https://fair2adapt-eosc.eu/

RO-Crates and ROHub platform

Fair Digital Objects (FDOs) enabling services in EOSC

Daniel Garijo (UPM), Raúl Palma (PSNC)

Sebastian Luna-Valero (EGI)

Funded by the European Union

The FAIR2Adapt Consortium

Grant No: 101188256

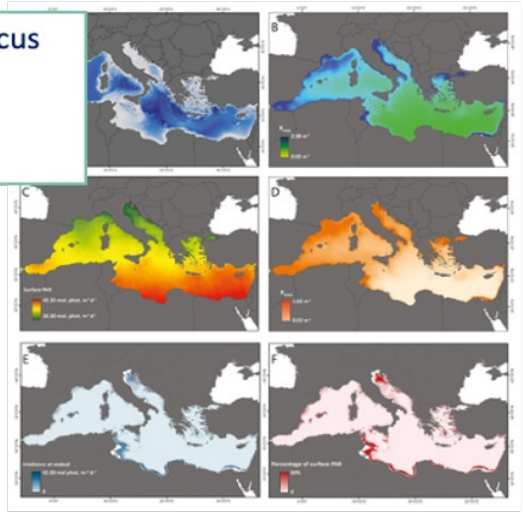
CC BY

How Scientists usually work (Example from Earth Sciences)

Data

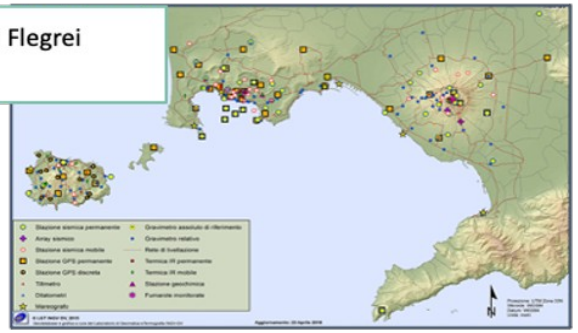
- Remote sensing data

Copernicus
NASA
WOA
...



- In-situ data from local networks

Vesuvius - Campi Flegrei (Italy)



Software/methods

- Data processing → Commercial and/or Personal tools (Matlab®, ENVI-IDL®, Python, R, etc.)
- Data modelling → Personal tools (Python, R, Matlab®, ENVI-IDL®, etc.)
- Post-processing → GIS (ArcGIS®, QGIS), GMT, etc.

Computing facilities

- Personal workstations
- Institutional HPC

Data – results management & dissemination

- Papers publication containing results and/or linked to data
- Results presented at meetings with ppt/poster stored
- Personal storage

Collaboration

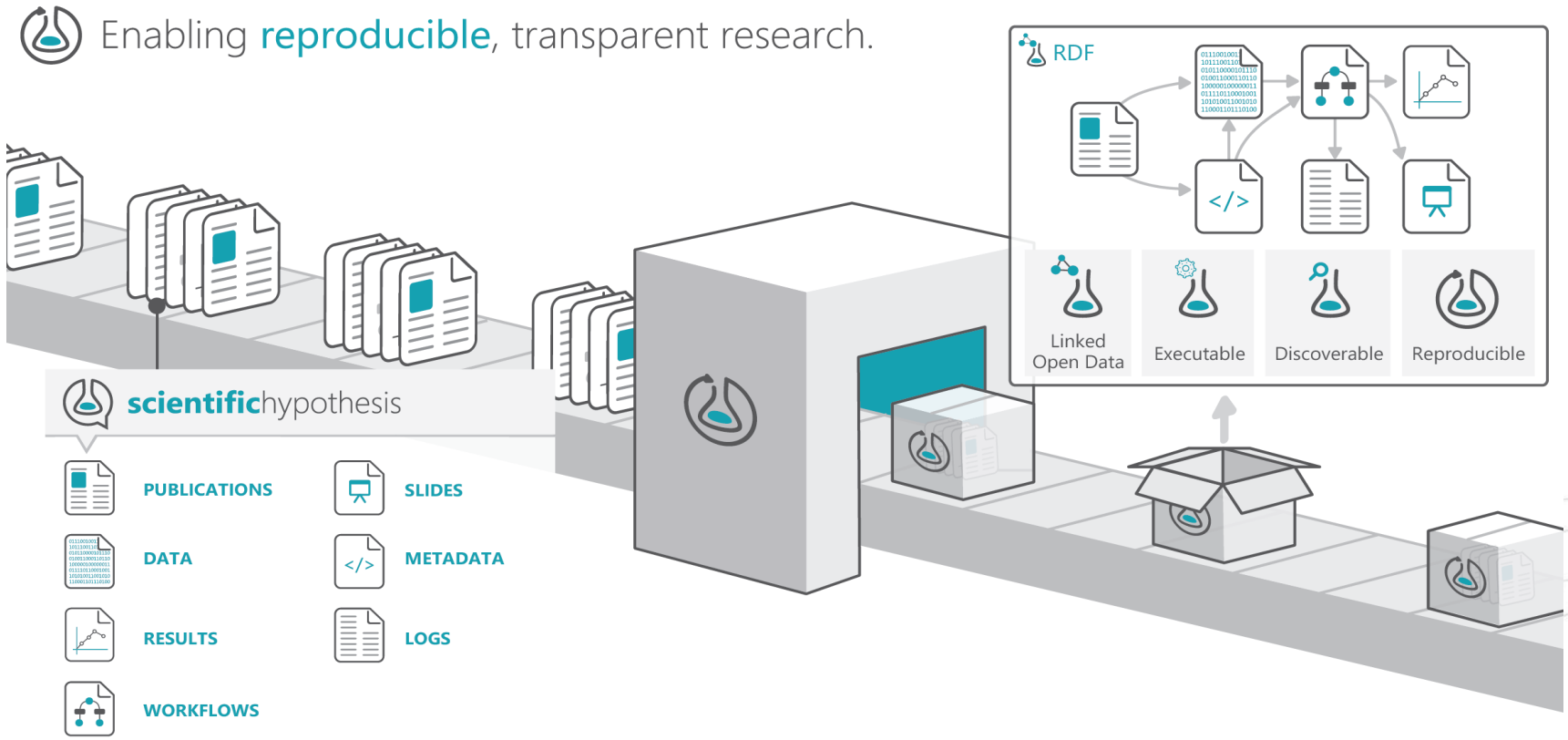
- In person, telecon
- Sharing by drive/cloud (Google Drive - institutional)

How to describe all these linked resources?

A. Fouilloux, F. Foglini, E. Trasatti. FAIR Research Objects for realising Open Science with RELIANCE EOSC Project

Research Objects (https://www.researchobject.org/ro-crate/)

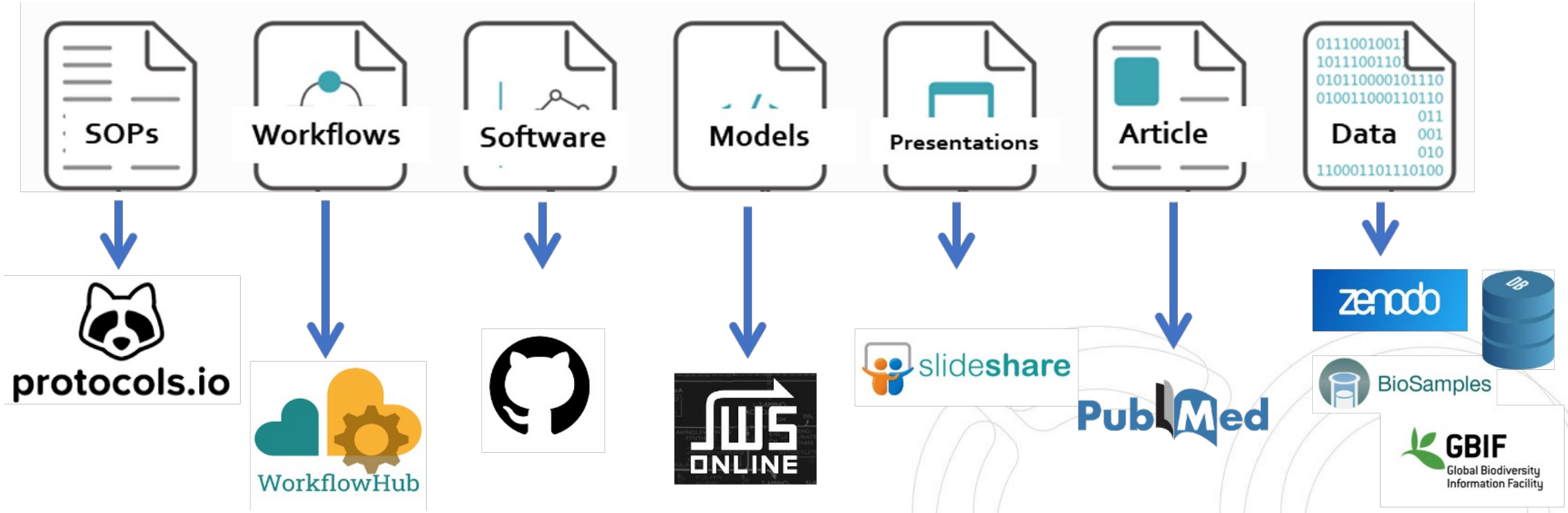
Goal: Account, describe and share everything about your research, including how those things are related



Research outcomes and related resources

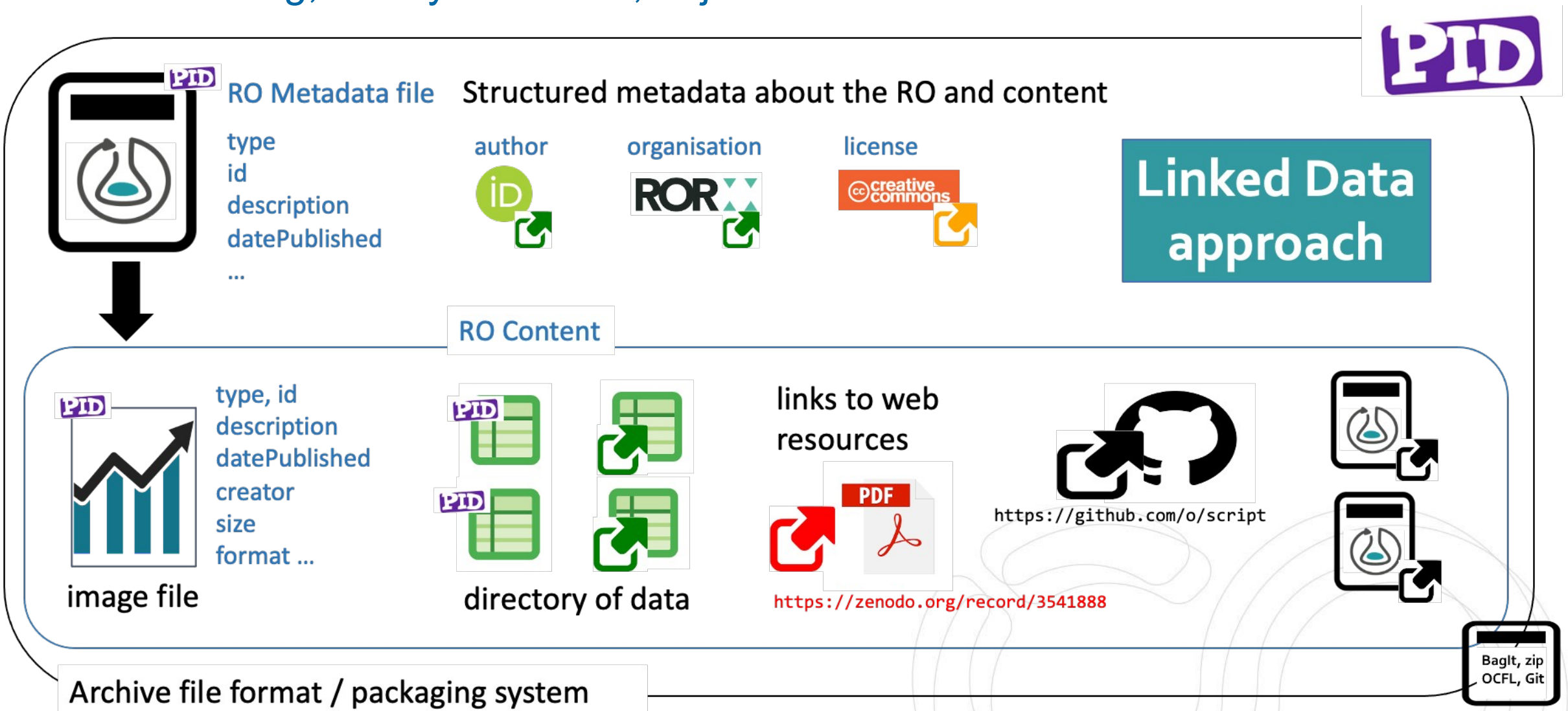
All are first class citizens and are required to make research FAIR

Each object has its own metadata and host repositories



[source RO-Crate: A framework for packaging research products into FAIR Research Objects]

Self-describing, chiefly metadata, objects

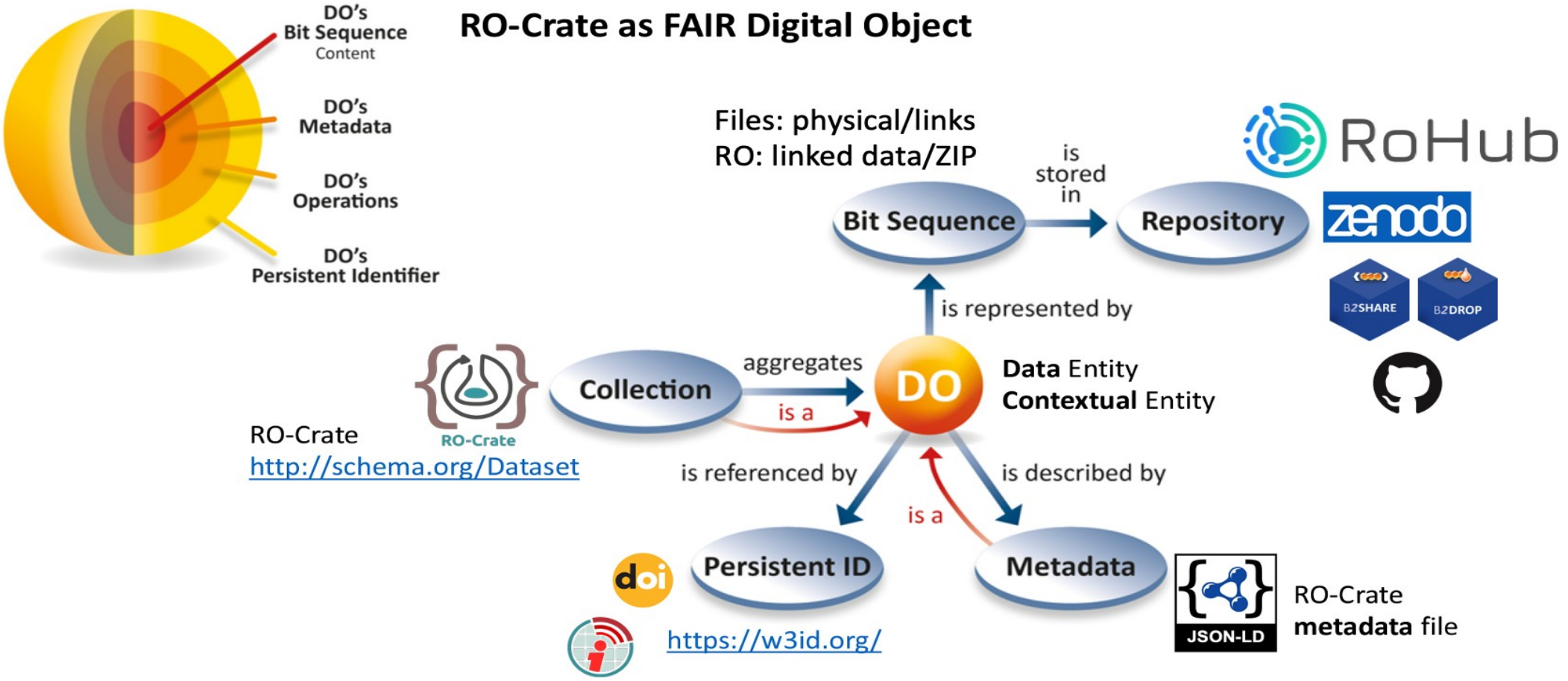


CC BY NC SA

[source RO-Crate: A framework for packaging research products into FAIR Research Objects]

RO-crate is FAIR Digital Object!

implementation of FAIR Digital Objects using RO Crate



CC BY SA

C. Goble, S. Soiland-Reyes. RO-Crate: A framework for packaging research products into FAIR Research Objects

Adapted from
https://doi.org/10.3390/publications8020021

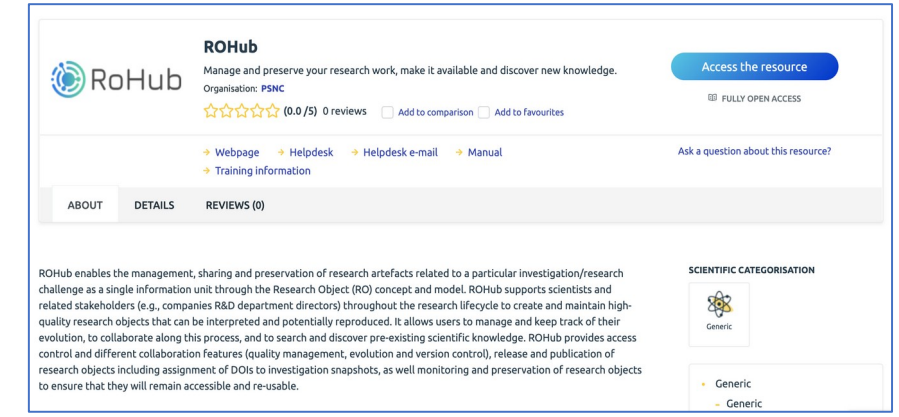
ROHub platform

2010-2013

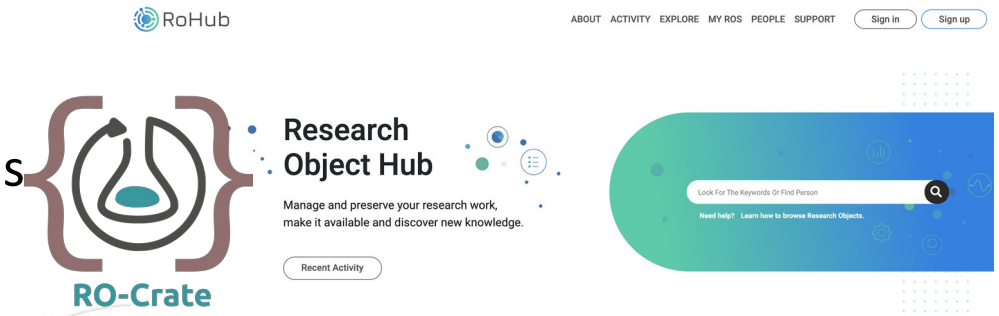
2014-2019

2020+

Onboarded and integrated in EOSC



- Holistic solution for research objects management
- Reference platform
  - Implements RO-Crate as the default RO exchange format
  - Support different stakeholders, with the primary focus on scientists, researchers, students and enthusiasts
  - Provides the backbone to a wealth of RO-centric applications and interfaces across different scientific communities
  - Integrates several added value services to deliver a rich VRE supporting the adoption of Open Science principles



