Specialists, Generalists, and Technical Repository Service Providers

An Update on the CoreTrustSeal Position Paper

Webinar, 7 October 2020, 13:00 UTC

Rorie Edmunds, World Data System-International Programme Office / Head of CoreTrustSeal Secretariat

Hervé L’Hours, Vice-Chair CoreTrustSeal Board. UK Data Service, UK Data Archive, University of Essex.
Overview

1. Introduction
2. The Consultation on Specialists, Generalists, and TRSPs
   a. Why this process?
   b. Current scope of CoreTrustSeal
   c. Defining the entities
   d. Process up to now
3. Overview of Responses
4. Remaining Challenges and Questions
5. Questions and discussion
Introduction
Introduction

- CoreTrustSeal is a community-based nonprofit organization that promotes sustainable and trustworthy data infrastructures by offering professional certification tools and services for data repositories and preservation-focussed institutions around the globe.

- Because CoreTrustSeal is a community certification, it is the Board’s duty to constantly gain consensus from the community to decide what is at the core-level and must be included in the certification.
Introduction

• Traditionally **domain data repositories** have sought certification, and formed the majority of applicants for the WDS and Data Seal of Approval standards, as well as early applicants for CoreTrustSeal.

• The **diversity of organizations** applying for certification has been **increasing**, including by infrastructure providers, repository software providers, bit-level replication services, and national archives, and commercial services.

• CoreTrustSeal wishes to help **provide certification tools** to such applicants, and is **examining the characteristics** of organizations to understand how we can **best serve** the community.
Introduction

• Notion of a Trustworthy Data Repository (TDR) traces back to the OAIS Reference Model, and refers to the responsibilities to a Designated Community of users. Although certification standards have focussed on specialist repositories, the OAIS Reference Model is not restricted to them.

• Similarly, the principal target of the CoreTrustSeal certification has been domain data repositories (i.e., ‘specialist repositories’), but it does not certify them as being a TDR for a specific discipline(s).
Introduction

Data Producer

Negotiate/Accept Information

Repository Team

Documented Policies & Procedures

Sufficient Control for Long-term Preservation

Define Designated Community

Make Independently Understandable

Make Available/Assure Authenticity

Designated Community
The Consultation on Specialists, Generalists, and TRSPs
Why is this Process Important?

• Demand for definitions of what is a ‘specialist’ repository, what is a ‘generalist’ repository, and what are the differences between them. In particular, there is need for a definition of a specialist Trustworthy Data Repository (TDR).

• Need to ensure all entities in the data ecosystem designed to preserve and protect the world’s digital legacy are appropriately recognized as being trustworthy.

• Digital preservation is carried out in increasingly complex partnerships; functions and responsibilities are distributed and shared among different organizations.
CoreTrustSeal’s Current Scope

Who Should Apply?

• Any organization with a mission that includes the *curation and long-term preservation* of a specified data collection is in scope for CoreTrustSeal certification. The organization must:
  1. Have *expertise in the domains* from which the data originate and the types of data deposited.
  2. Hold sufficient *oversight and management rights* over the data holdings to be able to take steps necessary to respond to changes in (a) technologies, and (b) the knowledge base of its (well-defined) Designated Community.

• The organization, as a data steward, may *outsource* some repository functions to third parties. However, such outsourcing roles and relationships should be *clearly defined*, and all parties must *provide evidence* related to all of the functions or processes they help undertake.
CoreTrustSeal’s Current Scope

What is Assessed?

- CoreTrustSeal does not evaluate data quality, but operational quality. In practical terms, the CoreTrustSeal Trustworthy Data Repositories Requirements examine an organization's ability to provide appropriate infrastructure—documented policies, skills, workflows, and technologies—that ensure its data holdings are sufficiently preserved, and remain accessible and reusable over time.
Defining the Entities for Certification

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. 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.
Defining the Entities for Certification

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).
Defining the Entities for Certification

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 Providers would **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 possibilities.

- These definitions will be reviewed and revised based on the feedback received.
The Process and Next Steps

• The large demand for certification from generalist repositories must be met in a clear and standardized way to ensure trustworthiness in a complicated and evolving landscape.

• Following an increasing number of applications outside of the scope for CoreTrustSeal certification, in late 2018, a Working Group (WG) was created to look at certification of generalist repositories and TRSPs. Representatives from interested parties were invited to be part of this WG, which then had several teleconference and developed a discussion document.

