Modelling curation and preservation levels for trustworthy digital repositories

Dr. Jonas Recker, GESIS – Leibniz Institute for the Social Sciences | Hervé L'Hours, UK Data Archive, UK Data Service, University of Essex | Mari Kleemola, Finnish Social Science Data Archive, Tampere University

Broadly defined digital curation practices ensure the accessibility, usability, and understandability of digital assets for a defined community of users. Preservation practices ensure this continues for as long as necessary. Ideally these practices are managed and performed in sustainable organizational settings with clearly defined responsibilities, are governed by policies, and require that the repository actively addresses any factors – legal, organizational, financial, technical, or cultural – which might put access and use of digital assets at risk.

Which specific practices are required to provide successful digital curation and preservation depends on many factors including the needs and 'knowledge base' of the users, the type of digital object to be preserved, the overall heterogeneity of collections, and (re)use scenarios. There is no 'one-size-fits-all' approach, and the details of implementation will vary. However, community agreement on the definition of the levels of "care provided and the degree of responsibility taken by a repository or other data service"[1] would provide an important reference point for digital objects' depositors, funders and (re)users and for collaborations and partnerships communally offering (meta)data services. Transparency of responsibility, and any changes to the level of care, are critical to networked, federated and interoperable research infrastructures.

There are numerous extant and ongoing efforts to define standards, requirements, and characteristics for repositories.[2] Some of these focus on technological aspects of digital preservation such as the NDSA Levels of Preservation[3], while others include organizational and data management aspects such as the CoreTrustSeal levels of curation and preservation (see below).

Envisioned as a "benchmark against which an organization may evaluate their digital preservation repository's capabilities and plan for further enhancement and growth" towards a

[1] CoreTrustSeal Standards and Certification Board. (2022). Curation & Preservation Levels: CoreTrustSeal Discussion Paper (v01.00). Zenodo. https://doi.org/10.5281/zenodo.6908019, p. 3.

[2] L'Hours, Hervé, & Bell, Darren. (2023). Repository & (meta)data Services Functions & Activities: Crosswalk (v01.00). Zenodo. https://doi.org/10.5281/zenodo.7690658.

[3] National Digital Stewardship Alliance. 2019. "Levels of Preservation (LoP)". https://ndsa.org/publications/levels-of-digital-preservation/

mature preservation program (NDSA 2019), the NDSA approach is intended to support repositories in discussing and improving their current performance.

The focus of the CoreTrustSeal is on certifying 'Trustworthy Digital Repositories' (TDRs), which includes assessing the organizational framework (mission, resources, policies), in addition to digital object management practices and technological infrastructure. To be considered a TDR, CoreTrustSeal requires that the repository offers active preservation. The purpose of the CoreTrustSeal curation and preservation levels is to allow applicants and reviewers to assess quickly if a repository's practices of digital object management put this repository within the "active preservation" remit and thus in scope for CoreTrustSeal certification. As the reach and impact of the CoreTrustSeal has broadened, there have been further questions and a need for clarification of the current levels of curation.

The current CoreTrustSeal curation and preservation levels (see table) assume that:

> "(1) initial deposits are retained unchanged and that edits are only made on copies of those originals, (2) metadata that enables the Designated Community to understand and use the data independently (i.e., without having to consult the original creator) is present at deposit or added by the repository, and (3) ongoing measures for active preservation are in place for the greater part of the collection(s)."[4]

Table: CoreTrustSeal Levels of Curation[5]

Level	**Description**
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 during ingest, enhancement of documentation and metadata
D.	Data-level curation – as in C above, but with additional editing of deposited data

The CoreTrustSeal seeks to provide certification at a community-agreed 'core' level.[6] The CoreTrustSeal Board (appointed from a community of peer reviewers from previously certified repositories) does not seek a final authority role on emerging issues and instead provides an

[4] CoreTrustSeal Standards and Certification Board. (2022). CoreTrustSeal Requirements 2023-2025 (V01.00). Zenodo. https://doi.org/10.5281/zenodo.7051012, p. 7.

[5] Ibid., p. 6.

[6] https://www.coretrustseal.org/about/.

open and transparent consensus model for developing and revising the 16 CoreTrustSeal Requirements. At the last community revision[7] specific questions were asked about the validity and comprehension of the current 'curation levels', including how they relate to preservation strategies and actions.

