EOSC Preservation:
Overview Discussion Paper

Long Term Data Preservation Task Force
December 2022

EOSC Association AISBL

Rue du Luxembourg 3, BE-1000 Brussels, Belgium
+32 2 537 73 18 | info@eosc.eu | www.eosc.eu
Reg. number: 0755 723 931 | VAT number: BE0755 723 931

Table of contents

EOSC Association & Task Forces

The EOSC Association (EOSC-A)[1] Advisory Group (AG)[2] “Sustaining EOSC”[3] contains two task forces designed to work on sustaining the European Open Science Cloud as it scales: the Long-Term Data Preservation Task Force (LTDP-TF)[4], and the Financial Sustainability TF. Bob Jones is the liaison to the EOSC Association Board of Directors for both TFs.

The LTDP-TF examines several interrelated areas with the objective to:

- Create a shared understanding and vision for sustainable preservation in EOSC.
- Agree on the functions of repositories and data services necessary to create, store, curate, preserve, and engage with data (re)users.
- Map and promote the roles, responsibilities and accountability of the preservation and related actors within EOSC, including the financial aspects and possible business models.
- Provide recommendations on engagement through a European Trusted Digital Repository (TDR) network to mature preservation and FAIR enabling practices across disciplines and geographies.

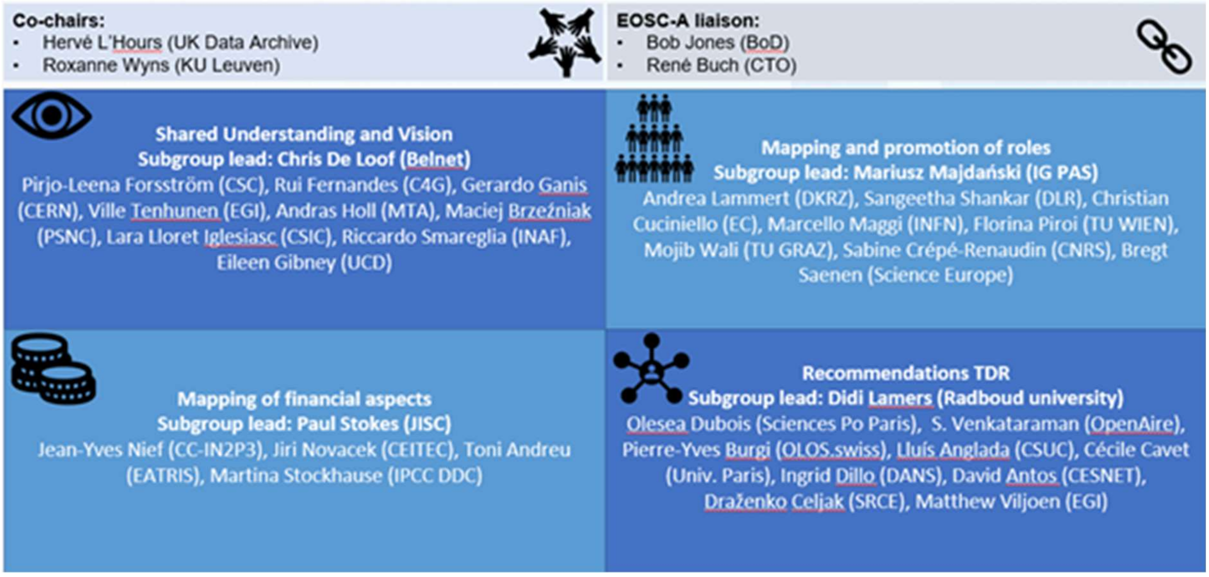
The LTDP TF work is taken forward by four subtasks covering vision, roles, finance and trust network. The subtask leads coordinate their area of work and provide updates at the periodic LTDP TF meetings. TF meetings provide an opportunity for subtask members and task force co-chairs to share their work and align with the vision. Insight into the ongoing work of the EOSC-A is provided through the attendance of Bob Jones (Board liaison) and René Buch (EOSC-A CTO).

[1] https://eosc.eu/

[2] https://eosc.eu/advisory-groups

[3] https://eosc.eu/sustaining-eosc

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



Co-chairs:

- Hervé L'Hours (UK Data Archive)
- Roxanne Wyns (KU Leuven)

EOSC-A liaison:

- Bob Jones (BoD)
- René Buch (CTO)

Shared Understanding and Vision
Subgroup lead: Chris De Loof (Belnet)
Pirjo-Leena Forsström (CSC), Rui Fernandes (C4G), Gerardo Ganis (CERN), Ville Tenhunen (EGI), Andras Holl (MTA), Maciej Brzeźniak (PSNC), Lara Lloret Iglesiasc (CSIC), Riccardo Smareglia (INAF), Eileen Gibney (UCD)

Mapping and promotion of roles
Subgroup lead: Mariusz Majdański (IG PAS)
Andrea Lammert (DKRZ), Sangeetha Shankar (DLR), Christian Cuciniello (EC), Marcello Maggi (INFN), Florina Piroi (TU WIEN), Mojib Wali (TU GRAZ), Sabine Crépé-Renaudin (CNRS), Bregt Saenen (Science Europe)

Mapping of financial aspects
Subgroup lead: Paul Stokes (JISC)
Jean-Yves Nief (CC-IN2P3), Jiri Novacek (CEITEC), Toni Andreu (EATRIS), Martina Stockhause (IPCC DDC)

Recommendations TDR
Subgroup lead: Didi Lamers (Radboud university)
Olesea Dubois (Sciences Po Paris), S. Venkataraman (OpenAire), Pierre-Yves Burgi (OLOS.swiss), Lluís Anglada (CSUC), Cécile Cavet (Univ. Paris), Ingrid Dillo (DANS), David Antos (CESNET), Draženko Celjak (SRCE), Matthew Viljoen (EGI)

Subtask leads and members

Other EOSC task forces cover aspects that interact with and impact the LTDP-TF including human (e.g. TF Data Stewardship and Curricula), technological (e.g. TF FAIR Metrics and Data Quality) and financial (e.g. TF Financial Sustainability) dimensions. The TF also tracks relevant European and international projects and initiatives identified as part of the ongoing stakeholder mapping exercise. Presentations from relevant stakeholders take place at TF and subtask level. This engagement is crucial to achieving an overview of LTDP initiatives and provides a discussion forum for the challenges remaining and the recommendations they imply.

The LTDP-TF will prepare recommendations on actions and policy for the EOSC-A and the European Commission. The TF will share intermediate documents for consultation and feedback by the wider community with final recommendations expected in April 2023.

