

The EOSC EDEN Core Preservation Processes (CPPs)

EOSC EDEN | Workshop
1st & 2nd October 2025 – Leuven, Belgium



**Funded by
the European Union**

04 | Feb | 2025 by Micky Lindlar (they/them)



Who's the „we“ in this? The CPP core writing group



Arkivum

... with input / comment from all other project partners

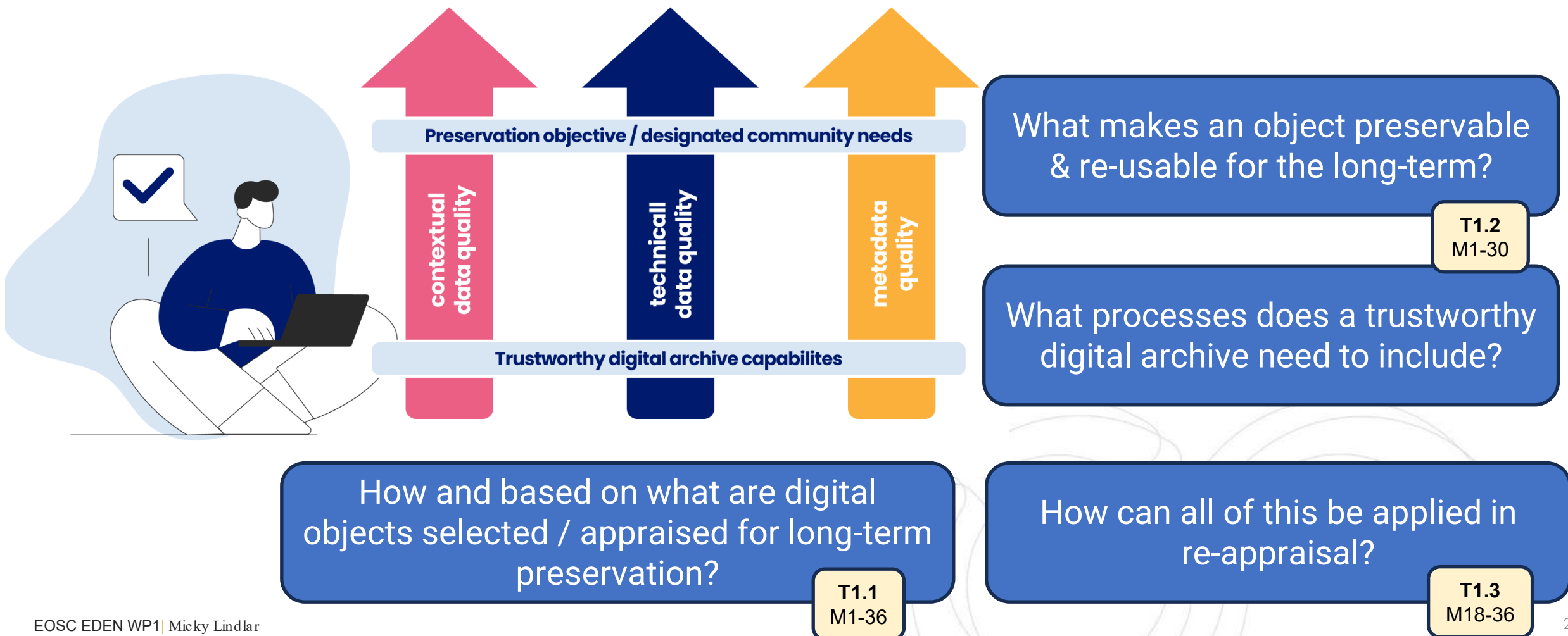
M1.1 Report on Identification of Core Preservation Processes

EOSC EDEN T1.2 ; Lindlar, Micky (Work package leader)¹ ;
Caron, Bertrand (Project leader)¹ ; Benauer, Maria (Project member)¹ ;
Kylander, Johan (Project member)² ; Dekeyser, Kris (Project member)³ ;
Addis, Matthew (Project member)⁴ ; Levlin, Mattias (Project member)² ;
Laukkanen, Mikko (Project member)² ; Lehtonen, Juha (Project member)² ;
Burger, Felix (Project member)¹ ; Koho, Tiina (Project member)² ;
Schwab, Franziska (Other)¹ ; Molloy, Laura (Other)⁵ ;
Zhang, Fen (Other)³

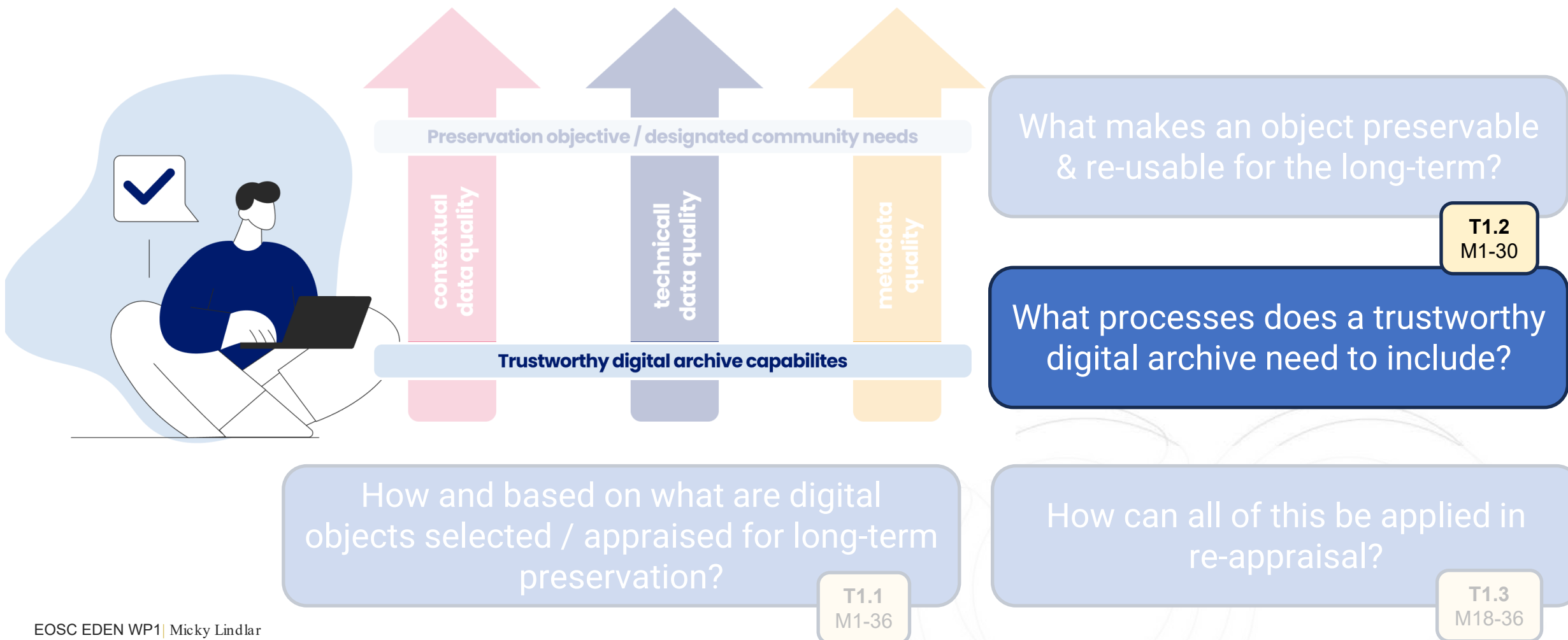
<https://doi.org/10.5281/zenodo.16992451>