3203 Research Objects

351 Users

15088 Resources

57744 Annotations

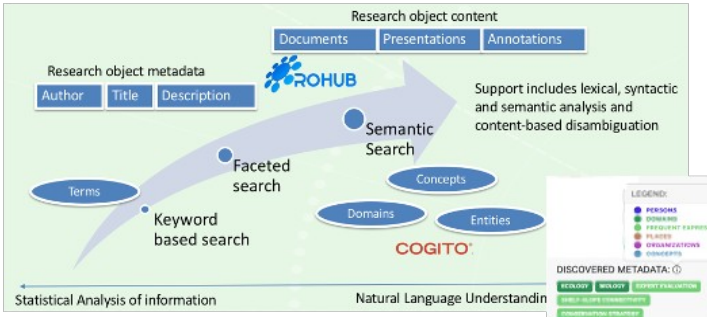
https://www.rohub.org/

CC BY NC SA

ROHub with RO added value services

Semantic enrichment

readability, discoverability, reuse



Recommendation

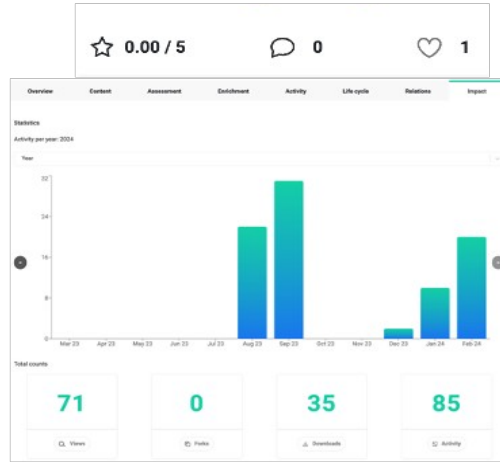
content-based, concentric spheres

Comprehension

Claim analysis, challenges & solutions
questions & answers, novelty

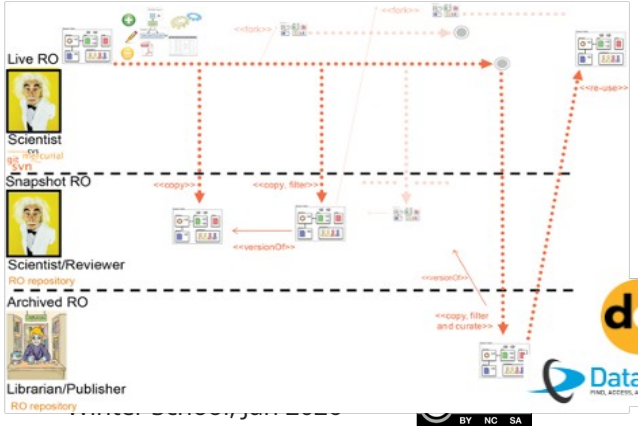
Impact

Sharing, rating



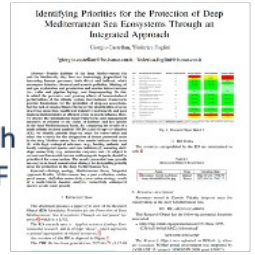
Research lifecycle & scholarly communication

collaboration, publication, citation, validation



RO-Crate

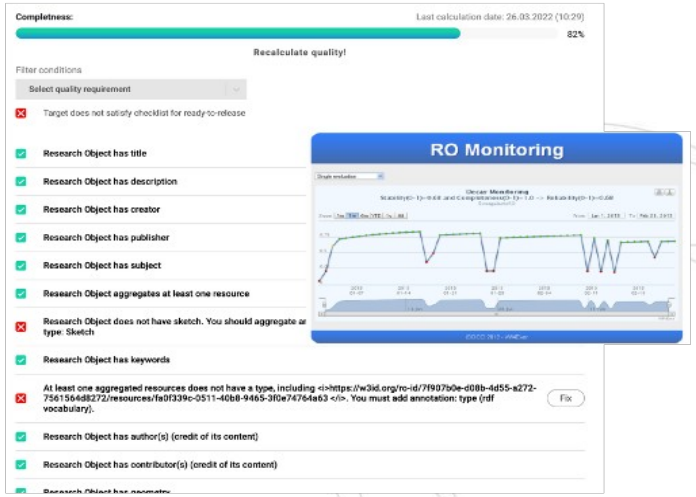
Publish RO-crate in EOSC



Publish as PDF

Completeness assessment

monitoring & preservation



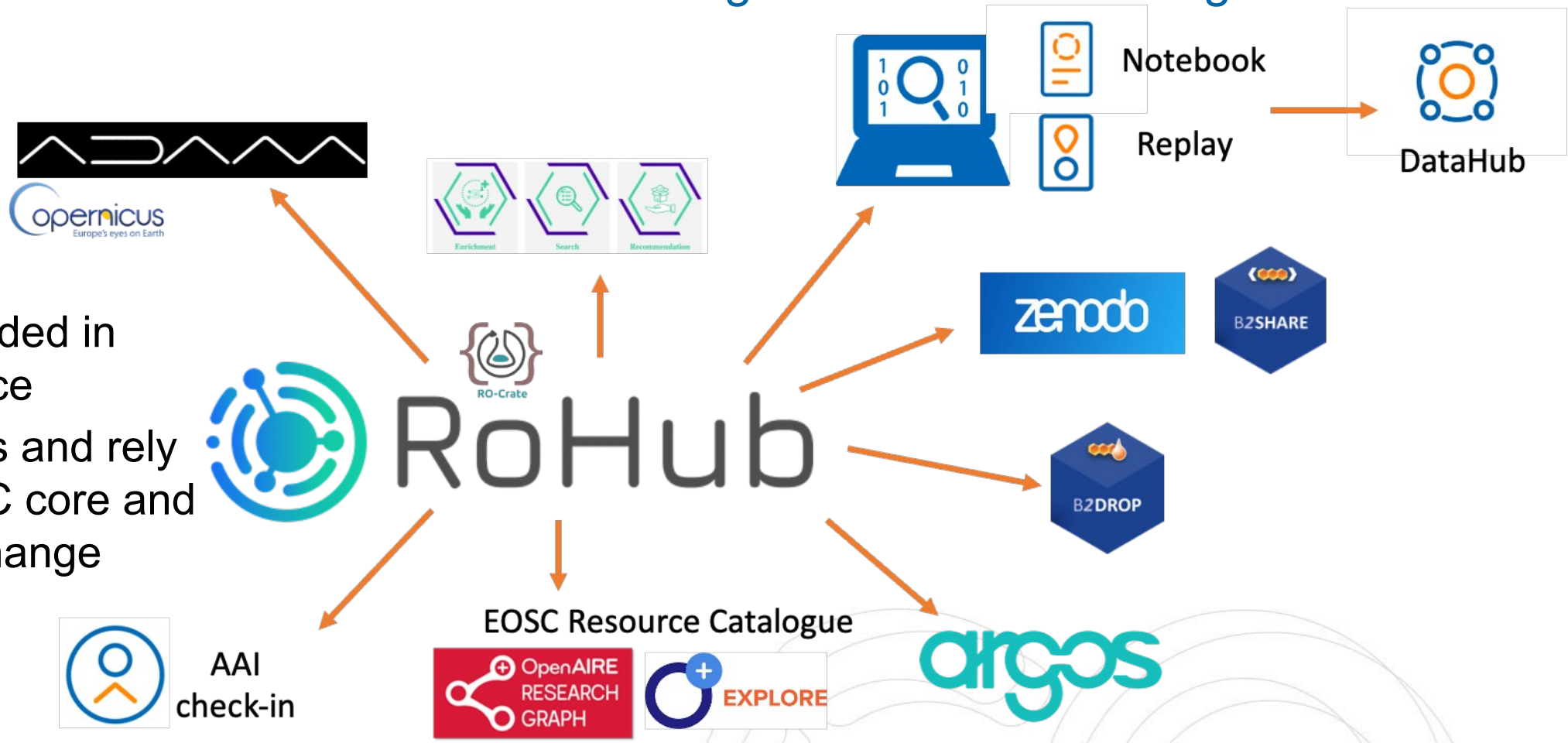
FAIR assessment

RO and components level

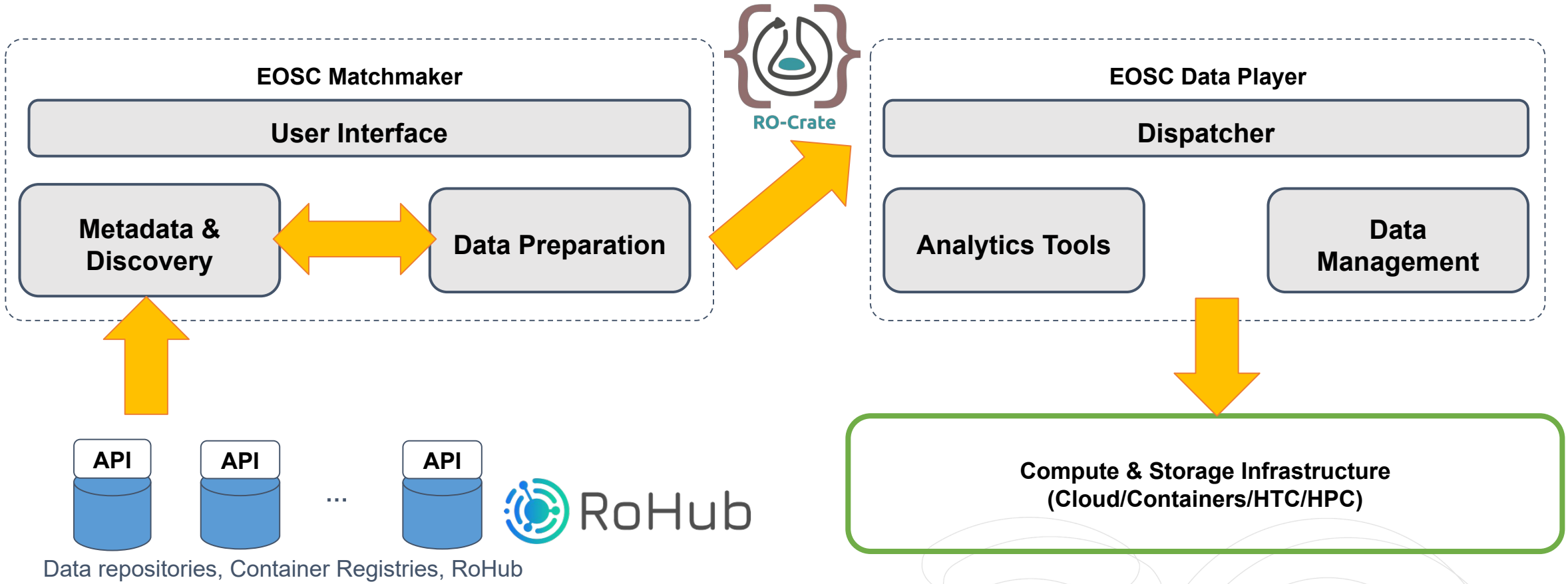
FAIROs

ROHub connections with EOSC Core and Exchange services for building VRE

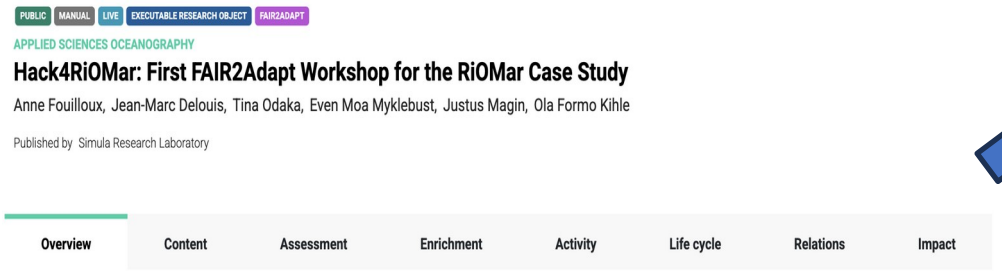
- ROHub is onboarded in EOSC marketplace
- ROHub integrates and rely on different EOSC core and other EOSC Exchange services

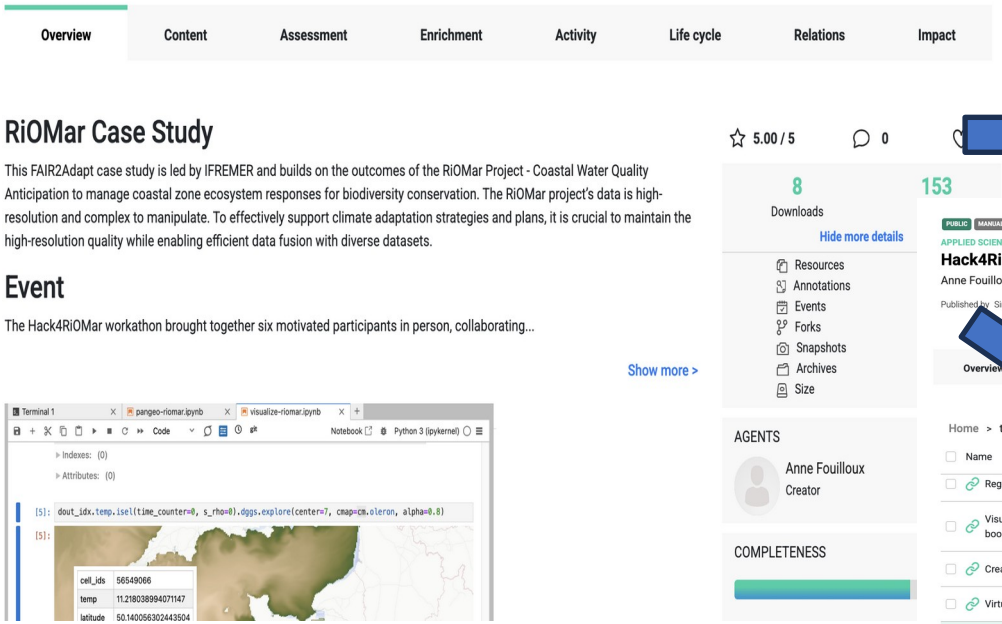


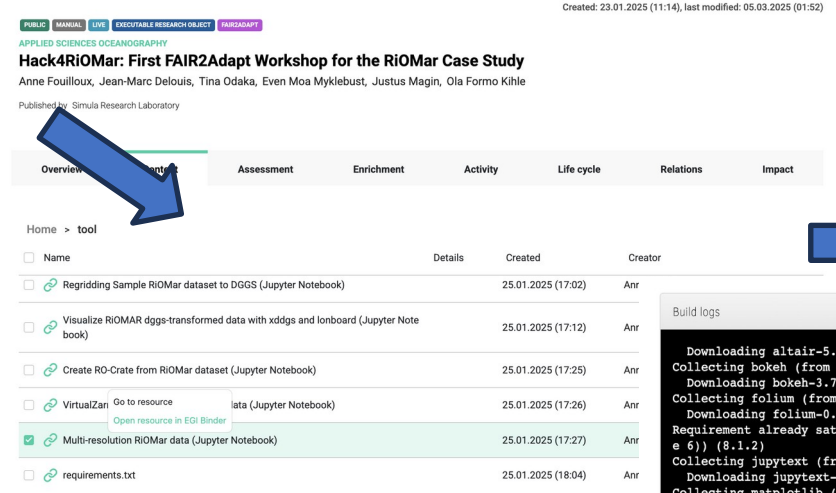
RO-Crate and ROHub adoption

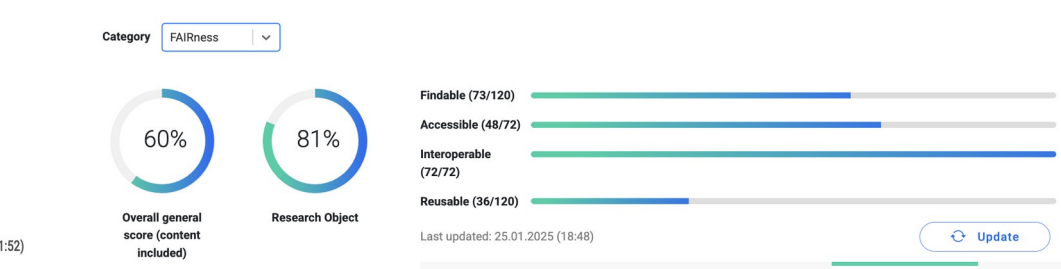


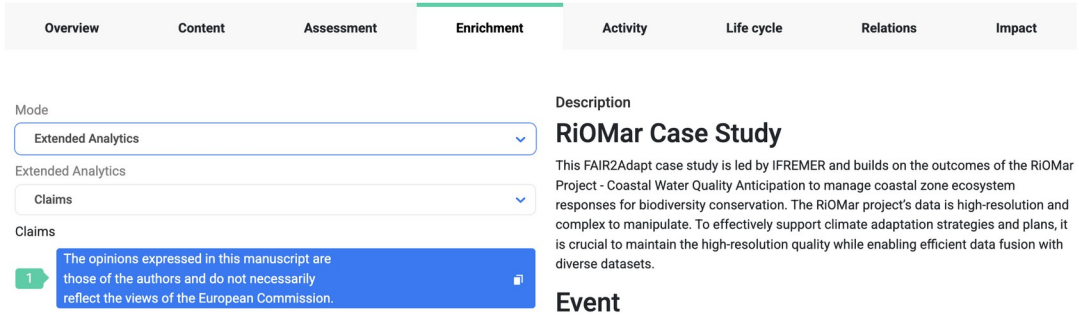
ROHub in practice

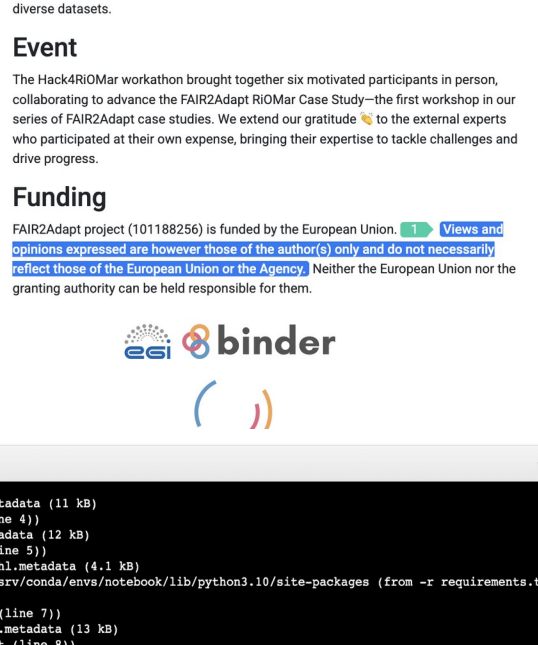


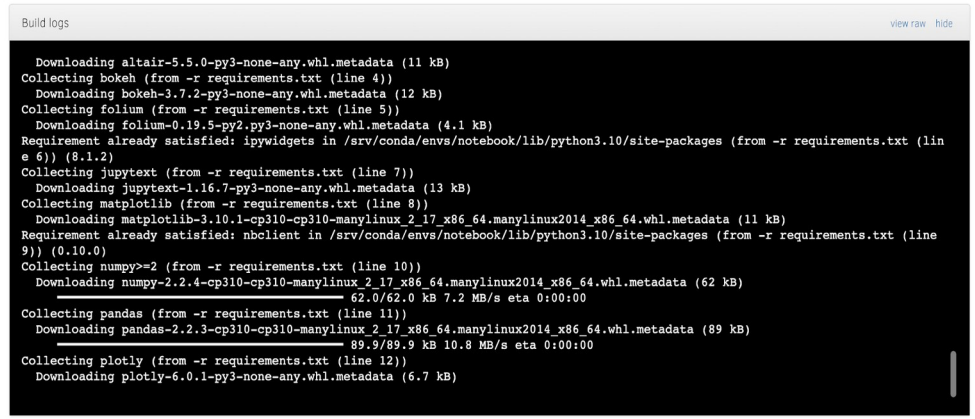












https://w3id.org/ro-id/a0007726-0f5b-4ccb-a160-f350d669805e

CC BY NC SA

https://fair2adapt-eosc.eu/

RO-Crates and ROHub platform

Fair Digital Objects (FDOs) enabling services in EOSC

Daniel Garijo (UPM), Raúl Palma (PSNC)

Sebastian Luna-Valero (EGI)

Funded by the European Union

The FAIR2Adapt Consortium

Grant No: 101188256

CC BY

National Czech Programme | eosc

Czech Academy of Sciences

J. Heyrovský Institute

Populating the EOSC Federation with Data.

The role and requirements of Electronic Lab Notebooks (ELNs) and workflows.

