Case Statement: Community-based catalogue of requirements for trustworthy Technical Repository Service Providers

Working Group

1. Charter

The Working Group will address a significant need in the research community relating to trustworthy data repositories and associated services worldwide: how to extend the umbrella of trust, as exemplified by formal certification authorities\(^1\), to a wider group. The WG will focus on three aspects of extension, reflecting the needs and expectations of the community:

- Many generalist repositories cannot currently be certified by any of the existing authorities - with typical impediments often involving the absence of specialised curation and appraisal capabilities, lack of long-term preservation services, and repository sustainability. Some repositories will adequately provide some elements of what can be expected of a trustworthy repository, and it is in the interest of the community to be informed about the status of the repositories for these elements (e.g., criteria-level modular assessment or certification).
- Some elements of service - for example bit-level and format preservation - can be outsourced, and there are benefits to this - economy of scale being an obvious one. Repositories seeking certification at present, should they make use of such services, have to certify services individually for each application event. It is more efficient to certify the service and for repositories using such a service to reference such a certification.
- The level of performance expected by certification authorities sets the bar at a high level of maturity, and this results in an inability to recognise and encourage repositories that are in a process of improving performance. The community has no information on the status of repositories’ work towards these criteria since the certification processes are binary and encompassing (i.e., all included criteria must be met).
- Certification of repositories currently addresses only datasets, while there is a clear need to provide trustworthy repository services for other research outputs (e.g., code, semantic artefacts).
- Finally, the cost and burden associated with the certification process is a deterrent in many cases.

To address these concerns, this working group intends to:

\(^1\) ISO 16363, nestor, and CoreTrustSeal
• Establish a set of modular repository features and characteristics, confirming and assessing the needs and expectations of the community with respect to the trustworthy repository ecosystem.
• Determine which of these features can be delegated to data repository service providers.
• Define and describe alternatives to formal certification that improve the appraisal and selection of both repositories that are not formally certified, as well as external services contributing to the repository features expected by the community.

In doing so, we hope to contribute to the following outcomes:

• An improvement in the number and scale of successful certifications.
• Extend the scope of certification beyond data repositories.
• Help the community make better-informed decisions about which infrastructures to use while the above is under way.

The main outputs expected for the working group are the following, based on three sets of community consultations:

1. Community consultation 1: “Inventory of Expectations”
   a. Through community consultation and a scoping review, compile an inventory of performance expectations for repositories and related services, and map this inventory to the portfolio of existing criteria applied by the repository certification authorities.
   b. Based on the results, develop a conceptual model of the actors and service providers in the ecosystem.
2. Community consultation 2: “Prioritisation”
   a. Use the generated mapping developed earlier to:
      i. verify importance of individual features to the community by way of surveys,
      ii. identify gaps between expectations and current certification options,
      iii. determine the criteria whereby an appropriate repository or service will be selected based on the context of the end user (scientific discipline, institutional affiliation, country, data formats, etc.).
   a. Determine which of the selected criteria should apply to generalist repositories, to services and other actors, and what level of performance can be expected of different actors in the ecosystem.
   b. Define the ways in which the level of performance can be verified, including but not limited to formal certification.
   c. Identify metadata that can be associated with the implementation of each criterion to facilitate modular, decentralised assessment and certification.

---

2 The current portfolio of criteria for all authorities are based on OAIS-RM. There may be additional criteria that are important - e.g., expected by CARE, privacy concerns, and reproducibility - that are also worth evaluating in the same process.
The working group recognises similar needs associated with other research outputs (e.g., code, semantic artefacts, etc.). These outputs need to be considered when developing recommendations - so as to maximise the reusability of criteria and levels of performance for services and repositories that manage mixed research outputs, or focus on these other outputs.

The working group will publish modular trustworthiness performance expectations for data repositories and related services as a set of recommendations, together with supplementary materials derived from the scoping review and community consultation.