WP1: EDEN's core digital preservation work package



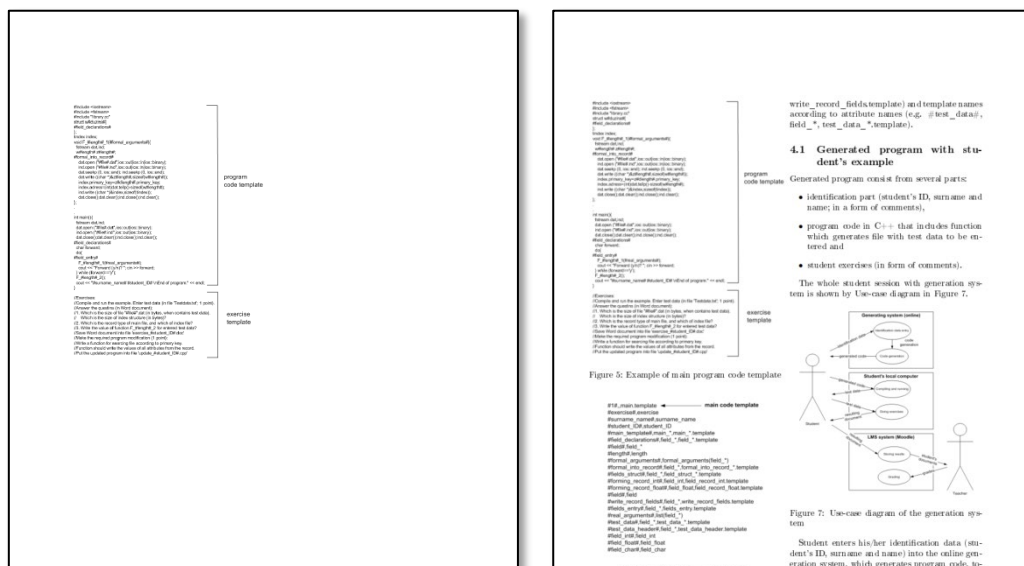
The EOSC EDEN Core Preservation Processes (CPPs)



Why should I care?

Real life examples of what can go wrong

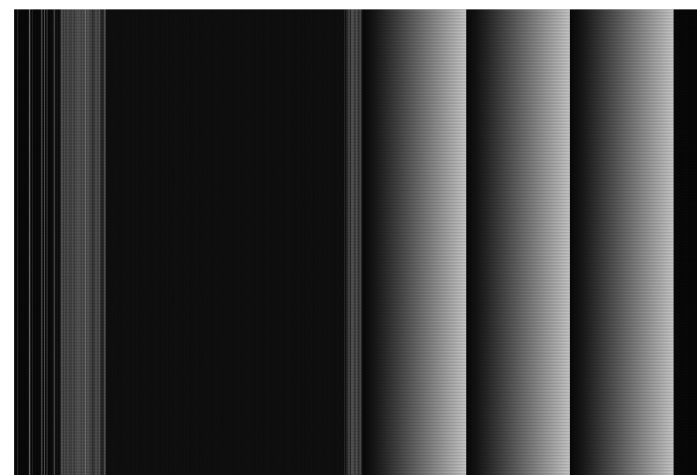
Apart from the usual reasons (sustainability, responsibility, do it once, do it right) digital preservation processes help improve the data today!



<https://openpreservation.org/blogs/validation-ok-they-said-fixing-the-rendering-of-a-so-called-valid-pdf/>

$$\int_0^1 \frac{1}{s} r^3(s) \sqrt{1-s^2} ds = \int_0^1 \frac{1}{s} r^3(s) \sqrt{1-s^2} ds \quad (6)$$

<https://openpreservation.org/blogs/trouble-shooting-pdf-validation-errors-a-case-of-pdf-hul-38/?q=109>



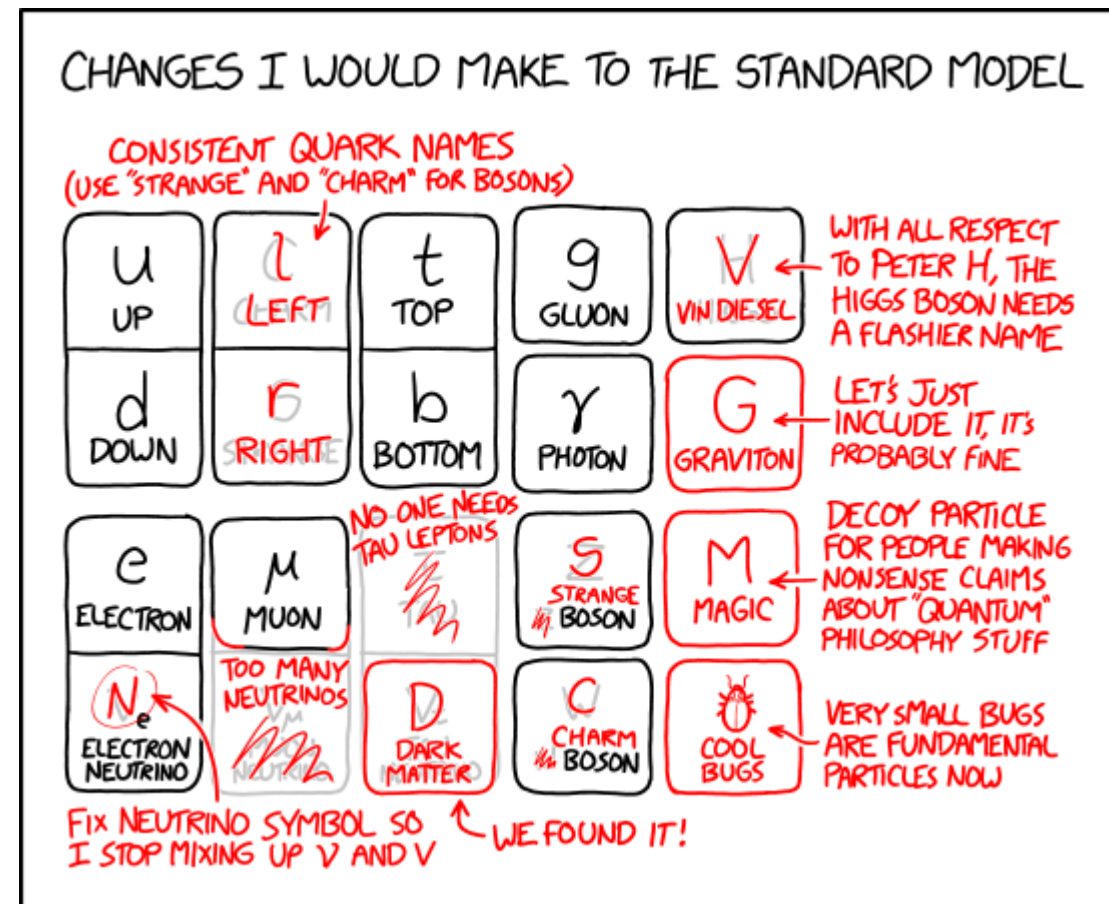
<https://openpreservation.org/blogs/when-stars-and-strips-align-fixing-the-stripoffsets-tiff-tag/?q=1>

