

User perspectives on long-term data preservation: Reducing pain points for researchers, repositories and archives

2nd EOSC EDEN webinar, hosted by Helene N. Andreassen¹, Giacomo Cannizzaro² & Philipp Konzett¹

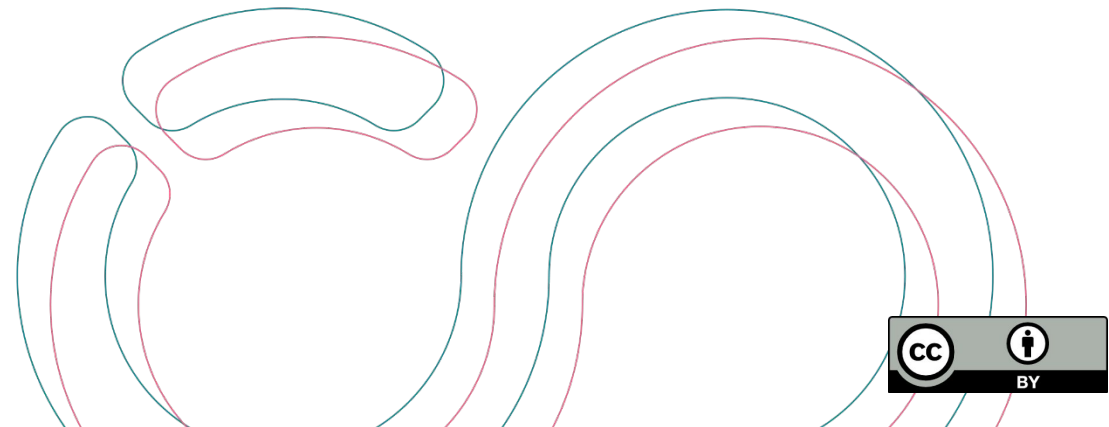
¹UiT The Arctic University of Norway, ²SURF

Tuesday 27th January 2026



**Funded by
the European Union**

Grant agreement 101188015



Welcome and introduction (Philipp)



Housekeeping

- Please note that this **webinar is being recorded** (until the Q&A), and the recording will be available on the EOSC EDEN YouTube channel with links to related resources.
- Please use the **Q&A panel** to submit your questions and comments during the presentations. They will be addressed in the panel or in the Q&A session at the end of the webinar.



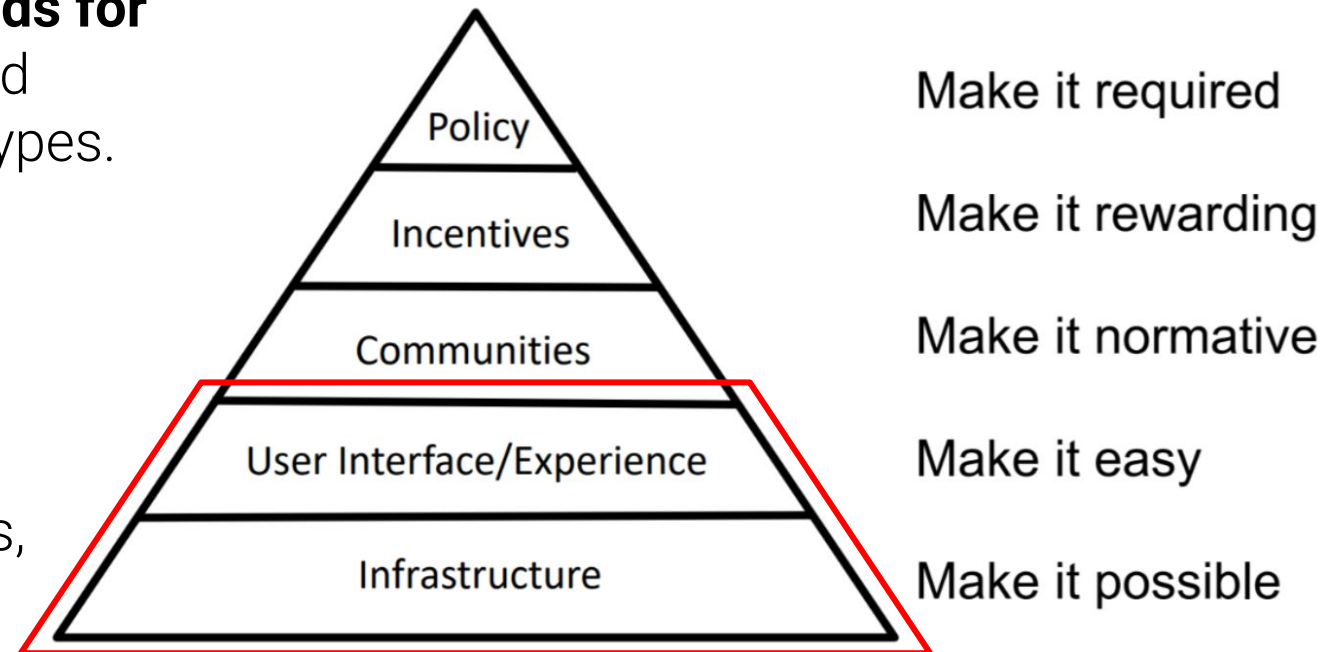
Agenda

Welcome and introduction	Philipp
Discipline Requirements and Needs	Helene
User Journey Maps	Giacomo
Q & A	All
Wrapping up	Philipp



Goal and focus of today's webinar

- Present results from our ongoing work on the identification of **requirements and needs for long-term data preservation** within and across scientific disciplines and data types.
- Focus on **user perspectives**, and what strategies and tools we have been and will be using to identify user needs and **reduce pain points** for researchers, repositories, and archives.



Motivational factors and driving forces behind change of research culture. From Nosek (2015).

EOSC EDEN Project

Summary

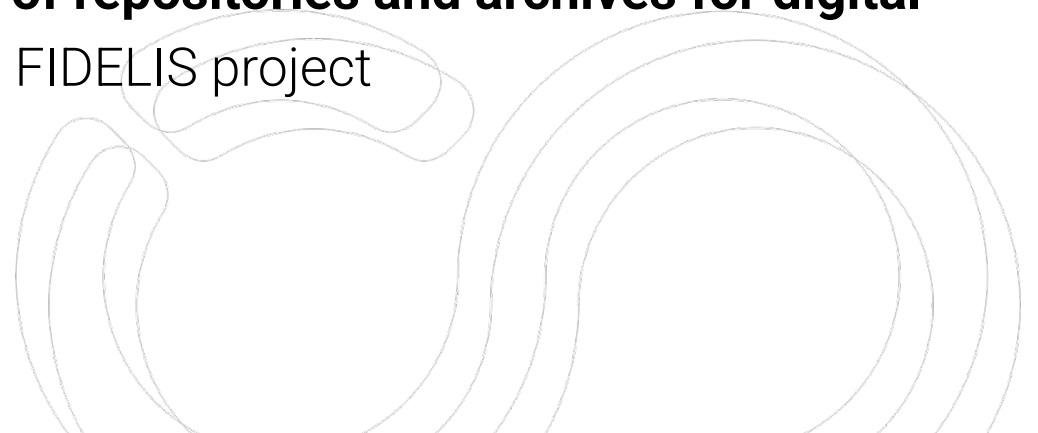
The EOSC EDEN project is funded by the European Union under the Horizon Europe framework programme and aims to **support and promote digital preservation practices and standards** at the European and national levels.

This support will be complemented by the **development of user-centric tools, services, and standards** designed to facilitate the creation of a distributed European infrastructure for digital data preservation, retention, and access.

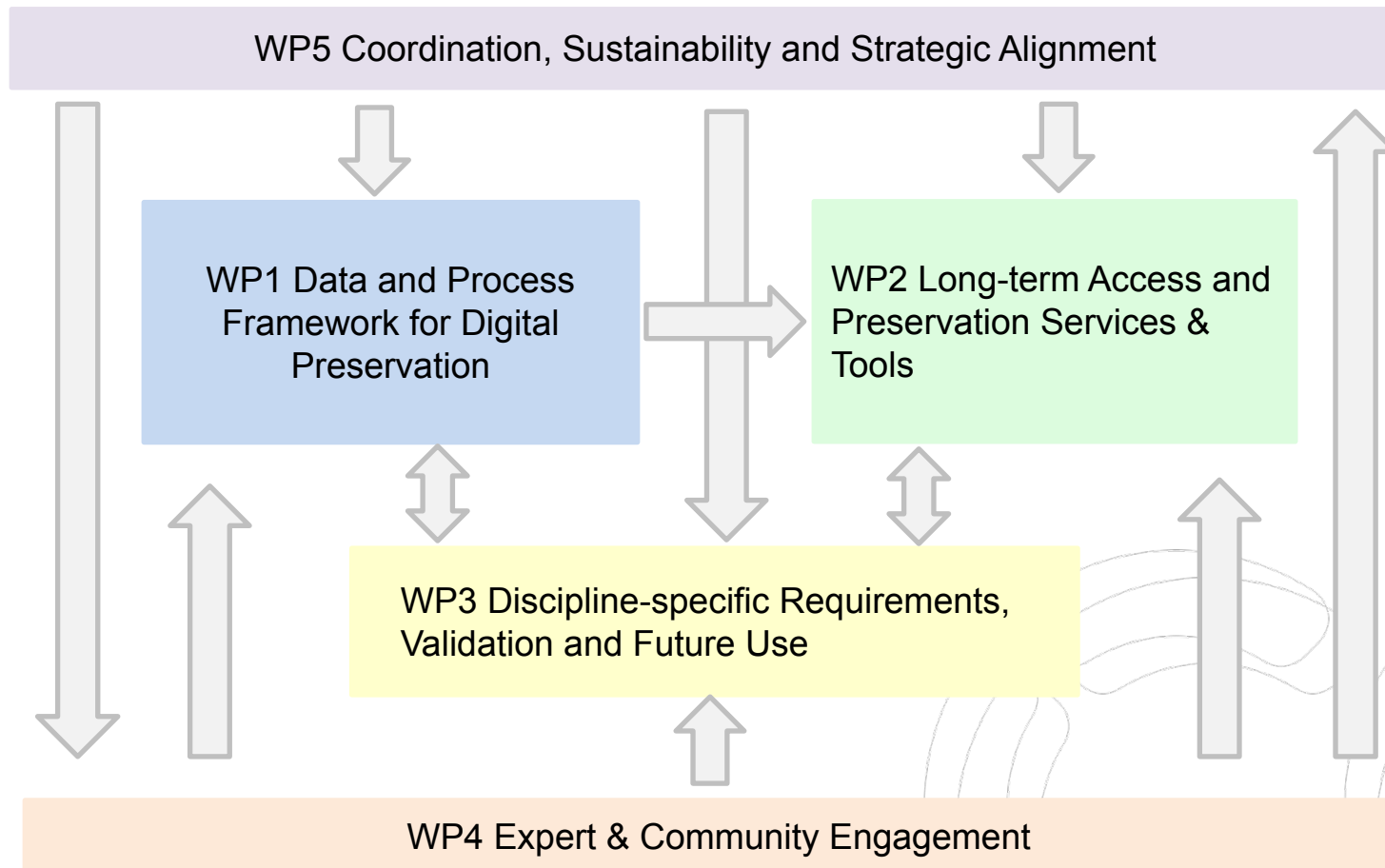


EOSC EDEN Objectives

- **Objective 1:** To establish a general **framework and practices** to support the creation of curation, digital preservation, and access strategies in Europe
- **Objective 2:** To **enrich EOSC with tools** to store and access digital data for long periods, automate and federate certain specialised curation and preservation tasks
- **Objective 3:** To increase **adoption** of curation, digital preservation and access practices within different **scientific disciplines**
- **Objective 4:** To boost the **data curation and quality in Europe**
- **Objective 5:** To identify and consolidate a **network of repositories and archives for digital preservation** within EOSC in collaboration with the FIDELIS project



Project Structure and Work Packages



WP3 Discipline-specific Requirements, Validation and Future Use



Links the project to discipline and data-type specific needs and requirements



Iteratively provide discipline-specific, and cross-disciplinary / data type-specific requirements for digital preservation and data quality to WP1 and WP2

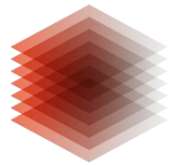


Validate and enhance the new digital preservation framework (WP1) and tools (WP2) via pilot testing



Provide a support-kit to empower discipline and data type-oriented communities to adopt, extend and use the new digital preservation framework (WP1) and fit-for-purpose tools (WP2)

EOSC EDEN Consortium



TIB LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY



Arkivum



Swiss Institute of
Bioinformatics

The early-adopter disciplines

- Provide expertise in defining requirements and needs
- Pilot test and validate the framework (WP1) and tools (WP2)
- Adopt and adapt the project outcomes
- Contribute to the creation of a support-kit to guide and support users in adopting and adapting the framework and tools



Climate Simulations



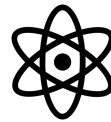
Earth & Environmental Sciences



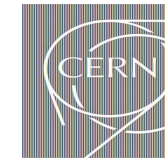
Universität
Bremen



Food Sciences



High-Energy Physics



Life Sciences and Bioinformatics



Swiss Institute of
Bioinformatics



Linguistics



Social Sciences



UK Data Service

Meeting our audience



Please answer our two questions in this poll!

