Data Repository Attributes WG Case Statement | RDA

RDA Data Repository Attributes WG Case Statement CoreTrustSeal Board Feedback

December 23, 2021

DOI: 10.5281/zenodo.5801501

The CoreTrustSeal Board is grateful for the opportunity to provide feedback on this draft case statement for a Data Repository Attributes working group. We fully support the development of clearer and more standardised use cases for repository information that result in consistent and validated repository metadata being made available, and maintained in machine-actionable ways. This aligns with the CoreTrustSeal goal to identify scenarios where repository metadata (either self-declared, or validated in some way) can be integrated into the repository review and certification process.

We are glad to see that the feedback provided by CoreTrustSeal, World Data System, COAR, European University Association Science Europe,[1] related to the draft paper on Data Repository Selection: Criteria That Matter[2] were taken into account in the emergence of Repository Features to Help Researchers: An invitation to a dialogue[3]. Some of these concerns remain important for the conduct of the working group, but we will not reiterate them here.

Below we provide some specific comments and suggestions for the text of the case statement. We would also like to raise several, more general points.

The proposed outputs make no reference to their technical implementation in the registries concerned, or the user experience of how repository metadata could be used to present filtered lists of 'recommended' repositories. Neither does it suggest that the publishers will clarify how they intend to use the repository metadata related to their own use cases. Potential tension between stakeholders' implementation based on their use cases is important. For example, a publisher-driven recommendation that focuses on "*facilitating the data peer review process for publishers and their authors*" could do so to the exclusion of arguably more important repository features, such as the ability to provide assurances of long term accessibility and understandability of the data. The impact of that type of use case would be quantifiably different from those designed to support "*integration with funder review processes*". Ideally all use cases

[1] https://doi.org/10.5281/zenodo.4649136
[2] https://doi.org/10.5281/zenodo.4084763
[3] https://doi.org/10.5281/zenodo.4683794

would be mapped to the "descriptive attributes" that they intend to use so that impact could be assessed.

Clarifying publisher use cases is important and valuable. We might question whether "*as a publisher, I would like to inform journal editors and authors of what repositories are appropriate to deposit their datasets"* is the best wording here. The publisher use case for a journal article or paper may simply be that it remains unchanged over time, the data users' use case for the underlying data may be more focussed on the kind of curation that necessitates managed change. This provides another example where impact can only be evaluated if use cases for publishers and other stakeholders) and the attributes they depend upon are mapped together by the working group.

The ability of repository metadata (particularly self-declared metadata) to allow a user (by which we assume a data depositor or data reuser) to "*evaluate the fitness for their use of the repository and the data that it stewards*" may be limited and we suggest some rephrasing.

We would note that "*the harmonization of a common set of repository attributes*" may provide a common, generic and high level set of metadata elements, but not provide for "*the needs and requirements of different communities*". This is an important distinction as efforts like FAIR data strive to define more specific disciplinary and domain needs for (meta) data and (meta) data curation and preservation.

The close co-dependency of metadata and data are relevant here. CoreTrustSeal seeks to identify and credit 'Trustworthy Digital Repositories' (TDR), but as we work towards better interoperability through valuable work such as that provided by re3data and FAIRsharing we also need trustworthy metadata registries . We consider that many of the organisational, technical and digital object management variables applied to repositories could, and perhaps should, be applied to metadata registries. We would suggest that the evaluation of the applicability of attributes to data repositories is extended to the metadata registries themselves. This should not present a challenge, given the co-Chairs involved.

Action 6 sees the adoption of the outputs of the WG as a RDA Recommendation as the starting point for "*broader community input, review, revision, and adoption*". We would suggest that the emergence of such a recommendation might be construed by some as an end point, especially with no stated intention or methodology for ongoing input or review.

The working group does not explicitly address the challenges for repositories of defining, maintaining or exposing these variables for harvesting/publishing them into compliant systems, nor does it envisage any criteria for the validation or maintenance of this metadata once received. Gaps between specification of repository attributes and repository compliance could present costs to repositories in terms of resources and reputation. Re3data and FAIRsharing, as

key actors in the repository registry space, and as co-Chairs of this proposed working group, do not undertake to adopt the recommendations, share proposals for their use at the user experience level, or provide transparent schemas and associated update and change management processes for their use in their systems.

Some clarity on how the main proponents, publishers and maintainers of these variables will act in the future would be valuable to the working group case statement, and to the wider community of data services users. An ongoing mechanism, such as an RDA maintenance group, would provide assurance of ongoing engagement and curation of these outputs.

Case Statement: Data Repository Attributes Working Group

1. Charter

The Data Repository Attributes Working Group seeks to produce a list of common attributes that describe a research data repository and to provide examples of the current approaches that different data repositories are taking to express and expose these attributes.[1]

The working group will produce two documentary outputs over the course of 18 months and four Research Data Alliance (RDA) plenary meetings; they are:

1. a list of common descriptive attributes of a data repository with
    1. a definition of each attribute,
    2. a rationale for the use and value of each attribute,
    3. the feasibility of its implementation,
    4. a gap analysis of its current availability from data repositories, and

2. a selection of examples that illustrate the approaches currently being taken by repositories to express and expose these attributes to users and user agents.

The list of descriptive attributes of a research data repository will be submitted for review and endorsement to become an RDA Recommendation, and the selection of exemplars will be submitted for consideration as an RDA Supporting Output. This work is planned to take place over 18 months between January 1, 2022 and June 30, 2023.