Let's not do this ...



<https://xkcd.com/927/>

.... or this



<https://xkcd.com/2351/>

CPP Definition – the abridged version

See the workshop „Readings & Ressources“ or <https://doi.org/10.5281/zenodo.16992452> for the longer version

Core Preservation Process (CPP) = a specific action that every Trustworthy Digital Archive (TDA) should undertake (either directly or through an associated service)

- focus in operational activities required by digital preservation
- are limited to generic processes (any content, domain, discipline)



We consider **Trustworthy Digital Archive** and **Trustworthy Digital Repository** as roles that an entity can take on. An entity can take on one or both roles.



Strategic/managerial activities like staffing or **secure IT infrastructure management** such as sensitive data protection **is out of scope**

The CPP Origin Story

Parent A:

„Collection of frameworks, guidelines and best-practices“

- ☑ CoreTrustSeal requirements
- ☑ EOSC LTDP TF recommendations
- ☑ FAIR Implementation Profiles
- ☑ ISO16363
- ☑ nestor Seal requirements (DIN31644)
- ☑ NDSA Levels of Preservation
- ☑ dpc RAM (Rapid Assessment Model)
- ☑ RDA TRUST Framework

Parent B:

„Core requirements for a digital preservation system“

- dpc publication (2022) as part of the „Procurement Toolkit“
- 10 core requirements, additionally described in 41 must/should/could statements

example

5. The system must have the facility to assess the characteristics of ingested digital content and record them in associated metadata“

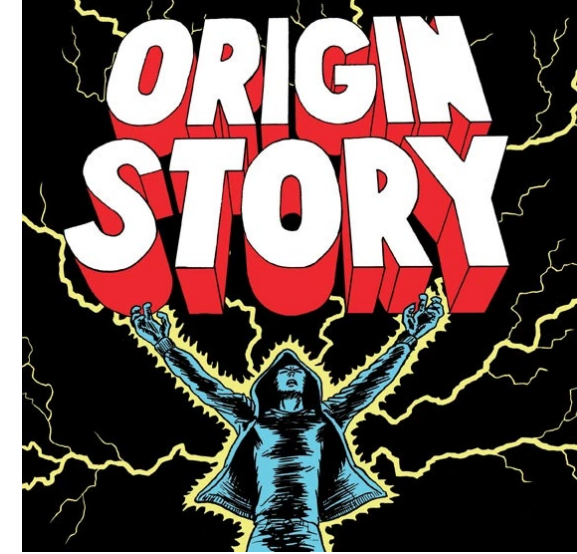
5.1 The system must **identify file formats to the level of specific file format version and reference appropriate registries of further information such as PRONOM and/or Wikidata.**

5.2 The system must extract technical characteristics (such as size, image dimensions, video codec, audio run time, creating application).

5.3 The system should identify content that cannot be rendered, such as broken, badly constructed, or encrypted content.

5.4 The system should validate file formats against file format specifications or customised profiles.

5.5 The system should capture external dependencies, where content (or software) not present in the digital object is vital for it to be rendered or used (such as non-embedded fonts, non-embedded media such as YouTube videos or software libraries).



The final CPP candidate list

CPP-001 Checksum Generation & Recording

CPP-002 Checksum Validation

CPP-003 Integrity Checking

CPP-004 Data Corruption Management

CPP-005 Identifier Management

CPP-006 AIP Batch Export

CPP-007 Virus Scanning

CPP-008 File Format Identification

CPP-009 Metadata Extraction

CPP-010 File Format Validation

CPP-011 Replication

CPP-012 Risk Mitigation

CPP-013 Object Management Reporting

CPP-014 File Migration

CPP-015 Emulation & Rendering Tools

CPP-016 Metadata Ingest & Management

CPP-017 Disposal

CPP-018 Community Watch

CPP-019 Data Quality Assessment

CPP-020 Rights Management

CPP-021 AIP Versioning

CPP-022 Significant Properties Definition

CPP-023 Risk Definition and Extraction

CPP-024 Enabling Discovery

CPP-025 Enabling Access

CPP-026 File Normalisation

CPP-027 File Repair

CPP-028 Creation of Derivatives

CPP-029 Ingest

CPP-30 Refreshment

Read the [M1.1 report](#)
for the blood, sweat and
versioning story behind
it



Isn't that all the same thing?

CPP-001 Checksum Generation & Recording

CPP-002 Checksum Validation

CPP-003 Integrity Checking

CPP-004 Data Corruption Management

CPP-005 Identifier Management

CPP-006 AIP Batch Export

CPP-007 Virus Scanning

CPP-008 File Format Identification

CPP-009 Metadata Extraction

CPP-010 File Format Validation

CPP-001 Checksum Generation & Recording

- Data may be received with no integrity information or with integrity information in algorithm not supported by TDA (e.g., MD5 but TDA uses SHA-256 internally)
- Checksums need to be generated & recorded in algorithm supported by TDA

CPP-002 Checksum Validation

- Checksum validation is an **event-/lifecycle-stage driven process**, typically conducted at **ingest, access** or after **preservation action**
- CPP-001 is a pre-requisite for CPP-003

CPP-003 Integrity Checking


- Integrity checking is a typically **ongoing/scheduled process**, closely connected to **storage management policy**
- CPP-001 is a pre-requisite for CPP-003

When to use the descriptions

The descriptions are a great starting point if you want to:

- understand why this process is necessary
- learn more about a specific process and read up on reference implementations
- check how the process relates to certification processes
- Follow up on reference implementation

When reading the descriptions, the [Glossary](#) is a helpful additional resource!