<https://nettskjema.no/a/590045>



Discipline Requirements and Needs (Helene)



WP3 Discipline-specific Requirements, Validation and Future Use



Links the project to discipline and data-type specific needs and requirements



Iteratively provide discipline-specific, and cross-disciplinary / data type-specific requirements for digital preservation and data quality to WP1 and WP2



Validate and enhance the new digital preservation framework (WP1) and tools (WP2) via pilot testing



Provide a support-kit to empower discipline and data type-oriented communities to adopt, extend and use the new digital preservation framework (WP1) and fit-for-purpose tools (WP2)

T3.1 Define discipline-specific requirements and needs

Objective

- Iteratively provide discipline-specific, and cross-disciplinary, and digital object-type specific requirements for long-term preservation and digital object quality to WP1 and WP2
- Three methods: desk-based mapping, interviews, survey questions
- The analyses will address the entire data stewardship lifecycle

Primary focus in year 1 of the project

- The seven early-adopter disciplines



Data collection Y1

- Desk-based mapping of existing requirements & identification of gaps and needs

- *Where?* Repositories and other discipline-specific communities.
- *How?* Examination of online documentation relevant for long-term preservation and data quality: Policy and strategy documents, online protocols, formally published documents, etc.

Discipline FORD lvl 2	Technical Data Quality				
FORD Classification LVL 2	Data type	File Format	File size Dataset size	File Quantity	checksums / data integrity
1.5 Earth and related environmental sciences	dataset, data collection, publication, software, poster, presentation	Strongly recommended: Network Common Data Format (NetCDF). GRIB, Binary (GRIB), CSV, ASCII, Zarr (zipped). Only open source formats for preservation that are well established in the climate science community.	<=8GB, Dataset: TB	project quota	yes, CF checker encouraged before data deposit, CF check at (meta)data review
1.5 Earth and related environmental sciences	geospatial	NetCDF	File: GB, Dataset: TB	up to 1000	yes
1.5 Earth and related environmental sciences	climate related datasets, forecasts, analyses	Netcdf, WMO-GRIB, SQL in relational DB (I)			
1.5 Earth and related environmental sciences	Broad range of data types, field observation and measurements, binary content, time-series data, experimental data etc.	If possible data is transferred to the relational system (tabular data), community specific binary data also accepted and archived as is; Accepted file formats are diverse: https://wiki.pangaea.de/wiki/Format	Not defined	Not defined	"...for binary files also checksums and file size, absolute location in bucket store..."
1.5 Earth and related environmental sciences	environmental data from ENVRI RIs				"...Further work on curation has considered also other, wider, aspects. In particular:

Data collection Y1

- Interviews to collect user stories, community norms and expectations
 - *Who?* Repository managers and researchers
 - *How?* Questions thematically structured around the data lifecycle, with separate sets developed for the two stakeholder groups

A: Overview interviews	
Researcher 1	
Discipline	Partner
Climate Simulations	DKRZ
Earth & Environmental Sciences	U Bremen
Food Sciences	Premotec
High-Energy Physics	CERN
Life Sciences & Bioinformatics	SIB
Linguistics	UiT
Social Sciences	UKDS
Researcher 2 (if time/available)	
Discipline	Partner
Climate Simulations	DKRZ
Earth and environmental sciences	U Bremen
Food Sciences	Premotec
High-Energy Physics	CERN
Life Sciences & Bioinformatics	SIB
Linguistics	UiT
Social Sciences	UKDS
Repository manager 1	
Discipline	Partner
Climate Simulations	DKRZ
Earth & Environmental Sciences	U Bremen
Food Sciences	Premotec
High-Energy Physics	CERN
Life Sciences & Bioinformatics	SIB
Linguistics	UiT
Social Sciences	UKDS
Repository manager 2 (if time/available)	
Discipline	Partner
Climate Simulations	DKRZ
Earth & Environmental Sciences	U Bremen
Food Sciences	Premotec
High Energy Physics	CERN
Life Sciences & Bioinformatics	SIB
Linguistics	UiT
Social Sciences	UKDS

T3.1 Interview guide

1. Setting the context for the interview

- 1.1 Researcher Context
- 1.2 Repository Context

2. Interview questions

2.1 Researchers

- 2.1.1 Starting question/"ice breaker"
- 2.1.2 The incentives for data deposit
- 2.1.3 Long-term preservation of data
- 2.1.4 Interaction with repositories
- 2.1.5 The purpose of data deposit in trustworthy repositories
- 2.1.6 Reuse of data: Knowledge, norms & attitudes
- 2.1.8 End of interview

2.2 Repositories

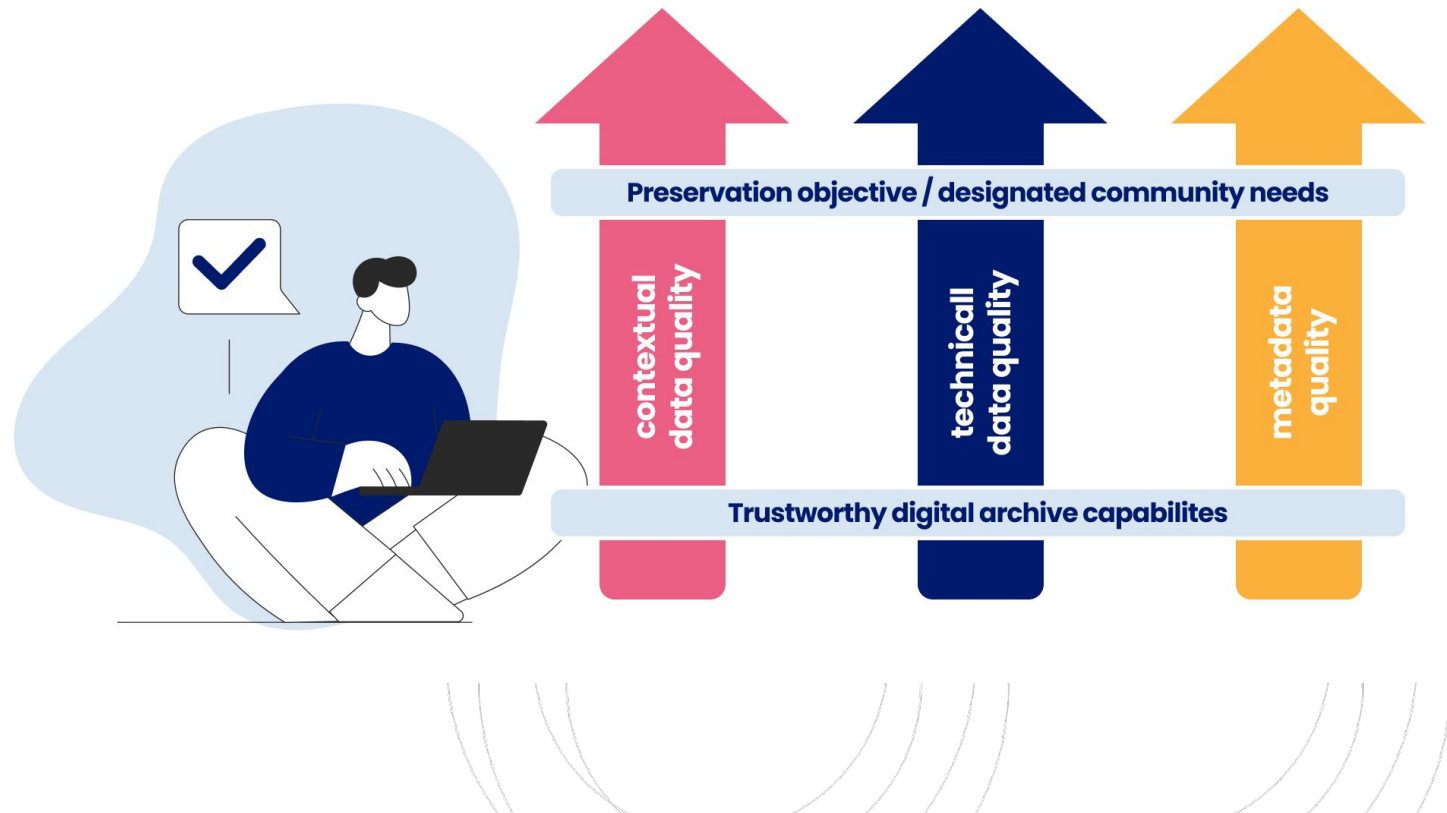
- 2.2.1 Starting question/ "Ice breaker"
- 2.2.2 Long-term preservation policy
- 2.2.3 Governance & resources
- 2.2.4 Community watch
- 2.2.5 Technology watch
- 2.2.6 End of interview

Data organisation and analysis

Categorise - combine - reuse

Structured around the **Re-Use fitness model** - a core component of EOSC EDEN.

