Ref. Ares(2017)532434 - 31/01/2017

D8.3
Interoperable cataloging and harmonization for environmental RI projects: system design

WORK PACKAGE 8 – DATA CURATION & CATALOGUING

LEADING BENEFICIARY: IFREMER

Author(s):	Beneficiary/Institution
Thomas Loubrieu, Frederic Merceur	IFREMER
Andre Chanzy, Christian Pichot[1], Damien Boulanger, François André[2], Maggie Hellström[3], Barbara Magnana[4], Abraham Nieva De La Hidalga[5], Zhiming Zhao, Paul Martin[6]	INRA[1], CNRS[2], LU[3], LTER[4], University of Cardiff[5], University of Amsterdam[6],

Accepted by: Keith Jeffery (WP 8 leader)

Deliverable type: REPORT

Dissemination level: PUBLIC

Deliverable due date: 31.01.2017/M21

Actual Date of Submission: 31.01.2017/M21

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

ABSTRACT

Short description of the document.

This technical document describes the functions required for a catalog system in the ENVRIplus context. It also provides architecture and design recommendations for the implementation of this system.

The catalogs host the descriptions of various entities, digital or physical (persons, equipment, datasets, provided services…). The catalogs are hence meant to enable discovery of the RI resources, as well as their curation, especially data curation, and provenance management.

Starting from the overall ENVRIplus information management framework in theme 2 (requirements, technical review, reference model and architecture), the deliverable specifically describes this catalog system implementation. From the overall context, priorities and options have been proposed to match the ENVRIplus schedule and funding constraints, and to optimally re-use the assets and expertise of Environmental RI and ICT partners.

The document is organized into three sections:

- Users, functional requirements and priorities
- Proposed architecture
- Detailed components and interfaces description

Project internal reviewer(s):

Project internal reviewer(s):	**Beneficiary/Institution**
Leonardo Candela	CNR/ISTI
Ingemar Häggström	EISCAT
Keith Jeffery	EPOS

Document history:

Date	**Version**
23.10.2016	Draft for comments.
28.10.2016	Keith Jeffery and Ingemar Häggström reviews integrated.
06.10.2016	Leonardo Candela review and Keith Jeffery add-on integrated.
27.11.2016	Review after ENVRIweek catalog workshop integrated.
12.12.2016	Review from Barbara Magnana, Chrsitian Pichot, Ingemar Häggström.

13.12.2016	Keith Jeffery minor edits.
19.12.2016	Review From Zhiming Zhao and Andre Chanzy.
25.01.2017	Review from Maggie Hellström and Abraham Nieva De La Hidalga. Alignment with latest D5.4 version.
30.01.2017	Keith Jeffery review.
30.01.2017	Paul Martin proof-reading

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the authors (Thomas Loubrieu, Thomas.loubrieu@ifremer.fr)

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 for 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

ABSTRACT 2
DOCUMENT AMENDMENT PROCEDURE 3
TERMINOLOGY 3
PROJECT SUMMARY 3
TABLE OF CONTENTS 5
1 Introduction 7
2 Users and functional requirements 8
2.1 Users 8
2.2 Functional requirements 9
2.2.1 Function Overview 9
2.2.2 General requirements for catalogs and metadata 11
2.2.3 Requirements for item-specific catalogs (data services, acquisition services …, physical samples…) 15
3 Proposed architecture 28
3.1 Architecture principles 28
3.2 Flagship data service catalog architecture 30
4 Summary 31
5 Appendices: Detailed components and interfaces description 32
5.1 Flagship Data Service Descriptions: Encoding guidelines 32
5.1.1 ISO19139 32
5.1.2 Dublin-Core 36
5.1.3 DCAT 38
5.1.4 Schema.org 38
5.1.5 Other recommended encoding, openAIRE, dataCite 39
5.1.6 OIL-E and CERIF mapping 39

Interoperable cataloguing and harmonization for environmental RI projects: system design

1 Introduction

For ENVRIplus, an interoperable catalog system aims at organizing the maintenance and access to descriptions of resources and outcomes (datasets, equipment, persons…) of multiple RIs in a framework which implements a number of functions on these descriptions. As defined in the Reference Model, maintenance of a catalog is a strategic component of the curation process and the descriptions maintained in the catalog support the acquisition, publication and use of data. The system hence must provide to users a function for seamless **discovery** of the description of resources in the RIs, encoded using a standardized format. The multi-RI context of ENVRIplus implies that, in addition to the descriptions usually available within each RI, resources may also have to be described at a higher granularity so to provide context. It must as well support data **curation** activities and guarantee the long term preservation of the data collected in an RI, for example by enabling identification of datasets stored in obsolete format. Finally, the catalog should document the quality of each RI's products by managing the **provenance** and lineages of the resources, from sensor calibration events or laboratory sampling procedures to scientific publication—one should be able to eventually answer the question: "does the data result which led to the conclusion of this paper come from properly calibrated sensors?"

The technical proposal described in the current document is elaborated from two major drivers:

- The theme 2 framework established in WP5 (Reference model guided RI design).
- Short term challenges for ENVRIplus catalog system and partnership in task 8.2 with assets and expertise provided by RI and ICT partners.

Theme 2 provides foundation for the current activity, as follows. Deliverable 5.1 (A consistent characterisation of existing and planned RIs) is useful to analyse assets and requirements of the environmental RI in the catalog scope. The document also provides a technical review in this field. Deliverable 5.4 (A development plan for common operations and cross-cutting services based on a network of data managers and developers) provides a framework for the technical architecture where the catalog system is embedded. At last, the Reference Model, developed in ENVRI and improved in ENVRIplus contributes to control and validate the completeness of the catalog functions in the science data lifecycle.

In the context of ENVRIplus, with a bounded time scale and limited resources, the strengths and assets of environmental RI and ICT partners involved in task 8.2 and the short term challenges identified for the catalog system also drive our work. Especially, the short-term challenge for the catalog system in ENVRIplus is to be proven useful and relevant by the RIs for handling faithful descriptions of their RI's own resources (datasets, equipment, etc.). This is especially critical because the contribution and expertise of each RI is required to populate the catalogs with their resource's descriptions at the expected quality and completeness level.

At last, complementary and usefulness of the catalogs for the use cases developed in WP9 need to be considered, although the diversity and poor maturity in this field poses challenges.

