Data Cataloguing

Erwann Quimbert[1](✉), Keith Jeffery[2], Claudia Martens[3], Paul Martin[4], and Zhiming Zhao[4]

[1] Ifremer, BP 70, 29280 Plouzané, France
erwann.quimbert@ifremer.fr
[2] Keith G Jeffery Consultants, 71 Gilligans Way, Faringdon SN FX, UK
keith.jeffery@keithgjefferyconsultants.co.uk
[3] German Climate Computing Center [DKRZ], 20146 Hamburg, Germany
martens@dkrz.de
[4] Multiscale Networked Systems, University of Amsterdam,
1098XH Amsterdam, The Netherlands
paulmartin.research@google.com, z.zhao@uva.nl

Abstract. After a brief reminder on general concepts used in data cataloguing activities, this chapter provides information concerning the architecture and design recommendations for the implementation of catalogue systems for the ENVRIplus community. 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 offer 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.

Keywords: Catalogue · Metadata · Data · Interoperability · Standard · ISO · OGC · Format · Schema

1 Introduction

Data catalogues have been used in data management for a long time. Under the impetus of European regulations, the number of metadata catalogues has been growing steadily over the last decade, and more specifically thanks to the Inspire Directive [1], which has made it mandatory for public authorities to create metadata more easily and to share them more widely. Data catalogues provide information about data concerning one or many organizations, domains or communities. This information is described and synthesised through metadata records. Data catalogue centralised metadata is gathered in one location, usually accessible online through a dedicated interface. In this chapter, we will focus on data catalogues related to environmental sciences.

A common definition is that metadata is "data about data". Metadata provide information on the data they describe to specify who created the data, what it contains, when it

Z. Zhao and M. Hellström (Eds.): Towards Interoperable Research Infrastructures for Environmental and Earth Sciences, LNCS 12003, pp. 140–161, 2020.
https://doi.org/10.1007/978-3-030-52829-4_8

was created, why it was created, and in which context. Metadata can be created automatically or manually and they are structured to allow easy and simple reading by end-users and by automated services.

As proposed by Riley [2], metadata can be classified into 3 categories:

1. **Descriptive metadata** give a precise idea about the content of a resource. Descriptive metadata may include a title, a description, keywords and one or many points of contact (creator, author, and editor). These metadata elements allow end-users to easily find a resource and to know if this resource fits their purpose and their research needs.
2. **Administrative metadata** include technical metadata (providing information about the format, file size, how they have been encoded, and software used), rights metadata (including user limitations, access rights, intellectual property rights and copyright constraints) and provenance metadata (lineage of the data, why this data has been created, by whom, and in which context).
3. **Structural metadata** provide information about the files that make up the resource and specify the relationships between them.

To complete this classification, it is often accepted that good metadata is metadata that is able to answer the 5 W's: *Who, What, Where, When* and *Why*.

RDA (Research Data Alliance) has developed agreed principles concerning metadata discussed in (Chapter 7) including the assertion that there is no difference between metadata and data except the use to which it is put. A library catalogue card used by a researcher to locate a scholarly paper is metadata when among other cards used by a librarian to count articles on river pollution it is data.

The purpose of data catalogues is multifold. One of its biggest benefits is to organise and centralise the metadata in one location which greatly facilitates data discovery for end-users and make data more accessible for different types of users (data consumers, data scientists or data stewards).

Data catalogues also avoid duplication of data.

Data catalogues exist to collect, create and maintain metadata. These records are indexed in a database and end-users should access the information through a user-friendly interface. This interface should offer common data search functionalities allowing users to narrow down their search according to different criteria: keywords (controlled vocabularies), geographic location, temporal and spatial resolution, and data sources.

Data catalogues have become an important pillar in the data management lifecycle. Indeed, almost every step of the data lifecycle is described in the metadata fields or accessible through the data catalogue online interface. Curated data are described by effective and structured metadata (cf. Riley's list above) providing information about data collection (e.g. metadata automatically produced about sensors/instruments) data processing (data lineage, software used, explanations of the different steps of data construction), data analysis (description of methods applied), data publishing (discovery metadata, policies for access, reuse and sharing) and data archiving (preserving data).

2 Metadata Standards and Interoperability Between Data Catalogues

2.1 Metadata Standards

"Metadata is only useful if it is understandable to the software applications and people that use it" [2]. We often speak about schema to illustrate the metadata structure. To facilitate this understanding metadata generally follow standardised schemas implementing recommendations from international organizations such as ISO[1] (International Organization for Standardization). There are several metadata standards widely used in the environmental science domain. It will not be possible to fully describe them in this chapter but a short description is given explaining in which community they are commonly used. To simplify integration within systems metadata, a machine-readable language is often used such as XML or RDF or even JSON-LD.

Metadata Standards versus Metadata Schemas
The terms 'schema' and 'standard' are used in an interchangeable way, but all refer to "the formal specification of the attributes (characteristics) employed for representing information resources" [3]. Yet another definition for 'metadata schema' is a "logical plan showing the relationships between metadata elements, normally through establishing rules for the use and management of metadata specifically as regards the semantics, the syntax and the optionality" [ISO/TC46, 2011] whereas 'syntax' describes the structure of a schema (language, rules to represent content) and 'semantics' describe the meaning of its elements, properties or attributes. Following Haslhofer and Klas [4] a metadata schema could be seen as a set of elements with a precise semantic definition and optionally rules how and what values can be assigned to these elements; a metadata standard then is a schema which is developed and maintained by an institution that is a standard-setting one. Hence a standard is a standard insofar as there is an institutional or organizational standardization unit developing and maintaining a standard - whereas all parties and persons involved agree this institution to be trustworthy and reliable. Some relevant standards are mentioned below.

ISO 19115 [5] is an internationally adopted schema for describing geospatial data. As indicated in their website "it provides information about the identification, the extent, the quality, the spatial and temporal aspects, the content, the spatial reference, the portrayal, distribution, and other properties of digital geographic data and services."

DataCite[2] [6] is an international consortium founded in 2009 with an emphasis to make explicitly *research data* citable, giving them a 'value' during the scientific process: "a persistent approach to access, identification, sharing and re-use of datasets" [6][3]. DataCite promotes the use of Persistent Identifiers for Digital Objects in order to unambiguously identify a digital resource, established as DOIs[4].

[1] https://www.iso.org/.
[2] https://schema.datacite.org/.
[3] https://schema.datacite.org/meta/kernel-4.1/doc/DataCite-MetadataKernel_v4.1.pdf.
[4] http://www.doi.org/index.html.

