CoreTrustSeal: Specialists, Generalists, and Technical Repository Service Providers

A request for community feedback

Communications with a number of stakeholders have presented the CoreTrustSeal Board with a challenge to continue to meet the needs of the domain/subject-based repository community while addressing the demand for assessment, peer review and recognition from a wider group of actors delivering data curation, storage, and access services. The CoreTrustSeal has received input and requests from a number of more generalist repositories and from those providing technical infrastructure and services to repositories.

This paper is the first step in a number of community engagements and requests for comment which will feed into the formal review of CoreTrustSeal in 2022. We are also consulting on some proposed immediate actions to respond to demand from the CoreTrustSeal community.

Our first step is to respond to demand for certification from repositories beyond the domain/subject-based curators which formed the majority of previous applicants. We need to recognise the need for domain/subject-based curation as recommended by policy makers including Science Europe[[1]](#footnote-0) while acknowledging that resources do not permit all data to be curated to this level. Generalist repositories provide essential services for governmental, cultural and scientific data including critical contributions to data management from galleries, archives, libraries and museums (GLAM).

We propose to define each successful CoreTrustSeal applicant as either representing a specialist (e.g. domain or subject-based) or a generalist repository. Between now and the next revision of the CoreTrustSeal requirements we will consult on whether these two repository types/curation levels need to be more clearly differentiated through logos, badging or some other means. We will also cooperate with repository registries to ensure appropriate organisational metadata is made available to those seeking repository service providers.

Our second step is to engage with the providers of products and services which either directly provide technical systems for repositories or otherwise provide partnership and support. These actors have always been a critical part of the data ecosystem, but we need to directly address assessments which depend on increasingly co-dependant partnerships.

To address this we need to reach some community consensus on who can apply directly for the CoreTrustSeal as it stands, and who can seek recognition for providing compliant supporting systems and services.

**Applicant:** In current practice, CoreTrustSeal status is attached to an applicant with direct responsibility for the curation and preservation of their collection objects’ data and metadata. This is reflected in the following draft statement on ‘Who can apply’:

“An official representative acting on behalf of an organisation whose mission includes the curation and long-term preservation of a specified data collection, i.e. the provision of the appropriate infrastructure (documented policies, people and skills, workflows and technologies) to ensure the data are sufficiently preserved, remain accessible, and re-usable over time. To fulfill this mission, the organisation must have sufficient expertise and hold sufficient oversight and management rights over the data to respond to changes in the technologies and the knowledge base of a well-defined Designated Community (Glossary) of users. The applicant may outsource to third parties. Outsourcing roles and relationships should be well defined, and all parties must provide evidence related to all of the functions or processes they help undertake.”

The existing CoreTrustSeal Requirements and evidence expectations may be elaborated or extended to allow the assessment process to differentiate more clearly between:

* A specialist (i.e. domain/subject-based) repository serving a defined designated community with a clear knowledge base
* A generalist repository serving a broader community without defining a detailed knowledge base for different types of objects under curation/preservation
* A Technical Repository Service Provider which could support bit-level preservation for either of the above across the data lifecycle (for example, deposit, secure storage, access).

And more generally:

* Other Insource/Outsource partners supporting various repository functions/ activity areas with their services
* Software or other products used by repositories without a specific partnership relationship.

The differences in curation responsibility and in the expectations of services provided must be clear to all stakeholders: repositories, reviewers, depositors, users, and funders. These are aligned with the concepts of OAIS[[2]](#footnote-1) mandatory responsibilities:



**Diagram: a simplified overview of mandatory OAIS responsibilities**

**Specialist Repositories:** The key recommendation from many data policy makers is that the value of data assets is maximised for the long term if deposited in a domain or subject-based repository([Glossary](http://doi.org/10.5281/zenodo.3632563)). These repositories are the predominant membership of the CoreTrustSeal community. These repositories must ensure that their stated area of expertise is evidenced in terms of meeting specialist (e.g. domain, disciplinary) standards as required by their designated community. They must have the skills and the processes to support data, depositors and end users from that community. They can be multi-specialist (e.g. multi-disciplinary) but this is not the same as ‘discipline agnostic’. A specialist repository may also offer generalist repository services. Subject-based repository requirements are a superset of those applied to generalist repositories.

**Generalist Repositories**: Generalist Repositories with a potentially heterogeneous collection and a non-specialist designated community provide a critical curation and preservation role for a vast proportion of data assets. Generalists can claim a broader (including public) designated community and can therefore apply more general restrictions on formats, metadata standards etc. Generalist repositories may not be expected to provide the kind of granular metadata, discovery or support that a specialist would. They are expected, however, to have considered and defined the knowledge base of their designated community (e.g. language and computer skills).

