified research and programs, funded wholly or in part by NIST, except for Standard Reference Data, free of charge in publicly accessible databases."*[10]

Recommendations for Repository to meet/improve compliance with this Requirement:

- Licenses should be reviewed by the NIST Office of Chief Counsel periodically to maintain accuracy and alignment with current policies.

Rating: 4 (The guideline has been fully implemented in the repository)

[8] https://www.nist.gov/topics/data/public-access-nist-research/copyright-fair-use-and-licensing-statements-srd-data-and

[9] American Innovation and Competitiveness Act, Pub. L. No. 114-329, § 208 (2017). https://www.congress.gov/bill/114th-congress/senate-bill/3084/text

[10] https://www.nist.gov/open

3. Continuity of access

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

Guidance:
This Requirement covers the governance related to continued operation of the repository over time and during disasters, as well as evidence in relation to succession planning; namely, the measures in place to ensure access to and availability of data holdings, both currently and in the future. Reviewers are seeking evidence that preparations are in place to address the risks inherent in changing circumstances, including in mission and/or scope.

Response: Plans for continuation of service and operation of the repository is a team effort between the Office of Data and Informatics (ODI), the Office of Information Systems Management's (OISM's) Application Systems Division and Research Services Office, and the Information Services Office (the NIST Research Library). Members of this group are named in the Open Access to Research (OAR) Charter. The OAR Charter, which guides development of the repository and ancillary services, explicitly lays out the roles and responsibilities of all project participants and stakeholders over time and in case of disasters. Full funding of both the OISM team, which supports the repository infrastructure, and ODI, as the repository developers, is in place. At this point there is no plan for relocation or transition to another body nor a plan for returning data to their creators in the unexpected event of a loss of funding or a shift in organizational priorities.

Further risk mitigation and redundancy plans are incorporated into the IT system security plan for 600-01 Cross-Laboratory Programs Applications System. Security is assessed annually by NIST's Office of Information Systems Management. This document is covered specifically in Requirements 7 and 16.

Recommendations for Repository to meet/improve compliance with this Requirement:

- A plan should be developed for relocation or transition to another organization or return of the data to data creators in the unexpected event of a loss of funding or a shift in organizational priorities.

Rating: 3 (The Repository is in the implementation phase)

4. 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.

Guidance:

Adherence to ethical norms is critical to responsible science. Disclosure risk—for example, the risk that an individual who participated in a survey can be identified or that the precise location of an endangered species can be pinpointed—is a concern that many repositories must address. Evidence should demonstrate that the repository has good practices for data with disclosure risks, including guidance for depositors and users. This is necessary to maintain the trust of those agreeing to have personal/sensitive data stored in the repository.

Response: NIST operates primarily in unclassified, non-proprietary, non-human-subject research fields. Currently, data that includes personally-identifiable information (PII) or business-identifiable information (BII) is not allowed in the repository, though it is not immediately obvious that this is the case when researchers use the MIDAS system to describe and upload data. The PDR team is planning to move (and is in the process of doing so at the time of this writing) beyond a fully open repository to provide gated public access for specific datasets. This type of gateway would allow for registration of users and mediated download of certain datasets.

Data undergoes a mandatory review before it is made public. NIST Suborder 1801.02 *Review of Data Intended for Publication*, requires that data being published be accompanied by information that will facilitate understanding and re-use.[11] Reviewers look for the following before approving the public availability of a dataset:

- Any PII/BII
- No proprietary information is included
- A commercial product disclaimer is included (if necessary).
- Data is provided in a Section 508-compliant format
- Instructions for use and appropriate explanations are provided (e.g., in a readme file)
- Experimental method/methodology is described sufficiently to allow replication
- Experimental method/methodology is appropriate for the study and conclusions that are drawn.

NIST representatives serve on the National Science and Technology Council (NSTC) Subcommittee on Open Science (SOS) working group that is developing recommendations for risk management. A NIST directive that is based on discussions at an NSTC SOS workshop in July 2019 provides guidance for inclusion of risk management in Data Management Plans (DMPs); this document is currently in internal review.