Dublin Core Metadata Initiative[5] [7] was founded in the aftermath of a World Wide Web conference during a workshop at the OCLC[6] (an organisation for a global digital library providing technology) headquartered in Dublin, Ohio (USA), aiming at achieving "consensus on a list of metadata elements that would yield simple descriptions of data in a wide range of subject areas for indexing and cataloguing on the Internet" [7]. Dublin Core was originally developed mainly by librarians, where 15 (initially 13 but extended when additional attributes were required) 'core' metadata elements[7] contain resource descriptions (contributor, coverage, creator, date, description, format, identifier, language, publisher, relations, rights, source, subject, title, and type). As these descriptions have been regarded as not sufficient, they were refined to 'qualified DC' by 55 'terms'[8]. DC has been represented progressively over time by text, HTML, XML and - recently - RDF. Only in this latter form does it approach the requirement for formal syntax and declared semantics.

CERIF[9] is a data model recommended by the European Union to the Member States for research information. It is described in some detail below.

DCAT [8] is a W3C recommendation 'data catalogue vocabulary' and has the advantage of being conceived natively with qualified relationships and use of RDF triples. It is currently undergoing revision by the DXWG (Data Exchange Working Group)[10].

Schema.org[11] is an initiative from Google and Microsoft now a community activity. It essentially provides a list of attributes, some with related vocabularies, for entities. In this way it is like CERIF: schema.org has entities for person and organisation, product and place for example. It may be encoded in RDF or JSON-LD.

All have some relevance to ENVRI. RIs are encouraged to choose a schema that has the capability to describe their 'world of interest'. Only rich metadata schemas (such as CERIF) can provide a unifying data model to which the others may be converted in a lossless manner.

Specification versus Interoperability

While Dublin Core and DataCite are generic metadata standards that aim to provide a minimum of metadata elements for describing a digital resource, ISO19115/19139 [8] is a standard especially for georeferenced data. The question is how to find an equilibrium between 'general' information that is sufficient to search and access research data across scientific disciplines on the one side and 'specific' information describing resources from certain research communities on the other side is not clearly answered yet (and maybe can´t be answered at all). RDA (Research Data Alliance) is working on a set of common metadata elements (each with syntax and semantics) linked by qualified references to act

5 https://www.dublincore.org/.
6 https://www.oclc.org/.
7 http://dublincore.org/documents/dces/.
8 http://dublincore.org/documents/dcmi-terms/.
9 https://www.eurocris.org/cerif/main-features-cerif.
10 https://www.w3.org/2017/dxwg/wiki/Main_Page.
11 https://schema.org/.

as rich metadata set for FAIR (Findability, Accessibility, Interoperability, Reusability) [9] with the aim of overcoming this problem[12].

2.2 Data Catalogues Tools

There are many tools used by scientific communities to create data catalogues. Two example tools used by the environmental and Earth science research communities are GeoNetwork and CKAN.

GeoNetwork[13] is an open-source software allowing the creation of customised catalogue applications. This tool is mainly used for describing and publishing geographic datasets and is related to ISO 19115/19139.

CKAN[14] is an open-source Data Management System widely used in the world of open data. It uses essentially some Dublin Core metadata elements[15] but allows for an infinite extension of additional attributes thus making interoperation difficult. EUDAT B2FIND uses CKAN for its frontend.

Independently of the software used, protocols exist for sharing metadata between data catalogues, in particular OGC-CSW[16], OAI-PMH[17], SPARQL[18] and others.

3 Design for ENVRI

3.1 ENVRIplus Context

Data cataloguing is a key service in the data management lifecycle of ENVRIplus [18–20]. For ENVRIplus, an interoperable catalogue system aims at organizing the maintenance and access to descriptions of resources and outcomes of multiple Research Infrastructures in a framework which implements a number of functions on these descriptions. As defined in the ENVRI Reference Model (Chapter 4), maintenance of a catalogue is a strategic component of the curation process and the descriptions maintained in the catalogue support the acquisition, publication and re-use of data. The system must provide to users a function for the seamless discovery of the description of resources in the Research Infrastructures, encoded using a standardised format. The multi-Research Infrastructures context of ENVRIplus implies that, in addition to the descriptions usually available within each Research Infrastructure, resources may also have to be described at a higher granularity so to provide context.

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

[12] https://drive.google.com/drive/folders/0B8FnM3PsoL2dd2RnYVBmcjRMYXc.
[13] https://geonetwork-opensource.org/.
[14] https://ckan.org/.
[15] https://ckan.org/portfolio/metadata/.
[16] https://www.opengeospatial.org/standards/cat.
[17] https://www.openarchives.org/pmh/.
[18] https://www.w3.org/TR/rdf-sparql-query/.

A Top-Down approach has been used with the aim of showcasing the products of the Research Infrastructures 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 Research Infrastructure data products.

3.2 RIs Involved in the Flagship Catalogue

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

- AnaEE[19] (Analysis and Experimentation on Ecosystems) focuses on providing innovative and integrated experimentation services for research on continental ecosystems.
- Euro-Argo[20] 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[21] is a pan-European Research Infrastructure for marine biology and ecology research.
- EPOS[22] (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[23] (In-Service Aircraft for a Global Observing System) is a European Research Infrastructure for global observations of atmospheric composition using commercial aircraft.
- ICOS[24] is a pan-European research infrastructure for quantifying and understanding the greenhouse gas balance of Europe and its neighbouring regions.
- LTER[25] (Long Term Ecological Research) is an essential component of world-wide efforts to better understand ecosystems.
- SeaDataNet[26] is a pan-European infrastructure to ease the access to marine data measured by the countries bordering the European seas.
- Actris[27] is the European Research Infrastructure for the observation of Aerosol, Clouds, and Trace gases.

[19] https://www.anaee.com/.
[20] https://www.euro-argo.eu/.
[21] http://www.embrc.eu.
[22] https://www.epos-ip.org/.
[23] http://www.iagos-data.fr/.
[24] http://www.icos-ri-eu.
[25] http://www.lter-europe.net/.
[26] https://www.seadatanet.org/.
[27] https://www.actris.eu/.

Four kinds of users were identified for this flagship catalogue:

- Users outside a Research Infrastructure, researching data-driven science.
- Users inside a Research Infrastructure, such as data managers, coordinators, and operators as well as data scientists.
- Stakeholders, decision-makers and funders of the Research Infrastructures who need to have a broad picture of the Research Infrastructure resources in the European landscape to control their efficiency and complementarity.
- Policymakers, using ENV RI information for government policy and laws.

3.3 Proposed Architecture

At the beginning of the project, it was decided to not create a new metadata model. The requirements on product description were defined by adopting the metadata elements of the RDA metadata interest group[28]. We noticed that this schema gathers most of the common properties among different data models exposed above. The idea is to automatically map the metadata model from each Research Infrastructures to a canonical schema. We also encouraged the use of existing controlled vocabularies.