→ Serves as **frame** for assessing aspects that are important for the evaluation of a repository's strategies for **identification, selection and (re)appraisal** of digital objects for long-term digital preservation



Step 1

Collection and categorisation of data, per entry/interviewee

High-Energy Physics

- DBM 7 entries
- 4 interviewees

Technical Data Quality				metadata quality			
	File size Dataset size	File Quantity	checksums / data integrity	Metadata standard (e.g. DublinCore, DataCite)	Controlled Vocabulary (Repository required)	License	PID
ended: Network Common	<=8GB, Dataset: TB	project quota up to 1000	yes, CF checker encourage	ISO 19115, DataCite, DublinCore	CF conventions	CC0	DOI, PID
IB, SQL in relational DB (File: GB, Dataset:TB		yes		CF	yes, various licenses not further specified	DOI
transferred to the relation	Not defined	Not defined	"...for binary files also checked"	ISO 19115	Essential Climate Variables	https://wiki.pangaea.de/wiki/License	DOI ... (ORCID)
1915 compatible	not defined	not defined	"...Further work on curation"	CERIF, ISO19115/INSPIRE, DC, DCAT, CKAN, Sensor	Used: "...ChEBI, EnvO, WoRMS, ITIS or QUDT..."		PID in general
s, but generally .xlsx, .t	dataset 2 GB			ISO 19136 (INSPIRE)	GEMET	not specified CC as example	
0, SEED, stationXML, JS	not defined or depending on provide	not defined	unknown	Goldstein et al., 2014 https://ecl.eearthchem.org/view/StationXML , DataCite (some stations have DOI)		CC0 for metadata, CC-BY for data	DOI
ity of nodes	not defined			CERIF, DC, DCAT, ISO 19115		SeisComp Licenses (see: https://orfeus.eu)	DOI
HDF5, CSV, images	large files supported (TB)	unlimited		ISO 19115 compliant via Geonetwork		CC BY 4 recommended	DOI
ended: RINEX, mSEED, c				ISO19115, NASA GCOMD DiF, DataCite	GCOMD vocabulary	CC BY 4 recommended	DOI, IGSN, ORC
				EuroFIR standard	EuroFIR thesauri		
			"...Further work on curation"	METROFOOD standard	METROFOOD thesauri		
				EuroFIR standard	EuroFIR thesauri		
	depends on the user	depends on the user		EuroFIR standard	EuroFIR thesauri	FoodCASE license	
				Own standard			
	(M) <50GB		(M) Zenodo includes checks	yes (M) - metadata follows zenodo format and exportabl		(M) - users chooses license	(M) DOI provided
ion: Addition to publicati	HEPDATA:						CERN Open Data
							DOI
							DOI
L, auto convert into . CSV	"The volume of data, both actual and 0.5 to 10 PB	No direct reference but implies big data se	"Host institutions should mo	States that data validation i		Not specified but recommends open acco	DOI for policies :
osition repository	N/A. It is not a deposition repository	N/A. It is not a deposition repository		community standards	LIPID MAPS classification system	Creative Commons Attribution 4.0 Int	yes
					yes, ChEMBL schema, https://ftp.ebi.ac.uk/pub/databases	Creative Commons Attribution-Share Alike	yes
osition repository	N/A. It is not a deposition repository	N/A. It is not a deposition repository			ChEBI Ontology	Creative Commons License (CC BY 4.0)	yes
TL, JSON, ZIP, GZ, TAR N/A. It is not a deposition repository	N/A. It is not a deposition repository	N/A. It is not a deposition repository	but data not informed, but all data is https://www.ebi.ac.uk/ontology/	UBERON, HGNC gene names, NCBI taxonomy, Confider		Creative Commons Zero license (CC0)	yes
IML, JSON, GTF, GZ, TA N/A. It is not a deposition repository	N/A. It is not a deposition repository	N/A. It is not a deposition repository	yes and all data are curated https://www.ebi.ac.uk/ontology/	HGNC		Ensembl imposes no restrictions on access	yes
						Creative Commons Attribution (CC BY) i	yes, includes DC
						CC0, CC BY and CC BY-NC. In general yes, includes DC	yes
						Creative Commons Attribution 4.0 Int	yes
					Repository-developed ontology or controlled vocabulary	Not specified. However, no use restrictio	yes
					yes, Rhea schema	Creative Commons Attribution-ShareAlike	yes
					yes, Controlled vocabulary is provided for each metadata	Creative Commons Attribution 4.0 Intern	yes
					logue of Life	reative Commons Public Domain (CC0)	yes, DOI
					NCBI Taxonomy, Gene Ontology (GO), the Nomenclature	CC0 1.0 Universal (CC0 1.0)	own persistent ic
					thub.com/EE	Experimental Factor Ontology, GWAS Catalog's standard	CC0 unless otherwise stated. GWAS Ca
					NCBI, Gene Ontology (GO), AGI (Arabidopsis Genome In	https://creativecommons.org/licenses/by/4.0/	yes
					EC (Enzyme Commission) number, Ensembl, HGNC, UniP	https://creativecommons.org/licenses/by/4.0/	yes, EC number
						https://creativecommons.org/licenses/by/4.0/	yes
					Taxonomic names based on global taxonomies, BIN syste	Open access with citation requirements	Yes, BIN and DC
					Multiple, dependent on checklist and attribute (https://gencc-1.0)	CC0-1.0	
					GenBank accession numbers, LPSN and the NCBI taxon	Creative Commons License (CC BY 4.0)	yes, DOI
					Entrez Gene ID, HGNC, Ensembl ID...	https://creativecommons.org/licenses/by/4.0/	

community st	[on researcher's motivation to deposi	-	"Well, data is stored -
scribe data qu	[On reasons to deposit]: "The first it is	"First, we have s	"For the project? It -
	Repositories mu		Data should be reta
updated with n	Metadata completeness (e.g., coordinati	Multiple data cer	Minimum 10 years (Compliance with World

metadata quality	Trustworthy digital archive capabilities
Access ± Embargo ± Restrictions	COMMENTS
Deposit criteria	Choice of repository
Backups	retention policies
Compliance with frameworks/policies	
-	[On adherence to community st
-	[On researcher's motivation to deposi
-	[On reasons to deposit]: "The first it is
Moratoriums on data access should	Metadata should describe data qu
-	Older datasets are updated with n
license needed	A food compiler, this is how we ca
It's very well collection is done by t	Most phd researchers and even just the
Mentions of implementing embargo	Challenges in maintaining metada
Restricted content is harvested for	Metadata quality is manually impr
Data is usually released six years	Metadata collection is automated
Open Data	
Embargo can be set as needed, af	for GEO, it's a good compromise t
Embargo until paper is out	(about the license) "If it was for me
4 access levels. Form to fill out and	-
Four levels of access. 1) open; 2)	-
-	-
-	-
"we have a three tier license and a	-
"most of my data is not secure or c	-

Desk-based mapping

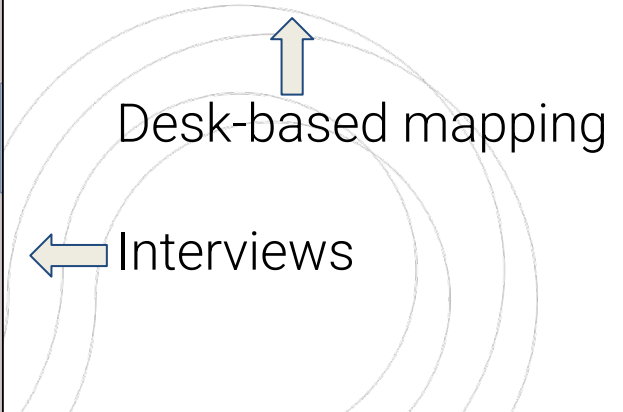
Interviews

Step 2

Synthesis of data per component, per discipline

- Requirements/existing practices
- Knowledge gaps/needs

Trustworthy Digital Archive Capabilities		Preservation Objective / Designated Community Needs	
Existing requirements/practices	Future Needs / Knowledge Gaps	Existing Requirements/practices	Future Needs / Knowledge Gaps
locations. JSON-LD are used on the landing pages, so metadata can be harvested by aggregators. The repository does not see funding as being endangered, but there would be issues transferring the data (over 8 petabytes) if funding was lost. [Researcher] A big motivation for the researcher to deposit their data was because of the cost of producing it, so would like it to be used by others as well. There was also requirements from funders. They did believe there were issues in the community of researchers not sharing data unless required to.		publication process. Also have participated in outreach activities (like booths at conferences) to encourage researchers to archive their data. After the end of the 10 year preservation period, data is supposed to be reappraised. [Researcher] Researcher was aware of the need for strong documentation and metadata to be able to reuse data over time. Researcher noted issues following the criteria for deposit across different repositories in the domain.	
Metadata completeness and suitable file formats are required by repository. Researcher demands repositories to ensure long-term funding, persistent identifiers, and compliance with FAIR principles. Multiple data centers and tape archives are used for backups by repository. Data is expected by researchers to be retained for at least 30 years due to its scientific value. Repositories guarantee 10 years, but have records intact for more than 25 years. Researcher follow FAIR principles and uses CC BY licencing while repository is compliant with World Data System standards and has established API interfacing. Repository ensure long-term support and financing by two host institutions under a legally binding cooperation agreement.	[Gap] Versioning issues were highlighted as problematic in commercial software. [Need] Repositories should provide web map services for geographic data discovery.	Researcher uses open-source software like MB-System for web map services. Repository tries to stay on top of new developments and implement more efficient tools and systems. Motivation for long-term archiving due to the irreplaceable nature of bathymetry data and its scientific value. Repository has a long history of datasets still intact. Researcher addresses the need for repositories to provide clear guidelines on metadata requirements and data formats to ensure proper data submission. Repository provides community workshops, a wiki page for guidance, and videos on YouTube to train users on data preparation and retrieval.	
Researchers don't deposit their data, they don't use the repositories (but some of the data/articles are published due to the phd/organisational requirements), as a trusted places they choose platforms recommended by experts or well-known institutions, regardless of formal certification or metadata. Repo persons use the tools and practices depending to the repository, usually login and password is the most important aspect from their side. Guidelines exist by repositories, and researchers follow these as well as optimizing data size prior to deposition. Generally acceptance for choosing a main data portal in the discipline. Data is expected by researcher to exist past the experiment lifetime with repositories being able to be taken down data on request. Governance of repositories through robust frameworks which are understood by researcher. But without mentions of financial burden plans. Service-level agreements in one repository in case of service cease to exist, the other does not have a strategy.	[Gap] Not mentioned through the interview but there should be some connections and communication between researchers and repository people. [Need] Storage costs mentioned as an issue by one repository. [Need] Exit strategy not existing in one repository. [Gap] Researcher dependent on good communication with repository to learn efficient deposition.	Data is backed up, preserved, and older versions are retained alongside current data to support long-term analysis. In food discipline there is a user group which meet from time to time and give their feedback about their needs and requirements. Researchers can be a part of this. Repositories is either focusing on preservation or a combination including content curation. From the researcher perspective the data curation is thorough. Repositories have internal routines which involve programs to perform various tasks like file conversion. Tools are also used by the researcher. Repositories provide manual guidance for deposition. However, this is seen as beneficial for the researcher as the repositories give effective guides on what and how to deposit data.	
The choice of repository for depositing data is mostly community-driven (repositories are already established for the different types of data), other times for convenience (e.g. Zenodo which can host generic data, which would not fit one of the specialised repositories). One researcher mentioning the lack of dedicated repositories for imaging data (high-resolution, therefore large image files). Repositories usually have a multi-site redundancy scheme, e.g. spread of over US / Europe / Japan. Researchers would require preserving data "forever" (one researcher mentioning the concrete example where a dataset was reused and highly important 15 years after its initial publication), however repositories face great challenges in ensuring long-term funding to sustain long-term data preservation, with no funding scheme allowing more than 4-to-maximum-5-year span. Even (e.g. Elixir) Core Data Resources face the same challenges. Therefore long-term preservation of data is tied to the funding cycles. The current prospect on funding is worsening (e.g. due to situation at the NIH). In case of a complete lack of funding, the repositories would migrate data to e.g. FTP, but its value would quickly decrease due the lack of active curation (data on proteins is constantly evolving with newly published research, therefore a snapshot of a protein data repository can quickly become outdated).		Data in repositories is curated by specialists (highly trained, specialised curators), with community curation also playing a (small) role, e.g. via contact forms provided by the repository. Repositories note the importance of training for researchers to become aware of all the available data they can (re)use, as the growth in complexity makes it extremely difficult even for specialists to be aware of all the types of data being offered by a (large) repository.	



