Compliance of ScholarsArchive@OSU with requirements for quality research data repositories

Cara Key and Clara Llebot

October 4, 2023

Contents

1 Introduction 2

2 Overview of existing requirements, good practices and frameworks to evaluate good quality repositories 2
  2.1 FAIR principles ............................................. 4
  2.2 Federal funding guidance and requirements for repositories .......................... 6
    2.2.1 The Holdren OSTP Memorandum ................................ 6
    2.2.2 The Nelson OSTP Memorandum ................................ 7
    2.2.3 Desirable Characteristics of Data Repositories ..................... 8
    2.2.4 Federal funding agencies guidance ................................ 9
  2.3 The CoreTrustSeal requirements ...................................... 9

3 Assessment of the ScholarsArchive@OSU institutional repository as a data repository 11
  3.1 About the repository ................................................ 12
  3.2 Free and open scholarship ........................................... 13
  3.3 Organizational Context ................................................ 14
  3.4 Research Data Services .............................................. 15
  3.5 Data curation ......................................................... 15
  3.6 Content management ................................................ 17
    3.6.1 Content restrictions ......................................... 17
    3.6.2 File formats ................................................ 18
    3.6.3 Documentation files .......................................... 19
    3.6.4 Metadata ...................................................... 20
    3.6.5 Identifiers ..................................................... 22
    3.6.6 Versioning of files ........................................... 23
    3.6.7 Licenses ....................................................... 24
  3.7 Analytics .......................................................... 25
  3.8 Preservation ......................................................... 26
    3.8.1 Retention ...................................................... 26
  3.9 Technical infrastructure ............................................... 27
1 Introduction

Oregon State University Libraries and Press is invested in providing high quality research data repository services through the OSU institutional repository, ScholarsArchive@OSU. As the product owner for ScholarsArchive@OSU and the lead of the OSULP Research Data Services program, respectively, we present the following report assessing that quality based on widely accepted standards of measurement for research data repositories. It is our intention that this work may be used in several ways: to articulate the strengths of the repository for data sharing; to identify directions for future development for data repository services; and to communicate transparently about the scope of research data services to the OSU research community.

In section 2, we give an overview of selected resources for requirements and recommendations for research data repositories. In section 3, we describe characteristics of ScholarsArchive@OSU both in general and as a research data repository; within each subsection, we identify relevant criteria from the resources in section 2 and make a determination of our current compliance level. In section 4 we discuss the strengths and highlights of research data services in ScholarsArchive@OSU, as well as some challenges and limitations, and we make several recommendations where OSULP may focus development efforts in the repository for the benefit of research data services. We conclude with some notes on ScholarsArchive@OSU within the context of concurrent projects at OSULP and trends in the broader research data services landscape.

2 Overview of existing requirements, good practices and frameworks to evaluate good quality repositories

There are multiple sources with guidance and requirements for evaluating repositories. The FAIR principles have emerged as a global guidance for sharing research data, and they can be used to guide the development of repositories
(Dunning et al., 2017). Funders all over the world have published recommendations so that researchers can choose an evaluate the quality of repositories. See for example the Desirable Characteristics for Data Repositories document (The National Science and Technology Council, 2022) in the United States, or the Requirements for Open Access Repositories (cOAlition S, n.d.) in Europe.

Organizations representing open repositories have published similar documents, outlining good practices, recommendations and characteristics so that repositories can evaluate and improve their current operations. The Confederation of Open Access Repositories, a global repository network that has 157 members from 51 countries, has published a Community Framework for Good Practices in Repositories, (Repositories, 2022), and the U.S. Repository Network has published a document detailing Desirable Characteristics of Digital Publication Repositories.

Another source of criteria to evaluate and guide the development of repositories can be found in Certifications of repository quality. There are several certification options, with varying levels of requirements and intensity. The Core level is the most basic certification level, represented by the CoreTrustSeal certification. It was created by merging two certificates, the Data Seal of Approval and the World Data System. The CoreTrustSeal is based on 16 requirements that a group of reviewers evaluates based on a self report from the institution. The Extended certification level is similar to the core certifications, but more involved, with a longer list of criteria that need to be met (nestor Working Group Trusted Repositories - Certification, 2009). The Extended certification level is represented mostly by German certifications (DIN 31644/Nestor Seal) and repositories. Formal certification is the most demanding certification, and involves a formal audit, in which reviewers come to the organization. It is represented by the ISO 16363 standard, which is based on the TRAC criteria (The RLG- National Archives and Records Administration Digital Repository Certification Task, 2007).

In order to keep this document focused we will use three of these sources to evaluate the quality of ScholarsArchive@OSU as a research data repository.

1. The FAIR principles, recognizing that these principles have emerged as international guidelines to guide any data sharing effort, including repositories.

2. The Desirable Characteristics for Data Repositories document published by the U.S. Office of Science and Technology Policy, since ScholarsArchive@OSU is a repository serving a public U.S. university that receives most of its funding from federal agencies.

3. The CoreTrustSeal, a certification of trustworthiness that ScholarsArchive@OSU pursued and obtained in 2022.

Sections 2.1, 2.2 and 2.3 summarize each of the requirements, but please check the original sources (cited under each respective section) for a complete description.
2.1 FAIR principles

The FAIR principles, described in 2016 by Wilkinson and collaborators, are a guide for the reusability of research data. These can be applied to datasets, but also to other digital objects (for example algorithms, tools, and workflows) that are part of the research data life cycle. Since data repositories play a big role in making shared data reusable, we will evaluate here how ScholarsArchive@OSU supports FAIR data.

Because the goal of the FAIR principles is to ensure reusability, there are some aspects that are part of a good quality repository that are not emphasized, like digital preservation or data quality. The FAIR principles are broad, and can be implemented in very different ways. This is a strength, because they can be adapted to every case and discipline, but it is also a challenge.

Many articles have been published on the implementation of FAIR principles, in general and for specific applications. See, in particular, Jacobsen et al. (2020) for broad reflections about the application of the FAIR principles. The authors explain that “accessible, robust, widespread and consistent” implementation is needed, but that misinterpretations of the principles (Mons et al., 2017) and misinterpretations of the scope of the principle, are making the task harder. To facilitate the implementation task Jacobsen et al. (2020) interpret each of the principles. Similar interpretations can also be found in the Go FAIR Foundation website\(^1\). The explanations included here after each of the principles are written based on these interpretation notes, focusing on the role of repositories in making data FAIR. Information on how ScholarsArchive@OSU addresses each principle can be found in section 3.

**FINDABLE**

**F1** (Meta)data are assigned a globally unique and persistent identifier. A universally unique identifier that the repository must commit to ensure is persistent for a period of time, even if the resource should no longer exist. The identifier should be resolvable by machines, and ideally link to the resource.

**F2** Data are described with rich metadata. In order to find a resource with a search or filtering, metadata must be as generous and comprehensive as possible, including in cross-domain, interdisciplinary search situations. The minimum requirement is metadata needed to cite the data (creator, title, publication date, publisher, and identifier). Generic and domain-specific descriptors that can be indexed by relevant search facilities should be provided.

**F3** Metadata clearly and explicitly include the identifier of the data they describe. The descriptor should explicitly and unambiguously say what resource is describing, by means of a known predicate, so that the metadata can be found with a search for the identifier by humans and

\(^1\)https://www.gofair.foundation/interpretation
machines. It is a good practice to avoid semantic meaning in identifiers, as these meanings can change and the metadata should contain this information.

F4 (Meta)data are registered or indexed in a searchable resource. There are many searchable resources; these do not index all possible metadata fields in all domains, and there is no uniform way to execute a search, especially a search without human intervention. These challenges are outside of the control of the repository, so a choice must be made about what search engine and metadata to prioritize.

ACCESSIBLE

A1 (Meta)data are retrievable by their identifier using a standardized communications protocol. An identifier must allow the user to retrieve the record, with no additional barriers. An example of a standardized communications protocol is http. It could be non-mechanized as long as it is clearly defined.

A1.1 The protocol is open, free, and universally implementable. Well defined an open reduce the cost of accessing the data, free ensures equitable access to specifications, and universally implementable means that is available to all, without restrictions.

A1.2 The protocol allows for an authentication and authorization procedure, where necessary. Some data must be protected and require additional steps to be accessed. A FAIR repository must have a protocol that allows for authentication and authorization for human or machine requesters.

A2 Metadata are accessible, even when the data are no longer available. Ideally data will be kept long term, but due to financial or legal requirements, the lifespan of some data may be limited. These data may have been used by others, so there should be actionable metadata to understand the nature and provenance of the data.

INTEROPERABLE

I1 (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. A lot of work is generally needed to understand and design ways to combine digital resources, for both humans and computers. Producers of digital resources must use a language that allows for mechanized interpretation, with information about entities and the relationship between entities.

