Eye on CoreTrustSeal

Recommendations for Criterion Ro from Digital Preservation and Research Data Management Perspectives

Michelle Lindlar
TIB - Leibniz Information Centre for Science and Technology
Germany
michelle.lindlar@tib.eu
0000-0003-3709-5608

Pia Rudnik
ZB MED - Information Centre for Life Sciences
Germany
rudnik@zbmed.de
0000-0003-4081-9646

Abstract – The CoreTrustSeal (CTS) is an accepted trustworthy digital repository certification process for both, research data management and digital preservation communities alike. But does it build on concepts known and understood by both of these communities? We take an in-depth look at the CTS requirement *Ro - Background Information / Context*, in which the applicants are asked to define their repository type, designated community and level of curation performed. By extracting information from the publicly available assessment reports and cross-checking these against available supporting information, we reflect on CTS from three viewpoints: the process, the institutional, and the community view. We distill concrete recommendations, which will be fed back to the CTS Board as part of the 2019 public call for review.

Keywords – trustworthy repository certification, standardization, community watch, CoreTrustSeal

Collaboration: a Necessity, an Opportunity or a Luxury?; Exploring New Horizons

I. Introduction

Within the digital preservation (DP) and research data management (RDM) communities, institutions partake in certification for different reasons: e.g., to assure that best practice is followed; to identify gaps and improve processes; or to benchmark against other certified repositories. As certification processes like CoreTrustSeal (CTS) are self-assessment based, they can be understood as a form of institutional self-reflection. Publicly available assessment reports form a valuable documentation of this self-reflection process. CTS can be looked at as a collaborative product of and for both the DP and the RDM communities. But do both understand the underlying process in the same way and are the results of equal use to them? In this paper, we take a critical look at the CTS process via the available supporting information, analyzing understandability and usability. Identified gaps and ambiguities may be of use to the CTS Board to improve guidance documentation, thus potentially achieving broader community acceptance in the future as the current de-facto standard CTS is on its way to become an accredited European Technical Standard [1].

Our analysis is limited to CTS criterion *Ro - Background Information / Context* and based on extracted information from 40 publicly available assessment reports and available Supporting Information. The background information requested in Ro includes descriptions of *Repository Type, Designated Community, Level of Curation Performed, Outsource Partners* and *Other Relevant Information*. Ro is undoubtedly of high importance as the context information gathered within serves as a central characterization of the repository. However, unlike R1-R16, Ro does not include a self-identified compliance level, making it unclear if any objective review criteria are applied.

CTS guidance documentation describes Ro checklist options as "not considered to be comprehensive" and states that comments "may be used to refine such lists in the future" [2]. This call for recommendations is one of the main motivations for our "Eye on CoreTrustSeal". In taking the viewpoint of different actors of certification, this paper addresses the conference topic *Exploring New Horizons*, in particular the sub-theme of community self-reflection and healthy critique regarding new or improved digital preservation standards / approaches, as well as the topic *Collaboration: a Necessity, an Opportunity or a Luxury*, in particular the reciprocal support of Research Data Management and Cultural Heritage domains.

II. Related Work

CTS is considered the basic certification level in a tiered certification landscape which is followed by extended level certification via DIN 31644 / nestor Seal and ISO 163637 based formal certification. CTS is the successor to the Data Seal of Approval (DSA). For a detailed introduction into CTS history, other certification processes and the European Framework for Audit and Certification of Digital Repositories we point to available literature which discusses this in detail [3],[4],[5].

As a basic or core level certification, the CTS process provides a "minimum set" of requirements, distilled from other national and international guidelines, such as

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

DRAMBORA, the Trustworthy Repositories Audit & Certification (TRAC) Checklist and the nestor Catalogue of Criteria for Trusted Digital Repositories. Dillo and de Leuw describe CTS to be "perceived as a less 'threatening', detailed and time-consuming procedure than [...] ISO or TRAC" estimating self-assessment to take "two to four days" [1]. However, a recent study put forth that the required time for basic certification self-assessment is often underestimated [4].

CTS "emphasizes the need to conform towards the OAIS standards" [1]. In explicitly referring to ISO 14721 and reusing its terminology, CTS underlines its place in a trustworthy digital repository landscape. However, this could lead to the impression that both RDM and DP communities have an equal understanding of OAIS. Using the example of the term "data", Corrado points out that this is not always the case. While CTS leverages the OAIS definition of the term - meaning any form of digital object including documents as well as images, datasets or anything else - it is, in fact, frequently misunderstood as "research data" only [6].

Regarding R0, it is notable that most literature only mentions 16 CTS requirements, therefore specifically excluding R0 [1],[7],[6]. One notable exception is a CTS Board Member's presentation from 2019, which explicitly lists R0 as a "Core TDR Requirement" [8], being in-line with CTS extended guidance, which states that "...all Requirements are mandatory and are equally weighted, standalone items" [2]. This is also underlined by Langley, who points out that there is no equivalent to R0 as a context requirement in other processes [9]. Hence, information extracted from R0 forms a unique data basis.

Only few examples could be found which exploit the data basis of certified institutions. A survey conducted by the Netherlands Coalition for Digital Preservation (NCDD) amongst DSA-certified repositories in 2016 mainly focuses on experiences made in regards to perceived clarity of instructions / compliance requirements / reviewers' comments as well as in regards to perceived benefits [10]. Along similar lines, Donaldson et al. conducted semi-structured interviews in 2017, asking representatives from different institutions about the perceived value of acquiring DSA certification [11]. Husen et al. compiled a list of DSA, ICSU-WDS, nestor, TRAC and ISO 16363/TDR certified repositories and the respective process' requirements comparing them against repositories recommended by stakeholders [12]. While all of these examples exploit available data of certified repositories, none of them looked at publicly available assessment reports as a data basis for comparison. Furthermore, none of the examples specifically dealt with the CTS, most likely due to its rather young nature dating back to the end of 2017.