• To make significant progress an initial CoreTrustSeal viewpoint was needed to stimulate discussion. The Board had two brainstorming sessions in late 2019, resulting in the position paper released to the community in June 2020, and with a four-week request for feedback.
The Process and Next Steps

• The position paper focusses on the certification of generalist repositories in comparison to the specialist curators that formed the majority of previous CoreTrustSeal applicants. It also starts to look at the definition of TRSPs, the categorization of which is complex due to their and which needs to be broken down.

• CoreTrustSeal might be in a position to support all TRSP types, but clarity is needed before we can move on to developing a robust methodology for their certification.

• The feedback seems to show a consensus that service providers might be usefully certified, while software providers would not. However, the Board is considering more broadly how all external TRSPs can support applicants by providing evidence about what they offer.
The Process and Next Steps

• In early August, all responses were consolidated and categorized according to feedback type and what it referred to. The Board was then requested to reviewed the feedback, with an eye to
  • Revising the document according to community consensus
  • Identify open questions/challenges that require more thought by CoreTrustSeal, by the community, or by both.

• Now in the process of developing an overall response to the feedback, and a **second version of the paper** will be released incorporating the community consensus. **Individual responses** will be given to specific feedback concerning an organization/situation.

• After another round of consultation with the community, it is expected that a **final version** of the position paper can be released. Under discussion is to also publish the paper as a **journal article** with some additional explanation.
Overview of Responses
Statistics

- 37 different individuals/groups/organizations responded
- 552 individual comments/pieces of feedback
Responses suggest no need to revise requirements to differentiate between generalist and specialist repositories or TRSPs

- However, clarifications were suggested, esp. regarding the repository type list, insourcing/outsourcing, as well as curation levels.
Organizational Infrastructure (R1-R6)

- Respondents considered R1-R5 equally relevant to generalist and specialist repositories as well as TRSPs
  - i.e. our expectations for implemented practices and policies should not fundamentally differ between these different types
- R6 “Expert Guidance” was considered less relevant to TRSPs, where the need for contracts and evidence relating to service delivery was highlighted.
Digital Object Management (R7-R14)

- In this area, respondents saw a need to distinguish specialist and generalist repositories practices and hence evidence:
  - the more specialized a service and the Designated Community needs it responds to, the more complex/specialized its practices become
Examples

- **Authenticity** becomes more complex, the more actively the data is curated (R7)
- **Preservation plan and strategies** may need to be more complex for specialist repositories (e.g. specialized file types, metadata requirements) (R10)
- Different levels of **quality checking** expected from TRSP, generalist and specialist repositories (‘technical’ vs ‘formal’ vs ‘semantic’ quality) (R11)
- Specialist repositories are expected to provide more specialized means of **discovery** (R13)
- Measures to enable **data re-use** are more complex for specialist communities (R14)
Technology (R15-R16)

- Same level of practices and evidence for all repository types and TRSP expected.
- However, especially security practices need to be informed by data type (e.g. if sensitive data is stored)
Remaining Challenges and Questions
Some Remaining Questions / Challenges

• Feedback/comments are all **valuable** and make a lot of **sense**, but still uncertain whether all questions will be solved by CoreTrustSeal, or if the **community will work with us** to solve them.

• Community must become familiar with the concept of having **three tiers of certification**, and that the tiers are **not independent**; they build upon the lower tiers.

• Two biggest difficulties in developing a three-tier certification are the definitions of ‘**preservation’** and ‘**Designated Community’**. CoreTrustSeal must ensure that we get these two definitions correct.
Some Remaining Questions / Challenges

• CoreTrustSeal allows for outsourcing partnerships, but the outsourcing of long-term preservation to others must ensure a level of oversight and control by the applicant. This is something that will be explored further with real scenarios in mind.

• The Board is now determining the next steps for both CoreTrustSeal and the wider community. The position paper is only the start, and continued discussions are needed with the community. Anything in the feedback that impacts the CoreTrustSeal Requirements must form part of the next official review of CoreTrustSeal in 2022.
Questions and Discussion
Thank you!

Contact us at info@coretrustseal.org or visit https://www.coretrustseal.org for further questions and comments.