Curation & Preservation Levels[1]

CoreTrustSeal Board **Discussion Paper v2.0**

Context

Applying the appropriate levels of curation and preservation to digital objects maximises the return on investment in data assets over time. Successful curation and long-term preservation services depend on a repository having the rights and taking the responsibility to provide an effective organisational infrastructure, digital object management and technical/security environment. Clearly communicating the levels of *curation* a repository offers[2] is an essential part of seeking certification against the CoreTrustSeal Requirements 2023-2025[3]. To be in scope for CoreTrustSeal applicants must take responsibility for *active long-term digital preservation* for a defined ('designated') community of users. As the issues of curation, preservation[4] and certification are receiving more attention from a wider range of actors, the need for clearer specification of preservation levels has become clear. The CoreTrustSeal Board sees this as an important issue for the data management community, for defining which applicants are in-scope for certification, and as a step toward the improved definition of all data and metadata services, including those that do not offer active preservation.[5]

Version 1.0 of this discussion paper was published during the 2022 revision of the CoreTrustSeal Requirements. Community feedback was received via direct comments on the public version of the

[1] Cite as: CoreTrustSeal Standards and Certification Board. (2023). Curation & Preservation Levels (v02.00). Zenodo. https://doi.org/10.5281/zenodo.8083359.

[2] See Appendix: CoreTrustSeal 2023-2025 Levels of Curation

[3] CoreTrustSeal Standards and Certification Board. (2022). CoreTrustSeal Requirements 2023-2025 (V01.00). Zenodo. https://doi.org/10.5281/zenodo.7051012

[4] https://www.eosc.eu/advisory-groups/long-term-data-preservation

[5] CoreTrustSeal: Specialists, Generalists, and Technical Repository Service Providers https://doi.org/10.5281/zenodo.3964071

document, via social media and email, as well as in the form of responses from the UK Data Service[6], the Digital Preservation Coalition (DPC)[7] and the EOSC Association's Long-Term Digital Preservation Task Force. Appendix A summarizes the comments received. This version (2.0) addresses the feedback and is made available for further community discussion and comment as we progress towards consensus.

Introduction

In addition to delivering Trustworthy Digital Repository (TDR) certification at a 'core' level, the CoreTrustSeal seeks to align with and contribute to the wider data lifecycle and landscape. For disciplinary and generalist data repositories, and across the (meta)data product and service provider ecosystem, the level of curation and preservation delivered for each object must clearly be communicated to data users, and to other stakeholders, including policy makers and funders.

If curation can be understood as the actions that deliver an *immediate* benefit to digital objects, then preservation includes these, and other steps to ensure data and metadata remain accessible, usable and understandable *into the future*. Preservation takes account of ongoing changes to the knowledge base of the user community and the surrounding technical context. Long-term does not have to mean 'forever'. Objects may be reappraised over time and their level of curation or preservation may change. Long-term preservation means that organizational measures, infrastructure, and policies are in place to actively preserve digital assets for *as long as necessary*. Minimum periods of *retention* are important and should be clear, but these do not equate to *active preservation*. To qualify for the CoreTrustSeal a repository must deliver active preservation while meeting a number of mandatory responsibilities (see Fig. 1).

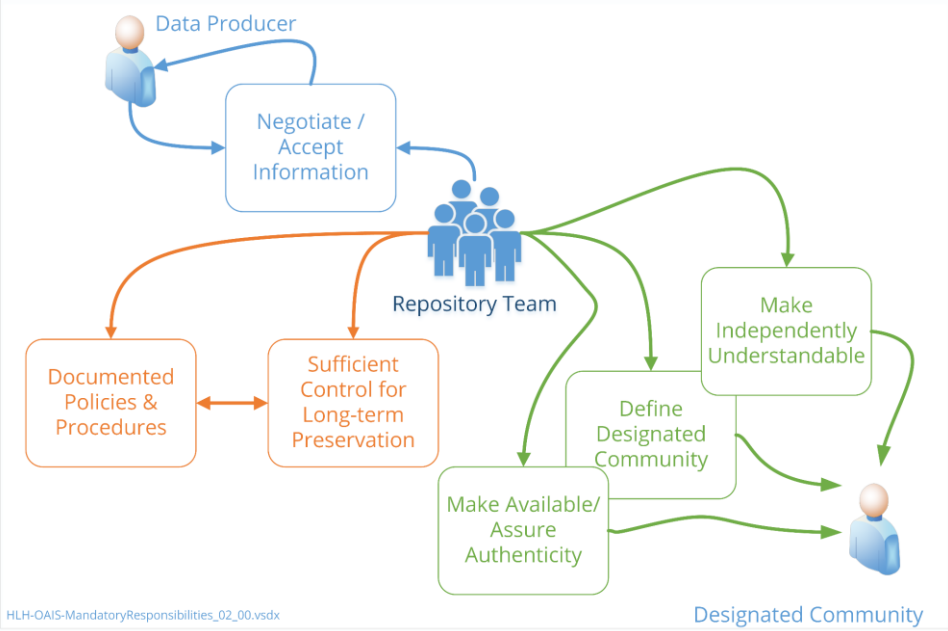

Figure 1: Mandatory Responsibilities[8]