III. Methodology

The Data Basis for our analysis are responses given to R0 as documented in 40 publicly available assessment reports. Within sections A and B, we briefly describe our data gathering, cleansing, and normalization. Section C gives a short overview of different explanatory materials provided by the CTS, which we refer to as Supporting Information. The different viewpoints from which we discuss the Data Basis and the Supporting Information are introduced in section D.

A. Data Basis - Extracted information

Each report contains an "Assessment Information" cover page, where general information such as *Name of Institution*, *Country*, *Guideline Versions Applied*, *Year Achieved* and *Previously Acquired Seal* are listed. While we extracted that information as a general overview which is contained within the Data Basis [13], the main focus was on information from R0.

In R0, the applicants are asked to describe their repository by selecting 1-n applicable *Repository Type* (list-choice incl. optional comment section), to give a *Brief Description of the Repository's Designated Community* (text box), and to select 1-n *Levels of Curation Performed* (list-choice incl. optional comment section). Furthermore, applicants can add *Other Relevant Information* (text box) as well as information about *Outsource Partners* (text box) they are cooperating with.

Our Data Basis [13] consists of data extracted from the R0 section of 40 CTS Assessment reports that were publicly available on January 15th 2019[1]. The following information was gathered from each reports' R0 section:

- *Repository Type* incl. list choice answers and comment section.
- *Brief Description of the Repository's Designated Community*
- *Level of Curation Performed* incl. list choice answers and comment section.
- *Outsource Partners*
- *Other Relevant Information*
- *Reviewer Entry*: While Dillo and de Leuw [1] point out that peer review comments are part of the publicly available documentation, only few published Reviewers' Entries could be found in R0. As such entries may include helpful information about the review procedure, they are nevertheless part of our analysis.

B. Data Cleansing / Normalization

The extracted information was, where possible, normalized to allow for a quantitative analysis. Normalization could only be achieved for *Repository Type* and *Level of Curation Performed* due to underlying checklists. Since normalization was not possible for the widely differing descriptions of the repositories' *Designated Communities*, these were exemplary mapped against definitions

[1] As of March 15th, the number of certified repositories has reached 49, however, assessment reports made available after January 15th 2019 could not be considered.

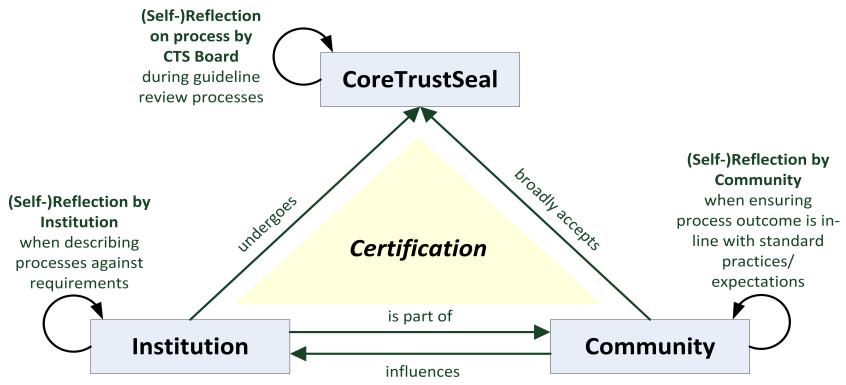

Figure 1: Three Actors of Certification

and examples provided by the CTS within the Supporting Information. The results shall highlight whether the applicants share an understanding of *Designated Community* and whether the assistance given by the CTS serves as a basis for this understanding, leading to comparable descriptions. We did not analyze extracted texts from *Outsource Partners* and *Other Relevant Information* further, as not every institution provided information, and as where given, it could neither be normalized, nor mapped against existing concepts. Due to this we do not consider *Outsource Partners* and *Other Relevant Information* to be as meaningful as the other categories.

C. CoreTrustSeal Supporting Information

There are three main sources of explanatory materials provided by CTS to which we collectively refer to as Supporting Information. As opposed to the CTS Board, which states that Supporting Information is "[...] primarily intended for giving reviewers guidance" [2], we primarily look at its benefit to the applicants' understanding of the requirements. The first Supporting Information consists of brief guidance on questions as well as checklists for *Repository Type* and *Level of Curation Performed* within the requirements themselves. The second piece of Supporting Information is the *Core Trustworthy Data Repositories Extended Guidance* [2], which includes general information on the process and more detailed information for each requirement. The third piece of Supporting Information is the *Core Trustworthy Data Repositories Requirements: Glossary* [14], which includes definitions for the terminology used in the CTS.

Regarding R0, we focused on Supporting Information about *Repository Type*, *Designated Community* and *Level of Curation Performed* to see if it can lead to an unambiguous understanding of these concepts. We examine whether the information given is easy to understand, sufficient and compliant to existing standards. In a second step, we map our Data Basis against the Supporting Information to infer if the applicants followed them and gave compliant answers, or if the assistance was misunderstood.

D. Three Actors of Certification

Trustworthy repository certification is shaped by three actors with their own views on the process: the *Institution* who undergoes the certification process, the specific certification standard itself, in our case *CoreTrustSeal*, personified by the board which signs responsible for the standard as well as the reviewers, and the *Community* which the institution is part of and who accepts the respective process as a (de-facto) standard (see Fig. 1).

