D8.4
Interoperable cataloguing and metadata harmonisation for environmental RIs: prototype

WORK PACKAGE 8 – DATA CURATION AND CATALOGUING

LEADING BENEFICIARY: NERC

Author(s):	Beneficiary/Institution
Erwann Quimbert	Ifremer
Keith G Jeffery	NERC
Claudia Martens	DKRZ
Damien Boulanger	CNRS
Thierry Carval	Ifremer
Margareta Hellström	Lund University
Harry Lankreijer	Lund University
Johannes Peterseil	LTER
Christian Pichot	INRA
Zhiming Zhao	University of Amsterdam

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654182

Accepted by: Keith G Jeffery (WP 8 leader)

Deliverable type: DEMONSTRATOR

Dissemination level: PUBLIC

Deliverable due date: 31.10.2018/M42

Actual Date of Submission: 06.11.2018/M42

ABSTRACT

This document describes the technical implementation for a catalogue system in the ENVRIplus framework. After a brief reminder on the general requirements detailed in D8.3, this report provides information concerning the architecture and design recommendations for the implementation of this catalogue system.

The main objective of this catalogue is to offer a unified discovery service allowing cross-disciplinary search and access to data collections coming from Research Infrastructures (RIs). This catalogue focuses on metadata with a coarse level of granularity. It was decided to present metadata representing different types of dataset series. Only metadata for so called flagship products (as defined by each community) are covered by the scope of this catalogue. The data collections remain within each RI. For RIs, the aim is to improve the visibility of their results beyond their traditional user communities.

Two candidate standards have been recommended for the implementation of RI catalogues in D8.3: CKAN (e.g., EUDAT/B2FIND) and CERIF (e.g., in EPOS). These two standards have been selected to implement this so called 'Flagship catalogue'.

Flagship catalogue is a demonstrator of the capabilities of ENVRIplus to enable cross-disciplinary discoverability of services provided by RI, organized in a cluster. The update and operation of this demonstrator is not foreseen after the duration of the ENVRIplus project, although both B2FIND supported by EUDAT and CERIF supported by euroCRIS and used by (among others) EPOS continue beyond the project.

The document is organized into three sections:

- A short summary of functional requirements and priorities defined in D8.3
- Potential flagship-catalogue architectures
- Flagship catalogue implementations with detailed components and interfaces description

Project internal reviewer(s):

Project internal reviewer(s):	**Beneficiary/Institution**
Robert Huber	UniHB
Malcolm Atkinson	University of Edinburgh

Document history:

Date	**Version**
19.09.2018	Outline for comments
19.10.2018	Draft for comments sent to Keith Jeffery and Zhiming Zhao

21.10.2018	Keith Jeffery and Zhiming Zhao reviews integrated.
28.10.2018	Report sent to Task 8.2 partners
29.10.2018	Report sent to reviewers
29.10.2018	Review from Robert Huber
29.10.2018	Review from Christian Pichot
05.11.2018	Review from Malcolm Atkinson
06.11.2018	Final check and to project office

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the authors (Erwann Quimbert, Erwann.quimbert@ifremer.fr or Keith G Jeffery, Keith.Jeffery@keithgjefferyconsultants.co.uk)

TERMINOLOGY

A complete project glossary is provided online here:

https://envriplus.manageprojects.com/s/text-documents/LFCMXHHCwS5hh

PROJECT SUMMARY

ENVRIplus is a Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, projects and networks together with technical specialist partners to create a more coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe. It is driven by three overarching goals: 1) promoting cross-fertilization between infrastructures, 2) implementing innovative concepts and devices across RIs, and 3) facilitating research and innovation in the field of environment for an increasing number of users outside the RIs.

ENVRIplus aligns its activities to a core strategic plan where sharing multi-disciplinary expertise will be most effective. The project aims to improve Earth observation monitoring systems and strategies, including actions to improve harmonization and innovation, and generate common solutions to many shared information technology and data related challenges. It also seeks to harmonize policies for access and provide strategies for knowledge transfer amongst RIs. ENVRIplus develops guidelines to enhance transdisciplinary use of data and data-products supported by applied use-cases involving RIs from different domains. The project coordinates actions to improve communication and cooperation, addressing Environmental RIs at all levels, from management to end-users, implementing RI-staff exchange programs, generating material for RI personnel, and proposing common strategic developments and actions for enhancing services to users and evaluating the socio-economic impacts.

ENVRIplus is expected to facilitate structuration and improve quality of services offered both within single RIs and at the pan-RI level. It promotes efficient and multi-disciplinary research offering new opportunities to users, new tools to RI managers and new communication strategies for environmental RI communities. The resulting solutions, services and other project outcomes

are made available to all environmental RI initiatives, thus contributing to the development of a coherent European RI ecosystem.

TABLE OF CONTENTS

I. INTRODUCTION

Data catalogue is a key service in the data management of ENVRIplus. This report describes the work undertaken in the task 8.2 and the current development status from the initial DoA (Description of Activity) on cataloguing (D8.3). As noted in D8.3, CERIF and CKAN are both candidate standards recommended for prototyping an ENVRIplus community catalogue for RI flagship data products. This activity (T8.2 within WP8) was led by Ifremer [1]with contributions from all task 8.2 partners.

To streamline the implementation of this flagship catalogue, it was decided to start with the EUDAT/B2FIND demonstrator. The demonstrator on CERIF has also been developed jointly with EPOS[2] and other relevant projects, e.g., VRE4EIC[3].

The goal of the Flagship catalogue is to expose and highlight products that best illustrate the content of RIs catalogues. This demonstrator also aims to provide a better overview to users of existing catalogues and resources, mostly data, indexed by these catalogues.

Progress since D8.3

In D8.3, two approaches were introduced for the flagship catalogue development:

- Top-Down approach

This approach aims at showcasing the outcome of the RI so that they reach new inter-disciplinary and data science usages. The homogeneous and qualified descriptions provided in a single seamless framework is a tool for stakeholders and decision makers to oversee and evaluate the outcome and complementarity of RI data products.

- Bottom-up approach

This approach was presented in D8.3 to highlight the possibility to create a fine-grained catalogue which would facilitate the deployment of the Theme 2 solutions. Nevertheless, the first conclusion revealed that it was really complicated to achieve this kind of catalogue in the context of ENVRIplus. The findings of D8.3 report was to conclude that in view of the maturity of RIs regarding the management of their data services, the level of standardization in this field (ISO19XXX series[4], Dublin-Core[5]…) and availability of homogeneous technical solutions partially adopted by the RIs (Geonetwork[6], CKAN[7]…), it was only feasible to develop an integrated service for flagship data product discovery.

[1] http://www.ifremer.fr
[2] https://www.epos-ip.org/
[3] https://www.vre4eic.eu/
[4] http://www.isotc211.org/
[5] http://dublincore.org/
[6] https://geonetwork-opensource.org/
[7] https://ckan.org/

To streamline the implementation of this catalogue, it was decided to focus on data products which consist of a description of a type of products the RIs want to first communicate about.

The task 8.2 has designed and developed a flagship data products catalogue which provide this top-down view and entry point for hierarchical discovery of the ENVRIplus RIs' resources.

For a first version, the following RIs have been targeted as first priority to have their resources described in ENVRIplus catalogue system.

- AnaEE[8] (Analysis and Experimentation on Ecosystems) focuses on providing innovative and integrated experimentation services for research on continental ecosystems.
- Euro-ARGO[9] is the European contribution to the ARGO program. Argo is a global array of 3,800 free-drifting profiling floats that measures the temperature and salinity of the upper 2000 m of the ocean.
- EMBRC[10] is a pan-European Research Infrastructure for marine biology and ecology research.
- EPOS[11] (European Plate Observing System) is a long-term plan to facilitate integrated use of data, data products, and facilities from distributed research infrastructures for solid Earth science in Europe.
- IAGOS[12] (In-Service Aircraft for a Global Observing System) is a European Research Infrastructure for global observations of atmospheric composition using commercial aircraft.
- ICOS[13] is a pan-European research infrastructure for quantifying and understanding the greenhouse gas balance of Europe and its neighbouring regions.
- LTER[14] (Long Term Ecological Research) is an essential component of world-wide efforts to better understand ecosystems.
- SeaDataNet[15] is a pan-European infrastructure to ease the access to marine data measured by the countries bordering the European seas.
- Actris[16] is the European Research Infrastructure for the observation of Aerosol, Clouds, and Trace gases.

As indicated in Introduction the main effort started with B2FIND.

[8] https://www.anaee.com/
[9] https://www.euro-argo.eu/
[10] http://www.embrc.eu/
[11] https://www.epos-ip.org/
[12] http://www.iagos-data.fr/
[13] The general Webpage for ICOS RI is http://www.icos-ri-eu, the webpage for the ICOS data is http://www.icos-cp.eu
[14] http://www.lter-europe.net/
[15] https://www.seadatanet.org/
[16] https://www.actris.eu/

Requirements

In D8.3 three kinds of users were identified for this flagship catalogue:

- Users outside an RI, researching data-driven science

- Users inside an RI, such as data managers, coordinators, and operators as well as the domain scientists

- and the stakeholders, decision makers and funders of the RIs who need to have a broad picture of the RI resources in the European landscape to control their efficiency and complementarity.