[6] L'Hours, Hervé, & Bell, Darren. (2023). UK Data Service (UKDS) Response to the CoreTrustSeal Curation & Preservation Levels Discussion Paper (v01.00). Zenodo. https://doi.org/10.5281/zenodo.7828046.
[7] https://www.dpconline.org/
[8] See section 3.1 of Reference Model for an Open Archival Information System (OAIS), https://public.ccsds.org/pubs/650x0m2.pdf

To remain 'understandable' and in line with principles such as FAIR[9] (Findable, Accessible, Interoperable and Re-Usable) it is necessary to ensure that supporting metadata also remains fit for the purposes of designated community use and ongoing preservation. Both data and metadata need to be preserved through managed changes that address, for example, the knowledge or skill set of the designated community. Common approaches to active preservation can include transformations to new data formats and metadata schemas, and updates to (meta)data content so that digital objects remain understandable and technically usable by the community.

The designated community's needs and preferences must be considered when determining the preservation actions to be applied. This depends on monitoring the knowledge base and technology needs of the community, and an understanding of wider technical risks[10] as well as technical developments that may impact how digital objects can be used. For digital objects with specialist characteristics and users (e.g. disciplinary) the active preservation of (meta)data can be more challenging and require additional expertise. A more generalist approach may not preserve those characteristics or meet those specialist needs.

Digital Objects as the Focus of Active Preservation

Efforts to unite different perspectives on information management are not new[11]. Active digital preservation ensures the continued use and understanding of digital objects for a defined designated community.

> "Every digital object is a physical object, a logical object, and a conceptual object, and its properties at each of those levels can be significantly different. A **physical** object is simply an inscription of signs on some physical medium. A **logical** object is an object that is recognized and processed by software. The **conceptual** object is the object as it is recognized and understood by a person, or in some cases recognized and processed by a computer application capable of executing business transactions".[12]

The intellectual **conceptual** entity depends on a **logical** entity rendered through a given hard- and software environment, based on bits and bytes stored on a **physical** medium. To ensure the continued use and understanding of digital objects by a defined designated community a Trustworthy Digital Repository must provide active preservation at the physical, logical and conceptual level.

The draft levels presented below address conceptual and logical curation and preservation[13]. A single organisation may hold digital objects that are curated and preserved at different levels. These levels must be sufficiently specific and granular to communicate the care provided and the degree of responsibility taken by a repository or other data service at the object level.

---

[9] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[10] Covered in detail in FAIR + Time: Preservation for a Designated Community https://doi.org/10.5281/zenodo.4783116

[11] https://www.zachman.com/ea-articles-reference/58-conceptual-logical-physical-it-is-simple-by-john-a-zachman

[12] https://chnm.gmu.edu/digitalhistory/links/pdf/preserving/8_37e.pdf ; see also https://www.naa.gov.au/sites/default/files/2020-01/An-Approach-to-the-Preservation-of-Digital-Records.pdf and OAIS for similar typologies.

[13] All assume that effective physical measures are in place, including back-ups and multi-format, multi-location, multi-copy redundancy and integrity.

Curation and Preservation Levels (v2.0)

The levels of curation and preservation below provide an initial basis for describing repository services[14]. The levels are cumulative as they progress from D to A. A repository may offer different levels of service for different digital objects.

In addition to describing the range of services offered at an organisational level this approach would enable repositories to describe and document the level of care at individual object level, thereby contributing to documented audit and provenance trails.

From the perspective of the CoreTrustSeal, Levels Z, D and C alone are not in scope for CoreTrustSeal certification as they do not entail active long-term preservation and hence do not provide a long-term perspective beyond bit storage. However, a repository may apply a workflow to an object that includes deposit compliance (D), and initial curation (C), followed by active preservation (B and A). Agreement on the definition of the levels will support further discussion on how they should be applied and what supporting evidence should be provided in each level.

Z. Level Zero. Content distributed as deposited. Unattended deposit-storage-access.

Data content and supporting metadata are stored for a given time period, or indefinitely. This may include multiple copies and monitoring of bitstreams for integrity. Data content and supporting metadata are distributed to users exactly as they are provided by depositors. Beyond these measures, there is no appraisal, curation or long-term preservation.

