eosc | Long-term Data Preservation Task Force

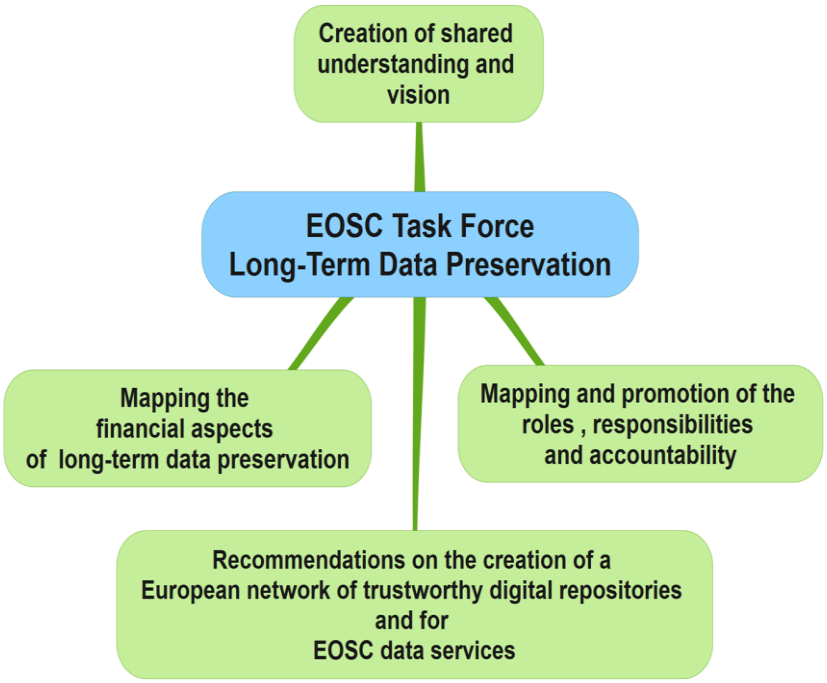
Preservation in the context of EOSC

Sustainable repositories curating digital objects from a long term FAIR enabling perspective