The catalogue requires the following functions:

- Enable discovery of resources via an English interface. Multilingualism is not prioritized in the initial version.
- Enable curation by providing information regarding the preservation of the data (storage format, host).
- Provenance of the data products should also be covered (see D8.6)
- Evaluation of descriptions will be useful for RIs to evaluate the quality of their contributions (e.g. missing fields, broken reference links…).

The requirements on product description were defined by adopting those of the RDA metadata interest group[17]. This list is kept simple on purpose.

II. Flagship catalogue architectures

In this section, the architecture of the two systems designed for flagship catalogue is described.

II.A B2FIND

B2FIND[18] is a discovery service for research data distributed within EOSC-hub [19]and beyond. It is a basic service of the pan-European collective data infrastructure EUDAT CDI and will become the central indexing tool for EOSC-hub[20]. For this a comprehensive joint metadata catalogue was built up that spans a wide spread scope of diverse metadata collected from heterogeneous sources – from Climate Research to Social Science and Particle Physics to Economics.

Therefore, different metadata formats, schemas and standards are homogenized on the B2FIND metadata schema, allowing users to search and find research data across scientific disciplines and research areas - thus enabling an interdisciplinary perspective.

B2FIND provides powerful search functionalities (Figure 1): metadata records can be found with a free-text-search, results may be narrowed down using different facets (e.g. geospatial or timeline search). Referencing data identifiers will either link to a landing page or to the resource itself.

[17] https://rd-alliance.org/groups/metadata-ig.html
[18] https://eudat.eu/services/b2find
[19] http://www.eosc-hub.eu/
[20] https://eosc-hub.eu/

B2FIND also endorses data providers in the creation, dissemination and quality assurance of their metadata. In addition, B2FIND offers detailed training that guides interested parties through the entire life cycle of metadata, thus being a supporting service for research communities concerning technical and content related issues.

B2FIND allows users to:

- Find collections of scientific data quickly and easily, irrespective of the data's origin, discipline or community
- Get quick overviews of available data
- Browse through collections using standardized facets

Communities publishing metadata in B2FIND benefit from improved visibility and searchability of their research data within an interdisciplinary and pan-European scope.

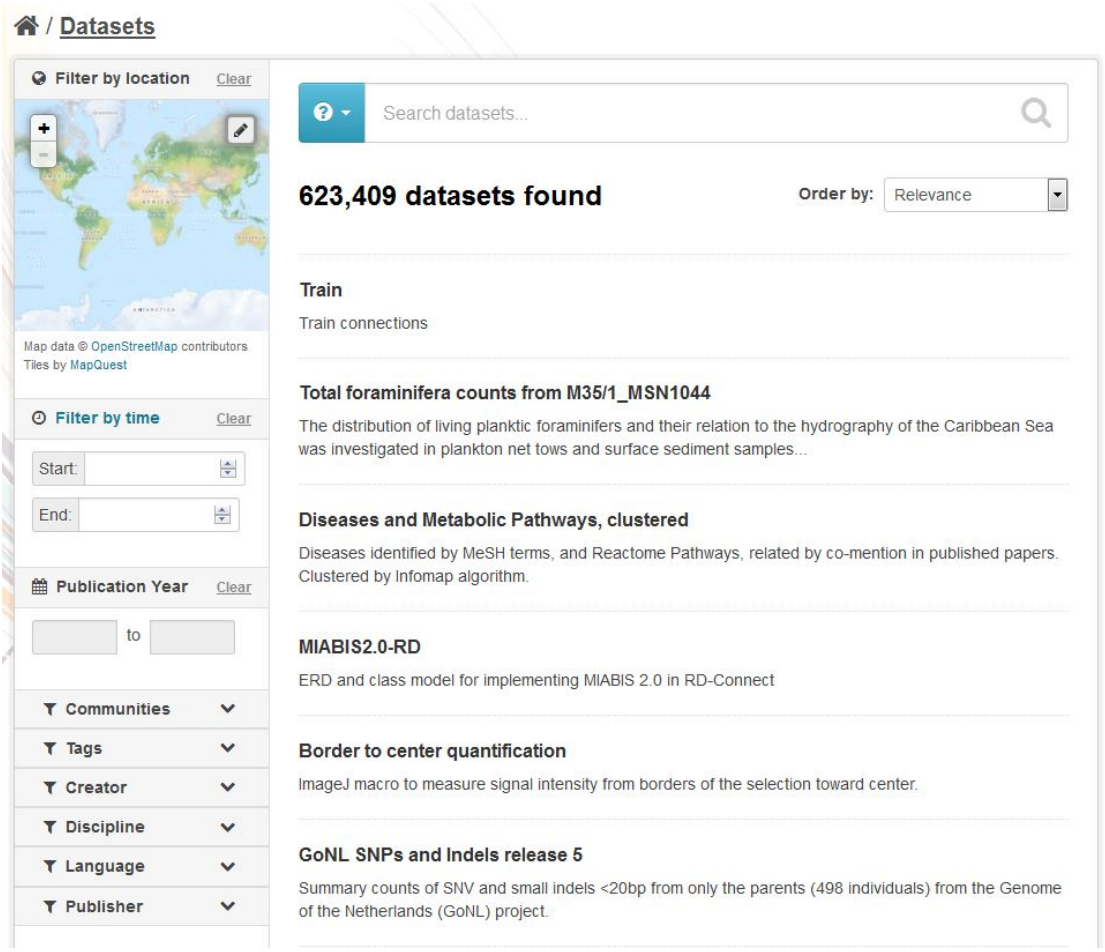

Figure 1 – Metadata search results page

How does it work?

The implementation of B2FIND was guided by the FAIR principles[21], including the establishment of common standards and guidelines for data providers as well as end users. Hereby, a close cooperation and coordination with scientific communities, Research Infrastructures and other initiatives dealing with metadata standardisation (e.g. OpenAire Advance[22], several RDA interest and working groups concerning metadata as e.g. RDA-DDPIG [23] and the EOSCpilot 'Data Interoperability' [24]) is essential in order to support standards that are both reasonable for community specific needs and enhanced exchangeability.

B2FIND is able to harvest other catalogues or systems by using the following protocols or formats: OAI-PMH, JSON API, CSW and DCAT (natively used by CKAN).

The modular metadata ingestion workflow of B2FIND consists of three steps: harvesting metadata, mapping to B2FIND schema (including format conversion and semantic mapping) and indexing and uploading them into the web portal. In particular standardized interfaces and protocols are used that aim for interoperable and easy to use service:

- Harvesting: Preferably B2FIND uses the standardized protocol OAI-PMH[25] to harvest metadata from data providers. Additionally other API's like OGC-CSW[26], SparQL[27] or JSON-API[28] are supported by B2FIND.
- Mapping: The community specific metadata records, formatted in domain specific formats and schemas, are mapped by B2FIND to a unique and generic target metadata schema based on the DataCite4.1 metadata schema. The mapping of the metadata elements and facets is based hereby on standardized vocabularies and ontologies (e.g. the field 'Language' is mapped on the ISO 639 library[29] and research 'Disciplines' are mapped on a standardized closed vocabulary). B2FIND supports different community specific metadata schemas, such as DataCite, DublinCore, MarcXML, DDI, CMDI, ISO19115/19139 and INSPIRE.
- Uploading and Indexing: The search portal and the graphical user interface is based on the open source portal software CKAN[30], which comes with the Apache Lucene SOLR[31] Servlet allowing indexing of the mapped JSON records and performant facetted search.

[21] https://www.force11.org/group/fairgroup/fairprinciples
[22] https://www.openaire.eu/advance/
[23] https://www.rd-alliance.org/groups/data-discovery-paradigms-ig
[24] https://www.eoscpilot.eu/events/eoscpilot-62-data-interoperability-meeting-technical-meeting
[25] https://www.openarchives.org/pmh/
[26] http://www.opengeospatial.org/standards/cat
[27] https://www.w3.org/TR/rdf-sparql-query/
[28] https://jsonapi.org/
[29] https://www.loc.gov/standards/iso639-2/php/code_list.php
[30] https://ckan.org/
[31] http://lucene.apache.org/solr/

CKAN was created by the Open Knowledge Foundation (OKFN[32]) in the United Kingdom. CKAN is a powerful data management system that makes data accessible – by providing tools to streamline publishing, sharing, finding and using data.

Metadata is harvested from metadata providers who decide which metadata format and flavour is made available for B2FIND. A sophisticated framework ensures that metadata providers are harvested regularly to always display complete and up to date information. B2FIND provides an optimized translation from community metadata schema to standard facets in the B2FIND metadata catalogue.

The EUDAT B2FIND team provides support by setting up the necessary data provider services on the metadata provider site, if required. The semantic mapping of the harvested metadata uses an elaborate and flexible software stack. This allows clearly formulated and easy implementation of the mapping rules according to specific needs.

System architecture

B2FIND is one service of the B2 Service suite (Figure2).

