RDA Common Descriptive Attributes of Research Data Repositories

Michael Witt (0000-0003-4221-7956), Matthew Cannon (0000-0002-1496-8392), Allyson Lister (0000-0002-7702-4495), Washington Segundo (0000-0003-3635-9384), Kathleen Shearer (0000-0001-8617-5781), Kazu Yamaji (0000-0001-6108-9385) and the Research Data Alliance Data Repository Attributes Working Group (DRAWG)

December 1, 2023 | CC-BY 4.0 International | DOI: 10.15497/rda00103

Introduction and Impact

A complete and current description of a research data repository is important to help a user find the repository; to learn the repository's purpose, policies, functionality and other characteristics; and to evaluate the fitness for their use of the repository and the data that it stewards. Many repositories do not provide adequate descriptions in their websites, structured metadata and documentation, which can make this challenging. Even fewer make this information available in a machine-readable or actionable manner, which hampers interoperability. Descriptive attributes may be expressed and exposed in different ways, making it difficult to compare repositories and to integrate repositories with other infrastructures such as registries. They can be difficult to navigate and find: they may be locked behind authentication, obscured within workflows, or buried in myriad documentation and web pages.

Well-described data repositories provide value and impact to a broad set of stakeholders such as researchers, repository managers, repository developers, publishers, funders, registries and others to discover and utilize data repositories. Some motivating use cases for the development of these attributes include:

- As a researcher, I would like to be able to discover repositories where I can deposit my data based on attributes that are important to me.
- As a repository manager, I would like to know what attributes are important for me to provide to users in order to advertise my repository, its services, and its data collections.
- As a repository developer, I would like to understand how to express and serialize these attributes as structured metadata for reuse by users and user agents in a manner that can be integrated into the functionality of my repository software platform.
- As a publisher, I would like to inform journal editors and authors of what repositories are appropriate to deposit their datasets that are associated with manuscripts that are being submitted.
- As a funder, I would like to be able to recommend and monitor data repositories to be utilized in conjunction with public access plans and data management plans for the research that I am sponsoring.
- As a registry, I would like to be able to easily identify and index attributes of data repositories to help users find the best repository for their purpose.

In this guide, a set of common, concise, high-level descriptors are defined that represent information that can be useful in describing a data repository along with examples of how each attribute can be expressed as metadata in different schemata, limitations and potential complications that each might pose in harmonization, a brief rationale for why each attribute is important and a gap analysis of how easy or difficult it may currently be to locate this information from a data repository. The attributes are conceptual with examples provided for illustration purposes without endorsement of any particular approach, standard or implementation.

RDA Common Descriptive Attributes of Research Data Repositories

Each attribute is structured following this pattern:

> Attribute [Label, Description, Examples, Schemata, Notes, Rationale, Gap Analysis]
> Gap Analysis Scale: 1 - easy to find, 2 - difficult, 3 - very difficult or doesn't exist

1. Label: **Repository Name**
   Description: The name that the repository provides and what users commonly call the repository.
   Examples: Data Repository for the University of Minnesota; DRUM
   Schemata: re3data:repositoryName & re3data:additionalName
   Notes: Repositories often have multiple names such as a full name and an acronym, or the repository may be a part of a larger data service or suite of repositories and share a name with it. Less frequently, a repository name may change over time, and some users may continue to use an older name for it. The name of the repository can sometimes be conflated with the name of the organization that provides the repository. Names can also be expressed in different languages. The recommended usage is to refer to a repository by the name that it uses to advertise itself, e.g., on its website or in a provided citation.
   Rationale: The repository name is one of the most essential and principal attributes for all stakeholders, because it provides them with something to call a repository, and a repository builds a reputation and awareness of its collections and services around its name. All stakeholders will commonly search for and reference a repository by its name.
   Gap analysis: 1 - Easy to find

2. Label: **URL**
   Description: The uniform resource locator (URL) that serves as the homepage and primary entry point for accessing the repository on the World Wide Web.
   Examples: https://datadryad.org; https://www.pangaea.de
   Schemata: IETF RFC 1738 Uniform Resource Locators
   Notes: The URL represents a link to the main landing page on the web for the repository. A repository may have other URLs that represent alternate points of entry such as web pages with information about the repository, its services, or its collections as well as web-based machine interfaces.