CERIF and CKAN frameworks are both chosen candidates for prototyping an ENVRIplus community catalogue for Research Infrastructures flagship data products.

To streamline the implementation of this flagship catalogue, it was decided to start with the EUDAT/B2FIND[29] demonstrator. The demonstrator on CERIF has also been developed jointly with EPOS and other relevant projects, e.g. VRE4EIC[30].

4 Cataloguing Using B2FIND

4.1 B2FIND Description and Workflow

B2FIND[31] is a discovery service for research data distributed within EOSC-hub and beyond. It is a basic service of the pan-European data infrastructure EUDAT CDI (Collaborative Data Infrastructure)[32] that currently consists of 26 partners, including the most renowned European data centres and research organisations. B2FIND is an essential service of the European Open Science Cloud[33] (EOSC) as it is the central indexing tool for the project that constitutes the EOSC (EOSC-Hub).

Therefore a comprehensive joint metadata catalogue was built up that includes metadata records for data that are stored in various data centres, using different meta/data formats on divergent granularity levels, representing all kinds of scientific output: from huge netCDF files of Climate Modelling outcome to small audio records of Swahili

28 https://rd-alliance.org/groups/metadata-ig.html.
29 http://b2find.eudat.eu/.
30 https://www.vre4eic.eu/.
31 http://b2find.eudat.eu/.
32 https://www.eudat.eu/eudat-cdi.
33 https://www.eosc-portal.eu/about/eosc.

syllables and phonemes; from immigrant panel data in the Netherlands to a paleoenvironment reconstruction from the Mozambique Channel and from an image of "Maison du Chirugien" in ancient Greek Pompeia to an xlsx for concentrations of Ca, Mg, K, and Na in throughfall, litterflow and soil in an Oriental beech forest.

In order to enable this interdisciplinary perspective, different metadata formats, schemas and standards are homogenised on the B2FIND metadata schema[34], which is based on the DataCite schema extended with the additional element <Disciplines>, allowing users to search and find research data across scientific disciplines and research areas. Good metadata management is guided by FAIR principles, including the establishment of common standards and guidelines for data providers. Hereby a close cooperation and coordination with scientific communities, Research Infrastructures and other initiatives dealing with metadata standardisation (OpenAire Advance, RDA interest and working groups and the EOSCpilot project to prepare the EOSC including a task on 'Data Interoperability'[35]) is essential in order to establish standards that are both reasonable for community-specific needs and usable for enhanced exchangeability. The main question still is how to find a balance between community-specific metadata that serve their needs on the one side and a metadata schema that is sufficiently generic to represent interdisciplinary research data but at the same time is specific enough to enable a useful search with satisfying search results.

Harvesting

Preferably B2FIND uses the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to harvest metadata from data providers. OAI-PMH offers several options that make it a suitable protocol for harvesting: a) possibility to define diverse metadata prefixes (default is Dublin Core), b) possibility to create subsets for harvesting (useful for large amounts of records, resp. divergent records e.g. from different projects or sites or measurement stations) and c) the possibility to configure incrementally harvesting (which allows to harvest only new records). Nonetheless, other harvesting methods are supported as well, e.g. OGC-CSW, JSON-API or triples from SPARQL endpoints.

Mapping

The mapping process is twofold as it includes a format conversion as well as a semantic mapping based on standardised vocabularies (e.g. the field 'Language' is mapped on the ISO 639 library[36] and research 'Disciplines' are mapped on a standardised closed vocabulary). Therefore, entries from XML records are selected based on XPATH rules that depend on community-specific metadata formats and then parsed to assign them to the keys specified in the XPATH rules, i.e. fields of the B2FIND schema. Resulting key-value pairs are stored in JSON dictionaries and checked/validated before uploaded to the B2FIND repository. B2FIND supports generic metadata schemas as DataCite and Dublin Core. Community specific metadata schemas are supported as well, e.g.

[34] http://b2find.eudat.eu/guidelines/mapping.html.
[35] https://www.eoscpilot.eu/content/d69-final-report-data-interoperability.
[36] https://iso639-3.sil.org/code_tables/639/data.

ISO19115/19139 and Inspire for Environmental Research Communities or DDI[37] and CMDI[38] for Social Sciences.

Upload and Indexing
B2FIND´s search portal and GUI is based on the open-source portal software CKAN, which comes with Apache Lucene SOLR Servlet allowing indexing of the mapped JSON records and performant faceted search functionalities. CKAN was created by the Open Knowledge Foundation (OKFN) and is a widely used data management system. CKAN has a very limited internal metadata schema[39] which has been enhanced for B2FIND while creating additional metadata elements as CKAN field "extra". B2FIND offers a full text search, results may be narrowed down using currently 11 facets (including spatial/temporal search and facets <Discipline>, <ResourceType>, <Publisher>, <Contributor>, <Language>, <Community>, <Tags> and <Creator>). "Community" here is the data provider where B2FIND harvests from.

4.2 B2FIND and FAIR Data Principles

FAIR data principles [9] are recommended guidelines to increase the impact of data in science generally by making them findable, accessible, interoperable and reusable. While these principles are increasingly recognised, specific elements need to be clarified: how to implement FAIR data principles during the data lifecycle? How to measure "FAIRness"? By whom? Currently, supporting FAIR data principles are done in varying ways with different methods[40]. The approach of B2FIND to these guidelines may be characterised as supporting 'Findability' by offering a discovery portal for research data based on a rich metadata catalogue, supporting 'Accessibility' by representing Persistent Identifiers for unique resolvability of data objects, supporting 'Interoperability' by implementing common standards, schemas and vocabularies and finally supporting 'Reusability' by offering licenses, provenance and domain-specific information. However, while FAIR principles refer to both data and metadata, B2FIND may manage only the metadata aspect.

4.3 Flagship Implementation

The implementation of ENVRIplus Flagship catalogue in B2FIND faced two main challenges: 1) how to integrate metadata records that are representing Research Infrastructures rather than Datasets, and 2) how to represent these RIs as part of ENVRIplus

[37] The Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioural, economic, and health sciences. https://ddialliance.org/.

[38] The Component MetaData Infrastructure (CMDI) provides a framework to describe and reuse metadata blueprints. Description building blocks ("components", which include field definitions) can be grouped into a ready-made description format (a "profile"). https://www.clarin.eu/content/component-metadata.

[39] https://docs.ckan.org/en/ckan-1.7.4/domain-model.html#overview.

[40] GO FAIR initiative is a good example, therefore: one aim is to support 'Implementation Networks', whereas these networks define in how far they are FAIR. See therefore: https://www.go-fair.org/.

within the B2FIND architecture. These questions concerned both the technical level and content-related issues and are described below. The implementation process itself revealed challenges that may be seen as exemplary: how to deal with persistent identifiers and how to deal with granularity issues.