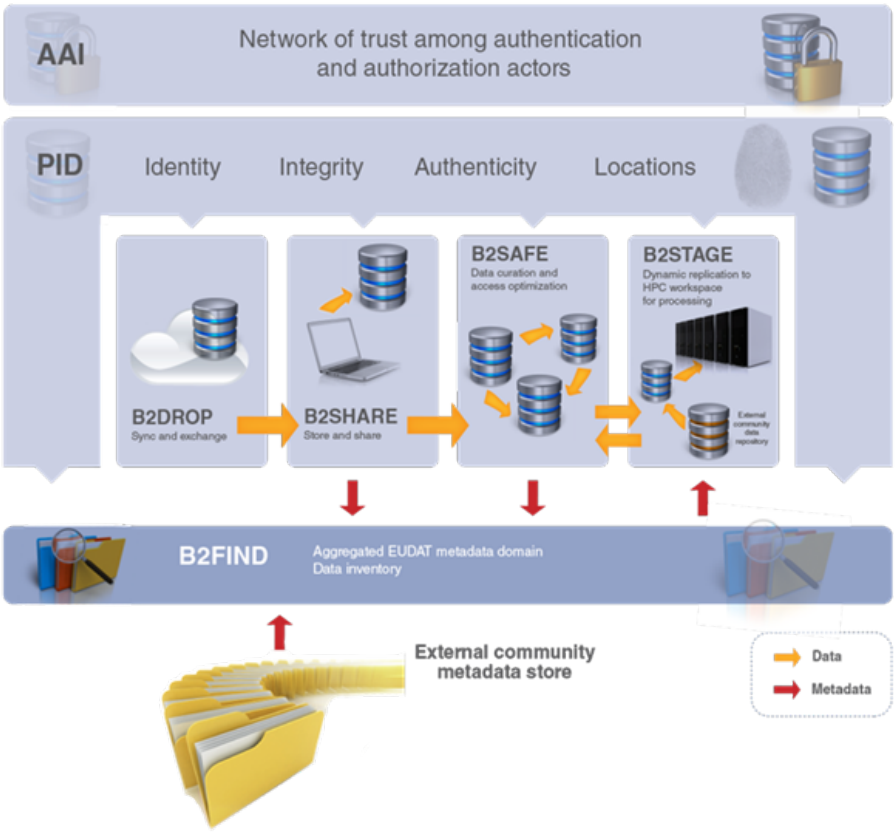

Figure 2. B2FIND in the B2 Service Suite

[32] https://okfn.org/

B2FIND has published its new version 2.4 during the Digital Infrastructures for Research Conference in Lisbon on October 11th 2018[33] which includes new design of the graphical user interface, improvements of search performance, enhancement of harvesting and mapping functionalities and extension of the metadata schema.

Some existing use cases

Currently[34] metadata records of fourteen Research Communities (Figure 3) are displayed whereas the ingestion process will proceed continuously for more than twenty communities to come.

Figure 3 – Communities using B2FIND

CLARIN - CMDI structure (Component Metadata Infrastructure)

CMDI was developed by members of CLARIN (Common Language Resources and Technology Infrastructure) to establish metadata exchange within the CLARIN infrastructure. It defines different levels of description (granularity): from complete corpora, subcorpora or corpus components, to individual resources, e.g. a recording of a dialogue (sound file + transcript).

For illustration we show the CMDI hierarchy schema as used in The Language Archive (see figure 4) and as the CMDI profile is applied in the project JASMIN (see figure 5).

[33] https://eudat.eu/news

[34] 2018-10-18, latest status to be seen here: http://b2find.eudat.eu/group

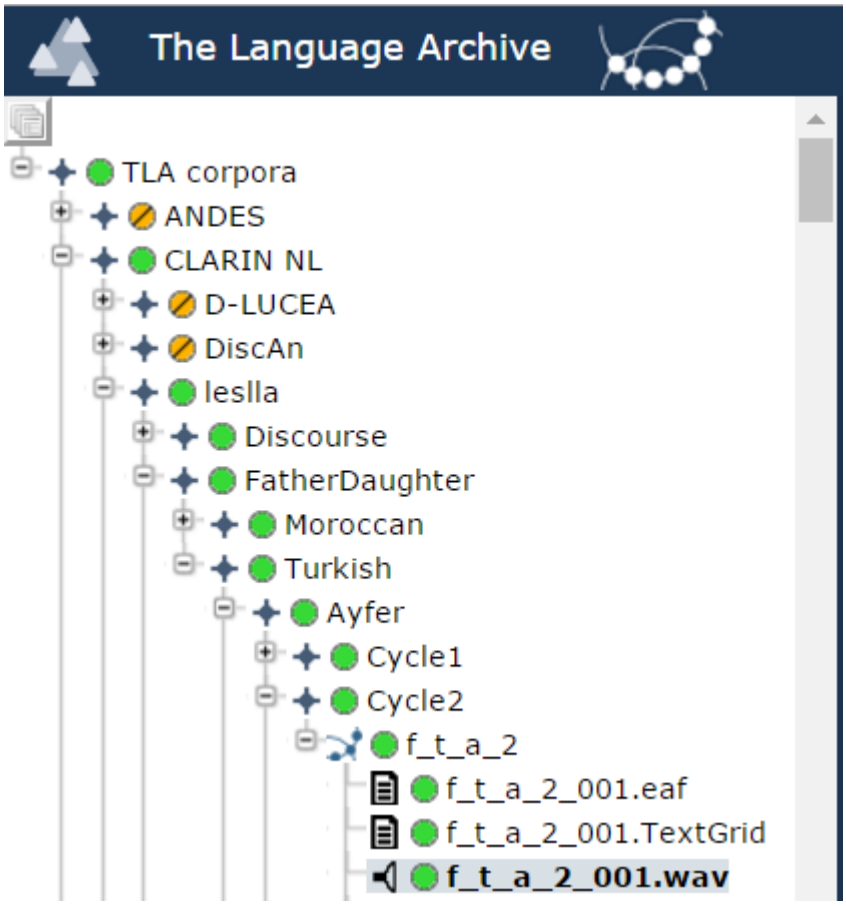

Figure 4: Hierarchy in the browser of The Language Archive (see at https://corpus1.mpi.nl/ds/asv/?4#

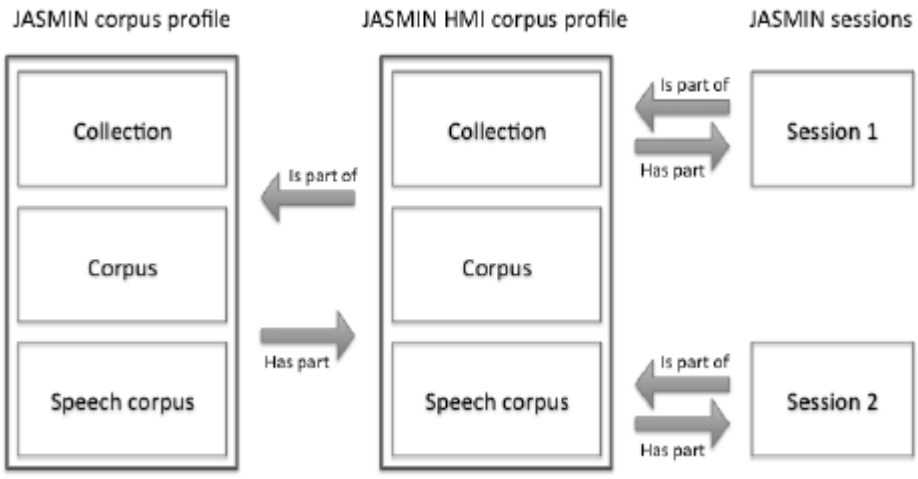

Figure 5: Hierarchy and granularity in JASMIN's CMDI profiles (see at https://www.researchgate.net/figure/228732331_fig2_Figure-2-Hierarchy-and-granularity-in-JASMIN-profiles)

DKRZ - LTA (Long Term Archive)

Data in the long-term archive of the DKRZ (DKRZ-LTA) is organized in multiple layers with a hierarchical structure. These layers are defined in the associated CERA2 model as Projects, Experiments, Dataset groups and Datasets. The relationships between these different layers are shown as a hierarchical tree structure in Figure 6.