2. Value Proposition

Extension of the scale of certification will become feasible via this working group, improving the current situation where less than 5% of registered data repositories are certified, and no technical service providers are. We will achieve this through the following:

1. Efficiency of certification will improve, since (1) component services can be certified once, and that certification can be referenced by potentially many repositories in addition to (2) modular, decentralised certification based on repository metadata may be possible for some criteria;
2. Modular certification allows communities flexibility in identifying the types of criteria that are most important for building trust in repositories within given communities without assuming that all criteria are weighted equally across communities;
3. Appropriate criteria that can be made applicable to non-data repositories can accelerate the movement towards assessment and future certification of such repositories;
4. The WG will work towards community consensus on selecting the most appropriate repository or service for a given context, which will promote the investment and (re)use of the highest-performance service that meets user needs within that context.

As a result of these outputs, it will be possible to present end users with more information on a wider choice of target repositories for their context, since it will become feasible to provide information that reflects partial compliance with full certification. In many cases this will be adequate, providing data depositors and data requestors with additional information and guidance to identify the most appropriate repository for their context.

Repository funders, journals, and researchers interested in depositing or reusing data will benefit from mechanisms that publicise performance assessment at the criteria level even though certification is considered or in process. Metadata-driven criteria-level modular performance assessments can also

---

3 For example, in FAIRSharing, re3data, and COAR.
4 This consideration requires careful balancing: on the one hand, it is clearly beneficial to know more about non-certified repositories and services, but one also has to retain the driving force towards formal certification.
reduce the cost of certification and the time between repository actions and certification of those actions.

The broad community engagement envisaged by the Working Group will serve as valuable inputs to the successful bidders in the relevant Horizon Europe projects, commencing in 2025, and will provide useful input for Certification Authorities.

3. Engagement With Existing Work in the Area

This group will build on related work done in multiple contexts. Specifically, we will consider the work of several groups who have developed schema for assessing “trust” in repositories specifically and in scholarly infrastructure in general such as CoreTrustSeal, Nestor, POSI, and ISO. We will create a side-by-side comparison of these criteria that will be further extended through landscape analysis and community consultation.

A practical starting point for criteria to consider is CoreTrustSeal, given its status and use in the data repository community. CoreTrustSeal provides a core set of criteria that define the base requirements for data repositories providing long-term preservation to be considered trustworthy. The requirements were first developed within the RDA Repository Audit and Certification Data Seal of Approval – World Data System Partnership WG, now the CoreTrustSeal Maintenance WG. CoreTrustSeal as an entity began in 2017, and the CoreTrustSeal certification continues to be a World Data System (WDS) membership requirement for data repositories.

There are recurring community concerns around the importance, implementation, and scope of the CoreTrustSeal certification, and CoreTrustSeal routinely accumulates and incorporates community feedback. A consistent topic is the inclusion and recognition of repository service providers, and how they should be considered when a data repository applies for CoreTrustSeal certification while outsourcing some subset of its services. While the CoreTrustSeal criteria may be applied to data repository service providers broadly, CoreTrustSeal certification is restricted to data repositories that provide active data preservation and therefore most service providers (and some data repositories who use their services) are not eligible for CoreTrustSeal as they are distinctly out of scope. It is therefore of interest to the data community and CoreTrustSeal to address the gap of how to define a trusted repository service provider, so that more data repositories who are trustworthy can be appropriately recognized as such and to accelerate the process of data repository certification. This WG

5 https://www.coretrustseal.org/
6 https://www.langzeitarchivierung.de/Webs/nestor/EN/Standardisierung/standardisierung.html
7 https://openscholarlyinfrastructure.org/
8 https://www.iso.org/standard/87472.html
9 https://rd-alliance.org/groups/repository-audit-and-certification-dsa%E2%80%93wds-partnership-wg.html
10 see https://www.coretrustseal.org/why-certification/meeting-community-needs/
12 https://doi.org/10.5281/zenodo.4568875
13 https://doi.org/10.5281/zenodo.8083359
will actively collaborate with CoreTrustSeal towards this goal, and similarly with other certification bodies in this landscape.

Within the RDA community, there are multiple working groups that are engaged in complimentary work. The RDA/WDS Certification of Digital Repositories IG\(^\text{14}\) has a direct connection, given their active engagement with CoreTrustSeal as well as emerging assessment criteria and considerations for data repositories, such as the TRUST principles\(^\text{15}\). In the same vein, engagement with the RDA/WDS Trust Principles Outreach and Adoption WG\(^\text{16}\) would provide a base to build a community of data repositories and service providers that may be interested in being a part of the process for developing trusted repository service provider criteria.