Preservation in the Context of EOSC and FAIR

A simplified overview of preservation in the EOSC context can be defined in terms of the outcomes, systems and actions related to any digital object:

> **Preservation Outcomes**: these concern digital objects that, having been curated for FAIRness and other desirable characteristics, are *maintained* to retain those characteristics for as long as necessary. These outcomes are targeted at a defined and 'designated community' of users.
>
> **Preservation Systems**: accept the deposit of digital objects for storage and access (making them 'repositories' in the broad sense) and also curate them with a long term FAIR-enabling perspective on the objects and their designated community (making them candidate to become 'trustworthy digital repositories'). The designated community's knowledge base and technological needs are understood and monitored over time. These systems have sufficient resources, including personnel and financial resources, to be sustainable in terms of their organisational, technology and security infrastructure.

These systems may involve multiple partners undertaking different roles across storage, curation, preservation, discovery and access etc. Defining the scope of responsibilities is critical to defining which entity is a candidate to become a trustworthy digital repository.

> **Preservation Actions**: are the changes to digital objects' (metadata and data) that are intended to keep them FAIR over time. These actions include technical steps such as emulation or transformations to modern, in-demand, file formats and metadata schemas; but also, actions to ensure that the conceptual content of data and metadata, including semantic artefacts such as ontologies, continues to be understood and re-usable.

Data and metadata *storage* to agreed standards and the *curation* of digital objects to meet agreed criteria are critical foundations for preservation but alone, these cannot ensure preservation. Preservation *outcomes* depend on *actions* by *systems* with a sustainable, long-term perspective and repositories that meet all these requirements may be candidates for assessment and certification as Trustworthy Digital Repositories (TDR).

Note that despite the focus on trustworthy repositories and FAIR digital objects, there are no strict criteria related to these for engagement with EOSC. Both trust and FAIR are an ongoing journey for many repositories and digital objects: transparency over current levels of FAIRness or trustworthiness are important to identify the current status and to plan for improvement. However, many less FAIR digital objects retain 'value' to their designated communities and may need to be preserved even if circumstances or resources do not allow for increased FAIRness. In general, assessment of value is important when making choices about investment, but definitions and assumptions will vary across communities.

Repositories may target *generalist* designated communities with commonly used file formats[5] and basic metadata. Designated communities which include researchers in a specialist discipline or domain may also need less common data formats and more complex metadata for relevant digital objects to be found, accessed, interoperated with and reused.

One 'best case' scenario, reflected in some recommendations[6,7] might be that all digital objects should be preserved for the long-term in a certified disciplinary or domain TDR. But limitations on resources (number of TDRs, skills, technical and financial) all require a more nuanced analysis of the current situation, future vision and a roadmap to reach our goals.

This initial discussion paper from the LTDP TF seeks community input across these issues around the vision, roles, finance and networks needed to deliver preservation for FAIR digital objects in the EOSC. The task force final recommendations to stakeholders will be targeted at the European, national and institutional level.

Approach

As described in the section 'EOSC Association & Task Forces', the work of the TF is organised in four subtasks: vision, finance, roles and trust networks.

> **Integrated vision of LTDP in the context of EOSC**: Initial baseline of visions and assumptions to be iteratively integrated with the outcomes of the subtasks.
>
> **Finance**: the finance subtask looks into the costs of preservation, which are strongly linked to human and other resources and mapped to processes/functions, outcomes (object characteristics) and time (e.g through guaranteed retention or preservation periods). This topic is strongly related to and affected by roles.
>
> **Roles**: the roles subtask seeks to identify the types of actors necessary to ensure that preservation is delivered by people with appropriate skills. The challenge is to address the ongoing debate on terms and boundaries around roles, and their responsibilities for processes, functions, outcomes, at an appropriately granular level, to ensure that preservation is supported by sufficient human resources. Fulfilling these roles has a clear mapping to the cost-centres addressed by the finance subtask. The permissions, prohibitions, obligations, duties and constraints[8] that govern roles are critical to accountability and to machine-actionable support.

[5] File formats are referenced for simplicity. The benefits and challenges of linked open data through serialisations such as RDF remain in scope for this task force.

[6] "It is always recommended to refer to broadly recognised discipline-specific or certified repositories" Practical Guide to the International Alignment of Research Data Management - Extended Edition https://doi.org/10.5281/zenodo.4915862

[7] "curated in trusted domain repositories whenever possible" Commitment Statement to Enabling FAIR Data in the Earth, Space, and Environmental Sciences https://doi.org/10.5281/zenodo.1451971

[8] Cf: Open Digital Rights Language (ODRL) Ontology https://www.w3.org/ns/odrl/2/ODRL20.html

TDR Network: A common vision on the concepts of trustworthy repositories versus repositories versus "trustworthy" data services needs to be developed. The scope and purpose of a European network of TDRs need to be clarified, as well as its audience and membership requirements and management.

The subtasks follow parallel but aligned processes which include:

- desk research to identify relevant reports and project outputs
- draft recommendations
- open consultation (webinars/workshops)
- revised recommendations taking into account feedback from the consultation

The work occurs alongside a range of engagements with and presentations from relevant stakeholders at the TF and subtask level. In seeking to deliver a LTDP vision and common understanding that integrates each of the subtask topics, the task force will address:

An overview of definitions and glossaries: desk research on current LTDP and related terminology in which we look for and refer to those definitions that are most closely aligned to LTDP in the context of EOSC. Where possible, we engage with preservation communities and provide suggestions on the improvement and refinement of those definitions[9].

A context and position paper: developing a problem statement to scope the higher level who, what, when, why and how of preservation within the EOSC as a direction and alignment of the work and interaction between the subtasks.

Understanding the current status: Identifying the 'As Is' situation including the identification of relevant stakeholders, projects and papers. In-scope areas include the European and national landscapes, the current state-of-the-art, norms, standards and existing resources.

Gap analysis: identification of key gaps and issues for LTDP in the context of EOSC

Recommendations: identifying the 'To Be' (ideal target scenario) and an action plan of desirable steps, improvements and timelines to reach it.

The LTDP TF is presented with the challenge of addressing the complex issue of long-term preservation with the EOSC and designing clear, implementable recommendations at the European, national and institutional levels.