Performing active digital preservation requires an ongoing, managed effort to maintain the accessibility and usability of the digital resources preserved. Yet the current CoreTrustSeal curation and preservation levels strongly focus on curation actions taken at the ingest stage. And while such actions are an important element of (preparing for) active preservation throughout the lifecycle of a digital object, taken on their own they are not sufficient in this regard.

The feedback received during the community revision was not sufficient to propose an update of 'levels of curation' to the 2023-2025 requirements. Instead, the Board developed a community discussion document[8] proposing an approach that would provide aligned, stepped and contiguous tiers of retention, curation and preservation suitable as high level descriptors for a wide range of repositories and other data services, whether or not they were in scope for CoreTrustSeal Certification (see Appendix 1 for more details on the levels):

- Z. Level Zero. Content distributed as deposited. Unattended deposit-storage-access.
- C. Basic Compliance and/or curation.
- B. Logical-Technical Curation.
- A. Conceptual preservation for understanding and reuse.

This discussion paper received valuable support and feedback, including from the EOSC Association's Long Term Data Preservation Task Force (LTDP-TF)[9], Digital Preservation Coalition (DPC), and an internal review by the UK Data Archive[10]. The revised tiers and extended proposals below reflect this feedback and will be subject to review by the CoreTrustSeal Board and further public consultation.

In response to the proposed level C of curation "C. Basic Compliance and/or curation" one item of feedback[11] 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

[7] https://www.coretrustseal.org/why-certification/meeting-community-needs/

[8] CoreTrustSeal Standards and Certification Board. (2022). Curation & Preservation Levels: CoreTrustSeal Discussion Paper (v01.00). Zenodo. https://doi.org/10.5281/zenodo.6908019.

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

[10] 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.

[11] András Holl,Library and Information Centre of the Hungarian Academy of Sciences, member LTDP-TF https://eoscsecretariat.eu/eb-profiles/andr%C3%A1s-holl

> 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[12]. 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 important input on the need to clarify the purpose of the levels and the degree to which they are prescriptive. This feedback noted that in the examples of formats and format migration provided for Logical-Technical curation

> “the missing factor here is the technical environment in which the data is used. Updating/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”[13]

This overt focus is noted and corrected in the proposed revision below. Paul similarly had concerns about “strongly steering towards a particular preservation approach”, in particular “if they codify file format normalisation”, 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”.[14]

The CoreTrustSeal sees a broad range of applications including a wide variety of deposit criteria and initial curation services as well as different approaches to technical and conceptual preservation. Applicants justify their approaches in terms of preservation goals and in terms of meeting their users’ needs. Any decision to take action on (file) formats at any point in the lifecycle should demonstrate awareness of the balance of risks and benefits in place. The purpose of these levels is to define a range of possible service offerings, not to mandate a

[12] Also see Lacagnina, Carlo et al. (2023). TOWARDS A DATA QUALITY FRAMEWORK FOR EOSC (1.0.0). Zenodo. https://doi.org/10.5281/zenodo.7515816 for a discussion on data quality assessment and standards.

[13] https://docs.google.com/document/d/1UlT1BHconkuuNQNKOQVhlIX7zu2RUnKN70_uOB1waYo/edit?usp=sharing.

[14] Also see Wheatley, Paul. 2022. “File format recommendations - I wouldn’t say they are unacceptable, but I wouldn’t recommend them either.” DPC Blog. https://www.dpconline.org/blog/file-format-recommendations for a more in depth discussion of the raised issues.

particular preservation approach; any repository or CoreTrustSeal applicant would specify their own deposit, curation and preservation criteria.

It may be helpful to use the example provided by the Long Term Preservation Task Force that differentiates between preservation *outcomes*, *actions* and *systems*. A repository system that *only* sets criteria on deposit and/or initial curation with no longer-term undertaking to take actions *if* necessary,is not offering active preservation. Those repository systems that offer active preservation are providing a service that monitors the need for action; if no action is required to achieve preservation outcomes, then no action needs to be taken.

Level definitions must be specific enough to determine whether they correspond to a given set of curation and preservation practices. This does not equate to prescribing a specific preservation approach as this needs to be considered in relation to the characteristics of the digital objects to be preserved and the needs of users among other factors.