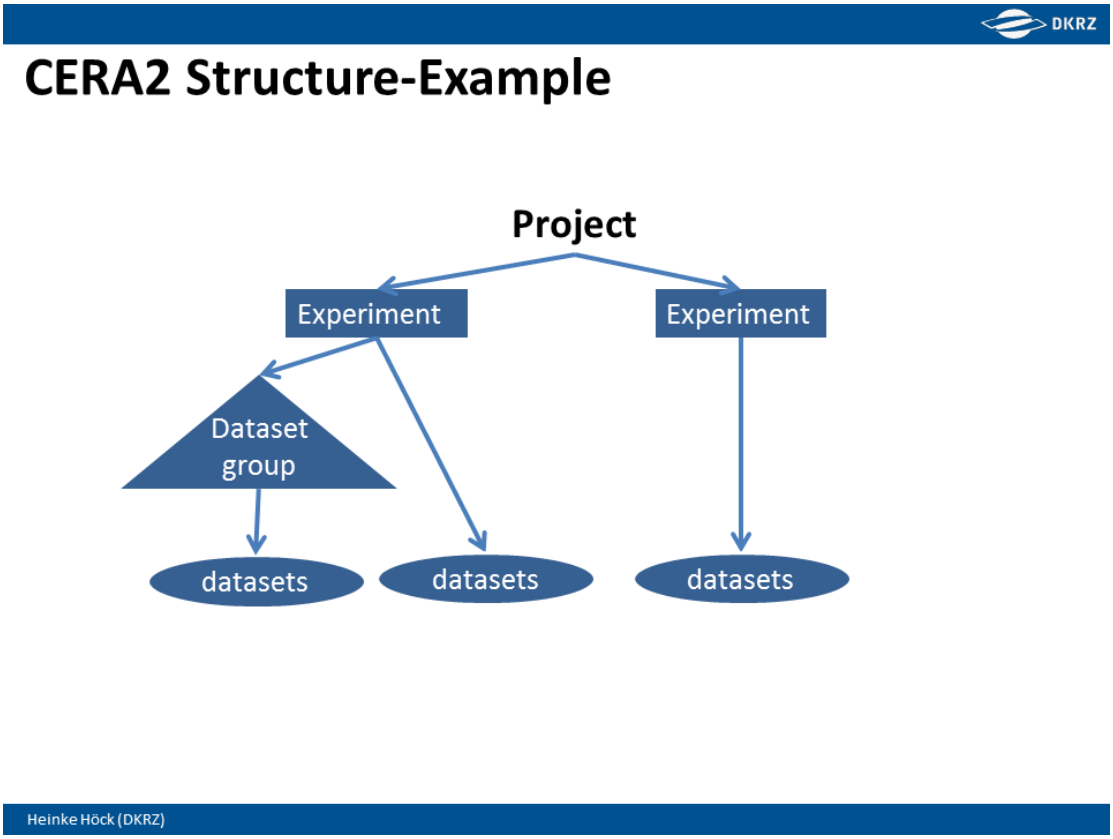

Figure 6: CERA2 Hierarchical Structure

Among other climate data, the DKRZ-LTA archive stores the experiments of the model intercomparison project CMIP5 1.This data and its associated metadata are kept in and accessible via the CERA database. B2FIND harvests metadata records corresponding to CMIP5 data collections from the CERA database, but only records those for which a DOI is assigned.

This means that the CMIP5 experiments are related to entities of aggregation level 3 in B2FIND, because they are citable aggregations of datasets and each metadata record refers, via a DOI, exactly to the landing page of the associated experiment in the DKRZ-LTA database CERA. There a user can browse through the datasets of the experiment and download the datasets they are interested in.

These examples illustrate the importance of supplying sufficient information about the structure and granularity of the research data to B2FIND.

Life+Respira

Life+Respira is a project aimed at showing that through the use of new technologies, along with urban planning and management, like promoting sustainable mobility and bicycle use, it is possible to improve air quality and reduce air pollution. The project has the added benefits included in socially engaging members of the community; some of the actions will be based on the help provided by a team of volunteer cyclists. This way citizens will become the driving force and main

beneficiaries of the results. All the steps proposed in the project are directly related to the improvement of air quality, urban environment and health.
This community uses several EUDAT services, such as B2DROP, B2SHARE and B2FIND. Data collections are published in B2SHARE where they get a unique PID. As B2FIND is harvesting B2SHARE incrementally those records are converted from MarcXML resp. XML to JSON files, mapped onto the B2FIND schema and uploaded to the web portal.

II.B CERIF

CERIF (Common European Research Information Format) is an EU Recommendation to Member States for research information. In 2000 the European Commission requested euroCRIS to maintain, develop and promote CERIF as a standard. It is a data model (Figure 7) based on EERT (extended entity-relationship modelling with temporal aspects).

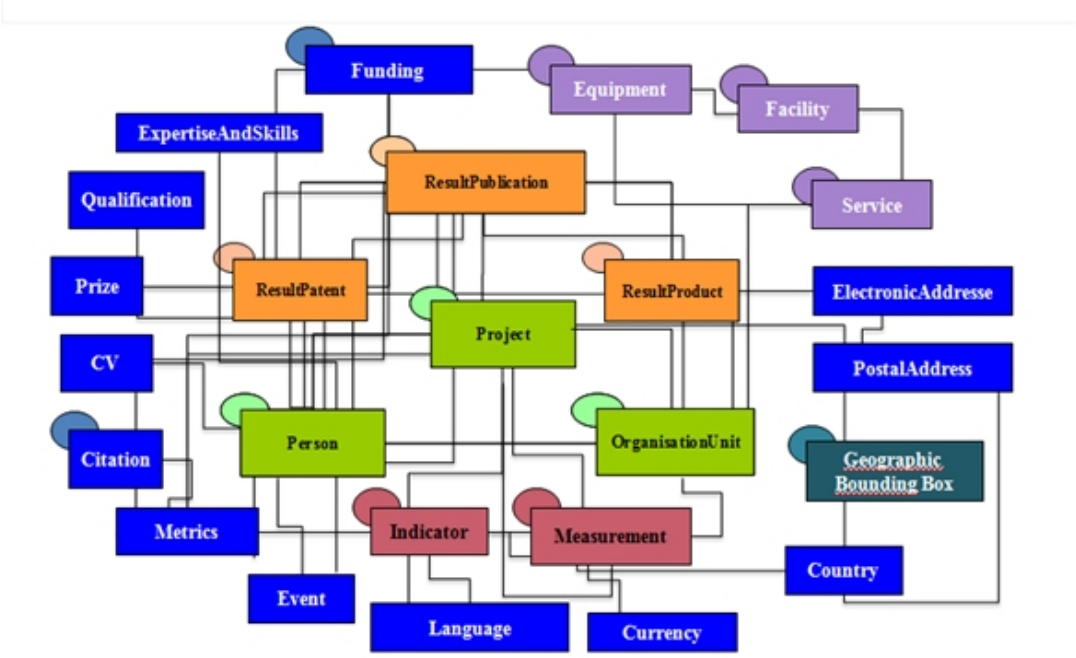

Figure 7. CERIF Data model

How does it work?

Although the model can be implemented in many ways (including object-oriented, logic programming and triplestores), most often it is implemented as a relational database but with a particular approach thus ensuring referential and functional integrity. CERIF has the concept of base entities representing real-world objects of interest and characterized by attributes. Examples are project, organization, research product (such as dataset, software), equipment and so on. The base entities are linked with relationship entities which describe the relationships between the base entities with a role (such as owner, manager, author) and date time start and end so giving the temporal span of the relationship. In this way versioning and provenance are 'built-in'.

CERIF also has a semantic layer (ontologies). Using the same base entity/relationship entity structure it is possible to define relationships between (multilingual) terms in different ontologies. The terms are used not only in the 'role' attribute of linking relations (e.g. owner, manager, author) but also to manage controlled lists of attribute values (e.g. ISO country codes). CERIF provides for multiple classification schemes to be used – and related to each other.

Mappings have been done from many common metadata standards (DC, DCAT, ISO19115/139, eGMS, DDI, CKAN(RDF), RIOXX…) to/from CERIF, emphasizing its richness and flexibility.

Some existing use cases

EPOS uses CERIF for its catalog because of the richness for discovery, contextualisation and action and because of the built-in versioning and provenance, important for both curation and contextualisation. The architecture of the software associated with the catalog (ICS: Integrated Core Services) is based on microservices (**Figure 8**).

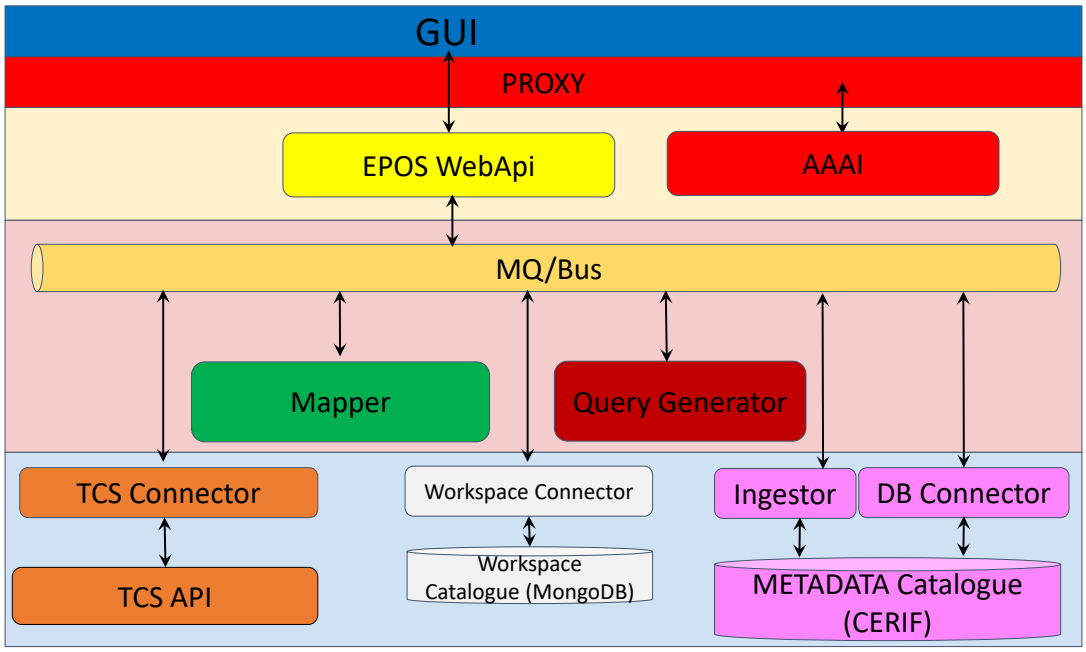

Figure 8 EPOS ICS Architecture

The implementation uses Postgres as the RDBMS and has been demonstrated on numerous occasions (**Figure 10Figure**). A mechanism for harvesting metadata from the various domain groups of EPOS (TCS: Thematic Core Services) and converting from their individual metadata schemes to CERIF has been implemented including an intermediate stage using EPOS-DCAT-AP (**Figure 9**).