Marek Cebecauer

Heyrovsky Institute. Czech Academy of Sciences. Prague. Czech Republic

EOSC Winter School 2026. Nice-France. 27-29.1.2026

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

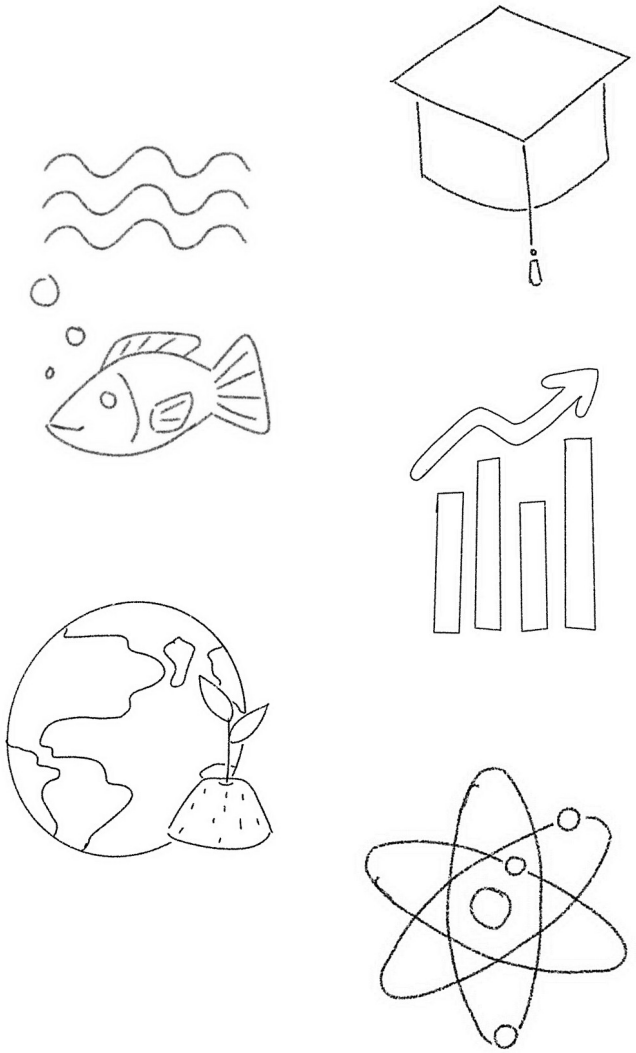
cesnet

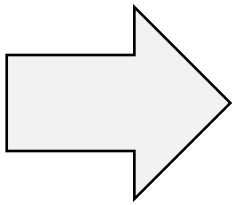
Registrační číslo projektu NRP

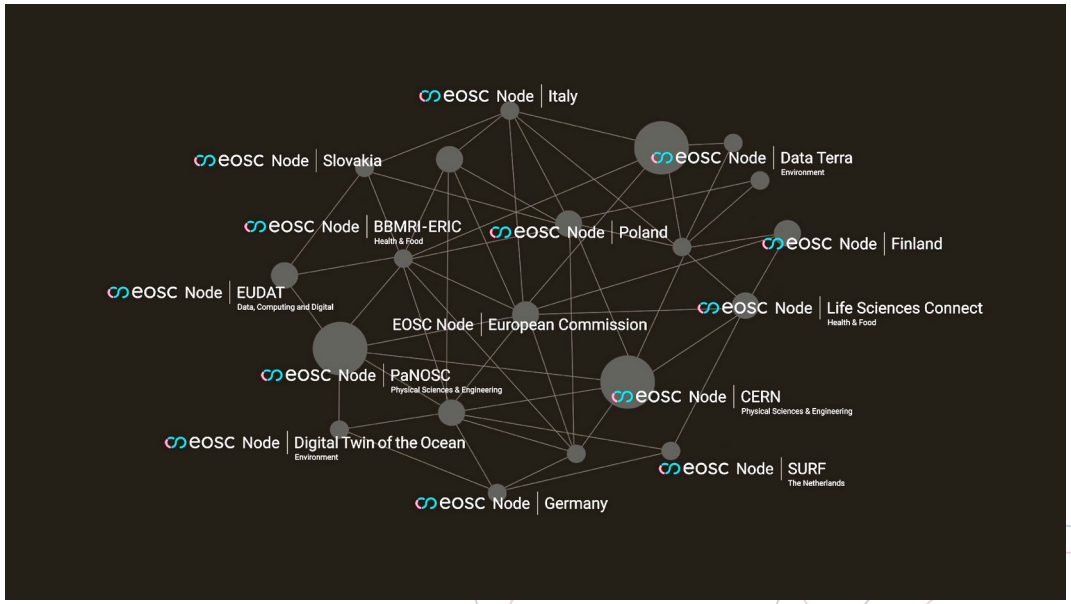
CZ.02.01.01/00/23_014/0008787

National Czech Programme

Populating the EOSC Federation with Data



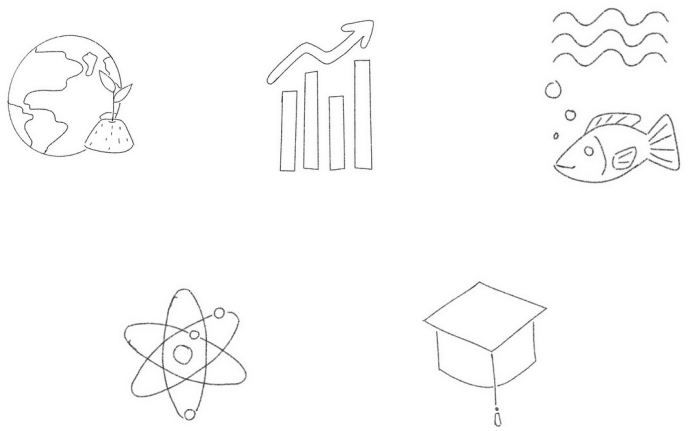


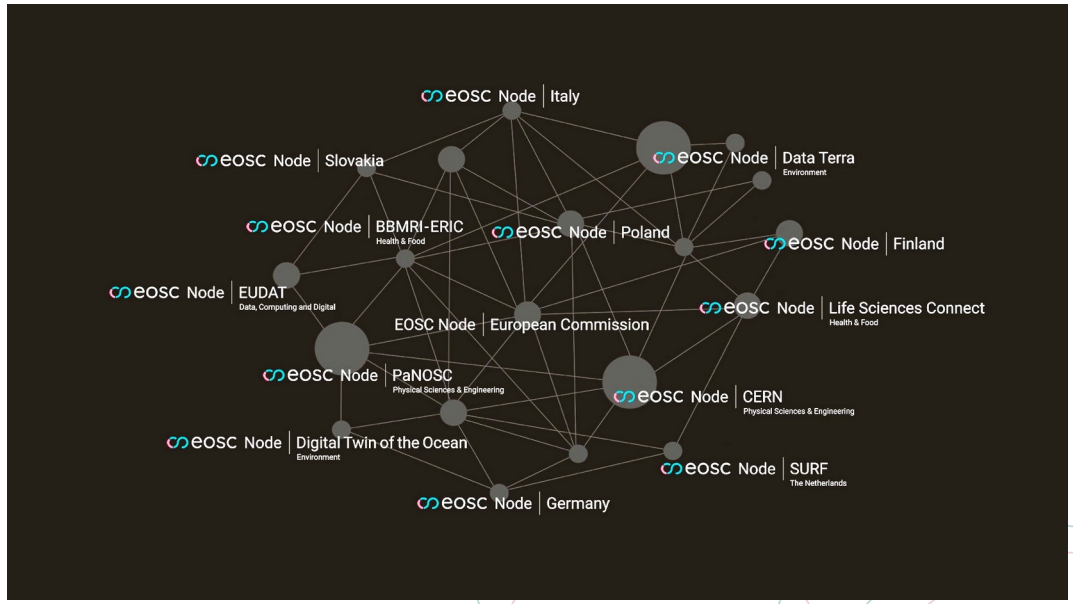


The EOSC Federation

National Czech Programme

Populating the EOSC Federation with Data

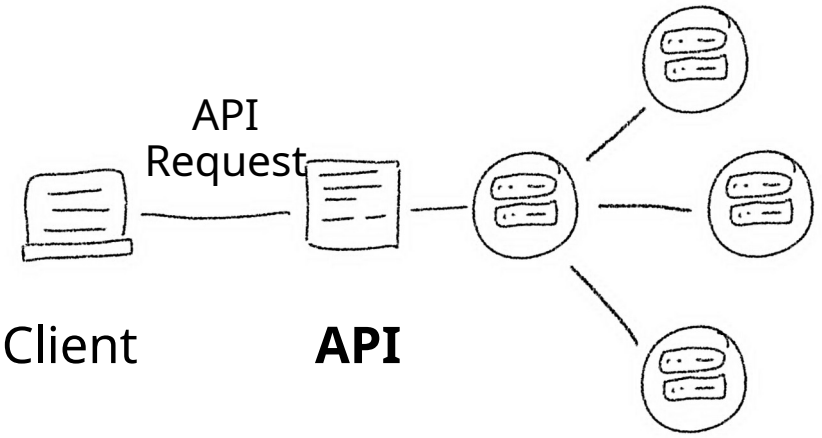


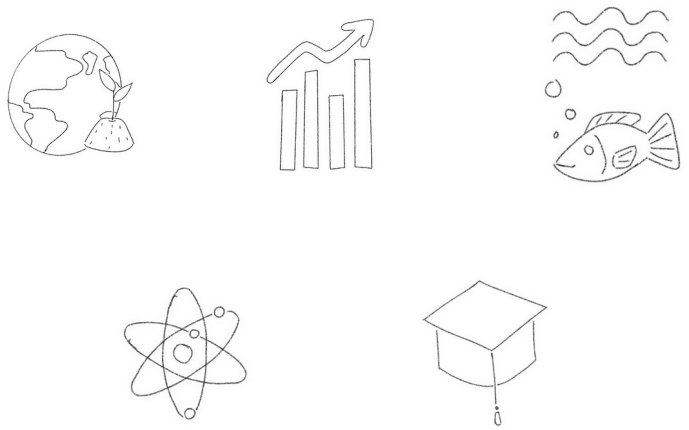


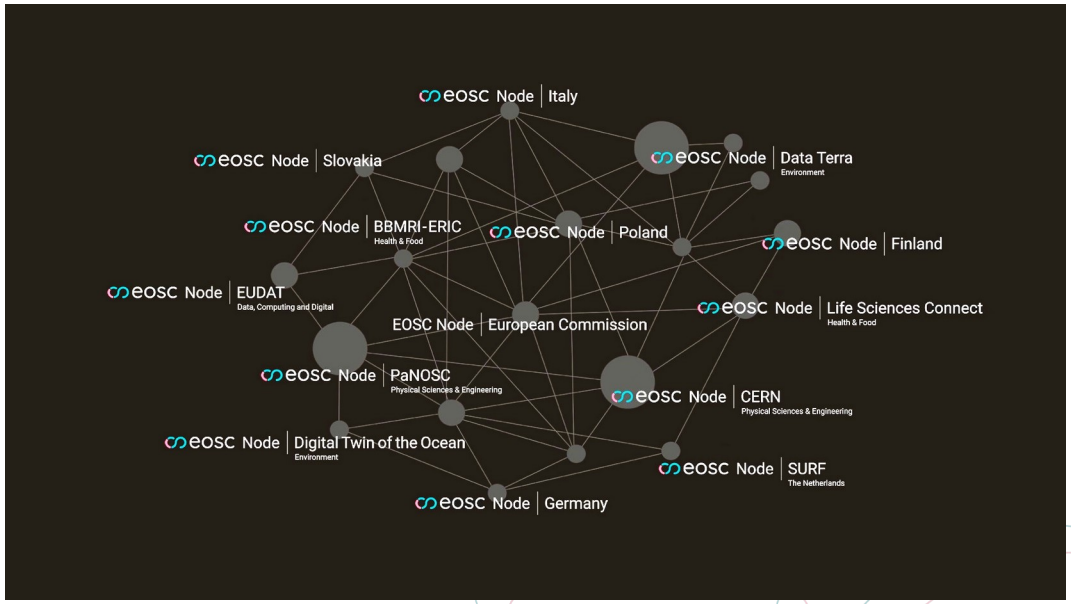
The EOSC Federation

National Czech Programme

Populating the EOSC Federation with Data



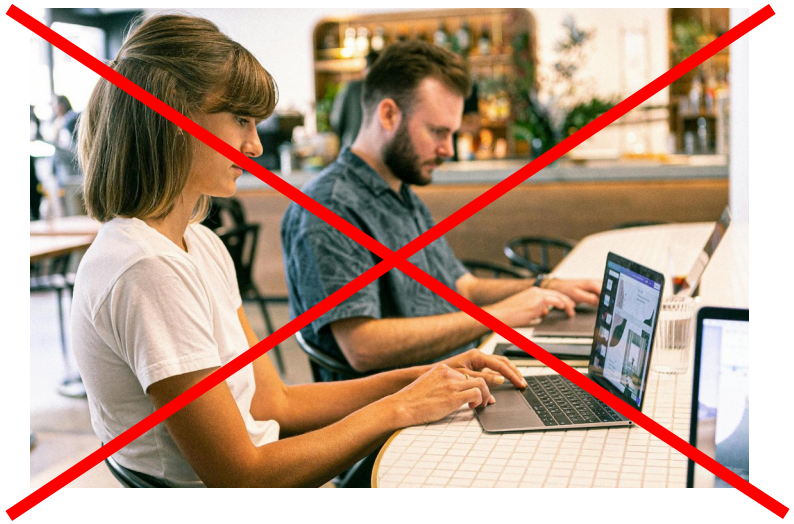


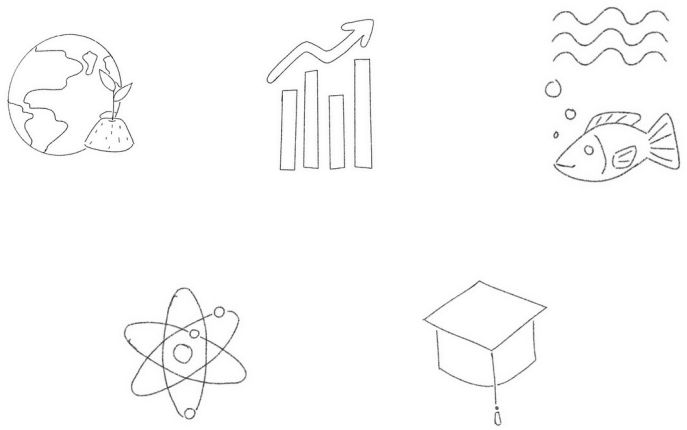


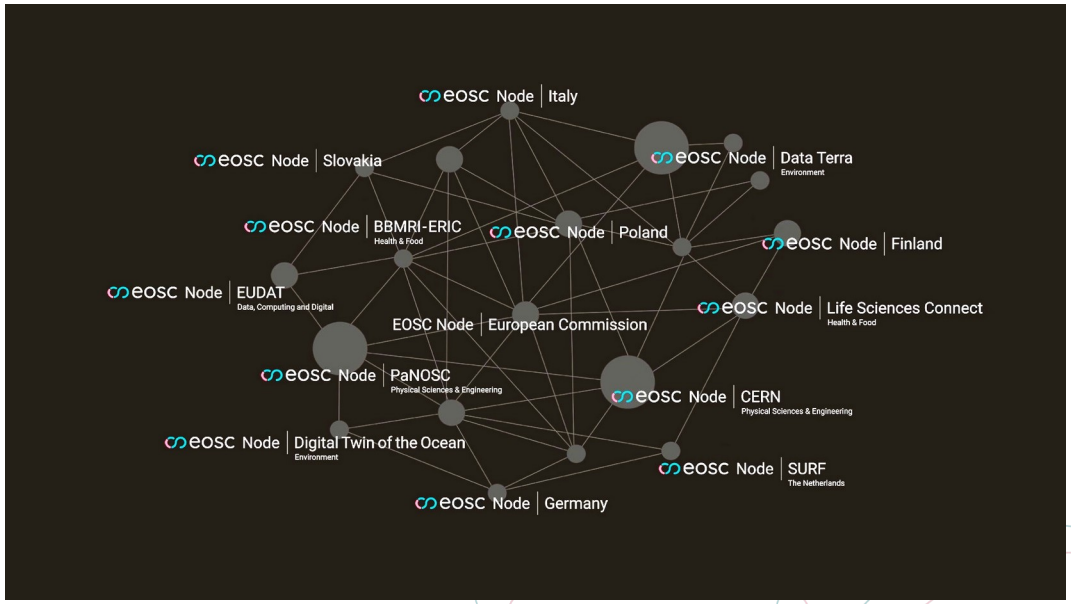
The EOSC Federation

National Czech Programme

Populating the EOSC Federation with Data



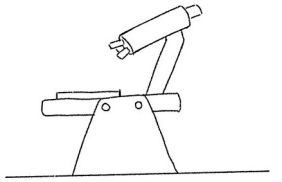




The EOSC Federation

National Czech Programme | eosc

Tools for day-to-day data management



- **Electronic Laboratory (Field) Notebooks** digitally capture and manage laboratory records, experimental and computational* data, and workflows in a structured, searchable and shareable form.

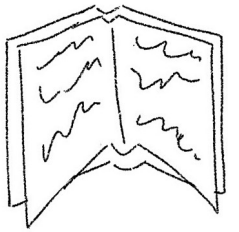
- **Workflow orchestrators** coordinate, schedule, and manage the execution of multi-step processes (experimental and computational) across different services and devices.

- There are **other systems** and services, which can replace ELNs and workflow orchestrators helping with data collection, processing and automation of data management in general.

** Not very common yet.*

National Czech Programme

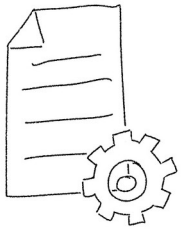
Electronic Laboratory (Field) Notebooks (ELNs)



- Facilitates day-to-day data management for humans (workflows for automation).
- Help with standardisation and FAIRification of data and other research outputs (templates, link to ontologies, …).
- Generate machine-readable or -actionable outputs.

National Czech Programme

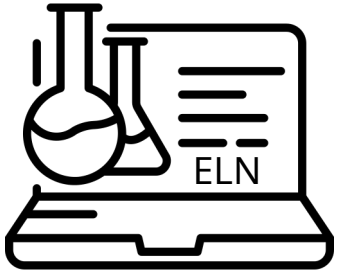
Data Formats and Structures

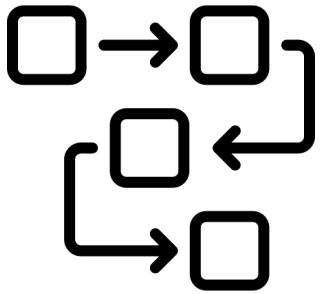


- ELN outputs: specific XML, JSON, YAML, … frequently PDF!!
- Some systems support export to RO-Crate (.ELN file).
- Workflow orchestrators: Diverse spectrum of output formats, including RO-Crates or other data containers.
- Other data systems: see Workflow orchestrators.

National Czech Programme

From ELN (Workflow) to the EOSC Federation





Workflo
w

RO-Crate/.ELN

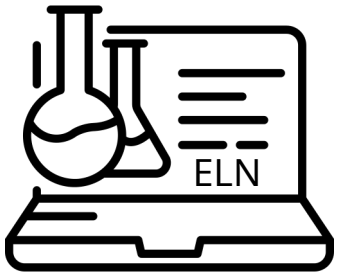
JSON schema* derived JSON-LD

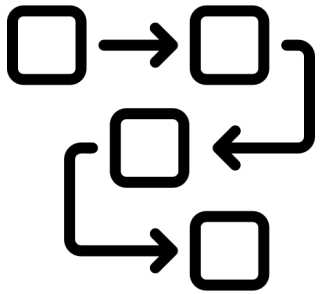
The EOSC Federation

*DCAT-AP based

National Czech Programme

From ELN (Workflow) to the EOSC Federation

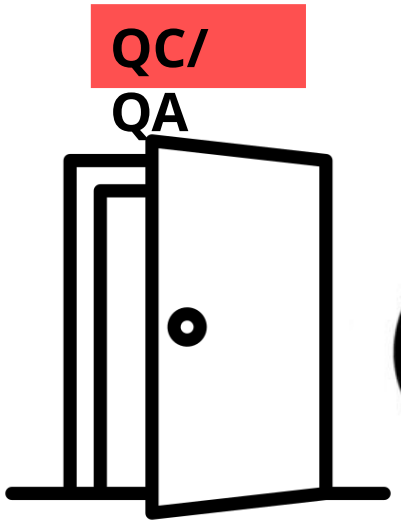




Workflo
w

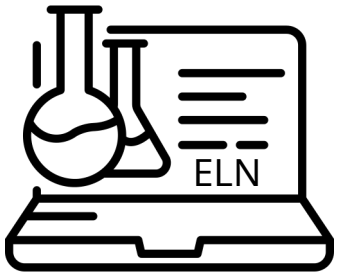
RO-Crate/.ELN

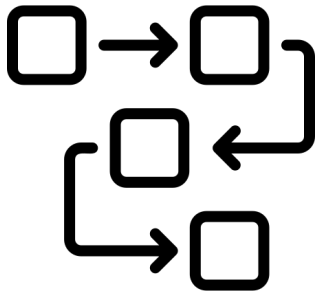
JSON schema derived JSON-LD



National Czech Programme

From ELN (Workflow) to the EOSC Federation



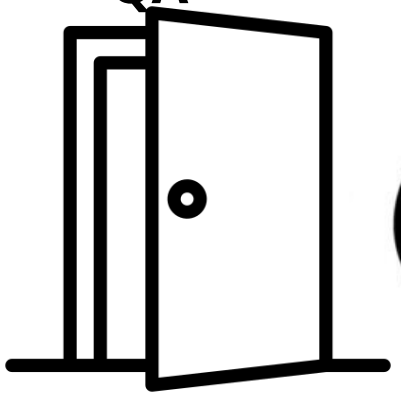


Workflow

RO-Crate/.ELN

JSON schema derived JSON-LD

QC/ QA



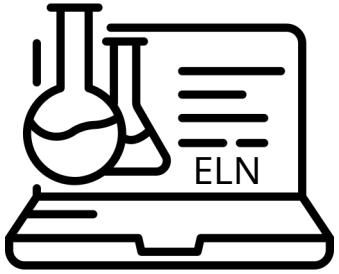
SHACL evaluation

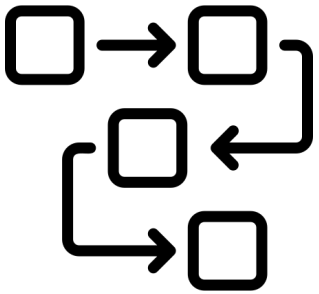
Curation, annotation

The EOSC Federation

National Czech Programme | eosc

From ELN (Workflow) to the EOSC Federation



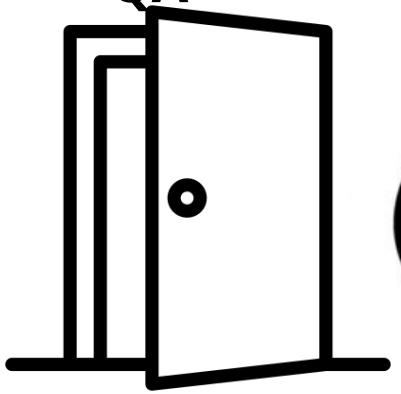


Workflow

RO-Crate/.ELN

JSON schema derived JSON-LD

QC/ QA



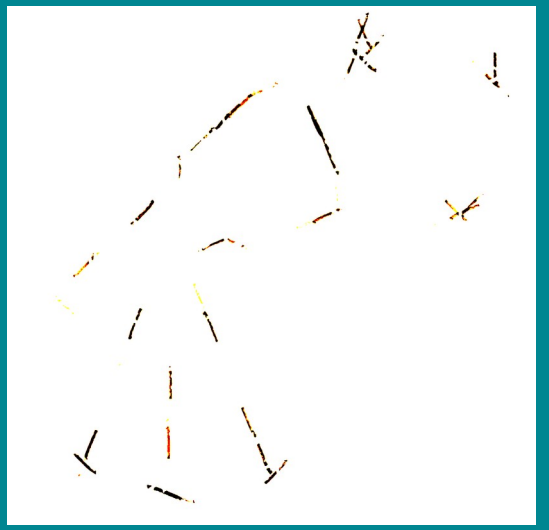
SHACL evaluation

Curation, annotation

The EOSC Federation

GraphDB, Triple Store, Document DB

National Czech Programme | eosc



thank.you.for.your.attenti

marek.cebecauer@jh-inst.cas.cz

Illustrations by **Alma Marie Cebecauer** and technical icons by Fr

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

J. Heyrovský Institute

Czech Academy of Sciences

Registrační číslo projektu NRP
CZ.02.01.01/00/23_014/0008787

eosc

Winter School 2026

27 - 29 January 2026
Nice, France

Data Quality Knowledge Graph of standards definitions, a concept for the EOSC Federation ?

Chris Schubert
TU Wien, Austria
OEA3
Chair of CODATA Task Group Data Quality Management

Wednesday, 28 Jan 2026

Data Quality an underestimated issue[1]

- Data Quality often remains an "underestimated pillar of research integrity" and “boring” burden within an increasingly "data-intensive, automated and cross-disciplinary" research data ecosystem
- What do we actually mean by Data Quality, so can we get on the same page, what we are talking about?
- More than good & bad data, do we speak the same language in our research domains?

Data Quality an underestimated issue[1]

Q 1 Are we aligned on Data Quality meaning?

for MetaData | Raw Data | Data Set | FAIR | Accuracy

Q 2 Data Quality as a burden or an accelerator ?

for trustworthy data
productivity

Q 3 How to express DQ assessment results ?

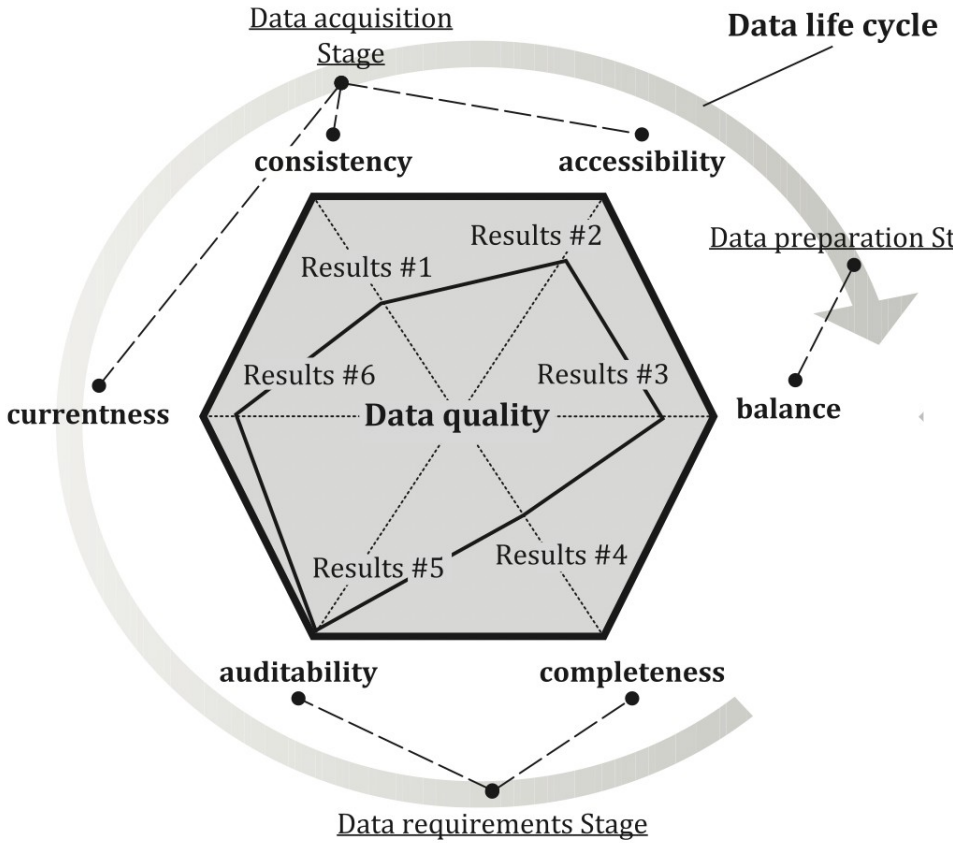
good | very good | passed | failed | compliant to ...