Vision in Progress

As noted in the *Preservation in the Context of EOSC and FAIR* section above, the ideal scenario *might* be that all digital objects of interest to EOSC researchers should be preserved for the

[9] E.g. CODATA RDMT Public Review 2022 https://codata.org/initiatives/data-science-and-stewardship/rdm-terminology-wg/

long-term in a certified domain TDR. However, on the reasonable assumption[10] that infinite resources are not available to secure this outcome, the emergent vision and direction from the TF must be more nuanced.

A more granular understanding of the situation to support a more refined vision depends on several key issues. These include:

> **The range of desirable 'outcomes' for EOSC digital objects**. The TF resources and timescales mean that digital objects will be addressed at a broad level. 'Digital objects' are not limited to 'research data', they extend to include metadata, software, semantic artefacts, publications, metadata related to physical samples, standards and schemas, in fact any digital objects that are of ongoing importance to researchers (and by extension to funders, policy makers and the public at large). The resources and timescales of the TF mean that these different types of digital object cannot all be addressed individually and in detail, but it should be understood that they are all in scope for preservation. Defining and obtaining these desirable outcomes implies that the digital objects are known (have an identity and location so they can be found and accessed), and can be acted upon (have associated rights). Deciding the appropriate outcome implies a process of appraisal and selection against some agreed criteria for 'value'. Outcomes could include retention (implying a need for effective storage), short-term curation to meet agreed criteria (e.g. FAIRness) or active long-term preservation at a generalist or specialist level. Preservation outcomes may also be influenced by other aims including the enabling of Open Data and Research, the provision of information security for sensitive data, or alignment with other principles such as CARE[11].
>
> **The range of systems that deliver the desirable outcomes**. All data and metadata-related services depend on effective storage. A wide range of organisations and partnerships organised on an institutional, national and European level work to deliver data and metadata services across the research lifecycle and between parts of the EOSC. However, the levels of expertise and the curation services provided differ across repositories. Not all repositories offer active long-term preservation and many long-term repositories have not achieved CoreTrustSeal or other certification[12.]
>
> **The range of actions within the systems.** *Systems* deliver *outcomes* through their *actions*. The TF has identified a need to align the work on Roles (and their associated responsibilities) and on Finance. The actions, or groups of actions collated into

[10] Based on discussions of the TF with different stakeholder groups such as COAR (https://www.coar-repositories.org/).

[11] https://datascience.codata.org/articles/10.5334/dsj-2020-043/

[12] CoreTrustSeal provides certification at the Core level, Nestor seal (based on DIN3166416) provides Extended certification, and ISO1636317 provides Formal certification.

> workflows, standard operating procedures and functions, require roles to fulfil them and imply a need for specific funding. Early work on the Finance subtask noted that if these functions and processes "*can be defined and widely understood (with some flexibility) then we have the starting point to make cost estimations comparable*". The costs of human resources, and the availability of skills implies that finance and roles would benefit from a common view of actions and functions.

Initial discussions within the Roles subtask identified the need to scope activities in terms of current and future digital objects. The development and promotion of pan-European rules for good practice around data being developed now, accompanied by clear practice for each research field (including recommended formats[13],[14]), and training has the potential to deliver immediate benefits. For older, existing data, retrospective appraisal may identify cases (e.g. physical experiments) where it is possible, or even more practical to gather the data again, than to curate them to modern standards, particularly in cases where researchers and technicians involved in the original production are not available. Reproducing the data may be cheaper, more accurate and less subject to historical rights issues. In other cases, these older digital objects may be irreplaceable and of sufficient value to warrant bringing them up to modern standards.

The broad identification of these wider contextual issues will set the stage for defining preservation systems, actions and outcomes including the finance issues and roles necessary to address past and future digital objects. A TDR undertakes many activities, and therefore requires roles, and incurs costs that are common to all data services and therefore not explicitly or exclusively 'costs of preservation'. These may include storage, technical infrastructure, information security measures, deposit and access services, or resource discovery systems.

Iterative discussions across the topics of finance, roles and the overall vision will inform and be informed by the emerging specifications of a network of trustworthy repositories and data services capable of representing curation and preservation actors in the wider landscape of research and research data management.

With an understanding of the wider actions, systems and outcomes we may be able to set our vision for preservation in a more specific real-world context. For example:

> Digital objects that act as inputs to, or outputs from, research are identified, findable and accessible in environments that support good storage practice. These objects are subject to appraisal, and reappraisal over time, to assess their value, their impact and the associated costs, risks and benefits. Ongoing appraisal informs the level of investment in the retention, curation and long-term preservation of digital objects. The

[13] https://www.loc.gov/preservation/resources/rfs/
[14] https://openpreservation.org/news/new-community-resource-international-comparison-of-recommended-file-formats/

levels of care, and changes to levels of care, provided by repositories and assigned to digital objects are transparent to (meta)data funders, depositors and users.

The EOSC LTDP TF welcomes wider community input to these ongoing proposals and discussions. The eventual recommendations for outcomes, systems and actions will target European, National and institutional audiences. The sections below briefly outline ongoing topics and issues within the subtasks of the TF.

Roles

It is reasonable to expect that the subtask on roles will face a number of iterations during the timeline of the TF as the scope, focus and granularity are developed. These roles must support proposals for specific solutions, good practices, training and development of processes that match the specific needs of repositories or scientific organisations. In addition to existing gaps in understanding of the roles of various actors, there are problems with identifying and financing those with the skills to take responsibility for digital objects' curation and preservation. As organisations update their procedures and practices to make their digital objects more FAIR they may need to identify and fill newer roles. Direct curatorial and preservation roles also need to be supported by specialised training roles that provide both general and specialist guidance. As noted above, the specific roles may be influenced by whether they address newly created digital objects or take account of the legacy of historical digital objects.

We acknowledge that there are different actors that provide and use digital objects, and those who manage them. These stakeholder groups and their associated roles must all be identified and addressed. There may be general agreement that data and metadata should be FAIR, or meet other criteria such as being Open, to ensure trust in science and provide positive outcomes for society. It may be generally agreed that digital objects of value should be preserved to maintain their value over time, but this must take into account a wide variety of different standards and expectations that impact costs and the roles required.

For example, digital object creators (including scientists and technicians providing 'raw' data) may feel that meeting standards such as FAIR or providing data and metadata that are 'preservation ready' is an imposition on their already limited time. Digital objects' owners (including research organisations responsible for data and metadata) know that additional responsibilities for data creators, hiring and training new technicians and data stewards, or upgrading storage and infrastructure, all incur costs.

Some of these additional costs will be unavoidable but they will be mitigated by guidance that supports a coherent and strategy-led approach to policy-making on research infrastructures in Europe. Such guidance would also provide a clearer focus on the roles needed to support these activities.