2 Users and functional requirements

2.1 Users

The users targeted for the ENVRIplus catalog systems are:

- Users outside the RI, **researching data science**: RIs are keen to benefit from ENVRIplus to make their products (datasets, experiment infrastructures…) useful beyond their traditional targeted users in their community. The development of data science raises new trans-disciplinary use cases for the RI's datasets. The "How do mosquito-borne diseases emerge and what are trends?" use case developed in WP9 (Service validation and deployment) is one good example of trans-disciplinary research supported by ENVRIplus.
- Users inside the RI, such as **data managers, coordinators, and operators**: pre-existing RIs have developed expertise and assets to manage their own resources. ENVRIplus benefits from these pre-existing features and the ENVRIplus development needs to nicely complement the RI assets and preserve the stable procedures already run by RI operators in their field of expertise. The RIs, depending on their status, have mentioned a number of functions they would like to develop or improve together in ENVRIplus. The ENVRIplus catalog system should contribute to some of them (data curation and citation, provenance and standardization of registries of equipment).
- The **stakeholders, decision makers** and funders of the RI need also to have a broad picture of the RI resources in the European landscape to control their efficiency and complementarity.

From the task 8.2 partnership and participation of additional RIs in the "IC_8 Catalog, curation, provenance" use case promoted by WP9 (Service validation and deployment), the following RIs have been targeted as first priority to have their resources described in ENVRIplus catalog system. Each has its own pre-existing catalog and data management infrastructures:

- ANAEE: see the discovery function for ANAEE France at http://w3.avignon.inra.fr/geonetwork_anaee/srv/eng/catalog.search#/home
- Euro-ARGO: see the data access service at http://www.euro-argo.eu/Activities/Data-Processing/Access-to-data
- EMBRC: see the service catalog at http://www.embrc.eu/services
- EPOS: see the service catalog at https://www.epos-ip.org/data-services/community-services-tcs/seismology
- IAGOS: see the data services at: http://www.iagos.fr/web/rubrique3.html
- ICOS: see https://www.icos-cp.eu/
- LTER: see the repository for research sites and datasets DEIMS at https://data.lter-europe.net/deims/
- SeaDataNet: for example the catalog of data products: http://sextant.ifremer.fr/en/web/seadatanet

2.2 Functional requirements

2.2.1 Function Overview

From Description of work, the catalogs are required to support the following functions:

- **Publication** of metadata to enable discovery of the resources descriptions (provided services, datasets, sensors, persons, software, documentation, …)
- Support **curation** of resources (dataset, software, sensors…) by enabling their maintenance, usage scheduling and preservation. The catalog should actually support acquisition by supporting the curation of its contributing components (e.g. sensors), beyond its outputs (e.g. datasets, samples). This is the most efficient and trustful way of maintaining up to date descriptions of components which are critical for the reproducibility of research and provenance.
- Support science activities by registering usage and processing of the resources, by citing and qualifying the results with **provenance** or implementing solution other specific use cases.

From the analysis provided by D5.1, section 2.3.3, the resources which are considered for cataloguing are:

1. Environment Research Infrastructure specific items:
   - Integrated data services (collections, synthesis…) and individual datasets (single observation time series, images…)
   - Acquisition services and systems (devices, sensors, equipped sites, networks…)
   - Features of interest
   - Processes (Software, workflow, web services)
   - Physical samples
2. Generic items:
   - Activities (Observation, processing, usages)
   - Contacts (persons, teams, organizations)
   - Reference terms & definitions (thesauri)
   - Documents (Publication, procedures)

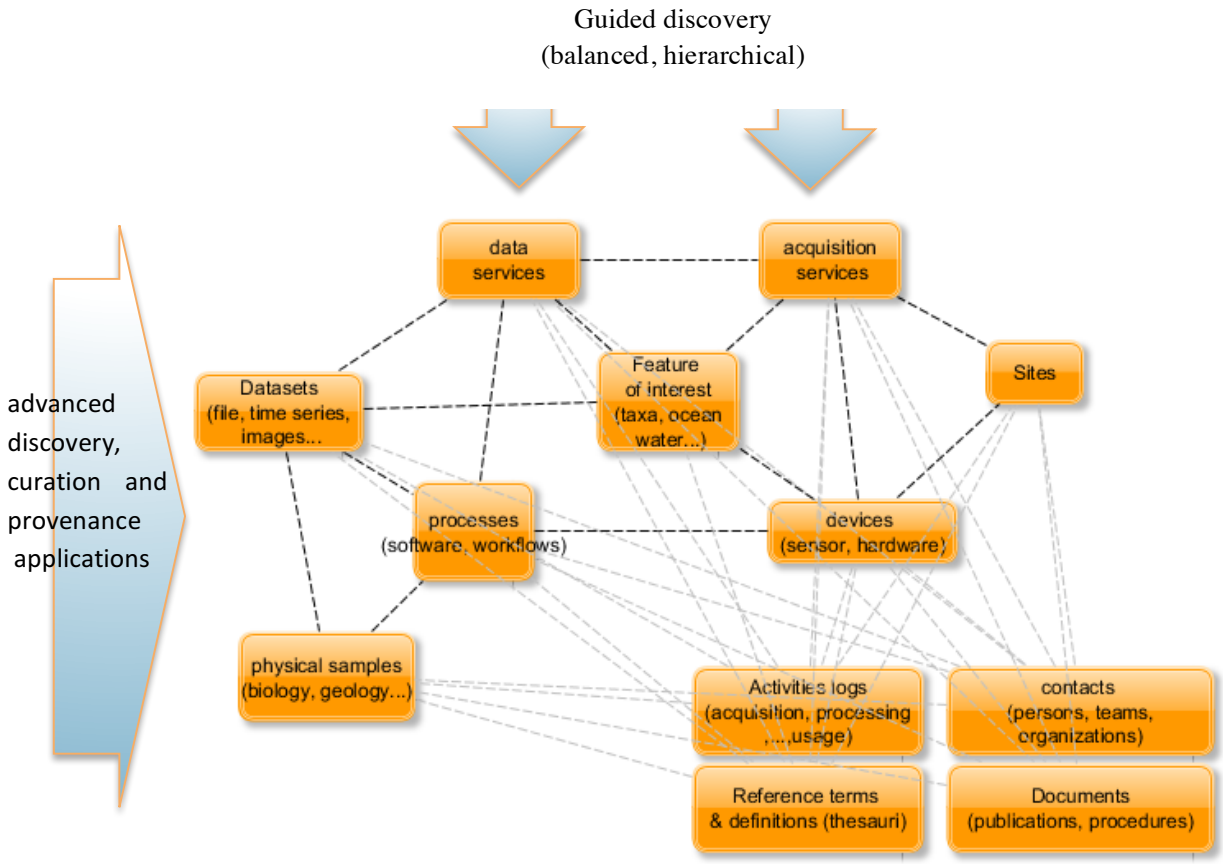

FIGURE 1: CONCEPTS DESCRIBED IN CATALOGS, LINKS BETWEEN THEM (BLACK) AND WITH GENERIC REFERENCE CONCEPTS (GREY)