[11] See Appendix A: Directive S 1801.02 *Review of Data Intended for Publication*

Recommendations for Repository to meet/improve compliance with this Requirement:

- The Repository and the related MIDAS system should be clearer about what data can and cannot be stored by the system. This could be achieved by explicitly calling out the FAQ and About pages that data containing PII and BII are prohibited.
- Further, the NIST Directive on risk management in DMPs will be a welcome addition to the guidance on this subject.
- The ability for the Repository to mediate access to specific datasets will allow for inclusion of more sensitive data.

Rating: 3 (The Repository is in the implementation phase) / 2 (The Repository has a theoretical concept – inclusion of sensitive data)

5. 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.

Guidance:

Repositories need funding to carry out their responsibilities, along with a competent staff who have expertise in data archiving. However, it is also understood that continuity of funding is seldom guaranteed, and this must be balanced with the need for stability.

Response: The Repository staff work for the Office of Data and Informatics (ODI) in the Material Measurement Laboratory (MML); in the Office of Information Systems Management (OISM), Division of Application Development Services; and in the Information Services Office (the NIST Research Library). Staff in these offices are recognized experts in technology infrastructure, data management, and preservation.

Costs for storage in and retrieval of data from the PDR are fully supported at the NIST level. ODI has guaranteed funding for maintaining and managing the operational repository, and the division has a fully funded budget. The IT resources brought to bear by OISM are also fully funded. These budgets include funding for conference attendance, training, and professional development as needed for staff.

The experience of the team working on the Repository is both deep and wide; this is because the team is made up of members from across NIST and not just from one organizational unit. ODI staffers come with decades of experience in managing and maintaining community repositories. Contract staff and those from OISM are highly skilled programmers working in various languages. The librarians who consult on the project have worked in the archival and data management side of information science for upwards of 20 years.

In addition to maintaining a credible relationship with researchers within the designated community, Repository staff are members of many international and national bodies promoting research data management such as the Research Data Alliance and the Research Data Framework initiative. Membership in these groups has led to not only staff sharing their expertise, but also the ability to network with many others with valuable experiences.

Recommendations for Repository to meet/improve compliance with this Requirement:

- An assessment of the team’s skills and identifying any skill gaps will allow for succession planning and future additions to the team.

Rating: 4 (The guideline has been fully implemented in the repository)

6. Expert guidance

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

Guidance:

An effective repository strives to accommodate evolutions in data types, data volumes, and data rates, as well as to adopt the most effective new technologies in order to remain valuable to its Designated Community. Given the rapid pace of change, it is therefore advisable for a repository to secure the advice and feedback of expert users on a regular basis to ensure its continued relevance and improvement.

Response: Since the inception of the Open Access to Research (OAR) project in early 2015 the OAR Organizational Unit (OU) Advisory Group has been the source of advice, feedback, and commentary for the various parts of the OAR Project including the PDR. The OAR OU Advisory Group is made up of representatives from each of the six NIST Laboratory OUs under the chairmanship of the NIST Open Access Officer. This group not only provides feedback on various workflows and requirements but is also one way the concerns of the NIST researcher are heard; other ways include feedback directly in the MIDAS system, a "Contact Us" form in the SDP Help page, email, and via the NIST Service Now platform. Because NIST is a federal bureau within the U.S. Department of Commerce, NIST's activities are also guided by the NSTC SOS and the Department of Commerce's Commerce Data Governance Board.

Outreach to the NIST research community occurs through training events (both in person and recorded) and an internal website. The need for improvement in this area is recognized by Repository staff.

Recommendations for Repository to meet/improve compliance with this Requirement:

- Create an outside user or partner group to solicit opinions and feedback. These users are the researchers who reuse the data, industry partners, academics, and the public. These stakeholders will add a needed point of view because currently only "in-house" users (NIST researchers and management who interact with the system from the submission side) are part of the feedback loop.

Rating: 3 (The Repository is in the implementation phase)

Digital Object Management

7. Data integrity and authenticity

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

Guidance:

The repository should provide evidence to show that it operates a data and metadata management system suitable for ensuring integrity and authenticity during the processes of ingest, archival storage, and data access. This Requirement covers the entire data lifecycle within the repository.

To protect the integrity of data and metadata, any intentional changes to data and metadata should be documented, including the rationale and originator of the change. Measures should be in place to ensure that unintentional or unauthorized changes can be detected, and correct versions of data and metadata recovered.