The roles required will be updated over time as the range of 'in-scope' outcomes, systems and actions are identified. This will include the different roles required at institutional, national and European level such as repository actors and preservation professionals, but may also depend on other actors across the research lifecycle.

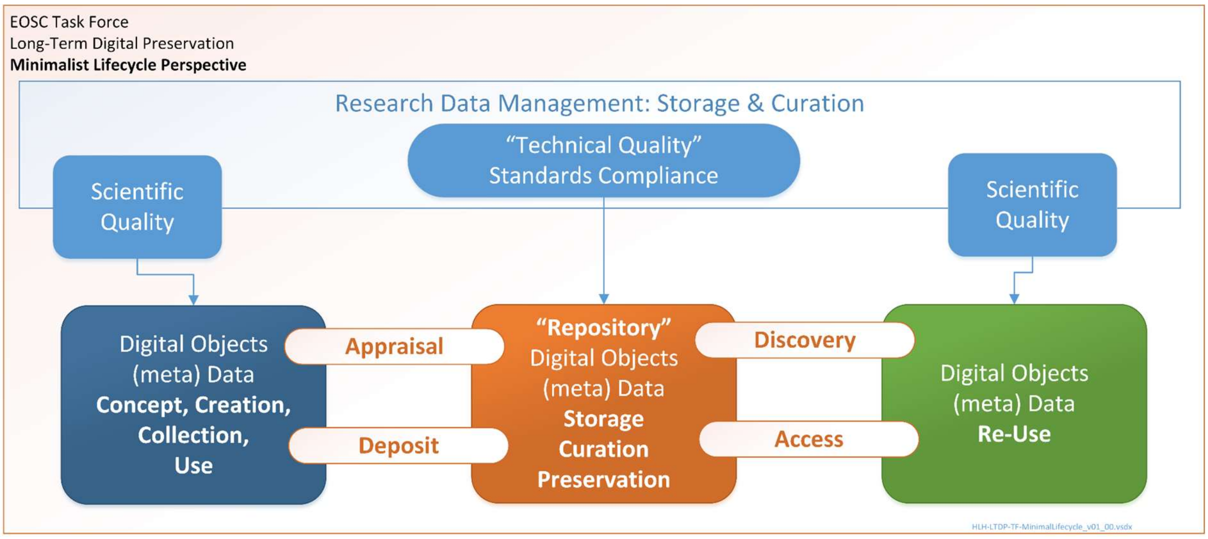

EOSC Task Force
Long-Term Digital Preservation
Minimalist Lifecycle Perspective
Research Data Management: Storage & Curation
"Technical Quality" Standards Compliance
Scientific Quality
Scientific Quality
Digital Objects (meta) Data **Concept, Creation, Collection, Use**
Appraisal
Deposit
"Repository" Digital Objects (meta) Data Storage Curation Preservation
Discovery
Access
Digital Objects (meta) Data **Re-Use**
HLH-LTDP-TF-MinimalLifecycle_v01_00.vsdx

Diagram: Simplified Lifecycle Overview

We expect that defining a common set of key functions that are in scope for preservation and other data services will help focus the assignment of roles and the associated definition of the human and financial costs they imply.

Finance

Mapping of the financial aspects of long-term data preservation, including the cost of digital preservation, financial responsibilities of the different stakeholders on an institutional, national and European level, and possible business models for repositories is necessary to deliver the systems that ensure preservation outcomes. This subtask is examining financial issues facing LTDP and what can be done to mitigate or solve them. The goal is not to build new cost models, but to collate and examine the available options and how these can be incorporated into an overall research infrastructure vision that preserves data and metadata for the long-term. Outputs from past projects such as the Curation Costs Exchange[15] allows semi-standardised addition of costs and comparison with peers but more recent work on preservation-specific aspects of costs are limited. It is important that work on the costs of (meta)data are aligned with an understanding of the value of digital objects.

An early assumption might be that costs increase as they progress through from effective storage to basic curation, and eventually active preservation. Curation and preservation may

[15] curationexchange.org

become more costly as they seek to comply with more detailed and specific domain or disciplinary standards; but if these standards are not met then the potential value of the digital objects may not be realised.

Only a portion of the *total cost* of a TDR is a is preservation-specific. The other side of this equation is a cost/benefit analysis of inaction. What are the costs to the community as a whole if a digital asset that has the potential for reuse is not cared for in a way that enables its value to be realised?

Agreement on the wider functions of repositories and data services is necessary to define a subset of ‘preservation costs’. Without this we cannot effectively calculate the human resource costs of preservation. Another argument for iterative alignment with the roles subtask is that the beneficiaries (the designated community accessing and using preserved data for the public benefit) may not be the same actors who have invested in the creation and preservation of digital objects.

Preservation actions that deliver preservation outcomes depend on the ongoing sustainability of the repository systems that care for them. Short-term project funding is not an effective basis for sustainable long-term systems and most preservation costs will be incurred after completion of projects. A full lifecycle perspective is helpful as a failure to invest in quality, FAIRness or other criteria among one group of actors or processes will simply transfer, and potentially increase, that cost to another part of the system.

Storage and other resources are not infinite so some process of appraisal and reappraisal over time is critical, but this will incur its own costs. Practices will differ between generalist and more specialist repositories, but there is room for cooperation on many aspects of engagement with designated communities of users. Different models for the centralisation of functions, group (collective) bargaining/negotiation for shared services, or the federation of existing services must be examined for their wider impact including interoperability and expertise at the repository and object level. Identifying areas with the potential for automation at scale must be balanced with the need to maintain the resources of human expertise. Shared registries of information may also provide a method of sharing costs.

Identifying the likely costs of preservation, at a level of granularity sufficient to support strategic planning and implementation, is challenging. Even more challenging can be to identify the costs of inaction and the risks incurred. The value of digital objects as digital assets is the basis for making a credible business case[16] for their preservation. If early investment reduces cost over time this presents an argument for targeted investment.

[16] https://jisc.onlinesurveys.ac.uk/cost-of-data-loss

Network of Trust

Achieving a future vision for long-term preservation within the EOSC, supported by appropriate financial investment and roles with appropriate skills, can only be achieved through ongoing engagement with curation and preservation professionals. A number of groups and communities already exist, but the EOSC Association has acknowledged the need for a body that integrates the expertise of and provides a voice for trustworthy digital repositories.

The subtask on creating a network of trustworthy digital repositories is reviewing the relevant stakeholders and current initiatives. The purpose of a future network will be considered, including how certification standards and the FAIR Principles might be aligned.