Any resource described in the catalogs, as basic metadata elements, should fulfil a number of **general requirements** (creation, update, curation and publication) which are detailed in section 2.2.2. These recommendations should be used as general guidelines for Research Infrastructures to provide catalogs which will be suitable for the use cases developed in WP9 (provenance, sensor registry, virtual research environments…) and for the discoverability of ENVRIplus resources.

Then depending on the concepts described in catalogs (see Figure 1), some **specific requirements** should be fulfilled. However, the detailed resource descriptions (datasets, sensors, samples…) in each RI are too complex, inconsistent in their specificity, or not mature enough. The science or operations expectations at a cross-disciplinary level are not well defined yet. This situation prevents us from attempting to provide a comprehensive specification for each detailed catalog within ENVRIplus project. However a strategy for the development of cross-disciplinary solutions and ensuring the most efficient outcome for ENVRIplus catalogs implementation is provided at section 2.2.3.

2.2.2 General requirements for catalogs and metadata

The metadata management should fulfil the following requirements. They are classified into 3 categories:

- Create and update metadata
- Curate metadata
- Publish metadata

This section is intended to converge with the "metadata" object recommendation given in the Reference Model – information viewpoint when refined.

The reference model defines metadata as "Data about data, in scientific applications is used to describe, explain, locate, or make it easier to retrieve, use, or manage an information resource."

In the context of this document we attempt to be more concrete on metadata. Metadata can be actually about data (i.e. datasets or data services). In addition, to properly describe the data, the metadata should also describes physical entities or resources involved in the processes of acquisition, processing or usage of the data, for example sensors, computation systems, persons, physical samples…

So, in the following section, we are considering metadata as either description of datasets or data services or description of physical entities or resources involved in the process of creating or using the data as detailed in Figure 1.

The following sections provides guidelines for management of these metadata. These best practices have been formalized as requirements.

2.2.2.1 Create and update the metadata

Expertise closeness

The person, team or system in charge of the resource should create or update its description.

Temporal closeness

The creation or update of the resources description should be done at the most appropriate time regarding the actual activity described. It is for example counter-productive to edit the description of a sensor deployment when the dataset is archived in a long-term data centre, after scientific paper publication, years after the observation has been done.

As a consequence, it is useful to re-use content from resource management information systems (for example sensor maintenance database, operation logger on field…) to automatically feed the catalogs.

2.2.2.2 *Curation of the metadata*

In a distributed environment, the information and specifically the catalog contents can be replicated in various repositories. To properly preserve the information while allowing free and flexible access and use of it, a few simple rules should be applied.

Two types of repositories, with 2 different sustainability mode can be considered:

- Application information systems (for acquisition, processing or publication)
- Long-term archival

The application (acquisition, processing or publication) information system sustainability is related to the sustainability of the application system infrastructure, for example the observatory. For example, the ENVRIplus flagship service catalog has the same sustainability as the ENVRIplus project (4 years). A virtual research environment should be also considered as an application information system.

When required, the long-term archive provides a support for sustaining "forever" the information.

The workflow for the metadata should comply with Figure 2.

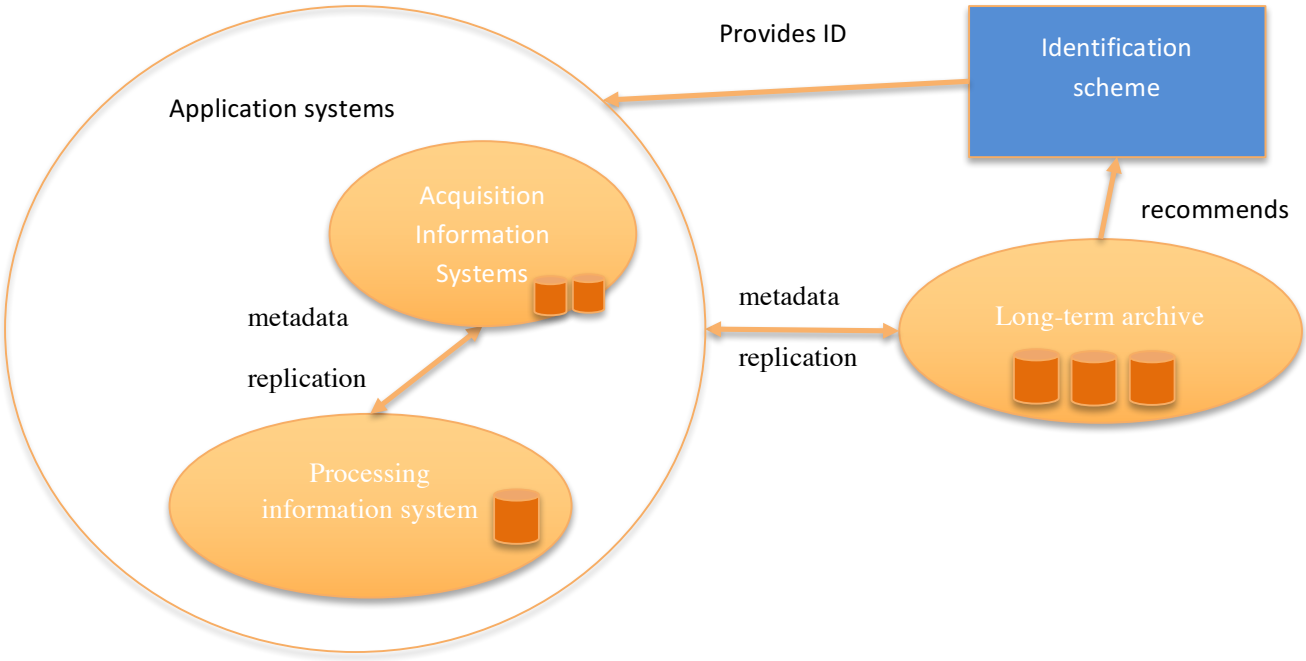

FIGURE 2: CATALOG REPOSITORIES INTERACTING WITH EACH OTHER

As written in 2.2.2.1 the metadata should be created as close as possible in time and expertise to the described resource or activity (sensors, processing).