Authenticity covers the degree of reliability of the original deposited data and its provenance, including the relationship between the original data and that disseminated, and whether or not existing relationships between datasets and/or metadata are maintained.

Response: The integrity of the datasets maintained by the Repository is ensured through the ingest and preservation workflows. This repository system is highly influenced by the Open Archival Information Systems (OAIS) repository model and creates Archival Information Packages (AIPs) for storage in the preservation system, and Distribution Information Packages (DIPs) served to the public via the Landing Page Service. The primary Submission Information Package (SIP) currently supported by the repository takes the form of an initial Project Open Data (POD) record and author-provided data files that are submitted via the Repository's publishing service (PS).

When a MIDAS record is published it is sent as JSON to the PS along with MIDAS-produced checksums. This JSON record includes a list of files and their directories. The PS takes the initial metadata, transforms it into the NIST Extensible Resource Data mode (NERDm) schema, and collocates it with copies of the dataset files. The PS also calculates a checksum and matches it against the MIDAS checksum. These checksums are put into the file metadata. An AIP is created via a preservation service by storing all the metadata and files in a bag based on the BagIt standard.

Before that bag is placed in long-term storage an integrity check is performed to make sure the bag is compliant with the BagIt specifications and to rerun the checksums. The metadata is validated against the POD schema and the NERDm. All files that are listed in the POD record are checked to verify that they are included in this bag or in a previously published bag. The bag is then split into multiple zip files, per the multi-bag BagIt profile. At that point a hash is calculated for each zip file and stored separately. The zipped files and hashes are delivered to a directory owned by OISM and replicated on Amazon Web Services (AWS); these are the preservation "buckets." The hashes are delivered back to MIDAS, which periodically checks the files in the public bucket to ensure the checksums still match.

When changes are made to a MIDAS record, they are captured in a new bag (a "change" bag or "errata" bag). This bag contains the new or changed data, all of the metadata, and a log of changes made to metadata as they are detected in interactions with MIDAS. More logging could be done to track provenance. The developers commented that they would like to implement PREMIS to track preservation events.

Once the bags are deposited in buckets in the AWS Data Management and Integrity Service, any changes are managed by the OISM team. There the system tracks "S3 Events" (uploads, changes, deletions, etc.) and sends that information to the OISM team who can then make sure it is a legitimate request (based on where the request originated). It is an automated system, so manual changes are automatically suspect.

The over-arching strategy for handling data changes is not made explicit to depositors, though a chart is available that lets them know their responsibilities and review requirements for several types of changes.[12] Versioning has a standard practice, but again this is not documented for users at this time. When there is a change, each dataset is changed using a three-field version number. If only metadata is changed the rightmost number is changed. If data is changed then the middle number is incremented. The leftmost field is only incremented upon request of the author. Users can see this versioning on the landing page both before publication (preview) and after.

Provenance information comes in the form of metadata from MIDAS (the POD record) and BagIt bags. Audit trails are logs from the long-term storage service. Further upstream, in future versions of the code will know when users are logged in and what changes they make in the record. This will improve provenance metadata collection.

Links to metadata and other datasets are maintained through both the POD and NERDm schemas. POD has fields for related identifiers that can link together datasets, and NERDm has linking fields that can describe parent/child relationships. Further, linking between papers and datasets is done in both the POD record and the metadata sent to DataCite for DOI registration.

Different versions of the same file are determined by the unique ModifiedDate field. When an update is submitted, the system determines whether it constitutes only a metadata change or a data file change (which also implies a metadata change). At the moment, the latter is assumed if data files are included in the SIP. At this time the system does not determine whether those data files have, in fact, changed (i.e., whether they result in a different checksum), nor is it determined what metadata has changed.

The Repository is a closed system, and the authenticity of data is not explicitly checked but is based on trust of the submitters in the closed NIST system. Before ingesting the data, sets are reviewed by Division Chiefs for adherence to NIST policy and can go through technical review as well. The system does not currently check the identities of depositors beyond the need of a NIST account to access the frontend system.

[12] See Appendix B: Directive PR 1801.01 Revising and Removing Information from NIST Inventories and Repositories

More information on integrity and availability can be found in the System Security Plan for 600-01 Cross-Laboratory Programs Applications System which is described in Requirement 16.