A. **RI Dataproducts**

As described above B2FIND is first and foremost a search portal for research data that should be findable across scientific disciplines. It is not primarily meant to be a search portal for other information as e.g. funding bodies, site information or research infrastructure descriptions. Concerning RIs that are part of ENVRIplus, most of them have their own search interface and some of them have already made their repositories harvestable. Thus, the flagship implementation started with harvesting already existing RI endpoints (DEIMS[41], NILU[42], EPOS, SeaDataNet, Euro-Argo, AnaEE, ICOS Carbon Portal[43]) and integrating them as "Communities" into a B2FIND testing machine[44], which means representing their data as e.g. "DEIMS". One challenge on B2FIND side was to develop the software stack[45] in order to be able to harvest from CSW endpoints. On the Data Provider side, the proper CSW configuration has been a task insofar as CSW does not yet allow the creation of Subsets (which would enable harvesting of just one subset for testing) and resumption token. Another issue concerned incrementally harvesting: OAI-PMH allows to exchange information of 'record status' and 'timestamp', which means that it is possible to harvest just those records that are not e.g. 'deleted' or those from a certain period of time (e.g. every week). CSW does not yet support these features. Creating a mapping for each "Community" has been relatively simple as all RIs use either DublinCore or ISO19139 as their metadata standard and usually XML as an exchange format. The only exception is ICOS that expose their metadata as triples. The decision to use the Flagship Catalogue for representing Data products (which means records that describe the *services* offered by the RIs rather than their *data*) compelled the RIs to create metadata records that fitted this purpose and expose them in a way that enabled B2FIND to ingest them.

B) **B2FIND/Flagship architecture**

Initially, the Flagship catalogue should have been visible in a way that would display both ENVRIplus as the main project and each RI as a part of it. CKAN allows to create "Groups" and "Subgroups"; however, B2FIND is constructed as CKAN "Group" and its "Communities" as CKAN "Subgroups" which means that a further distinction between ENVRIplus and RIs could not be implemented. In order to enable a search for RIs the decision was to create a 'Community' = ENVRIplus and use the metadata element

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

[42] https://www.nilu.no/en/.

[43] The data centre of ICOS, https://icos-cp.eu/.

[44] http://eudat7-ingest.dkrz.de/dataset.

[45] B2FIND uses CKAN only for GUI and search interface while the backend is developed B2FIND code, it´s Open Source on GitHub: https://github.com/EUDAT-B2FIND.

<Contributor> as a distinctive feature (Fig. 1). As the flagship implementation enforced B2FIND to enhance its metadata schema (to enable a faceted search via <Contributor>) it was implemented on a test machine at DKRZ[46]. The demonstrator may be seen here: http://eudat7-ingest.dkrz.de/dataset?groups=envriplus.

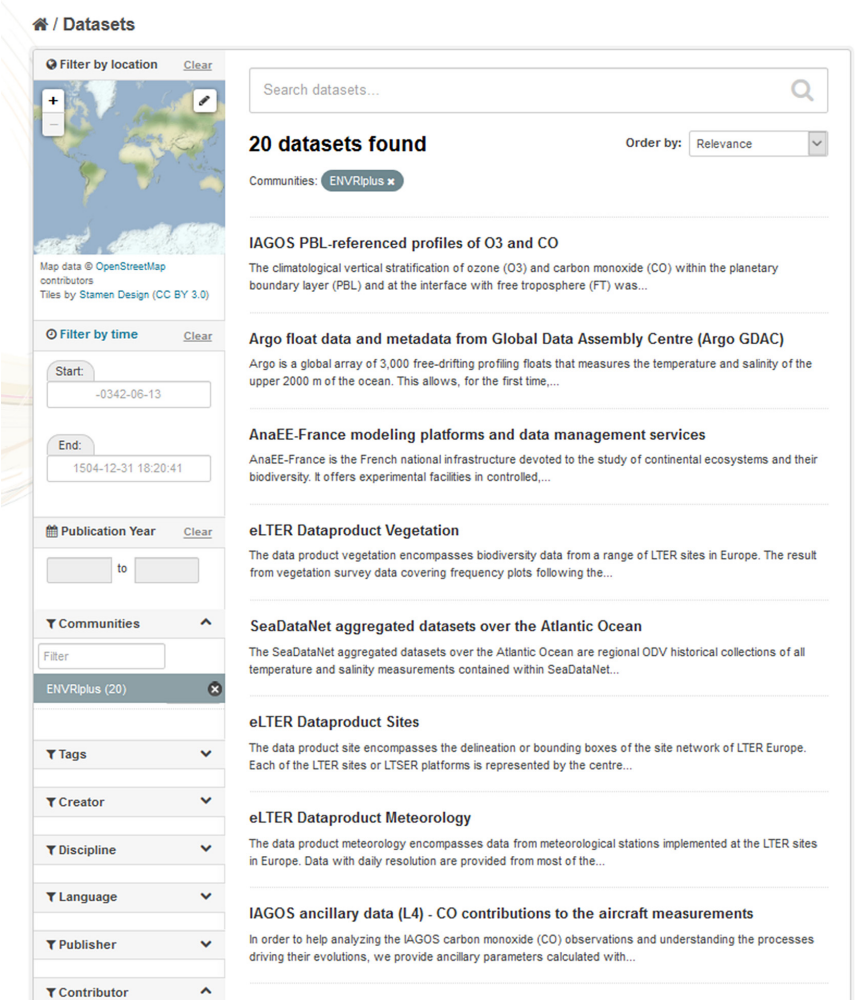

Fig. 1. Flagship catalogue in B2FIND: partial search result page.

As described above B2FIND links to a certain resource by using persistent identifiers (if offered within the metadata) in order to increase the reliability of a digital resource (Fig. 2). Therefore, an internal 'ranking' is used: if a DOI is provided it will be displayed, both as a link to the Landing page and additionally as a small icon on the single record

[46] https://www.dkrz.de/about-en.

entry page. If no DOI but another PID (e.g. a Handle) is offered this one will be shown, both as a link and as an icon. If none (DOI or PID) is given, B2FIND will represent any other URN or URL.

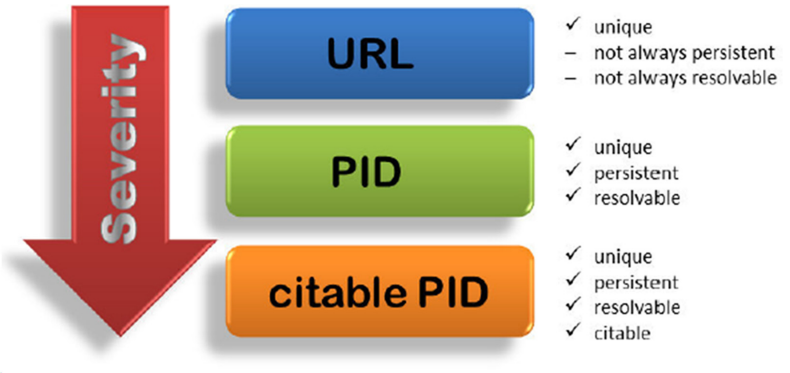

Fig. 2. Consistency of identifiers.