**Technical Repository Service Providers (TRSP)**: Software providers and providers of technical infrastructure and associated services which support trustworthy digital repositories are vital components of the data ecosystem. These tools and services do not take direct responsibility for the selection, curation, appraisal and access condition of the data they hold on a temporary or permanent basis. Like any insource/ outsource provider, the Technical Repository Service Providerswould need to offer evidence for the functions/ activity areas which they support. We’ll use the informal TRSP abbreviation for brevity but we acknowledge that this may represent a wide range of possible.

Each of the statements above, and each statement and question covered under the 16 CoreTrustSeal Requirements below is open to community comment and debate. Feedback will guide both the immediate actions and future recommendations for the requirements and assessment process. For the current version of the CoreTrustSeal TDR Requirements, please refer to <http://doi.org/10.5281/zenodo.3638211>.

## **R0. Context**

### **R0. Repository type**

“Repository Type. Select all relevant types from:

● Domain or subject-based repository

● Institutional repository

● National repository system, including governmental

● Publication repository

● Library

● Museum

● Archive

● Research project repository

● Other (Please describe)”

An ‘institutional repository’, for example, could be a specialist or a generalist repository, or a self-deposit TRSP.

How should we expand or amend this list to differentiate specialist and generalist repositories?

### R0. Brief Description of Repository

Do we need to ask more specific questions about the digital objects and collections being curated under context?

If so, which object or collection characteristics are most relevant to the assessment?

### R0. Designated Community

A generalist repository may serve a broader designated community of users. A specialist repository must define the knowledge base of their community more specifically. All applicants must demonstrate clear interaction with and responsiveness to their primary designated communities. In defining the differences between a repository and a TRSP the designated community is key.

How can we best define and communicate the difference between designated communities for generalist and specialist repository providers?

Is there a minimum level of evidence repositories should provide to support their designated community statements?

### R0. Level of Curation

A. Content distributed as deposited

B. Basic curation – e.g., brief checking, addition of basic metadata or documentation

C. Enhanced curation – e.g., conversion to new formats, enhancement of documentation

D. Data-level curation – as in C above, but with additional editing of deposited data for accuracy

A TRSP might only apply level A curation. Other applicants would need to provide a greater degree of curation for some or all of their collection.

Do these levels meet the needs of both specialist and generalist repositories?

## R0. Context. Insource/Outsource Partners.

If a specialist or generalist repository shares some part of their work with a third party, including a TRSP, how should we best identify where the partner takes some responsibility?

E.g. Is addressing this per requirement sufficient, or should this section provide a mapping by function, or activity areas, or by some other means?

**Function**: e.g. deposit (R8), curation (R11), storage (R9), preservation (R10) , discovery (R13) access/use (R14).

**Activity Area**: e.g. rights management (R2), continuity of access (R3), integrity checks and authenticity measure (R7), workflow management (R12), identifier management (R13), technical infrastructure management (R15), security management (R16).

## R1. Mission

An applicant must have a mission which includes the long term provision of access to, including active preservation of data and metadata.

Should the mission statements of specialist and generalist repositories be expected to differ in any way?

Is there some value in seeking mission information from a TRSP?

## R2. Licenses

The applicant must be responsible for assuring that the licences agreed at appraisal (R8) define the level of curation quality and preservation of the data, and the permissions, prohibitions and duties applicable to data users. Technical implementation may be outsourced.

Should the evidence for licence management differ in any way between generalist and domain repositories?

## R3. Continuity of Access

Continuity of access includes the applicant's ability to ensure business continuity, disaster recovery and succession planning. Specialist repositories may face additional challenges in ensuring full continuity of more expert driven aspects of their services.

Dependencies on TRSP must form part of any continuity risk management. Additional risks might include the potential outcomes of non-payment to a third party service provider, or the lack of a data portability solution if a provider ceases to offer the service.

Should evaluation of this requirement be different for specialist and generalist repositories?

What additional evidence should an applicant provide on behalf of a TRSP that they rely upon? Is a contract or memorandum of understanding sufficient, or do we need to know more about the service providers own approach to risk?

## R4. Confidentiality and Ethics

The applicant must always take responsibility for addressing these issues.

Should a TRSP be expected to provide a confidentiality/ethics statement?

Should the expectations around confidentiality and ethics differ between generalist and specialist repositories?

## R5. Organizational Infrastructure

Understanding the organisation structure, governance and the availability of appropriately skilled resources is invaluable in the assessment process. Specialist repositories are expected to demonstrate the skills applicable to their stated domain/discipline/subject.

Many of the issues under organisational infrastructure will also apply to a TRSP.

Is there additional evidence that should be provided if an applicant relies on aTRSP?

## R6. Expert Guidance

The expert guidance to be demonstrated under requirement 6 is mainly focussed on the applicant seeking appropriate external scientific and technical guidance. This would apply equally to a specialist or generalist repository in line with the curation levels they offer.

A TRSP may provide some aspects of this expert guidance and could therefore provide supporting evidence.

What evidence for expert guidance would be appropriate for a TRSP?