Our discussion follows the actors shown in Fig. 1, while acknowledging that their boundaries are often transparent. From the institutional and the procedural viewpoint, we check if institutions are able to describe themselves against the given criteria within the requirement and if reviewers are able to judge these descriptions. This allows us to reflect on unclarities and different interpretations of requirements and to formulate suggestions for improving the CTS. While the CTS is an accepted de-facto standard in both, the RDM and the DP community, we specifically take the DP community viewpoint to check if CTS allows for a comparison against good practice.

IV. Findings

While general information about repositories is not the main scope of this paper, one aspect shall be mentioned as it might help to contextualize the answers: a comparison of available *Previously Acquired Seal* information put forth only 27.5% (11) were renewing certification from a previously held DSA. Many institutions therefore had no previous experience with basic level certification.

A. Repository Type

Repository Type is supported by a checklist which the applicant can choose all matching options from. List choices given for *Repository Type* within the Requirement, augmented here by their *Glossary* definition [14], are:

- *Domain or subject-based repository*: Specializes in a specific research field or data type
- *Institutional repository*: Generic, multi-subject repository serving a research performing institution.

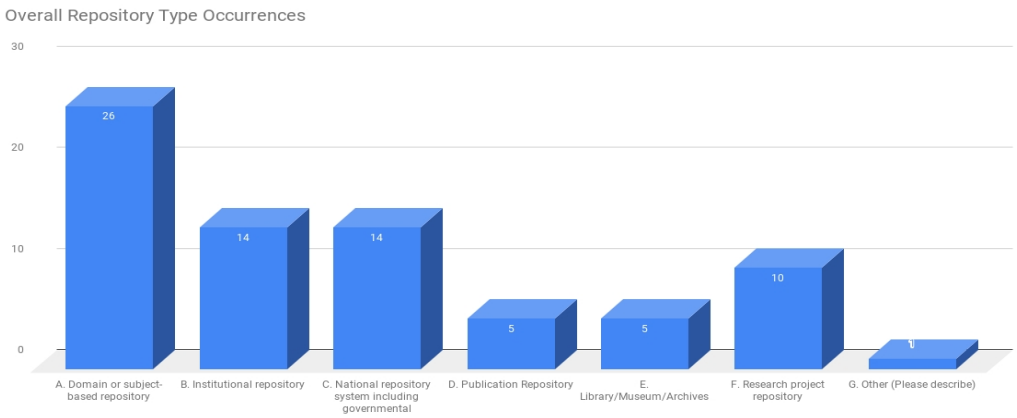

Figure 2: Overall Repository Type Occurrences

- *National repository system, including governmental*: Multidisciplinary, national infrastructure. Has a legal mandate for certain (public or governmental) data types.
- *Publication repository*: Generic, multidisciplinary repository, focussing on data linked to publications.
- *Library/Museum/Archives*: <No definition given in *Glossary*>
- *Research project repository*: Capture research results that require a deposit mandate by a funder or organization as a 'record of science'. Often tied to a specific (multi)disciplinary project
- *Other (Please describe)*

The included *Repository Types* are based on work by Armbruster and Romary [15] and were further developed by the Data Seal of Approval - World Data System partnership working group. CTS adapted this and added *Library/Museums/Archives* as well as *Other (Please describe)* [16] to the list, the latter with regard to aforementioned possible refinements in the future.

Our analysis shows that while Armbruster and Rosary use *Publication repository* as a super-classification of other repositories, CTS moved it to the level of a specific type of repository, thereby changing the definition of *Publication repository* from a rather content-type generic one to one "focusing on data linked to publications" [15],[14]. Surprisingly, all *Repository Types* with the exception of *Library/Museum/Archive* have a *Glossary* entry - here, only the OAIS definition of *Archive* can be found: "An organization that intends to preserve information for access and use by a Designated Community" [14]. The *Brief Description of Repository* text field serves several purposes: for chosen option *Other* further elaboration seems to be expected in text field, additionally, the requirement text itself points out that the field may be optionally used if the applicant wants to provide further information to contextualize the *Repository Type* for the reviewer. In contrast, the *Extended Guidance* points out that a description should be given if more than one type is chosen, providing details on how these roles are fulfilled [2] - a pointer easily missed by applicants who do not carefully read through all Supporting Information.

The analysis of the Data Basis put forth 4 different types of answers for *Repository Type*. 20% (8) answered using just the list choices available, 52.5% (21) used list choices and added additional comments, 25% (10) did not explicitly state list choices first, but instead answered in prose form including keywords that allowed a mapping to list choices. 2.5% (1) did not answer the question at all. In a first step, we normalized the answers by mapping prose answers to list choices to allow for a quantitative comparison of assigned *Repository Types*.

While *Repository Type* offers a multiple choice option, 55% (22) of the assessed institutions identified against one *Repository Type*, 15% (6) against 2, 12.5% (5) against 3, 10% (4) against 4 and 5% (2) against 5 *Repository Types*. An interesting finding is that for institutions with multiple *Repository Types*, all but two combinations included the option *Domain or subject-based repository*. Overall, only 35% (14) of the certified repositories did not identify as a *Domain or subject-based repository*. Fig. 2 shows the overall occurrences of *Repository Types* in the Data Basis. Two reasons come to mind for the comparatively low representation of *Publication repository* and *Library/Museum/Archives*: either, the actual number of assessed institutions do not identify against these types -or- these two options were not as comprehensible as others and therefore not chosen as often. The latter is supported by the fact that especially the option *Library/Museum/Archive* is not, like all other choices, a classification for repositories, but rather an organizational one. Further insights into the institutional view and why different *Repository Types* were chosen shall be gained by an analysis of comments given in 31 of 40 assessment reports. Three overarching themes could be identified in those comments, in some cases more than one were mentioned. Distribution of comment types across occurrences are as follows:

- Further describing domain content (11)