Preservation Objective / Designated Community Needs		Contextual Data Quality	
Existing Requirements	Future Needs / Knowledge Gaps	Existing Requirements	Future Needs / Knowledge Gaps
1 = Generalist vs. Specialist curator 2 = Level of Curation A-Z 3 = Reassessment Schedule 4 = Original/Access copies		1 = Required Documentation 2 = Data Rights / Consent held for Publication 3 = Language 4 = GDPR / Sensitive data evaluation 5 = Data Reproduction proof	
3 entries. 2 IDs (23,66) providing information. Specialist curators providing some level of curation (A & C). One ID provides reassessment.	GAP: Sparse coverage in this discipline, in particular Reassessment and if Original/access copies of digital objects are retained NEED: Consistent Level of curation NEED: Better documented preservation objectives overall	3 entries (ID23, ID57 and ID66): - Required documentation: 2 empty, 1 yes. - Data Rights / Consent held for publication: 2 empty and 1 stating: "terms of use" - Language: 1 empty, 1 stating "English", 1 stating "English & German" - GDPR/ Sensitive data: 2 empty, 1 stating "no sensitive data" - Reproduction proof for data (provenance, raw data, etc.) : 1 empty, 1 stating "yes", 1 stating "provenance".	[Yes] Gap Required documentation [Yes] Gap Data Rights / Consent [Some] Gap Language [Yes] Gap GDPR/ Sensitive data (low relevance in discipline?) [Some] Gap Reproduction proof for data [Need 1] The discipline Climate Simulations needs to present explicit requirements on Required documentation, Gap Data Rights / Consent, GDPR/ Sensitive data [Need 2] The discipline Climate Simulations needs to supplement the explicit requirements on Language and Reproduction proof for data
	GAP: No information on Reassessment checks or if	8 entries (ID32, 33, 60, 61, 62, 63, 64, and ID65) - Required documentation: 2 empty, 4 yes, 1 stating "terms of use", 1 explaining that only discipline metadata is required, no readme. - Data Rights / Consent held for publication: 5 empty, 1 stating "via CC license selection on submission form", 1 comment stating: "terms of use", and 1 stating "information on data rights as ideal (so needed)". - Reproduction proof for data (provenance, raw data, etc.): 1 yes, 1 stating "ISO Lineage (Provenance)"	[Some] Gap Required documentation [Yes] Gap Data Rights / Consent [Some] Gap Language [Yes] Gap GDPR/ Sensitive data (low relevance in discipline?) [Yes] Gap Reproduction proof for data [Need 1] The discipline Earth & Environmental Sciences needs to present explicit requirements on Data Rights, GDPR and Reproduction proof [Need 2] The discipline Earth & Environmental Sciences needs to supplement the explicit requirements on Required documentation and Language

Step 3

Summary of each of the five components, across disciplines

- Identification of commonalities and specificities in the existing requirements
- Identification of emerging needs in the discipline-specific repositories and communities

4.2 Commonalities and Specificities in Existing Requirements

Within the area of *Contextual Data Quality*, all seven early-adopter disciplines highlight the provenance information along with contextual metadata (e.g. a full description of the data, its origin, reuse and method replication). Another commonality in this category is the need for disciplines that handle sensitive data regularly (Linguistics, Social Sciences, etc.) to have mechanisms. Yet, ensuring proper ethical handling through e.g. anonymization is challenging. One specificity that stands out in the Contextual Data Quality category is Energy Physics ownership of data is often transferred to repository operators.

Regarding *Technical Data Quality*, a commonality across the disciplines is the need for standardized file formats. This move away from proprietary formats is particularly relevant for Physics, and, in the case of a climate data repository, non-proprietary formats. Numerous accepted file formats are listed within the existing requirements, with the most common being CSV, FASTA, JSON, NETCDF, PDF, ROOT, SIF, etc. Another commonality is the use of YAML, and also DOCX and XLSX as proprietary file formats. A specific guideline is the lack of an explicitly stated restriction on the use of sensitive data. Whether this means an actual lack of restriction or is – more precisely – a need for clear guidelines. Another issue in this area of Technical Data Quality is highlighted by the Life Sciences & Bioinformatics (other disciplines), where duplicated data is a problem to be solved.

Regarding *Metadata Quality*, the use of persistent identifiers is common across all disciplines, but also other types of persistent identifiers. Another commonality is the need for various metadata standards (both discipline-specific and agreed across disciplines) to attain consistency in repositories. Almost all disciplines, and software licenses – when used – are open source. A specific guideline is the use of known within Life Sciences & Bioinformatics as well) the use of sensitive data is highlighted, which is less emphasized in other disciplines.

In the area of *Trustworthy Digital Archive Capabilities*, most repositories highlight deposit criteria that stipulate data must be discipline- and for

4.3 Emerging Needs in Discipline-Specific Repositories and Communities

In this section, we draw attention to the main needs observed across the desk-based mapping and interviews.

4.3.1 Contextual Data Quality

The most critical needs for this category are ensuring that robust data documentation can support preservation and data utility. There is also a strong need to ensure that sensitive data – including data that person-identifying, contain commercial links, or have security concerns –, as well as data involving Indigenous communities, receive appropriate consideration. Although all disciplines indicate that dataset-accompanying documentation is required, both the form and content of this documentation is undefined. It is strongly recommended to provide standardised, accessible, and comprehensive methods for applying documentation to datasets, intended to exist alongside standardized metadata schema. Arising specifically from the interview results is a conflation of metadata documentation vs. data-specific documentation, where the latter provides detailed information on data provenance (e.g. methodologies, codebooks, re-use), such as a stand-alone ReadMe file. There is a strong need to create an inclusive, standardized documentation tool (e.g. interactive “ReadMe” document builder) that has the possibility to be scaled both within, and across, discipline-specific needs.

- The tool mentioned above should also be comprehensible for both depositors and repositories, contain explicit requirements for ascertaining data rights and obligations, noting that different disciplines will have different obligations and rights agreements.
- The robustness of metadata standards in older datasets, or datasets involving unique data capture contexts (e.g. recordings), must be improved and brought to a standardized level that allows for long-term understandability, and access that is as open as possible but as closed as necessary. Automated tools that are able to process older datasets plus their documentation – including potentially fragmented and incomplete metadata and stand-alone documentation files – and subsequently equip them with robust, standardized documentation, metadata and licences, are required.
- Awareness about GDPR concerns surrounding sensitive data, as well as consideration of the CARE principles must increase, especially for disciplines focussed in the Humanities and Social Sciences. However, universal tools to ensure GDPR-compliance and/or CARE-adherence are recommended to be rolled out across all disciplines. Some disciplines state that sensitive data is not handled and therefore legal guidelines such as GDPR requirements are not applicable. Similarly, the impact of sc

The screenshot shows the Zenodo project page for 'D3.1 - Report on Discipline Requirements and Needs'. The page is published on July 1, 2025, and is version V1.0. It is marked as a 'Project deliverable' and is 'Open'. The authors listed are: Andreassen, Helene N.¹; Flügel, Anna-Lena²; Klemetsen, Terje¹; Smart, Kathleen¹; Benauer, Maria³; Cannizzaro, Giacomo⁴; Ciesla, Malwina⁵; Conzett, Philipp¹; Huber, Robert⁶; Heier, Svein¹; Laik, Arijit⁴; Lammert, Andrea⁷; Le Meur, Jean-Yves⁸; L'Houers, Hervé^{9, 10, 11}; Lindlar, Micky³; Mehl, Florence¹²; Mendes de Farias, Tarcisio¹²; Middlebos, Wesley⁸; Parkes, Oliver^{11, 10, 9}; Presser, Karl¹³; Sima, Ana Claudia¹⁴; Molloy, Laura (Other)¹⁵; Snyder, Kyle Patrick (Other)⁴; Märkälä, Anu (Other)¹⁶. A 'Show affiliations' button is visible. The project description states: 'Report presenting methodology and results from the first EOSC EDEN data collection within WP3, aimed at describing discipline and data-type specific requirements. Data are collected among the seven early adopter disciplines in the project. Recommendations, based on observed commonalities and needs, are structured around the Re-use Fitness model.'

(Andreassen et al., 2025, all data available, except for first stage of interview transcript analysis)

A heterogeneous information landscape

- **Well-established & mature**

- Clear stand-alone documents or wikis adhering to established standards
- Data scientist experts, developers, digital preservationists

- **Operation on reduced scale**

- Often dependent on other units for IT and archival services
- Uneven presence of services and implementation of standards
- Often scholar-led, close connection with research community but with reduced capacity for development

May excel on specific areas that are critical for the type of data they work with, e.g. discipline-specific metadata, GDPR compliance

The challenge of representativity

We have chosen a wide approach to capture as **diverse practices/requirements and needs** that can be related to long-term preservation and data quality, as possible.

This approach has put some **limits** on the extent of representativity and the level of detail; we cannot report the complete diversity of existing requirements, needs and gaps within a given discipline. Still, the data offer a **broad starting point** that can be built on in other tasks and phases of the project (and beyond).



Commonalities and specificities in existing requirements

Component	Common requirement	Specific requirement Linguistics*
Contextual data quality	Data provenance information and contextual metadata important for data reuse	Consent mechanisms
Technical data quality	Preference for open and standardised file formats	-
Metadata quality	Common use of persistent identifiers	Multiple access levels & licenses to protect sensitive data
Trustworthy digital archive capabilities	Clear deposit criteria that stipulate that data must be discipline- and format-specific	-
Preservation objective/designated community needs	Specialised curators, but with varying levels of long-term preservation expertise	Ethical and/or legal framework for preserving sensitive data

*To a large extent also relevant for the Social Sciences → data type-specific requirements

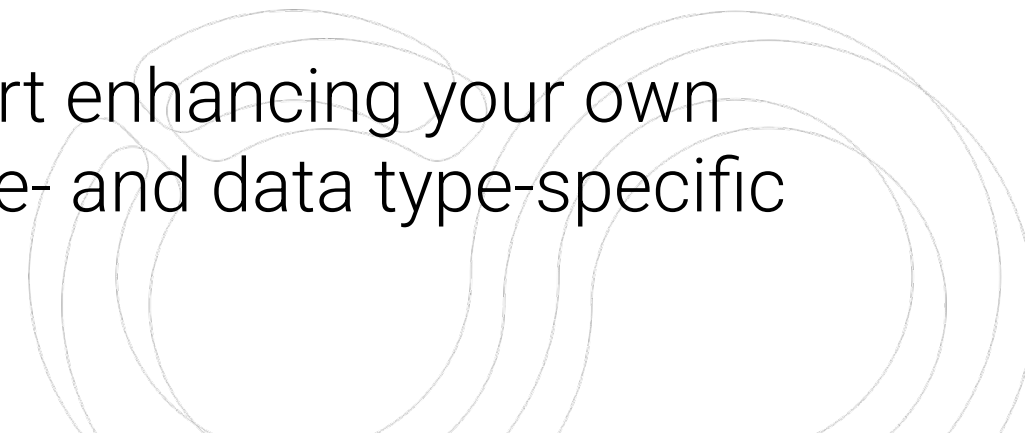
Emerging needs

Component	Need	Possible measure
Contextual data quality	Robust data documentation to support preservation and data utility	Standardised, accessible, comprehensive (and scalable) methods for applying documentation to datasets
Technical data quality	Effective management of file formats to ensure preservation, accessibility and interoperability	Systematised monitoring of acceptable file formats
Metadata quality	Researchers' understanding/adherence to metadata standards; metadata standards interoperability across disciplines	Flexible, discipline-agnostic metadata standards to facilitate interdisciplinary research, data discovery and data transfer
Trustworthy digital archive capabilities	Standard, comprehensive and mandatory policies	Retention policies, financial planning, exit strategies, versioning support, ...
Preservation objective/designated community needs	Strategies for long-term preservation and re-appraisal	Standardised and comprehensive strategic reappraisal and reassessment schedule and guidelines

Y2 of T3.1:
More fine-tuned and narrow data collections

Up next: How to systematise, visualise and contextualise user needs and pain points

As a warm-up: By joining our Slido, start enhancing your own knowledge on how to identify discipline- and data type-specific needs!





