FAIRCORE4EOSC
Developing EOSC-Core components to enable a FAIR EOSC ecosystem

Presentation title: FAIRCORE4EOSC - Extending the EOSC Platform to support FAIR

22 | 11 | 2023  Tommi Suominen, FAIRCORE4EOSC Coordinator at CSC – IT Center for Science
Slides by Tommi Suominen, Joonas Kesäniemi & Rumana Quazi

Funded by the European Union

faircore4eosc.eu
Twitter: @FAIRCORE4EOSC
LinkedIn: company/faircore4eosc
Youtube: FAIRCORE4EOSC
FAIRCORE4EOSC in a nutshell
- today at month 17/36

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.
Context

Enhancing FAIRness in the EOSC ecosystem

The European Open Science Cloud (EOSC) is an ecosystem of research data and related services that will enable and enhance seamless access to and reliable re-use of FAIR research objects (including data, publications, software, etc.).

The Strategic Research and Innovation Agenda (SRIA) for EOSC was created in 2021, as a roadmap for future development. 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).
Implementation Challenges (SRIA) addressed

**FAIRCORE4EOSC** develops 9 new EOSC CORE components to address gaps identified in the SRIA. Our concrete service development work furthers the realisation of the priorities highlighted in the SRIA, that are the Minimum Viable EOSC (MVE) and web of FAIR data.

- **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.
The 9 FAIRCORE4EOSC components supporting FAIR

**RDGraph (F, A)**

EOSC Research Discovery Graph

EOSC Research Discovery Graph (RDGraph) is a flexible and federated EOSC search service across EOSC repositories that extends EOSC Research Catalogue.

**PIDGraph (F, A)**

EOSC PID Graph

Services for providing access to the PID Graph, which is made up of links and records gathered from persistent identifier (PID) authority data sources.

**MSCR (I, R)**

EOSC Metadata Schema and Crosswalk Registry

Support publishing, discovery and access of metadata schemas and crosswalks and provide functions to operationalise metadata conversion by combining crosswalks.
The 9 FAIRCORE4EOSC components supporting FAIR

**DTR (I, R)**
EOSC Data Type Registry

Provide user friendly and machine actionable Interfaces for the registration and usage of Data Types and Kernel Information Profiles.

**PIDMR (F, A)**
EOSC PID Meta Resolver

Provides users with a common interface to resolve different types of PIDs regardless of their originating system. The PIDMR either resolves to a given URI or provides Kernel Information Profiles if available.

**CAT (F, I)**
EOSC Compliance Assessment Toolkit

The Compliance Assessment Toolkit will support the EOSC PID policy with services to encode, record, and query compliance with the policy.
The EOSC RAiD will mint PIDs for research projects, which will allow authorised EOSC users and services to manage information about project-related participants, services, and outcomes.

Ensure the long-term preservation of research software in different disciplines. APIs and connectors will be developed to interconnect research outputs infrastructures with the Software Heritage universal source code archive, using the CodeMeta standard, and the Software Heritage intrinsic identifiers (SWHID).

Equip EOSC with a mirror of the Software Heritage universal source code archive. In order to prevent information loss, a mirror of Software Heritage will be established by GRNET to serve the EOSC community and will be updated regularly to follow the growth of the universal source code archive.
Case Studies

**Social Sciences and Humanities**
This case-study will focus on improving the discoverability of CLARIN data through the integration of the Digital Object Gateway (DOG), a crucial component for the interoperability of the CLARIN infrastructure, Language Resource Switchboard and Virtual Collection Registry tools.

**Climate Change**
ENES supports climate modellers in their work, in particular in the area of data management. In this case study we demonstrate how the developed EOSC-Core components can improve the discoverability and re-use of research results from the ENES community.

**Mathematics**
zBMath Open & swMATH projects aggregate significant scientific advances in mathematics and related disciplines supporting researchers in finding relevant publications and data. The case study will increase the discoverability of the zBMath Open and swMATH data and services in the mathematical and EOSC community.

**European Integration of National-Level Services**
The case study will showcase how the developed components can enrich the content of the national research information systems displaying international connections to research objects and improve their interoperability.