Enhancing Digital preservation strategies at European and National level

Project Number: 101188015 Start Date of Project: 01/01/2025 Duration: 36 months

M1.1 Report on Identification of Core Preservation Processes: Glossary

Dissemination Level	Public
Due Date of Deliverable	31/08/2025, M8
Actual Submission Date	29/08/2025
Work Package	WP 1 - Data and Process Framework for Long-Term Digital Preservation
Task	1.2
Version	V. 1.0
Number of Pages	p.1 – p.23

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Agency. Neither the European Union nor the granting authority can be held responsible for them. This deliverable is licensed under a Creative Commons Attribution 4.0 International License.

 Funded by the European Union

EOSCEDEN has received funding from the EU's Horizon Europe research and innovation programme under Grant Agreement no. 101188015.

CPP Relationship visualization

Classification: Logical/Strategic ▾

View: Graph ▾

Select All Deselect All ☒ Preservation Planning ☒ Dissemination ☒ Bit-level Preservation ☒ Generation of New Files ☒ Other Activities ☒ Lifecycle Management

☒ Characterisation

Select All Deselect All

Total Visible: 513

Procedural

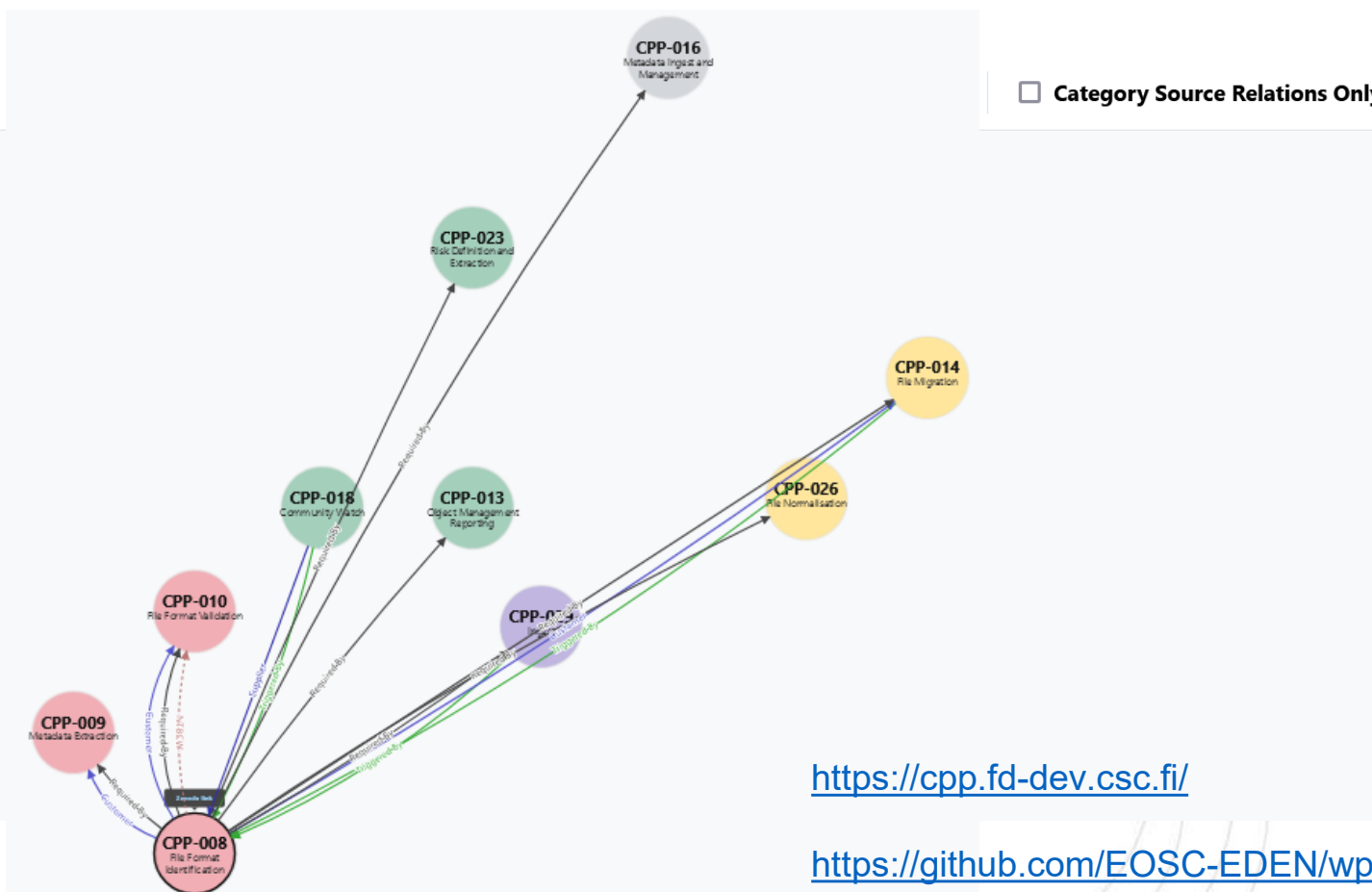
- ☒ Triggered By (85) ☒ Triggers (8)
- ☒ Supplier (91) ☒ Customer (67)
- ☒ Alternative To (2)

Dependencies

- ☒ Requires (84) ☒ Required By (80)
- ☒ May Require (16) ☒ May Be Req. By (16)

Logical

- ☒ Affects (2) ☒ Affected By (2)
- ☒ Facilitates (3) ☒ Facilitated By (3)
- ☒ Affinity (36)