I2 (Meta)data use vocabularies that follow FAIR principles. Vocabularies includes flat or hierarchical vocabularies, data models, structured ontologies, etc. These must be findable, accessible, interoperable
and reusable so that users (humans or machines) can fully understand metadata, including the intent and context, to act on it.

I3 **(Meta)data include qualified references to other (meta)data.** (Meta)data must be connected to resources, specifying clearly the nature of the relationship. This can be done by referencing the external resource’s persistent identifier. For example, providing links to other versions of a dataset, using a FAIR named relations such as ”prior version” or ”next version”.

**RE-USABLE**

R1 **(Meta)data are richly described with a plurality of accurate and relevant attributes.** Digital objects must have metadata that enables users to decide if a discovered resource is appropriate for intended reuse, given a specific task. Detailed metadata with operational instructions for re-use. The metadata author should be as generous as possible.

R1.1 **(Meta)data are released with a clear and accessible data usage license.** A license describing the conditions under which the digital resource can be found must be present without exception. The license must be easily findable. Data and metadata may have different licenses. The lack of a license does not indicate open, it only creates legal uncertainty that in many cases will prevent reuse. The license should be as open as possible to facilitate reuse.

R1.2 **(Meta)data are associated with detailed provenance.** Provenance includes information such as how the digital object was generated, under what condition, the sources that were used to generate it, who should be given credit, processes that were used to clean or filter the data, etc.

R1.3 **(Meta)data meet domain-relevant community standards.** Several domains have standards regarding the minimal information needed to assess the quality of the data and to facilitate reuse. These should be followed, while describing the data with rich metadata.

2.2 Federal funding guidance and requirements for repositories

2.2.1 The Holdren OSTP Memorandum

In 2013 the Office of Science and Technology Policy issued a memorandum (Office of Science and Technology Policy & Holdren, 2013) that required “all Federal agencies with more than $100M in R&D expenditures to develop plans to make the published results of federally funded research freely available to the public within one year of publication and requiring researchers to better account for and manage the digital data resulting from federally funded scientific research”. In a report published in 2021 (Lander, 2021) the White House announced that
all of the more than 20 agencies covered by the memorandum, plus a few other agencies that were not covered by the memorandum, had implemented policies as a consequence of the memorandum. These policies require researchers to make articles publicly available no later than 12 months after publication, and require researchers to write data management plans explaining how the research data will be shared with the public and how privacy and security concerns will be addressed. As a result of this memorandum 18 federal agencies generated policies that required researchers to write data management plans in which they must explain how they will share data. Most of these policies can be found in the SPARC website\(^2\). The report outlined successes, and also pointed to problems that still persist that should shape future policy. The main items to be addressed are (1) the embargo period for articles, (2) a lack of coordination among agencies and a lack of alignment Federal-wide public access efforts with current global open science policies, and (3) alignment of "Federal-wide public access efforts with emerging priorities related to equity, research security and infrastructure for open science".

### 2.2.2 The Nelson OSTP Memorandum

In 2022 a new memorandum -the Nelson memorandum- was released (Office of Science and Technology Policy & Nelson, 2022). The Nelson Memorandum builds on the Holdren Memorandum, and affects all Federal agencies, including agencies with small R&D expenditures that were not affected by the 2013 Memorandum.

- This new policy removes the 12 month embargo for ensuring public access of peer-reviewed scholarly publications.

- There is also a requirement to make scientific data underlying peer-reviewed scholarly publications freely available and publicly accessible by default at the time of publication. Data must be protected when needed according to federal laws and OMB policies.

- Federal agencies must also provide guidance to ensure that the repositories used to share scientific data align with the document “Desirable Characteristics of Data Repositories for Federally Funded Research” (The National Science and Technology Council, 2022).

- Federal agencies must outline policies to establish researcher responsibilities on how federal data will be managed and shared. This includes how to manage potential limitations to the use and sharing of the data due to legal, privacy, ethical, technical, intellectual property or security reasons. It also includes maximizing appropriate sharing of the data (e.g. establishing limited data access), and specifying an appropriate online digital repository where the data will be deposited.

\(^2\)http://researchsharing.sparcopen.org/data
2.2.3 Desirable Characteristics of Data Repositories

The Desirable Characteristics of Data Repositories for Federally Funded Research document (The National Science and Technology Council, 2022) is designated by the OSTP Nelson memo as a reference to use for federally funded research. We expect that the new and updated policies that will be published in 2023 and 2024 as a result of the OSTP Nelson memo will refer to this document. Therefore, this report uses the Desirable Characteristics report to describe ScholarsArchive@OSU.

The main goal of the Desirable Characteristics document is to improve consistency across federal agencies, so that their repository requirements are aligned. The FAIR principles were considered in the creation of this document as the desired result: data that will be findable, accessible, interoperable and reusable.

The report describes characteristics that repositories should have to deposit federally funded research data. There are 14 expectations for all repositories in general, and 7 expectations that apply only to repositories storing human data:

1. **Free and easy access**: providing broad, equitable and maximally open access to datasets and their metadata.

2. **Clear use guidance**: the dataset documentation includes information about the terms of dataset access and use.

3. **Risk management**: the repository can ensure administrative technical and physical safeguards for its content, and its capabilities are well documented.

4. **Retention policy**: the repository has a retention policy.

5. **Long-term organizational sustainability**: there are plans for long-term management of the data.

6. **Unique persistent identifiers**: the repository assigns persistent identifiers to each dataset.

7. **Metadata**: metadata accompanies the dataset to facilitate discovery, reuse, and citation of the data.

8. **Curation and quality assurance**: the repository provides a data curation service in which an expert curator improves the accuracy and integrity of the dataset and its metadata.

9. **Broad and measured reuse**: the metadata that accompanies the dataset in the repository describes the terms of reuse and provides tools to measure the use of the dataset, like number of citations.

10. **Common format**: the data formats used in the repository are widely used, preferably non-proprietary, and consistent with standards used in the discipline.
11. **Provenance**: the repository records the origin, chain of custody, version control, and any modifications experienced by each dataset.

12. **Authentication**: the repository supports authentication of data submitters and the association of persistent identifiers of authors, datasets, and other works.

13. **Long-term technical sustainability**: the repository has a stable technical infrastructure and funding plans that ensure its capability to manage the data long term.

14. **Security and integrity**: the repository engages in documented practices to prevent unauthorized access, modification and release of data, according to the levels of security needed for each dataset.

ScholarsArchive@OSU does not store datasets with any type of sensitive data that may need to be protected for privacy or confidentiality reasons.

### 2.2.4 Federal funding agencies guidance

After the publication of the 2013 memo all the federal agencies that were affected by the memo, plus some that were not, published and implemented data sharing policies. These policies were compiled by SPARC at https://researchsharing.sparcopen.org/data/. Updates on these policies due to the 2022 memo are being documented by SPARC at https://sparcopen.org/our-work/2022-updated-ostp-policy-guidance/.

### 2.3 The CoreTrustSeal requirements

To become CoreTrustSeal certified a repository must fulfill 16 requirements. The following is a summary of these requirements as explained in two documents published by the CoreTrustSeal board (CoreTrustSeal Standards And Certification Board, 2022a, 2022b). The certification is based on a self-assessment, so the fulfillment on the criteria relies heavily on policies and publicly available documentation. Note that ScholarsArchive@OSU obtained its certification in 2022, so the criteria that we had to comply with were the CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022 (CoreTrustSeal Standards And Certification Board, 2019). The self-assessed progress is assigned one of two compliance levels: “In Progress: the repository is in the implementation phase” or “Implemented: the requirement has been fully implemented by the repository”.

**R01 Mission and scope**: repositories makes clear to depositors and users that its mission is to provide access and the active preservation of digital objects.

**R02 Rights management**: the repository communicates all rights covering data and metadata deposit, storage, preservation access and use, and explain how it is managed for humans and machines. The repository must
obtain all necessary rights from the depositor, and have controls in place to monitor compliance with these.

R03 **Continuity of Service**: the repository has a plan to ensure ongoing services (deposit, storage, preservation and access of its data and metadata). It includes management of repository services over time, but also the response when services have problems, and succession planning.

R04 **Legal & Ethical**: data and metadata are created, curated, preserved, accessed and used in compliance with legal and ethical norms. The repository identifies and manages relevant legal and ethical standards (e.g. privacy and confidentiality), and complies with them. Sometimes this requires requesting information from depositors.

R05 **Governance & Resources**: the repository has adequate funding and staff managed through a clear system of governance to effectively carry out the mission.