Q 4 Is there a lack of DQ rules ?

in which domain | best practices | open standards

Q 5 Scoring or assisting in DQ processes ?

Knowledge transfer | awareness | role of ML & AI



<A data quality model for analytics and ML>

ISO/IEC 5259-1:2024, Artificial intelligence — Data quality for analytics and machine learning (ML)

[1] The Data Quality Challenge – February 2020- Recommendations for Sustainable Research in the Digital Turn, Rat für Informations Infrastrukturen 2021, https://rfii.de/download/herausforderung-datenqualitaet-november-2019/

Data Quality Definitions

One example from ISO, the Online Browsing Platform (OBP) (https://www.iso.org/obp/) provides **eight** distinct definitions given for the term "Data Quality" or the closely related term "data - " and "data information quality"

> Data quality is defined as the **characteristics of data that relate to their ability to satisfy stated requirements.** And some minor variation uses the phrasing: **characteristic of data that bears on their ability to satisfy stated requirements.**
>
> Data quality is the **degree to which a set of inherent characteristics of data fulfils requirements** (ISO 8000)
>
> Data quality is the **degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions** (ISO/IEC 25012/25024)
>
> etc.

that is not a criticism, it is based by purposes, historical development and domain pillars

Data Quality Definitions

or from CODATA:

„Reliability and application efficiency of data. Perception or assessment of a dataset's fitness to serve its purpose in a given context. Aspects of data quality include: Accuracy, Completeness, Update status, Relevance, Consistency across data sources, Reliability, Appropriate presentation, Accessibility. Data quality is affected by the way data are entered, stored and managed.“

IRI: https://terms.codata.org/rdmt/data-quality

does EOSC has a common meaning on DQ ?

not needed, just being aware that the research disciplines control the terms

Data Quality Dimensions Definitions

- Data Quality dimensions (can we define *a priori* Data Quality requirements across disciplines?)
  - Three C's: Completeness, Correctness and Context (various authors, since 1980s)
  - Accuracy, Relevancy, Representation, Accessibility (Wang & Strong, 1996)
  - Completeness, Uniqueness, Timeliness, Validity, Accuracy, Consistency (Data Management Association UK, 2013)
  - Five C's of Sherman (2015): Clean, Consistent, Conformed, Current and Comprehensive
  - Conformance, Completeness and Plausibility (Kahn *et al.*, 2016)
  - etc.

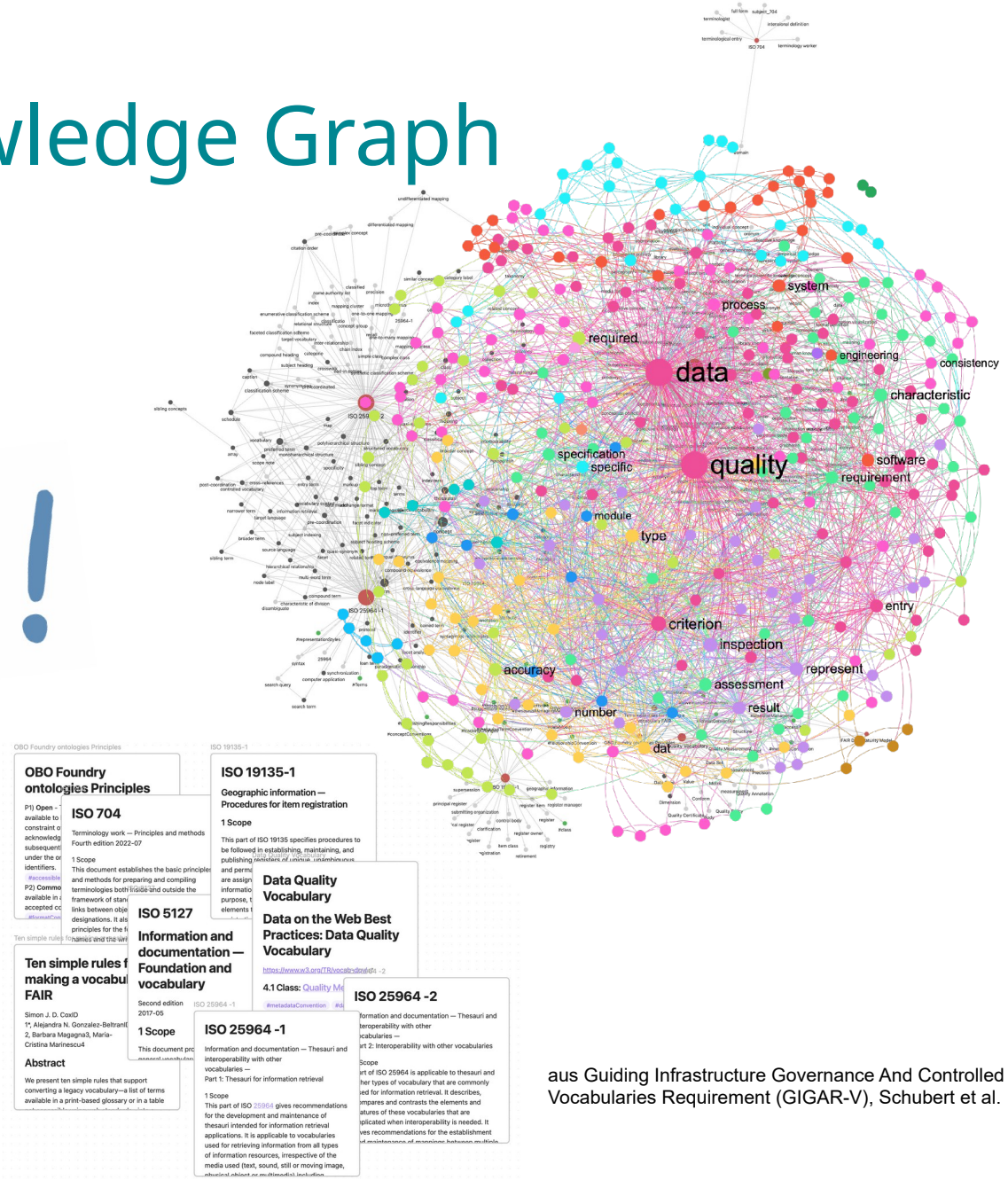
Data Quality fit for use?

Data Quality Knowledge Graph

Is a knowledge graph an appropriate representation to cover definitions relating to data quality, its dimensions, indicators and measures in the associated domain?

Thank You !

Chris Schubert
TU Wien, University of Technology Vienna,
chris.schubert@tuwien.ac.at



aus Guiding Infrastructure Governance And Controlled Vocabularies Requirement (GIGAR-V), Schubert et al.

Conceptual tool towards long-term data retention

Hervé L’Hours and Jacques Flores

Introduction

The **Retention, Reappraisal Iteration Logic tool** (RRAIL) is a document, currently under review, put forth by the Long-Term Data Retention Task force whose aims incorporate the following:

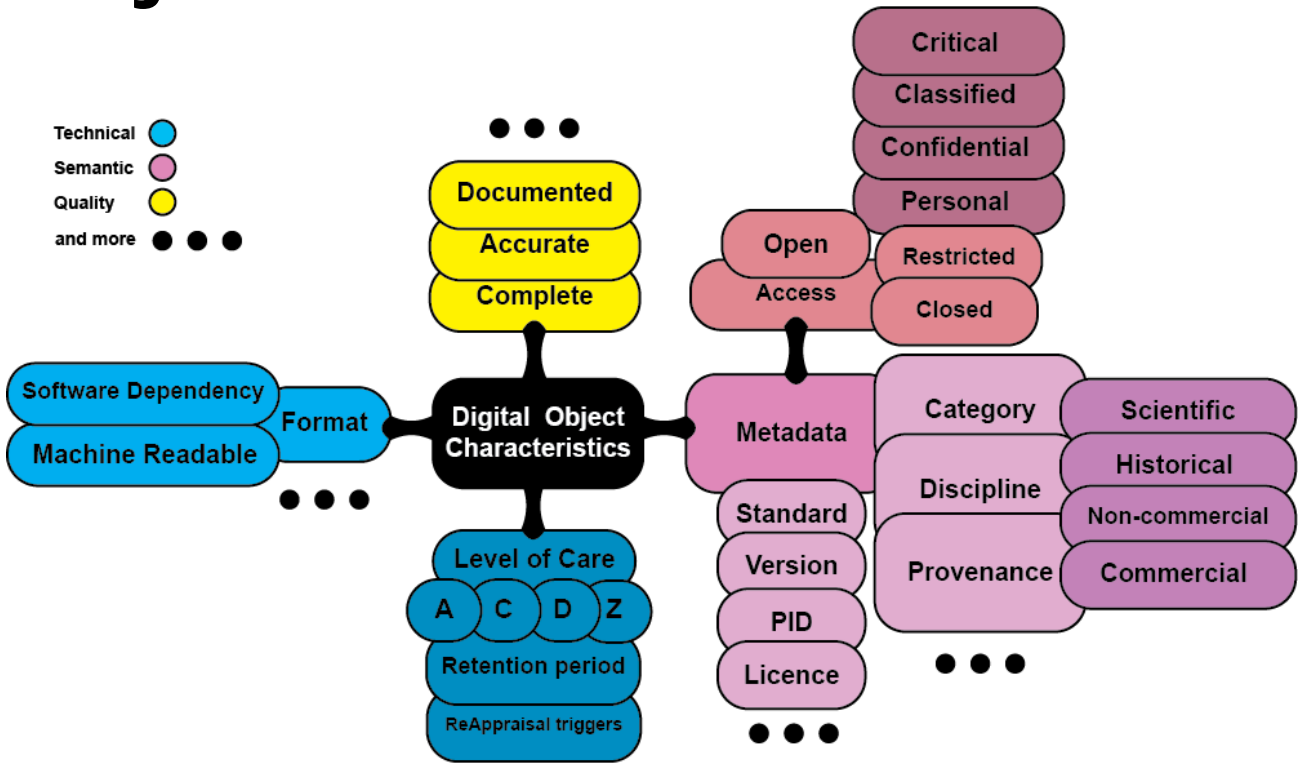
- Define common terminology in the field of long-term retention
- Depict the stages of a digital object as it moves through:
  - Appraisal and Reappraisal
  - Retention and Preservation
- Map /assimilate recommendations from both EDEN and FIDELIS on core preservation principles and repositories activities and functions (respectively)

Key Components

ReRAIL is dependent on the following 4 building blocks:

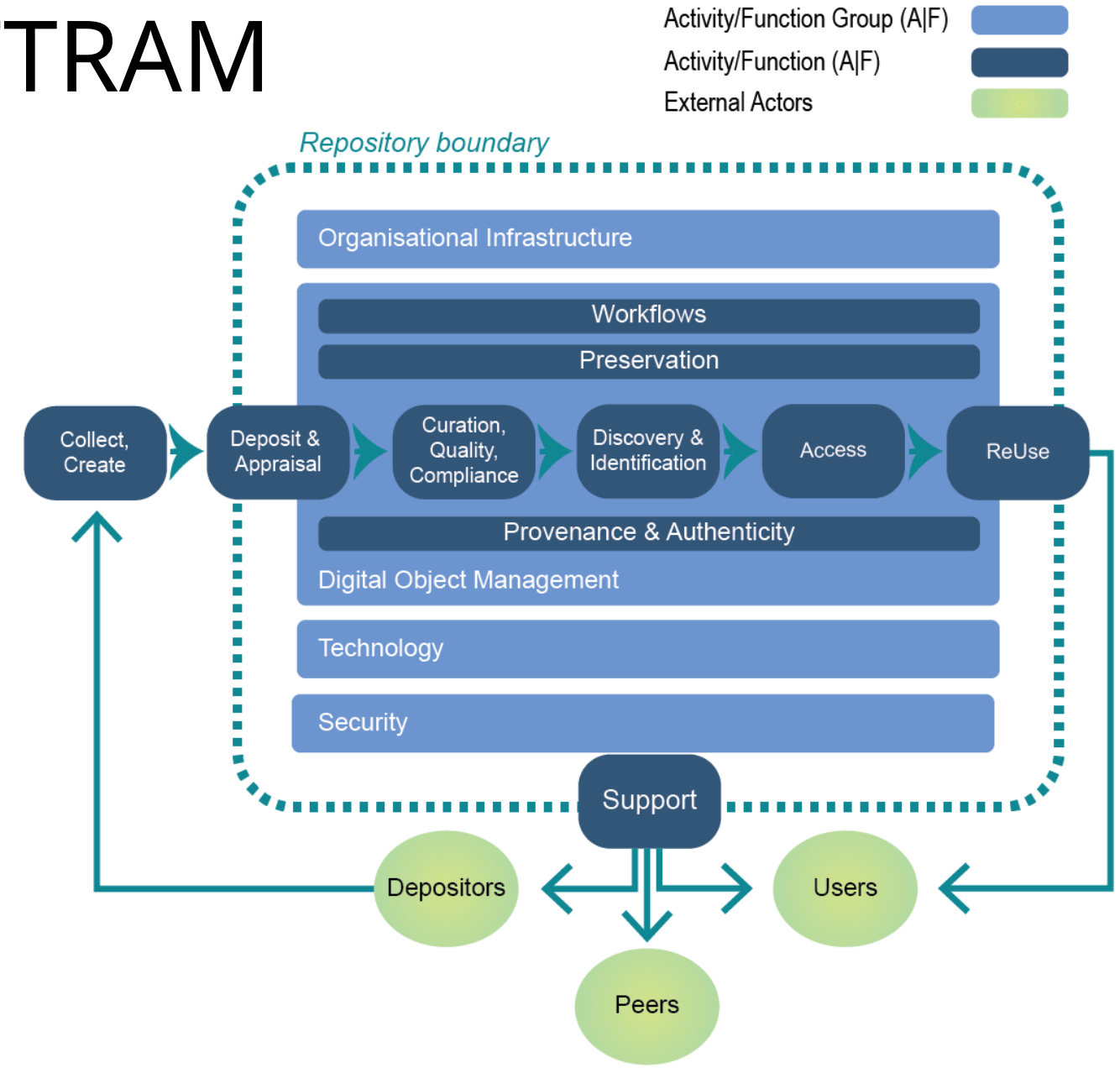
- Digital Object Characteristics
- Transparent Trustworthy Repository Attribute Matrix (TTRAM : FIDELIS)
- Level of Retention, Curation and Preservation (LoRCAP)
- Core Preservation Processes (CPP: EDEN)

Digital Object characteristics



Example Object Characteristics. This image shows some of the possible characteristics a digital object may have when submitted into a repository. The actual characteristics may differ depending on the type/scope/field of the repository. The classification of object characteristics into Semantic, Technical and Quality still requires further development.

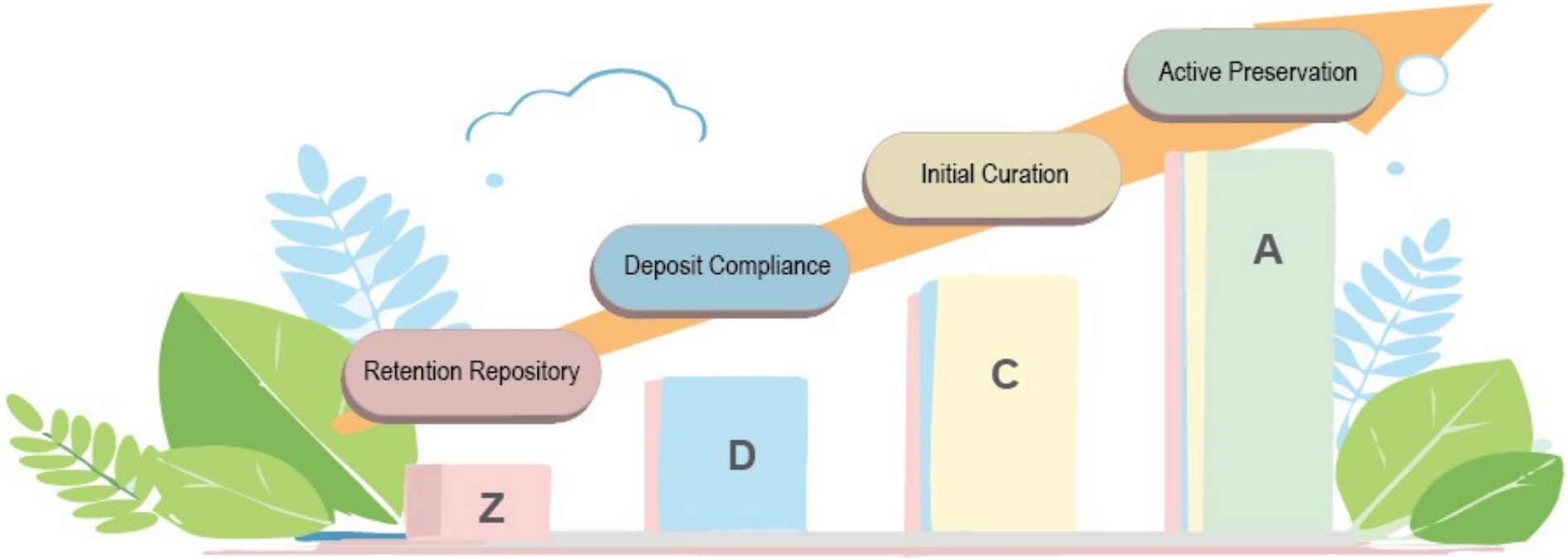
TTRAM



TTRAM Digital object Management Activities & Functions (A|F) workflow.

[FIDELIS TTRAMatrix v01.00 Introduction and Overview]

LoRCAP

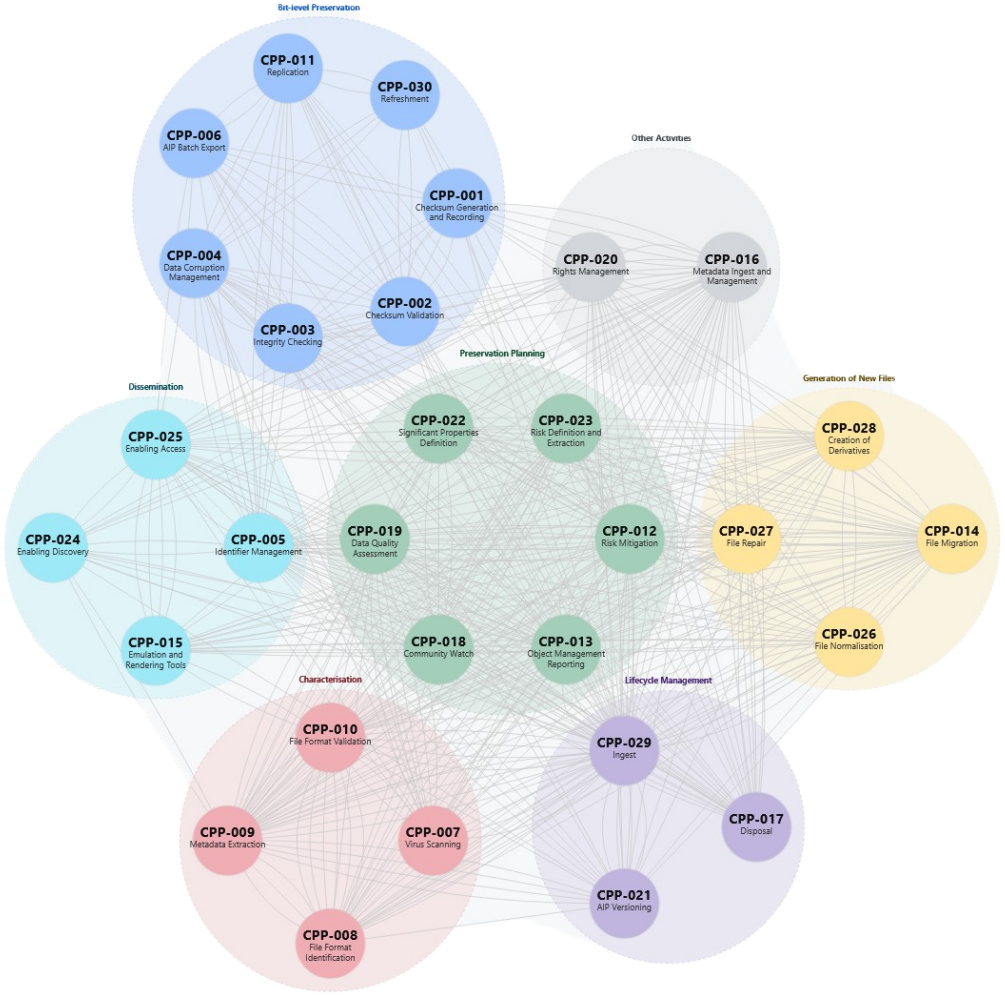


LoRCAP Levels of Retention, Curation And Preservation. The levels build upon each other demonstrating that initial curation will also incur deposit compliance and retention but also showing that an active preservation repository may or may not also offer initial curation. Levels of Retention, Curation And Preservation - Zenodo

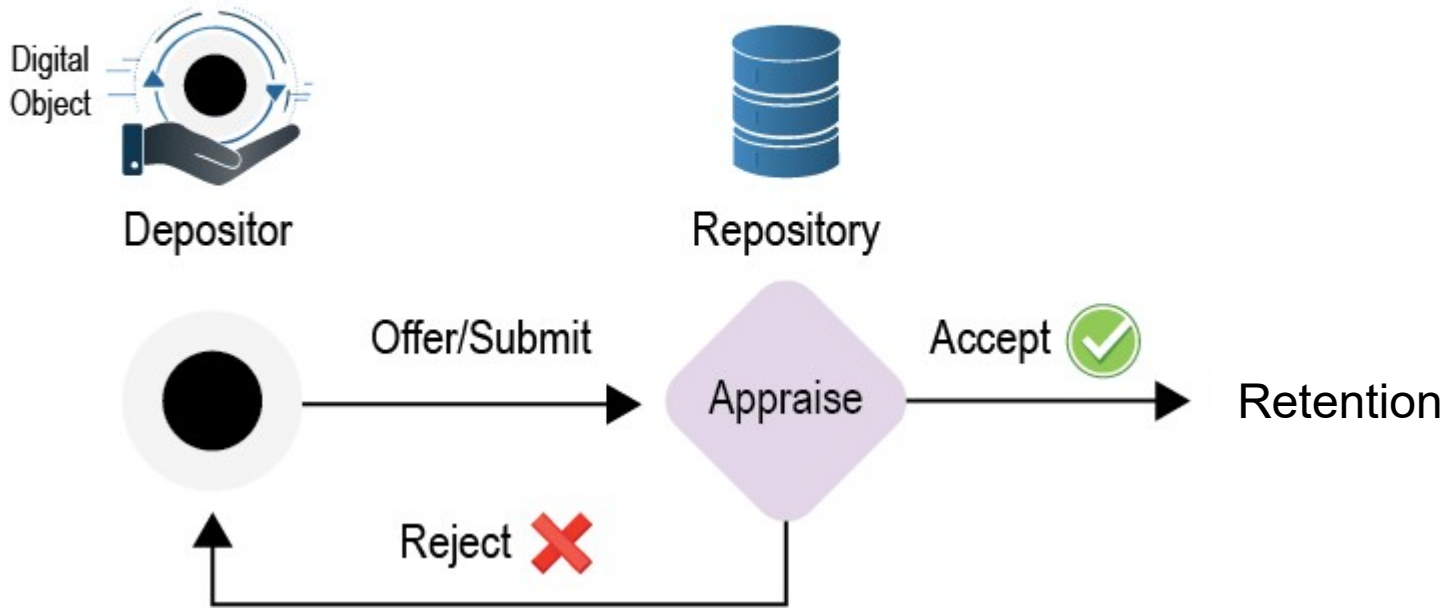
Core Preservation Processes

Core Preservation Processes Visualised. Each node depicts one of the 30 CPP's. The links between CPP's represent procedural relationships (e.g. "trigger"), dependencies (e.g. "Required By") and logical relationships (e.g. "Not to be confused with"). Link to CPP visualizer: https://cpp.fd-dev.csc.fi/.