PV2023

04 | 05 | 2023 by Hervé L'Hours & Roxanne Wyns

Task force structure



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

EOSC-A liaison: Ignacio Blanquer (BoD)
EOSC-A support officer: Celia Alvarez

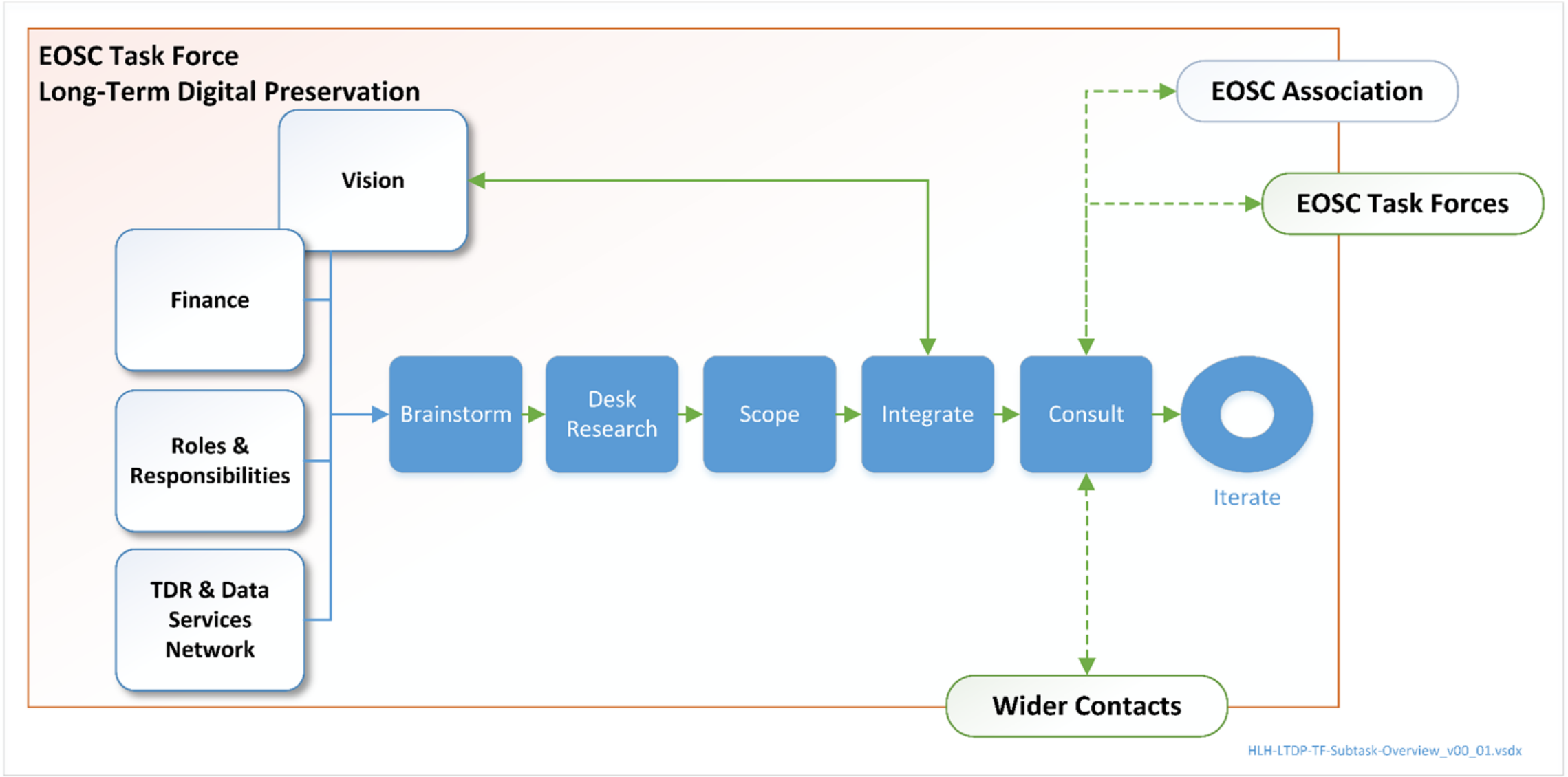
Shared Understanding and Vision
Subgroup lead: Chris De Loof (Belnet)
Pirjo-Leena Forsström (CSC), Rui Fernandes (C4G), Gerardo Ganis (CERN), Ville Tenhunen (EGI), Andras Holl (MTA), Maciej Brzeźniak (PSNC), Lara Lloret Iglesiasc (CSIC), Riccardo Smareglia (INAF), Eileen Gibney (UCD)

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)

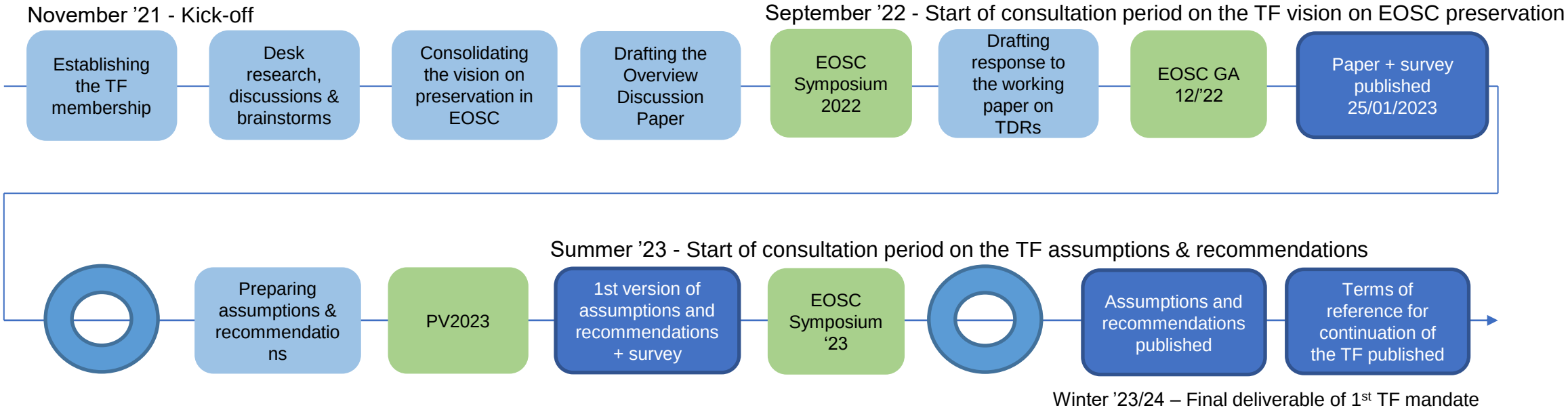
Mapping of financial aspects
Subgroup lead: Paul Stokes (JISC)
Jean-Yves Nief (CC-IN2P3), Jiri Novacek (CEITEC), Toni Andreu (EATRIS), Martina Stockhause (IPCC DDC)