Permanent identification

The identification scheme used by information systems creating new metadata should provide commonality in a distributed environment where different operators assign identifiers for same types of resource (uuid, pid….).

The primary identification scheme should comply with long term sustainability requirements for the resource description. The same unique identifier should be kept as a label on its description, from the creation of the description, across all the different repositories where the description is replicated. This is mandatory for proper duplicate management.

Additional secondary identifiers might be added to the description during the replication. They are useful to identify with codes the resources in a specific context, for example station or sites in a given network.

Replication for optimal sustainability

The metadata should be replicated with as little brokering or translation as possible to avoid risk associated with brokering or translating information by individuals who are not experts on the resource in question.

The metadata should be replicated to enable good availability of the information at every time scale, within the application sustainability period or for the long term.

Cross-links for dedicated repositories within the same sustainability scale

Within one sustainability time scale (acquisition system, processing or long-term) links or external references can be used to cross different repositories of catalogs with different scopes (sensors, datasets…).

Example: The Euro-ARGO RI has a catalog for profiling float platforms. Each platform should be identified with a unique identifier for sensing devices (e.g. PID or UUID, taking into account guidelines proposed in WP6). When the platform metadata is replicated in processing and publication systems (Copernicus MEMS, EMODNET) or long-term archive (SeaDataNet), the identifier is kept for traceability and duplicate management in a wider scope.

2.2.2.3 Publication

Non-public information

Information available in catalogs might be restricted or simply useless outside (or in some circumstances even within) the RI. In these cases, limited publication (with Authentication and Authorization Infrastructure access control) or no publication of the catalog can apply.

Machine and human interfaces

The catalogs should be discoverable (using filters or lists) via machine and human interfaces.

User friendly and shareable

The human interface should be user-friendly and resources should be reachable with shared URLs.

Search-engine optimization

The user-friendly resource description page should be optimized for general public search engines. The http header attributes (for example max-age) should be used carefully to properly propagate the updates.

Citable resources

For provenance management or simply acknowledgement of contributing resources, citation guidelines should be proposed together with the resource description in the catalog. The proposed citation should use stable PID (e.g. DOI) for identification. They resolve to a landing page describing the cited resource. This requirement, listed for data in Computational Viewpoint of the Reference Model, can be extended to other types of descriptions (large equipment, observation program…). The citation of these resources is a strategic necessity for RIs which must report on their impact on their respective communities.

FIGURE 3: "HOW TO CITE" SECTION IN A DATASET DESCRIPTION (HTTP://WWW.SEANOE.ORG)

Community and content-agnostic standards

The machine interface should implement community standards for the resources described in the catalog.

The machine interface should also implement one content-agnostic protocol and format (auto-descriptive), generic for every type of resource. This will allow semantic linking between different concepts in managed in different catalogs.

Actionable descriptions

The resource description should be actionable by providing references to linked resources or services enabling the user or applications to actually access the described resource. For example visualization or download access from dataset description.

2.2.3 Requirements for item-specific catalogs (data services, acquisition services …, physical samples…)

2.2.3.1 Strategy for ENVRIplus catalogs development

Among the concepts described in the catalogs, we can consider the individual resources (datasets, persons, processes…) and the resources aggregated into services (data and acquisition).

The RI's data services (observation collections, synthesis…) and acquisition services (integrated observatories, observation networks…) need to be shown in a synthetic catalog which will display and promote, beyond their traditional communities, the high quality services provided by the RI.

In addition, the individual resource descriptions need to be compliant with recommended standards which should be harmonized across Research Infrastructures so that trans-disciplinary discovery, curation and provenance tools or technology driven demonstrations (e.g. the theme 2 catalog of solutions) are applicable across the cluster of RIs.

Thus two main targets are identified for the ENVRIplus catalog system. They both participate in a complementary top-down and bottom-up approaches:

Top-Down

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 will also be a tool for stakeholders and decision makers to oversee and evaluate the outcome and complementarity of RI data products.

The top-down approaches can be compared, in news business to traditional newspapers where an editorial policy selects the news to meet public expectation, with a certain amount of accepted subjectivity.

FIGURE 4: TRADITIONAL NEWSPAPER WITH TOP-DOWN "MANNED" EDITORIAL POLICY

The task8.2 will design and develop a “flagship data and acquisition service catalog” which will provide this top-down view and entry point for hierarchical discovery of the ENVRIplus RIs’ resources.

This development will support Use Case IC_8 (catalogs) in WP9.

Bottom-Up

This approach will provide fine-grained catalogs which will facilitate the deployment of the Theme 2 solutions (for example for provenance management) throughout the interested RI. This can also enable fine grained discovery and development of trans-disciplinary applications to browse resources across the RI.

The detailed resources description (datasets, sensors, samples…) in each RI, is too complex and sometimes too specific to expect a comprehensive specification for each to be produced within the ENVRIplus project. It will also challenge the capabilities of aggregating applications to provide relevant results out of these heterogeneous inputs.

The bottom-up approach can be compared, to the use of social media by news businesses where automated algorithms aggregate the news to meet individual expectations with a certain amount of accepted heterogeneity in results (verified and un-verified news, advertisements…).