- Further describing domain content & mission / history (4)
- Further describing domain content & services provided (4)
- Further describing domain content & mission history & services provided (12)

While Supporting Information asks for additional information when more than one *Repository Type* is chosen, two institutions chose several options without describing the roles further. Another 3 applicants used the comments to specifically refine *Repository Type* roles. An example for this is the following answer by the Finish Social Data Archive (FSD) [13]: "Domain or subject based repository: A domain-based repository with focus on research data from social sciences; National repository system, including governmental: A national service resource for research and teaching; Library/Museum/Archives: Social science data archive".
The FSD serves as an epitome for what we see as a major point of confusion regarding *Repository Type*. It described itself against three different types, thankfully providing further input on why these specific types were chosen. The FSD's additional information can be interpreted as follows: *Domain or subject-based repository* describes the repository's depth (intellectual content focus), *National repository system* describes the repository's width (of audience), and *Library/Museum/Archive* describes a function the repository performs, namely archiving. While a description against such a 3-dimensional matrix (depth / width / function) would be highly beneficial and allow an adequate comparison of repositories against each other, it actually goes against the definitions of *Repository Types* as per CTS Supporting Information. To exemplify: FSD's answer to be simultaneously a *Domain or subject-based repository* for research data from social sciences -and- a *National repository system* is not legitimate when taking the *Glossary* into account, as per definition such a repository is multidisciplinary - whereas FSD is limited to the social science discipline. FSD is not the only example, where descriptors for domain-specific and multidisciplinary were mixed - *Domain or subject-based repository* and *National repository system* were assigned - sometimes in combination with even further types - by a total of 11 repositories.

In a second step, we searched the extracted comments for the keywords "preserv*", "archiv*", "long-term" and "OAIS". A total of 11 assessment reports included these keywords within comments to *Repository Type*, ranging from generic statements such as providing "the opportunity for individuals and organizations to deposit collections that are considered worthwhile preserving for future generations and research projects" to specific statements such as being an "OAIS compliant web-based platform". Out of these 11 repositories, only two classified their repository as being of type *Library/Museum/Archives*, despite the fact that they are fulfilling archiving / preservation functions, according to their comments.

From a DP community point of view, the mixture of depth, width and function within *Repository Types* creates a confusing classification schema, resulting in a difficult to compare data basis. Additional stumbling blocks exist in non-intuitive *Glossary* definitions, as per which, e.g., a *Publication repository* does not contain publications but rather focuses on data linked to publications. Going back to Corrado's comment regarding data as defined by OAIS, one valid reading of "data linked to publications" could then be "publications linked to publications". It remains unclear what is meant by the CTS definition of *Publication repository*, and, furthermore, how assessed institutions understood it. Another unclarity exists on a most fundamental level: As further questions within Ro are built on OAIS concepts such as *Designated Community*, we wonder whether *Repository* is seen as a technological platform or rather understood as the *Archive* in the sense of the OAIS.

B. Designated Community

Asking the applicants to give a *Brief Description of the Repository's Designated Community* is another clear CTS reference to OAIS. OAIS defines *Designated Community* as an "identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time" [17]. The *Designated Community* concept is central to OAIS, as the so-called *Representation Information* provided by an *Archive* is dependant on both, the *Designated Community's* specific needs as well as its ability to understand information. Within OAIS, concrete preservation actions performed by an *Archive* are always connected to the *Designated Community* [17]. However, even within the DP community the concept of *Designated Community* is controversially discussed: its central importance within OAIS is felt to stand in stark contrast to its rather abstract and speculative nature, which leads to what is often perceived as a lack of practical use [18],[19],[20].
The question at hand is, whether CTS offers sufficient assistance to applicants, resulting in a clear understanding and adequate description of their *Designated Community*. What stands out first by looking at the Supporting Information is that it provides the OAIS definition of *Designated Community*, but initially contradicts this definition by referring to a *Repository's Designated Community*, whereas according to OAIS, only an *Archive* can have a *Designated Community* [17]. While CTS does not further explain this discrepancy, we assume that it happened unintentionally. Despite this impreciseness, CTS underlines the importance of *Designated Community* by explicitly referring to it in various requirements and their corresponding guidance (R4, R5, R6, R8, R10, R11, R14, R15, R16), as well as by requesting precise information from the applicants regarding the *Designated Community* [2]: "Please make sure that the response is specific—for example, 'quantitative social science researchers and instructors". The following description within CTS Supporting Infor-

mation should further "help elucidate" the meaning of *Designated Community* [2]: "A clear Designated Community demonstrates that the applicant understands the scope, knowledge base, and methodologies—including preferred software/formats—they are targeting. [...] A repository with a very specific, narrow Designated Community might state the expected knowledge base (e.g., degree/level of understanding of statistics or genetics). A very broad designated community (e.g., the general public) would imply that the repository has a wider range of contextual documentation to ensure its data can be understood by everyone".
In other words, the CTS wants the applicants to specifically define their *Designated Community* using terms like "broad" or "narrow" and concepts like "scope", "knowledge base" (degree of understanding; only applicable for a narrow *Designated Community*), "methodologies" (e.g., software/format preference) and necessary "contextual documentation" (the extent of which depends, according to CTS, on the broadness of the *Designated Community*). While the CTS seemingly expects the applicants to use these concepts, they are, at least in our opinion, not sufficiently and specifically enough introduced and defined within the Supporting Information. The OAIS concept *Knowledge Base* [17], for example, is neither included in the *Glossary* [14], nor elaborated on in any other piece of Supporting Information. Furthermore, it remains unclear whether it is the *Designated Community* itself or its "scope" that should be described as being either broad or narrow. Does a narrow *Designated Community* consist of few people, whereas a broad *Designated Community* means "many people"? And, what exactly is meant by "scope"? Is "scope" a further specification of a domain (e.g., "social sciences"), of a profession (e.g., "researchers and instructors"), or both? Finally, it remains unclear whether the term "contextual documentation" is used synonymous to the OAIS concept of *Representation Information* or if it means something else.
These ambiguities result in the following question for both the process as well as the institutional view: How could applicants who are not familiar with the terminology used in the Supporting Information meet the reviewers' requirements and give an adequate and "specific" definition of their *Designated Community*?