R06 **Expertise & Guidance**: the repository adopts mechanisms to secure ongoing expertise, guidance and feedback, either in-house or external. This involves identifying the skills necessary to deliver the services, maintaining those skills, and adapting them to the evolution of data types, volumes or rates. Use of the most effective new technologies.

R07 **Provenance and authenticity**: the repository guarantees the authenticity of the digital objects and provides provenance information. Changes to data and metadata should be documented. Provenance information includes the relationship between the deposited object and those provided at the point of access.

R08 **Deposit & Appraisal**: the repository has criteria to evaluate the selection of content (digital objects are relevant) and to ensure that the digital objects are appropriate for preservation and can be understandable to the Designated Community. This criteria includes collection development policies, a documented deposit process, required metadata, preferred formats, etc.

R09 **Preservation Plan**: The repository assumes responsibility for long-term preservation, and has documentation so that the Designated Community understand the level of responsibility undertaken. This should include format migration, emulation, ensuring bit level integrity, minimum retention or preservation periods, and the approach to deletion of data and metadata from the repository.

R10 **Quality Assurance**: the repository complies with quality standards, and documents these for the end users. Some examples of these technical quality criteria for digital objects are acceptable format, metadata schema, metadata content, and links to other digital objects.
R11 **Workflows**: the workflows from deposit to access in the repository are well defined, documented, and change-managed.

R12 **Discovery and Identification**: the repository enables discovery of digital objects. Once discovered the repository should facilitate referencing the object through full citations and persistent identifiers.

R13 **Reuse**: the repository must ensure that the digital objects can be reused over time. Data and metadata must be understandable into the future despite changes in technology. The repository must engage with their Designated Community of users to identify their needs, so that the metadata schemas, controlled vocabularies, ontologies, etc. are useful.

R14 **Storage & Integrity**: the repository should have measures in place to detect unintentional or unauthorised changes to data and metadata, and to recover the correct versions. The storage locations of the repository, the repository’s strategy for multiple copies, and risk management techniques should be documented.

R15 **Technical Infrastructure**: the repository operates on reliable and well managed, stable core infrastructure. The repository must demonstrate that the hardware and software is well managed and appropriate to the needs of the repository functions and Designated Community of users.

R16 **Security**: the repository protects the facility and its data, metadata, products, services and users. The repository plans for potential scenarios that include malicious actions, human error, or technical failure, and decides which risk levels are acceptable.

### 3 Assessment of the ScholarsArchive@OSU institutional repository as a data repository

In the following section we provide a comprehensive overview of characteristics of the ScholarsArchive@OSU repository, particularly those aspects pertaining to its usage as a data repository. Descriptions of repository characteristics are drawn from a combination of the authors' work directly with the repository, public user guides and policies (see section 3.11), local documentation, and consultations with other OSULP staff. Much of this information was compiled in the course of obtaining CoreTrustSeal certification and was furnished in our application (Boock et al., 2022).

For each characteristic, we identify relevant criteria from each of the source documents described above and give a rating of the repository’s compliance with those criteria. The ratings are self-assessed by the authors based on our understanding of the criteria as defined, and informed by the ratings assigned for corresponding 2020-2022 CoreTrustSeal requirements in our certification application. Notes on the FAIR Principles: These are vague and not always obvious
as a standard of measurement. We have relied on the expectations defined by the CoreTrustSeal and NSTC to define compliance with these. The principles that do not map well to these other criteria have been evaluated taking the context of the repository and our user community in mind. For example, ScholarsArchive@OSU is a discipline agnostic repository for which flexibility is important, and will not meet the principles that require discipline specific expertise, resources, standards, etc. The rating scale used is:

- **Fully implemented**: The repository fully meets the relevant aspects of this criterion.
- **Partially met**: The repository addresses a substantial part of this criterion but is lacking in some dimension(s).
- **In planning stages**: The repository has taken initial steps to comply with this criterion.
- **Not met**: The repository has not taken steps to comply with this criterion.
- **N/A**: There is no applicable criterion from this source related to this repository characteristic.

For criteria rated as less than fully implemented, gaps in compliance and potential actions are detailed in section 4.

### 3.1 About the repository

ScholarsArchive@OSU (https://ir.library.oregonstate.edu) is Oregon State University’s institutional repository for the research, scholarship, and publications of the university (Oregon State University Libraries and Press, n.d.). The ScholarsArchive@OSU Collection Policy specifies the types of content that are accepted into the repository, with primary collection areas including faculty research articles, theses and dissertations, and technical reports and white papers (Oregon State University Libraries and Press, 2023).

ScholarsArchive@OSU uses Hyrax\(^3\), an open-source repository system built by the Samvera community and used primarily by libraries, archives, and museums. The Library Information Technology department at OSU Libraries and Press (OSULP) manages ScholarsArchive@OSU and is responsible for customizing features and maintaining the technology stack. The role of OSU Libraries to support the institutional repository and the context of the repository in relation to the Library’s mission are asserted throughout the publicly available ScholarsArchive@OSU Policies, notably in the Terms of Service section, stating: “The Site is maintained by Oregon State University Libraries [...] in support of our mission to disseminate information and research to scholars, educators and the public” (Oregon State University Libraries and Press, 2023).

- **FAIR**, N/A.

---

\(^3\)https://hyrax.samvera.org
• NSTC, N/A.
• CoreTrustSeal, Mission & scope (R01): Fully implemented.

3.2 Free and open scholarship

ScholarsArchive@OSU is an open access repository. Most content is publicly available to anyone with an internet connection at no cost. Users do not need to register or login to view or download the majority of items. A small percentage of items are available only to authenticated members of the OSU community, at the discretion of the author. As described in the ScholarsArchive@OSU User Guide, “Embargoes & Visibility” section 4, depositors may select a limited visibility option, either restricting access to users authenticated via Oregon State University institutional credentials or limiting the view of their work to metadata only (Oregon State University Libraries and Press, 2022c). Depositors may also set an embargo period for their work, after which access restrictions will be limited. These options are described fully in the deposit agreement 5 that depositors must read when they submit their work.

OSU students, faculty, and staff may submit work to ScholarsArchive@OSU. Authentication using an OSU Network Identifier (ONID), a university-level identity account managed by OSU University Information Technology, is required to submit a work. Submitted works are subject to different mediated review and approval workflows depending on the type of content. There is no cost to deposit content in the repository.

ScholarsArchive@OSU provides for direct search and retrieval of repository objects from the repository’s home page. Advanced search options include the ability to search within specific fields and across multiple fields with a single search. All ScholarsArchive@OSU content with public visibility is discoverable in Google Scholar, Google, and other internet search engines. Following DOI registration through DataCite, datasets are also findable through DataCite Search 6, DataCite Commons 7, and aggregator services that harvest DataCite metadata such as OSF SHARE 8.

• FAIR
  – The protocol is open, free, and universally implementable (A1.1): Fully implemented.
  – (Meta)data are registered or indexed in a searchable resource (F4): Fully implemented.

• NSTC
  – Free and easy access: Fully implemented.

4https://guides.library.oregonstate.edu/Scholars-Archive/Embargoes
5https://ir.library.oregonstate.edu/agreement
6https://search.datacite.org
7https://commons.datacite.org
8https://share.osf.io/
– **Authentication**: Fully implemented, with regard to the recommendation for authentication of data submitters; also see section 3.6.5.

- **CoreTrustSeal**
  
  – **Discovery & identification (R12)**: Fully implemented; also see section 3.6.5.

### 3.3 Organizational Context

Oregon State University was established in 1868 and is the state’s largest university. As the institutional repository for Oregon State University, ScholarsArchive@OSU is recognized as a significant contributor to the university’s mission, and accordingly has reliable staffing and financial resources. OSULP intends to maintain the institutional repository and its contents in perpetuity, but repository policy indicates that, “In the unlikely event that OSU and OSULP are unwilling or unable to continue to offer ScholarsArchive@OSU as a service to the University community, staff will work with content contributors and supervising curators to identify other curatorial institutions, within or outside the OSU system, willing to take on future custodial responsibility for the content” (Oregon State University Libraries and Press, 2023).

The repository is operated by OSU Libraries, which has explicit responsibility “for collecting, maintaining, preserving, and providing access to the items in ScholarsArchive@OSU” (Oregon State University Libraries and Press, n.d.). Daily operations and technical maintenance for the repository are managed by the Library Information Technology department. Routine workflows are determined by the department head and the Digital Repository Librarian and supported by a full-time Institutional Repository Specialist. A team of data curators supports research data submissions, as detailed in section 3.5. The ScholarsArchive@OSU User Group is responsible for reviewing and suggesting changes to repository workflows, policies, and procedures. Higher order governance of ScholarsArchive@OSU is within the purview of library administration, and significant changes to policies are subject to approval by library administration before becoming codified.