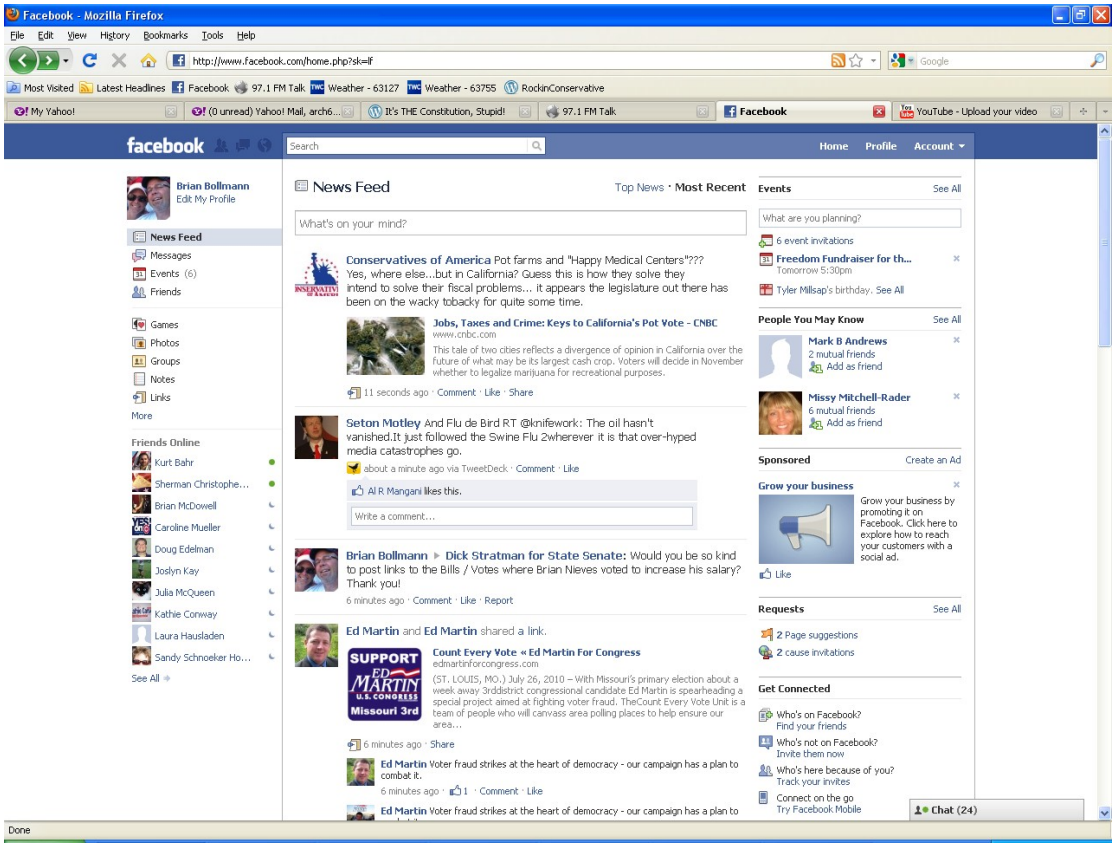

FIGURE 5: FACEBOOK WALL WHERE INPUTS ARE AGGREGATED BY ALGORITHMS DEPENDING ON PERSONNAL CONFIGURATION OF THE USER

In the short term, the fine-grained catalogs are useful and can be used as back-end for software solutions proposed in theme 2 (see use cases in WP9, for example TC_4: sensor registry). The specific fine-grained catalogs will then be designed and developed depending on these use cases.

Besides the general requirements given (see 2.2.2), the current document will not give further requirements on fine-grained catalogs. Actually maximum flexibility and reactivity in the development of these catalogs is desirable to fulfil the requirement of the different use cases.

Nevertheless, for resources such as contacts, documents and keywords which are useful for describing flagship resources (see 2.2.3.2), some requirements are given in section 2.2.3.4.

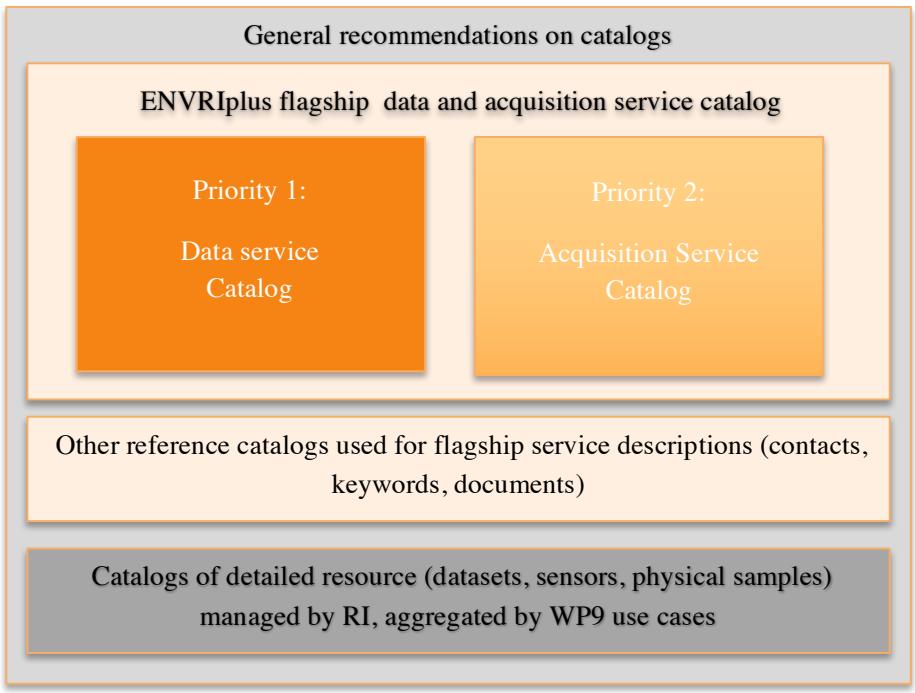

FIGURE 6: STRATEGY FOR ENVRIPLUS CATALOG DEVELOPMENT

2.2.3.2 Flagship data and acquisition service catalog

Data Services

As seen in D5.1 technical review, The maturity of RIs regarding the management of their data services in catalogs, the level of standardization in this field (ISO19XXX series, Dublin-Core…) and availability of homogeneous technical solutions partially adopted by the RI (Geonetwork, CKAN, CERIF) enables the development of an integrated service for flagship data product discovery by the end of the ENVRIplus project. To streamline the implementation of the data product catalog, it has been decided to concentrate on so-called flagship data services which are the datasets the RIs want to first communicate on. Examples of flagship data services are, "historical data collection of quality assessed water column temperature and salinity in Mediterranean Sea" for SeaDataNet and "ARGO profiling float observations general data assembly portal" for Euro-ARGO.

Acquisition services

However, RIs do not only produce datasets, but also provide services on top of platforms, infrastructures or sites organized in networks. These well-maintained infrastructures are organized to host experiments for scientists sometimes called principal investigators (PIs). For example ANAEE provides a suite of experimental equipment from ecotrons to *in natura* sites where researchers can carry out experiments and collect results. ARGO provides a fully integrated network of ocean drifting and profiling platforms with data management—researchers focused on ocean properties or location can benefit from this infrastructure, for deployment operations and data management. To complement the exposition of data services in the catalog, the publication of descriptions of acquisition services is required. As the current level of maturity and adoption of

standards regarding the management of infrastructure descriptions is not as good as for datasets yet, this task is being considered as a second priority.