Our Data Basis shows a variety of *Brief Descriptions of the Repository's* (sic!) *Designated Community*. We searched the descriptions for the keywords "designated community", "scope", "knowledge base", "method*" and "OAIS" to check if the applicants explicitly referred to the Supporting Information's terminology. In total, only 35% (14) use the term "designated community" in the prose text (we excluded the sub-headline *Brief Description of Repository's Designated Community* from this analysis). Instead, terminology such as "users" or "target community" are used within the answers. The term "scope" is mentioned by 10% (4), but not in the sense of "*Designated Community's* scope". Concerning the *Designated Community's* knowledge base, only the Finnish Social Science Data Archive (2.5%) explicitly refers to it. The term "method*" is used 4 (10%) times in total, but only 7,5% (3) matching responses actually use it to describe their community's methodologies. Finally, only 5% (2) explicitly refer to OAIS in their description of *Designated Community*. These results show that neither the terminology provided by the CTS via the Supporting Information, nor the OAIS terminology seem to be accepted or understood by the majority of CTS applicants.

As a quantitative analysis based on normalized data is not possible for the *Designated Community* responses, we now take a look at three representative examples for *Designated Community* descriptions [13]:
The Commonwealth Scientific and Industrial Research Organisation (CSIRO) gives an example of a specific description of a broad *Designated Community* according to CTS. What stands out first is that CSIRO does not describe its own, but the *Designated Community* of its repository, the Data Access Portal. While this is in-line with CTS Supporting Information, it is not, as mentioned above, compliant to OAIS, which says that only an *Archive* can have a *Designated Community*. CSIRO defines the *Designated Community* as "broad" and as "data users [that] may include: general public / industry specific groups such as agriculture / policymakers / students / researchers". Further, the institution underlines the broadness by listing 17 different research disciplines which are represented in its collections, e.g. "Environmental Sciences". Although not explicitly referring to the terms "scope", "knowledge base" or "methodologies", CSIRO's answer seems to be specific enough for the CTS reviewers considering the fact that no *Reviewer Entry* has been made. This is consistent with the argumentation in the Supporting Information, that it might not be possible to state the *Knowledge Base* of a broad *Designated Community*. Therefore it can be presumed that, from the CTS's viewpoint, CSIRO gives a specific description of a broad *Designated Community*. Despite the answer apparently fulfilling CTS expectations, we cannot fully agree with it, due to the above-mentioned inconsistency to OAIS.
Another example of a broad community is given by the Norwegian Centre for Research Data (NSD). NDS's response is accompanied by a *Reviewer Entry*. This is an important aspect for our analysis, as the entry represents the CTS's point of view and gives us an impression of what CTS considers an unspecific answer. What initially stands out is that the NSD does not refer to a "Designated Community", but to a "target group". This leads to the impression that the NSD is not aware of the prospective and speculative nature of the concept of *Designated Community*, and therefore alternatively presents its current target group. This group is defined as "research community" represented by "Researchers and students / Research institutions / Finance providers / National authorities". The NSD further describes "Norwegian social science research" as its main scientific focus, but also refers to its "competence to several fields of research" and its "multi- and interdisciplinary mandate". Though the NSD points out that it is not limited to Norwegian social sciences, it does not, in contrast to CSIRO, name

concrete alternative scientific disciplines. This impreciseness is criticised in the *Reviewer Entry*: "It may be helpful to add to the Designated Community the domain specific scope ('the social sciences, humanities and medicine and the health sciences' like listed in the Strategic Plan)". Not only does this feedback clarify the meaning of "scope" as domain-specific (not profession-specific), it also demonstrates that the scope is apparently important to CTS reviewers, whereas the above-mentioned contradiction to OAIS by describing actual, not potential consumers, is not mentioned at all. This leads to the impression that in this case the CTS reviewer was not aware of this important aspect within the OAIS concept of *Designated Community*. Alternatively, it is possible that the CTS is aware of the discrepancy, but has agreed to ignore it per se without having described this approach in the Supporting Information.

We conclude the *Designated Community* analysis with an example of a specific description of a narrow *Designated Community* according to CTS. Due to the specialised scientific focus of the Institute for Computational Linguistics "Antonio Zampolli", its *Designated Community* can be regarded as narrow. This is supported by the lack of a *Reviewer Entry*, leading to the assumption that the CTS regarded this description as specific. Generally stating that its *Designated Community* "is constituted by the scholars of disciplines where language plays a central role", the Institute Zampolli continues more concretely by explicitly naming "producers" and "consumers" as representatives of its *Designated Community*. The institute includes both, their domain-specific (e.g. "Humanities and the Cultural Social Sciences") as well as their professional scope (e.g. "Computational Linguists, Information and Communication Technologies (ICT) experts and Language Engineers") within the description. Furthermore, examples of methodologies that are used by its *Designated Community* are included: Producers "produce language data and digital tools to work with such data", whereas consumers "are interested in analyzing language data and using text processing tools available in the CLARIN infrastructure." According to CTS Supporting Information, the narrowness of its *Designated Community* should allow the Institute "Zampolli" to state the community's *Knowledge Base*. From our viewpoint, the above-mentioned detailed descriptions can be seen as an implicit description of *Knowledge Base*, even though the term itself is not mentioned. Apparently, CTS reviewers agree and regard the Institute's response as compliant to the CTS process. But, from a community viewpoint, does the response correspond to OAIS? At least, the Institute seems to be familiar with this digital preservation standard: "We ensure long term preservation [...] according to the definition of Preservation Description Information (PDI) given in the OAIS reference model". Taking this into consideration, it is even more striking that, like CTS and CSIRO, the Institute Zampolli also contradicts OAIS by mistaking *Repository* for *Archive*: "The aim of a CLARIN repository is to preserve research data sets and make them available for a Designated Community".