Recommendations for Repository to meet/improve compliance with this Requirement:

- The Repository should improve the detail of documentation for all integrity and authenticity workflows and checks.
- Improving the tracking of changes, including logging who is making changes, should be added to the Repository's list of projects.

Rating: 4 (The guideline has been fully implemented in the repository)

8. Appraisal

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

Guidance:

The appraisal function is critical to evaluate whether data meet all criteria for selection and to ensure appropriate management for their preservation. Appraisal and reappraisal over time ensure data remain relevant and understandable to the Designated Community.

Response: Appraisal is generally in the hands of the Organizational Unit (OU) leadership, and that has translated into a series of Directives that act as the collection development policy. The policy encourages that any data that is "publishable" or that underpins results reported in a published paper be deposited; metadata from Standard Reference Datasets are required to be deposited.

A record cannot pass from the edit to the review state (and therefore cannot be published) if required metadata is not provided. The quality of the metadata is checked in the review process. Subsequently, if the librarian responsible for the curation step feels that metadata is missing, they will suggest, but cannot require, changes to the record for the author to consider prior to publishing.

Because of the broad scope of the designated community there are not preferred file formats, but staff are encouraged to deposit data in machine-readable, machine-actionable, non-proprietary formats. Data management training and guidance are provided by various NIST offices. At this time all file formats are accepted in the repository.

When a deletion is requested it must be for one of the reasons outlined in NIST Directive PR 1801.01.[13] These include data that are no longer supported, erroneous data, or data that may no longer be public but available on request. In each case a "tombstone" page is created that describes why the data are no longer available.

Recommendations for Repository to meet/improve compliance with this Requirement:

- The Repository team should work with OU leadership to spell out collection development policies at the OU level. This guidance would complement any at the NIST-wide level including the extant directives and the data taxonomy.

Rating: 4 (The guideline has been fully implemented in the repository)

[13] See Appendix B: Directive PR 1801.01 Revising and Removing Information from NIST Inventories and Repositories

9. Documented storage procedures

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

Guidance:

Repositories need to store data and metadata from the point of deposit, through the ingest process, to the point of access. Repositories that perform digital preservation must offer "archival storage" in OAIS terms.

Response: The processes to document procedures for the Repository are in early stages. Several Repository procedures are only documented via the code that executes them (the Submission Information Package (SIP) to Archival Information Package (AIP) process, for example). Others like the BagIt specifications and profiles are extensively documented as they follow a common standard.

The archival storage function is not explicitly documented but occurs as defined by the OAIS model.[14] The AIP (BagIt bags) are accepted by the AWS Storage and replicated in three storage areas: a Gold copy, a Public copy, and a long-term archival storage copy. Each storage area includes both the current AIP and all previous versions of the AIP.

Whenever an AIP is ingested, a hash is generated and checked at each step. Then the Gold copy's hash value is checked daily versus a hash repository stored on an AWS server. The Gold and Public copy hashes are checked against each other daily. If a Gold hash is bad, then it is pulled from a previous version. If a Public hash is bad, then a new copy is pulled from the Gold copy. The original AIP is available for 30 days, and during that time the Repository can retrieve the original submission if the hashes on the Gold and Public copies are both bad.

Data deterioration is not currently accounted for beyond storing multiple copies of an AIP; essentially, the Repository has put its trust into Amazon's storage data protection. Amazon claims that their services have 99.99999999% durability.[15] Further protections for data assets are ensured through Amazon's use of redundant storage across multiple facilities and retention of one copy on physical media.

Recommendations for Repository to meet/improve compliance with this Requirement:

- While the work of the team to implement archival workflows is nearly complete there is little documentation.
- A comprehensive, end-to-end documentation file would move this rating from a 3 to a 4.

Rating: 3 (The Repository is in the implementation phase)

[14] "Archival Storage function: stores, maintains, and retrieves AIPs. It accepts AIPs submitted from the Ingest function, assigns them to long term storage, migrates AIPs as needed, checks for errors, and provides requested AIPs to the Access function" – OAIS Repository from Wikipedia https://en.wikipedia.org/wiki/Open_Archival_Information_System

[15] https://aws.amazon.com/glacier/, https://aws.amazon.com/s3/?nc2=h_ql_prod_fs_s3

10. Preservation plan

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

Guidance:

The repository, data depositors, and Designated Community need to understand the level of responsibility undertaken for each deposited item in the repository. The repository must have the rights to undertake these responsibilities. Procedures must be documented, and their completion assured.