Rationale: The URL is the primary way that users will access and bookmark the repository.
Gap analysis: 1 - Easy to find

3. Label: **Country**
   Description: The country in which the repository operates.
   Examples: Japan, Brazil, United States
   Schemata: GeoNames
   Notes: In some cases, a repository may be owned, managed, or otherwise associated with an organization that spans multiple countries or a distributed infrastructure such that components of the repository reside in different countries. The country may be determined by the federal, legal jurisdiction that most directly affects the operation of the repository.
   Rationale: Essential for national repository services and providing clarity and regulatory compliance for data governance.
   Gap analysis: 1 - Easy to find

4. Label: **Language**
   Description: The native language of the user interface of the repository.
   Examples: EN-US, Cantonese, Hindi
   Schemata: dcterms:language
   Notes: Many repositories may be indexed in English or another language than the native user interface of the repository. Some repositories present interfaces in multiple languages. The language of the interface may be different than the language of the metadata or datasets in the repository. Translation software can be useful for some users but may not accurately represent all functionality.
   Rationale: It is important for users to know if the repository presents functionality in a language that they can read, and it may be useful for enabling national infrastructures.
   Gap analysis: 1 - Easy to find

5. Label: **Organization**
   Description: The organization responsible for the data repository.
   Examples: Chinese Academy of Sciences; figshare; University of Auckland
   Schemata: ROR, schema.org:Organization
   Notes: One or more organizations can be responsible for a repository, individually or as part of a collaboration, and in a variety of different capacities such as providing funding, infrastructure, administration, governance, curation, and commercial services. The hierarchy and relationships among organizations can be complex and opaque to an end-user. For example, an organization may be a division or unit of a larger organization. Some users may be required or influenced to use a specific repository based on their organization, and the organization/s behind a repository must be transparent.
   Rationale: The organization and its reputation and commitment to data management may influence a user’s trust and decision to use a data repository.
   Gap analysis: 2 - Difficult to find

6. Label: **Contact**
   Description: Contact information for the data repository.
   Examples: ICPSR-help@umich.edu; tel:+49-123-4567890; https://snd.gu.se/en/contact
   Schemata: schema.org:Contact Point
   Notes: Contact may be an email address, mailing address, telephone number, or a form. Many data repositories have multiple points of contact for general inquiries, technical support, or administration. Some contacts may be restricted to registered users. Some information about contacts may be out of date, and some data repositories may not be able to respond or respond promptly to every inquiry.
   Rationale: A user may need to communicate with the staff of the data repository for inquiries or support. Contact information is an important starting point for building trust.
   Gap analysis: 1 - Easy to find

7. Label: **Description**
   Description: A general description of the data repository.
   Examples: “Edinburgh DataShare is an online digital repository of multi-disciplinary research datasets produced at the University of Edinburgh.”
   Schemata: FAIRsharing:Description
   Notes: Descriptions provided by repositories are highly variable, and sometimes do not even exist. There can be multiple descriptions across the repository website, and it can be difficult to know which one is the “canonical” or most complete description. The description will typically include information about the institution and the scope of the data collected in the repository as well as references to its functionality and policies.
   Rationale: Users need to understand basic information about a data repository to determine if it is appropriate for them to further explore for their use. A robust description can also be indexed to improve findability of the repository by search engines.
   Gap analysis: 2 - Difficult to find

8. Label: **Research Area**
   Description: The subject classification of datasets in a repository.
   Examples: Biomedical Science; Geochemistry; Demographics; Humanities
   Schemata: LCSH, Web of Science Subject Classification; DFG
   Notes: Datasets are often organized at a high level by a fixed subject classification in addition to keywords. While free text can be more flexible, detailed, and messy, controlled vocabularies provide consistency and organization. Research areas can be associated with disciplines or communities. The granularity and shared understanding of what a research area constitutes can vary depending on the level of specialization of the repository, geography, and the research domain. Generalist repositories may struggle to represent their multidisciplinary collections in some classification schemes.
   Rationale: It is important for users to understand the subject matter of the datasets in order to know if the repository is appropriate for use for research interest or in their discipline. Classification by research area using controlled vocabularies helps to organize datasets for improved browsing, findability, and interoperability.