For the flagship implementation the RIs 'Dataproducts' did not all provide a DOI or PID (except for IAGOS, see Fig. 3) but an identifier that links to the described resource. Some effort was needed to define where the 'Source' information is given - some RIs presented internal identifier within the metadata element <identifier> (such as UUIDs) which are not automatically resolvable, sometimes this information was given in <alternateIdentifier> or <relatedIdentifier> attributes or within the header. To solve this problem a specific map file for each RI was created that defined the XPATH rules for each metadata element in order to map it onto the B2FIND schema.

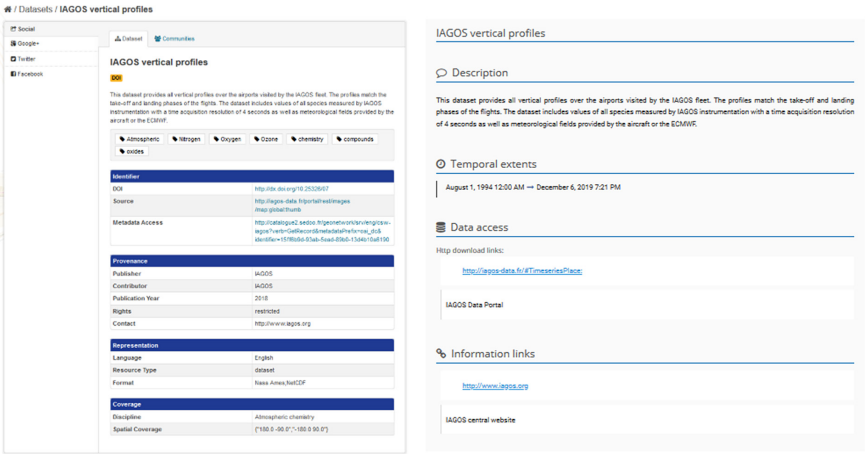

Fig. 3. B2FIND single record entry which links to IAGOS Landing page.

The effort spent on implementing the flagship product catalogue was useful as it initiated concrete technical developments on both sides (e.g. regarding CSW harvesting or enhanced B2FIND schema including <Contributor>). Nonetheless, it is questionable whether B2FIND is an adequate catalogue for ENVRIplus RI 'data products' as it is first and foremost a search portal for research data (and not services).

5 Cataloguing Using CERIF

5.1 EPOS Implementation

CERIF[47] (Common European Research Information Format) is an EU Recommendation to the Member States for research information since 1991. In 2000 CERIF was updated to a richer model, moving from a model like the later Dublin Core to the CERIF as used today: an extended-entity-relational-temporal model. The European Commission requested euroCRIS to maintain, develop and promote CERIF as a standard. It is a data model (Fig. 4) based on EERT (extended entity-relationship modelling with temporal aspects).

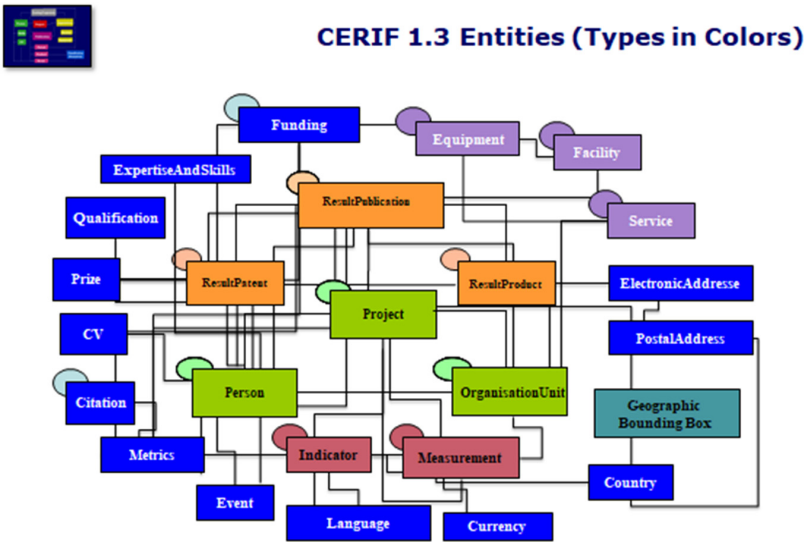

Fig. 4. CERIF Data Model showing entities (boxes) and relationships (lines) (Acknowledgement Brigitte Jörg).

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

[47] An introductory presentation on CERIF: https://www.eurocris.org/cerif/main-features-cerif Tutorial: https://www.eurocris.org/community/taskgroups/cerif.

with a particular approach thus ensuring referential and functional integrity. CERIF has the concept of base entities representing real-world objects of interest and characterised 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 and 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, ISO19115/19139, eGMS, DDI, CKAN(RDF), RIOXX and others) to/from CERIF, emphasizing its richness and flexibility.

Some Existing Use Cases

EPOS uses CERIF for its catalogue 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 catalogue (ICS: Integrated Core Services) is based on microservices (Fig. 5).

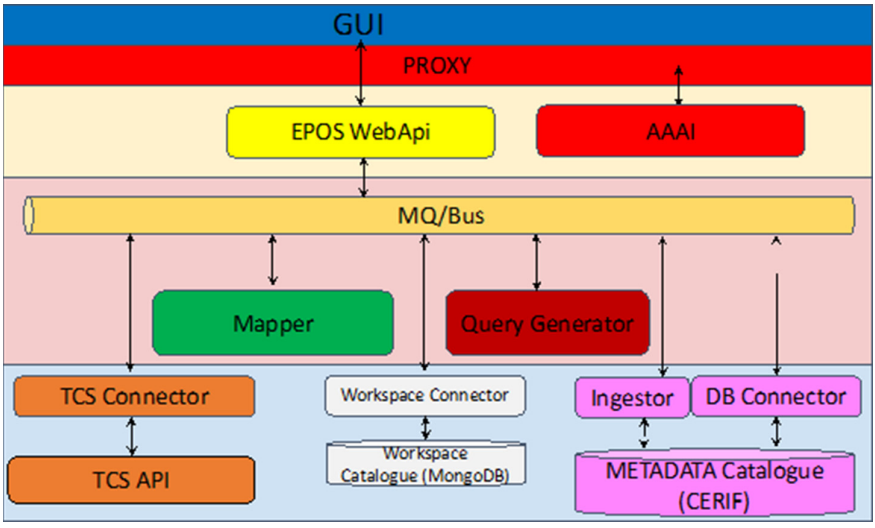

Fig. 5. EPOS ICS architecture.