D. Deposit Compliance

Data content and supporting metadata deposited are checked at the point of deposit for compliance with defined criteria, e.g. data formats, metadata elements, and compliance with legal and ethical norms.[15]

C. Initial Curation

In addition to Level D above, if these criteria are not met the digital objects are curated by the repository to meet the defined criteria. This initial curation for access and use may include, e.g., the correction or enhancement of metadata and/or data content, or the creation of dissemination formats.

B. Logical-Technical Curation

In addition to D and/or C above the repository takes long-term responsibility for ensuring that the data and metadata can be rendered as required by the designated community.

This entails the responsibility for updating hard- and software environments, archival and dissemination formats of digital objects, and metadata in response to the threat of technological obsolescence and/or to accommodate changing needs of the Designated Community.

[14] And potentially for describing curation- and preservation-related information at object-level.

[15] The actions that follow these checks are determined by the repository. For example, a repository may choose to return (meta)data that does not meet the deposit criteria to the depositor, or to ingest the (meta)data and document non-compliance, or to undertake initial curation to ensure compliance.

A. Conceptual preservation for understanding and reuse

In addition to B above, the repository takes long-term responsibility that the data content and metadata can be independently understood by the designated community.

This entails the responsibility for updating the content of metadata elements and other semantic artefacts such as controlled vocabularies and ontologies if necessary. It may include responsibility for editing the structure and content of deposited data, for example in response to changes in legal regulations.

Concluding Thoughts and Next Steps

This revised version of the discussion paper is shared with the community for comment with a view to future iterations and the development of a CoreTrustSeal position paper that may influence future versions of the Requirements. For the 2023-2025 version of the Requirements the current levels of curation (see Appendix) have been retained. In the future integrated curation and preservation levels that have been agreed by the community would provide a valuable reference point for communicating the degree of care a digital object receives and which actors take responsibility for that care. This would then provide insights into how those offering different levels of curation could be assessed and evaluated.

Appendix A: Feedback Received to Version 1.0 of the Discussion Paper[16]

In response to the original proposed level C of curation "C. Basic Compliance and/or curation" one item of feedback[17] noted:

> "There is another possible case, where each data set is peer reviewed at deposit by a field expert, but then no further active preservation happens [...]. So that might be a sub-level of "C", where not only formats, metadata are checked, but it is ensured that the data is really meaningful."

This feedback reflects two important points. The first is that curation and preservation criteria may include some validation of the 'content quality' of resources; this is seen as separate from the 'standards compliance' quality measures often applied by repositories and which are the focus of TDR certification standards such as CoreTrustSeal. The second point is that the proposed level "C" conflates two service scenarios that might be completely separate: setting criteria for accepting or refusing deposits, versus providing curation services to meet a defined set of criteria. This point was also noted in the UKDS analysis and has led to a separation of the proposed levels into levels C and D.

Paul Wheatley from the DPC provided some important input on the need to clarify the purpose of the levels and the degree to which they are prescriptive. Their feedback noted that the examples of formats and format migration provided for Logical-Technical curation disregarded the fact that

> "[u]pdating/changing the environment (e.g. using different rendering/processing/execution/ analysis software) or recreating/packaging the original environment and software (e.g. using an emulation approach) might be equally or indeed more valid in examples precluded."

This overt focus is noted and corrected in the proposed revisions. The DPC feedback similarly had concerns about "strongly steering towards a particular preservation approach", noting the risks of a "process of file format migration / normalisation so that data meets 'compliance'" and of "asking the depositor to perform ad hoc file format migration without any oversight, documentation or evaluation of accuracy".

The purpose of the CoreTrustSeal curation and preservation levels is to define a range of possible service offerings. The CoreTrustSeal and the proposed levels are not intended to mandate a particular approach, as specific preservation methods must be defined in relation to the characteristics of the digital objects and the needs of users among other factors. Level definitions must be clearly differentiated and granular enough to be applied to specific curation and preservation practices. This does not equate to prescribing a specific preservation approach and the text has been revised to clarify this.

[16] See also Recker, Jonas; L'Hours, Hervé and Mari Kleemola (2023). "Modeling curation and preservation levels for trustworthy digital repositories". PV2023 Conference, Geneva. https://indico.cern.ch/event/1188041/contributions/5309462/.

[17] András Holl, Library and Information Centre of the Hungarian Academy of Sciences, member EOSC Association-Long Term Data Preservation Task Force (LTDP-TF) https://eoscsecretariat.eu/eb-profiles/andr%C3%A1s-holl

Appendix B: CoreTrustSeal 2023-2025 Levels of Curation (unchanged from 2019-2022)

“Level of Curation Performed. Select all relevant types from:

- 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”