Gap analysis: 2 - Difficult to find

9. Label: **Persistent Identifiers**
   Description: The repository provides or utilizes persistent identifiers.
   Examples: DOI, ORCID, RRID, CSTR
   Schemata: DataCite Metadata Schema 4.4; ORCID Record Schema
   Notes: Persistent identifiers create a precise reference for datasets and actors associated with a data repository, including identifiers for digital objects, content creators, research organizations, and the repository itself. Identifiers can be important for the purposes of citation and attribution, versioning, measuring impact, and making associations with other datasets or actors. There are a wide variety of different kinds of identifiers, some of which have been deprecated. It is possible for the same dataset or actor to have multiple identifiers of the same or different kinds. Some identifiers may not be globally unique or resolvable outside of the repository.
   Rationale: The use of persistent identifiers is a best practice that has been reinforced by research funder requirements, publisher guidelines, and repository certifications. Users may require identifiers in order to get proper credit for sharing their data.
   Gap analysis: 1 - Easy to find

10. Label: **Machine Interoperability**
    Description: Application Programming Interfaces (APIs), service endpoints, and other protocol interfaces that enable machine access to a repository.
    Examples: OAI-PMH, FTP, REST, SPARQL, SWORD, OpenDAP, RDFa
    Schemata: FAIRsharing:Data Processes
    Notes: Machine interfaces extend the functionality of a repository beyond a person using a web browser to a client application that can be operated by a user or user-agent. This may include embedded information that is made available through HTTP but is intended for machine consumption. APIs may be used to harvest data for the repository, or for researchers to be able to interact with the underlying data without using the repository front-end. Other APIs and endpoints may transfer data or perform operations using the datasets themselves that enable value-added functionality for client-side applications.
    Rationale: Users who are writing code or using applications beyond a web browser need to know if their programs can interact with the repository and what protocols are supported. Interoperability is key to furthering the FAIR Principles, in particular, for user agents.
    Gap analysis: 2 - Difficult to find

11. Label: **Metadata**
    Description: Format/s of the metadata that describes datasets in a repository.
    Examples: Dublin Core, NetCDF, DDI
    Schemata: re3data:metadataStandard
    Notes: The format of the metadata will typically drive a user's ability to search for datasets within the repository and help discovery services to properly harvest and index metadata. It also helps data producers understand how to prepare their data for deposit.

A repository may support one or more metadata standards or it may implement its own custom format. Different communities and different kinds of data employ different standards. In some cases, metadata may be serialized into files that are included in a dataset as documentation that are not available to the search functionality of a data repository. Other important metadata may track the provenance of datasets and maintain version control, links to other related objects, usage statistics, and citations.
Rationale: Making it clear to the user explicitly how datasets are uniformly described will help them properly prepare their data for deposit and understand how to better search and explore datasets in the repository. Structured metadata is also important for machine interoperability and indexing.
Gap analysis: 2 - Difficult to find

12. Label: **Curation**
Description: Curatorial services performed by repository functionality or personnel that enhance or otherwise add value to datasets and create purposeful collections of data.
Examples: CoreTrustSeal:Levels of Curation
Schemata: re3data:qualityManagement
Notes: Curation is an expansive term that may include a wide range of cross-disciplinary activities that could entail actions that are performed by the repository or repository personnel beyond the initial effort of the data producer who deposits a dataset. These may include enhancing metadata, performing quality assurance, reviewing of datasets, selection and deselection of datasets, organizing datasets into collections, reformatting datasets, creating new points of access, or other activities to improve, enhance, and ensure collected datasets. Curation captures additional context and enhances the discovery and use of the data in ways that create value and enable the long-term fitness of the data. While there is a broad consensus that curation is important, there are different understandings of what curation entails and how to articulate it, which present challenges. The definition of curation and curation as a professional repository practice are currently evolving and difficult to measure.
Rationale: Users of a repository need to understand what actions may be taken to data that they deposit, what benefits curation services offer them and reuse of their data, and be informed of any additional time it may take.
Gap analysis: 3 - Very difficult to find or does not exist