The implementation uses PostgreSQL as the RDBMS and has been demonstrated on numerous occasions (Fig. 7). A mechanism for harvesting metadata from the various domain groups of EPOS (TCS: Thematic Core Services) and converting from their

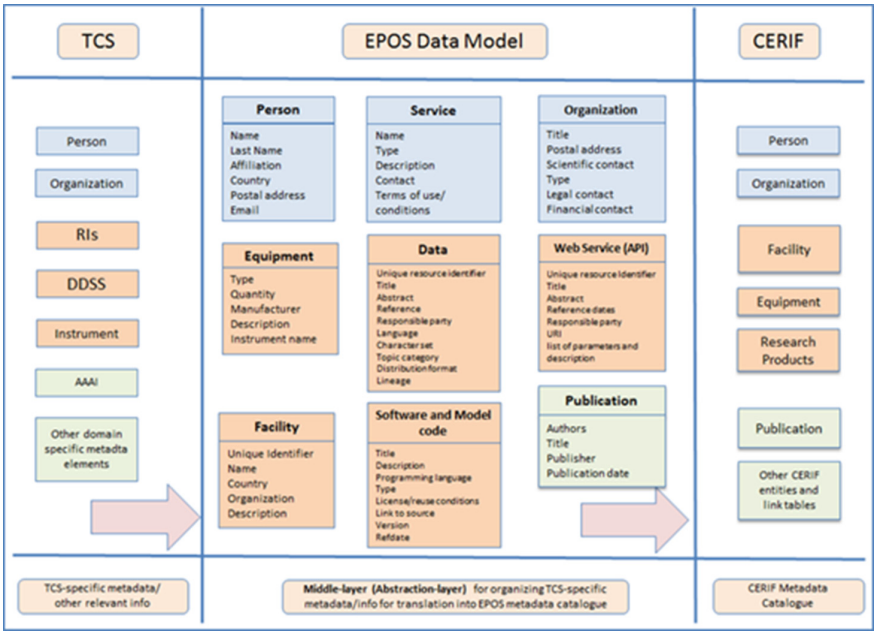

Fig. 6. EPOS metadata harvesting architecture.

individual metadata schemes to CERIF has been implemented including an intermediate stage using EPOS-DCAT-AP - a particular application profile of the DCAT standard [11]. (Figure 6).

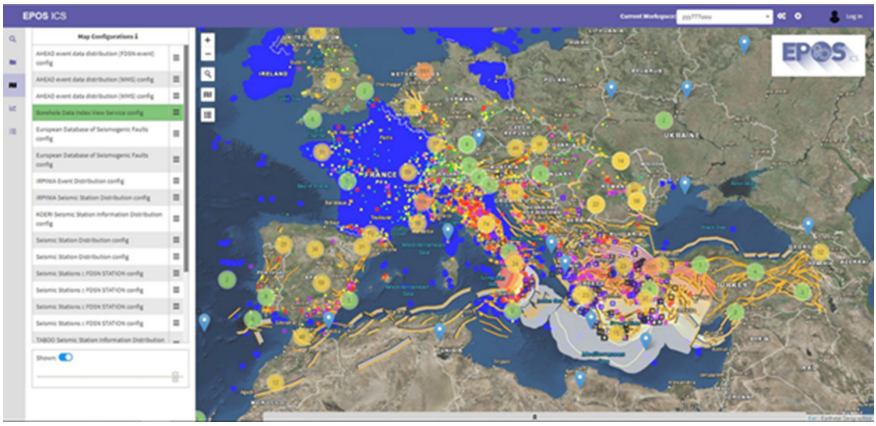

Fig. 7. EPOS user interface.

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

The integration of metadata from different domains within EPOS is accomplished by a matching/mapping/harvesting/conversion process: to date 17 different metadata 'standards' from the RIs within EPOS have been mapped. The mapping uses 3 M technology[48] (from FORTH-ICS[49]) 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 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, ISO19115/19139, 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 enrichment as the domains progressively provide richer metadata as needed for the processing they wish to accomplish.

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

The CERIF schema is documented[50] with a navigable model in TOAD[51].

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. OpenAIRE[52] uses the CERIF data model and it has influenced strongly the data model of ORCID[53].

The EPOS CERIF catalogue content has been loaded into an RDBMS at IFREMER which demonstrates portability and ease of set-up. 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 the metadata format based on DataCite and integrated with CKAN used at EUDAT for inclusion in the EUDAT B2FIND catalogue. Unfortunately, conversion from the B2FIND catalogue (based on CKAN) to CERIF is not possible because the records cannot be made available by the hosting organisation, largely due to resource limitations.

CERIF is natively FAIR since it supports all four aspects of the FAIR principles. Because of its referential and functional integrity, formal syntax and rich declared semantics CERIF is more machine-actionable than most metadata standards which usually require human intervention to interpret the metadata e.g. for the composition of workflows.

5.2 VRE4EIC and ENVRI

To prototype the use of CERIF as a joint catalogue service combining datasets from multiple RIs for use by a single VRE, a collaboration was established between ENVRIplus

[48] https://www.ics.forth.gr/isl/index_main.php?l=e&c=721.
[49] https://www.ics.forth.gr/.
[50] https://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/CERIF1.3_FDM.pdf.
[51] https://www.eurocris.org/Uploads/Web%20pages/CERIF-1.5/MInfo.html.
[52] https://www.openaire.eu/.
[53] https://orcid.org/.

and the VRE4EICproject. VRE4EIC concerned itself with the development of a standard reference architecture for virtual research environments, as well as the prototyping of exemplar building blocks as prescribed by that reference architecture. In particular, the project consortium developed VRE4EIC Metadata Service to demonstrate how data from multiple RIs might be harvested using a variety of protocols and techniques and then provided via a common portal. X3 ML mappings [12] from standards such as ISO 19139 [10] and DCAT to CERIF [13] were used to automatically ingest metadata published by different RIs to produce a single resource catalogue.

The VRE4EIC Metadata Service was developed in accordance with the e-VRE Reference Architecture [14], providing the necessary components to implement the functionality of a metadata manager as prescribed by the architecture [17]. The purpose of the resulting portal was to provide faceted search over a single CERIF-based VRE catalogue containing metadata harvested from a selection of environmental science data sources. The search was therefore based on the composition of queries based on the context of the research data, filtered by organisations, projects, sites, instruments, and people as shown in Fig. 8.