<https://cpp.fd-dev.csc.fi/>

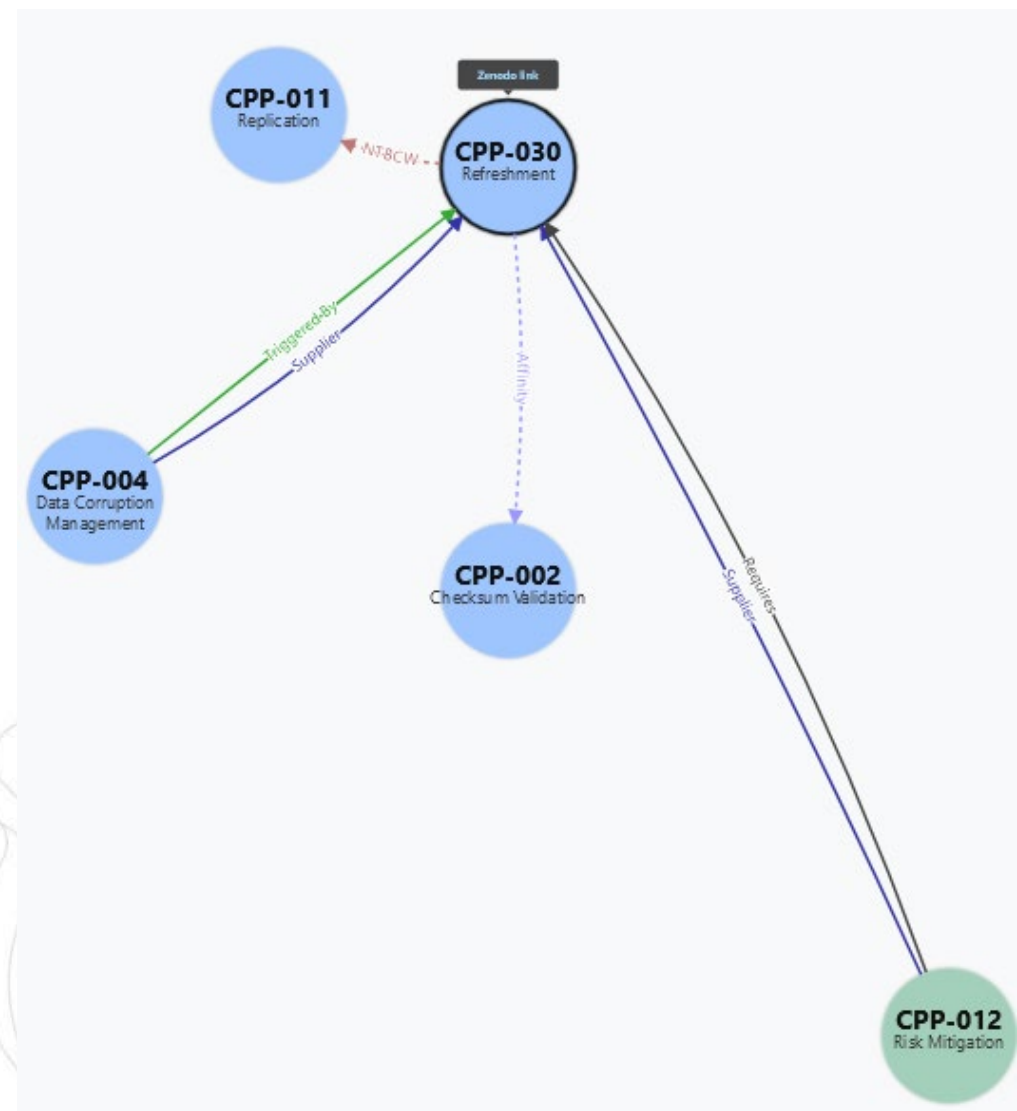
<https://github.com/EOSC-EDEN/wp1-cpp-visualization>

When to use the visualization tool

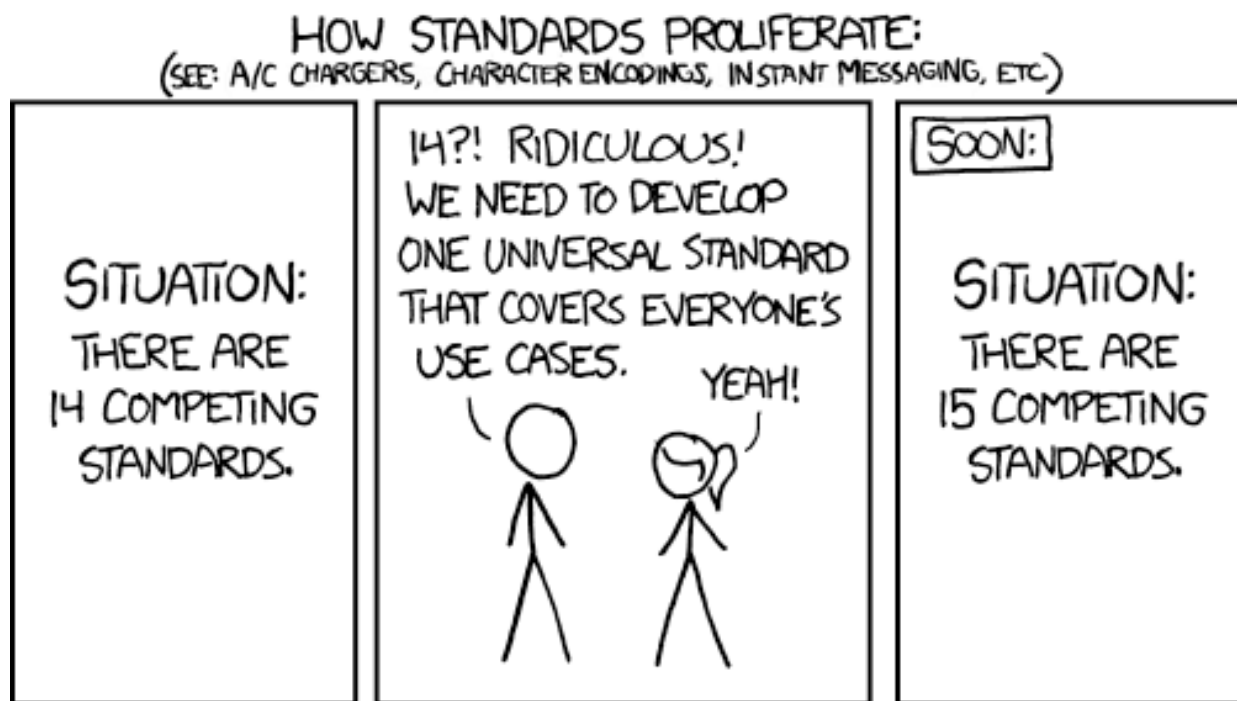
The visualization tool is a great starting point if you want to:

- See the relationships between different CPPs (dependencies, triggers, not-to-be-confused-with, etc.)
- Explore how you can extend your workflows to include further core preservation processes
- Understand how core preservation processes are connected in workflows