Work initiated by the RDA/WDS Certification of Digital Repositories IG and adopted by the FAIRCORE4EOSC project as a basis for the consistent encoding of assessment and certification efforts will be of interest, both as a means of standardising the representation of best practices, criteria, metrics, and benchmarks, and as a case study for further refinement and validation of the work done in FAIRCORE4EOSC\(^\text{17}\).

A broader set of community expectations and best practices in respect of research infrastructure is expressed and summarised in the Global Open Research Commons-related work, initiated by the RDA GORC IG. These will serve as a framework for features and characteristics of trustworthy repository infrastructure\(^\text{18}\).

The RDA FAIRsharing WG\(^\text{19}\) will provide valuable integration with a wealth of repository metadata across all research areas; this information can be used within landscape analyses as well as possible future integration of the certification attributes produced by this WG into human-readable and machine-actionable form via curation of FAIRsharing repository records. The RDA Data Repository Attributes WG\(^\text{20}\) and the Global Open Research Commons International Model WG\(^\text{21}\) also address criteria or considerations for data repositories to improve their usability and function, which are directly linked to repository trustworthiness. Additionally, the WG will work with the RDA-OfR Mapping the Landscape of Digital Tools WG\(^\text{22}\), where the distinction between tools providers and service providers is not distinct. Collaboration with these groups to ensure minimal overlap will be paramount.

\(^{14}\) https://www.rd-alliance.org/groups/rdawds-certification-digital-repositories-ig.html
\(^{16}\) https://www.rd-alliance.org/groups/rdawds-trust-principles-outreach-and-adoption-working-group
\(^{20}\) https://www.rd-alliance.org/groups/data-repository-attributes-wg
\(^{21}\) https://www.rd-alliance.org/groups/gorc-international-model-wg
\(^{22}\) https://www.rd-alliance.org/groups/rda-ofr-mapping-landscape-digital-research-tools-wg
There are several RDA outputs and recommendations that will be reviewed for relevant criteria as part of our landscape review, and an initial library of material to review can be found on Zotero\(^{23}\). Depending on the result of the review, collaborations with additional RDA IGs and WGs will occur. This WG will be cautious to consider criteria at the repository and service level and not at the dataset level, which is out of scope.

The intermediate and final outputs of this working group will contribute to the HORIZON-INFRA-2024-EOSC-01-03\(^{24}\) and HORIZON-INFRA-2024-EOSC-01-04\(^{25}\) projects, among others. Specifically, the landscape analysis and comparison of existing certification and assessment criteria will directly contribute to identifying common and desired “practises, standards and tools for long-term preservation” globally, which can then be tailored to the needs of EOSC as required by HORIZON-INFRA-2024-EOSC-01-04. Moreover, the intended deliverable of this working group to identify and vet a set of criteria for trustworthy repository service providers would directly enable the identification and classification of trustworthy repositories who are currently not eligible or unable to acquire trustworthiness status, thus supporting the creation of “[a] European network of trustworthy repositories” and contributing to “[t]he concept and requirements of trustworthy repositories are harmonised”, both outcomes of HORIZON-INFRA-2024-EOSC-01-03.

The EOSC Association Task Force on Long-Term Data Preservation (LTDP-TF) has produced recommendations to be taken into account, and will continue for an additional period from mid-2024. The working group will engage the task force directly to share information and effort where possible.

Additionally, this WG will work with national and pan-national service providers themselves, such as ARCHIVER\(^{26}\) and 4Science\(^{27}\) in the EU, to identify characteristics of repository service providers and how trustworthiness may be defined.

Many research infrastructures, such as ELIXIR and CLARIN, have developed criteria over time that either extend or replace formal criteria published by e.g. CoretrustSeal, and these have to be considered too.