Note that the flagship data and acquisition services will be complemented with a science and engineering view point, the synthetic RI description already provided in the ESFRI roadmap with a management and organizational view point (see Figure 7).

FIGURE 7: ESFRI SYNTHETIC RI CATALOG WITH ANAEE FACT SHEET[1]

Detailed functional requirements are given in the following sections.

2.2.3.3 ENVRIplus Flagship data and acquisition service catalog

2.2.3.3.1 Flagship data service catalog

1. Introduction

The section aims at defining the requirements for the description repository of ENVRIplus RI flagship data services.

Example of products are:

- Historical data collection of quality assessed water column temperature and salinity in Mediterranean Sea (data service, SeaDataNet).

- ARGO profiling float observations general data assembly portal (data service, Euro-ARGO).

[1] https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri-roadmap, https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=mapri_european

The granularity of data services considered here is coarse; to balance RI visibility, each should provide 5 to 20 data services. The data services described in this catalog must have reached the published status as defined by the Reference Model (See Data Publishing in Research Data Lifecycle).

The functions provided by the repository are:

- Enable **discovery** of resources in English interface. Multilingualism is not supported in the initial version.
- Enable **curation** by providing information regarding the preservation of the data (storage format, host).
- **Provenance** of the dataset should also be covered. However, the task 8.3 which defines the requirements for provenance had not started when this document was written.
- **Evaluation of descriptions** will be useful for RI to evaluate the quality of their contributions (e.g. missing fields, broken reference links…). This will as well help external users to quickly evaluate the result they got from their catalog request.

The following sections define, for the flagship data service catalog:

- The detailed data service description required for discovery and curation.
- The evaluation of description function.
- The request criteria capabilities.

2. Data Service description

The product description content (the fields or attributes) is defined here after.

External references for vocabularies and contacts are required, guidelines for these reference information are given in 2.2.3.3.2.

The requirements on product description are defined by adopting those of the RDA metadata interest group [https://rd-alliance.org/groups/metadata-ig.html]. This list is kept simple on purpose. For this initial version of the catalog, we need to quickly achieve consensus between RIs which prevent us from considering sometimes important details (for example depth or altitude coverage). The European INSPIRE directive on geo-spatial data discovery (based on ISO19115 conceptual model, see http://inspire.ec.europa.eu/documents/Network_Services/TechnicalGuidance_DiscoveryServices_v3.1.pdf) is also considered.

When necessary they are adapted for the ENVRIplus context.

RI

The research infrastructure responsible for the current data service.

For example: ICOS

Id

PID for the current data service resolving to a landing page. Ideally, a DOI (digital object identifier).

For example: http://dx.doi.org/10.12770/a61129f0-afbc-4bfa-8307-00f37d37d98a

Title

Title in English for the data service. The rule advised by W3C for a good title in an HTML page can be applied: https://www.w3.org/QA/Tips/good-titles.html

It especially helps to show well the results produced within a search engine.

For example: North Atlantic Ocean - Temperature and salinity observation collection V2

Abstract

Description of the current product with no layout directive except carriage return.

Quicklook

(Added to RDA recommendation for user friendliness of the catalog.)

URL to an image illustrating the current data service.

The image should be shown in thumbnail version with height and width dimensions in between 130 and 180 pixels.

Keywords

Free keyword or keyword with a definition provided in a vocabulary service as an URI as recommended in section 2.2.3.4.1.

Keywords might encode a wide range of information:

- Research sites
- Typology of data services (raw observation, spatio-temporal synthesis…)
- Temporal resolution (single measurement, hourly, daily…)
- Spatial sampling geometry (single point, transect sampling horizontal…)
- Observed properties (temperature, pressure…)
- Feature of interest (atmosphere, sea water, biota…)
- Status (nominal, obsolete, superseded)

Temporal window

Period during which the data is applicable for. When the data service is continuously updated with new observation data, the « end » boundary is left blank.

Encoding is ISO8601.

For example:

Start date: 1900-01-01

End date: 2013-12-31

Geo-spatial bounding box

In WGS84, westernmost and easternmost longitude (-180, 180 or 0,360 for coverage cross meridian 180) and southernmost and northern most latitude (-90, 90) where the data is applicable.

For example:

Westernmost Longitude: 27.50

Easternmost Longitude: 42.00

Southernmost Latitude: 40.50

Northernmost Latitude: 47.50

Although not strictly required for cost-effectiveness purpose, a more flexible encoding of the spatial coverage of the dataset will much better represent the actual fitness of the dataset. The various geometries of the dataset single point, multi-point, poly-line (trajectory) or polygon are indeed loosely synthetized as a bounding box. GeoJSON or GML encoding provide this flexibility and are supported by CKAN; as much as possible this feature will be implemented.

The vertical coverage is also very important to document data services. However, the various coordinate reference systems in this field makes it challenging to manage numerically at a trans-disciplinary level. The keywords and especially feature of interest (water column, sea bed or land surface, solid earth, atmosphere… see **Keywords**) might be an interesting alternative to document the vertical coverage of data services.

Location

URL where the data service is curated. Generally a landing page.

Format

Format as the product datasets is curated. The format should be listed in the MIME/TYPE reference list. Either as standardized by IANA (http://www.iana.org/assignments/media-types/media-types.xhtml) or shared by a community (e.g. application/x-netcdf).

For example: application/x-netcdf

Contacts

Contacts relevant for the current resource. The contacts have roles and identifiers (as PID).

The URI of the PID must comply with recommendation in section 2.2.3.4.2.

The roles for data services are as defined in [ISO19115:2003], among:

	Name	Domain code	Definition
4.	owner	003	party that owns the resource
5.	user	004	party who uses the resource
6.	distributor	005	party who distributes the resource
7.	originator	006	party who created the resource
8.	pointOfContact	007	party who can be contacted for acquiring knowledge about or acquisition of the resource
9.	principalInvestigator	008	key party responsible for gathering information and conducting research
10.	processor	009	party who has processed the data in a manner such that the resource has been modified
11.	publisher	010	party who published the resource
12.	author	011	party who authored the resource

FIGURE 8: CONTACT ROLES AS DEFINED IN ISO19115

For example: originator

Related documents or resources

Related documentation (user manual, quality report, scientific paper…) should be listed with PIDs (ideally DOIs) and free-text short description.