**EOSC Service Providers**
The case study aims to meet domain-specific requirements of research communities for common data services that improve discovery, access and reusability of research data. Leveraging the EUDAT services, the case study will act as a rule model for other service providers to increase the adoption of the developed components.
FC4E Implementation timeline

- Phase 1: Technical specifications
- Phase 2: Beta release
- Phase 3: Cross-Integration BETA release

Collection of requirements and technical specifications (initially M1 - M8 with requirements continually being collected until M24)

Demonstrators (M4 - M24)

EOSC CAT, RDGraph, PIDGraph, PIDMR, RAID, MSCR, DTR, SWHM, RSAC

Case studies definition

(Early) testing

Final Integration

EOSC-Core enriched with the new FAIRCORE4EOSC components

Phase 4: Production release & final documentation

Case studies ready for showcase and operation

November 2023
Metadata schema and crosswalk registry in integration with the data type registry
Main features

- Register source schema
- Register target schema
- Register Crosswalk
- Create Crosswalk editor
- Use Crosswalk
- Create Transformation engine
- Create target data
MSCR Content

Schema description formats

- CSV, JSON Schema, XML Schema
- SKOS
- OWL
- SHACL
- Any file (e.g. PDF) – cannot be part of a crosswalk

Crosswalk

- Any file (e.g. XSLT) or reference to implementation
- Created with the Crosswalk editor (native format)
- Imported SSSOM (converted into internal format)

Hosted or referenced

- Hosted content is maintained with the MSCR (repository)
- Referenced content is maintained somewhere else (e.g. Github)

Versioned
MSCR Home Page

Metadata Schema and Crosswalk Registry

MSCR allows registered users and communities to create, register and maintain schemas and crosswalks with FDs. The published content can be searched, browsed and downloaded without restrictions. MSCR provides an API to facilitate the transformation of data from one scheme to another via registered crosswalks.

**Schemes**

<table>
<thead>
<tr>
<th>Name</th>
<th>Namespace</th>
<th>Status</th>
<th>Revision</th>
<th>PID</th>
</tr>
</thead>
<tbody>
<tr>
<td>MARC to MODS 1</td>
<td><a href="http://example.org/space/3.0">http://example.org/space/3.0</a></td>
<td>DRAFT</td>
<td>1.1(+5 other)</td>
<td>1</td>
</tr>
<tr>
<td>DTC to EML</td>
<td><a href="http://example.org/space/3.0">http://example.org/space/3.0</a></td>
<td>DRAFT</td>
<td>1.1(+5 other)</td>
<td>2</td>
</tr>
<tr>
<td>Dataset to Dublin Core</td>
<td><a href="http://example.org/space/3.0">http://example.org/space/3.0</a></td>
<td>INCOMPLETE</td>
<td>1.1(+5 other)</td>
<td>3</td>
</tr>
</tbody>
</table>

**Crosswalks**

<table>
<thead>
<tr>
<th>Name</th>
<th>Namespace</th>
<th>Status</th>
<th>Revision</th>
<th>PID</th>
</tr>
</thead>
<tbody>
<tr>
<td>MARC to MODS 1</td>
<td><a href="http://example.org/space/3.0">http://example.org/space/3.0</a></td>
<td>DRAFT</td>
<td>1.1(+5 other)</td>
<td>1</td>
</tr>
<tr>
<td>DTC to EML</td>
<td><a href="http://example.org/space/3.0">http://example.org/space/3.0</a></td>
<td>DRAFT</td>
<td>1.1(+5 other)</td>
<td>2</td>
</tr>
<tr>
<td>Dataset to Dublin Core</td>
<td><a href="http://example.org/space/3.0">http://example.org/space/3.0</a></td>
<td>INCOMPLETE</td>
<td>1.1(+5 other)</td>
<td>3</td>
</tr>
</tbody>
</table>
Authorized user can login and create a personal account
  • For the potentially messy content

Groups are be managed by system level admins
  • Process for requesting a new group required
  • Roles
    • Group admin
    • Group member
  • Can be used for more authoritative content

API keys can be generated for programmatic use
MSCR Home Page

Metadata Schema and Crosswalk Registry

MSCR allows registered users and communities to create, register and version schemas and crosswalks. MSCR provides an easy-to-use interface for managing metadata schemas and crosswalks. Schemas and crosswalks can be created and modified online. Users can specify the name of the schema and the namespace for the schema. The namespace is used to identify the schema within the MSCR registry.