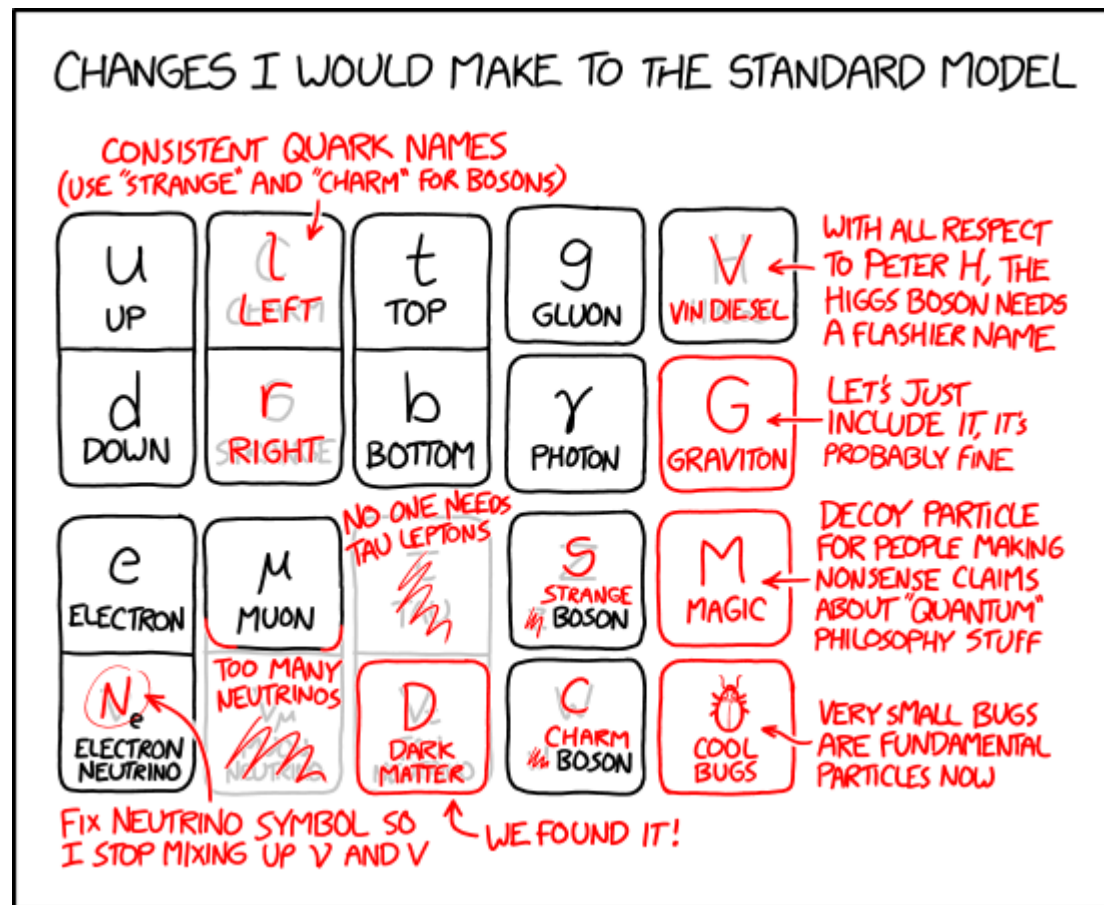
The visualization tool can be used as a dynamic table of contents; each bubble links to the CPP description document.



So did we succeed in not doing this?



<https://xkcd.com/927/>

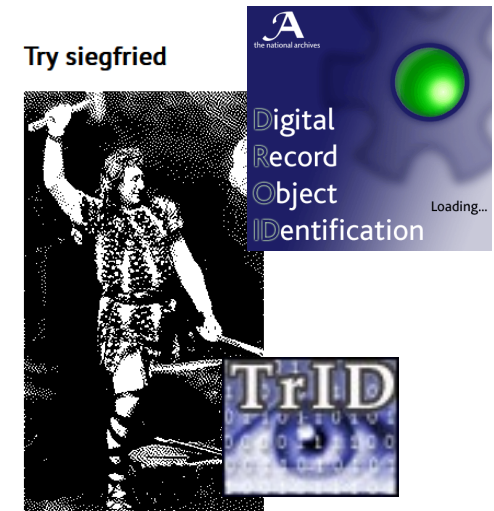


<https://xkcd.com/2351/>

An Example: File Format Identification

File Format Identification = A standard digital preservation process

- ([automatic](#)) identification of a digital object's file format
- Methods:
 - „magic number“ = first (2) bytes of file
 - syntax / semantic check against pattern registry
 - file format extension
 - heuristic combining several factors
 - opening file format in tool
- Tools typically return file format name / version and identifier used in a standard file format registry (e.g., [PRONOM](#), [Library of Congress Sustainability of Digital Formats](#), [Wikidata](#))



Drag a file on to Siegfried's anvil!

nestor_Praktikertag_2021.pdf
0.1 MB

ns: pronom
id: [fmt/19](#)
format: Acrobat PDF 1.5 - Portable Document Format
version: 1.5
mime: application/pdf
basis: extension match pdf, byte match at [[0 8]
[118014 5]]
warning:

LIBRARY
LIBRARY OF CONGRESS



The **technical registry**
PRONOM

Certification & File Format Identification

Deposit & Appraisal (R08)

R08. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for users.

- A list of preferred formats.
- Checks in place to ensure that depositors adhere to the preferred for



[Entwurf für ein Landhaus by Antoine François Peyre](#) (Künstler_in) - Albertina, Austria - Public Domain.

4.2.5.1 The repository shall have tools or methods to identify the file type of all submitted Data Objects.

Examples of Ways the Repository can Demonstrate it is Meeting these Requirements

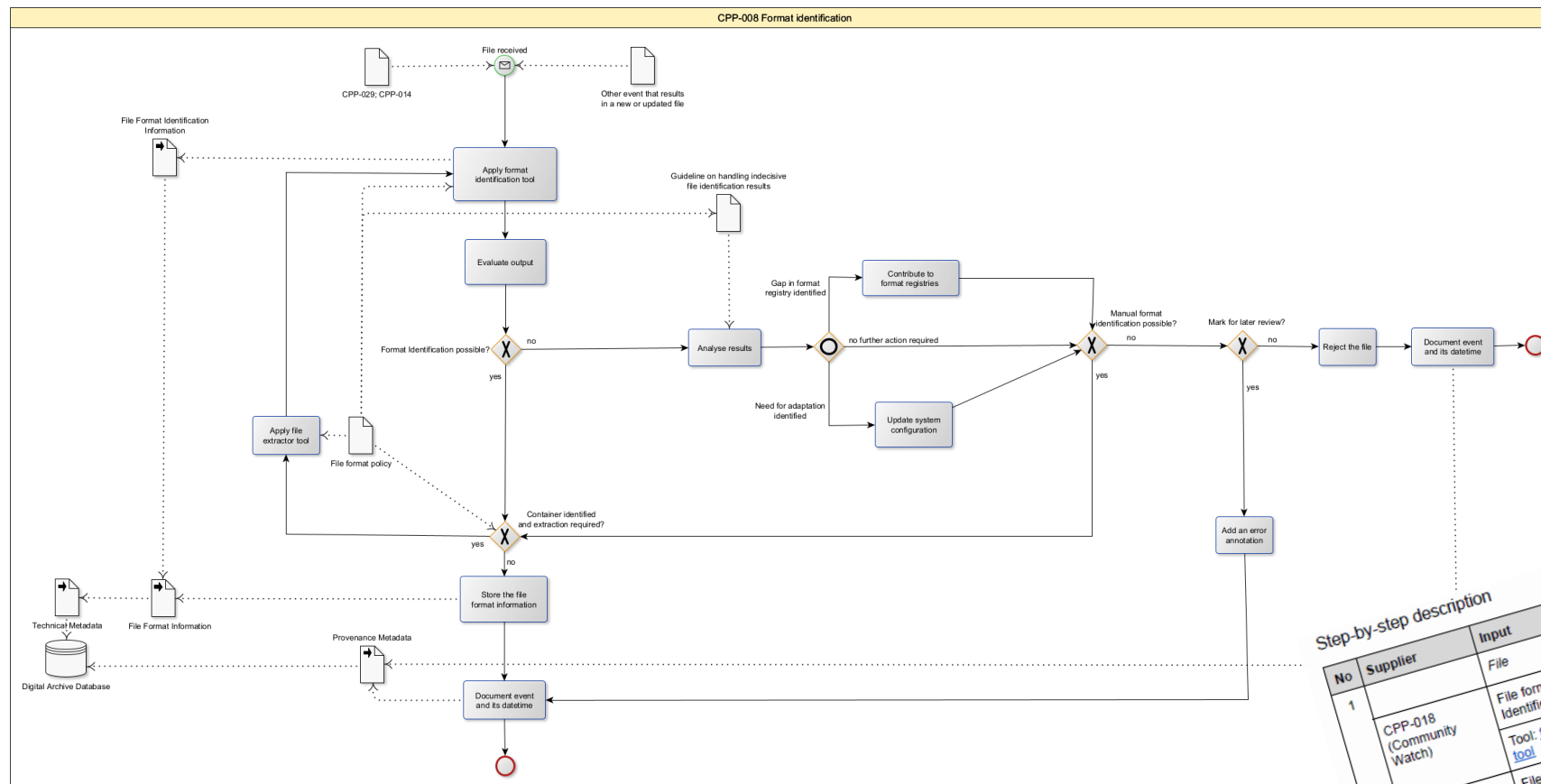
Subscription or access to registries of Representation Information (including format registries) viewable records in local registries (with persistent links to digital objects); database records that include Representation Information and a persistent link to relevant digital objects.