M1.1 Report on Identification of Core Preservation Processes

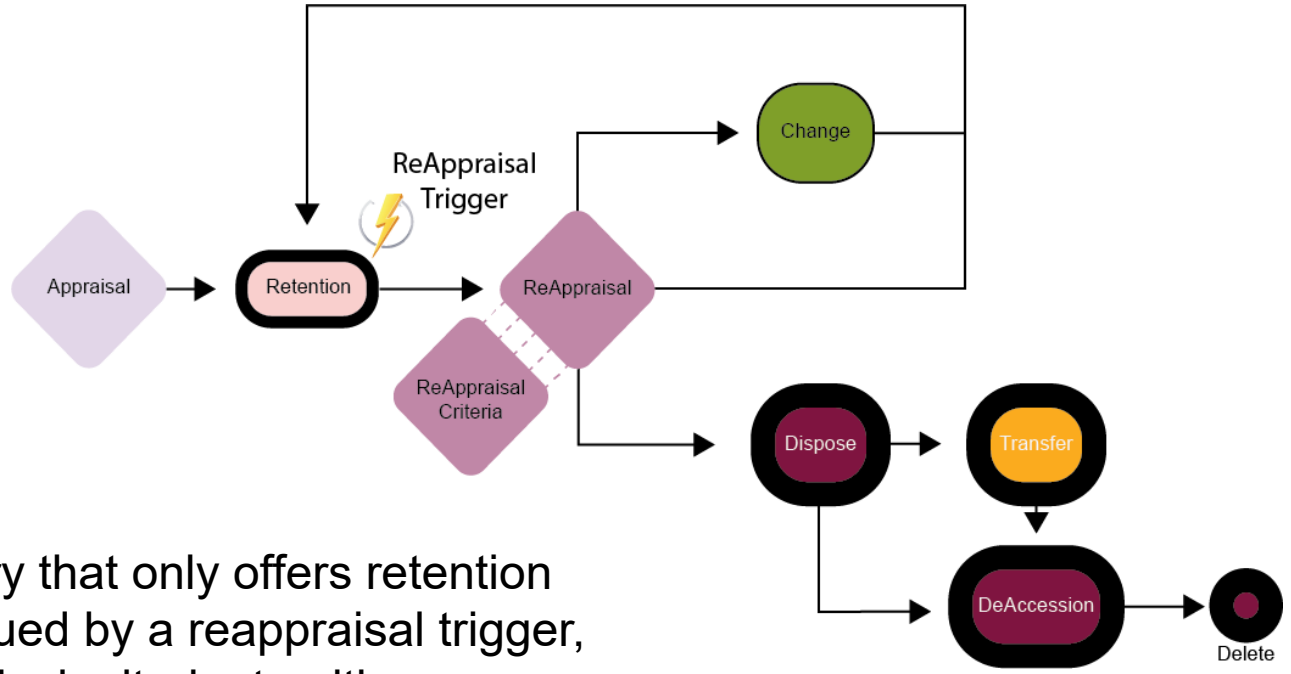


Submission and Appraisal



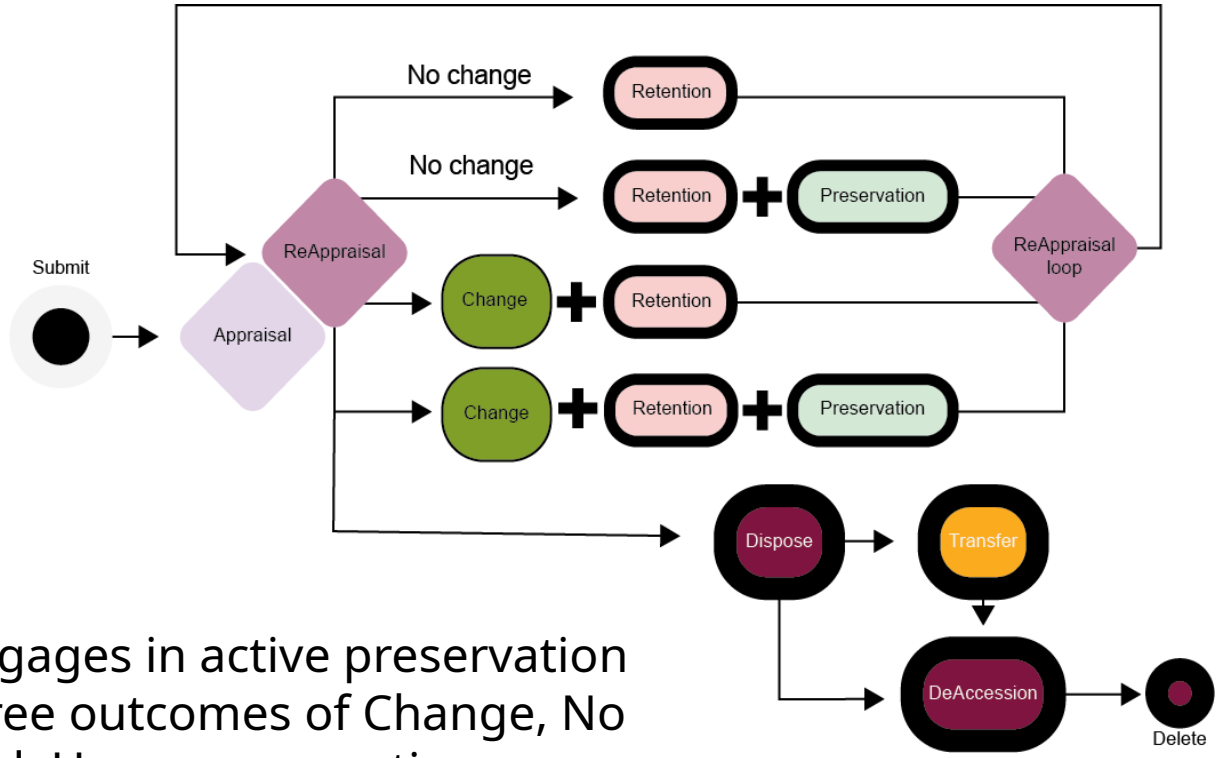
A repository may Accept or Reject the digital object being offered based on defined deposit compliance criteria. This assessment of digital objects' data and metadata may be machine-actionable or human-mediated. The option to 'resubmit' a digital object for deposit is implicit in 'reject' but excluded here to simplify the model.

Retention & ReAppraisal



After appraisal, a repository that only offers retention will at a moment in time, cued by a reappraisal trigger, choose, based on reappraisal criteria, to either continue to retain the DO with changes, continue to retain it without changes or dispose of the DO

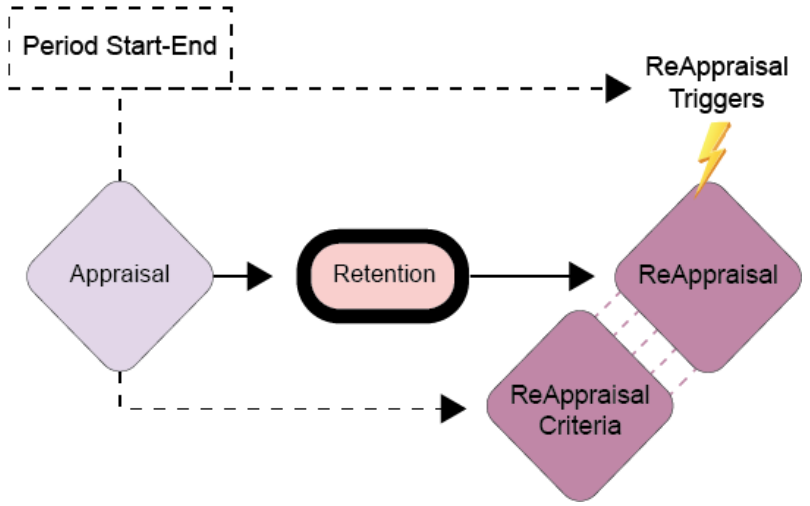
Preservation and Reappraisal



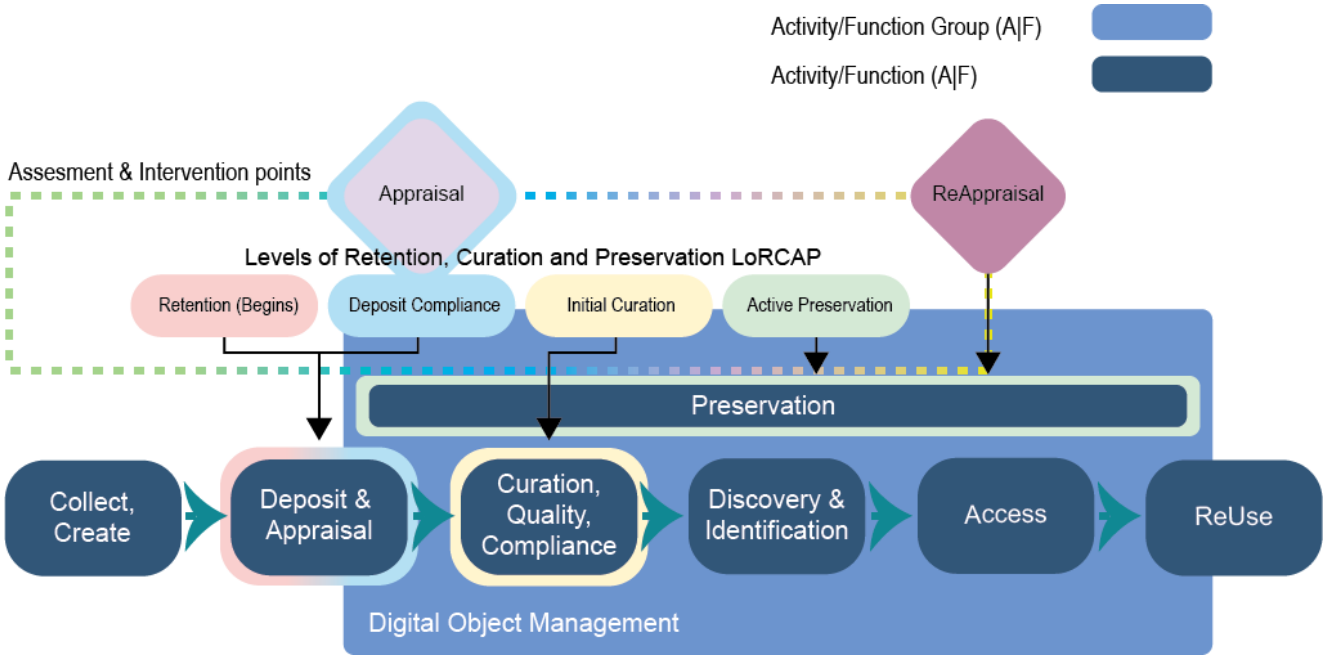
A repository that engages in active preservation will maintain the three outcomes of Change, No Change and Disposal. However, an active preservation repository enables a DO to change its status from active preservation into mere-retention and vice versa.

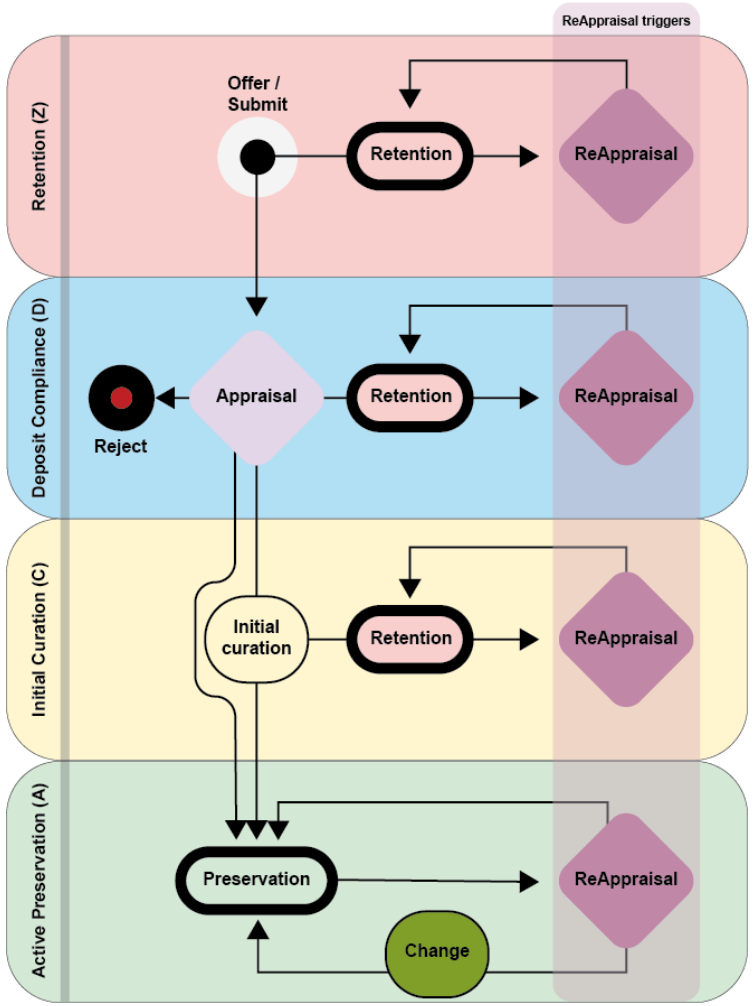
Reappraisal triggers and criteria

ReAppraisal occurs either at the end of an agreed retention period, or as a result of a trigger (e.g. closure of the repository) before that time is reached. There is a limited number of ReAppraisal outcomes. It is conceptually possible that a digital object will be automatically deleted at the reappraisal point with no further assessment (e.g. for legal reasons based on a records retention schedule).



Mapping it all together





eosc | Health Data Task Force

European Health Data Space

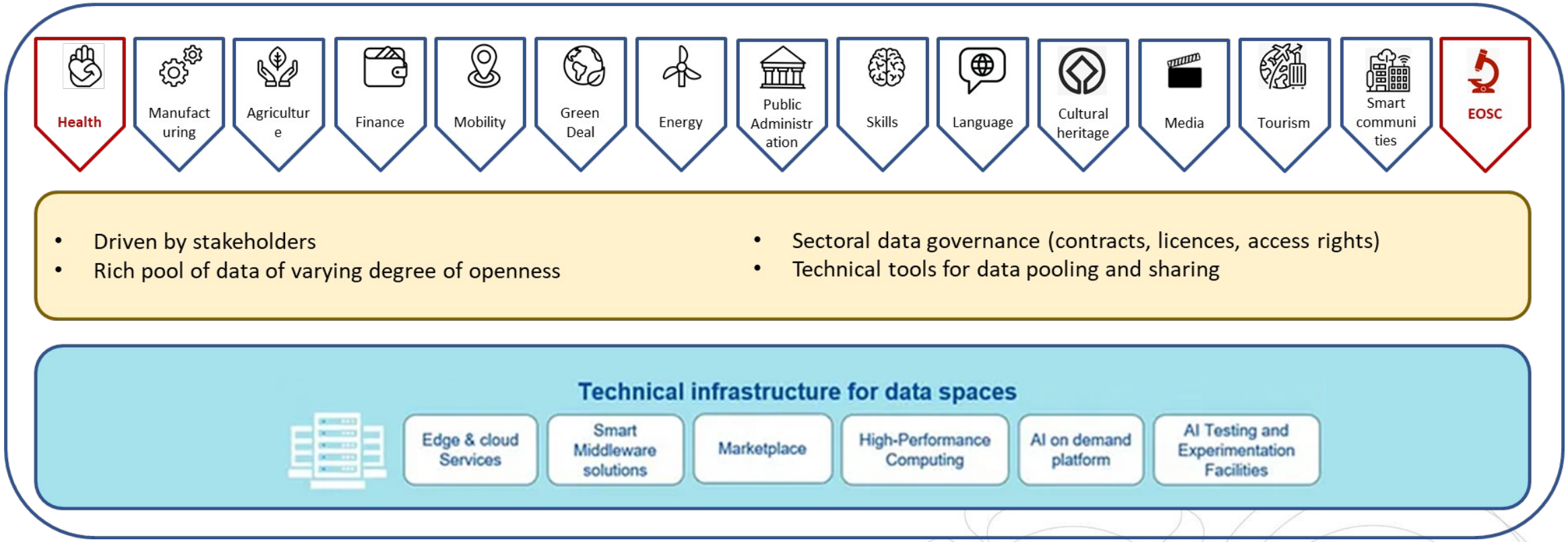
Overview of resources, services and architecture of the upcoming EHDS

- Matthias Löbe — IMISE U Leipzig, AA EOSC and EHDS synergies
- Wolmar N. Åkerström — Uppsala University, AA Interoperability liaison group
- Bernd Saurugger — TU Wien, AA EOSC Federation monitoring
- Juan Gonzalez-Garcia — IACS, TF chair
- Petr Holub — BBMRI-ERIC & Masaryk University, TF chair