- **FAIR**, N/A.

- **NSTC**
  
  – **Long-term organizational sustainability**: Partially met with regard to the long-term management of the data within the University; also see section 3.8

- **CoreTrustSeal**
  
  – **Continuity of service (R03)**: Partially met.
  
  – **Governance & resources (R05)**: Fully implemented.
3.4 Research Data Services

OSU Libraries provides research data services for faculty, academic units, and students to support sharing of data produced at OSU. Under the umbrella of data services, library faculty assist researchers with consultations, data curation and review, support for creating metadata and documentation, and writing data management plans (Oregon State University Libraries and Press, 2022b). These services are provided free of charge to members of the OSU community.

As part of OSU Libraries' research data services program, ScholarsArchive@OSU serves as a data repository option for OSU researchers’ datasets. ScholarsArchive@OSU accepts 20-30 datasets per year; the majority are associated with university theses and faculty research articles. Although any dataset created at OSU can be in scope, ScholarsArchive@OSU primarily accepts datasets that are not suited for a discipline-specific repository. ScholarsArchive@OSU also provides a good solution for researchers who are less familiar with the data sharing process and need guidance, or who prefer to have the institutional affiliation and/or local support for their work.

The expertise of OSULP’s research data services is supported by resources and opportunities available through a network of relevant organizations. OSU Libraries maintains institutional memberships to a number of organizations involved in research data initiatives, including the Scholarly Publishing and Academic Resources Coalition\(^9\), Coalition for Networked Information\(^{10}\), Digital Library Federation\(^{11}\), the Center for Research Libraries\(^{12}\), and regional consortia; and OSU is a Samvera Partner institution\(^{13}\). Additionally, OSULP research data services are supported by member programs including DataCite, ORCID, OCLC, and MetaArchive.

- **FAIR**, N/A.
- **NSTC**, N/A.
- **CoreTrustSeal**
  - *Expertise & guidance(R06)*: Fully implemented; also see section 3.5.

3.5 Data curation

In order to ensure that content is discoverable, downloadable, understandable, usable, and shareable, dataset submissions to ScholarsArchive@OSU are mediated by a library faculty Data Curator prior to publication. The Data Management Specialist holds primary responsibility for dataset review, with support from three liaison librarians serving a portion of their time in the Data Curator role, with dataset curation responsibilities aligning with their areas of expertise.

\(^{9}\)https://sparcopen.org
\(^{10}\)https://www.cni.org
\(^{11}\)https://www.diglib.org
\(^{12}\)https://www.crl.edu
\(^{13}\)https://samvera.org
They perform a quality check on dataset submissions to ensure that the data, metadata, and documentation meet the requirements specified in the Research Data Curation and Collection Policies (Oregon State University Libraries and Press, 2023).

For a given dataset, the Data Curator evaluates the data files for readability, completeness, organization, size, and format (see 3.6.2), and checks for sensitive information. The Data Curator verifies that a documentation file accompanies the dataset, and reviews the documentation file for completeness and accuracy (see 3.6.3). The Data Curator reviews the information included in the metadata record, with particular attention to the understandability of the abstract and the selection of an appropriate license (see 3.6.4 and 3.6.7). If any part of the dataset submission does not meet the requirements, the Data Curator will communicate with the depositor by email to ask for improvements, and will provide guidance and support where needed. Deposits are available in ScholarsArchive@OSU only after datasets and the accompanying documentation have been approved by a Data Curator as meeting ScholarsArchive@OSU’s quality standards. Documentation of the review procedures is maintained internally.

Data curation services are assessed periodically through faculty research surveys administered to OSU researchers. In Llebot and Van Tuyl (2019), OSULP faculty investigated how domain experts value the reusability of datasets in the repository, and evaluated the effectiveness of curation services in ScholarsArchive@OSU. In a forthcoming project, faculty are evaluating the main factors that researchers use to decide on a discipline-agnostic repository to use when sharing data.

- **FAIR**
  - *Data are described with rich metadata (F2)*: Fully implemented;
  - *Meta(data) are richly described with a plurality of accurate and relevant attributes (R1)*: Partially met;
  - *(Meta)data are released with a clear and accessible data usage license (R1.1)*: Fully implemented.

- **NSTC**
  - *Curation and quality assurance*: Fully implemented.

- **CoreTrustSeal**
  - *Expertise & guidance(R06)*: Fully implemented; also see section 3.4.
  - *Deposit & appraisal (R08)*: Fully implemented; also see section 3.6.
  - *Quality Assurance (R10)*: Fully implemented; also see section 3.6.
  - *Reuse (R13)*: Fully implemented with regard to seeking input from the Designated Community; also see sections 3.11 and 3.6.4.
3.6 Content management

Works deposited to ScholarsArchive@OSU, including research datasets, are expected to meet repository standards with respect to content, file formats, and metadata description. Research datasets are also required to be submitted with a documentation file. These standards are detailed in the Collection Policy (Oregon State University Libraries and Press, 2023), and written instructions for depositors are provided through the ScholarsArchive@OSU User Guide (Oregon State University Libraries and Press, 2022c). Repository and workflow features that support content management include assignment of unique persistent identifiers, license management, and version control.

The Hyrax software used by ScholarsArchive@OSU provides for mediated and unmediated deposit workflows. Built in software features supporting a mediated workflow include limited admin access for reviewers, with user roles assigned according to work types; automatic notifications to depositors and/or reviewers at any change in status of a submitted work; and an internal messaging system where communications between depositor and reviewer remain attached to the work. As suggested in section 3.5, dataset submissions are subject to a mediated workflow, and are only approved for publication once quality standards are met. The dataset review workflow is documented internally, and updates to the workflow are discussed and agreed upon by Data Curators and repository administrators.

- **FAIR**, N/A.
- **NSTC**, N/A.
- **CoreTrustSeal**
  - *Deposit & appraisal (R08)*: Fully implemented, with specific aspects covered in subsections 3.6.4, 3.6.2, and 3.6.1 below; also see section 3.5.
  - *Quality Assurance (R10)*: Fully implemented, with specific aspects covered in subsections 3.6.1, 3.6.2, 3.6.3, and 3.6.4 below; also see section 3.5.
  - *Workflows (R11)*: Fully implemented.

3.6.1 Content restrictions

Unrestricted, sensitive and confidential data are defined by OSU’s Office of Information Security in the *Infosec Guidebook*. Of these, only unrestricted data - “data intended for appropriate general use within the university” - may be eligible for deposit to ScholarsArchive@OSU. ScholarsArchive@OSU is not an appropriate repository for data subject to FERPA, HIPAA/HITECH regulation of sensitive clinical or medical data, or other personally-identifiable information (PII) with disclosure risk. Confidential or sensitive data are not accepted in

14https://uit.oregonstate.edu/infosec/infosec-guidebook/data-classification-data-element
ScholarsArchive@OSU in any form. Researchers interested in depositing unrestricted human subjects data must communicate that the dataset has human subjects data during the submission process. It is the depositor’s responsibility to redact or anonymize content containing personally identifiable information prior to submission to ScholarsArchive@OSU. They will work with a Data Curator and, when needed, OSU’s Office of Research Integrity to ensure that the dataset complies with all the regulations, laws, and ethical standards applicable to the dataset. Datasets contributed to ScholarsArchive@OSU are monitored by a Data Curator for conformance to these obligations. For datasets with human subject data, Data Curators also confirm that there are no direct identifiers and that the Institutional Review Board approved the sharing of the dataset.

- **FAIR**, N/A. However, this text speaks about why we do not need a protocol that allows for an authentication and authorization procedure, where necessary (A1.2) that would protect sensitive data.

- **NSTC**
  - *Security and integrity*: Fully implemented with regard to preventing unauthorized release of data “according to the levels of security needed for each dataset,” as only unrestricted data are eligible for deposit; also see section 3.10.
  - *Management of human subjects data*: Fully implemented, in the form of a prohibitive policy, though it is ultimately up to the depositor to ensure compliance within the data.

- **CoreTrustSeal**
  - *Legal & ethical (R04)*: Fully implemented, though it is ultimately up to the depositor to ensure compliance within the data.
  - *Quality Assurance (R10)*: Fully implemented; see section 3.6.