C. Level of Curation Performed

The second R0 question which is supported by a checklist is *Level of Curation Performed*. Like *Repository Type*, *Level of Curation* is also a multiple-choice option, asking the applicant to select any combination of the four offered choices. An optional comment section exists, allowing the applicant to provide further details on the *Levels of Curation* chosen. List choices provided for *Level of Curation* can be seen in Table 1. The different *Levels of Curation* were developed by the RDA-WDS Cost Recovery Interest Group and used in a survey of over twenty data centers to understand different income streams and cost recovery models in place [21].

Supporting Information clarifies that curation actions are to take place on copies of the data, while original deposits should remain unchanged and should be in-line with producer-repository license agreements. Additionally, CTS expects any edits made to the data to be conducted and documented by respective (domain) experts, recommending to further describe this in the comment section to allow for a better understanding by reviewers. Especially in case of Level *D*, "reviewers will expect a higher level of formal provenance, integrity, and version management (change logs etc.)" [2]. The same type of answer classification previously used for *Repository Type* was applied for the following data analysis. For *Level of Curation Performed*, 22.5% (9) answered using just the list choices available, 55% (22) used list choices and additional comments, 20% (8) did not explicitly state list choices, instead answering in prose form included keywords that allowed a mapping to list choices. 2.5% (1) gave an answer which could not be mapped to a *Level of Curation Performed* option. Thus, all except 1 answers were matched to the *Level of Curation Performed* choices to allow for a quantitative comparison. Despite the fact that multiple answers are possible, 55% (22) identified against only one specific *Level of Curation*. Here, it is interesting to note that the most frequent stand-alone choice was *D. Data-level curation* (11), whereas no institution identified against just *A. Content distributed as is*.

Table 1: Level of Curation Performed with Answer Frequency

Level of Curation	Answer Frequency
A. Content distributed as deposited	9
B. Basic curation - e.g., brief checking, addition of basic metadata or documentation	22
C. Enhanced curation - e.g., conversion to new formats, enhancement of documentation	17
D. Data-level curation - as in C above, but with additional editing of deposited data for accuracy	20

It can therefore be said that every assessed institution performs at least a basic *Level of Curation* on some of its data. 42.5% (17) of the institutions chose 2-3 different *Curation Levels*, 2.5% (1) stated that all four options are applied to some (sub-)sets of data. Table 1 shows the overall occurrences of *Level of Curation Performed* in the Data Basis.

Further insights into the assessed institutions' understanding of the different *Curation Levels* were sought via the 31 comments provided. Here, two major themes could be identified: additional information on the process and additional information on applicability of different levels chosen. 14 comments included further refinements of applicability of levels chosen, e.g., pointing out that *Data-level curation* will only take place when additional funding is available or that a lower level applies to legacy data. It seems that the commenting institutions were unsure how to answer if different *Curation Levels* apply. A full understanding of answers is therefore only possible when taking additional comments into consideration.

One example could be identified, where an institution chose one *Level of Curation* (*D. Data-level curation*), but commented that only some data is checked against linked data for accuracy - a fact which furthermore underlines the need for an additional qualifier for curation frequency. It can be assumed that the institution understood the question as asking for the highest *Level of Curation* applied, therefore omitting the level which applies to the rest of its holdings. As a consequence, the quantitative data analysis as shown in Table 1 does not necessarily provide a basis to benchmark against: it allows us to know how many institutions perform specific levels, but it does not allow us to derive if this is done as a standard procedure for the majority of data or as an exception for sub-sets under specific circumstances only.

Another question is whether the assessed institutions actually had a shared understanding of the different level choices available - we approach answering this question by analyzing the different comments given. Taking further comments for *Level D* as an example, process descriptions range from "Metadata on variable level is displayed in the online catalogue" over "include e.g. creation of new formats, enhancement of documentation and also additional editing of deposited data for accuracy" to "[...] staff review all incoming data files and apply specialized curation activities such as quality assurance, file integrity checks, documentation review, metadata creation for discoverability, and file transformations into archival formats. We work closely with authors to ensure that the data is in a format and structure that best facilitates long-term access, discovery, and reuse" [13]. Based on the wide range of different process descriptions, it appears that the definition of the different levels is not clear - something that was already noted during a testbed run of the CTS process in 2016, where regarding *Level D* the point was raised that "data accuracy" may be misunderstood as pertaining to the quality of research / intellectual data accuracy, when instead a "technical quality in terms of good data" including metadata is meant [16]. Unfortunately, it seems that definitions of the different levels continue to be unclear to institutions going through self-assessment.

Of high interest from a community point of view is the fact that a definition of *Curation* is currently missing in the *Glossary*; this was already pointed out during the testbed run, where the following definition was suggested: "Activities required to make deposited data preservable or usable now and in the future. Depending on technological changes, curation may be required at certain points in time throughout the data lifecycle" [16]. The implementation of this definition into the *Glossary* would be extremely helpful for several reasons. First off, the connection between *Curation* and *Preservation* is not always clear. As Lazorchak points out, "Curation is a useful concept for describing the evolving whole-life view of digital preservation, but concentrates on underpinning activities of building and managing collections of digital assets and so does not fully describe a more broad approach to digital materials management" [22]. Within the RDM community, typical curation life-cycle activities include assigning persistent identifiers and curating metadata, whereas DP focuses on life-cycle activities such as preservation metadata creation and file-format based processes like file format characterization and migration or emulation. An analysis of the comments shows that only one institution mentions technical metadata creation, and while 4 comments mention OAIS terminology such as *Submission Information Package*, the majority of comments seem uniformed of DP practice.