The possible scope, activities and obligations of the network will be examined alongside the requirements for repositories to join the network, onboarding, monitoring continued compliance with membership criteria and the promotion of broad geographical and disciplinary inclusion.

Topics for inclusion in the recommendations include:

- Functional aspects e.g. training and support for repositories, knowledge exchange, recommendations, standards, etc.
- Lobby functions that provide a forum and a “voice” for repositories to be heard by stakeholders in the research data management community, policy makers and funders.
- Technical elements e.g. federating a network of technically coupled repositories through EOSC.
- Gaps and practical solutions to overcome them in the short and longer term.
- The role of the EOSC-A.
- Phases of development and scaling the network.
- Ongoing sustainable funding for the network.

Appendix: Definitions, Concepts and Discussion Points

This appendix is extracted from task force work in progress that seeks to identify current definitions and raise key concepts and discussion points. It will be expanded to address FAIR digital objects, the differentiation of preservation and other (meta)data services, and concepts such as appraisal and research artefacts over time.

Later discussion papers and recommendations from the task force will formalise and standardise the use of definitions and concepts for consistency.

General LTDP Definitions

Curation

The CASRAI Dictionary and the current proposed CODATA RDMT that derives from it provide the following definition for Curation:

> The activity of managing and promoting the use of data from their point of creation to ensure that they are fit for contemporary purpose and available for discovery and reuse. For dynamic datasets this may mean continuous enrichment or updating to keep them fit for purpose. Higher levels of curation will also involve links with annotation and with other published materials.

And an overlapping definition of Data Curation (which could equally apply to metadata curation).

> A managed process, throughout the data lifecycle, by which data/data collections are cleansed, documented, standardised, formatted and inter-related. This includes versioning data, or forming a new collection from several data sources, annotating with metadata, adding codes to raw data (e.g., classifying a galaxy image with a galaxy type such as “spiral”). Higher levels of curation involve maintaining links with annotation and with other published materials. Thus a dataset may include a citation link to publication whose analysis was based on the data. The goal of curation is to manage and promote the use of data from its point of creation to ensure it is fit for contemporary purpose and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Special forms of curation may be available in data repositories. The data curation process itself must be documented as part of curation. Thus curation and provenance are highly related.

Data management infrastructure (https://casrai.org/rdm-glossary/)

An infrastructure used to provide data management and enforce data management policies. A data management infrastructure should include resources such as a **data repository** and an information catalogue.

Repository (https://casrai.org/rdm-glossary/)

A repository **preserves**, manages, and provides access to many types of digital materials in a variety of formats.

Data repository (https://www.rdc-drc.ca/glossary/original-rdc-glossary/)

A data repository is an archival service providing long-term care for digital objects with research value. *The standard for such repositories is the Open Archival Information System (OAIS) reference model (ISO 14721:2003).*

Designated Community (OAIS reference model (ISO 14721:2003)

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.

Trusted Digital Repository (https://www.dpconline.org/handbook/glossary)

A trusted digital repository has been defined as having "a mission to provide reliable, **long-term access** to managed digital resources to its designated community, now and into the future". The TDR must include the following seven attributes: compliance with the reference model for an Open Archival Information System (OAIS), administrative responsibility, organisational viability, financial sustainability, technological and procedural suitability, system security, and procedural accountability. The concept has been an important one particularly in relation to certification of digital repositories.

Digital Preservation (https://www.dpconline.org/handbook/glossary)

Data / Digital Preservation refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary. It refers to all of the actions (aka "digital curation") required to maintain access to digital materials beyond the limits of media failure or technological and organisational change.

Long-Term Data/Digital Preservation (https://www.dpconline.org/handbook/glossary)

The Long-Term Data / Digital Preservation is an activity that guarantees continued access to digital materials, or at least to the information contained in them, indefinitely. Medium-term data preservation is the continued access to digital materials beyond changes in technology for a defined period of time but not indefinitely.

Key Concepts and Topics

The items below reflect topics of internal discussion within the task force.

Digital vs Data. The TF name refers to 'data'. LTDP or trustworthy repository acronym (TDR) may use either 'data' or 'digital'. The TF scope covers 'digital objects' (see below).

Digital: any information encoded as zeros and ones. In the TF context it is necessary to differentiate digital/data repositories from non-digital repositories. For LTDP or TDR the non-digital holdings of a gallery, library, museum or blood-sample banks are beyond scope, except in as far as they are described digitally through data and metadata.

Data: the definition and boundaries of what should be referred to as data can be endlessly debated. Discussions include typologies of data as the outcome of collection or creation, variations on the interpretations of raw vs processed data, and questions over whether other artefacts (e.g. software) should be treated as 'data'. In this TF context and at this level it is clearer and more inclusive to use the term: digital objects.

Digital Objects: a collation of data and metadata. Different actors will place different 'boundaries' around data and metadata to define different 'objects'. An object might be easily defined as a collection of data and metadata files in a single directory, but might equally be geographically distributed. With linked data (RDF), the boundaries become even less defined. Objects may be linked to other related objects, may have different versions, or may depend on other digital objects (e.g. semantic artefacts such as ontologies). Different researchers and different archives will have different perspectives on what constitutes 'a digital object'. But when a group of objects is cared for together by an organisation it may be described as a 'collection' and that organisation may be described as a 'repository' (see below). A repository collection may contain objects under different levels of curation and preservation. Object metadata does not come with a 'level of curation/preservation' attached that would allow us to identify its current level of care, or to highlight when that level of care changes.

Metadata: any data about data, but often used in the broad sense of structured metadata defined by e.g. schemas in XML. When discussing 'metadata' it can be helpful to define its purpose: identification, description, administration, provenance, technical etc. Metadata and the schemas and ontologies must also be FAIR and require preservation.

Repository: any entity that accepts, stores and provides appropriate access to objects (including digital objects) may be described as a 'repository'. Any data store, business records management system, library, museum, archive etc.

Trust vs Trustworthy. Trust (confidence and belief in the reliability or truth of something) is offered, accepted, sought and earned between parties. When we choose to 'trust', we may turn out to be right or wrong. But **Trustworthiness** in the TDR (see below) context is about **demonstrating** that practices meet a set of standard requirements through an assessment process.