### 3.6.2 File formats

ScholarsArchive@OSU is technically format-agnostic, but for purposes of accessibility and preservation, files should adhere to several guiding principles (Oregon State University Libraries and Press, 2023). According to repository policy, content files must be presented in formats that are non-proprietary; for example, CSV should be used instead of XSLX. Data files must be actionable and usable in an analysis application; for example, tabular data should never be shared in PDF or image formats. Data files should also be platform-independent and accessible in all operating systems. A guide to ScholarsArchive@OSU Preferred File Formats document is provided to assist depositors with understanding and selecting appropriate formats (Oregon State University Libraries and Press, 2022c). File format evaluation and recommendations are based on resources such as Library of Congress’s *Recommended Formats Statement*\(^\text{15}\), Open

\(^{15}\) [https://www.loc.gov/preservation/resources/rfs/index.html](https://www.loc.gov/preservation/resources/rfs/index.html)
Preservation Foundation’s *International Comparison of Recommended File Formats*\(^{16}\), the Data Curation Network primers\(^{17}\), and the disciplinary expertise of the submitter.

As mentioned in 3.5, dataset curation involves a review of submitted file formats, and Data Curators provide support for evaluating and converting files to alternative formats as appropriate. Data Curators will approve file formats that do not comply with the file format recommendations only after consultation with the depositor, when recommended formats are not available for the file type, or when formatting of the recommended file format makes the data file substantially less usable. In some cases, non-compliant formats may be submitted alongside recommended formats to facilitate near-term reuse of data.

- **FAIR**
  - *(Meta)data meet domain-relevant community standards (R1.3)*: Partially implemented. It is more likely that a data file format approved by ScholarsArchive@OSU will follow domain-relevant community standards, but the criteria are built with preservation in mind, not community standards.

- **NSTC**
  - Common format: Fully implemented.

- **CoreTrustSeal**
  - Quality Assurance (R10): Fully implemented; see section 3.6.

### 3.6.3 Documentation files

To promote reuse of deposited datasets by other researchers, ScholarsArchive@OSU requires a documentation file(s) to accompany all datasets. Documentation files are submitted and stored alongside the data files, to enhance understanding and assessment of datasets in the repository. The data-level documentation is intended to be more granular and flexible than the bibliographic metadata record described in 3.6.4, and is distinct from the descriptive metadata describing the dataset as a whole.

Documentation files must be submitted in a non-proprietary format, such as TXT, PDF, or XML. A generic ReadMe template is available to depositors for use and modification in documentation files, linked from the ScholarsArchive@OSU User Guide for research data deposits (Oregon State University Libraries and Press, 2022c). Use of the template is not required, and researchers are encouraged to adopt a research metadata standard appropriate for their discipline.

\(^{16}\)https://openpreservation.org/resources/member-groups/international-comparison-of-recommended-file-formats/

\(^{17}\)https://datacurationnetwork.org/outputs/data-curation-primers/
such as Ecological Metadata Language (EML). Documentation files should provide project-level information including title, abstract, creator name(s), primary contact information, license information, preferred citation, data collection methods, and an overview of files. Where repeated, these values should be consistent with the bibliographic metadata. Data-level documentation should also be included, such as a data dictionary, data origin, data type, acquisition details, and processing methods.

- **FAIR**
  - *Meta(data) are richly described with a plurality of accurate and relevant attributes (R1): Partially met. Researchers are required to fill up a template, but the depth and thoroughness of the metadata is up to the researcher.*
  - *(Meta)data are released with a clear and accessible data usage license (R1.1): Fully implemented.*
  - *(Meta)data are associated with detailed provenance (R1.2): Partially met. The documentation template asks researchers to document the provenance of their data.*
  - *(Meta)data meet domain-relevant community standards (R1.3); (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation (I1): Partially implemented. Adoption of a research metadata standard appropriate for the discipline is encouraged, but very few researchers use these standards. Most of the documentation files in ScholarsArchive@OSU are created following the template provided by the repository, which is a plain text file. The template is a version of a commonly used document by other institutional repositories. See section 3.6.4 for a discussion of the bibliographic record.*

- **NSTC**
  - *Metadata: Fully implemented; also see section 3.6.4.*

- **CoreTrustSeal**
  - *Quality Assurance (R10): Fully implemented; see section 3.6.*
  - *Reuse (R13): Fully implemented; also see section 3.6.4.*

3.6.4 Metadata

ScholarsArchive@OSU provides for rich metadata description in order to maximize discovery, understandability, and reusability of content. Separate from the data-level documentation described in section 3.6.3, the bibliographic metadata describes the complete dataset, serving to situate the work within the repository. A prescribed set of descriptive metadata is defined for each type of work to be deposited in ScholarsArchive@OSU, including a specific definition
for dataset deposits. ScholarsArchive@OSU uses metadata properties from the Dublin Core Metadata Element Set and other established vocabularies. A title, creator, resource type, and rights statement are required for all resource types; deposits do not reach the deposit review stage without information in these fields. Use of linked data for metadata properties and for values in several controlled vocabulary fields clearly defines the meaning of those fields and promotes consistency between works; and mechanisms such as drop-down menus and type-ahead suggestions are provided to assist researchers with selecting appropriate values. The schema is updated as needed when a desired additional property is identified, or as properties are changed or deprecated. Technical and administrative metadata, including file characterization and provenance information for the repository object (associated username and timestamp), are generated automatically by the repository software and stored in the repository at the time of deposit and whenever modifications are made.

Depositors fill out metadata fields in a web form at the time of submission; user guidance for entering descriptive metadata is provided to depositors in the ScholarsArchive@OSU User Guide, with dedicated guidance by type of deposit (Oregon State University Libraries and Press, 2022c). Dataset depositors are encouraged to consult with Data Curators if they are unsure how to use any metadata field. A Data Curator checks the metadata as part of the deposit review process. Repository staff reserve the right to edit user-submitted metadata in order to correct or normalize values, to make updates in keeping with changing metadata standards, or to enhance descriptions for greater discoverability. In cases where additional metadata is needed, the depositor is notified and must supply that additional metadata before the work is approved for publication in the repository.

ScholarsArchive@OSU metadata is transformed for use in several ways. Metadata values for selected properties are mapped to bibliographic metatags so they can be parsed by search and indexing services such as Google Scholar. The repository exposes Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) compliant structured metadata for harvesting by external services and organizations. Metadata for datasets is also crosswalked to the DataCite 4.0 schema for DOI registration.

- **FAIR,**
  - (Meta)data are assigned a globally unique and persistent identifier (F1): Fully implemented. All objects in the repository have a citable, persistent url, and datasets are also assigned a DOI.
  - Data are described with rich metadata (F2): Fully implemented for metadata needed to find, use, and reuse the dataset. Domain specific descriptors are included in datasets in an unstructured way.
  - (Meta)data are registered or indexed in a searchable resource (F4): Fully implemented.
  - (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation (I1): Fully implemented with
regard to the bibliographic record. ScholarsArchive@OSU uses a metadata schema that is formal, accessible, shared. See section 3.6.3 for a discussion on documentation files.

- (Meta)data use vocabularies that follow FAIR principles (I2): Partially met. ScholarsArchive@OSU uses some controlled vocabularies (e.g. for rights statements), but not all the metadata fields make use of controlled vocabularies or identifiers.

- (Meta)data include qualified references to other (meta)data (I3): Not met. References to other metadata are included in ScholarsArchive@OSU through the “Related items” field. However, the field does not include a named relation. These references are submitted to DataCite using DataCite’s metadata schema, which includes these named relations that follow the FAIR principles, but it is only done for datasets, not for the other objects in the repository.

- (Meta)data are richly described with a plurality of accurate and relevant attributes (R1): Fully implemented, while recognizing that we could always incorporate more metadata.

- (Meta)data are associated with detailed provenance (R1.2): Partially met. The provenance of the object once it gets submitted to the repository is well documented, automatically. Provenance before the object was submitted can be included as a “Related items” or in the documentation files.

- (Meta)data meet domain-relevant community standards (R1.3): Partially implemented. It is met to the extent that it is applicable. ScholarsArchive@OSU uses community and library metadata standards, but since it is a domain agnostic repository it does not use any domain specific standard.

- NSTC:
  - Metadata: Fully implemented.
  - Provenance: Fully implemented; also see section 3.6.6.

- CoreTrustSeal
  - Quality Assurance (R10): Fully implemented; see section 3.6.
  - Reuse (R13): Fully implemented; also see section 3.11.

### 3.6.5 Identifiers

Persistent URLs are assigned to every item in ScholarsArchive@OSU. These persistent URLs are shown prominently at the top of item landing pages. In any future migration of content to a new repository platform, ScholarsArchive@OSU guarantees that these URLs will be preserved. In rare cases in which an item is removed from the repository (see 3.8.1), a “tombstone” will be created in the
location of the persistent URL. Persistent URLs for related items in ScholarsArchive@OSU as well as persistent or non-persistent URLs for resources in other systems, such as a publication associated with a given dataset, may be entered in an item’s metadata in the Related Items field.