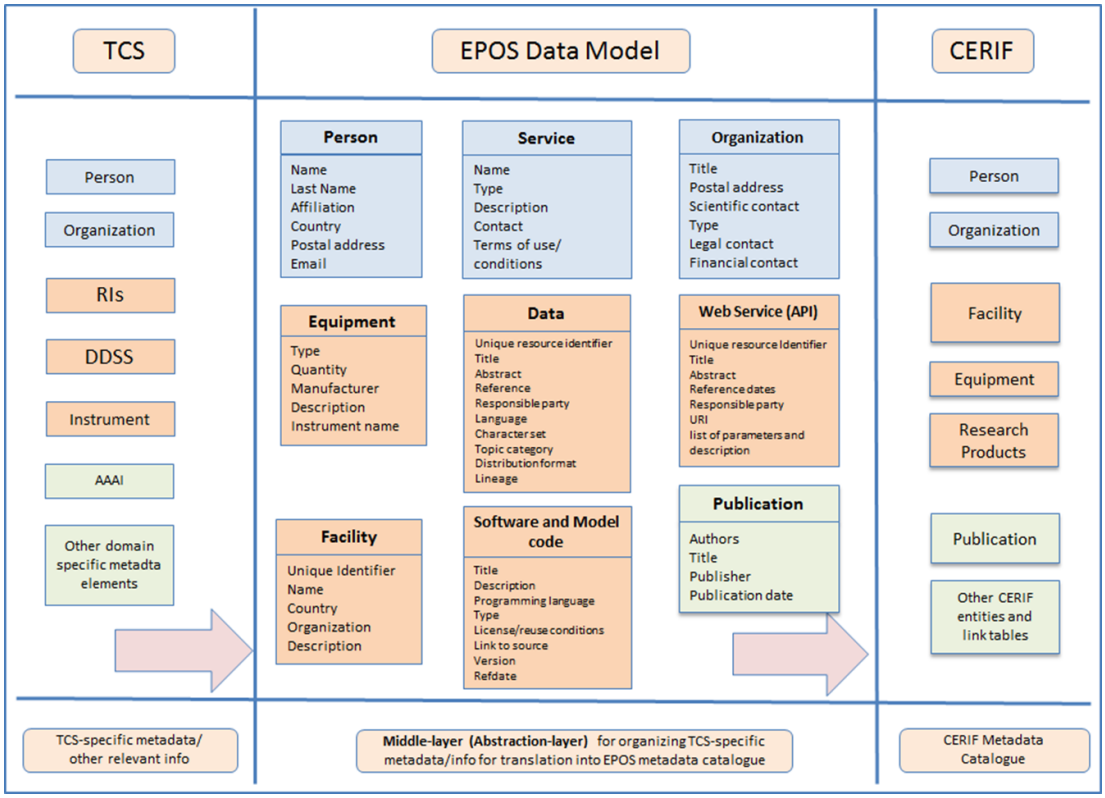

Figure 9: EPOS Metadata Harvesting Architecture

CERIF thus provides EPOS users with a homogeneous view over heterogeneous assets allowing cross-disciplinary research as well as within-domain research.

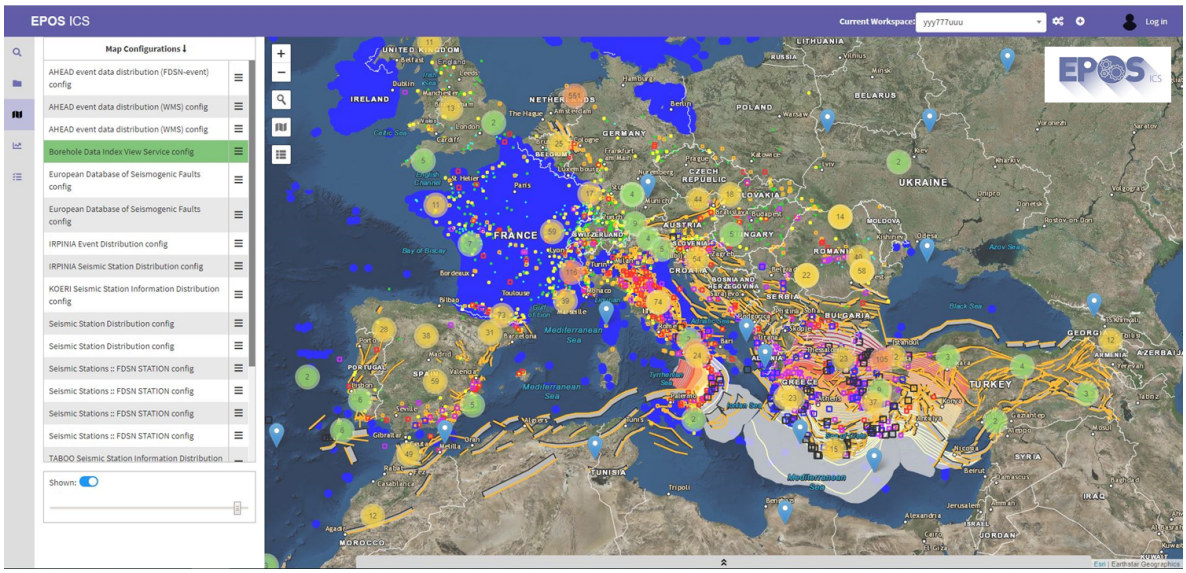

Figure 10: EPOS User interface

III. Flagship catalogue implementations

III.A B2FIND

Since the beginning of the project, an important integration work has been carried out with the support of the EUDAT team (B2FIND is configured and operated by DKRZ).

Introduction

During the implementation of the Flagship Catalogue several challenges occurred that had to be solved in close cooperation between RIs and EUDAT-B2FIND development team. One main issue is the granularity of the described products as B2FIND is meant to be an entry point for research data collections that should be found across scientific areas and one prerequisite has been to represent data products from ENVRIplus Communities rather than all individual research data records. Another big challenge was the technical implementation of a B2FIND group "ENVRIplus" that contains several metadata records from different ENVRIplus RIs.

Integration processes

For this demonstrator, RIs have followed the integration processes suggested by B2FIND (see Figure 11)

Each RIs should provide EUDAT with a document explaining how the catalogue from RI can be harvested by B2FIND. B2FIND metadata ingestion workflow is described in Figure 11.

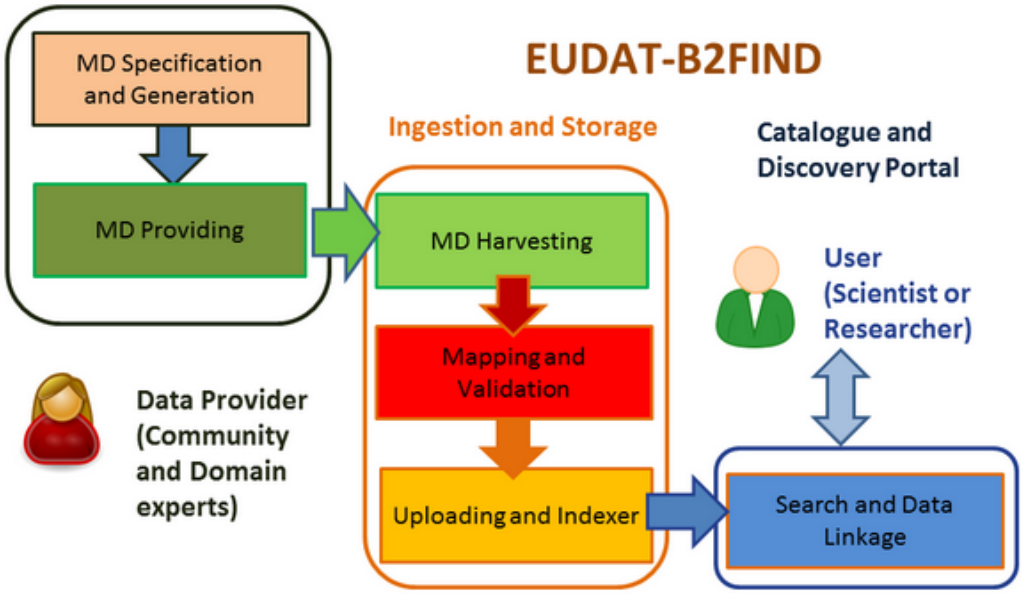

Figure 11. The B2FIND metadata ingestion workflow

B2FIND Ingestion Workflow comprises three main steps:

1. **Harvest**[35] community specific metadata from data provider (RIs)

There are only a few mandatory requirements for the metadata that is to be published on B2FIND.