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)

Task force approach



Task force timeline



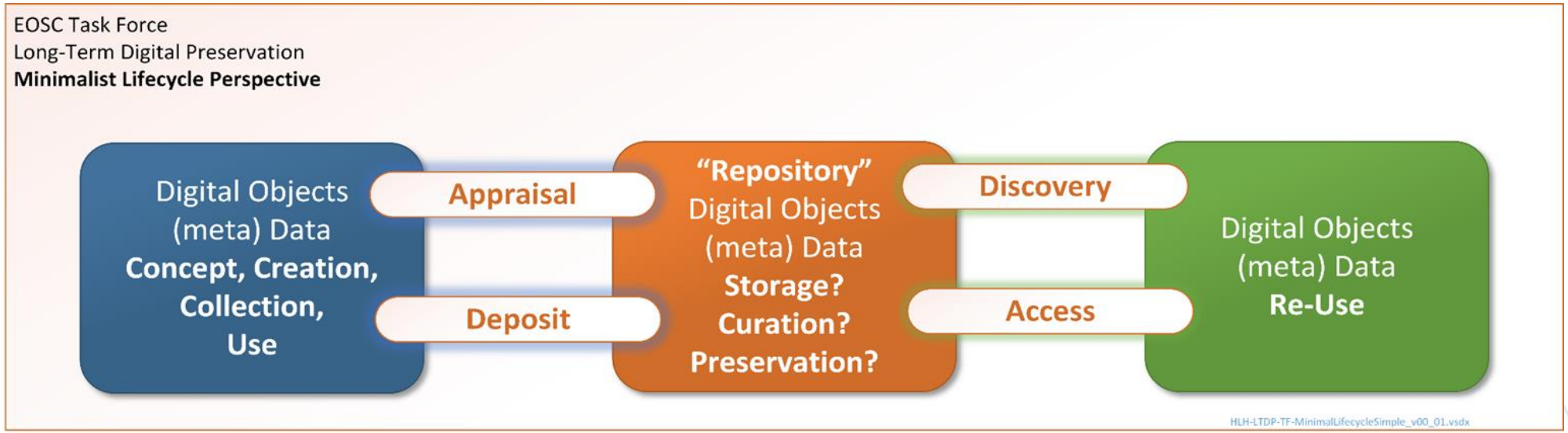
TF outputs

- Overview discussion paper https://doi.org/10.5281/zenodo.7516259 + survey
- Response to the working paper “Toward a European network of FAIR-enabling Trustworthy Digital Repositories”. https://doi.org/10.5281/zenodo.7568399
- Preservation in the context of EOSC. Sustainable repositories curating digital objects from a long-term FAIR enabling perspective, PV2023

Preservation in its wider context

Wide range of assumptions and perspectives around digital preservation

Contextualise preservation within the wider research life cycle and across the federated and consolidated network of partners that deliver scientific infrastructure (meta)data services



Care of digital objects

Strategic risk benefit approach is needed to identify financial, operational, reputational risks related to preservation action and inaction

This requires

- Skills to fulfil roles and deliver critical infrastructure services, legal frameworks etc.
- Risk registers that identify preservation specific risks to help quantify the challenges of ensuring digital objects storage, integrity, provenance and fitness for use over time

A failure to preserve digital objects as FAIR over time might result in the retention of digital objects that have reduced potential for reuse