All datasets in the repository are allocated DOIs. Dataset DOIs are registered through the DataCite DOI registration agency. ScholarsArchive@OSU metadata properties are crosswalked to the DataCite schema to complete the registration process, and ScholarsArchive@OSU datasets are indexed by DataCite. The DOI is displayed in the DOI metadata field on dataset landing pages.

The unique OSU Network Identifier (ONID) of the depositor and any person modifying an item is recorded in the repository system. Other personal identifiers for authors and contributors, such as ORCIDs, are not in use in ScholarsArchive@OSU at this time.

- **FAIR**
  - *(Meta)data are assigned a globally unique and persistent identifier (F1): Fully implemented;*
  - Metadata clearly and explicitly include the identifier of the data they describe (F3): Fully implemented.
  - *(Meta)data are retrievable by their identifier using a standardized communications protocol (A1): Fully implemented.
  - *Metadata are accessible, even when the data are no longer available (A2): Fully implemented.*

- **NSTC**
  - *Unique persistent identifiers: Fully implemented.*
  - *Authentication: Partially met, with regard to the recommendation for association of persistent identifiers of authors, datasets, and other works; also see section 3.2.*

- **CoreTrustSeal**
  - *Discovery & identification (R12): Fully implemented; also see section 3.2.*

### 3.6.6 Versioning of files

ScholarsArchive@OSU allows multiple versions of files to be added to a work. This feature can be used to update or roll back file versions, and is used in the case of relatively minor revisions. Attaching a new version of a file replaces the publicly visible and downloadable file with the new version, and also prompts a new file characterization process and generation of a new thumbnail. Administrative metadata including the username and a timestamp associated with the modification are automatically recorded. Previous versions of the file are stored in the system, and can be reverted by repository administrators. Researchers
are encouraged to include additional information about versions of the dataset within the ReadMe template. Record revisions, including version updates or changes to metadata, trigger updates to the administrative metadata stored with the work.

Datasets that have major revisions (e.g. including more data to the dataset), are versioned by creating a new repository object, and linking the versions in the “Related Items” field.

- **FAIR**
  - *(Meta)data include qualified references to other (meta)data (I3):* Partially implemented, with regards to versioning. The versioning of major revisions could be improved.

- **NSTC**
  - *Provenance:* Fully implemented; also see section 3.6.4.

- **CoreTrustSeal**
  - *Provenance & authenticity (R07):* Fully implemented; also see section 3.10.

### 3.6.7 Licenses

Depositors of research datasets are required to select an appropriate Creative Commons or software license for their dataset. Licenses are captured and stored as part of the metadata for the repository record. Additionally, a copyright statement from RightsStatements.org\(^{18}\) is required for every work in ScholarsArchive@OSU, and these are also stored as part of the metadata record. The integration of widely-used licenses and rights statements into item metadata, in the form of URIs, provides unambiguous and stable references directing end users to a reliable source for additional information if needed. Guidance for depositors selecting a rights statement and license is provided in the ScholarsArchive@OSU User Guide, in the “Copyright & Licensing Help” section (Oregon State University Libraries and Press, 2022c), and support is also available from the Data Curators. Depositors may describe more complex license terms, such as differing terms for supplementary files, within the documentation file that accompanies the dataset. All ScholarsArchive@OSU metadata records hold a CC0 public domain dedication, as stated in the repository Collection Policy (Oregon State University Libraries and Press, 2023).

All depositors to ScholarsArchive@OSU must agree to grant Oregon State University a non-exclusive distribution license, according to the deposit agreement that depositors must read when they submit their work. The license grants Oregon State University the non-exclusive right to reproduce, translate, make copies for the purpose of security, backup, and preservation, and distribute a

\(^{18}\)https://rightsstatements.org/en/
submission (including the abstract) worldwide in print and electronic format and in any medium. Depositors must agree that OSU may, without changing the content, translate the submission to any medium or format.

- **FAIR**
  - *(Meta)data are released with a clear and accessible data usage license (R1.1): Fully implemented.*

- **NSTC**
  - *Clear use guidance: Fully implemented.*
  - *Broad and measured reuse: Fully implemented, with regard to the recommendation for describing terms of reuse; also see section 3.7.*

- **CoreTrustSeal**
  - *Rights management (R02): Fully implemented.*

3.7 Analytics

Analytics for ScholarsArchive@OSU are tracked using repository tools and by Google Analytics. Usage statistics from Google Analytics are openly available as a visualization, accessible from repository item pages by clicking an "Analytics" button:

1. On the landing page for each work, displaying pageviews for the metadata record; and
2. On the landing page for each fileset attached to a work, showing both pageviews and downloads for the file.

Data shown in the analytics visualization include total counts over the lifetime of the item as well as daily counts. More extensive metrics pertaining to user activity, item counts, and collection statistics are available to administrators in the repository dashboard and through the Google Analytics interface. Data pertaining to reuse, such as citation counts, are not provided in ScholarsArchive@OSU, though citations and related articles are visible in Google Scholar for some ScholarsArchive@OSU content, depending on the level of indexing of the citing resource. Aggregated data from the repository reporting tools and Google Analytics are reported annually as part of the IPEDS Academic Libraries Component questionnaire for inclusion in the ACRL Academic Library Trends and Statistics Survey.

- **FAIR**, N/A.

- **NSTC**
  - *Broad and measured reuse: Partially met, with regard to the recommendation for providing tools to measure use; also see section 3.6.7.*

- **CoreTrustSeal**, N/A.
3.8 Preservation

The OSULP Digital Preservation Policy pertains to all digital content managed by OSULP, and describes OSULP's mission "to make accessible and hold in trust for future use" our digital resources (Oregon State University Libraries and Press, 2022a). This library-wide policy makes explicit the ScholarsArchive@OSU repository's commitment to preserving its digital resources through a comprehensive digital preservation program. The ScholarsArchive@OSU Preservation Policy details the specific preservation strategies for the institutional repository (Oregon State University Libraries and Press, 2023).

Of particular relevance to the preservation of datasets, ScholarsArchive@OSU is committed to providing bit-level preservation of all content contained in the repository regardless of the file format. ScholarsArchive@OSU relies on regular local and cloud storage backups (see 3.9) along with regular automatic fixity checking (see 3.10 to meet this bit-level guarantee. While MetaArchive is used for selective replication of certain ScholarsArchive@OSU content, no external preservation services are used to store research datasets at this time (Oregon State University Libraries and Press, 2022a).

OSU Libraries and Press guarantees that objects submitted to ScholarsArchive@OSU using recommended file formats will be retrievable in the future, specifically those that OSULP has designated as "Highest Confidence" in the Preferred file Formats guide ((Oregon State University Libraries and Press, 2022c). In the case of deprecation or obsolescence of any recommended file format(s), OSU digital repository staff will engage in format migration or other active preservation strategies as appropriate to provide continuing access to and usability of the objects. An annual preservation assessment includes a review of current file format recommendations as well as an inventory of the file types housed within the repository.

- **FAIR**, N/A.
- **NSTC**
  - *Long-term organizational sustainability*: Partially met; also see section 3.3.
- **CoreTrustSeal**
  - *Preservation Plan (R09)*: Partially met; also see section 3.8.1.

3.8.1 Retention

ScholarsArchive@OSU is committed to indefinite retention of content both in the spirit of transparency and open scholarship, and as a component of the official record of the university. Once deposited and approved, content will not be removed from ScholarsArchive@OSU, with limited exceptions. Under certain circumstances, items may be withdrawn, replaced, or restricted from view, but
the content will still remain in the repository to meet our commitments to persis-
tence and transparency. Special cases for access restrictions, exceptions that
can be eligible for removal, and other details are covered in the ScholarsArchive-
@OSU Withdrawal, Replacement, and Access Restriction Policy (Oregon State

- **FAIR**,  
  - *Metadata are accessible, even when the data are no longer available (A2):* Fully implemented.

- **NSTC**  
  - *Retention policy:* Fully implemented

- **CoreTrustSeal**  
  - *Preservation Plan (R09):* Fully implemented with regard to policies and documentation of “the approach to deletion of data and meta-
data from the repository.”

### 3.9 Technical infrastructure

ScholarsArchive@OSU is an instance of Hyrax\(^{19}\), which is a Samvera repository
solution. Samvera is built on well established components, including Fedora
repository architecture for digital asset management and Apache Solr to power
search and indexing. The Hyrax front end is actively developed by the Samvera
community, designed specifically to meet digital collection repository needs, and
is commonly used in libraries and similar organizations. The OSULP Library
Information Technology unit includes four analyst programmers who provide
custom feature development work to support local needs.

