FAIRCORE4EOSC project

Taking EOSC into the future

Tommi Suominen (CSC)

FAIRCORE4EOSC

Developing EOSC-Core components to enable a FAIR EOSC ecosystem

It focuses on the development and realisation of core components for the European Open Science Cloud (EOSC). Supporting a FAIR EOSC and addressing gaps identified in the Strategic Research and Innovation Agenda (SRIA). Leveraging existing technologies and services, the project will develop components aimed to improve the discoverability and interoperability of an increased amount of research outputs.

https://faircore4eosc.eu/

Tommi Suominen

eosc | FAIRCORE4EOSC

Core Components Supporting a FAIR EOSC

CAT
EOSC Compliance Assessment Toolkit

MSCR
EOSC Metadata Schema and Crosswalk Registry

DTR
EOSC Data Type Registry

PIDMR
EOSC PID Meta Resolver

RSAC
EOSC Research Software APIs and Connectors

SWHM
EOSC Software Heritage Mirror

RAiD
EOSC Research Activity Identifier Service

PIDGraph
EOSC PID Graph

RDGraph
EOSC Research Discovery Graph

FAIRCORE4EOSC in a nutshell

- Just completed the first year -

Full name: Developing EOSC-Core components to enable a FAIR EOSC ecosystem

Research and Innovation Action

Budget: 10 million EUR

Duration: June 2022 – May 2025

Consortium: 22 partners, coordinated by CSC – IT Center for Science

Coordinator Tommi Suominen (CSC), **Project Manager** Anu Märkälä (CSC) and **Technical Coordinator** Mark van de Sanden (SURF)

Website: faircore4eosc.eu

Key results: In response to the gaps identified in the SRIA, the project will develop nine new EOSC-Core components aimed to improve the discoverability and interoperability of an increased amount of research outputs.

eosc | FAIRCORE4EOSC

Amsterdam, Netherlands – Kick-off meeting, June 2022

Implementation Challenges (SRIA) addressed

FAIRCORE4EOSC is research and innovation action that focusses on concrete service development to further the realisation of the MVE and web of FAIR data.

Priorities highlighted in the SRIA are the establishment of the Web of FAIR data and a Minimum Viable EOSC (MVE) by 2027, that is the core components and functions to enable EOSC to operate (the EOSC-Core).

- *Identifiers:* Introducing new resource types; machine-actionable persistent identifiers (PIDs); establishing a PID meta-resolver; standardising PID graphs; PID compliance framework to ensure compliance to the EOSC PID policy and to ensure quality of service for PIDs;
- *Metadata and Ontologies:* Provide or embrace/stimulate existing registries of metadata schemas, ontologies and crosswalks, develop services that build on metadata registries and can facilitate the creation and sharing of crosswalks;
- *Interoperability:* Enable discovery of data sources available in different formats, making search tools available; Provide tools for quality validation of metadata records and of digital objects; Implement EOSC PID Policy;
- *Research Software*: metadata description standards for research software, automated deposit of new releases into a scholarly repository and Software Heritage.

FAIRCORE4EOSC develops 9 new EOSC CORE components to address gaps identified in the SRIA

EOSC Research Discovery Graph (RDGraph) to deliver advanced discovery tools across EOSC resources and communities.

EOSC PID Graph (PIDGraph) to improve the way of interlinking research entities across domains and data sources on the basis of PIDs.

EOSC Metadata Schema and Crosswalk Registry (MSCR) to support publishing, discovery and access of metadata schemas and provide functions to operationalise metadata conversions by combining crosswalks.

EOSC Data Type Registry (DTR) to provide user friendly APIs for metadata imports and access to different data types and metadata mappings.

EOSC PID Meta Resolver (PIDMR) to offer users a single PID resolving API in which any kind of PID can be resolved through a single, scalable PID resolving infrastructure.

EOSC Compliance Assessment Toolkit (CAT) to support the EOSC PID policy compliance and implementation.

EOSC Research Activity Identifier Service (RAiD) to mint PIDs for research projects, allowing to manage and track project related activities.

EOSC Research Software APIs and Connectors (RSAC) to ensure the long-term preservation of research software in different disciplines.

EOSC Software Heritage Mirror (SWHM) to equip EOSC with a mirror of the Software Heritage universal source code archive.

EOSC PID Meta Resolver

PIDMR

EOSC PID Meta Resolver

Provides users with a common interface to resolve different types of PIDs regardless of their originating system. Depending on options provided by each PID system, the PIDMR can resolve either to a landing page, metadata describing the digital object, or the underlying resource itself.