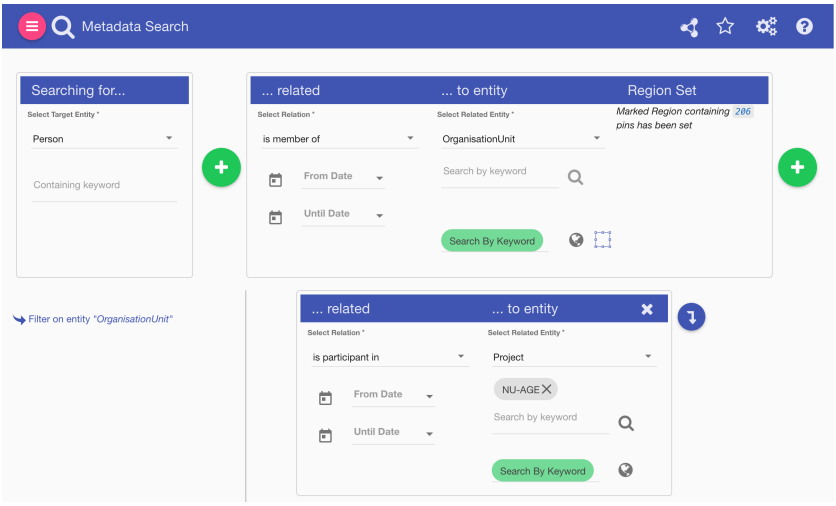

Fig. 8. The VRE4EIC metadata portal in action: searching for people that are members of an organisation which participated in the 'NU-AGE' project.

The portal (maintained at CNR-ISTI, Italy) supports geospatial search, export and storage of specific queries, and the export of results in various formats such as Turtle RDF and JSON. The CERIF catalogue itself was implemented in RDF (based on an OWL 2 ontology [15] using a Virtuoso data store[54], and was structured according to CERIF version 1.6[55]. Metadata harvested from external sources were converted to CERIF RDF

[54] https://virtuoso.openlinksw.com/.
[55] https://www.eurocris.org/cerif/main-features-cerif.

using the X3 ML mapping framework[56]; the mapping process itself was as illustrated in Fig. 9:

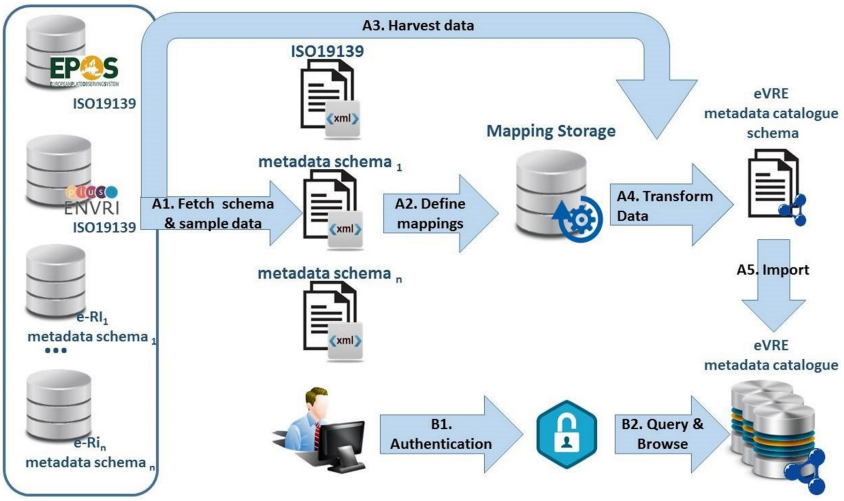

Fig. 9. e-VRE metadata acquisition and retrieval workflow: metadata records are acquired from multiple sources, mapped to CERIF RDF and stored in the VRE catalogue; authenticated VRE users then query data via the e-VRE.

1. Sample metadata, along with their corresponding metadata schemes, were retrieved for analysis. In addition to metadata from ENVRI and EPOS also records from CRIS (Current Research Information Systems which describe projects, persons, outputs, and funding) were harvested.
2. Mappings were defined that dictate the transformation of selected RDF and XML based schemas into CERIF RDF.
3. Metadata is retrieved from different data sources in their native formats, e.g. as ISO 19139 or CKAN[57] metadata (specifically as used in B2FIND within EUDAT in the context of ENVRI).
4. These mappings could then be used to transform the source metadata into CERIF format.
5. The transformed metadata was then ingested into the CERIF metadata catalogue.

Once ingested, these metadata became available to users of the portal, who could query and browse the metadata catalogue upon authentication via a front-end authentication/authorisation service. X3ML mappings were constructed using the 3M Mapping Memory Manager[58]. Among other functions, 3M supported the specification of generators to produce unique identifiers for new concepts constructed during translation of

[56] https://www.ics.forth.gr/isl/index_main.php?l=e&c=721.
[57] https://ckan.org/.
[58] https://github.com/isl/Mapping-Memory-Manager.

terms. Mappings into CERIF RDF were produced for Dublin Core, CKAN, DCAT-AP, and ISO 19139 metadata, as well as RI architecture descriptions in OIL-E.

The VRE4EIC Metadata Service demonstrated many desirable characteristics for a catalogue service, those being: a flexible model in CERIF for integrating heterogeneous metadata; a tool-assisted metadata mapping pipeline to easily create or refine metadata mappings or refine existing mappings; and a mature technology base for unified VRE catalogues. It was judged however that more development was needed in the discovery of new resources and the acquisition of updates through some automated polling/harvesting system against a catalogue of amenable sources. In this respect, RI-side services for the advertisement of new resources or updates to which a VRE can subscribe to trigger automated ingestion of new or modified metadata would be particularly useful.

A notable feature of CERIF is how it separates its semantic layer from its primary entity-relationship model. Most CERIF relations are semantically agnostic, lacking any particular interpretation beyond identifying a link. Almost every entity and relation can be assigned through a classification that indicates a particular semantic interpretation (e.g. that the relationship between a Person and a Product is that of a creator or author or developer), allowing a CERIF database to be enriched with concepts from an external semantic model (or several linked models).

The vocabulary provided by OIL-E (Chapter 6) has been identified as a means to further classify objects in CERIF in terms of their role in an RI, e.g. classifying individuals and facilities by the roles they play in research activities, datasets in terms of the research data lifecycle, or computational services by the functions they enable. This provides additional operational context for faceted search (e.g. identifying which processes generated a given data product) but providing additional context into the scientific context for data products (e.g. categorising the experimental method applied or the branch of science to which it belongs) is also necessary. Environmental science RIs such as AnaEE and LTER-Europe are actively developing better vocabularies for describing ecosystem and biodiversity research data, building upon existing SKOS vocabularies.

6 Future Directions and Challenges for Cataloguing

To demonstrate cataloguing capabilities a two-pronged approach was adopted.