Response: The preservation and custody transfer of datasets is done at the Organizational Unit (OU) level and policy and procedures for such are set broadly at the NIST level. The responsibility for various levels of preservation is outlined in NIST O 5701.00 Managing Public Access to Results of Federally Funded Research and its Appendix A (the NIST Data Taxonomy).[16] This chart sets out expectations for preservation, review, and discoverability for published results and Standard Reference Datasets, publishable results, derived data, and working data.

There is a social contract between the depositor and repository based on the various policies and records retention requirements to which NIST and its researchers adhere; in addition, employees agree to abide by NIST policies when they are hired. The policies relevant here are in the 5700 series related to providing public access to results of federally funded research and the Records Retention Schedules maintained by the NIST Records Manager. Because the work is federally funded the Repository has the rights to copy, transform, and store the items and provide access to them. Specific Award Conditions in grants require that data resulting from federally funded projects be made publicly available; in some cases, the data may be made available through the NIST repository (NIST Guidance 5702.01).

Transfer of custody and responsibility is not explicit and not clearly explained within the system. Responsibilities are outlined for all staff levels in NIST O 5702.00 Preservation and Maintenance of Published Research Data.[17] While the transfer of custody is clearer than transfer of responsibility, because the user may upload copies to MIDAS and "see them" later in the PDR, there is no explanation of what occurs.

Plans for future migrations and other obsolescence mitigation techniques are included in the design profiles and these plans are supported by the selection of the BagIt standard as the AIP unit. The motivation behind the BagIt profile is that everything associated with a data publication is self-describing. By unpacking a bag, one can understand anything that is in that bag or in related bags, and where to find associated data. Additional details for interpreting that information is documented in the BagIt profiles.

The actions relevant to preservation are not entirely documented, though some have been embedded in the Repository's code. As the Repository is based on community standards and the OAIS model architecture, there has not been a drive by the team to create a high level of

[16] https://www.nist.gov/system/files/documents/2019/11/08/final_o_5701_ver_2.pdf
[17] https://www.nist.gov/system/files/documents/2019/11/12/final_o_5702_1.pdf

documentation. The NIST intranet (INET) webpage *Research Data Infrastructure* is the currently used to lay out the standards and documentation.

All preservation actions are automated by the Preservation Service and the code that monitors integrity so that there is even, accurate, and efficient application of these actions.

Recommendations for Repository to meet/improve compliance with this Requirement:

- The Repository should complete documentation of the Preservation Service and associated workflows.
- An explicit notification of the transfer of custody and responsibility should be included in a "click-through" notification when datasets are submitted for review.

Rating: 4 (The guideline has been fully implemented in the repository)

11. 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.

Guidance:

Repositories must ensure there is sufficient information about the data for the Designated Community to assess the quality of the data. Quality assessment becomes increasingly relevant when the Designated Community is multidisciplinary, where users may not have the personal experience to make an evaluation of quality from the data alone. Repositories must be able to evaluate completeness and quality of the data and metadata.

Data, or associated metadata, may have quality issues relevant to their research value, but this does not preclude their use if a user can make a well-informed decision on their suitability through provided documentation.

Response: The PDR complies with NIST Guidelines, Information Quality Standards, and Administrative Mechanism ("NIST Guidelines") in accordance with Section 515, the Office of Management and Budget Guidelines (M-19-15),[18] and the Department of Commerce Guidelines (67 Fed Reg. 8451).[19] Data and metadata quality is assured through the required review and curation processes that all datasets go through before publication. A dataset must be reviewed by at least the author's Division Chief to confirm adherence to NIST policy and, in some cases, a technical review. Finally, the Research Library performs a curation step that checks readability, completeness, and functionality just before publication. These steps in combination mean that the metadata is well reviewed. The data itself is less strenuously reviewed when a technical review has not occurred. The requirement for a technical review is made at the Organizational Unit (OU) level.