A Trustworthy Digital Repository (TDR) is more specifically and narrowly defined than a 'repository'. There are several TDR standards[17,18,19] all of which inherit key concepts from the OAIS reference model[20] ISO standard which defines a conformant OAIS as having certain 'mandatory responsibilities[21]. The terms Trusted Digital Repository and Trustworthy Digital

[17] CoreTrustSeal https://doi.org/10.5281/zenodo.3632533

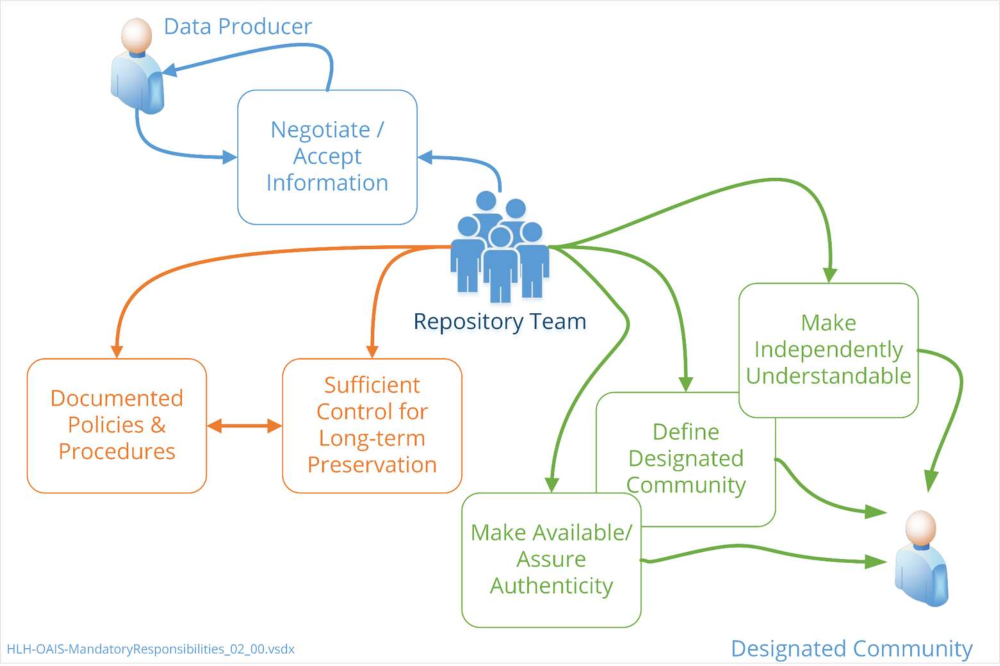
[18] Nestor Seal https://www.langzeitarchivierung.de/Webs/nestor/EN/Zertifizierung/nestor_Siegel/siegel.html

[19] ISO16363 https://public.ccsds.org/pubs/652x0m1.pdf

[20] https://public.ccsds.org/pubs/650x0m2.pdf

[21] *FAIR + Time: Preservation for a Designated Community* covers this for an audience that is less familiar with OAIS and TDR concepts. https://doi.org/10.5281/zenodo.5797776

Repository are often used interchangeably, but with the 'trust vs trustworthy' explanation above in mind it is clear that a 'trustworthy digital repository' is not just "a place containing digital objects that I trust".

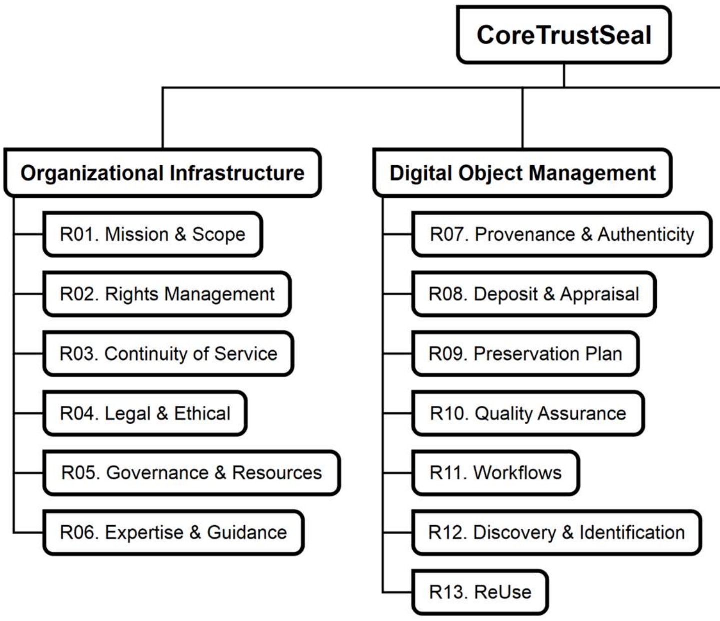

Data Producer
Negotiate / Accept Information
Repository Team
Documented Policies & Procedures
Sufficient Control for Long-term Preservation
Make Independently Understandable
Define Designated Community
Make Available/ Assure Authenticity
Designated Community
HLH-OAIS-MandatoryResponsibilities_02_00.vsdx

CoreTrustSeal was developed through an RDA working group and is an RDA endorsed output that provides extended guidance on its 16 requirements. Conversational CoreTrustSeal[22] provides a simplified (and unofficial) overview. The CoreTrustSeal seeks to provide a community agreed 'core' level of expectations for organisational infrastructure, digital object management and technology/security that together identify a repository as being trustworthy.

In the LTDP TF context we must use "TDR" in this formal sense. A TDR shares a large number of characteristics with other types of organisation and other data services. Any organisation responsible for digital objects may need to address changes in community, community needs, or technology at any time. What sets a TDR apart is the provision of organisational and technical infrastructures, and security designed to ensure 'preservation' for its designated community of users. Preservation is a promise and a process to ensure that risks to the continued **access**, **usability** and **understandability** of digital objects are minimised.

To be in scope for CoreTrustSeal and TDR status, applicants must have a mission (Requirement 01) to provide continued access to data and a Preservation plan (R09).

[22] https://doi.org/10.5281/zenodo.4033966


CoreTrustSeal
Organizational Infrastructure
R01. Mission & Scope
R02. Rights Management
R03. Continuity of Service
R04. Legal & Ethical
R05. Governance & Resources
R06. Expertise & Guidance
Digital Object Management
R07. Provenance & Authenticity
R08. Deposit & Appraisal
R09. Preservation Plan
R10. Quality Assurance
R11. Workflows
R12. Discovery & Identification
R13. ReUse
Information Technology & Security
R14. Storage & Integrity
R15. Technical Infrastructure
R16. Security

The 16 CoreTrustSeal Requirements 2023-2025

Characteristics and actions of TDRs include:

- Monitor Community (knowledge base, needs)
- Monitor Technology (avoid obsolescence, identify new opportunities)
- Create a preservation-ready infrastructure: appraisal, deposit, data and metadata enrichment, rights, risks, plans, formats, re-appraisal
- Take Preservation actions when necessary (format migration, emulation)

Some repositories may set minimum standards for deposit, such as required metadata, others may offer curation services to enrich or improve metadata before access. Preservation involves both of these and more.