### 4. UN Sustainable Development Goals (SDGs)

The underlying goal that is supported by this WG is providing evidence over time for policy and decision-making. Long-term preservation by trusted repositories and service providers will ensure the continued availability of data for reference from policy-making today and for reliable use in future policy making. Therefore, the outputs of this group will contribute indirectly to all of the UN SDGs by potentially further

\(^{23}\) [https://www.zotero.org/groups/5395078/rda_trsp_requirements_wg/library](https://www.zotero.org/groups/5395078/rda_trsp_requirements_wg/library)


\(^{26}\) [https://archiver-project.eu/](https://archiver-project.eu/)

\(^{27}\) [https://4science.com/](https://4science.com/)
enabling research, its storage, accessibility, and (re)use addressing these very goals as well as the policy and decision making that happens based on that research, now and in the future. More direct contributions can be identified for:

- “Peace, Justice and Strong Institutions”, where the certification of repository service providers will enable effective and efficient certification processes for data repositories, which in turn enables effective and efficient research workflows for researchers, funders, and other stakeholders;
- "Partnerships for the Goals", where the proposed criteria would enable increased visibility of trusted repository services, their use, and integration in the research ecosystem, enhancing the overall efficiency and effectiveness of repositories and affiliated research infrastructure.

5. Work Plan

The proposed Working Group is one of the Pilot Working Groups of the RDA TIGER project. The project aims to provide services to facilitate and support well-defined Working Groups between key European and international initiatives, resulting in concrete alignment, harmonisation, and standardisation of Open Science developments and technologies globally. The project’s Pilot WGs demonstrate and test the services provided through the project. As such, the proposed Group will benefit from a host of services the project provides, including facilitation, landscape analysis services, and communications support.

The members of the proposed Working Group had a preliminary meeting to discuss the overall aims, scope, and (potential) WG members and stakeholders of the Group and to formulate an application for a BoF session at the P21. At the P21 BoF session, the group engaged stakeholders, further discussed the aims of the Group, identified co-chairs, and revised this case statement before submission. A mailing list, and process for scheduling future meetings was established to engage interested community members. Co-chairs were identified in late 2023. Co-chair meetings occurred in December 2023, January 2024, and February 2024 to finalise the case statement from the feedback provided at RDA 21. In addition, feedback from a session at RDA 22 has now been incorporated into the case statement, and in terms of the scope of work, it is important to note the narrowing of focus through the phases of work.

Initial work to identify and inventorise the desirable characteristics of repositories and services will of necessity cast the net wide, but subsequent phases (prioritisation, selection, and categorisation of characteristics in respect of domain, object type, or service type will of necessity focus more. The final set of recommendations will be focused on TRSPs, while providing proposals in respect of other service and object types.
In parallel, once an inventory of desirable characteristics has been selected and prioritised, effort will be put into community consultation to identify mechanisms for appraisal of these, with a preliminary indication of whether and how levels of performance, benchmarking, or other types of contextualisation apply to each characteristic.

**Meeting schedule**

Once the Case Statement is accepted and the WG endorsed by RDA TAB and Council, virtual monthly meetings will be held to provide updates on WG tasks, with monthly agendas drawn up by co-chairs. The administrative aspects of the meetings will be organised by the RDA TIGER Facilitator, who will ensure that the meeting invites, zoom platform links, and agenda documents are circulated in sufficient time for the meeting. The facilitator will also follow up on any other tasks as necessary, including arranging for all group members to provide input to group activities (asynchronously if needed) to ensure consensus can be reached with all group decision-making.

**Milestones**

Below are the milestones the group will meet to produce its Recommendation within the 18-month WG lifespan.