MEDIZIN INFORMATIK INITIATIVE

nfdi 4 health

Common European Data Spaces

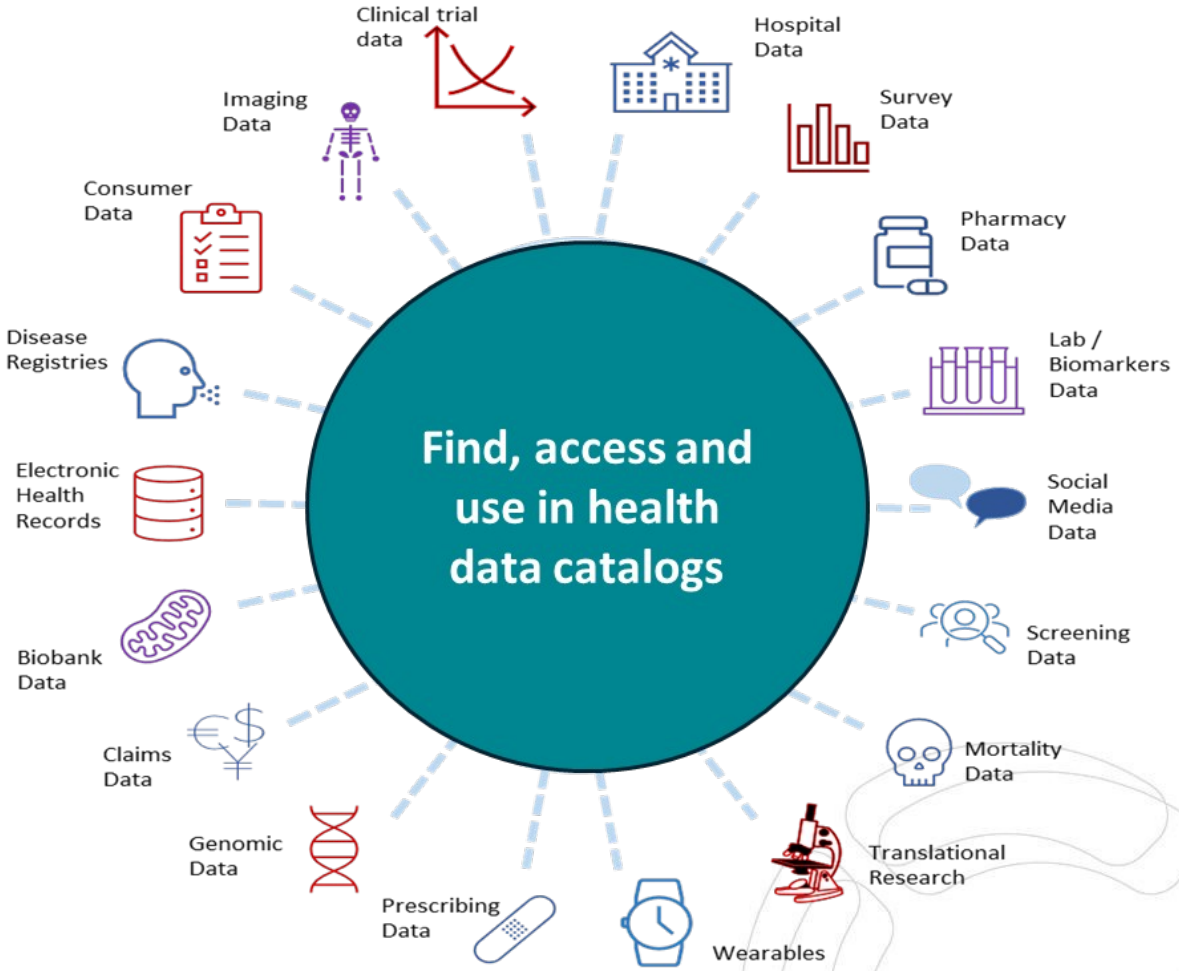


Source: EOSC

eosc | Health Data Task Force

Health Data Diversity

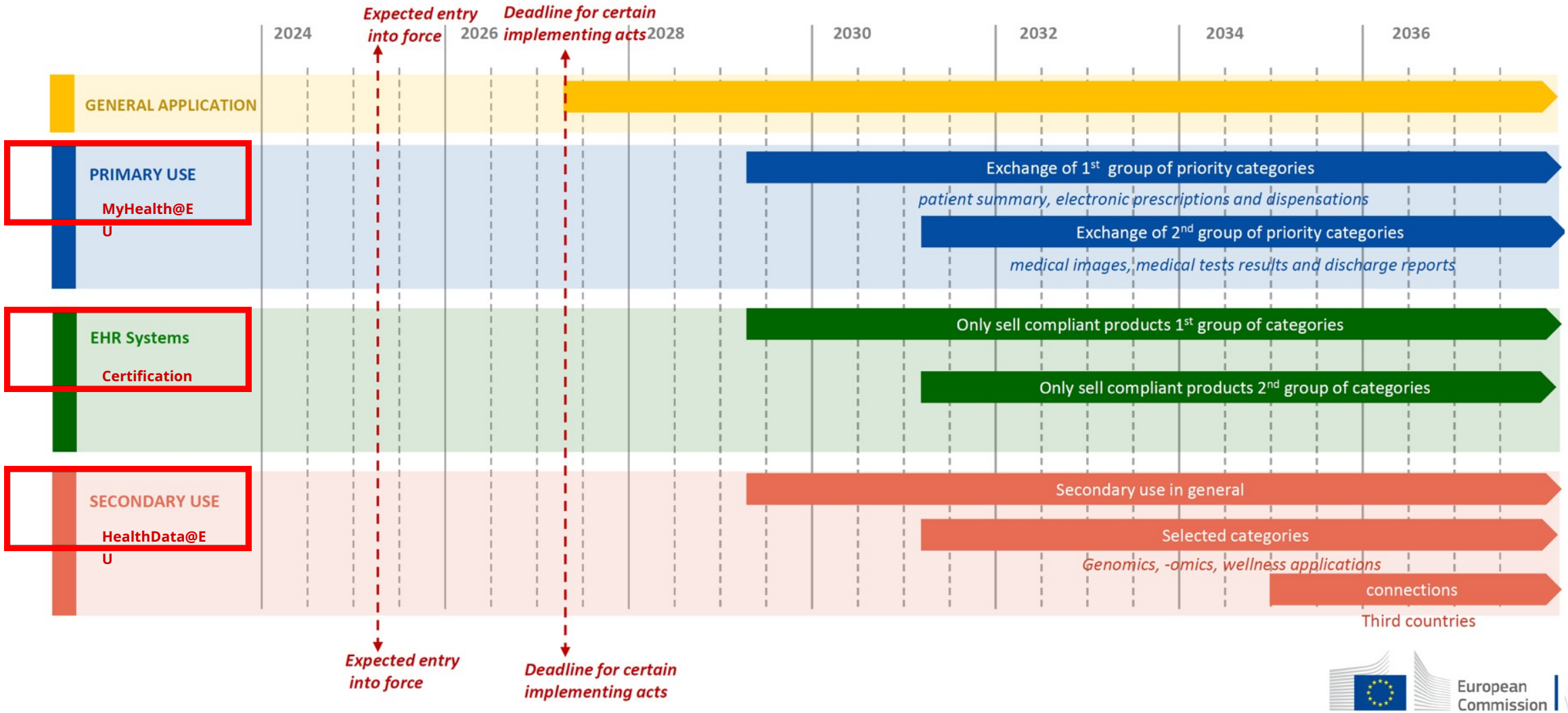
A lot of different sources provided by different stakeholders



Source: Aarestrup, F.M., Albeyatti, A., Armitage, W.J. *et al.* Towards a European health research and innovation cloud (HRIC). *Genome Med* 12, 18 (2020). https://doi.org/10.1186/s13073-020-0713-z

eosc | Health Data Task Force

EHDS — Overall Timeline



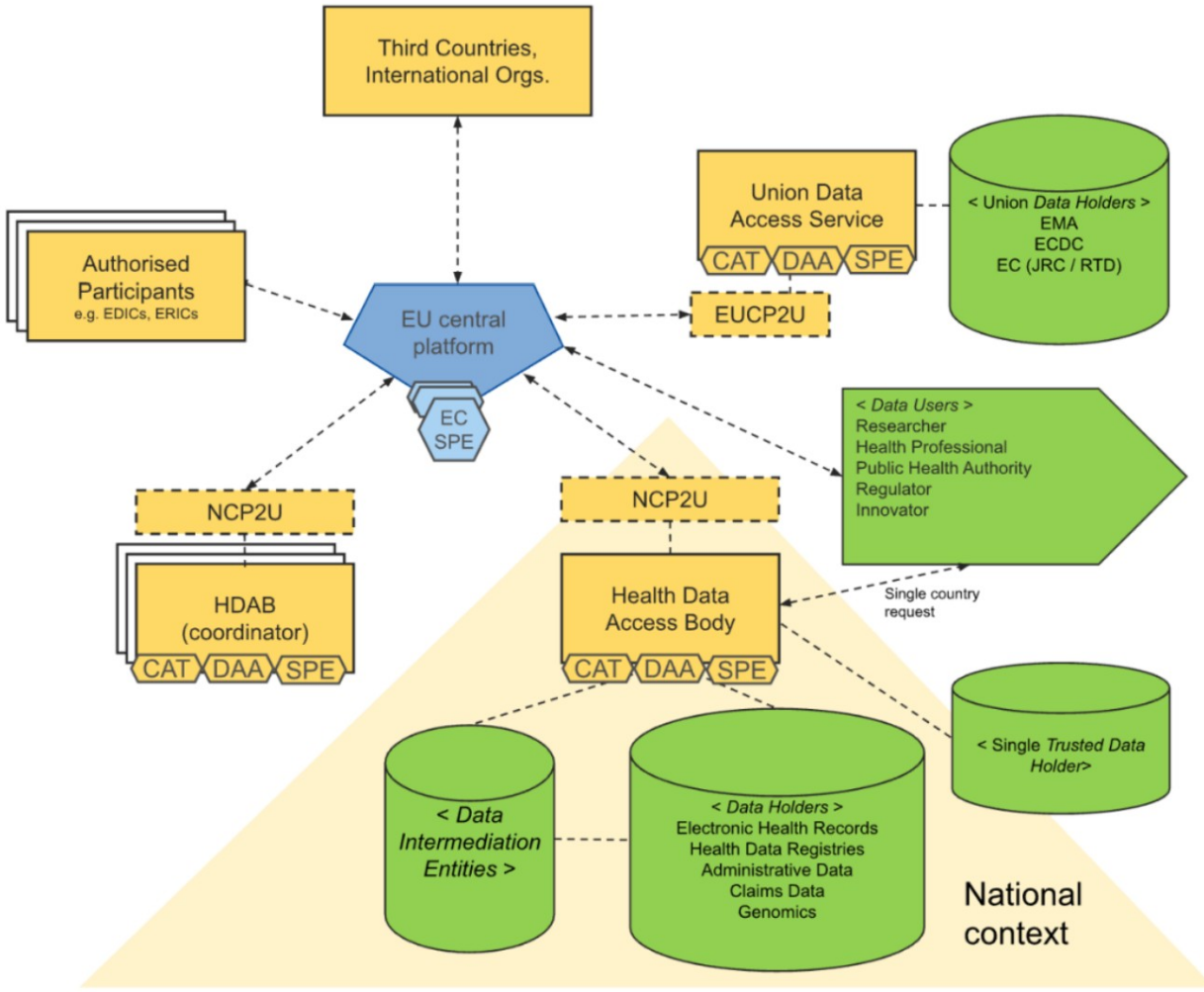
European Commission

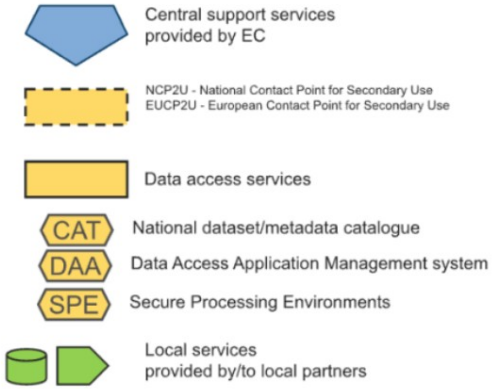
Source: European Commission

HealthData@EU EHDS infrastructure for secondary use of health data

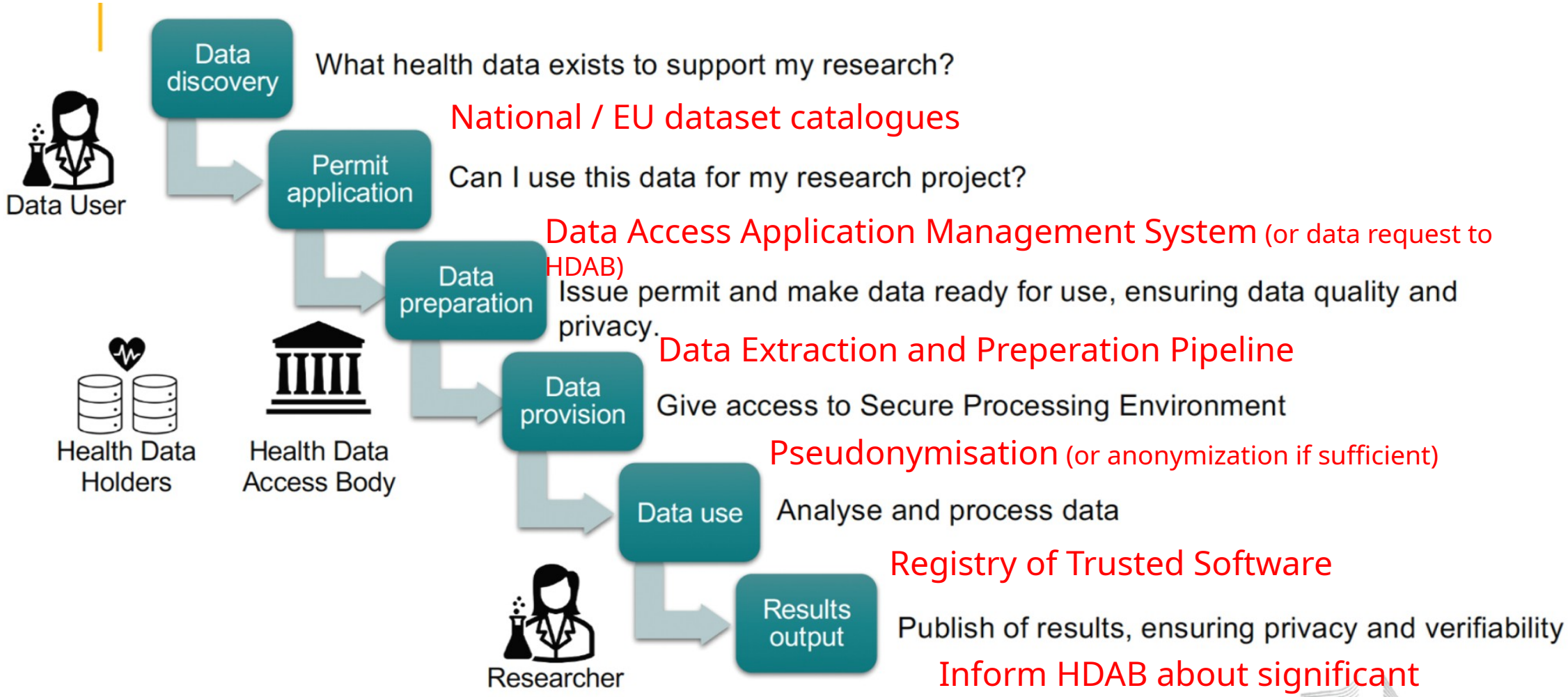
Cross-border secondary use infrastructure

HealthData@EU





Data User Journey



Source: European Commission

eosc | Health Data Task Force

Health Data Holders

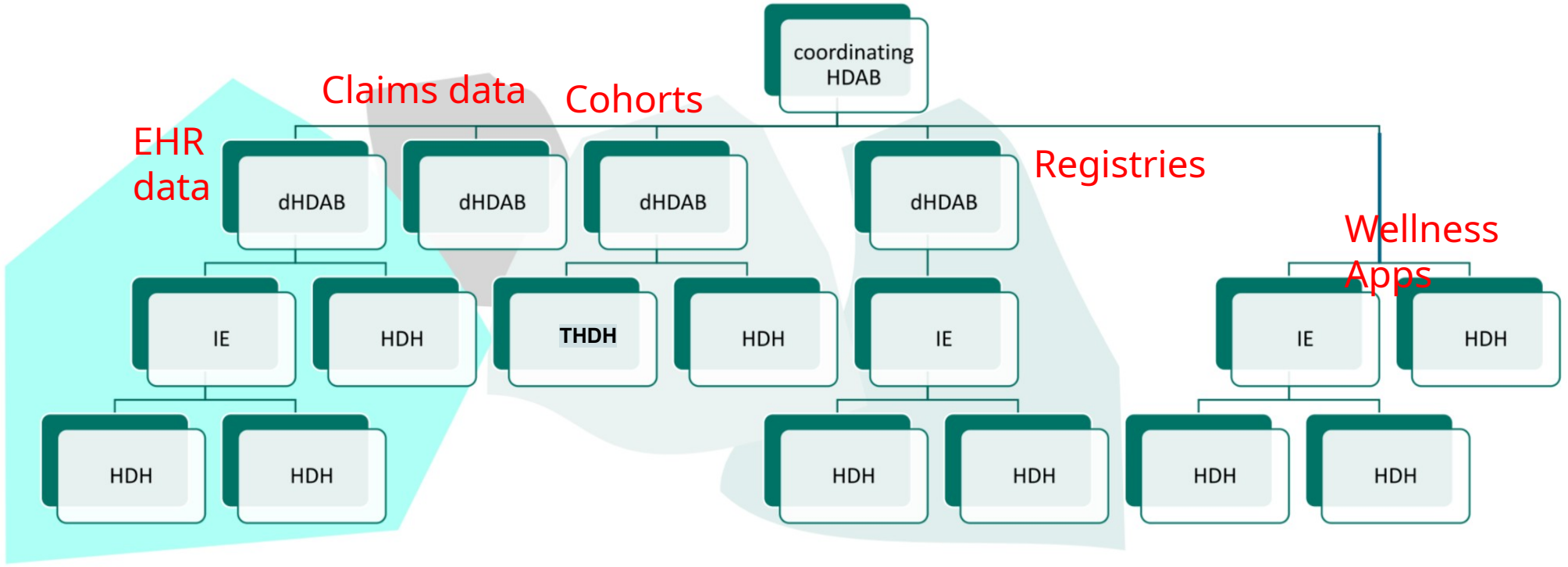
Affected by the EHDS regulation

Major stakeholders	Major activities	Major Information systems (HealthData@EU and member states)
● (Trusted) Data Holder ○ any organisation in healthcare, or reimbursement services, or developing health products/services, or research related to healthcare ○ individuals and micro-enterprises are exampted ● (per Member State) ○ National Contact Point ○ (Multiple) Health Data Access Bodies ○ Intermediation Entities ● EU ○ Central Catalogue and Portal	● Data Discoverability ● Data Access Management and Assessment ● Issuing permits ● Data Preparation ● Certify and audit SPEs ● Data Linkage ● Pseudonymisation ● Anonymisation ● Transparency and Publishing ● Opt-Out Management ● Significant Findings Reporting	● Data Catalogue ● Data Access Application Management System ● Secure Processing Environment ● Certified Electronic Health Record System ● Transparency Portal ● Consent Management ● Fee system

eosc | Health Data Task Force

Possible Ecosystem

Country-level, e.g. Germany



dHDAB - Domain-specific HDAB (optional)
IE - Intermediation Entities

HDH - Health Data Holder
THDH - Trusted Health Data Holder

Source: BfArM, Germany

EHDS and EOSC comparison (1)

Data

	EHDS	EOSC
Domain	● Primary use of health data for personal treatment ● Secondary use of health data for research	● In principle all kinds of data
Privacy	● Open data (EU open high-value data) ● Restricted data (EU Data Governance Act) ● **Sensitive Data (EHDS data)**	● **Open data** ● **Restricted data** ● Sensitive data
Discovery	● Health profile for DCAT-AP (HealthDCAT-AP)	● Domain-specific DCAT-AP profiles, schema.org, DataCite
Access	● Sensitive data will never be directly accessible	● Sensitive data can be shared directly with third parties if legislation allows
Focus	● Reusability	● Reusability and reproducibility

eosc | Health Data Task Force

EHDS and EOSC comparison (2)

Architecture

	EHDS	EOSC
Architecture	● Federation of data and service providers ● Homogeneous, established by the EU member states ● Central EU node	● Federation of data and service providers ● Heterogeneous, top-down ● Central EU node
Services types	● Services are defined in the regulation, certain freedom in shaping	● Service integrated in the EU node ● Onboarded services from third-parties ● Other services from third-parties
Authentification	● eID	● AARC (incl eID interface)
Middleware	● Web Services (AS4)	● SIMPL
Analyses	● Secure Processing Environment	● Trusted Research Environment

eosc | Health Data Task Force

EHDS and EOSC comparison (3)

Governance

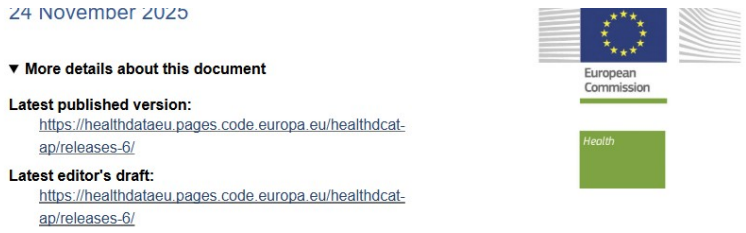
	EHDS	EOSC
Obligation	● Compulsory	● Voluntary
Realization	● Mainly by the member states	● Mainly through EU projects
Legal basis	● No consent required, but right to opt-out ● Existing legal grounds continue to be valid	● Research infrastructures have health data with different legal grounds
Timeline	● Starting March 2025 with the publication of the regulation ● 2027: implementation acts ● 2029: most health data types ● 2031: remaining data types (e.g. genomics)	● First node available since October 2024 ● Federation starting in 2025 ● Build-up phases
Fees	● Subject to a fee (cost model under development)	● Not decided
Transparenc	● All access applications and data permits will be shared publicly	● Commitment to FAIR data

eosc | Health Data Task Force

Data set description with HealthDCAT-AP

Subprofile of DCAT-AP

- Developed in the HealthData@EU project
- Release 6 from 24 November, 2025
- Currently being further developed and tested for applicability as part of the TEHDAS2 project
- Preliminary work:
  - 2022: M6.1 Landscape analysis of available metadata catalogues and metadata standards in use
  - 2023: M6.2 Technical working group on the transition from existing metadata templates to HealthDCAT-AP
  - 2024: D6.2 Recommendations on further development and deployment for possible EU-wide uptake
  - 2025: D5.1 Guideline for data holders on data description


24 November 2025

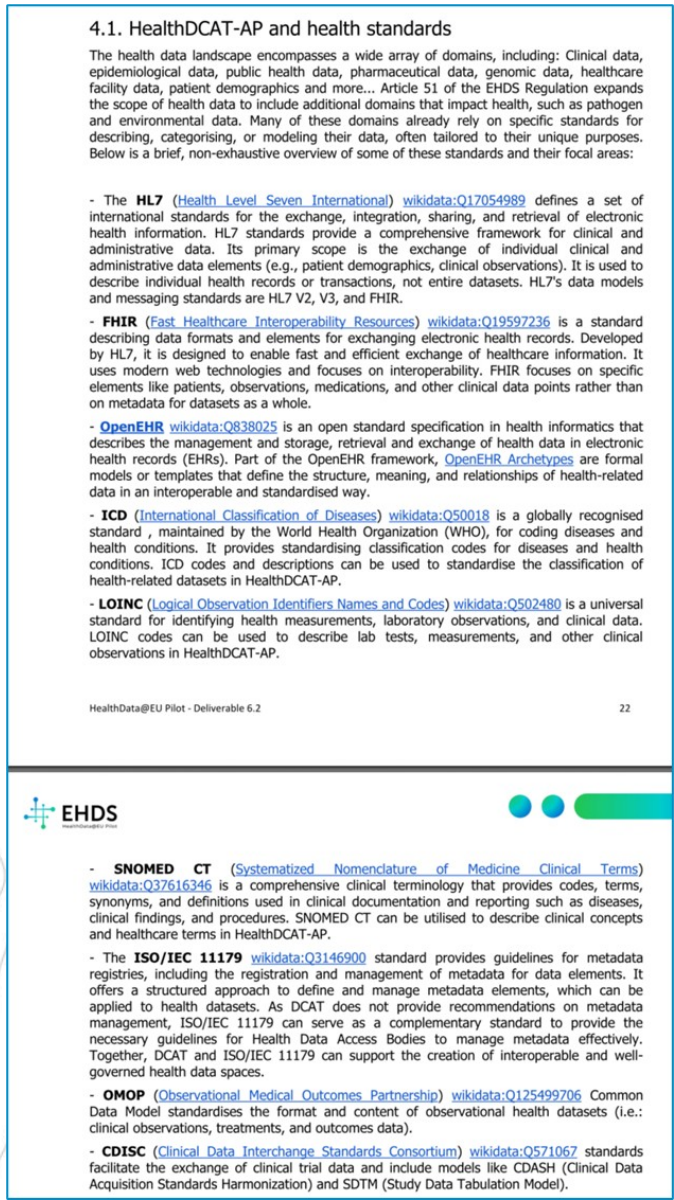
▼ More details about this document

Latest published version:
https://healthdataeu.pages.code.europa.eu/healthdcat-ap/releases-6/

Latest editor's draft:
https://healthdataeu.pages.code.europa.eu/healthdcat-ap/releases-6/

European Commission

Health


4.1. HealthDCAT-AP and health standards

The health data landscape encompasses a wide array of domains, including: Clinical data, epidemiological data, public health data, pharmaceutical data, genomic data, healthcare facility data, patient demographics and more... Article 51 of the EHDS Regulation expands the scope of health data to include additional domains that impact health, such as pathogen and environmental data. Many of these domains already rely on specific standards for describing, categorising, or modeling their data, often tailored to their unique purposes. Below is a brief, non-exhaustive overview of some of these standards and their focal areas:

- The **HL7** (Health Level Seven International) wikidata:Q17054989 defines a set of international standards for the exchange, integration, sharing, and retrieval of electronic health information. HL7 standards provide a comprehensive framework for clinical and administrative data. Its primary scope is the exchange of individual clinical and administrative data elements (e.g., patient demographics, clinical observations). It is used to describe individual health records or transactions, not entire datasets. HL7's data models and messaging standards are HL7 V2, V3, and FHIR.
- **FHIR** (Fast Healthcare Interoperability Resources) wikidata:Q19597236 is a standard describing data formats and elements for exchanging electronic health records. Developed by HL7, it is designed to enable fast and efficient exchange of healthcare information. It uses modern web technologies and focuses on interoperability. FHIR focuses on specific elements like patients, observations, medications, and other clinical data points rather than on metadata for datasets as a whole.
- **OpenEHR** wikidata:Q838025 is an open standard specification in health informatics that describes the management and storage, retrieval and exchange of health data in electronic health records (EHRs). Part of the OpenEHR framework, OpenEHR Archetypes are formal models or templates that define the structure, meaning, and relationships of health-related data in an interoperable and standardised way.
- **ICD** (International Classification of Diseases) wikidata:Q50018 is a globally recognised standard , maintained by the World Health Organization (WHO), for coding diseases and health conditions. It provides standardising classification codes for diseases and health conditions. ICD codes and descriptions can be used to standardise the classification of health-related datasets in HealthDCAT-AP.
- **LOINC** (Logical Observation Identifiers Names and Codes) wikidata:Q502480 is a universal standard for identifying health measurements, laboratory observations, and clinical data. LOINC codes can be used to describe lab tests, measurements, and other clinical observations in HealthDCAT-AP.