Preservation repositories prepare to address the next round of change and provide an environment capable of doing that in a sustainable way.

Preservation is not digitisation. Converting the analogue to digital does not (alone) guarantee its future.

Preservation is not retention. But it is helpful to know the minimum retention period for an object.

Preservation is not storage. Effective storage is critical to a TDR, even though we lack formal, minimal storage definitions (N copies, on N media in N locations etc). 'Bit-level preservation' is used but not defined as anything beyond effective storage practice.

To Preserve, or not to Preserve

The remit of repositories is usually to ensure that they, and the digital objects they care for, meet standard criteria. This is more a 'technical quality' compliance role than one that evaluates 'scientific' quality. But not all digital objects are equal, and resources are not infinite. So, not all data can or must be preserved whether in a generalist or a disciplinary repository. Appraisal and selection processes aligned with collections development policies support these decisions. Without preservation the value of 'digital assets' may not be maintained over time. Digital objects are subject to re-appraisal and changes in their level of curation and preservation (or even retention).

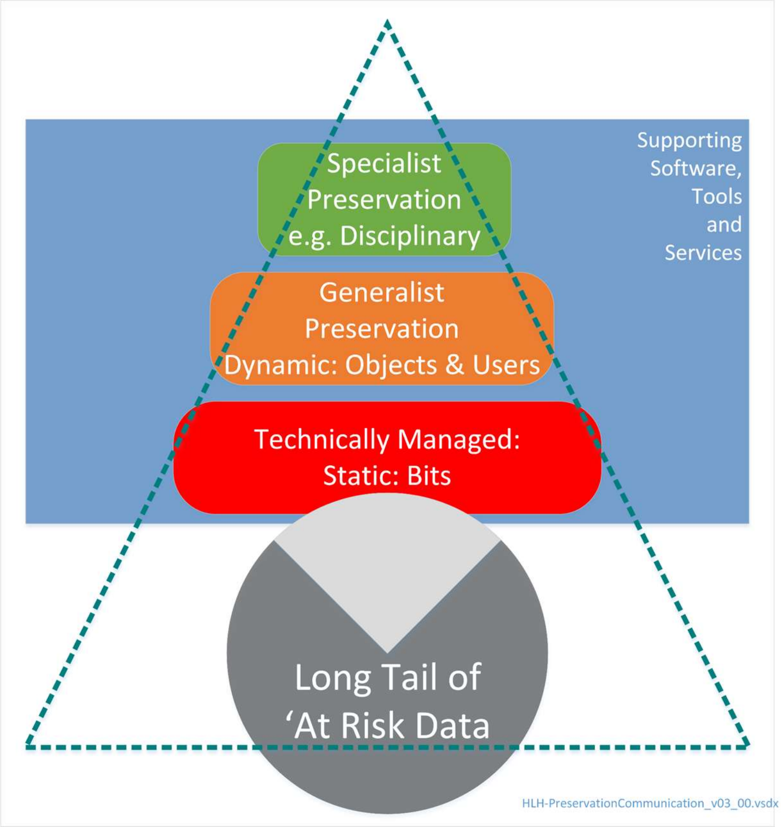
Preservation has costs, but exact costs can be hard to extract from other activities. Failing to preserve has costs, but these are often difficult to quantify until a preservation failure has occurred.

Disciplinary vs Generalist

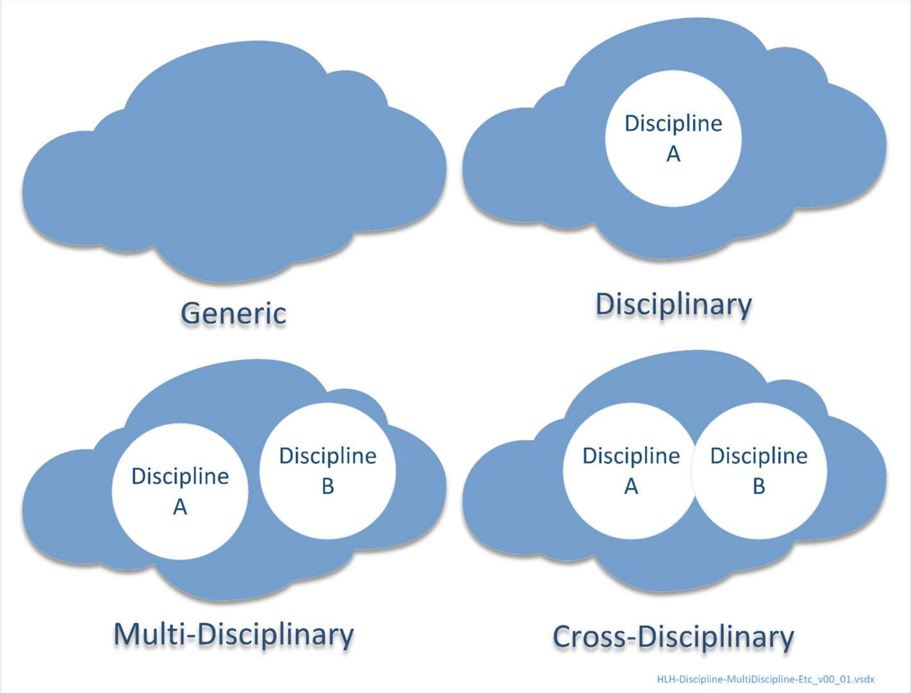
Many funders will prefer a TDR, and guidance (e.g. Science Europe[23]) may recommend one with disciplinary expertise. CoreTrustSeal is working to identify what disciplinary expertise a repository claims at the point of application, and will seek to include this in repository registry[24] information when certified. This will support searches for disciplinary TDRs. CoreTrustSeal expects the applicant to present relevant evidence related to its claimed discipline, including defining a clear designated community. But CoreTrustSeal does not certify **for** specific disciplines or domains.

[23] Practical Guide to the International Alignment of Research Data Management - Extended Edition https://doi.org/10.5281/zenodo.4915862

[24] https://www.re3data.org/



Defining and agreeing disciplinary expertise is not trivial, and will become even more challenging as repositories and data services claim to be interdisciplinary, multi-disciplinary or cross-disciplinary.