As said previously, metadata harvesting is preferably done in OAI-PMH, but B2FIND supports also other protocols such as Catalog Service for the Web (CSW) or JSON-API. Once one of these transfer methods has been successfully implemented, B2FIND first takes up a few test samples to analyse their content.

Data provider has to set up the Interface and expose metadata in a standardized format and schema.

2. **Map**[36] metadata on the common B2FIND schema[37]

The provided metadata must be mapped to the B2FIND schema in a meaningful way. Currently this is done in close cooperation between the data provider and the B2FIND team. By iteratively discussing the process a suitable solution is reached in each case.

This mapping is carried out by B2FIND along specifications given by RIs. It is performed on B2FIND side using XPATH rules and semantic mapping.

The implementation of the mapping is based on a detailed specification and documentation of the community-specific metadata. A spreadsheet template has been designed for gathering the required data.

To transform and reformat the harvested raw metadata records to datasets, which can be uploaded to the B2FIND catalogue and indexed and displayed in the B2FIND portal, the following processing steps must be carried out:

1. Select entries from the XML records, based on XPATH rules that depend on community-specific metadata formats
2. Parse through the selected values and assign them to the keys specified in the XPATH rules, i.e. fields of the B2FIND schema.
3. Store the resulting key-value pairs in JSON dictionaries.
4. Check and validate these JSON records before uploading to the B2FIND repository.

This mapping procedure (Figure 12) needs regular adaption and extensions according to the needs of the changing requirements of the communities.

[35] *more info in =>* *http://b2find.eudat.eu/guidelines/harvesting.html*

[36] *more info in =>* http://b2find.eudat.eu/guidelines/mapping.html

[37] *https://github.com/EUDAT-B2FIND/md-ingestion/blob/master/mapfiles/b2find_schema.json*

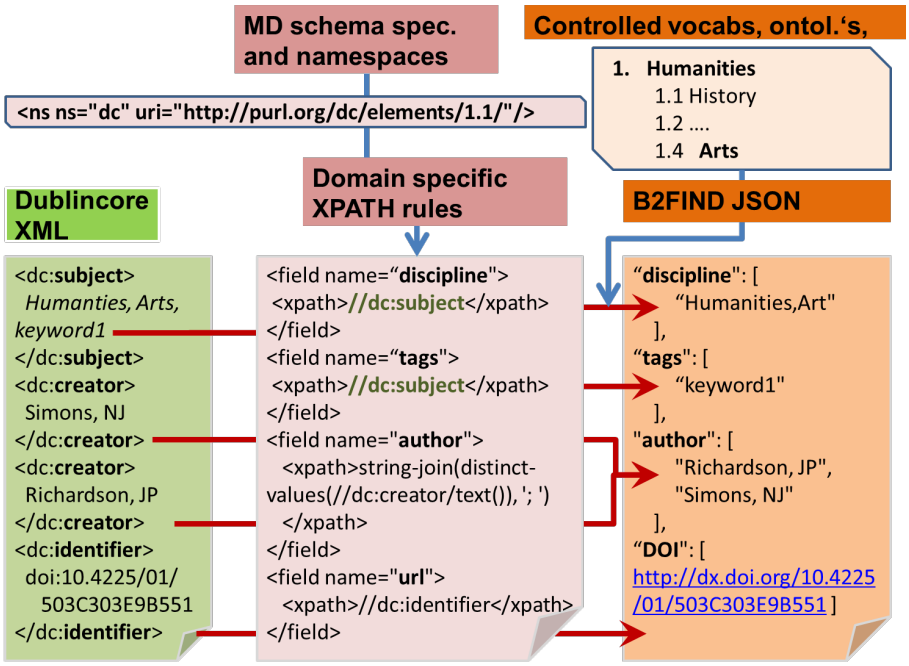

Figure 12. Mapping process

3. **Upload** and index metadata in the B2FIND MD catalogue and search portal

Finally the mapped and validated JSON records are indexed and uploaded as catalogue records to B2FIND.

Metadata schema

To allow a unique search space, B2FIND established a common, interdisciplinary metadata schema. This schema is based on the DataCite Metadata Schema 4.1[38] and therefore as well compatible with guidelines of other e-infrastructures as OpenAire[39], their schemas are based as well on the DataCite schema.

A new metadata schema [40](see Figure 13) has been established in October 2018.

[38] http://schema.datacite.org/meta/kernel-4.1/
[39] https://www.openaire.eu/
[40] https://github.com/EUDAT-B2FIND/md-ingestion/blob/master/mapfiles/b2find_schema.json

Group	B2FIND	Obligation/Occurrence	Description
General Information	Title	mandatory (1)	A name or title by which a resource is known.
	Description	recommended (0-1)	An additional information describing the content of the resource. Could be an abstract, a summary or a Table of Content.
	Tags	optional (0-n)	A subject, keyword, classification code, or key phrase describing the content.
Identifier	DOI	mandatory (1-3) (at least one resource identifier is mandatory)	A persistent, citable identifier that uniquely identifies a resource.
	PID		A persistent identifier that uniquely identifies a resource.
	Source		An identifier (URL) that uniquely identifies a resource. It may link to the data itself or a landing page that curates the data.
	RelatedIdentifier	optional (0-n)	A link to related resources or supplements
	MetaDataAccess	optional (0-1)	Link to the originally harvested metadata record (GetRecord request)
Provenance Information	Creator	recommended (0-n)	The main researchers involved in producing the data, or the authors of the publication, or the measurement or monitoring station that produces the data in priority order.
	Publisher	recommended (0-n)	The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource. This property will be used to formulate the citation, so consider the prominence of the role.

Group	B2FIND	Obligation/Occurrence	Description
	Contributor	optional (0-n)	The institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource.
	PublicationYear	recommended (0-1)	The year when the data was or will be made publicly available.
	Rights	optional (0-n)	Any rights information for this resource.
	OpenAccess	optional (0-1)	Is the dataset openly accessible or not.
	Contact	optional (0-n)	Any contact information for this resource.
Representation Information	Language	optional (0-1)	The primary language of the resource.
	ResourceType	recommended (0-1)	A general type of the resource.
	Format	optional (0-1)	Technical format of the resource.
Coverage	Discipline	recommended (0-n)	A scientific discipline the resource originates from. A closed vocabulary is used.
	Spatial Coverage	optional (0-1)	A geolocation the research data itself is related to. Content of this category is displayed in plain text. If a longitude/latitude information is given it will be displayed at the map.
	Temporal Coverage	optional (0-1)	Period of time the research data itself is related to. Could be a date format or plain text.

Figure 13. B2FIND metadata schema

Results – status of integration in B2FIND

Thus the ingestion process started with some specifically selected RIs in order to test to what extent different exchange formats and different metadata standards could be applied.

Results of this process have been deployed on one the B2FIND test machines which is used for exactly this case, testing the implementation of new Communities. It is openly accessible at http://eudat7-ingest.dkrz.de/dataset. The ingestion of metadata records have successfully been tested for AnaEE, Deims, ICOS, Nilu, Euro-Argo, SeaDataNet, EPOS and IAGOS which are shown as communities in B2FIND.

For the Flagship demonstrator prerequisite was to group these records into one community on B2FIND which is named "ENVRIplus" : http://eudat7-ingest.dkrz.de/dataset?groups=envriplus. Integration of all these records are still in progress.

AnaEE

The catalogue of the French national node (AnaEE-France) is available at http://w3.avignon.inra.fr/geonetwork_anaee/srv/eng/catalog.search#/home

This catalogue is not yet implemented at the AnaEE(-EU) level.

Metadata schema is ISO 19139an ***XML schema*** for the "Geographic ***Metadata*** (gmd)" specification "***ISO 19115***"

Four metadata records describing the main types of AnaEE-F 'products' have been integrated in the ENVRIplus flagship catalogue: i) the services provided for experimentation on continental ecosytems, ii) the modeling platforms and data management services, iii) the projects of experiments on ecosystems using the AnaEE services and iv) the datasets from experimentations using the AnaEE services. In the future the generic 'dataset record' will probably be split into sub-categories e.g. by ecosystem types (forest, freshwater, grassland or cropland).

Epos

The EPOS catalogue is under development at https://epos-ics-c-beta.brgm.fr/epos/epos-gui/master/

Metadata schema used by EPOS is CERIF CERIF https://www.eurocris.org/cerif/main-features-cerif

There is no records from EPOS in B2FIND for the moment because metadata are insufficiently rich for the EPOS user interface and associated processing. The work has concentrated on converting B2FIND CKAN records to CERIF

Euro Argo

Argo is a global array of 3,000 free-drifting profiling floats that measures the temperature and salinity of the upper 2000 m of the ocean. The array provides 100,000 temperature/salinity profiles and velocity measurements per year distributed over the global oceans at an average of 3-degree

spacing. Some floats provide additional bio-geo parameters such as oxygen or chlorophyll. All data collected by Argo floats are publically available in near real-time via the Global Data Assembly Centers (GDACs) in Brest (France) and Monterey (California) after an automated quality control (QC), and in scientifically quality controlled form, delayed mode data, via the GDACs within six months of collection.