### Schemas

<table>
<thead>
<tr>
<th>Name</th>
<th>Namespace</th>
</tr>
</thead>
<tbody>
<tr>
<td>MARC to MODS 1</td>
<td><a href="http://example.org/space/1.0/marc-to-mods-1">http://example.org/space/1.0/marc-to-mods-1</a></td>
</tr>
<tr>
<td>DTC to EML</td>
<td><a href="http://example.org/space/1.0/dtc-to-eml">http://example.org/space/1.0/dtc-to-eml</a></td>
</tr>
<tr>
<td>Datacite to Dublin Core</td>
<td><a href="http://example.org/space/1.0/datacite-to-dublin-core">http://example.org/space/1.0/datacite-to-dublin-core</a></td>
</tr>
</tbody>
</table>

### Crosswalks

<table>
<thead>
<tr>
<th>Name</th>
<th>Namespace</th>
</tr>
</thead>
<tbody>
<tr>
<td>MARC to MODS 1</td>
<td><a href="http://example.org/space/1.0/marc-to-mods-1">http://example.org/space/1.0/marc-to-mods-1</a></td>
</tr>
<tr>
<td>DTC to EML</td>
<td><a href="http://example.org/space/1.0/dtc-to-eml">http://example.org/space/1.0/dtc-to-eml</a></td>
</tr>
<tr>
<td>Datacite to Dublin Core</td>
<td><a href="http://example.org/space/1.0/datacite-to-dublin-core">http://example.org/space/1.0/datacite-to-dublin-core</a></td>
</tr>
</tbody>
</table>

### Add New Crosswalk

#### Add New Crosswalk Description

**Select source schema**

- Default schema

**Select target schema**

- Test schema 2

**Select a schema to see properties.**

**Name**

- Datacite to Dublin Core

**Description**

- This crosswalk will create suitable mapping

**Format**

- XSLT

**Information description languages**

Select information description languages in which the content of the data model is described.

- Select languages for information description

[English]
Edit operations

⚠️ Type mismatch between source and target

Source: AffiliationIdentifier
Source type: String
Description: Example value: xx-yyyy-yyyy
Source filter
   Add Filter
Source operation
   Search or select

Mapping operation

Target: AffiliationIdentifier
Target type: Int
Description: Example value: xxxxxxxx
Target filter
   Add Filter
Target operation
   Search or select

Save  Cancel
Mappings in MSCR

• Semantic mappings
  • Social security number → exact match → Personal identification number
  • Description → broader -→ Short description
  • <temperature> → is about → <http://vocabs/measures/temperature>

• Structural mappings
  • <coordinate>12.232,29.323</coordinate> → (exact match) →
    <coordinates><coordinate><lat>12.232</lat><long>29.323</long></coordinate></coordinates>

• Value mappings
  • “DRAFT” → “Not_ready_yet”
  • “DRAFT → <http://standards.org/status/draft>
  • Kesäniemi, Joonas → Joonas Kesäniemi
  • 20C → 36.8F
Why FAIR?

Findable
- Registered content is assigned a PID (Handle) as needed
  - pre-existing PIDs for referenced content are preserved and promoted
- mappings utilize part identifiers by default
  - considered highly contextual, still reusable for example for semantic mappings
- Support for harvesting
  - microformats?

Accessible
- Resolvable PIDs (ERIC handles)
- Published content is persistent by default
  - Each version has its own PID
- All metadata is publicly accessible
  - visibility/searchability of draft content can be toggled
Why FAIR?

Interoperable
- Final metadata schema of crosswalks schemas is still under development
- Support for DTR defined data types
- Crosswalk import/export format
  - SSSOM import and export for crosswalks/mappings that fit to specification
  - Default format will be custom format (JSON)
  - RML export for CSV/JSON/XML to RDF crosswalks (?)
- Registered schemas can always be downloaded in their original format
  - Internal format (SHACL) is also available

Reusable
- Licensing is part of the content’s metadata schema
- Published content cannot be changed
  - Metadata can always be modified
- Provenance information on a version level
- Schemas in common standards
  - XSD, JSON schema, OWL, SHACL, SKOS
Hands-on tutorials for adopting the FC4E services

Are planned to be organized as a workshop at the CRIS2024 Conference, which takes place 15-17 May, 2024 at the Technical University of Vienna (TU Wien)

The idea is to show interested adopters how to get started with using these services.
We are FAIRCORE4EOSC