- **Month 0-3**: Community consultation to identify types of data services that require assessment from the community and performance expectations for them. Desk research on current service providers and their context as well as current assessment models such as CoreTrustSeal, Nestor, ISO, and POSI. Two task groups are expected to be formed to tackle the respective desk research topics, and the co-chairs will lead the work.
• Month 4-7: A speaker series or group discussion space will be used to gather further information about service providers and performance expectations. We will merge results into an initial set of assessments, including mapping criteria between certification bodies.
  ○ D1: Inventory of performance expectations for repositories and related services, and map these to the portfolio of existing criteria applied by the certification authorities to verify importance to the community, and to identify gaps.
  ○ D2: Based on community input and desktop study, develop a conceptual model of the actors and service providers in the ecosystem.
  ○ D3: Within the same community consultation process, determine the criteria whereby an appropriate repository or service will be selected, based on the context of the end user (scientific discipline, institutional affiliation, country, data formats, etc.).
• Month 8-11: Test with experts from the community and the speaker series, and create a mechanism for benchmarking of performance.
  ○ D4: Through a process of community engagement, determine which of the selected criteria should apply to generalist repositories, to services, and what level of performance can be expected of different actors in the ecosystem.
  ○ D5: Determine, through stakeholder consultation, priorities for criteria implementation and identify metadata that can be associated with the implementation of each criterion to facilitate modular, decentralised certification.
• Month 12-15: Development of a testbed for peer review and certification based on the defined requirements and procedure. The testbed will provide practical insight, thus enabling iterative improvements to the requirements and procedure. The testbed will be driven by the CoreTrustSeal. A pool of reviewers will be set up to test the developed procedure.
  ○ Final versions of D1, D2, D3, D4 and D5
• Month 16-18: Outreach and communication to relevant actors to solicit input and promote adoption of the outputs

The Working Group will produce a catalogue of requirements, providing a list of criteria to assess trustworthy TRSPs. This catalogue will ideally be web-based, interactive, and machine-actionable. It may also be available in a descriptive document to provide multiple formats.

6. Adoption Plan

There are multiple organisations that have already expressed an interest in contributing to and/or using the outputs of the working group, including CoreTrustSeal and ELIXIR. The outputs will serve as landscape analysis materials for two EU-funded projects for which proposal evaluation is underway, and awards are expected towards the end of 2024. The working group will actively engage with the successful consortia in these two funded projects.
7. Initial Membership

Co-chairs:

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wim Hugo</td>
<td>DANS</td>
</tr>
<tr>
<td>Allyson Lister</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>Lauren Maxwell</td>
<td>University of Heidelberg</td>
</tr>
<tr>
<td>Mohan Ramamurthy</td>
<td>UCAR/Unidata</td>
</tr>
<tr>
<td>Max Wilkinson</td>
<td>ARDC</td>
</tr>
</tbody>
</table>

Members:

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Francoise Genova</td>
<td>Observatoire Astronomique de Strasbourg</td>
</tr>
<tr>
<td>Mark Leggott</td>
<td>Digital Research Alliance of Canada</td>
</tr>
<tr>
<td>Dimitri Szabo</td>
<td>INRAE</td>
</tr>
<tr>
<td>Mari Kleemola</td>
<td>Finnish Social Science Data Archive, Tampere University and CESSDA</td>
</tr>
<tr>
<td>Kyle Snyder</td>
<td>SURF</td>
</tr>
<tr>
<td>Ylva Berglund Prytz</td>
<td>Swedish Univ. of Agricultural Sciences (SLU)</td>
</tr>
<tr>
<td>CJ Woodford</td>
<td>World Data System International Technology Office</td>
</tr>
<tr>
<td>Meredith Goins</td>
<td>World Data System</td>
</tr>
<tr>
<td>Hervé L'Hours</td>
<td>UKDA</td>
</tr>
<tr>
<td>Roxanne Wyns</td>
<td>Catholic University of Leuven</td>
</tr>
<tr>
<td>Gilles Ohanessian</td>
<td></td>
</tr>
<tr>
<td>Dieuwertje Bloemen</td>
<td>Catholic University of Leuven</td>
</tr>
<tr>
<td>Liise Lehtsalu</td>
<td>RDA Europe</td>
</tr>
<tr>
<td>Name</td>
<td>Institution</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>Athina Papadopoulou</td>
<td>RDA Europe</td>
</tr>
<tr>
<td>Mark van de Sanden</td>
<td>SURF</td>
</tr>
<tr>
<td>Lesley Wyborn</td>
<td></td>
</tr>
<tr>
<td>Gonzalo Torderolo Silva</td>
<td></td>
</tr>
<tr>
<td>Susanna-Assunta Sansone</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>Rebecca Koskela</td>
<td></td>
</tr>
<tr>
<td>Isabelle PERSEIL</td>
<td></td>
</tr>
<tr>
<td>Didi Lamers</td>
<td></td>
</tr>
</tbody>
</table>