EOSC Research Activity Identifier Service

RAiD

EOSC Research Activity Identifier Service

RAiD is a new Persistent Identifier (PID) developed by the Australian Research Data Commons (ARDC) to mint persistent, unique and resolvable information for research projects. RAiD enables users and services to manage information about project-related participants, services, and outcomes. RAiD also collects related identifies (of funding, contributors, organisations, inputs, outputs, etc.), plus descriptive information about the project (title, description, subject, etc.), and stores them in an associated metadata record.

Problem addressed: RAiD connects existing persistent identifiers for researchers, institutions, outputs and tools with key project information to create a timeline of research projects, as well as facilitating project reporting and providing a project archive. The EOSC RAiD service will be part of a global database of research project information, which is accessible through a global API.

EOSC Research Software APIs and Connectors

RSAC

EOSC Research Software APIs and Connectors

- To identify software source code artifacts, the SoftWare Heritage ID (a.k.a SWHID) is an intrinsic identifier that is calculated from the content itself. It is not a registered identifier, such as the DOI.
- challenge: software is a complex digital artifact with many levels of granularity. Researchers need to Identify an algorithm in a file and a full project with many branches and releases. Also, identifying the metadata record and the code itself requires different types of identifiers.
- solution: adoption of the SWHID with the development of the RSAC sub-components provides a mechanism to reference Research Software at all granularity levels.
- The adoption of the SWHID was recommended by the EOSC SIRS report.
- In FC4EOSC, WP6 (T6.4) has launched the SWHID WG which drafted a first version of the specifications that will be submitted for ISO standardization. This will impact the SWHID adoption in the scholarly ecosystem.

EOSC (PID Policy) Compliance Assessment Toolkit

CAT

EOSC Compliance Assessment Toolkit

The Compliance Assessment Toolkit (CAT) will support the EOSC PID policy with services to encode, record, and query compliance with the policy. The CAT provides user and application interfaces to encode, record, and query compliance with the policy.

There is currently no unique way to assess the compliance of Persistent Identifier (PID) schemes in use. In order for EOSC to implement a consistent PID Policy, it needs to have an assessment tool in place to assess and verify compliance. This will be provided by the Compliance Assessment Toolkit (CAT).

Potential users: Repository (PID) managers and PID service providers aiming to evaluate the compliance of their own services and publish a public record of such compliance, or to query existing assessments for other services in the PID ecosystem. End users of persistent identifiers are supported with assessments of the actors in the ecosystem. Guidance and best practices, linked to compliance criteria, assists all actors with improvements to the use and provision of services.

EOSC PID Graph

PIDGraph

EOSC PID Graph

Challenge: Due to the diverse domains in scientific research, there exist many types of PID covering different aspects of the research lifecycle, and not necessarily easily accessible information about the relationships between individual PIDs of diverse types.

Addressed by: The creation of the EOSC PIDGraph service, providing a PID-focused graph allowing connections to be drawn between this variety of research PIDs and data exchange.

EOSC Work: Efforts to incorporate identifiers and metadata from research outputs and resources throughout the research lifecycle (over 30 different outputs and resource types), and *in particular identifiers and metadata for emerging types such as material samples, instruments, preprints and DMPs*. We continue to focus on expanding the PIDGraph to include additional nodes of identifiers for related outputs and resources.

EOSC Metadata Schema and Crosswalk Registry

MSCR

EOSC Metadata Schema and Crosswalk Registry

Maintenance of interoperability – PIDs enable versioning and provenance management

MSCR support publishing, discovery and access of metadata schemas and crosswalks and provide functions to operationalise metadata conversion by combining crosswalks. All of these objects will be issued PIDS; crosswalks, mappings and schemas and their versions. This allows to manage their provenance across different “physical resources”.

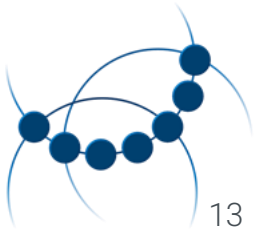
EOSC Data Type Registry

DTR

EOSC Data Type Registry

The DTR provides user friendly and machine actionable Interfaces for the registration and usage of Data Types and Kernel Information Profiles. Upon creation, each type is assigned a PID. The DTR can generate schemas corresponding to each type and use those to validate the content of PID records describing digital objects.

We are FAIRCORE4EOSC !



RAISE

eosc | RAISE
Research Analysis Identifier System

Promoting a transparent way of sharing and processing data

Provides the infrastructure for a distributed crowdsourced data processing system, moving from open data to open access data for processing. RAISE will provide the mechanism for sending the algorithm to the **dataset** instead of sending the data to the algorithm. The real value of open data for the research community is not to access them but to process them as conveniently as possible in order to reduce time-to-result and increase productivity.

https://raise-science.eu/

Gorka Epelde