13. Label: **Terms of Deposit**
Description: Policies that explain what datasets the repository will accept for deposit, from whom, and under what conditions, including costs.
Examples: https://researchdata.reading.ac.uk/submission_policy.html
Schemata: re3data:dataUpload
Notes: Terms of Deposit will typically cover who is able to submit data to a repository for deposit, what kinds of data will be accepted, if there are any size or other restrictions, any costs, and what rights the user is granting the repository. This information can be obscured in a broader set of policies such as Terms of Use, or it may be called by a different name such as a submission policy or conditions for deposit. In some cases, this

information is only visible to registered users of the data repository, and it may be presented to the user as a part of the deposit workflow. It is common while depositing a dataset that a user will ascribe a license for reuse of the dataset by others.
Rationale: Terms of Deposit are critical to help a user determine if a repository will accept their data, how long it may take, and if there will be an associated cost.
Gap analysis: 2 - Difficult to find

14. Label: **Terms of Access**
Description: Policies for who can view and access a dataset and under what conditions.
Examples: “Files may be deposited under closed, open, or embargoed access.” https://about.zenodo.org/policies/
Schemata: FAIRsharing:Data Access Conditions
Notes: Terms of Access include information about the data availability such as whether the data are openly available, can depositors limit access to data, if the repository supports peer review functionality, and if there is an embargo, registration or authentication or fee for users to access data. Data repositories may have different Terms of Access for different datasets or data collections. Datasets may be visible but not available if they have been embargoed or otherwise limited. Restrictions at the repository level may affect access to datasets; for example, a user may need to register or pay a fee to access a repository before datasets are made available to them. Other terms may be policies or functionality that ensure data protection, including security protocols. Once a dataset has been acquired by a user, their reuse may be further governed by a Data Use License.
Rationale: Users who need to comply with open access mandates need to understand if a repository fulfills this requirement.
Gap analysis: 2- Difficult to find

15. Label: **Dataset Use License**
Description: The terms of reuse of datasets that are provided by a repository.
Examples: Creative Commons, ODbL, GNU
Schemata: DUO, dcterms:license
Notes: Typically a license will be selected by the data producer when they deposit a dataset that will govern the reuse of the data by others. A repository may offer and support multiple, different licenses. When possible, it is helpful to link to a machine-actionable version of the license. The license should be clearly presented when a user downloads a dataset, ideally being included with the download. Some repositories obscure what licenses are available in their documentation, and they may only be presented in the later part of the deposit workflow.
Rationale: Users need to understand what they can do with data that they download, and users who may submit data for deposit need to know what license options are available to them.
Gap analysis: 2- difficult to find

16. Label: **Certification**

Description: Earning a certification, typically by means of a third party audit or community endorsement process, presents evidence that a repository meets a formal standard and adheres to a set of best, professional practices.
Examples: CoreTrustSeal, ISO 16363
Schemata: schema.org:credentialCategory
Notes: The most common certifications for data repositories relate to trustworthiness, which is to say that a repository advertises its commitments to data stewardship and provides evidence that has been audited to ensure that the commitment has been implemented in its systems and processes. Other certifications may relate to security, privacy and data protection, organizational membership, or other areas. Repositories that earn a certification may advertise their certification with a badge that can enhance their reputation and increase the level of trust users have in them. Most certifications are managed by a third party and require renewal after a period of time. Repositories that have invested in getting certified generally do a good job of advertising this to their users; however, the majority of repositories are net yet certified.
Rationale: Funders or publishers may require deposit of data in a certified repository, and certification helps a user trust and understand the repository's commitment to stewarding their data.
Gap analysis: 3 - Very difficult to find or does not exist

17. Label: **Preservation**
Description: Policies that explain the repository's commitment and processes that ensure the long-term preservation, fitness, and availability of datasets.
Examples: PURR, https://purr.purdue.edu/legal/digitalpreservation
Schemata: FAIRsharing:Data Preservation Policy
Notes: Preservation policies detail how the repository mitigates risks to data that it is archiving and helps users make decisions about file formats to use and other practices to help future-proof their data. Policies may include a digital preservation strategy plan, file format recommendations, preservation support policies, retention and transition policies, what tools and platforms are used for preservation, and an overview of the repository's preservation practices. Despite the variety of standards for evaluating preservation policies and practices, some repositories do not employ or do not clearly advertise their commitment to preservation. There may be different levels of preservation support for different kinds of data or groups of users. Preservation is generally considered to be active, evolving, and on-going practice.
Rationale: A clearly stated commitment to preservation enables trust that the repository will be a good steward for a user's datasets.
Gap analysis: 3 - Very difficult to find or does not exist