The TF needs to take care to use terms in a way that clarifies that a generalist/generic repository is not confused with one that is 'multi-disciplinary'.

Repositories and Data Services

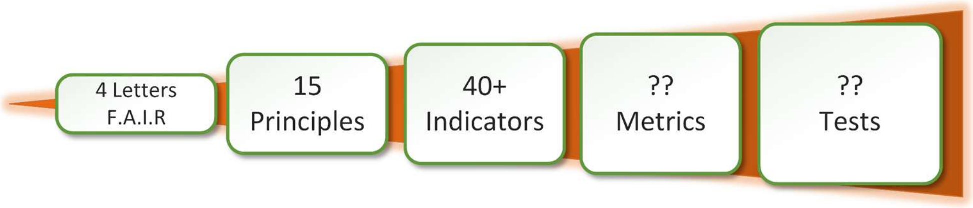
Not all repositories are candidates to become TDRs but still play a critical role in the data infrastructures and lifecycles. Federated infrastructures such as the EOSC depend on many types of data services. Many repositories, including TDRs, are composed of complex partnerships or outsource some functions to other data services. There are some key commonalities between services that 'touch' data and metadata including ensuring integrity (avoiding unintended change). If any changes to data and metadata are undertaken by a data service, then they should show provenance tracking (recording intended changes).

These other data services are not yet defined sufficiently for assessment (or certification)[25]. This is important as repositories are not the only actors, so we need trustworthy data services. One example is the development of registries, whose important role[26] makes them a dependency for the effectiveness of other data services. To undertake evaluations, it is important to be able to define the type of 'thing' (repository, service, registry, software etc) being assessed and to set standards, processes and governance.

[25] FAIRsFAIR D2.7 Framework for assessing FAIR Services https://doi.org/10.5281/zenodo.5336234
[26] As highlighted in Turning FAIR into reality https://data.europa.eu/doi/10.2777/54599

FAIR

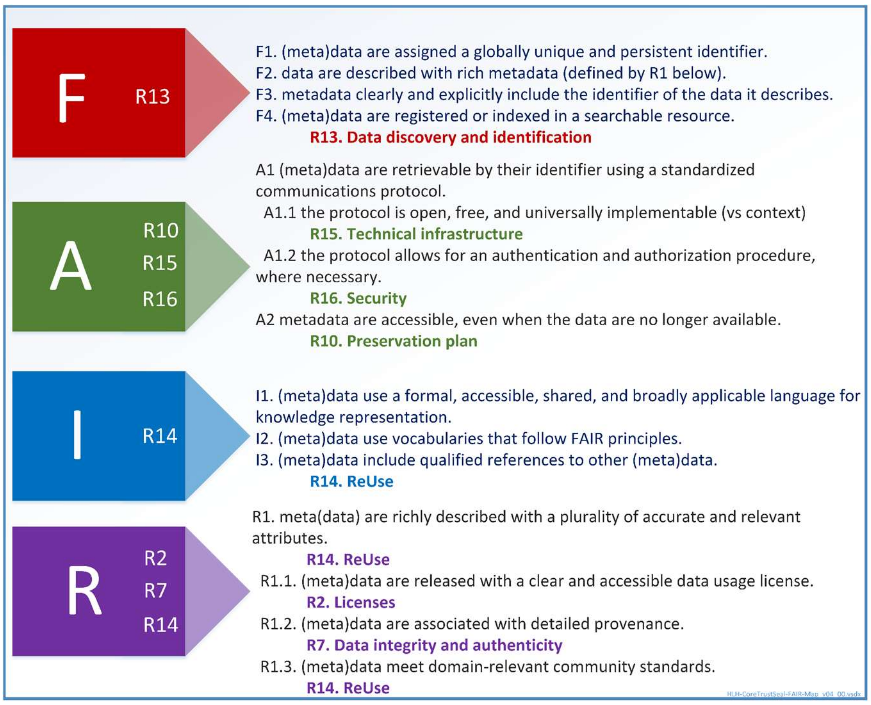
The development of FAIR expectations and the delivery of FAIR infrastructure and objects is an ongoing journey. We have the acronym and principles and we've evolved indicators but the development of metrics and tools that test against them is still in progress.



HLH-FAIR-AcronymPrinciplesIndicatorsMetricsTests_02_00.vsdx

There is more work to do on disciplinary FAIR definitions and to improve machine actionability of assessment.

For an object to be findable, accessible, interoperable and reusable we may have to undertake curation actions to ensure digital objects meet minimum criteria to support immediate access and reuse. But objects and their environments can become obsolete, so unattended objects are not guaranteed to remain FAIR over time. The problem of FAIR+Time is solved by preservation.



F1. (meta)data are assigned a globally unique and persistent identifier.
F2. data are described with rich metadata (defined by R1 below).
F3. metadata clearly and explicitly include the identifier of the data it describes.
F4. (meta)data are registered or indexed in a searchable resource.
R13. Data discovery and identification

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable (vs context)
R15. Technical infrastructure
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
R16. Security
A2 metadata are accessible, even when the data are no longer available.
R10. Preservation plan

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.
R14. ReUse

R1. meta(data) are richly described with a plurality of accurate and relevant attributes.
R14. ReUse
R1.1. (meta)data are released with a clear and accessible data usage license.
R2. Licenses
R1.2. (meta)data are associated with detailed provenance.
R7. Data integrity and authenticity
R1.3. (meta)data meet domain-relevant community standards.
R14. ReUse

HLH-CoreTrustSeal-FAIR-Map_v04_00.vsdx

FAIRsFAIR mapping of the Principles to CoreTrustSeal Requirements

EOSC

The provision of FAIRenabling Trustworthy Digital Repositories[27] containing automatically assessed FAIR digital objects is a desirable outcome. But there is recognition[28] that assessment of objects and certification of services should not yet be a gatekeeper to the EOSC; Trust, preservation and FAIR are a journey.

There is a long tail of unFAIR data to be addressed and many repositories will need ongoing support to reach trustworthy status. Delivering this at scale partially depends on machine actionability of metadata and assessment, which depends on clear expectations supported by the community, including disciplinary communities.

[27] M4.3 CoreTrustSeal+FAIRenabling, Capability and Maturity https://doi.org/10.5281/zenodo.5346822

[28] European Commission, Directorate-General for Research and Innovation, Slavec, A., Jones, S., Aronsen, J., et al., Recommendations on certifying services required to enable FAIR within EOSC, Genova, F.(editor), Publications Office, 2021, https://data.europa.eu/doi/10.2777/127253