Many repositories have heterogeneous collections and may choose different curation and preservation approaches for groups of assets based on archival value, mission, in-house expertise, available resources, etc. This fact presents a challenge; identifying the different levels of retention, initial curation and active preservation offered by a repository does not communicate what level of care a specific digital object is receiving. This factor, as raised by the UKDS feedback, would be important to address in future.

Revised Curation and Preservation Levels

The tiers of curation and preservation below provide a basis for describing repository service levels and curation- and preservation-related information recorded at object-level as suggested by the UKDS feedback. As service levels, the tiers are cumulative as they progress from D to A. A repository may offer a service that stops at a particular tier, or offer different tiers for different collections of digital objects.

As object-level information, the tiers are distinct but contiguous, enabling repositories to describe and document the care a given object in a collection has received, thereby contributing to documented audit trails.

From the perspective of the CoreTrustSeal, Levels Z, D and C 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, 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 distributed to users exactly as they are provided by depositors. 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[15]

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.

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.

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.

Conclusion

A clear and concise set of tiered curation and preservation levels supports all organisations holding data or metadata as a part of their service provision. The levels also provide a graduated reference point for repositories planning and developing their capacity (including preservation capacity) for different digital objects in their collections. The certification of trustworthy digital repositories such as that offered by the CoreTrustSeal requires the definition of a minimum level of curation and preservation that repositories must have reached to be in scope. A repository may apply different retention, curation and/or preservation levels to different objects. Clarity on

[15] As an element of a service level, checks at deposit compliance (D) could result in the digital objects being returned to the depositor for change and resubmission if the criteria are not met. If the repository also provides initial curation (C) curators may make the amendments on behalf of the depositor following the checks. Both deposit compliance checking, and initial curation can be essential services from repositories that then provide the additional tiers of active curation (B and A).

these levels at the repository and the object level provides transparency to service funders and users, and between service providers; they also provide helpful instruments in the development of preservation policies and practices across heterogeneous collections.

Clear service levels applied across communities of practice provide a valuable reference point for assessment but are not sufficient in themselves to ensure effective outcomes for digital objects or to confer trustworthy digital repository status. Exactly which practices can be considered as adequate for successful preservation strongly depends on the individual, often complex conditions in which a repository operates. These practices will vary depending on whether the services offered are generic, or specialist (e.g. disciplinary). The definition of further supporting information for each level (including deposit criteria, curation standards and preservation plans) and their inclusion in repository registry information has the potential to streamline and partially automate assessment and certification processes.

Beyond certification the application of equivalent metadata and linked information at the object level would provide transparency at the point of reuse at the level of care data and metadata are receiving. The alignment of digital object characteristics (received level of care) with repository service offerings could provide a rich resource for analysis if applied using semantic web and graph technologies.

Appendix 1: First Draft Example of Tiered Curation and Preservation

All of the levels below are options in real-world appraisal decision-making. Levels Z and C 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. Agreement on these levels will support further discussion on how they should be applied and how supporting evidence should be provided.

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

> Data content and supporting metadata are distributed to users exactly as they are provided by depositors. No curation or long term preservation.

C. Basic Compliance and/or curation

> Data content and supporting metadata deposited are checked at the point of deposit for compliance with defined criteria for data formats and metadata elements. If these criteria are not met the digital objects are returned to the depositor for change, or the repository undertakes the necessary curation steps to ensure they comply. Minimal curation for initial access and use, but no long term preservation.

B. Logical-Technical Curation

> In addition to C above the repository takes long-term responsibility for ensuring that the data and metadata are updated over time to newer standards and formats in response to:
>
> > i. technical risks (e.g. file format obsolescence) and/or
> >
> > ii. the changing needs of the designated community (e.g. newer alternate formats become necessary for reuse).

A. Conceptual preservation for understanding and reuse

> In addition to B and C above the repository monitors changes to the definition and demands of their designated community, including their knowledge base, and takes responsibility for the preservation actions that ensure digital objects can be understood and re-used. Usually this will involve updates to the content of metadata elements and other semantic artifacts such as controlled vocabularies and ontologies. For some repositories it may include responsibility for editing the structure and content of deposited data.