Beyond the review there is the ability to provide a data type which gives an idea of the quality of the documentation associated with the data. These types, in order of most reviewed to least, are SRD, data publication, and public data listing, and are defined in the NERDm documentation.[20] Finally, there are the NERDm/POD field dataQuality that covers data quality, but it is not routinely used by NIST.

Metadata and data can be commented on within the Data.gov system, but not within the PDR. This external system allows for commenting by any interested party and not just the PDR's Designated Community.

Researchers are encouraged to include links to their papers that are supported by the data, or Data Description Papers. These links, typically a DOI or other persistent ID, are included in the POD record, the record sent to DataCite, and in the PDR landing page. By using persistent identifiers, the PDR is helping to build out a network of interconnected research

[18] https://www.whitehouse.gov/wp-content/uploads/2019/04/M-19-15.pdf

[19] https://www.federalregister.gov/documents/2002/02/22/R2-59/guidelines-for-ensuring-and-maximizing-the-quality-objectivity-utility-and-integrity-of-information

[20] https://data.nist.gov/od/dm/nerdm/#sec:Resource See the @type field.

products. In addition, the linking of various research outputs strengthens users' ability to assess data quality.

Recommendations for Repository to meet/improve compliance with this Requirement:

- The Repository should assess if and how the POD field dataQuality could be implemented.

Rating: 4 (The guideline has been fully implemented in the repository)

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8341

12. Workflows

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

Guidance:

To ensure the consistency of practices across datasets and services and to avoid ad hoc actions, workflows should be defined according to the repository's activities and clearly documented. Provisions for managed change should be in place. The OAIS reference model can help to specify the workflow functions of a repository.

Response: The PDR is part of the overarching Open Access to Research (OAR) project that has built up research data infrastructure at NIST. The workflow from ingest to publication is as follows:

1) Researcher fills out MIDAS descriptive record and uploads data files. The files are stored in a networked file system until publication.

2) The record is reviewed at the division level; this is the first of two qualitative checks of the record.

3) After review the Research Library curates the record, deposits the DOI, and publishes the record. At this point the data files are transferred to the preservation system. This is the second qualitative check of the record.

4) The preservation system creates a standard BagIt bag containing both the data files and metadata and transfers it to the AWS Data Management and Integrity Service.

5) The AWS Data Management and Integrity Service places the bags in three S3 buckets: the Gold Bucket, the Public Bucket, and the Long-Term Storage Bucket (Amazon Glacier).

6) At this point the data record and files are available for public discovery on the Science Data Portal.

Documentation for each of these steps is in various stages of completeness. Steps 1 and 2 are well documented in the INET OAR pages, the MIDAS User Guide, and NIST Directives. Step three is not documented; the Research Library is responsible for that documentation of the curation step while the transfer to the preservation system is the duty of OISM. Documentation for the Preservation Service, the NIST use of BagIt profiles, and how the files are transferred can be pulled from the code for this service but has not been formally documented. Finally, documentation for step 5 has not been completed, though the infrastructure development team is aware that it is needed.

Further documentation is also required to fully communicate to the Designated Community the levels and measures in place to secure the system and maintain privacy when sensitive datasets are accepted. This is laid out in the relevant IT System Security Plan which is not a public document. Another piece of documentation that is missing is on data types and how different types of data will affect the workflow. Specifying the data types would also help communicate to depositors and users the data handling techniques used by the system.

Generally, the workflow is automated and standardized; some variation occurs at the review and curation steps, but the allowance for variation is consistent with NIST policy requirements and does not break the overarching workflow.

Recommendations for Repository to meet/improve compliance with this Requirement:

- Technical documentation should be completed for each step in this workflow.

Rating: 3 (The guideline is in the implementation phase)

13. Data discovery and identification

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

Guidance:

Effective data discovery is key to data sharing. Once discovered, datasets should be referenceable through full citations, including persistent identifiers to help ensure that data can be accessed into the future.