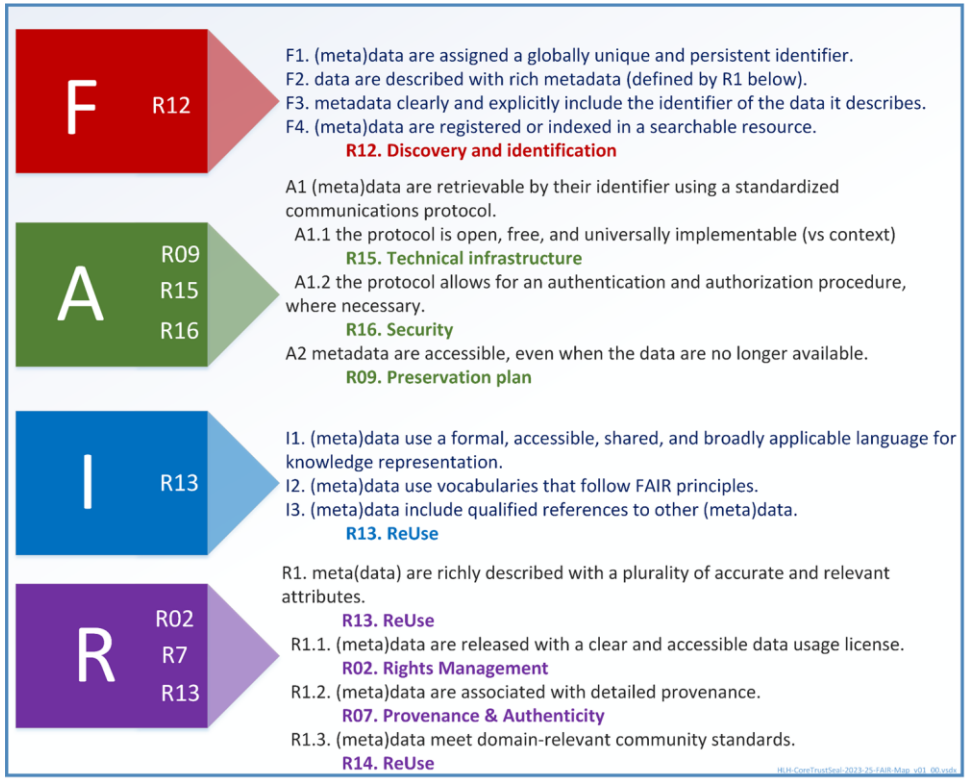
TRUST & FAIR

Trustworthy Digital Repositories cover transparency, responsibility, sustainability and technology

Digital objects with a high potential for reuse are represented by FAIR

The need for data and metadata to *remain* FAIR is not explicit in the FAIR principles

FAIR-enabling repositories can ensure that digital objects remain FAIR over time through preservation



Preservation in the context of EOSC

A successful EOSC necessitates engagement with a broad range of actors

Trustworthiness and FAIR digital objects require specialist and complex practices of different actors in one or more stages of the digital object management life cycle

Priorities for the TF

- Consider the roles and financial implications related to preservation
- Identify which research data management activities are unique to preservation
- Clarify how preservation works alongside other research data management activities

An ongoing journey

Definitions and expectations of data services, including preservation, are hard to define and agree upon

Standards, associated tests of, and credit for compliance (incl. certification) are necessary and desirable, but must be seen as part of an emerging journey towards TRUST and FAIR

The TF sees transparency on the status of repositories as a priority

- Organisational infrastructure
- Digital object management actions
- Technology
- Security
- Roadmap towards improvement of TRUST and FAIR

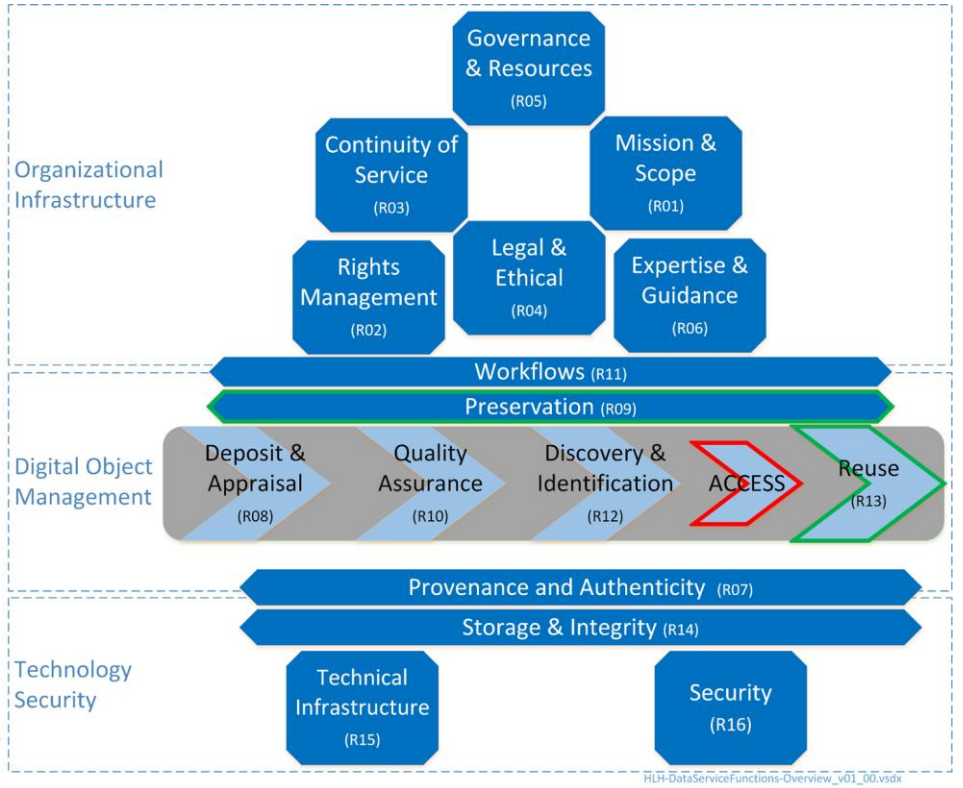
This will increase the knowledge base and the trust of funders, depositors and reusers