To the support staff: How do you engage with your community of users to identify their changing needs for long-term preservation over time ?



To all: After deposit, how long should the type of data you typically work with, be kept available? Could they be replaced or reproduced in the future?

User Journey Maps (Giacomo)



T3.2 Pilot methodology and validation



Links the project to discipline and data-type specific needs and requirements



Iteratively provide discipline-specific, and cross-disciplinary / data type-specific requirements for digital preservation and data quality to WP1 and WP2



Validate and enhance the new digital preservation framework (WP1) and tools (WP2) via pilot testing



Provide a support-kit to empower discipline and data type-oriented communities to adopt, extend and use the new digital preservation framework (WP1) and fit-for-purpose tools (WP2)

Testable outputs

WP1



Develops processes as requirements for implementation (WP2) and community adoption (WP3, WP4)



Existing practices for identification, selection and appraisal of data for digital preservation



Requirements in digital preservation processes for **re-use fitness** of digital objects



Framework to identify candidates to digital preservation based on use, benefit and quality



Model for re-appraisal points along data lifecycle

WP2



Delivers reference implementations and supporting tools to integrate repositories and services into the EOSC Federation



Registry of digital preservation services and tools



Publish machine-interoperable services for curation and preservation actions



Identify standards and protocols for submit and exchange Information packages



Use case implementation and testing of services and tools

WP3 Discipline-specific Requirements, Validation and Future Use



Links the project to discipline and data-type specific needs and requirements



Iteratively provide discipline-specific, and cross-disciplinary / data type-specific requirements for digital preservation and data quality to WP1 and WP2



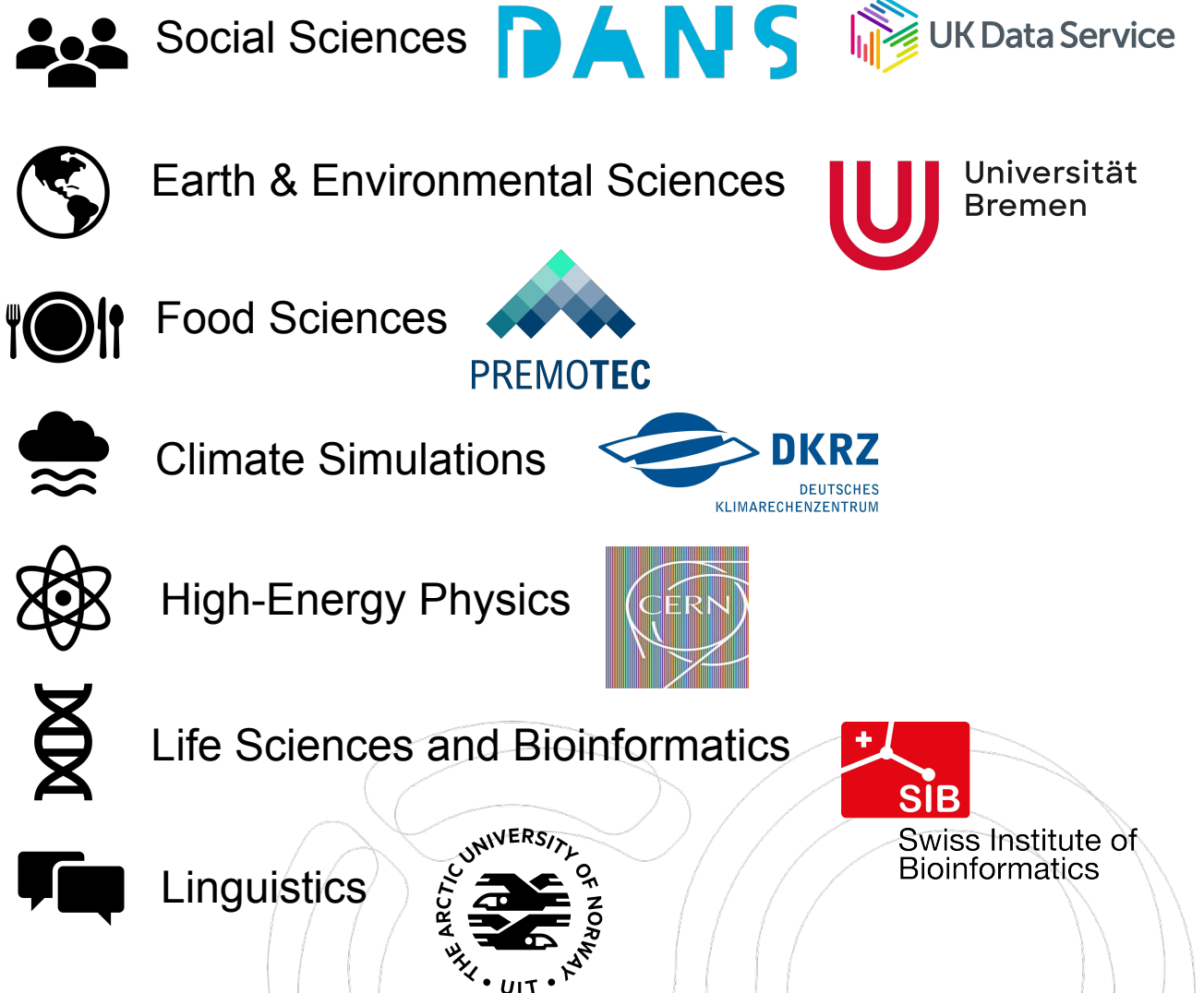
Validate and enhance the new digital preservation framework (WP1) and tools (WP2) via pilot testing



Provide a support-kit to empower discipline and data type-oriented communities to adopt, extend and use the new digital preservation framework (WP1) and fit-for-purpose tools (WP2)

Discipline-oriented pilots

- Requirements and needs from disciplinary perspective
- Testing and validation
- Early adoption
- Support kit to endorse wider adoption of the project results



Discipline-oriented pilots

- Requirements and needs from disciplinary perspective
- Testing and validation
- Early adoption
- Support kit to endorse wider adoption of the project results