All data are publically available from this global repository:

https://doi.org/10.17882/42182

Iagos

IAGOS data portal isn't a metadata catalogue (http://www.iagos-data.fr)

For now, IAGOS metadata records are only in the AERIS catalogue: https://www.aeris-data.fr/catalogue/

This a temporary address. AERIS will setup an official geonetwork server in the future.

http://catalogue2.sedoo.fr/geonetwork/srv/eng/csw-iagos

The AERIS metadata profile is an extension of the INSPIRE recommendation metadata profile. The model is presented in this google document:

https://docs.google.com/spreadsheets/d/1YQu_cSolc_q9HFmt4hxS0Zk3-J3DrDACnXLd8HfxFME/edit?usp=sharing

Six records are harvested by B2FIND:

IAGOS time series

IAGOS vertical profiles

IAGOS climatologies

IAGOS footprints (this one isn't available but it was easier for EUDAT to keep it than remove it, so we can ignore it for now)

IAGOS CO contributions

IAGOS PBL-referenced profiles of O3 and CO

IAGOS doesn't have a lot of datasets. Time series and vertical profiles (a subset of time series) are datasets of level 2 and the most important ones. The other ones are of level 3 and level 4. Those are products that took time to implement and that we want to be discoverable and then used.

ICOS

This is the address to the data portal of ICOS with access to all datasets: https://data.icos-cp.eu/portal/

The ICOS data products that are presented as flagship products in the B2Find catalogue are compilations of a time series of datasets. As such are the flagships not available in the portal, but the individual datasets are.

The URL given on the landing page of each (7) flagship products will point to the selection of datasets available in the ICOS data portal.

This is the URL to the ICOS SPARQL endpoint for harvesting datasets: https://meta.icos-cp.eu/sparqlclient/?type=CSV

Output is default to CSV format, but JSON, XML, TSV or Turtle are offered now as well. Endpoint is offering several predefined requests, but user specific search request can be applied as well.

Metadata schema is ICOS, WIGOS Metadata Standard based on ISO 19115/INSPIRE.

In the catalogue seven different flagship products from ICOS are presented. ICOS is a network of observational stations in three different domains: atmosphere, ecosystem and marine (ocean). In short each of those stations are delivering a timeseries of observational measurements. For atmospheric, high tower stations the focus is on atmospheric concentrations of greenhouse gases. For ecosystem stations the focus is on exchange of CO_2 between the biosphere and atmosphere. For the ocean station the focus is on the concentration of CO_2 in the ocean surface and the flux of CO_2. All stations also deliver a range of additional variables, including meteorological variables.

The first three of the seven flagship products are so called Near-real time (NRT) observational data, one each from the atmosphere, ecosystem and ocean domains. These datasets are available as quickly as possible soon after the actual time of measurement with only automated quality control applied. The second set of three data products, again from each domain, are finalized quality controlled and aggregated times series of observations. (Depending on the type of measurements, the collected data are quality checked and corrected e.g. for calibration of instruments.) The last data product is a model integration product, where atmospheric measurements are combined with an atmospheric transport model, resulting in footprint analysis results. This data product contains timeseries of the source area of observed atmospheric concentrations for selected locations in Europe.

LTER Europe

The DEIMS Site and Dataset Registry[41] (Dynamic Ecological Information Management System, DEIMS-SDR) provides a web based catalogue and editor to identify and document observation facilities (e.g. sites, stations, sensors). DEIMS-SDR is provided by the Umweltbundesamt GmbH and is used by LTER Europe, ILTER but also by a number of European scale projects dealing with long term observations. DEIMS-SDR aims to provide a viable option to fulfil metadata requirements for research projects and national networks if they lack their own system.

DEIMS-SDR is based on Drupal 7 and the current version of DEIMS (Version 2), is a branch of the DEIMS system developed by US LTER (Gries et al. 2010). In addition to the documentation of datasets provided by the DEIMS core branch, DEIMS-SDR also includes the documentation of research sites, networks and persons. For each of the research sites a landing page is provided containing information on the research sites, as well as related information (e.g. datasets and data products).

DEIMS-SDR implements the LTER dataset community profile (Kliment & Oggioni 2011) and allows the export to different metadata formats for datasets (e.g., EML 2.1.1, BDP, ISO19115, ISO19139).

[41] See https://deims.org/

Additionally, DEIMS-SDR includes an interface to directly upload datasets to the open eScience data sharing platform B2SHARE (Ardestani et al. 2015).

The four data products included in the ENVRI Flagship Catalogue are harmonized data which are and will be provided from the eLTER RI in future. Within the current eLTER H2020 project workflows and procedures are developed as well as the data are collected from the partners and harmonized. These data products encompass the following topics:

• eLTER Data Product Site covering the documentation of the site network of LTER Europe and the different national LTER networks. The location of the LTER facilities is provided as center location as well as INSPIRE EF record.

• eLTER Data Product Meteorology covering standard climatology at the different sites harmonized according to the eLTER VA data specification.

• eLTER Data Product Deposition covering atmospheric deposition from a range of sites following the manuals of the UNECE Integrated Monitoring harmonized according to the eLTER VA specification.

• eLTER Data Product Vegetation covering vascular and bryophyte species occurrence using permanent plots applying frequency measurement. Harmonization is done according to the eLTER VA specification.

The aim of the eLTER Data Products is to provide harmonized access to data. Harmonization is done on the level of data formats as well as the reference lists used. For vascular plants e.g. the R-package TLP is used to harmonize species names using a common taxonomy[42]. The data formats and variables to be used are described in the eLTER VA data specification.

The eLTER Data Product Vegetation encompasses terrestrial biodiversity data from a range of LTER facilities in Europe. The data result from vegetation surveys using frequency plots following the UNECE IM protocol. For frequency plots 5 × 5 m areas were used with replicates at each LTER sites. Within each of these areas, vegetation was recorded on two permanently marked 0.5 × 0.5 m subplots. For each subplot, the percentage ground cover of every vascular plant was determined. Harmonization of species names was done with the R-package TLP was used.

The datasets are provided for each LTER site contributing to the eLTER Data Product[43]

Seadatanet

All datasets are available in the web portal: https://www.seadatanet.org/

Catalogue: https://sextant.ifremer.fr/en/web/seadatanet/catalogue#/search?from=1&to=20

For the flagship catalogue three records have been realized to present datasets and products provided by SeaDataNet.

[42] http://www.theplantlist.org/
[43] https://deims.org/dataset/91fbdffa-0cac-46b6-9311-000aa58c6f33

- SeaDataNet is a standardized infrastructure for managing the large and diverse marine data sets collected at sea by the oceanographic fleets, the ships of opportunity and the automatic observation systems. A major objective and challenge in SeaDataNet is to provide an integrated and harmonised overview and access to these data resources, using a distributed network approach. This is achieved by developing, implementing and operating the Common Data Index service that gives users a highly detailed insight into the availability and geographical spreading of marine data across the different data centres across Europe. The CDI provides an ISO19115 - ISO19139 based index (metadatabase) to individual data sets (such as samples, timeseries, profiles, trajectories, etc) and it provides a unique interface to online data access. Data sets are available in ODV (Ocean Data View) and NetCDF (CF) SeaDataNet formats that can be imported to ODV software, which includes the Data Interpolating Variational Analysis software tool (DIVA).

- The SeaDataNet aggregated datasets are regional ODV historical collections of all temperature and salinity measurements contained within SeaDataNet database covering all the European sea basins (Arctic Sea, Baltic Sea, Black Sea, North Sea, North Atlantic Ocean, and Mediterranean Sea).

- SeaDataNet gridded climatologies are based on the SeaDataNet Temperature and Salinity historical data collection v1.1 for all the European sea basins.

How to use the flagship catalogue?

User manual is documented at: https://www.eudat.eu/services/userdoc/b2find-usage

This demonstrator is accessible on a test platform available at: http://eudat7-ingest.dkrz.de/group/envriplus

III.B CERIF

Integration processes

The integration of metadata from different domains within EPOS is accomplished by a matching/mapping/harvesting/conversion process. The mapping uses 3M technology[44] (from FORTH, GR) as used in the VRE4EIC project. The conversion is done in two steps, from the native metadata format of a particular domain to EPOS-DCAT-AP (from KNMI, [Trani *et al.* 2018][45]) and thence to CERIF. This is to reduce the burden on the IT staff in the particular domains since their metadata standards are typically DC, ISO19115/139, DCAT and so closer to DCAT than to CERIF. The onward conversion to CERIF not only permits richer discovery/contextualization/action but also provides versioning, provenance and curation capabilities while allowing metadata

[44] https://www.ics.forth.gr/isl/index_main.php?l=e&c=721

[45] [Trani *et al.* 2018] Luca Trani, Malcolm Atkinson, Daniele Bailo, Rossana Paciello and Rosa Filgueira, *Establishing Core Concepts for Information-Powered Collaborations*, FGCS vol. 89, 421-437, 2018.

enrichment as the domains progressively provide richer metadata as needed for the processing they wish to accomplish (**Figure 11**).