Response: The datasets within the Repository are discoverable in two primary locations and via search engines on the internet. The first primary location is the NIST Science Data Portal,[21] where datasets are indexed via their keywords, categories, and authors. Users can enter free-text strings or perform advanced searches. There is also a search API that users can access; it is the same API the SDP interface uses. There are also some high-value, legacy datasets that are housed on their own pages in the NIST domain. These are linked via HomepageURL and dataAccess metadata within a PDR record. The second primary location is https://www.data.gov/. The searchable metadata catalog behind the SDP is based on NIST Extensible Resource Data model (NERDm) which is a superset of the POD standard. In addition, users can use the DataCite search to locate NIST datasets and the DataCite supported R3Data catalog of repositories lists the PDR.

Persistent identifiers are supported throughout the system; internally Archival Resource Keys (ARKs) are used, and each public dataset is assigned a DataCite DOI. The repository facilitates machine harvesting via the metadata API. At this time, it is a "proprietary" API, but NIST is looking to expose it as OAI-PMH in the future.

Recommended custom data citations are offered on each landing page and more guidance is provided on the NIST website.[22]

Recommendations for Repository to meet/improve compliance with this Requirement:

- The Repository should move to the open OAI-PMH standard for exposing data.

Rating: 4 (The guideline has been fully implemented in the repository)

[21] https://data.nist.gov

[22] NIST Copyright, Fair Use, and Licensing Statements for SRD, Data, Software, and Technical Series Publications, https://www.nist.gov/topics/data/public-access-nist-research/copyright-fair-use-and-licensing-statements-srd-data-and#citations

14. 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.

Guidance:

Repositories must ensure that data continues to be understood and used effectively into the future despite changes in technology and the Designated Community's knowledge base. This Requirement evaluates the measures taken to ensure that data are reusable.

Response: Currently there is a lot of room for growth within this requirement. For example, while a user can access the complete metadata in JSON, the metadata record is not typically downloaded with a data file unless it is included in an optional documentation file (such as a README). This is because the method to download the data and metadata as an archive package is not clearly explained on the data landing page or in the help pages for the Repository.

The Repository ensures continued understandability of the data via the possibility to download archive packages with the metadata; this is not ideal because it does not address maintaining usefulness of an individual file. In general, the metadata that is captured is aimed at the dataset level, not individual files. Some metadata can be captured at the file level, from the distribution metadata but is not required by the deposit system in all cases.[23]

Due to the range of the Designated Community's disciplines a wide range of formats are contained by the repository, and none are expressly prohibited. The Repository encourages open formats but accepts whatever the user deposits, and, currently, there is no ability to require changes.

At the moment, other than recommendations encouraging the use of open formats, there are no measures taken to account for the evolution of formats over time.

Recommendations for Repository to meet/improve compliance with this Requirement:

- Datafiles should automatically download with a copy of the metadata record OR any creator-supplied documentation.
- A method to require metadata at the distribution level for all datafiles (even those submitted in bulk) should be explored.

Rating: 4 (The guideline has been fully implemented in the repository)

[23] Files that are uploaded one by one are required to include a title and description, but if the files are batch uploaded via a directory folder then that is not required.

Technology

15. 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.

Guidance:

Repositories need to operate on reliable and stable core infrastructures that maximize service availability. Furthermore, hardware and software used must be relevant and appropriate to the Designated Community and to the functions that a repository fulfils. The OAIS reference model specifies the functions of a repository in meeting user needs.

Response: The Repository is a home-grown system divided into three parts (the back end that ingests metadata and data files, document storage, and the front end that displays information to the public,) that uses the OAIS model for reference. Other standards that are in use are the BagIt standard for AIPs, metadata standards for descriptive metadata (POD and the local NERDm), JSON for defining and validating the metadata, and JSON-LD for associating metadata for standard semantics and RESTful APIs. The database is built on MongoDB while the services are implemented using Java or Python. The entire system runs on the Linux operating system. Finally, SAML is used to authenticate users during the publication phase. All of the software is open source, and most are community supported. Other community supported software used by the Repository team includes Angular TypeScript, the JavaSpring Framework, and javascript.

Storage is provided by Amazon S3 buckets and the Amazon Glacier service. The code is in a public repository, with institution-specific settings living in private repositories. The Repository is built around micro-services (the DOI Service, the Preservation Service, etc.) that can be changed or updated on their own without taking down the entire Repository for maintenance. Maintenance and infrastructure development are planned for and outlined in the OAR Charter. Current availability, bandwidth, and connectivity are sufficient to meet the needs of the Designated Community.