M3.1 User Journey Maps



D3.2 Pilot Methodology



Pilot Methodology

Triggered at each WP release

WP

Release

WP outputs
+
Supporting assets

Hands off testing
package

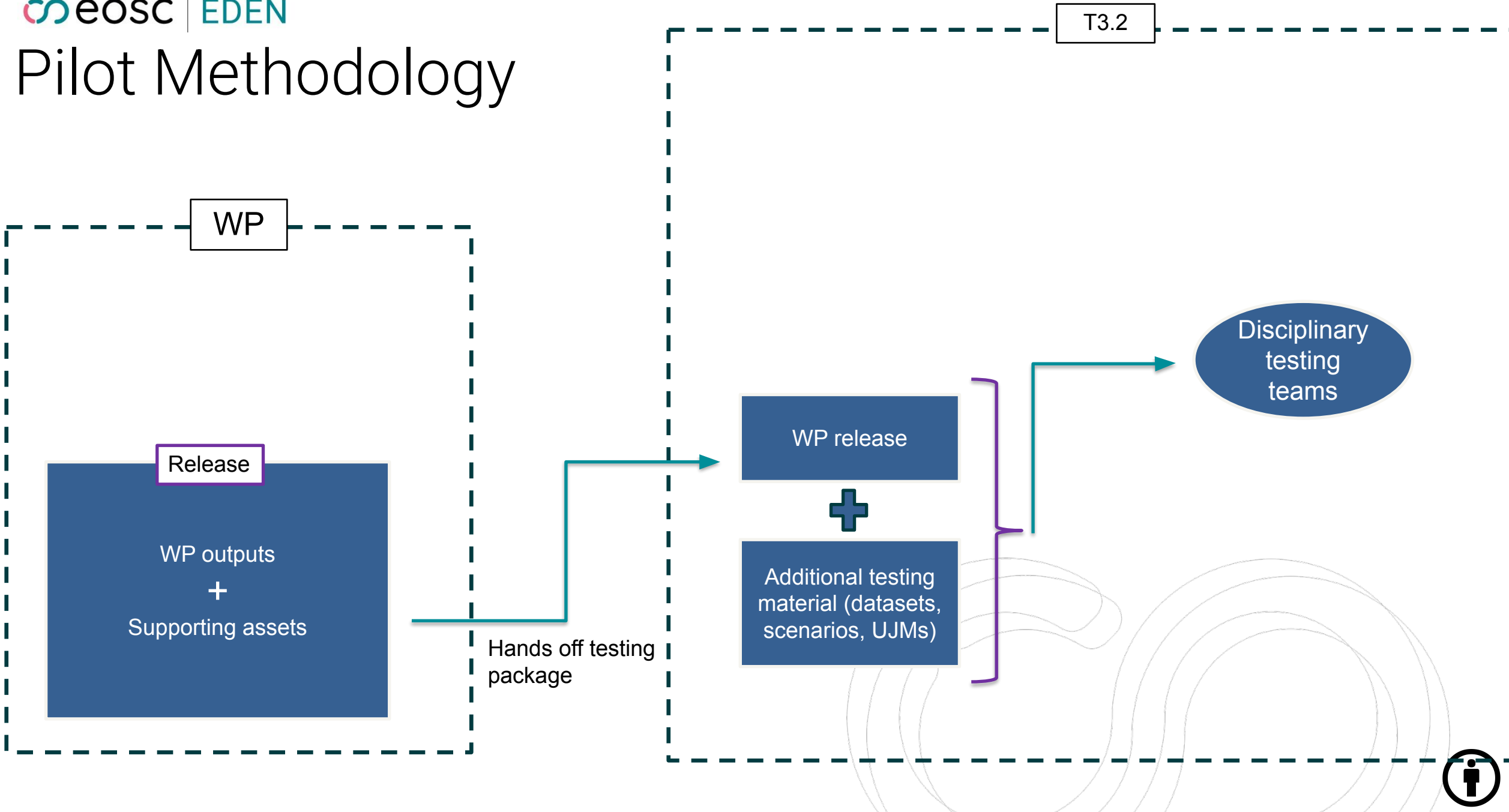
WP release

T3.2

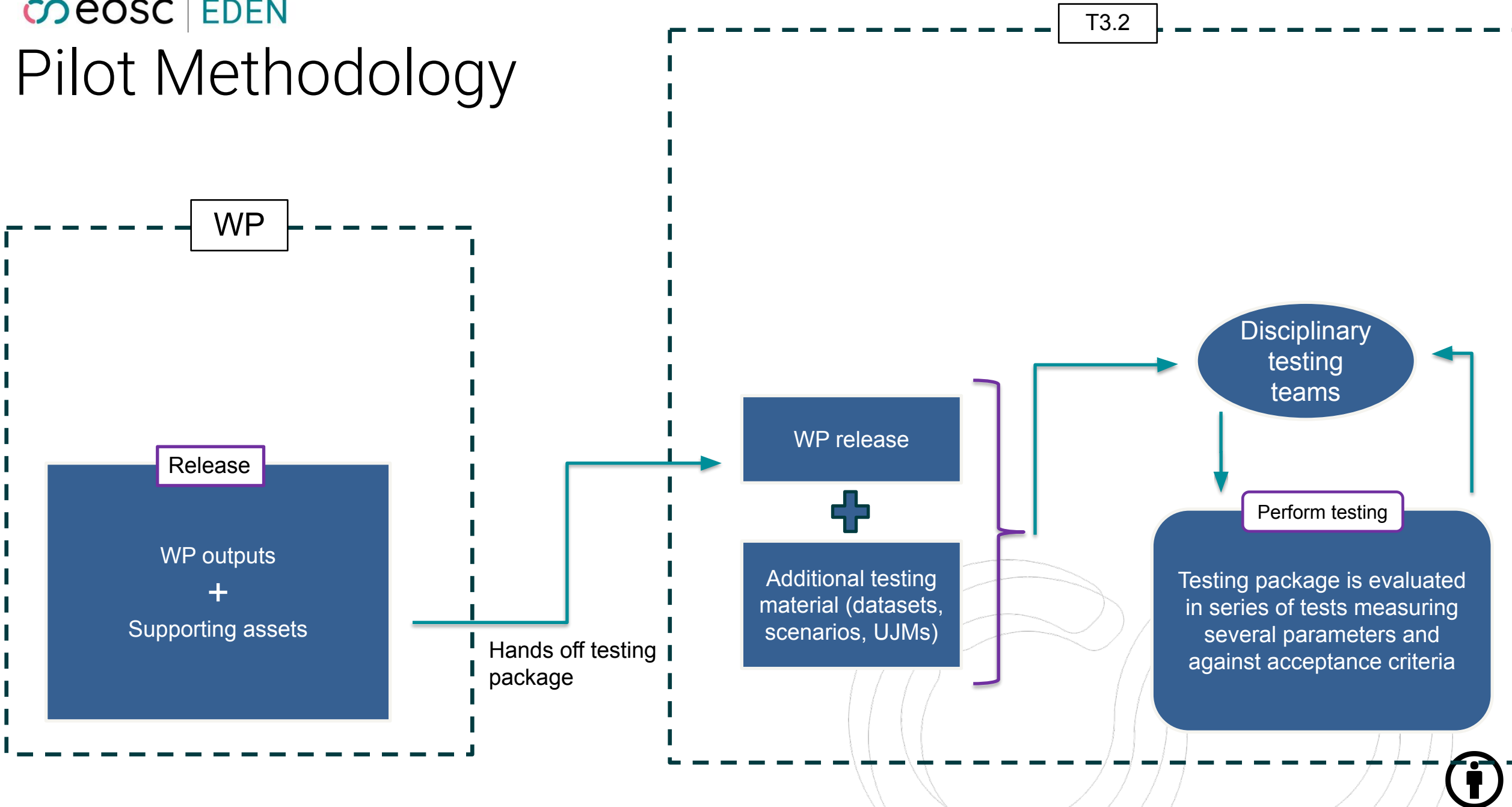
Disciplinary
testing
teams



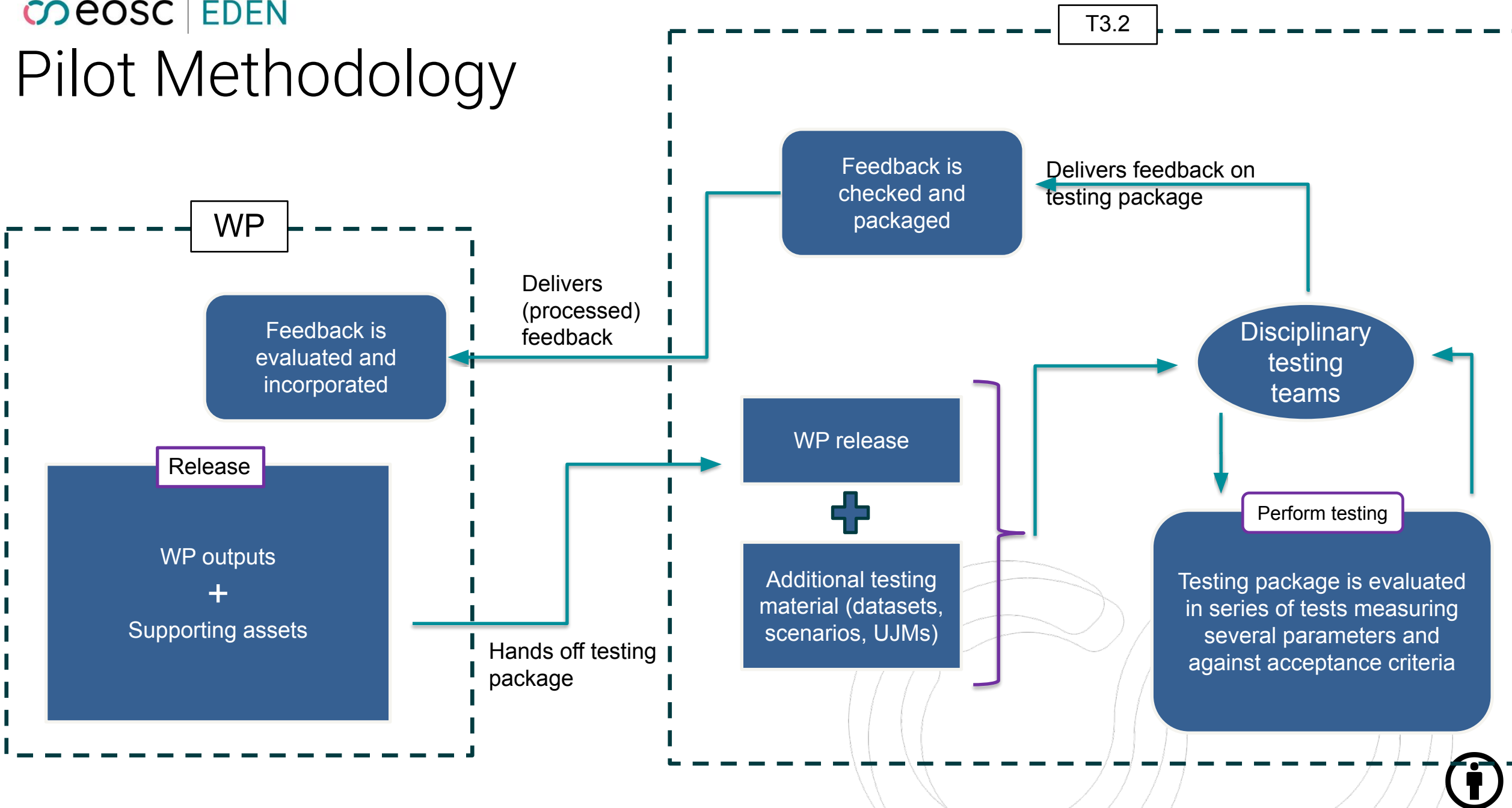
Pilot Methodology



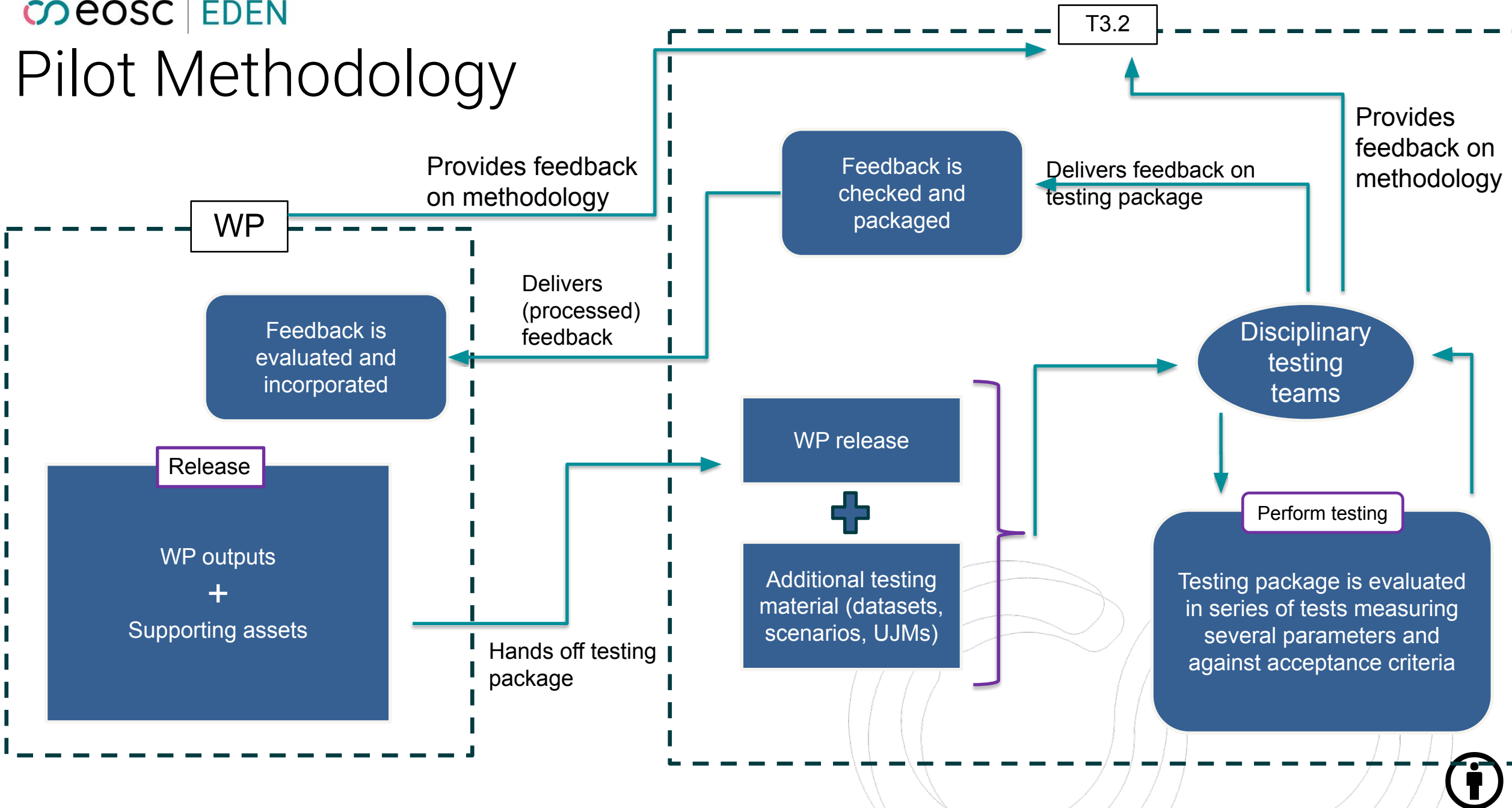
Pilot Methodology



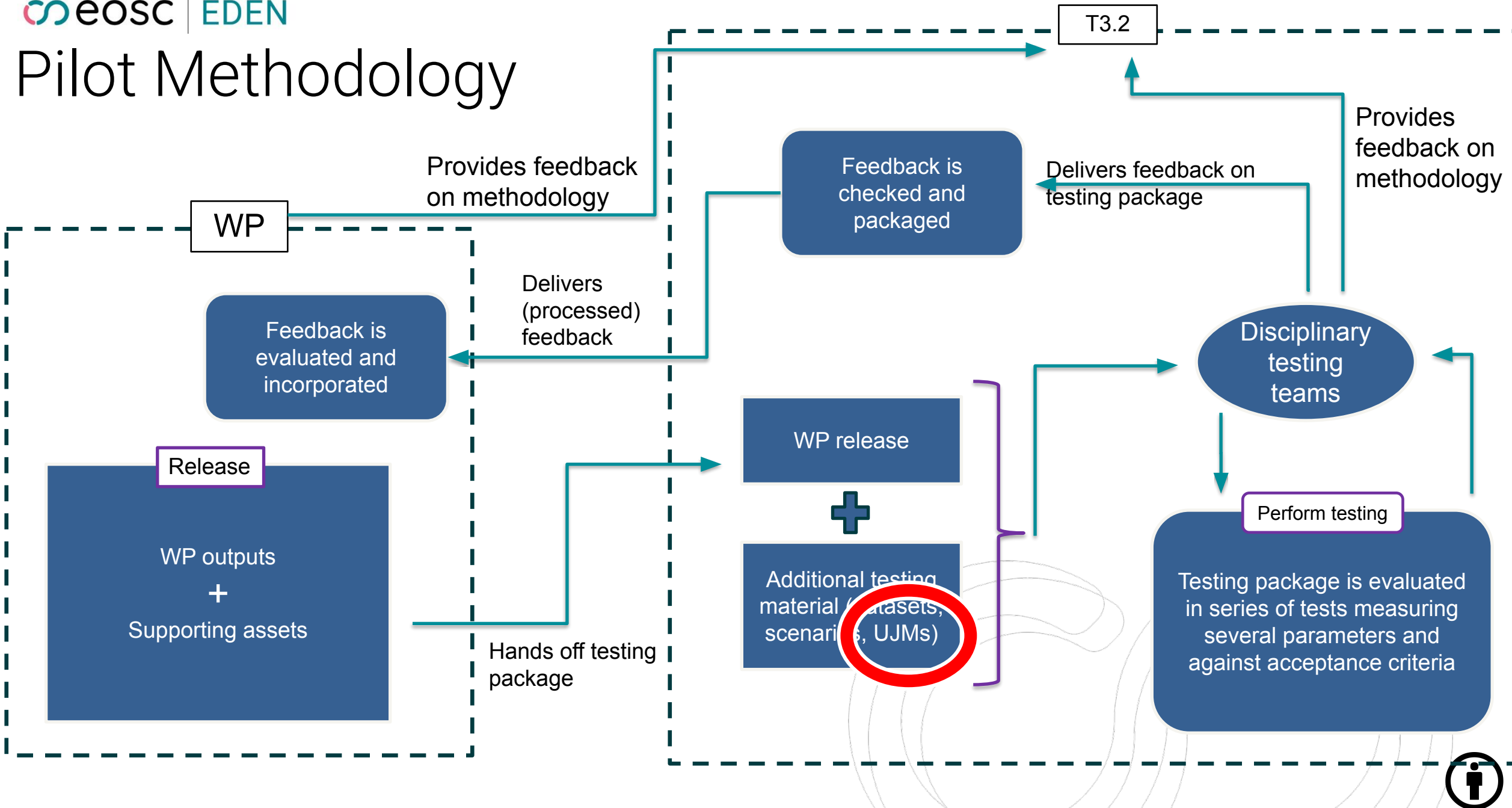
Pilot Methodology



Pilot Methodology



Pilot Methodology