Other resources, such as web services, web API, data visualization or sub setting portal should be listed with URL and free text short description.

For example:

http://archimer.ifremer.fr/doc/00153/26387/24482.pdf

Availability, licence

Terms of use of the data services are described in a document or licence which can be generic (e.g. creative commons) or specific. In both cases, the document is cited through a canonical and stable URL. The commitment of the RI or e-infrastructure to maintain the availability of the data service should be described in the term of use documentation. This can be done as a Service Level Agreement as recommended in ITIL (ISO20000) for operational IT services.

In a later version of the ENVRIplus Information System, with connection to AAI implementation, the list of users having accepted or obtained a license could be registered, in a licence catalog or in the AAI.

For example:

http://creativecommons.org/licenses/by/4.0/

or http://www.seadatanet.org/Data-Access/License/1.0

Status and update date
(Added to RDA recommendation so to enable sorting with latest update.)

The date when the specification of the data services has been updated. Note that if a product is a data service which data is continuously updated (e.g. in real time), the update date is not the date of the latest data submission but the date of the specification update for the data flow (e.g. new parameter, new quality control, …).

For provenance purposes, every dataset published once should be preserved and kept available as well as its description in the catalog. However for several legitimate reasons (storage volume, desire to prevent erroneous re-use etc.), the data itself might be made deleted or made unavailable. Therefore, together with the update, a status (managed in **Keywords**) should describe if the data service is obsolete or superseded. A comment in the abstract can give details on superseding products and motivation.

The date must be encoded in ISO8601.

For example: 2016-02-08T12:56:00

3. Evaluation function

The evaluation function aims at measuring the quality of the description (metadata). Unlike traditional checkers, which give a yes/no status or a number of errors or warning on the metadata record, the objective is to "continuously" quantify the quality of the description encoded in the metadata record.

This evaluation will introduce some flexibility in the metadata ingestion process of the catalogs: any record can be considered and ingested if they meet the standard encoding (ISO19XXXX, Dublin-core…). The measurement of the quality of the record then is useful with respect to two perspectives:

- The external users will be able to sort their request results according to the quality of the records as Google does for web pages.
- The metadata record provider will be encouraged to improve the quality of his/her contributions.

The proposed evaluation criteria are:

- Completeness of the description (%) regarding the attributes expected. Specific thresholds can be considered in the quality of the record with the following statuses: data curated, visualization service available, direct data access available, data cited.
- Accuracy of the description: the usage of external references as valid resolvable URL should increase the value of a completed attribute.

The evaluation function will need to be proof-tested against real metadata records and validated afterward.

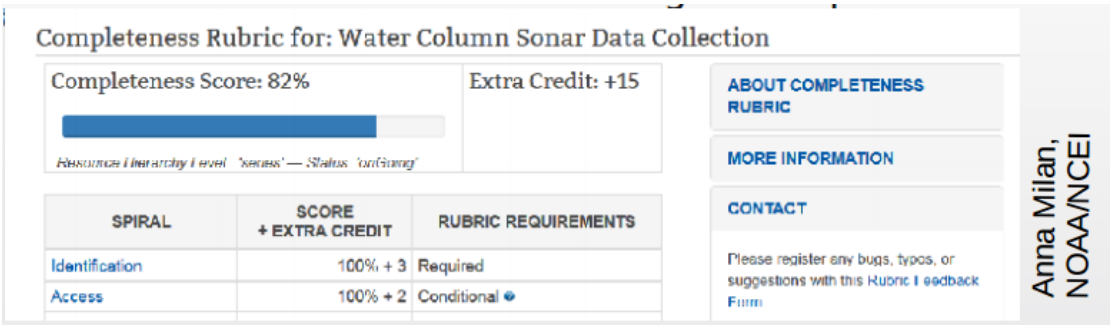

FIGURE 9: EXAMPLE OF METADATA EVALUATION DONE IN US/NOAA DATA CENTRES

4. Search Request criteria

For the first version, simple request criteria should be supported:

- Spatio-temporal criteria : x,y,z,t
- Keyword or full text
- Update date criteria (for latest updates)

The request only apply to metadata and not on the data content (observed quantities or observed objects).

2.2.3.3.2 Flagship acquisition services catalog

Although the maturity level for managing acquisition service catalogs homogeneously at cross-disciplinary level is not as good as for data services, some requirements can be described.

A **flagship acquisition services catalog** is useful for RIs to communicate on their services, beyond their traditional research community.

SOME RI ARE INDEED NOT DESIGNED FOR PROVIDING AN INTEGRATED DATA SERVICE BUT RATHER HIGH LOGISTICS AND SUPPORT TO HOST RESEARCH EXPERIMENT ON EQUIPPED SITES. TO COMMUNICATE ON CAPABILITIES, SOME RI ARE ALREADY PROVIDING ACQUISITION SERVICES DESCRIPTION SERVICES IN

CATALOGS (SEE

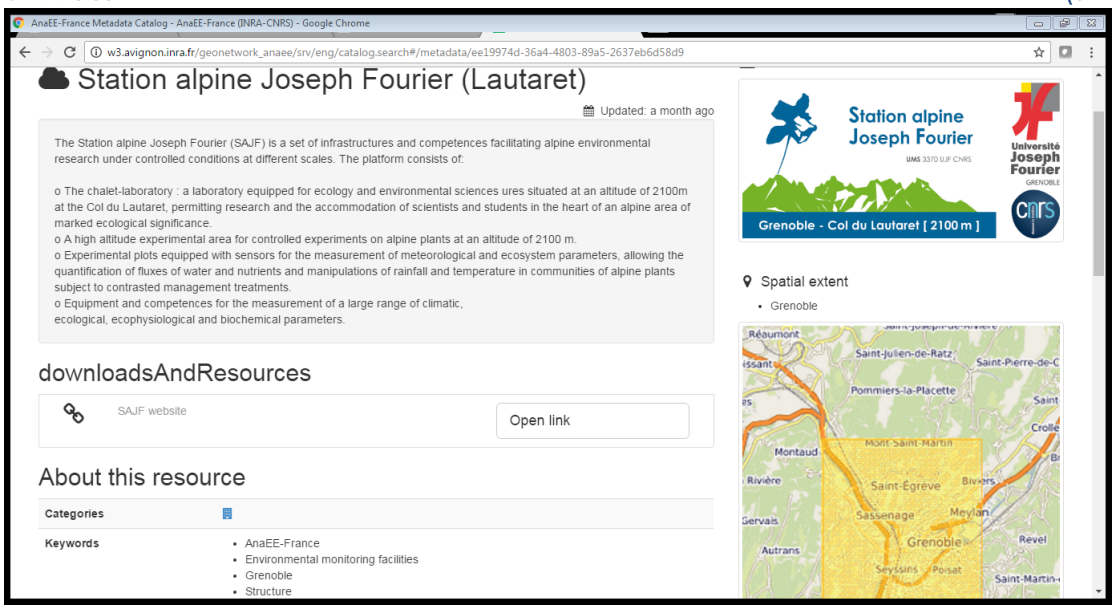

Figure 10).