ISO 16363:2025

Space data and information transfer systems — Audit and certification of trustworthy digital repositories

CCSDS 652.0-M-2
AUDIT AND
CERTIFICATION OF
TRUSTWORTHY DIGITAL
REPOSITORIES

CPP on File Format Identification



"Tierpselever gjorde bra byggarbete i Söderfors",
Uppland 1967 by Arbetarbladet, Tierp - Upplands
Museum, Sweden - CC BY-NC-ND.

Step-by-step description			Steps	Output	Customer
1	Supplier	File	Apply the format identification tool(s) on the File	File format identification information	
	CPP-018 (Community Watch)	File format policy - Identification		File format identification successful (step 3)	
		Tool: format identification tool		File format could not be identified (step 2a)	
2		File format identification information	Parse and evaluate its output	A gap in the format registry is identified (step 2a- optional)	
		File format identification information	Analyse the format identification information and decide on next steps	Need for configuration update identified (step 2a- optional)	
2a		File format identification information		Manual file format identification possible (step 3)	
				Mark the File for later review and add the reason for review to the information (step 3)	

What's next for the CPPs

Enrich the descriptions

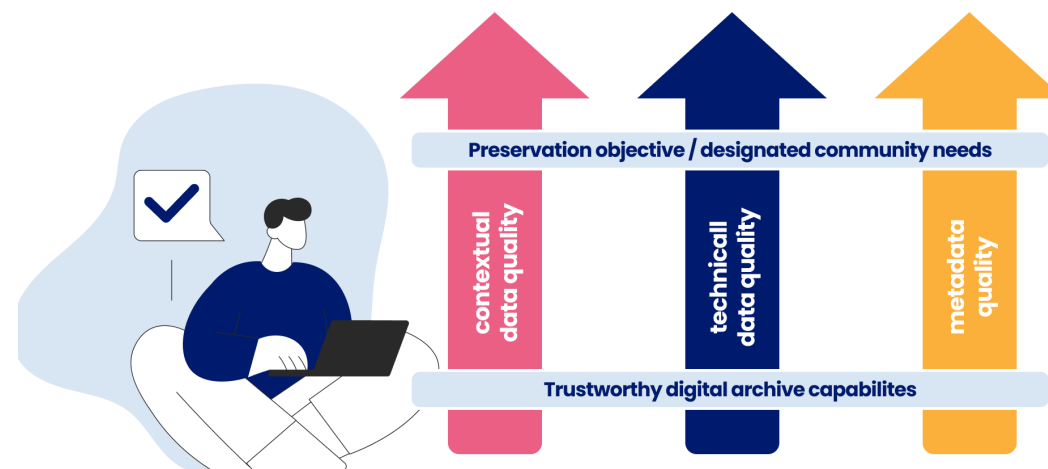
- Add more BPMN diagrams
- Add more use cases and reference implementations

Enhance usability

- Explore structured data representations
- Introduce semantic versioning

Embed in Re-Use Fitness Framework

- Test against discipline specific requirements
- Produce guidelines and training documents based on CPPs