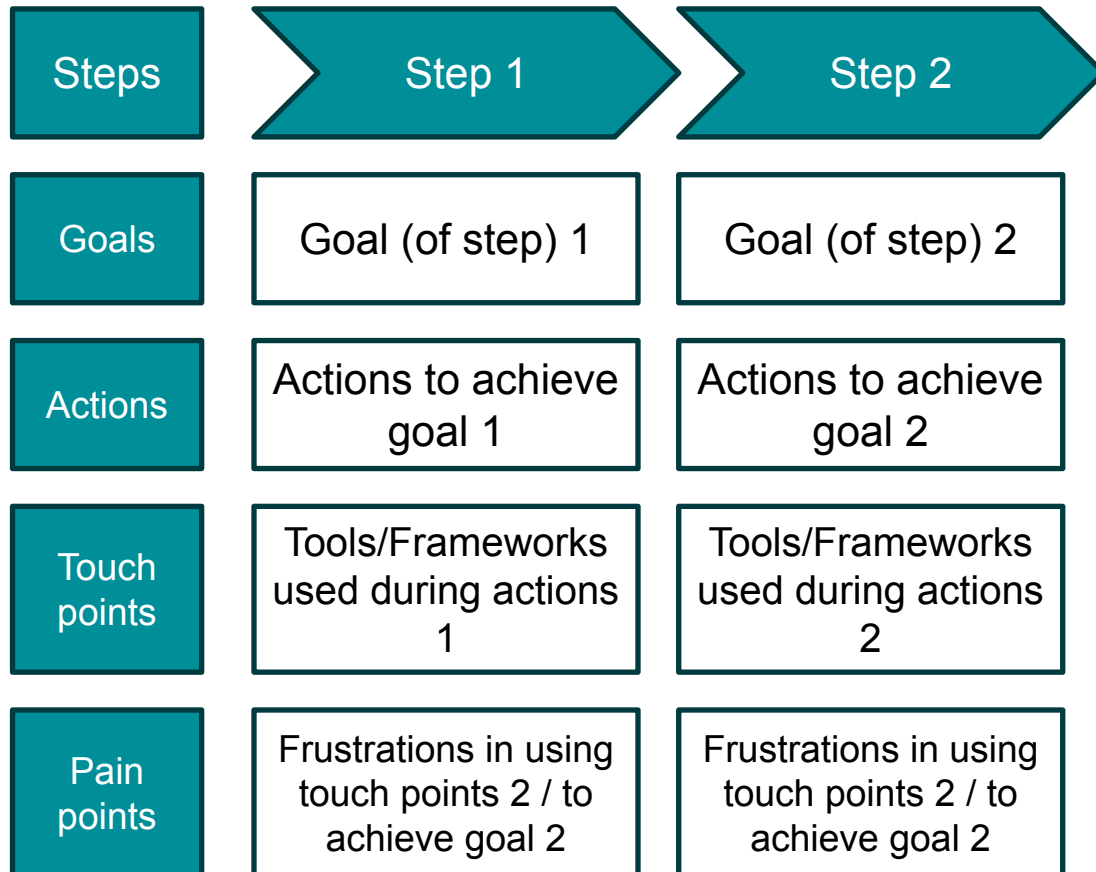
User Journey Maps

A visualisation of the **process** that a **user** goes through to accomplish a **goal**



User Journey Maps

A visualisation of the **process** that a **user** goes through in order to accomplish a **goal**



Clearly Visualise

- Steps
- Goals
- Actions
- Touchpoints
- Frustrations experienced in the process



EDEN M3.1 - User Journey Maps

UJM 1 "ingest and appraisal"

actions from repository staff and end-user sides

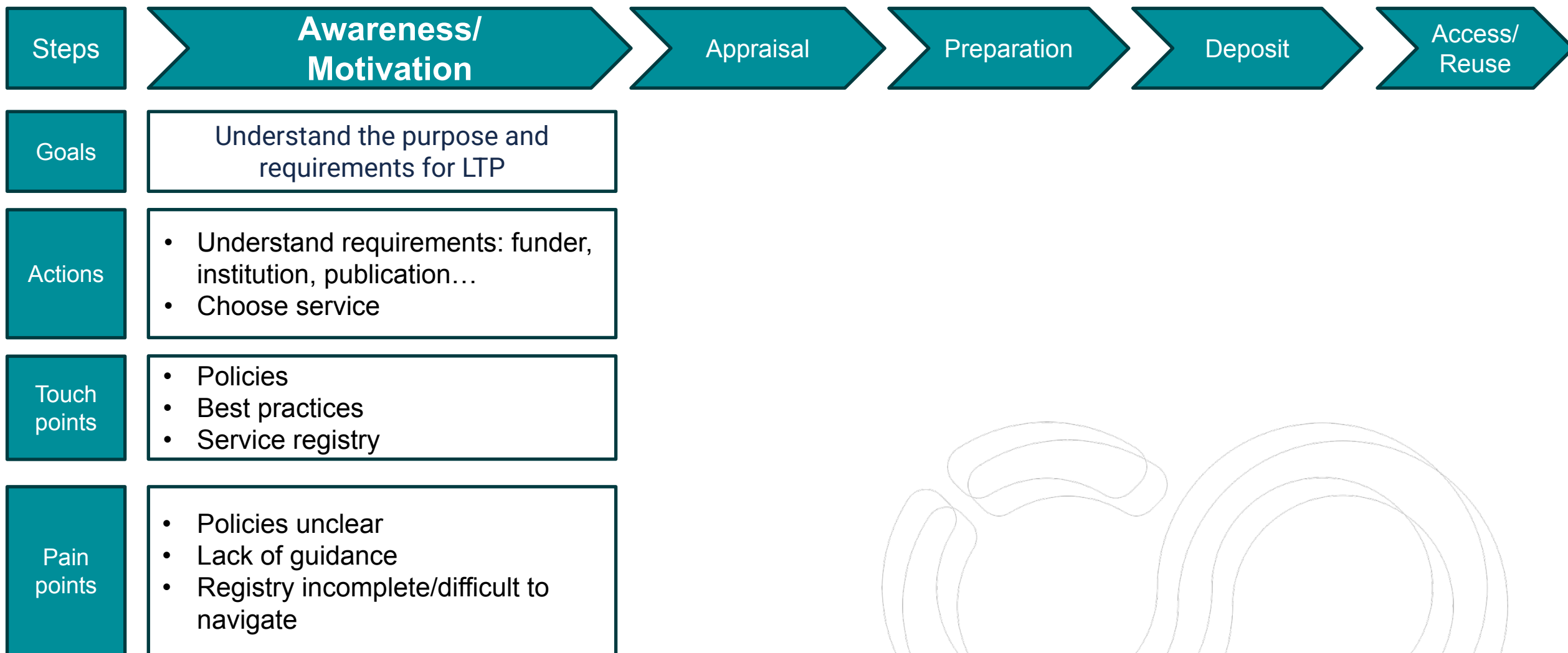


UJM 2 "reappraisal and de-accession"