The infrastructure development plan for the OSULP Library Information
Technology department includes running the repository in an on-premises Ku-
bernetes cluster, with continual replication to Amazon Web Services (AWS)
Simple Storage Service (S3) for long term backups. Backups of the filesystem,
individual repository objects, and the relational database are maintained on-
device using an OSU campus server, and off-device in the cloud and in another
timezone via AWS. Backups are run and compared with the originals nightly.
Additionally, snapshots are captured hourly, daily, and weekly for all compo-
nents.

Repository-level exports provide a mechanism for simple, consistent restores,
including regeneration of search indexes. Declarative infrastructure allows for
quick restoration of the services that present the data. All technologies in use are
reviewed at least annually and as needed by a Library Information Technology
department manager and two infrastructure analysts.

- **FAIR**, N/A.

\(^{19}\)https://hyrax.samvera.org
3.10 Data integrity and security

Data integrity is supported by the use of fixity checking at regular intervals. Upon ingest, a system FITS tool produces an original MD5 checksum for files, and Fedora later computes a SHA1 checksum. Fedora’s fixity service is used to record and periodically evaluate file integrity in order to monitor the consistency of repository objects. AWS also runs a fixity service to verify individual object consistency within backups. Fixity reports are reviewed as they are available, with relevant files restored from backups as needed. Reports are also analyzed as a whole during the annual preservation assessment to identify any underlying patterns. With regard to the physical data and media integrity, all data is stored on HDD/SSD locally, and AWS S3 is used for off-premises storage. No optical discs or magnetic tapes or other media that breaks down physically over time are used for storing Scholars Archive content, locally or off-premises.

Oregon State University Information Systems has overall responsibility for managing risks to the university’s data and infrastructure. The Office of Information Security is tasked with identifying threats, such as hackers and malicious software, to the data housed in OSU systems including the ScholarsArchive@OSU repository. Within ScholarsArchive@OSU, depositors must log in using university credentials with multifactor authentication, and the repository system enforces role-based limits on item revision. Provenance metadata is automatically captured as part of the deposit process and each time metadata is revised. This provenance metadata includes the name and email address of the depositor, the name and email address of the person who approved the deposit, the time these actions were taken (e.g. 2013-08-06T17:01:01Z (GMT)), the number of bitstreams added, the number of bytes and the checksum for each, the file names, and the date that the item was made available in the repository.

Physical servers on which the data are stored are housed in a locked and climate-controlled university facility. Access to computing facilities is controlled by university Information Services. Only personnel that need access are granted access according to established university procedures.

- FAIR, N/A.

- NSTC
  - Long-term technical sustainability: Fully implemented.
  - CoreTrustSeal
    - Technical infrastructure (R15): Fully implemented.

- NSTC
  - Risk management: Fully implemented.
  - Security and integrity: Fully implemented with regard to preventing unauthorized access and modification of data; also see section 3.6.1.
• **CoreTrustSeal**
  - *Provenance & authenticity (R07)*: Fully implemented; also see section 3.6.6.
  - *Storage & integrity (R14)*: Partially met.
  - *Security (R16)*: Fully implemented.

### 3.11 Public Documentation

Various types of written documentation about ScholarsArchive@OSU are maintained by OSULP and made publicly accessible for users who are preparing, depositing, or accessing research data.

- An extensive ScholarsArchive@OSU User Guide is provided as a LibGuide at [https://guides.library.oregonstate.edu/Scholars-Archive](https://guides.library.oregonstate.edu/Scholars-Archive). The User Guide gives step by step instructions for depositing different types of content and completing metadata fields, information for users to help manage their work in the repository, and help for searching in the repository. Sections of the user guide that are particularly relevant to research data include:
  - Instructions for dataset deposits, at [https://guides.library.oregonstate.edu/Scholars-Archive/Datasets](https://guides.library.oregonstate.edu/Scholars-Archive/Datasets)
  - Recommendations for preferred file formats, at [https://guides.library.oregonstate.edu/Scholars-Archive/PreferredFileFormats](https://guides.library.oregonstate.edu/Scholars-Archive/PreferredFileFormats)

- Within OSULP’s Research Data Services LibGuide, the "ScholarsArchive@OSU as a Data Repository" page ([https://guides.library.oregonstate.edu/research-data-services/SA-Data-Repository](https://guides.library.oregonstate.edu/research-data-services/SA-Data-Repository)) discusses some of the characteristics of the ScholarsArchive@OSU to help OSU researchers decide whether it is a good fit for depositing their datasets.

- The ScholarsArchive@OSU Policies page, at [https://wiki.library.oregonstate.edu/confluence/display/RP/ScholarsArchive@OSU+Policies](https://wiki.library.oregonstate.edu/confluence/display/RP/ScholarsArchive@OSU+Policies), contains repository policies and terms of service.

- The ScholarsArchive@OSU Deposit Agreement is available at [https://ir.library.oregonstate.edu/agreement](https://ir.library.oregonstate.edu/agreement), and is also presented as part of the deposit process. It explains the depositor’s and repository’s rights concerning works, and also covers visibility options.

- **FAIR**, N/A.
- **NSTC**, N/A.
- **CoreTrustSeal**, N/A.
4 Overall compliance of ScholarsArchive@OSU with research data repository criteria

The expectation is that federal funding agency policies will align closely with the *Desirable Characteristics of Data Repositories for Federally Funded Research* guidance, and therefore will not introduce widely diverging criteria from those evaluated here. However, we will actively monitor policies as they are released, and, if indicated, will produce an updated version of this report in 2025 after the policies are published.

In general, ScholarsArchive@OSU is a strong repository that meets or partially meets most requirements. It was certified by the CoreTrustSeal in 2022, which is an achievement that confirms the quality of the repository. Based on the preceding compliance assessment, we close with a discussion of the strengths of ScholarsArchive@OSU as a data repository, and recommend areas of development to improve compliance levels and overall services.

4.1 Strengths

- **Access:** ScholarsArchive@OSU is effective at its core mission of providing free and open access to OSU scholarship, including research data.
  - Content in ScholarsArchive@OSU is indexed by third party services, including Google Scholar, major search engines, and data aggregators. Assignment of unique identifiers – both DataCite DOIs and system-generated persistent identifiers – enables unambiguous citations and reliable linking to items. Data stored in the repository are widely discoverable through multiple avenues, contributing to broad impact for OSU research.
  - While many other repository options for research data sharing charge fees to deposit materials, OSU community members can deposit materials with ScholarsArchive@OSU at no cost. Furthermore, repository user accounts are integrated with university accounts using SSO authentication, reducing overhead for depositors while also providing a layer of data integrity and security.
  - Barriers to use and reuse are intended to be low. Just as there is no cost for sharing materials, there is also no cost to end users to access materials. No login is required to access or download the majority of ScholarsArchive@OSU content. Rights statements and licenses are consistently assigned and clearly communicated alongside datasets, so that users understand any restrictions on use of materials. Researchers are supported on choosing a license, and are encouraged to select permissive licenses.
  - The repository aims to respect researchers’ needs with regard to public access. Open public access is not mandated for OSU research
data deposits, and system-enforced visibility restrictions are available. However, see the recommendation below to disallow embargoes for any research that is funded by federal agencies.

- **Curation services:** Expert data curation directly contributes to the success of ScholarsArchive@OSU as a research data repository.
  - Curation services allow us to give personal support to our researcher community, and they promote quality and consistency in OSU’s openly shared data.
  - We can see positive results from curation work, both from the point of view of researchers (see Llebot and Van Tuyl (2019)); and for the efficiency of the preservation mission of the repository as evidenced by the regular usage of highest or medium confidence file formats.
  - Hyrax software provides for a mediated workflow to facilitate submission review, which keeps communications between depositors and reviewers attached to submissions. OSULP developers have added custom features to support our local dataset curation workflow.

- **Description:** All dataset deposits are accompanied by both a repository metadata record for the dataset as a whole and a documentation file that captures more granular details with fewer constraints.
  - At the dataset level, ScholarsArchive@OSU provides for rich metadata description with an extensive set of available metadata fields. The repository has attempted to strike a balance between the consistency and interoperability of linked open data and other controlled vocabularies on one side, and ease of use for self-deposit workflows on the other.
  - The metadata standard that is applied to datasets ensures an adequate minimum level of description to enable discovery, understanding, and reuse. This is enhanced by the requirement for data-level documentation.
  - For research data deposits, metadata and documentation review is part of expert data curation, which ensures quality and completeness especially in contrast to unmediated self-deposit systems.
  - This said, recommendations for metadata improvements are offered below, including stronger identifier support and streamlining metadata workflows.

- **Flexibility:** ScholarsArchive@OSU supports open access for many different types of OSU scholarship, and its flexibility should be noted as a strength.
  - As a data repository, ScholarsArchive@OSU meets a wide variety of needs. As a discipline agnostic repository, the collection of datasets it holds reflects the broad range of our research community.
While we make recommendations for formats, the repository is technically format agnostic and can accommodate many types of data. Additionally, documentation files are accepted in several formats.