The fact that even within the RDM community, uncertainty regarding *Curation* terminology exists, can be seen in yet another testbed comment. Initially, the example given for *Level C. Enhanced curation* was "e.g., creation of new formats", leading to the question whether this describes a (pre-)ingest / normalization process or a migration process. In order to make the answer clearer, the CTS Board decided to change this to "conversion to new formats", hoping to make the intended meaning clearer [16]. At least in our view, the meaning remains as unclear as it was before - only if read in conjunction with the (currently missing) *Glossary* definition is it understood to describe migration rather than normalization. Furthermore, the term *Levels of Curation* reminds one of *Levels of Digital Preservation* [23], a well-accepted framework within the DP community. However, neither can the concepts included in *Level of Curation Performed* be mapped to *Levels of Digital Preservation*, nor are they as concise and exhaustive as the latter.

V. Conclusion

Based on the available Supporting Information and our Data Basis, we have presented an in-depth analysis of the R0 criteria *Repository Type*, *Designated Community* and *Level of Curation Performed* from a procedural, institutional and community point of view. This analysis has put forth a number of weaknesses within the process,

which resulted in ambiguous assessment report answers. While one of the benefits of trustworthy repository certification is the ability to benchmark repositories against good practice as well as against other repositories, a data basis which is open to interpretation due to unclear results prohibits such benchmarking. Furthermore, we have taken a close look at the process from a DP community point of view, pointing out where CTS descriptions go against concepts which are standardized and accepted within the community. This final section presents concrete recommendations for the improvement of the CTS, which we derive from our findings. They are grouped by the respective R0 section and are preceded by a number of generic process recommendations. All recommendations will be passed on to the CTS Board as part of the current review of CoreTrustSeal requirements.

A. Generic Recommendations for R0

1. Heighten Relevance of R0 Assessment: Despite the importance of context information, R0 seems to be treated in a rather irrelevant manner by both reviewers and applicants alike, possibly due to the fact that there is no associated compliance level. While we acknowledge that compliance level are not necessarily applicable to R0, lacking information or incomplete answers should in our view be of consequence and not be disregarded during the review process.

2. Standardize Assessment Reports: Different templates for assessment reports exist. Some institutions have used the online form and others provided the self-assessment in a document form. These different handing-in procedures can lead to different forms of answers, where some institutions do not stick to the given list choices, instead writing prose answers which are hard to map and also hard to judge by reviewers. This should be standardized with clear routines checking for mandatory information, e.g. at least one checkbox ticked for *Level of Curation Performed*.

3. Introduce OAIS in Supporting Information: Our analysis has shown that many applicants seem neither familiar with the OAIS in general, nor with its core concepts like *Designated Community* in particular. This often leads to answers that are not compliant to the standard, not easy to compare, and difficult to assess by reviewers. We therefore recommend an implementation of a short introduction to OAIS in the Supporting Information.

4. Include concise Definitions and Examples: To avoid unclarity concerning terminology used in R0, we recommend that clear definitions and concrete examples are added to the Supporting Information. This applies to the list choices for *Repository Type* and *Level of Curation Performed* as well as to terms like "scope", "methodologies", "contextual documentation", and "knowledge base" used in the *Designated Community* section. Including definitions and examples directly within the requirements might significantly improve the applicants' understanding of R0 and omit the need to query different pieces of information.

B. Recommendations for Repository Type Section

1. Replace mixture of depth, width and function with 3-level approach: Instead of mixing different levels within a *Repository Type* list choice, institutions should describe themselves against the three levels separately: Depth/Content - ranking from domain-specific to multidisciplinary; Width/Audience - ranking from project specific via institutional to national; Function - determining whether archiving is included or not.

2. Define Repository's Boundaries: It is not clearly defined if *Repository* within CTS pertains to the technological concept or to an *Archive* in the OAIS sense. This definition should be clearly stated by the CTS, especially as further R0 questions and concepts such as *Designated Community* build on the OAIS understanding of *Archive*.

C. Recommendations for Designated Community Section

1. Change sub-headline "Brief Description of Repository's Designated Community": Given the fact that the expression "Repository's Designated Community" is not compliant to OAIS, the term *Repository* should either be deleted or replaced by the term *Archive*. This would avoid one inconsistency to OAIS which we observed in our analysis.

2. Stimulate formalized Descriptions of Designated Community: A formalized way to describe *Designated Community* would be helpful. It would lead to a better understanding of the concept itself, stimulate self-reflection and result in comparable answers. The inter-dependency of the concepts "scope", "methodologies" and "knowledge base" need to be exemplified, e.g., by referring to the Digital Preservation Coalition which states: "the broader the scope of the Designated Community, the less specialized the knowledge associated with that community" [24]. A questionnaire, e.g., on the domain-specific and professional scope of a *Designated Community*, would be a helpful orientation.

D. Recommendations for Level of Curation Section

1. Describe Conditions for Levels Applied: Few institutions follow one *Level of Curation* for all data. Adding a tiered model, where each applicable level is described more granular, e.g. as, "applies to (a) all objects (b) sub-collections based on depositor agreement (c) sub-collections based on external requirement / funding (d) sub-collections based on technical suitability" should lead to a meaningful assessment approach.