Some records describing ‘data products’ were created from several RIs and ingested by B2FIND. This exposed the effort of metadata mapping but also the capability of a catalogue with metadata from different domains with unified syntax (but not necessarily unified semantics). This catalogue certainly demonstrated the potential for a homogeneous view over heterogeneous assets described by their metadata converted to a common format. However, the relatively limited schema used in EUDAT B2FIND means that some richness from the original ENVRI RI metadata records was lost.

Separately the EPOS metadata catalogue of services was used as an exemplar of the use of CERIF for integrated cataloguing, curation and provenance and via the associated VRE4EIC project the harvesting, mapping and conversion to CERIF of heterogeneous assets from multiple sources was demonstrated. Furthermore, CERIF provided a richer metadata syntax and semantics although - of course - if the source ENVRI RI catalogue had only limited metadata the full richness could not be achieved. There was some

investigation in VRE4EIC of enhancing metadata by inferential methods since the formal syntax, referential and functional integrity and declared semantics of CERIF lend themselves to logic processing.

The objective of these two parallel exercises was to allow RIs to see what can be achieved – and what effort is necessary - in the 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 be searchable in the production B2FIND portal. Within EPOS 17 different metadata formats had to be mapped and converted to be ingested into the CERIF catalogue and made available for (re-)use and in VRE4EIC further heterogeneous assets were added. The effort of correct matching and mapping between metadata standards should not be underestimated but – once achieved – can provide homogeneous access over heterogeneous asset descriptions and hence support a portal functionality allowing the end-user to gain interoperability.

As indicated by K. Jeffery (see Chapter 7: 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). The need for rich metadata is becoming generally accepted. As mentioned by authors from the EOSC Pilot project [16] "Minimum and common metadata is useful for data discovery and data access. Rich metadata formats can be complex to adopt, but have the advantage of making data more "usable" by both humans and machines".

It is planned to continue – in the ENVRI community - with the EUDAT B2FIND catalogue (maintained by EUDAT) and also to continue the work with CERIF (maintained by EPOS), anticipating the need for richer metadata than the B2FIND schema for at least some of the ENVRI RIs. CERIF already can handle the functionality associated with services – and other RI assets - as required in the EOSC (European Open Science Cloud). In particular, EUDAT/B2FIND is concentrated on datasets whereas the EPOS CERIF catalogue - while also handling datasets, workflows, software, equipment and other assets - initially concentrated on services to ensure alignment with the emerging EOSC. A mapping between CERIF and the draft metadata standard for EOSC services has been done.

The overall strategy is to make cataloguing technology available to the ENVRI RIs for them to choose how they wish to proceed, considering also other International obligations for interoperability which may determine particular metadata standards. This means that it is likely for the foreseeable future that ENVRI will need to support a range of metadata standards - among the RIs, internationally and also to align with general efforts such as schema.org from Google and associated dataset search - but that to interoperate them a canonical rich metadata schema will be required. The work is open to be shared among any in the ENVRI community who wish to avail themselves of the software, techniques and know-how.

Acknowledgements. This work was supported by the European Union's Horizon 2020 research and innovation programme via the ENVRIplus project under grant agreement No. 654182.

References

1. DIRECTIVE 2003/4/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 28 January 2003 on public access to environmental information and repealing Council Directive 90/313/EEC. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:041:0026:0032:EN:PDF. Accessed 04 Dec 2019
2. Riley, J.: NISO: Understanding Metadata (2017). https://groups.niso.org/apps/group_public/download.php/17443/understanding-metadata
3. Alemu, G., Stevens, B.: An Emergent Theory of Digital Library Metadata - Enrich then Filter. Chandos Information Professional Series. Elsevier, Amsterdam (2015)
4. Haslhofer, B., Klas, W.: A survey of techniques for achieving metadata interoperability. ACM Comput. Surv. (CSUR) **42**(2) (2010). http://eprints.cs.univie.ac.at/79/1/haslhofer08_acmSur_final.pdf
5. ISO 19115-1:2014: Geographic information—Metadata—Part 1: Fundamentals. ISO standard, International Organization for Standardization (2014)
6. DataCite Metadata Working Group. DataCite Metadata Schema for the Publication and Citation of Research Data. Version 4.1 (2017). http://doi.org/10.5438/0015
7. Parnell, P., et al.: Dublin Core: An Annotated Bibliography (2011). https://pdfs.semanticscholar.org/a614/cfb06d53ed8f0829370eab47bef02639f191.pdf
8. Erickson, J., Maali, F.: Data catalogue vocabulary (DCAT). W3C recommendation, W3C (2014). http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/
9. Wilkinson, M., Dumontier, M., Aalbersberg, I., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
10. ISO 19139:2007: Geographic information—Metadata—XML schema implementation. ISO/TS standard, International Organization for Standardization (2007)
11. Trani, L., Atkinson, M., Bailo, D., Paciello, R., Filgueira, R.: Establishing core concepts for information-powered collaborations. FGCS **89**, 421–437 (2018)
12. Marketakis, Y., et al.: X3ML mapping framework for information integration in cultural heritage and beyond. Int. J. Digit. Libr. **18**(4), 301–319 (2016). https://doi.org/10.1007/s00799-016-0179-1
13. Jörg, B.: CERIF: the common European research information format model. Data Sci. J. **9**, CRIS24–CRIS31 (2010). https://doi.org/10.2481/dsj.CRIS4
14. Remy, L., et al.: Building an integrated enhanced virtual research environment metadata catalogue. J. Electron. Libr. (2019). https://zenodo.org/record/3497056
15. W3C OWL Working Group: OWL 2 web ontology language. W3C recommendation, W3C (2012). https://www.w3.org/TR/2012/REC-owl2-overview-20121211/
16. Asmi, A., et al.: 1st Report on Data Interoperability - Findability and Interoperability. EOSCpilot deliverable report D6.3. Submitted on 31 December (2017). https://eoscpilot.eu/sites/default/files/eoscpilot-d6.3.pdf
17. Martin, P., Remy, L., Theodoridou, M., Jeffery, K., Zhao, Z.: Mapping heterogeneous research infrastructure metadata into a unified catalogue for use in a generic virtual research environment. Future Gener. Comput. Syst. **101**, 1–13 (2019). https://doi.org/10.1016/j.future.2019.05.076

18. Zhao, Z., et al.: Reference model guided system design and implementation for interoperable environmental research infrastructures. In: 2015 IEEE 11th International Conference on e-Science, pp. 551–556. IEEE, Munich (2015). https://doi.org/10.1109/eScience.2015.41
19. Chen, Y., et al.: A common reference model for environmental science research infrastructures. In: Proceedings of EnviroInfo 2013 (2013)
20. Martin, P., et al.: Open information linking for environmental research infrastructures. In: 2015 IEEE 11th International Conference on e-Science, pp. 513–520. IEEE, Munich (2015). https://doi.org/10.1109/eScience.2015.66