A software inventory is in use and updated regularly by the development team. System documentation is programmatic or written into the APIs that is in line with the software. Supplemental documentation is available on Confluence Wiki pages. The developers have noted that more robust documentation would be useful.

There is a plan for recovering data, and the developers' software repository acts as a backup for the software. The system administrator has helped to develop the procedures required to restore the entire system from scratch in the event of catastrophic failure. Disaster and continuity plans have been written, and demonstrated to work, for full recovery of the system. This plan covers recovery of the system from both the NISTnet VM platform and the AWS platform systems and applications, and they include both a primary and secondary action plans. The procedures for these are documented in the Confluence Wiki. These plans detail recovery of everything, from the OS to the working system, including reloading all of the data from the deep archives. Business continuity is also covered to some extent in the Security Assessment Report (SAR).

Recommendations for Repository to meet/improve compliance with this Requirement:

- Documentation for all technical infrastructure should be completed. This could be structured around the repository functions as spelled out in the OAIS reference model.

Rating: 4 (The guideline has been fully implemented in the repository)

16. Security

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

Guidance:

The repository should analyze potential threats, assess risks, and create a consistent security system. It should describe damage scenarios based on malicious actions, human error, or technical failure that pose a threat to the repository and its data, products, services, and users. It should measure the likelihood and impact of such scenarios, decide which risk levels are acceptable, and determine which measures should be taken to counter the threats to the repository and its Designated Community. This should be an ongoing process.

Response: The PDR is included in the IT system security plan for 600-01 Cross-Laboratory Programs Applications System. The SSP documents security controls, the employees who have security roles, and the risk analysis undertaken. It is maintained and assessed annually by OISM.

Security is assessed annually by NIST's Office of Information Systems Management. Security is low (L) for confidentiality, integrity, and availability of all aspects of the cross-laboratory programs system except for integrity of the data preservation service, which is moderate (M).[24] This system supplements functionality provided by IT system security plan 183-01 for MIDAS, which houses NIST's enterprise data inventory. MIDAS is categorized as M/M/L for confidentiality, integrity, and availability, respectively. Specifically, that MIDAS functionality includes the ability to maintain/update the version of NIST's Enterprise Data Inventory that is serialized as a JSON file, which includes metadata for NIST's public datasets, the AWS S3 MIDAS Gold Bucket, which is primary storage location for NIST's public datasets, the AWS S3 MIDAS Public Bucket, which is where the public accesses those datasets, as well as the process for securely uploading datasets to those locations, along with the daily integrity checking process to ensure continued accurateness of those datasets.

The OAR Preservation Service is rated L/M/L, as it integrates closely with the existing MIDAS application. Specifically, the Preservation Service has direct access to the existing MIDAS NetApp Private Storage (NPS) volumes for purposes of creating a preservation package (known as a "bag") for approved datasets uploaded via MIDAS and generates the hashes for those packages that are compared via existing MIDAS hash comparison processes to ensure ongoing integrity of those packages.

Development of the Repository has followed all guidelines provided by the NIST IT Security Officer. If the Repository needs to deviate from a required standard, that is documented. Authentication and authorizations systems are used only on services internal to NIST such as MIDAS record creation and are based on SAML implemented in Java authentication. Authorization is done via tokens.

Recommendations for Repository to meet/improve compliance with this Requirement:

[24] More information about the Low/Moderate/High ratings for information security can be found in *Standards for Security Categorization of Federal Information and Information Systems, FIPS PUB 199*, Final version issued 2004-02-01, https://doi.org/10.6028/NIST.FIPS.199

- Regularly review and update the security plan currently in place to ensure standard practices are followed over time.
- Develop user-level auditing to detect malicious or erroneous use of the system before an embarrassing error occurs.

Rating: 4 (The guideline has been fully implemented in the repository)

Acknowledgments

The author would like to thank the Office of Data and Informatics team members Gretchen Greene, Raymond Plante, and Deoyani Nandrekar Heinis along with Jon Zhang from the Office of Information Systems Management and NIST Open Access Officer, Katherine E. Sharpless for their time and expertise in performing this self-audit.

Appendix A: Review of Data Intended for Publication

Appendix B: Revising and Removing Information from NIST Inventories and Repositories