2. Include Digital Preservation-centric Model: To understand the preservation functions the repository fulfills, a preservation focused model should be included in R0. This can be either the *Levels of Preservation* model, or, at the simplest level, by asking the institution which of the preservation levels bit-stream / logical / semantic are implemented.

References

[1] I. Dillo and L. de Leeuw, "CoreTrustSeal," *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare*, vol. 71, no. 1, pp. 162–170, 2018. DOI: 10.31263/voebm.v71i1.1981.

[2] CoreTrustSeal, *Core Trustworthy Data Repositories Extended Guidance*. 2018, Version 1.1. [Online]. Available: https://www.coretrustseal.org/wp-content/uploads/2017/01/20180629-CTS-Extended-Guidance-v1.1.pdf.

[3] Digital Preservation Coaliton, *Digital Preservation Handbook*. 2015. [Online]. Available: https://www.dpconline.org/handbook.

[4] M. Lindlar and F. Schwab, "All that Work ... for What? Return on Investment for Trustworthy Archive Certification Processes – a Case Study," in *Proceedings of the 15th International Conference of Digital Preservation*, 2018. DOI: 10.17605/OSF.IO/8A3SC.

[5] CESSDA, "Cessda user guide: Trusted digital repositories: Audit and certification," 2015. DOI: 10.5281/zenodo.2595238.

[6] E. M. Corrado, "Repositories, trust, and the coretrustseal," *Technical Services Quarterly*, vol. 36, no. 1, pp. 61–72, 2019. DOI: 10.1080/07317131.2018.1532055.

[7] E. Zhang and M. R. Moulton. (2018). Preparing for a Trustworthiness Assessment of the National Transportation Library's Digital Repository ROSA P, [Online]. Available: https://rosap.ntl.bts.gov/view/dot/34480.

[8] R. Jenkyns, "Building trust in scientific data: Certification & the coretrustseal," 2019. DOI: 10.5281/zenodo.2555289.

[9] S. Langley. (2018). Digital preservation maturity model, risks, trusted digital repository certification and workflow crosswalk [dataset], [Online]. Available: https://www.repository.cam.ac.uk/handle/1810/287009.

[10] K.-J. Waterman and B. Sierman, *Survey of DSA-certified digital repositories: Report on the findings in a survey of all DSA-certified digital repositories on investments in and benefits of acquiring the Data Seal of Approval (DSA)*, English. Netherlands Coalition for Digital Preservation, Nov. 2016.

[11] D. R. Donaldson, I. Dillo, R. Downs, and S. Ramdeen, "The perceived value of acquiring data seals of approval," *International Journal of Digital Curation*, vol. 12, no. 1, pp. 130–151, 2017. DOI: 10.2218/ijdc.v12i1.481.

[12] S. Husen, Z. Wilde, A. de Waard, and H. Cousijn. (2017). Data Set - Recommended versus Certified Repositories: Mind the Gap.

[13] M. Lindlar and P. Rudnik. (2019). Eye on core trust seal - data set, [Online]. Available: https://dx.doi.org/10.5281/zenodo.3267690.

[14] Data Seal of Approval and ICSU World Data System. (2016). Core Trustworthy Data Repositories Requirements: Glossary, [Online]. Available: https://drive.google.com/file/d/0B4qnUFYMgSc-REpsNVQwWDVfSkU/view.

[15] C. Armbruster and L. Romary, "Comparing Repository Types: Challenges and Barriers for Subject-Based Repositories, Research Repositories, National Repository Systems and Institutional Repositories in Serving Scholarly Communication," 2009. DOI: 10.2139/ssrn.1506905.

[16] DSA–WDS Partnership Working Group. (2016). DSA–WDS Partnership Working Group Results of Common Certification Testbed, [Online]. Available: https://www.rd-alliance.org/system/files/DSA-WDS_Common_Certification_Testbed_Report_V0_5_Final.pdf.

[17] CCSDS, *Reference Model for an Open Archival Information System (OAIS) - Magenta Book*, CCSDS, 2012. [Online]. Available: http://public.ccsds.org/publications/archive/650x0m2.pdf.

[18] R. S. Bettivia, "The power of imaginary users: Designated communities in the oais reference model," *Proceedings of the Association for Information Science and Technology*, vol. 53, no. 1, pp. 1–9, 2016. DOI: 10.1002/pra2.2016.14505301038.

[19] C. Keitel, "Der einzige Kompass, den wir haben: Zur Kritik der Designated Community," *Informationswissenschaft: Theorie, Methode und Praxis*, vol. 5, no. 1, 2018. DOI: 10.18755/iw.2018.4.

[20] D. Rosenthal. (2014). TRAC Audit: Lessons, [Online]. Available: https://blog.dshr.org/2014/08/trac-audit-lessons.html.

[21] RDA-WDS Interest Group on Cost Recovery for Data Centres, "Income streams for data repositories," 2016. DOI: DOI:10.5281/zenodo.46693.

[22] B. Lazorchak. (2011). Digital Preservation, Digital Curation, Digital Stewardship: What's in (Some) Names? Library of Congress, [Online]. Available: https://blogs.loc.gov/thesignal/2011/08/digital-preservation-digital-curation-digital-stewardship-what%E2%80%99s-in-some-names/.

[23] M. Philips, J. Bailey, A. Goethals, and T. Owens, "The NDSA Levels of Digital Preservation: An Explanation and Uses," in *Proceedings of the Archiving (IS&T) Conference*, 2013. [Online]. Available: {https://ndsa.org/documents/NDSA_Levels_Archiving_2013.pdf}.

[24] B. Lavoie. (2014). The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition), Digital Preservation Coalition, [Online]. Available: https://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file.