HealthData@EU Pilot - Deliverable 6.2 22

EHDS

- **SNOMED CT** (Systematized Nomenclature of Medicine Clinical Terms) wikidata:Q37616346 is a comprehensive clinical terminology that provides codes, terms, synonyms, and definitions used in clinical documentation and reporting such as diseases, clinical findings, and procedures. SNOMED CT can be utilised to describe clinical concepts and healthcare terms in HealthDCAT-AP.
- The **ISO/IEC 11179** wikidata:Q3146900 standard provides guidelines for metadata registries, including the registration and management of metadata for data elements. It offers a structured approach to define and manage metadata elements, which can be applied to health datasets. As DCAT does not provide recommendations on metadata management, ISO/IEC 11179 can serve as a complementary standard to provide the necessary guidelines for Health Data Access Bodies to manage metadata effectively. Together, DCAT and ISO/IEC 11179 can support the creation of interoperable and well-governed health data spaces.
- **OMOP** (Observational Medical Outcomes Partnership) wikidata:Q125499706 Common Data Model standardises the format and content of observational health datasets (i.e.: clinical observations, treatments, and outcomes data).
- **CDISC** (Clinical Data Interchange Standards Consortium) wikidata:Q571067 standards facilitate the exchange of clinical trial data and include models like CDASH (Clinical Data Acquisition Standards Harmonization) and SDTM (Study Data Tabulation Model).

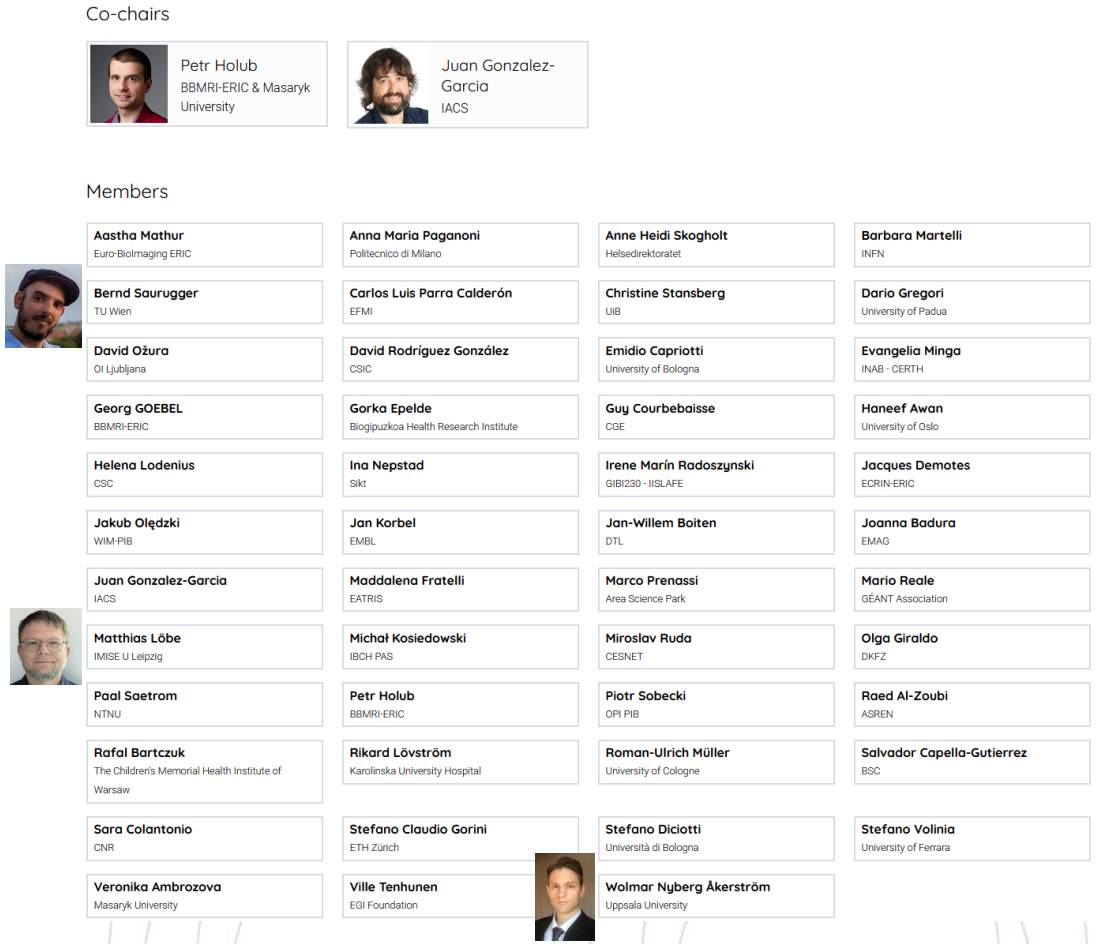
eosc | Health Data Task Force

Gap analysis, challenges and synergies

- The EHDS is being introduced via an EU regulation, i.e., it is not voluntary
  - Focus heavily on industrial use
- Data sets will be in the EHDS but also in other data spaces like EOSC
  - Interoperable metadata descriptors (Cataloguing)
  - Interoperable data standards (Data linkage)
  - Interoperable connectors
- Data will not be downloaded to the user
  - Overlap: TREs and SPEs
  - Governance models
  - Privacy: audits and certified software
  - Rules for anonymisation unclear
- Interoperability
  - Problems with missing “study-level” metadata from medical studies (study design)
  - Problems with required controlled vocabularies (WikiData)
  - List of predefined core variables unclear
  - Common data models (FHIR, OMOP, openEHR)

EOSC Health Data Task Force

- Website
- **Deliverable 1** — Identification of gaps, redundancies, and possible synergies regarding different user journeys of researchers in EHDS and EOSC
- **Deliverable 2** — Strategic view of EOSC supporting the adoption of EHDS (work in progress)
- Second Joint Action Towards the European Health Data Space (TEHDAS2)
- HealthData@EU Central Platform (Metadata Catalog)
- HealthDCAT-AP Specification (Release 6)
- EOSC & DG-SANTE Round Table on EHDS and EOSC (Oktober 2024)



Co-chairs

Petr Holub
BBMRI-ERIC & Masaryk University

Juan Gonzalez-Garcia
IACS

Members

Aastha Mathur Euro-BioImaging ERIC	Anna Maria Paganoni Politecnico di Milano	Anne Heidi Skogholt Helsedirektoratet	Barbara Martelli INFN
Bernd Saurugger TU Wien	Carlos Luis Parra Calderón EFMI	Christine Stansberg UiB	Dario Gregori University of Padua
David Ožura OI Ljubljana	David Rodríguez González CSIC	Emidio Capriotti University of Bologna	Evangelia Minga INAB - CERTH
Georg GOEBEL BBMRI-ERIC	Gorka Epelde Biogipuzkoa Health Research Institute	Guy Courbebaisse CGE	Haneef Awan University of Oslo
Helena Lodenius CSC	Ina Nepstad Sikt	Irene Marín Radoszynski GIBI230 - IISLAFE	Jacques Demotes ECRIN-ERIC
Jakub Olędzki WIM-PIB	Jan Korbel EMBL	Jan-Willem Boiten DTL	Joanna Badura EMAG
Juan Gonzalez-Garcia IACS	Maddalena Fratelli EATRIS	Marco Prenassi Area Science Park	Mario Reale GÉANT Association
Matthias Löbe IMISE U Leipzig	Michał Kosiedowski IBCH PAS	Miroslav Ruda CESNET	Olga Giraldo DKFZ
Paal Saetrom NTNU	Petr Holub BBMRI-ERIC	Piotr Sobecki OPI PIB	Raed Al-Zoubi ASREN
Rafal Bartczuk The Children's Memorial Health Institute of Warsaw	Rikard Lövström Karolinska University Hospital	Roman-Ulrich Müller University of Cologne	Salvador Capella-Gutierrez BSC
Sara Colantonio CNR	Stefano Claudio Gorini ETH Zürich	Stefano Diciotti Università di Bologna	Stefano Volinia University of Ferrara
Veronika Ambrozova Masaryk University	Ville Tenhunen EGI Foundation	Wolmar Nyberg Åkerström Uppsala University	

EXTRACT

A distributed data-mining software platform for
extreme data across the compute continuum

TASKA

Transient Astrophysics
with a Square Kilometre Array pathfinder

Baptiste Cecconi, Stéphane Aicardi and the EXTRACT collaboration

extract-project.eu

The EXTRACT Project has received funding from the European Union's Horizon Europe programme under grant agreement number 101093110.

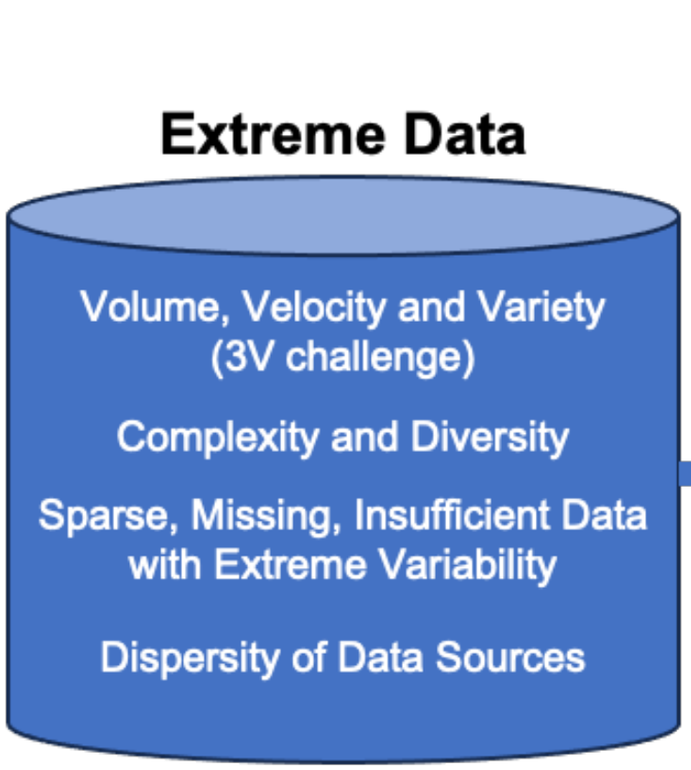
EXTRACT

A distributed data-mining software platform for extreme data across the compute continuum

EU project

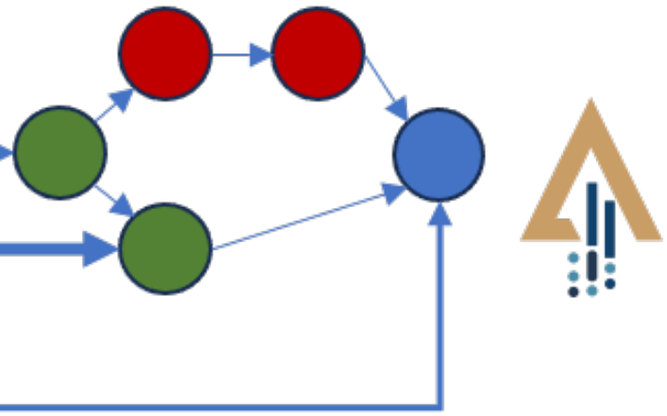
https://extract-project.eu/

EXTRACT aims to create a data-mining **software platform** for **extreme data** across the **compute continuum**



Data Mining Software Platform

- Data infrastructures and AI & Big-data frameworks
- Data-driven orchestration
- Interoperability

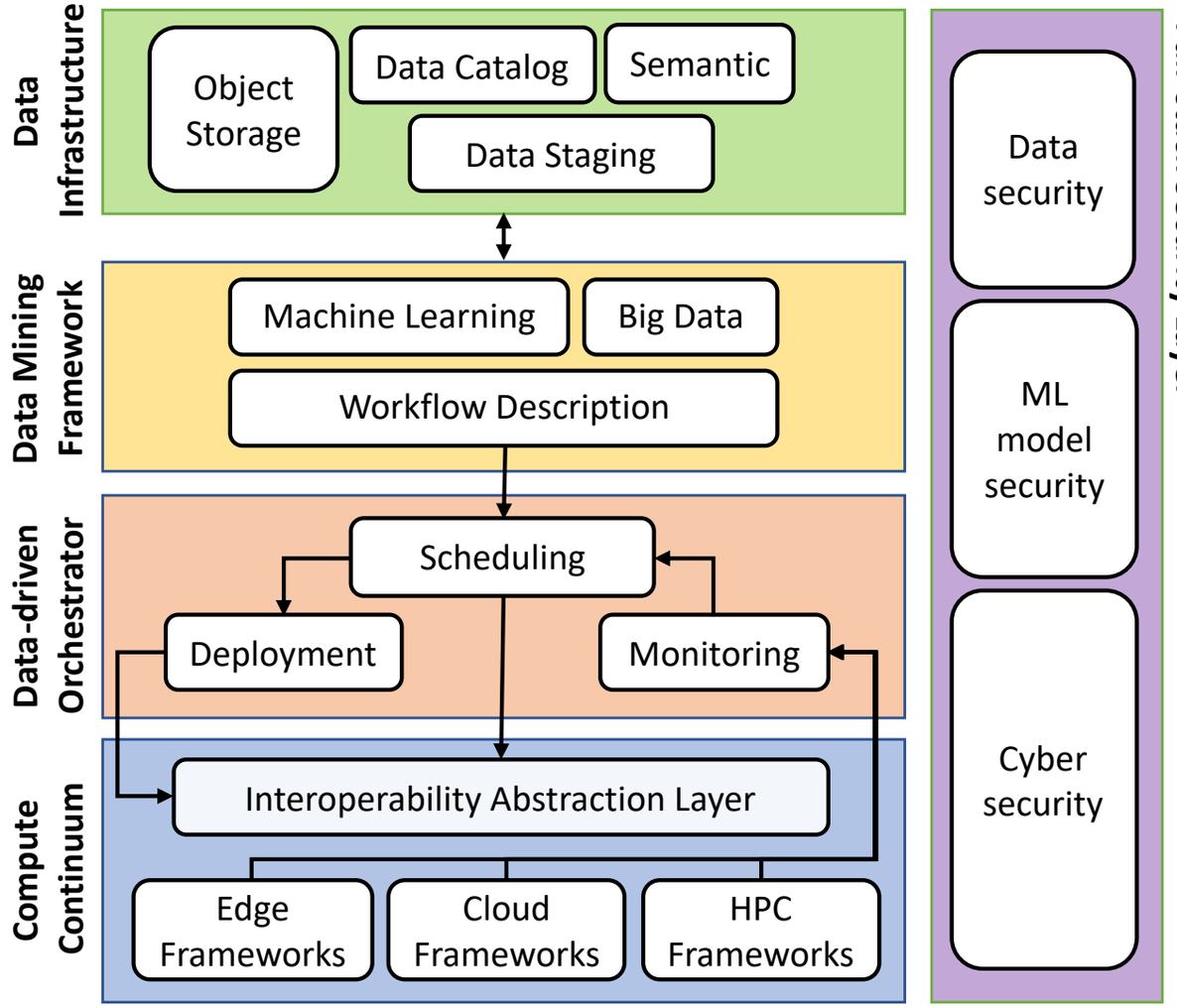


Compute Continuum
HPC
Edge
Cloud

EXTRACT

A distributed data-mining software platform for extreme data across the compute continuum

- Handle the complete lifecycle and value chain of extreme data
  - **Data collection** across highly distributed and heterogeneous sources
  - **Data mining** of meaningful, accurate, reliable and useful knowledge
  - **Secure and trustworthy used of knowledge** by applications and end users



- Everything looks local

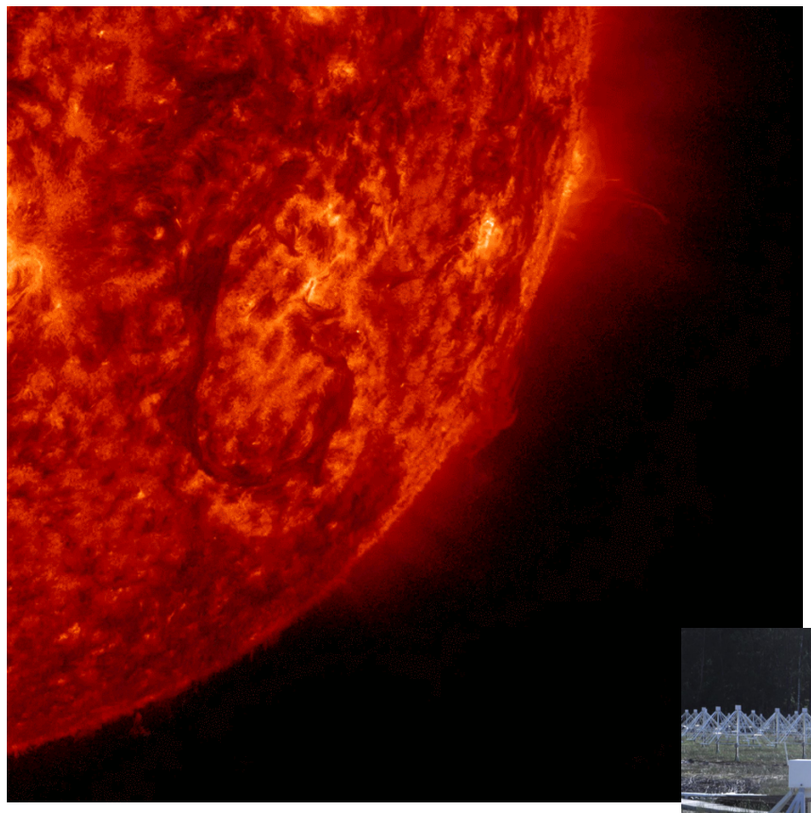
EXTRACT Use-Cases, sharing the same platform

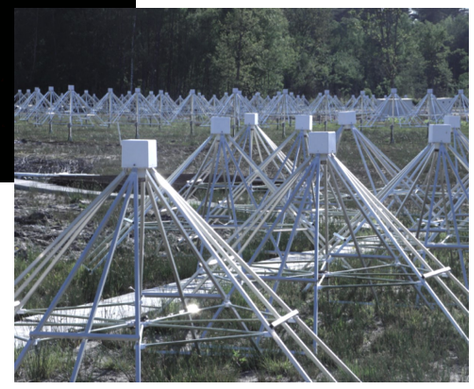
PER

Personalised Evacuation Route (PER)
in the City of Venice based on an
Urban Digital Twin and an AI engine

TASKA

Transient Astrophysics with the Square Kilometre Array pathfinder (TASKA)
NenuFAR generating high-volume and high-velocity data





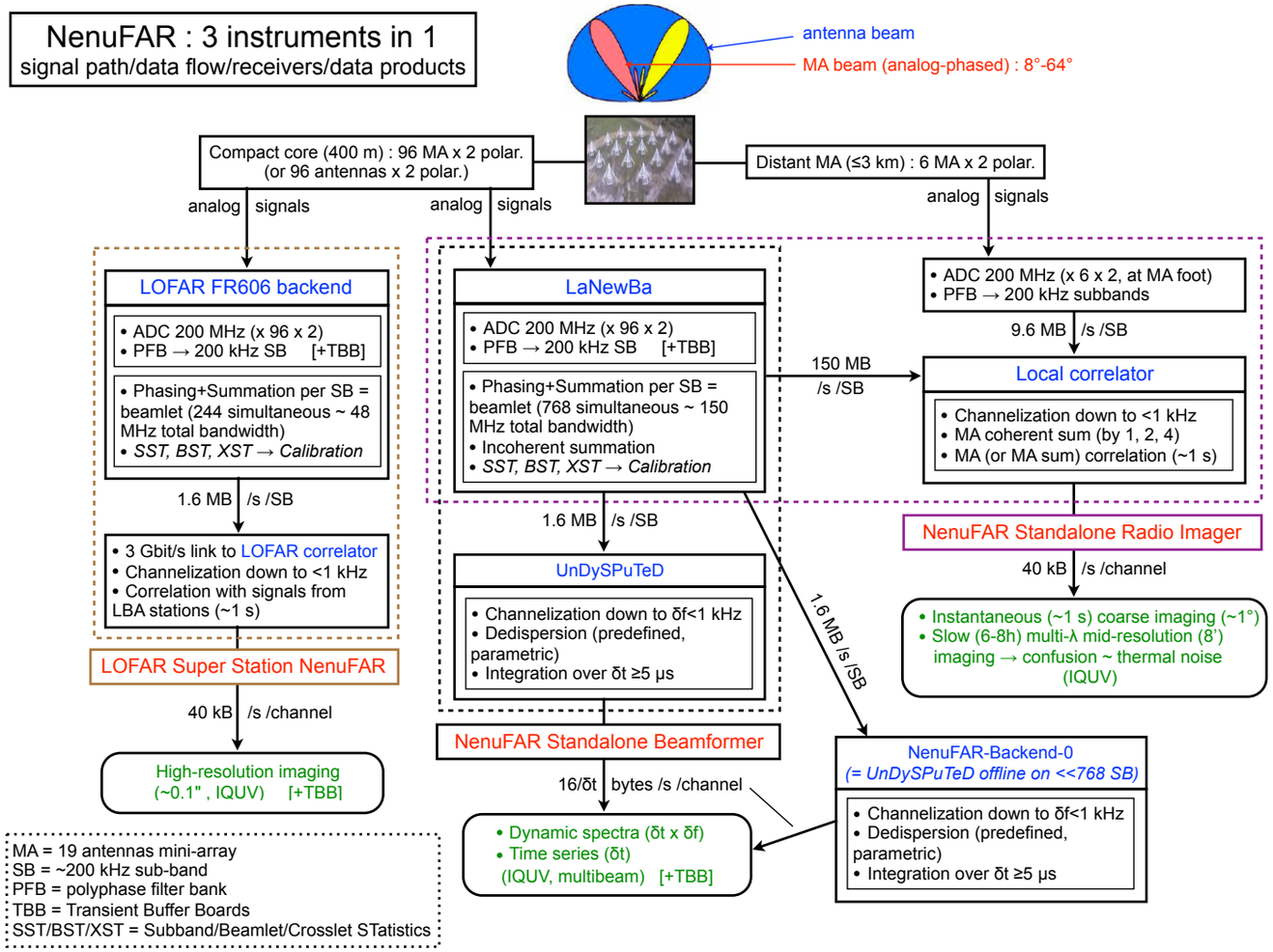
NenuFAR *New extension in Nançay Upgrading loFAR*