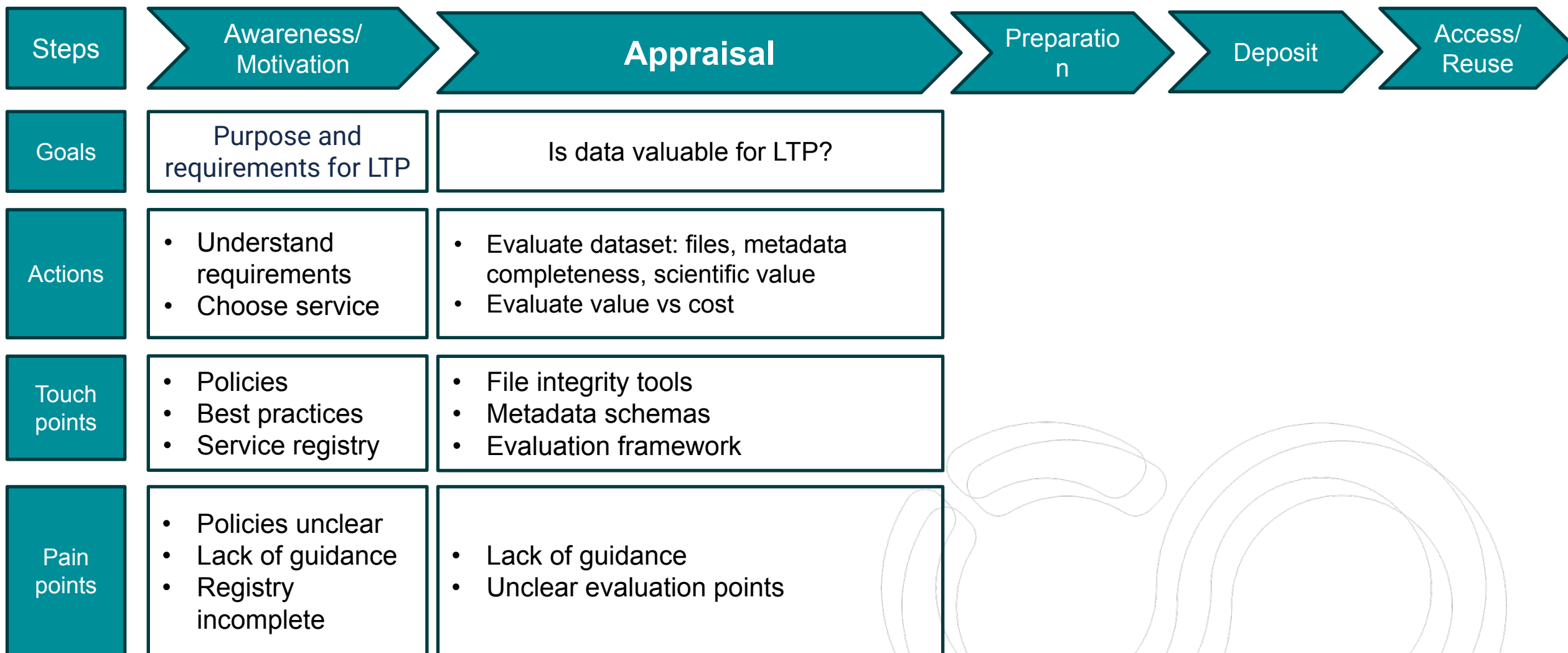


General UJM → Discipline-specific

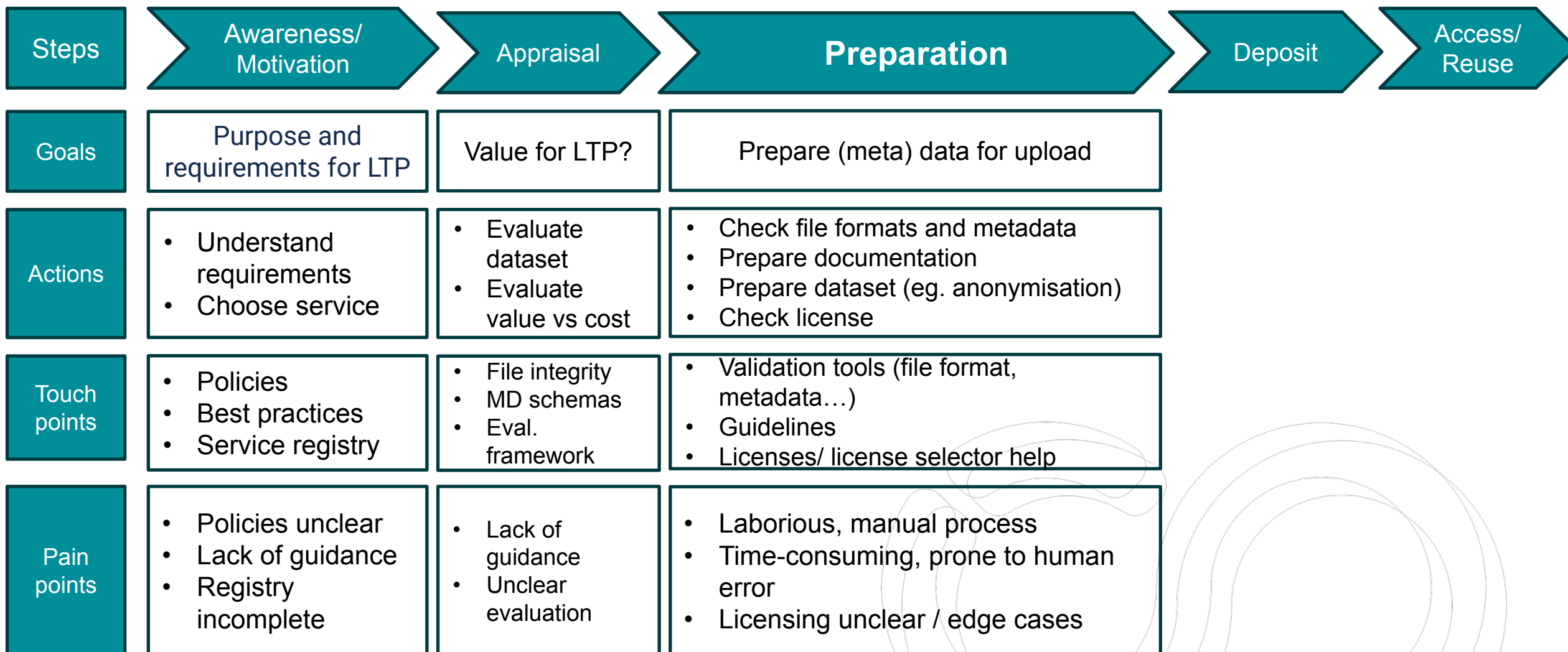
UJM 1 "ingest and appraisal" - User's point of view



UJM 1 "ingest and appraisal" - User's point of view



UJM 1 "ingest and appraisal" - User's point of view



UJM 1 "ingest and appraisal" - User's point of view

Steps	Awareness/ Motivation	Appraisal	Preparation	Deposit	Access/ Reuse
Goals	Purpose and requirements for LTP	Value for LTP?	Prepare data	Upload dataset with full information	
Actions	<ul style="list-style-type: none"> Understand requirements Choose service 	<ul style="list-style-type: none"> Evaluate dataset Evaluate value vs cost 	<ul style="list-style-type: none"> Prepare/check file formats, metadata, documentation 	<ul style="list-style-type: none"> Upload data Checksum Fill metadata Select license 	
Touch points	<ul style="list-style-type: none"> Policies Best practices Service registry 	<ul style="list-style-type: none"> File integrity MD schemas Eval. framework 	<ul style="list-style-type: none"> Validation tools Guidelines 	<ul style="list-style-type: none"> Repository web app or API File integrity checks License helptool 	
Pain points	<ul style="list-style-type: none"> Policies unclear Lack of guidance Registry incomplete 	<ul style="list-style-type: none"> Lack of guidance Unclear evaluation 	<ul style="list-style-type: none"> Laborious, manual process prone to human error 	<ul style="list-style-type: none"> Manual metadata entry Licenses incomplete/edge case Lack of automation Repository app is clunky 	

UJM 1 "ingest and appraisal" - User's point of view

Steps	Awareness/ Motivation	Appraisal	Preparation	Deposit	Access/Reuse
Goals	Purpose and requirements for LTP	Value for LTP?	Prepare data	Upload dataset	Access public dataset
Actions	<ul style="list-style-type: none"> Understand requirements Choose service 	<ul style="list-style-type: none"> Evaluate dataset Evaluate value vs cost 	<ul style="list-style-type: none"> Prepare/check file formats, metadata, documentation 	<ul style="list-style-type: none"> Upload data Checksum Fill metadata Select license 	<ul style="list-style-type: none"> Look for and find dataset Download and reuse dataset
Touch points	<ul style="list-style-type: none"> Policies Best practices Service registry 	<ul style="list-style-type: none"> File integrity MD schemas Eval. framework 	<ul style="list-style-type: none"> Validation tools Guidelines 	<ul style="list-style-type: none"> Repository app File integrity License help 	<ul style="list-style-type: none"> Data discovery services (PIDGraphs) Repository app
Pain points	<ul style="list-style-type: none"> Policies unclear Lack of guidance Registry incomplete 	<ul style="list-style-type: none"> Lack of guidance Unclear evaluation 	<ul style="list-style-type: none"> Laborious, manual process prone to human error 	<ul style="list-style-type: none"> Manual process edge cases Repository app is clunky 	<ul style="list-style-type: none"> Dataset not findable (PIDs?) Not easy to download (embargo, AAI issues) MD not complete -> difficult to reuse (provenance etc)

UJM 1 "ingest and appraisal" - User's point of view

Steps	Awareness/ Motivation	Appraisal	Preparation	Deposit	Access/Reuse
Goals	Purpose and requirements for LTP	Value for LTP?	Prepare data	Upload dataset	Access public dataset
Actions	<ul style="list-style-type: none"> Understand requirements Choose service 	<ul style="list-style-type: none"> Evaluate dataset Evaluate value vs cost 	<ul style="list-style-type: none"> Prepare/check file formats, metadata, documentation 	<ul style="list-style-type: none"> Upload data Checksum Fill metadata Select license 	<ul style="list-style-type: none"> Look for and find dataset Download and reuse dataset
Touch points	<ul style="list-style-type: none"> Policies Best practices Service registry 	<ul style="list-style-type: none"> File integrity MD schemas Eval. framework 	<ul style="list-style-type: none"> Validation tools Guidelines 	<ul style="list-style-type: none"> Repository app File integrity License help 	<ul style="list-style-type: none"> Data discovery services (PIDGraphs) Repository app
Pain points	<ul style="list-style-type: none"> Policies unclear Lack of guidance Registry incomplete 	<ul style="list-style-type: none"> Lack of guidance Unclear evaluation 	<ul style="list-style-type: none"> Laborious, manual process prone to human error 	<ul style="list-style-type: none"> Manual process edge case Repository app is clunky 	<ul style="list-style-type: none"> Dataset not findable (PIDs?) Not easy to download (embargo, AAI issues) MD not complete -> difficult to reuse (provenance etc)



- ✓ Clearer policies
- ✓ Registry of services and tools (WP2)

- ✓ Metrics for reuse (WP1)
- ✓ Guidance (WP3)

- ✓ Automated tools (WP2)
- ✓ Support-kit (WP3)



What is a pain point you may experience in your “ingest and appraisal” journey?

Next up: Q & A

Before we stop the recording:

- Thank you to our speakers, Helene and Giacomo!
- Thank you to the audience for attending and your interaction!
- Please keep yourselves up-to-date on EOSC EDEN by using any of our communication channels!



eden-fidelis.eu



[linkedin.com/company/eosc-eden](https://www.linkedin.com/company/eosc-eden)



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



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



<https://eden-fidelis.eu/#newsletter>



github.com/EOSC-EDEN



[EOSC EDEN Zenodo Community](https://zenodo.org/communities/eosc-eden)

[#EOSCEDEN](https://twitter.com/EOSCEDEN)

Reminder: Meeting our audience



Please answer our two questions in this poll!

<https://nettskjema.no/a/590045>



Q & A



Wrapping up

- Presented and discussed ongoing work within EOSC EDEN on the **identification of requirements and needs for long-term data preservation**.
- Our focus has been on **user perspectives**, i.e. requirements, needs, and pain points from a **discipline** and **data type** point of view.
- Thank you for your **feedback** and interaction! We'd be happy to **get you involved in EOSC EDEN beyond this webinar**. How? See our contact information on the next slide.



eden-fidelis.eu



[linkedin.com/company/eosc-eden](https://www.linkedin.com/company/eosc-eden)



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



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



<https://eden-fidelis.eu/#newsletter>



github.com/EOSC-EDEN



[EOSC EDEN Zenodo Community](https://zenodo.org/communities/eosc-eden)

[#EOSCEDEN](https://twitter.com/EOSCEDEN)

Thank you!



**Funded by
the European Union**



References

Andreassen, H. N., Flügel, A.-L., Klemetsen, T., Smart, K., Benauer, M., Cannizzaro, G., Ciesla, M., Conzett, P., Huber, R., Høier, S., Laik, A., Lammert, A., Le Meur, J.-Y., L'Hours, H., Lindlar, M., Mehl, F., Mendes de Farias, T., Middlebos, W., Parkes, O., ... Märkälä, A. (2025). *D3.1 - Report on Discipline Requirements and Needs* (V1.0). Zenodo. <https://doi.org/10.5281/zenodo.15789261>

Nosek, B. (2015). Shifting Incentives from Getting It Published to Getting it Right. <https://osf.io/zvp8k/>