2. Value Proposition

A complete and current description of a research data repository is important to help a user discover a repository; to understand 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. Descriptive attributes may be expressed and exposed in different ways, making it difficult to compare repositories and to enable interoperability among repositories and other infrastructures such as registries. Incomplete and proprietary repository descriptions present challenges for stakeholders such as researchers, repository managers, repository developers, publishers, funders, and registries to enable the discovery and comparison of data repositories. For example:

- As a researcher, I would like to be able to generate a list of repositories to determine where I can deposit my data based on a query of descriptive 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 know how to express and serialize these attributes as structured metadata for reuse by users and user agents in a manner that is 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 harvest and index attributes of data repositories to help users find the best repository for their purpose.
- As a data re-user...

While this is not an exhaustive list of stakeholders and potential use cases, the value of identifying and harmonizing a list of descriptive attributes of data repositories and highlighting current approaches being taken by repositories would help the community address these important challenges and move towards developing a standard for the description and interoperability of information about data repositories. The statements of interest below demonstrate that there is a significant interest in this work.

3. Engagement With Existing Work in the Area

Many sets of attributes have been identified by different initiatives with differing scopes and motivations.[2] These attributes have included information about data repositories such as terms of deposit, subject classifications, geographic coverage, API and protocol support, funding models, governance, preservation services and policies, openness of the underlying infrastructure, adherence to relevant standards and certifications, and more. The results of these efforts reflect the variety of stakeholders and the diversity of repository attributes of interest across different communities. The harmonization of a common set of repository attributes, accompanied by the rationale for these attributes, will provide the community with a clearer understanding of the needs and requirements of different communities, and this commonality can enable greater interoperability across repositories, registries, and other data infrastructures.

4. Work Plan

The proposed co-chairs of the working group have submitted a Birds of a Feather (BoF) session proposal for RDA P18 to engage stakeholders in further discussion around these issues and revision of this case statement, if necessary. It will meet monthly via Zoom throughout the 18 months with a rotation of co-chairs formulating and sharing agendas in advance and leading each meeting. All meetings will be open to the community and progress towards the two deliverables will be noted on the RDA wiki. Correspondence between meetings will take place using an RDA mailing list that will be archived and accessible to all RDA members. The working group will strive to achieve consensus in its decision-making through open and respectful discussion. All views will be recorded from the deliberations of the group (e.g., mailing list archive, wiki) for consideration and community review of its outputs.

The working group will refine its methods based on feedback from the review of this case statement and member input, but in general we anticipate taking these actions:

1. Identify current standards and approaches to describing data repositories
2. Define use cases/user stories for utilizing metadata about data repositories
3. Draft a list of attributes and the rationale for their use/value
4. Conduct focus groups to validate/refine list
5. Perform an environmental scan to identify exemplars of different approaches
6. Submit a list of data repository attributes as an RDA Recommendation for broader community input, review, revision, and adoption
7. Submit a selection of exemplars as a RDA Supporting Output
8. Outreach to present and promote the adoption of the outputs

Milestones will include:

1. BoF session at RDA P18 and approval of case statement
2. Monthly meetings commencing January 2022

3. Identification of current descriptive approaches, use case definitions, and first draft of attributes (Action items 1-3) before RDA P19
4. Environmental scan to identify exemplars, completion and submission of list of attributes as RDA Recommendation three months after RDA P20
5. Revision of list based on community input and submission of exemplars as RDA Supporting Output by June 20, 2023
6. After conclusion of working group, presentation of outputs and report of early adoption at RDA P21

Minimally, the working group will engage the Metadata Interest Group (IG), Domain Repositories IG, and Repository Platforms for Research IG, and we will explore joint sessions as needed. Other interest and working groups as well as stakeholders from outside of the RDA will also be welcomed and encouraged to participate.

5. Adoption Plan

Primary adopters:

1. Repository managers
2. Repository software developers
3. Registries - e.g., FAIRsharing, OpenDOAR, re3data

Consultation and input from both the working group and the broader stakeholder community will be undertaken to identify

1. the relevance of the descriptive attributes drafted as the first output of the group, and
2. the feasibility of adopting these characteristics from the perspective of implementation.

Consultation on attribute relevance will allow an iterative process of development. Stakeholders will identify the important functionalities they need from repositories (such as facilitating the data peer review process for publishers and their authors, or integration with funder review processes); provide the rationale for why these characteristics are important for their community; and clearly articulate the aspects and functions needed to support their use cases.

This consultation will ensure that the final version of the list of descriptive attributes will both represent those repository attributes that are already in use as well as those which are of most relevance to our stakeholders, who will benefit directly from a harmonized, common list of attributes, and who will ultimately lead in their adoption and implementation.

6. Initial Membership

The working group will be led by co-chairs who represent international perspectives from a variety of stakeholders, including a variety of repositories, registries, publishers, and librarians.
Co-chairs:

- Matthew Cannon, Taylor & Francis (UK)
- Allyson Lister, FAIRsharing.org, University of Oxford (UK)
- Washington Segundo, Instituto Brasileiro de Informação em Ciência e Tecnologia (Brazil)
- Kathleen Shearer, Confederation of Open Access Repositories (Canada)
- Michael Witt, re3data, Purdue University (USA)
- Kazu Yamaji, National Informatics Institute (Japan)

---

[1] This effort was catalyzed by the FAIRsharing WG session on Repository Features Across Initiatives" at the 17th RDA Plenary Meeting.
[2] E.g., Metadata Schema for the Description of Research Data Repositories, Repository Features to Help Researchers: An invitation to a dialogue, Identifying ELIXIR Core Data Resources, Core Trust Seal, Science Europe, The TRUST Principles for digital repositories, COAR Community Framework for Good Practices in Repositories, NIH: Selecting a Repository for Data Resulting from NIH-Supported Research, OpenDOAR Repositories and Metadata Practices, DCAT

Review period start:

Friday, 3 December, 2021 to Monday, 3 January, 2022

Documents