## R7. Integrity and Authenticity

An actor in any part of the data lifecycle with storage (see also R9) responsibility for data should be able to demonstrate integrity measures that ensure unintended change is monitored and mitigated. It would be beneficial for the community of practice to identify minimum and ideal levels of data integrity assurance that could be applied to a specialist, a generalist or a TRSP.

Increasing the level of pre-deposit authenticity information to maximise data provenance is generally desirable.

For the repository phase of the lifecycle assuring data integrity may become more challenging for higher curation levels i.e. the more intended changes that may be applied to data and metadata. Higher levels of curation also imply a need for more detailed records of change to support authenticity.

Though the ideal level of provenance may depend on the demands of the designated community it would be beneficial for the community to reach consensus on minimal provenance expectations.

A CoreTrustSeal applicant must retain responsibility for the integrity and provenance measures applied to data under curation. What evidence can a TRSP provide to demonstrate the application of these measures?

## R8. Appraisal

Demonstrating a clear collection development plan, including an appraisal process that supports the definition of clear curation and preservation actions must remain the responsibility of the applicant. A TRSP may offer evidence that it supports these measures.

## R9. Documented Storage Procedures

These documented storage procedures are addressed from the perspective of active curation and preservation measures, in contrast to the wider application of a technical infrastructure covered under R16.

Any specialist or generalist repository must document any storage procedures where they have control of the storage. Where they rely on aTRSP to provide storage the provider may provide supporting evidence. The detail required for procedures is likely to increase for higher levels of curation.

## R10. Preservation Plan

The applicant must take responsibility for an active preservation plan covering data and metadata in line with the expectations of the designated community. A generalist repository may accept a wider range of formats and curate for less detailed metadata than a specialist repository. Preservation actions may be outsourced to a TRSP, preservation plans may not. Bit-level “preservation” is important, but not sufficient.

The CoreTrustSeal acknowledges that recommendations usually focus on the need to deposit in a domain/subject-based repository. We support these recommendations and recognise that specialist data in non-specialist repositories may receive a less specialised level of curation. However, it is not the role of the CoreTrustSeal to make these judgements during the application process.

## R11. Data quality

The “scientific” quality of data is for funders, depositors and users to evaluate. A trustworthy digital repository ensures the technical quality and understandability of data through curation actions which ensure standards compliance. These actions start to be defined at appraisal (R8), are implemented at R11 and communicated at the point of data re-use (R14). The applicant may outsource actions to ensure data and metadata quality, but must take responsibility for defining those actions. A TRSP may provide evidence that actions have been implemented.

Can the community reach consensus on the minimal practice for data and metadata curation and preservation (for appraisal, curation quality and re-use) that is acceptable from a generalist repository? Is this a key differentiator between a generalist repository and a TRSP?

## R12. Workflows

All Specialist, generalist and TRSP must document the workflows that they undertake. Overall responsibility for ensuring that workflows are fit for the mission and meet the needs of the designated community remain with the applicant.

## R13. Data discovery and identification

Standards for discovery and measures for ensuring persistence must be the responsibility of the applicant. All repositories must ensure persistent identification and enable general searching locally or through third parties. In addition specialist repositories must address the more specific resource discovery needs of their designated communities.

A TRSP can provide supporting evidence when these actions are taken using their systems.

## R14. Data reuse

The level of contextual metadata required to enable access and re-use may be lower for a generalist than a specialist repository. But it is the applicant who must identify the needs of their designated community and demonstrate that they are met. Interactions with users including collection of usage statistics and feedback may be mediated by a TRSP.

## R15. Technical infrastructure

Provision of technical infrastructure provides the clearest link between specialist and generalist repositories and their TRSP. This is also the area where all repositories and their TRSP can cooperate to agree minimal and ideal practices.

As outsourcing elements of technical infrastructure becomes more common we will depend more on providers for evidence. But the applicant remains responsible for defining the expectations of the technical infrastructure and the relationship with any outsource partners.

How can we define the expectations from technical infrastructure in a way which clarifies the relationship between repositories and service providers and enables comparison across repositories?

## R16. Security

With Security, and to a lesser extent for technical infrastructure (R15) the limitations of a ‘core’ assessment with no site visit become clear. But more important may be the lack of clearly defined minimal and ideal practices to address different tiers of community risk (from open access, to user data, so sensitive personal datasets).

The focus of a CoreTrustSeal review of R16 for Security is to identify whether the repository demonstrates a clear understanding of the level of security risk presented by their collections, and clear responses to those risks. This expectation would be consistent across specialist and generalist repositories.

How can the CoreTrustSeal best support the community definition of clearer security expectations for data, including sensitive data?

## Final Comments

Thank you for your time. Please provide any additional comments here and return to info@coretrustseal.org.

1. <https://www.scienceeurope.org/media/jezkhnoo/se_rdm_practical_guide_final.pdf> [↑](#footnote-ref-0)
2. <https://public.ccsds.org/pubs/650x0m2.pdf> [↑](#footnote-ref-1)