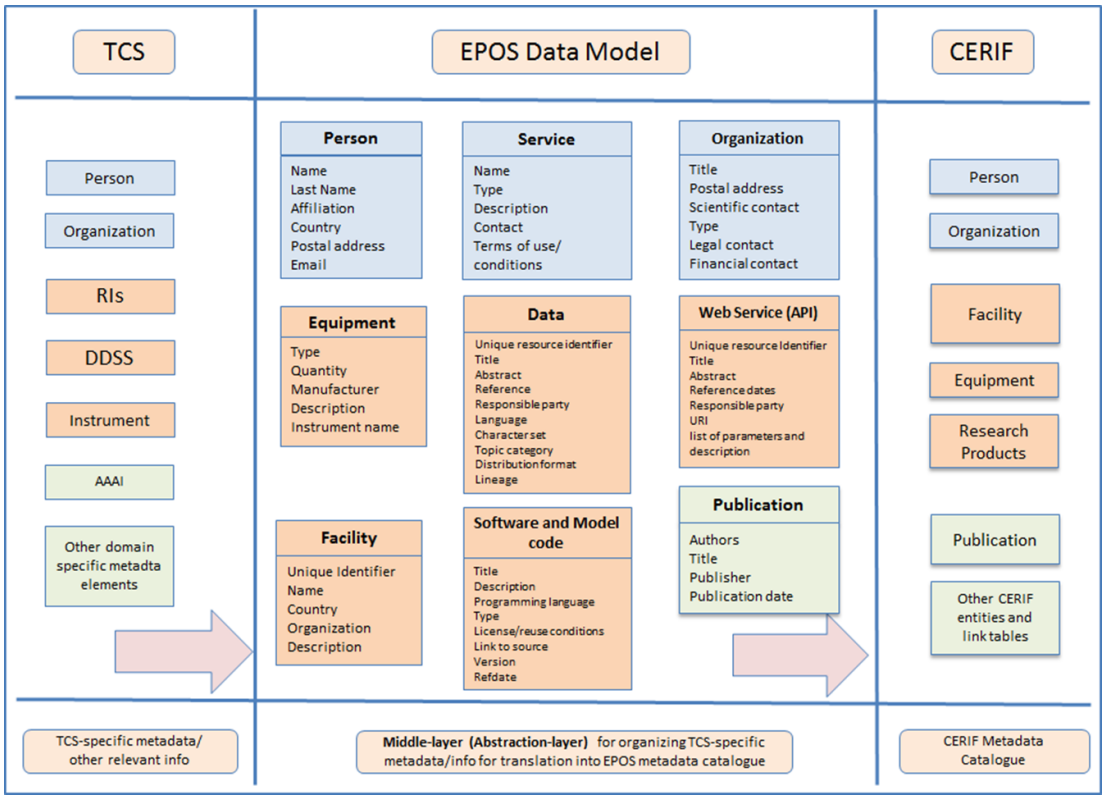

Figure 11 EPOS METADATA HARVESTING ARCHITECTURE

euroCRIS also provide a XML linearization of CERIF for interoperation via web services, as well as scripts for the commonly-used RDBMS implementations.

CERIF schema

The CERIF schema is documented at:

https://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/CERIF1.3_FDM.pdf

And a navigable model in TOAD is at:

https://www.eurocris.org/Uploads/Web%20pages/CERIF-1.5/MInfo.html

Results

CERIF has been used successfully within EPOS in the context of ENVRIplus. However, it is very widely used in research institutions and universities and in research funding organisations throughout Europe and indeed internationally. Of the 6 SMEs providing CERIF systems to the market, one has been taken over by Elsevier and one by Thomson-Reuters and thus incorporating CERIF in their products.

The EPOS CERIF catalog content has been loaded into a RDBMS at IFREMER. The current work is to provide the user interface software to be used at that location. In parallel work proceeds on (a) converting CERIF to CKAN for inclusion in the EUDAT B2FIND CKAN catalog and (b) converting CKAN to CERIF to include the records of other RIs from the EUDAT B2FIND CkAN catalog into CERIF. The work proceeds slowly because of resource limitations.

How to use CERIF (user manual)

An introductory presentation on CERIF is obtained by clicking on the image at

https://www.eurocris.org/cerif/main-features-cerif

A tutorial to be downloaded is also available at right-hand-side of

https://www.eurocris.org/community/taskgroups/cerif

CONCLUSIONS

T8.2 in ENVRIplus concerns cataloguing and specifically cataloguing of dataset assets for the purposes of discovery and download. The RIs within the ENVRI community each had catalogs of varying size, detail, stage of development and format. Few had any provision for associated curation or provenance tracking which are the other – related – tasks in WP8 requiring integrated catalog entries. After the first round of deliverables of WP8, RIs started to develop policies to address these issues. To demonstrate cataloguing capabilities a two-pronged approach was adopted. A few records were converted and uploaded from several RIs to a B2FIND test server and indexed for the search. This exposed the effort of metadata mapping but also the capability of a catalog with metadata from different domains. Separately the EPOS metadata catalog of services was used as an exemplar of the use of CERIF for integrated cataloguing, curation and provenance. The objective of these two parallel exercises was to allow RIs to see what can be achieved – and what effort is necessary - in integration of heterogeneous metadata describing assets to permit homogeneous cross-domain (re-)use of assets.

Further enhancements and improvements of the mapping (from various metadata formats used by the RIs to a canonical format) are necessary before the ENVRIplus records could be published and searchable in the production B2FIND portal. Within EPOS 16 different metadata formats had to be mapped and converted to be ingested into the CERIF catalog and made available for (re-)use. The effort of correct matching and mapping between metadata standards should not be underestimated but – once achieved – can provide a homogeneous access over heterogeneous asset descriptions. Furthermore, work on improved harvesting protocols allowing both push and pull initiation and incremental updates (especially for metadata describing streamed data) needs to be done.

As indicated in D8.2 the choice of the metadata elements in the catalogue (including their syntax and semantics) is crucial for the processes not only of curation but also of provenance and catalogue management and utilisation for dataset discovery and download. The RIs have different metadata formats and each has its own roadmap or evolution path improving metadata as required by their community. Unfortunately there are many metadata standards, some general (and usually too abstract for scientific use) and some detailed and domain specific (but not easily

mapped against other formats). Mapping every metadata standard to every other is a *n(n-1)* problem. Mapping to one canonical rich metadata format, sufficient to accommodate all of the others, is a problem of magnitude *n*. The need for rich metadata is becoming generally accepted; W3C is revising DCAT to make it much more expressive; currently RDA is working on a recommended set of metadata elements and the relationships between them; it is to be hoped this becomes a generally accepted reference canonical format as the target for metadata mappings.

T8.2 has provided as demonstrators catalogs (1) in EUDAT/B2FIND technology using CKAN metadata implemented as attributes in a RDBMs (Relational Database Management System) with imported and converted metadata records from several RIs; (2) in CERIF as used in EPOS (and elsewhere); in the EPOS case the current catalog covers services (that access and process datasets) and so is aligned with the trend in Europe - for example the services catalog of the EOSC. In EPOS, metadata describing datasets, equipment and software assets are now being converted and ingested to provide a full catalog of assets.

During T8.2 meetings, many discussions about the feasibility to implement a semantic linking synonym service took place. The objectives of this semantic linking service were to extend keyword search request with synonym keywords and possibly approximate synonyms within the same querying interface. The issue was to demonstrate that the use of different annotation system based on different referential bases can be exploited in a common platform. Further investigations will be required to establish a prototype. CERIF provides in its 'semantic layer' the capability for crosswalking between terms in different ontologies, even multilingually. CKAN does not possess natively this capability.

The cost of collecting, inputting and maintaining metadata describing assets is high. To reduce it requires (a) automation; (b) incremental collection of metadata along the workflow to avoid repeated input (with consequent effort and introduced inconsistency errors). The more that RIs in the ENVRI community can converge on a few standards (both for metadata syntax and semantics) and a few standard metadata management processes the more feasible becomes integration and interoperation.

Increasing energy utilisation in network traffic – with both cost and environmental implications – requires a re-think. It is no longer sensible to discover and download datasets; the cost (both economic and environmental) is too high. The emerging requirement is for an integrated approach with distributed processing of data either at the originating node, or – for multi-domain required processing - transmitting only selected subsets from the originating node to a central node (possibly a specialised computing facility such as HPC) for combined processing. This can be combined with data partitioning, fragmentation, replication (including caching) – all recorded with provenance and curation metadata. This plays into the European agenda for future cloud computing including Fog and Edge computing. Further work in the ENVRI community – partly in the project ENVRIFAIR commencing January 2019 – will address these issues.

It is planned to continue – in the ENVRI community - with the EUDAT B2FIND catalog (maintained by EUDAT) and also to continue the work with CERIF (maintained by EPOS), anticipating the need for richer metadata than the CKAN schema can carry for the functionality associated with services – and other RI assets - in the EOSC (European Open Science Cloud).