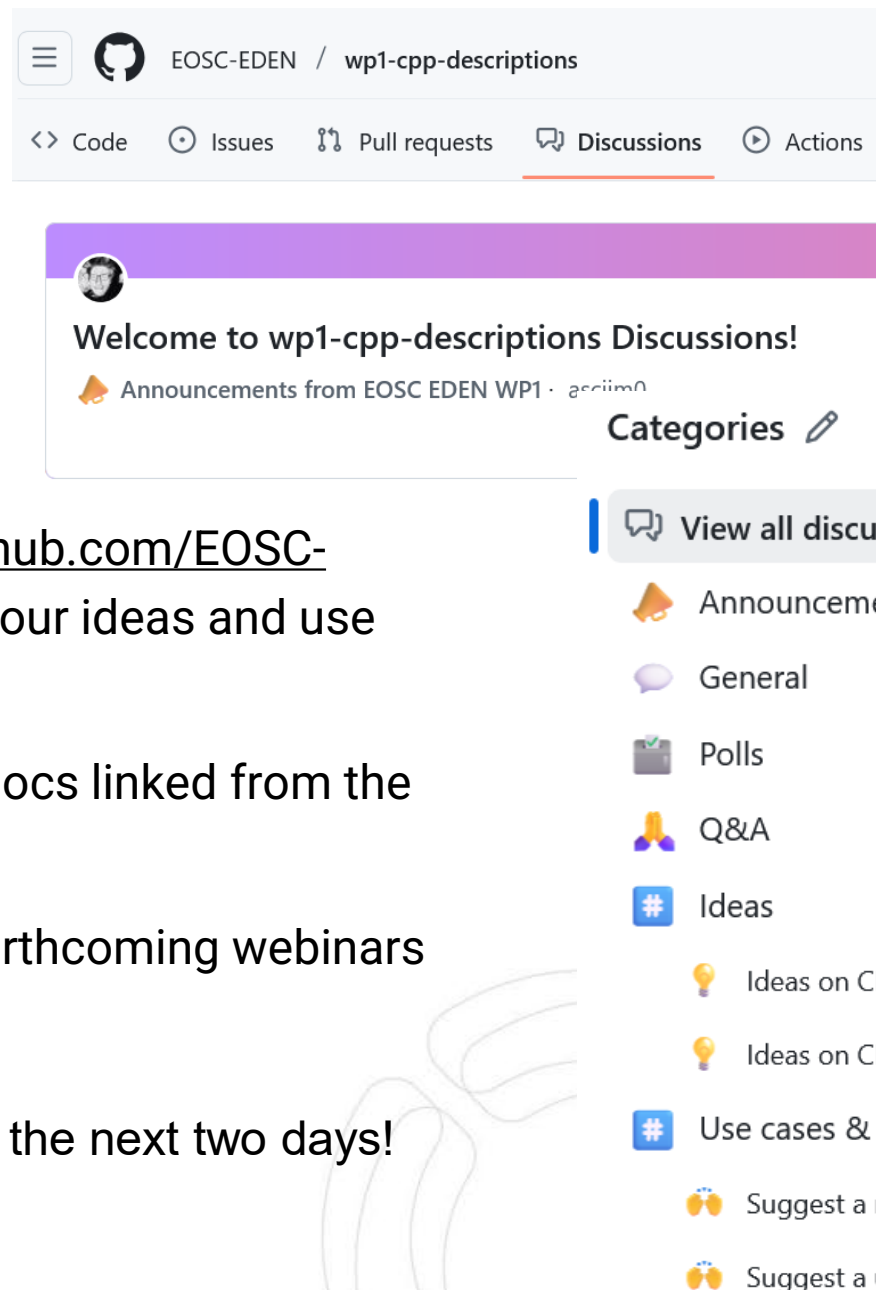


And, of course, hear & incorporate community feedback!

How to give feedback

- Use the „Discussions“ tab at <https://github.com/EOSC-EDEN/wp1-cpp-descriptions/> to share your ideas and use cases or to ask questions!
- Drop comments directly in the Google docs linked from the github repo
- Drop us an email or participate in our forthcoming webinars and other events

... and join the discussion over the course of the next two days!



Thank You



micky.lindlar@tib.eu



digipres.club/@mickylindlar



eden-fidelis.eu



linkedin.com/company/eosc-eden



[@eosc-eden.bsky.social](https://eosc-eden.bsky.social)



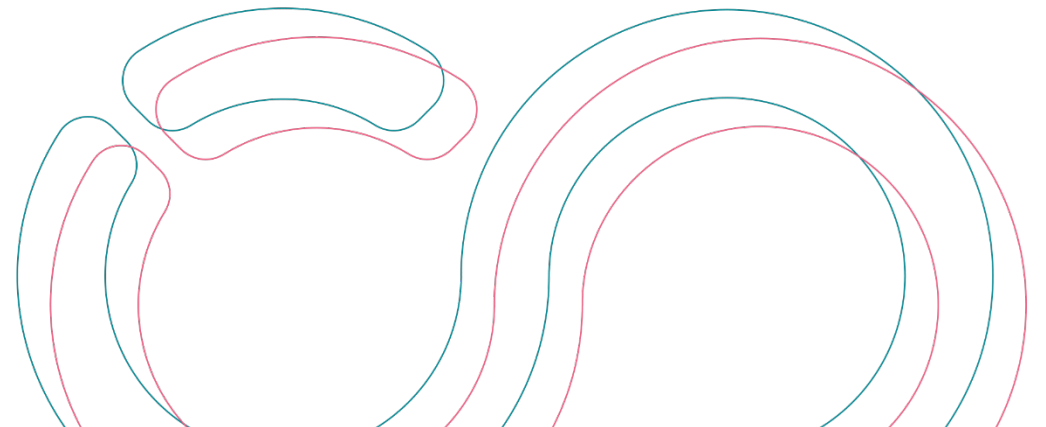
[@EOSC-EDEN](https://www.youtube.com/@EOSC-EDEN)

Let's finish with a quick 2-slide Mentimeter poll!



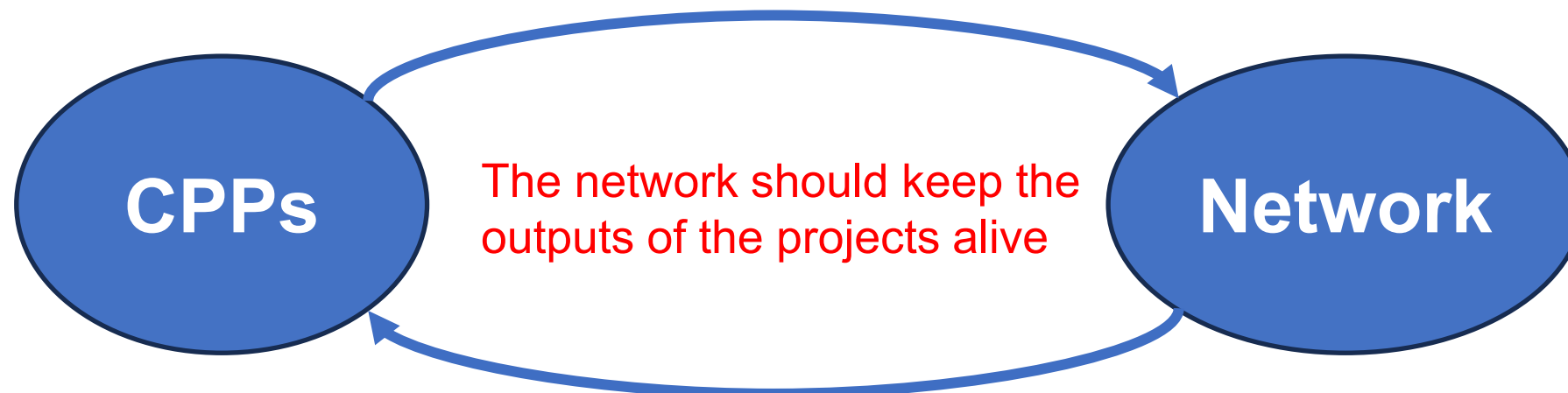
**Funded by
the European Union**

20/06/2025 by Micky Lindlar, TIB



What is the role of the CPPs in the network?

- Facilitate discussion on Usage and adoption?
- Training tools and guidelines for digital preservation practice
- Advocacy tool for TDA processes
- Fostering comparison between practices
- Gather disciplin-specific requirements for CPPs ?



What role does the network play in further development of the CPPs

- Updates, additions, removal of CPP descriptions (or any part thereof)
- How do we agree on proposed changes?
- Who signs off on changes?
- Regular review of descriptions?

CPP Breakout Group Pitch