Pathfinder de SKA (LOW) , Infrastructure de recherche F = 20-80 MHz

N_A~2000 antennes Fonctionnement en mode réseau phasé et interféromètre

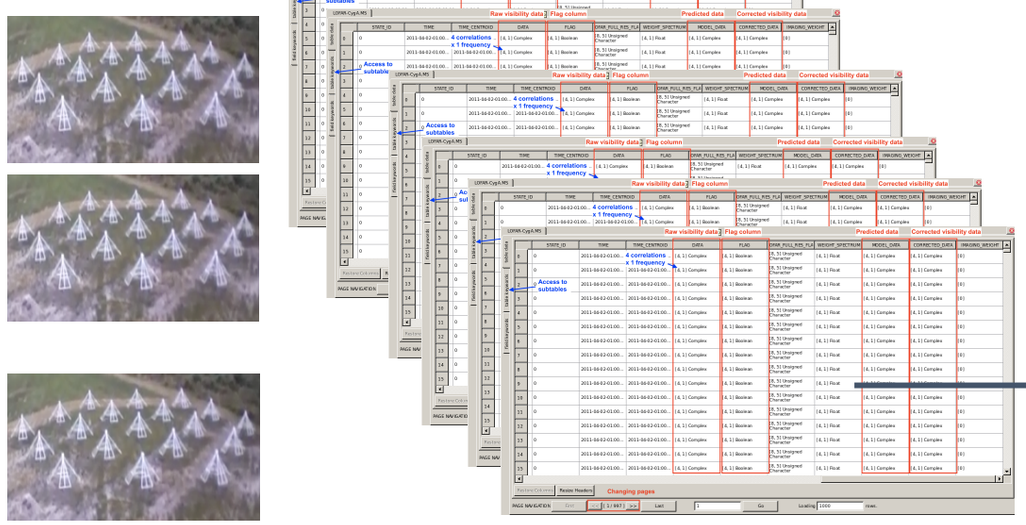
Extreme data with NenuFAR

- **~2000 antennas** (96 mini-arrays of 19 antennas + 6 remotes mini-arrays)
- **Beamform data:**
  raw data rate = 1.2 GB/s (~35 PB/yr)
- **Imaging data:**
  raw data rate = 8.6 TB/hr (~74 PB/yr)
- **Local data storage:** 3.5 PB
- **Science teams are reducing data** down to about 1 PB/yr (>1/100 of raw data rate)



- **Reduced data transferred to distributed datacenter**
  (currently: 3 PB in Meudon and 5 PB in Orléans)
- **Processing of imaging data on distributed datacenter:**
  data can't be moved easily, need to process where the data is located

Use-Case C: Workflow orchestration for radiointerferometry



Starting dataset: Visibilities
(Measurement Sets (MS) Format)

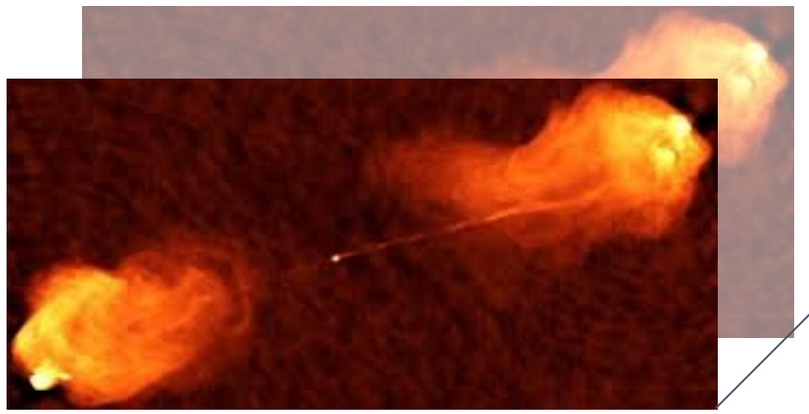
Optimal dataset distribution ?
(Multiple sites)

Data transfer
data processing

(Flagging, Rebining, Calibration, Imaging)

Flexible reduction?

(Multiple tools)
from combination of
known analytic "bricks"

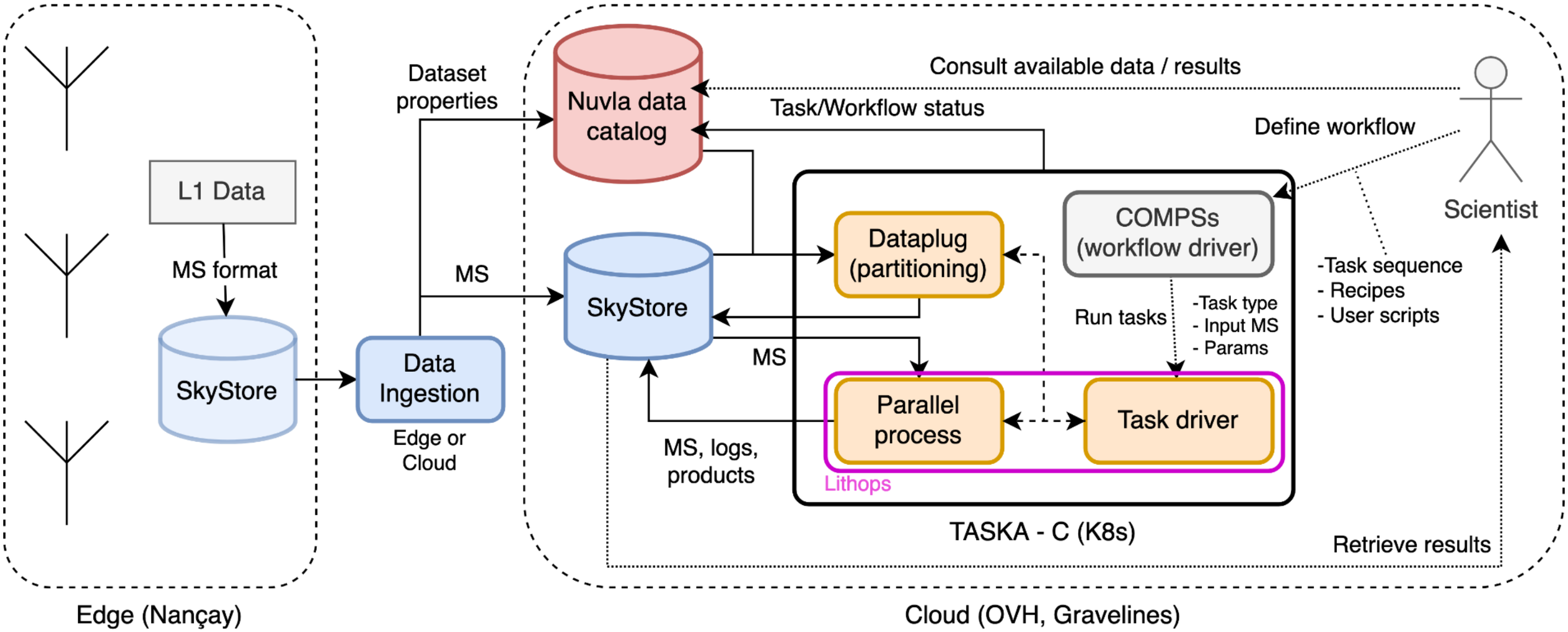


Final product: time/freq
Image cubes

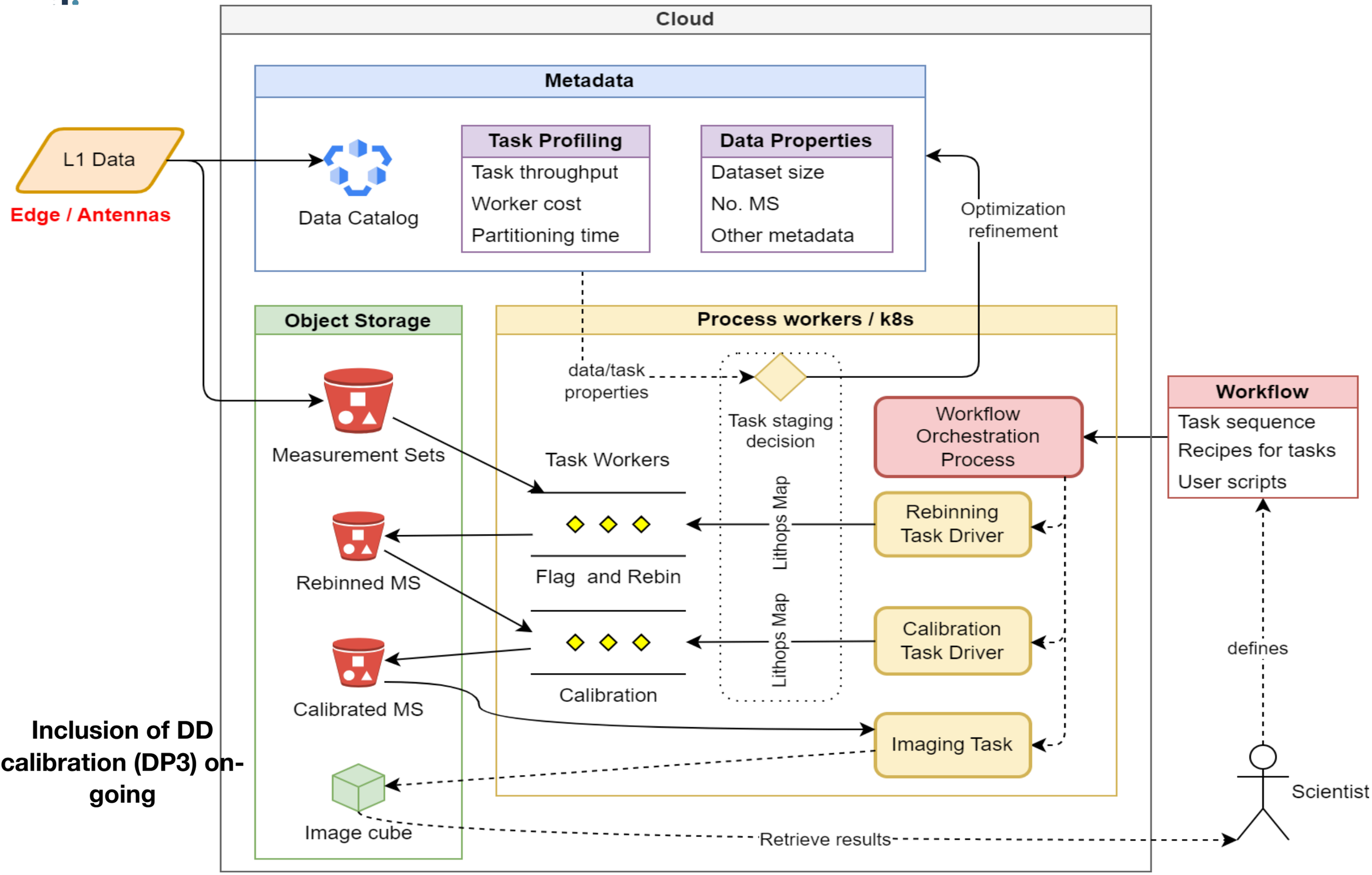
Analytics

Multiple tools
(scientific quality, fidelity)

Use-Case C: Architecture view



Use-Case C: User view

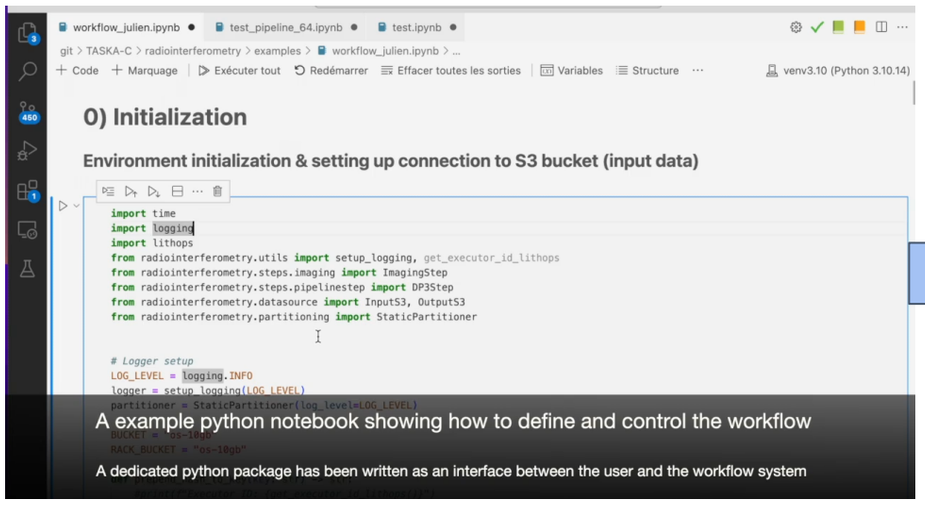


Inclusion of DD calibration (DP3) on-going

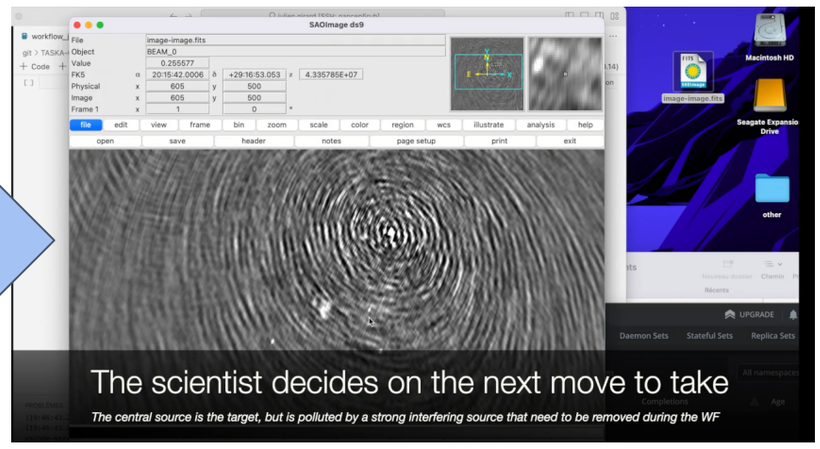
TASKA - MVP “Interactive” Workflow

- Built as a “wrapper” that interacts with the astronomy community tools
  High potential impact because of the platform deployment in other communities (security, medical, resource management, etc.)
- Easy to invoke, easy to code, easy to customize, easy to “chain”: *natively made for workflows*
- Each task has a “definition” block and a “run” block: *separating the workflow building from its running*
- Run as a python script or in a python notebook (cf. DEMO video)

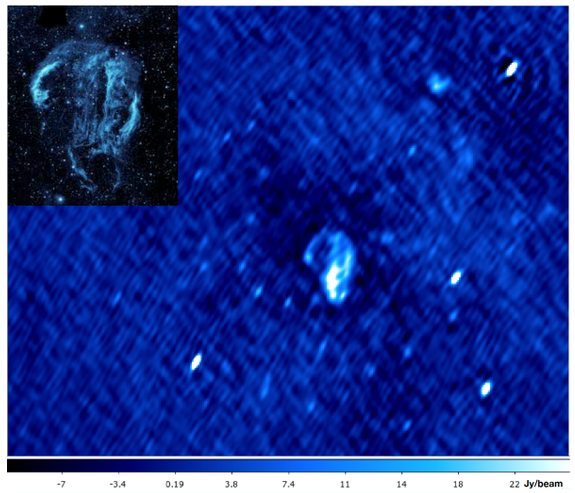
```
import time
import logging
import lithops
from radiointerferometry.utils import setup_logging, get_executor_id_lithops
from radiointerferometry.steps.imaging import ImagingStep
from radiointerferometry.steps.pipelinestep import DP3Step
from radiointerferometry.datasource import InputS3, OutputS3
from radiointerferometry.partitioning import StaticPartitioner
```



Controlled through a python notebook (S3, data partitioning, worker management, …)



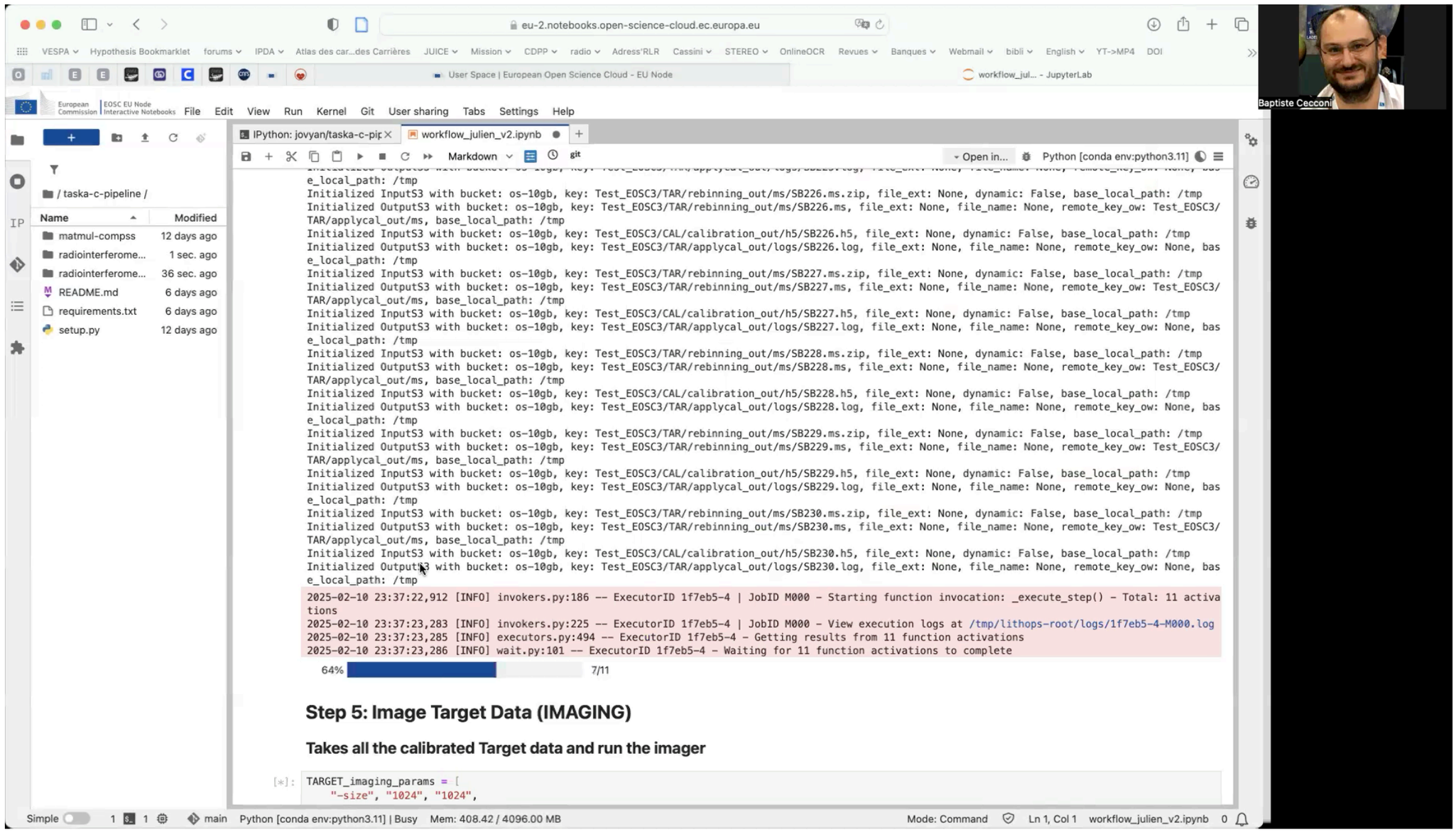
The final products are then retrieved on the scientist computer



…as if the process and data were **local**

TASKA - Demos

- Demo 1 (video): Running interactive workflow (EOSC-EU node Notebook service, using OVH K8S)
- Demo 2 (video): Running an automated workflow (EOSC-EU node Cloud Container)



Current Status

Running on K8S managed by OVH

- Full MVP implemented
- Interactive mode (Jupyter notebook)
- Data available on local S3 storage
- Fully remote processing

Other deployments:

- EGI K8S [Rancher @ CESNET]: tested ✅
- EOSC EU Node K8S [OKD]: tested ✅
- ObsParis K8S [OKD]: tested ✅
- BRGM K8S [Rancher]: next

Ongoing developments:

- Automated workflow mode
- Integrate data catalogue for automated workflow
- Integrate provenance management for efficient reprocessing
- Integrate new step with HPC - slurm
- Implement data staging and data clean-up steps

EXTRACT - TASKA - Summary

- We have developed a **framework for distributed data computing** on cloud clusters
- Currently validating
  - unsupervised/automated workflow
  - running a step on an HPC resource
  - running on a multi-cluster scale (data distributed in several data centers)

- Application on NenuFAR (SKA -Square Kilometer Array- pathfinder)
- Clear huge potential for SRCNet (SKA Region Center Network)

- EOSC tests:
  - deployment of framework on EOSC-EU node (Cloud Container Platform) : easy
  - for full deployment / operation, we miss:
    - S3 storage service
    - a data catalog (e.g., RUCIO)