Specific requirements for preservation

CoreTrustSeal applicability to wider data services

- Requirement for Preservation [R09] and ReUse [R13] unique to TDR candidate repositories
- Other CTS Requirements are applicable across all organisations related to data and metadata

CoreTrustSeal as a reference point for the task force, in alignment with the perspectives on preservation outcomes, systems, and actions



Review of CoreTrustSeal Applicability to non-Preservation (Trustworthy) Data Services https://doi.org/10.5281/zenodo.7646134

TF perspective on preservation

EOSC Preservation: Overview Discussion Paper https://doi.org/10.5281/zenodo.7516259

"**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.

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 reusable."

Curation and preservation levels

As proposes by CoreTrustSeal https://doi.org/10.5281/zenodo.6908019 and discussed with the TF

Level 0 Digital object is distributed unchanged from the state in which is was deposited

Level C delivers 'Basic compliance and/or curation'

- Data and metadata checked at point of deposit
- Necessary steps to make sure they comply
- Minimal curation for initial access and use but no long term preservation

The task force proposes to differentiate between deposit standards (D) and initial curation (C)

Level C is not providing active preservation as it provides no long term commitments or necessary actions for continued FAIRness of the deposited objects

Curation and preservation levels

As proposes by CoreTrustSeal https://doi.org/10.5281/zenodo.6908019 and discussed with the TF

Level B delivers 'Logical-Technical Curation' in addition to C

- Ensures that data and metadata are updated over time to newer standards and formats (technical risks, changing needs of the designated community)

Level A delivers 'Conceptual preservation for understanding and reuse' in addition to B and C

- Monitors changes to the definition an demands of their designated community
- Takes responsibility for preservation actions to ensure that digital objects can be understood and reused
- This involves updates to the content of metadata elements and other semantic artefacts

Different institutions, nations, domains and disciplines have different expectations from their trustworthy repository

Additional functions and costs

All repositories need the provision of deposit, storage, curation and access

Repositories delivering preservation have additional functions, activities, roles and costs

Not enough insight into the differentiate costs between initial curation and long term preservation

Active preservation may be assumed to represent additional costs over lower levels of care, but digital objects that lose their reuse value because of lack of active preservation also have a cost

Defining the value of digital objects is more challenging then defining the costs of their care (evolves over time, specific to communities, ease of reproduction …)

Value of digital objects might differ if we look at potential for reuse vs. ability to reproduce and validate

Transparency

The TF advocates transparency on the levels of care

- How long will a digital object be retained in its current state and whether its being preserved for ongoing usability
- How and when a decision might be taken to change the level of care

Ideally, decisions to reduce the level for a digital object would be transparent across communities so others can take over the responsibility for the object

Towards recommendations

The task force has been explicit in defining what preservation means in the context of EOSC

Systems claiming to deliver preservation outcomes, need to be sustainable and curate digital objects from a long-term FAIR enabling perspective

Criteria for FAIR digital objects aligned with those for trustworthy digital repositories such as CoreTrustSeal provide an important reference point for evaluating objects and services

TRUST and FAIR enabling practices should be seen as an end goal in an ongoing journey

Transparency around current levels of FAIRness and trustworthiness of repositories will lead to better understanding

The TF will focus on statements and recommendations that are generally applicable

- Broad policies and guidance where relevant bodies have influence
- Direct practice and infrastructure where those bodies have control


Thank you!
Questions?