The low barriers to depositing data, along with responsive guidance by data curators, especially helps to support early career researchers who may not have experience with data sharing.

4.2 Recommendations for future development

This compliance assessment indicates several areas of potential improvement for the repository. The following should be considered in ScholarsArchive@OSU strategy going forward and pursued where practical.

• **Long-term management of data:**
  
  – In order to fully meet NSTC’s *Long term organizational sustainability* criterion, we would develop a more robust preservation plan for datasets using external services such as MetaArchive, Preservica, or similar.
  
  – In order to fully meet CoreTrustSeal requirement R09, *Preservation Plan*, with regard to the guidance ”If preservation levels differ between classes or collections of items, the differences in preservation approach, and the criteria applied to determine the preservation level should be explained” – we would develop and publish a clear definition of preservation levels. Along with this, we might consider a non-permanent retention schedule for some preservation levels.
  
  – In order to fully meet CoreTrustSeal requirement R03, *Continuity of Service*, ScholarsArchive@OSU would need to develop a contingency plan, identifying one or more specific institutions willing to take on custodial responsibility, and establishing a formal written agreement.

• **Content management:**
  
  – In order to more fully meet FAIR principle I3, we would need to improve the versioning of major revisions. For instance, we might consider implementing a workflow using complex objects to represent major revisions of a dataset.
  
  – Given the directive in the OSTP Nelson Memo, we anticipate working with the OSU Research Office on a policy for ScholarsArchive@OSU to disallow any embargo for works derived from federally funded projects.

• **Metrics:**
  
  – In order to fully meet the NSTC recommendation for *Broad and measured reuse*, we would need to provide metrics for reuse of datasets, such as citation counts.
• Metadata:

– While we have adequate support for identifiers for works through automatic assignment of PIDs and manual registration of DOIs, we could do much more with linking identifiers in our metadata, improving compliance with the NSTC Authentication criterion. Adding support for researcher ORCIDs\(^{20}\), the Research Organization Registry (ROR)\(^{21}\), and more robust use of Crossref Funder Registry\(^{22}\) identifiers would help create clear relationships between datasets and related entities. Additionally, we are already considering ways to automate or otherwise streamline the process of DOI registration.

– In order to more fully meet FAIR principle I2, we would need to increase our usage of controlled vocabularies or machine readable data types in metadata field values. The present RDF metadata is somewhat machine readable, but converting and reconciling some text fields to use controlled vocabularies could increase machine comprehension and reduce ambiguity.

– In order to more fully meet FAIR principle I3, we could revise the metadata schema to incorporate qualified references to other related items, where the relationship is specified by an unambiguous predicate.

• Documentation:

– In order to fully meet CoreTrustSeal requirement R14, Storage and integrity, Library Information Technology should maintain (at a minimum) locally shared, up to date documentation of storage systems and procedures. The extended guidance recommends "data flow diagrams covering deposit, curation, and access locations (plus any access restrictions)."

• Curation and quality assurance:

– While we have assessed criteria related to expert data curation as fully met, and indeed consider curation services as a strength of ScholarsArchive@OSU, an institutional membership to the Data Curation Network would grant us access to a pool of specialized expertise as well as professional development opportunities that would further strengthen repository services. This would particularly enhance our alignment with NSTC’s Curation and quality assurance and CoreTrustSeal R06, Expertise & guidance criteria.

\(^{20}\)https://orcid.org
\(^{21}\)https://ror.org
\(^{22}\)https://www.crossref.org/services/funder-registry/
5 Conclusions

This report shows that ScholarsArchive@OSU has fully implemented most of the criteria that different sources define for good quality data repositories. Thus, ScholarsArchive@OSU is well positioned to support OSU researchers with their data sharing needs, including researchers that must comply with federal funding data sharing requirements. Researchers can be sure that OSU will be a good steward of their data, since the repository complies with mostly applicable industry standard metrics.

We recognize that best practices for data repository services are evolving, and that there is always more that we can do to support researchers. Assessing a criteria as "fully implemented" doesn’t mean that there are no possible service improvements we can make, or even that we aren’t interested in making them. Still, overall and with finite resources, it will be more impactful to apply our efforts to areas that are assessed as not fully implemented.

This report identifies challenges and recommendations for future development. These opportunities for improvement focus on the use of identifiers; strengthening of our preservation and sustainability plan; improvement of user facing information such as metrics for reuse of datasets, promotion of the repository, and public facing policies; metadata work to improve indexing, machine readability, and to document the relationship between datasets and other items.

Since ScholarsArchive@OSU is an instance of Hyrax, we plan on relying on developments from the Samvera community for the implementation of some of these recommendations. The community has shown interest in feature development to support data sharing, and in 2022 convened a working group to explore data repository needs for Hyrax in which they surveyed members of the community for input. For example, the complexity of including ORCID identifiers into the repository will be better addressed with official software releases.

For the rest of the recommendations, OSU Libraries is working, and must continue to work, in strengthening internal structures that support these tasks. Finding accountability mechanisms can be a useful strategy to move forward. For example, the CoreTrustSeal certification application helped us revisit our preservation strategy, improve policies, and bring together a digital preservation interest group, and a cross-departmental repository user group.

ScholarsArchive@OSU does not aspire to fully implement all criteria, because not all of the criteria align well with its scope and mission. A crucial characteristic of ScholarsArchive@OSU is that it is a discipline agnostic repository that hosts many types of objects in addition to datasets. Many of the criteria that refer to discipline specific services (e.g. the FAIR principles have many of these) are out of scope and not a high priority. That being said, there exist tools that ScholarsArchive@OSU should consider piloting. The Data Curation Network offers disciplinary expertise to curate datasets, and support for repositories to learn and develop expertise in data curation to their members. Becoming a member has a cost that OSULP should evaluate.

https://datacurationnetwork.org/
In the data repository landscape there are many repositories (besides institutional repositories) that are also discipline agnostic. See for example Dryad 24, Zenodo 25, Figshare 26 or Mendeley Data 27. Several of these discipline agnostic repositories are participating in a Generalist Repository Ecosystem Initiative with the goal to “develop collaborative approaches for data management and sharing through inclusion of the generalist repositories in the NIH data ecosystem” and “to better enable search and discovery of NIH-funded data in the generalist repositories” 28. These repositories are generally larger, better funded, and specialized in research datasets, which means that, often, they can offer better features, and have robust software that works well with other tools commonly used by the research community (e.g. like GitHub). The role of our institutional repository should not be to compete against these repositories, but to identify its niche. It is important to identify clearly what is the mission of our repository, and where are our strengths, so that we can leverage what we do well and be intentional in where we put our limited resources. We believe that institutional repositories have a role, but we need to identify it and advertise the role more effectively.

Diana Park, Jane Nichols, Cara Key and Clara Llebot are already working on a project to identify these strengths (and opportunities for improvement) from the point of view of OSU researchers and students. One of our hypotheses is that institutional repositories excel at developing a relationship with researcher, doing personalized (although not discipline specific) data curation, and educate researchers that are not very familiar with data sharing.

The sustainability of ScholarsArchive@OSU as a data repository must be considered too. Data sharing mandates are getting stricter and it is likely that we will see increase in datasets submitted in the repository. Having a defined scope for our services should help us to identify these who will benefit the most from depositing their research data into ScholarsArchive, and to know when to refer researchers to other repositories to keep our services doable. If the scale of data sharing will overwhelms our repository there are many paths that we could follow to move forward. It is a good moment to explore these, to be informed and prepared. For example, several of us have been talking to Dryad, a non profit repository that offers a paid subscription model that allows researchers of participating institutions to submit datasets free of charge. The datasets are curated, and their software is well integrated with many publishers, offering a streamlined peer review experience. An other avenue that should be explored is collaboration with other OSU colleges and offices. The support that OSU offers to researchers for their data storage and computing needs are generally fragmented. There may be opportunities for collaboration that should be explored, especially related to the creation of the new OSU Data Center, which is scheduled for 2025.

24https://datadryad.org/
25https://zenodo.org/
26https://figshare.com/
27https://data.mendeley.com/
In conclusion, we have invested, and at this time we have a high quality data repository, and expertise in our staff to manage it. This expertise is helpful not only for managing the repository, but to provide services to the university and OSU researchers and students. It is a good moment to think about the directions to move forward.

References


CoreTrustSeal Standards And Certification Board. (2022a). CoreTrustSeal Requirements 2023-2025. https://doi.org/10.5281/ZENODO.7051011


Oregon State University Libraries and Press. (n.d.). ScholarsArchive@OSU. Retrieved March 15, 2023, from https://ir.library.oregonstate.edu/