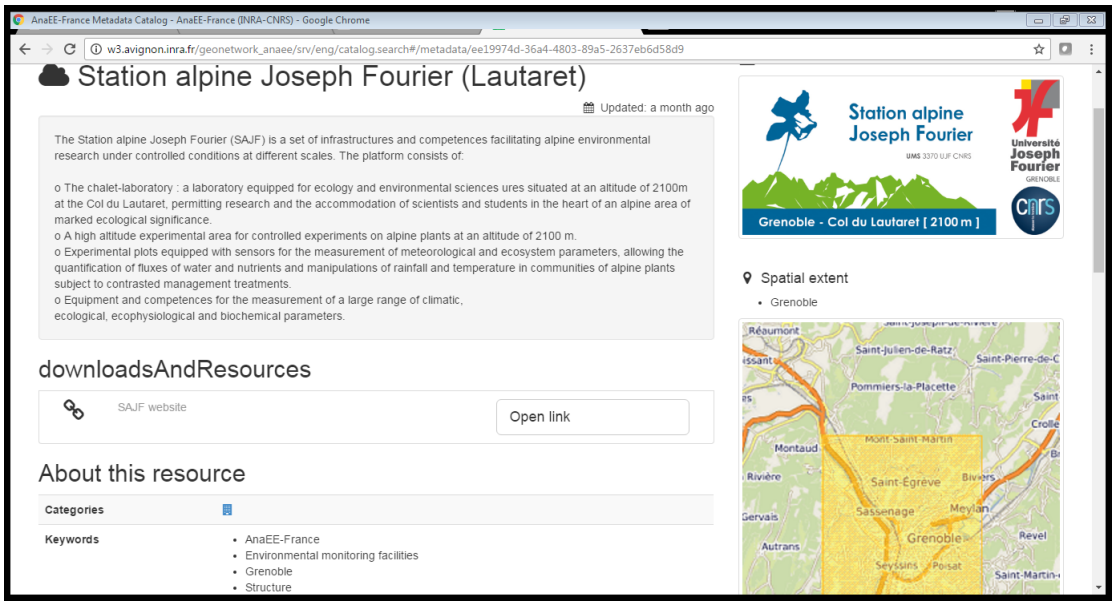

FIGURE 10: STATION ALPINE JOSEPH FOURIR (LAUTARET) ACQUISITION SERVICE FROM ANAEE RI

The acquisition service description should contain the following attributes:

- RI
- Identifier
- Title
- Abstract
- Quicklook
- Keywords
- Location(s)
- Contacts
- Related Document&resources

2.2.3.4 Reference catalogs: recommendation for use

In two specific cases, for dedicated attributes, values are references to external catalogs. The following sections give recommendations:

For usage of controlled vocabularies, enumeration for classification, see section 2.2.3.4.1.

For citation of persons, teams of organizations, see section 2.2.3.4.2.

2.2.3.4.1 Controlled vocabularies, enumerations

Keywords should be encoded as labels and as much as possible as links (URI) to well-maintained, stable references. The systems curating these references must manage deprecated and superseded operations of the terms without threatening the permanence of the references.

These keywords can be linked to full ontologies linking concepts together but simple lists of well-defined terms are also acceptable.

The reference system should provide multilingual labels and definitions.

The system should also support content negotiation so to show human or machine readable interfaces depending on the requesting client. The content negotiation principles are detailed in https://www.w3.org/TR/swbp-vocab-pub/#negotiation.

Examples of vocabulary services are:

- NERC Vocabulary services: http://www.bodc.ac.uk/products/web_services/vocab/
- EU-FP7 project ESPAS: http://espas.spaceweatherservices.com/index.php
- EnvThes: http://vocabs.ceh.ac.uk/evn/tbl/envthes.evn

For encoding these reference catalogs, SKOS (https://www.w3.org/2004/02/skos/) or ADMS can be considered (https://www.w3.org/TR/vocab-adms/).

2.2.3.4.2 Contacts: Persons, teams and organizations

This proposition is based on the OpenAIRE Guidelines[2] for Data Archives.

1. Name

The name of the contact (occurrences: 1): could be a name of a person or the name of a team, an organization, a project, etc.

For example:

- John Taylor
- Ifremer
- Argo

2. Email

The email of the contact (occurrences: 0-n).

[2] https://guidelines.openaire.eu/en/latest/data/field_creator.html

3. Identifier

The identifier of the contact (occurrences: 0-n).

Identifier

An identifier that uniquely identifies the contact (occurrences: 0-1). The format depends upon the scheme. For ORCID, one example is:

http://orcid.org/0000-0001-8737-4678

Identifier's scheme

The name of the identifier scheme (occurrences: 0-1). The ORCID scheme is recommended.

Identifier's scheme URI

The URI of the scheme (occurrences: 0-1). Examples:

- http://www.isni.org
- http://orcid.org

4. Affiliation

The organizational or institutional free-text affiliation of the contact (occurrences: 0-n). Examples:

- Université de Brest, CNRS, IRD, UMR 6539 LEMAR, IUEM; Technopôle Brest Iroise, Place Nicolas Copernic, 29280 Plouzané, France
- Institute of Marine Biology; National Taiwan Ocean University; Keelung 20224,Taiwan

3 Proposed architecture

3.1 Architecture principles

The overall targeted architecture is described in deliverable D5.4 (A Development Plan for Common Operations and Cross-Cutting Services based on a Network of Data Managers and Developers). The catalog architecture will fit in this overall architecture.

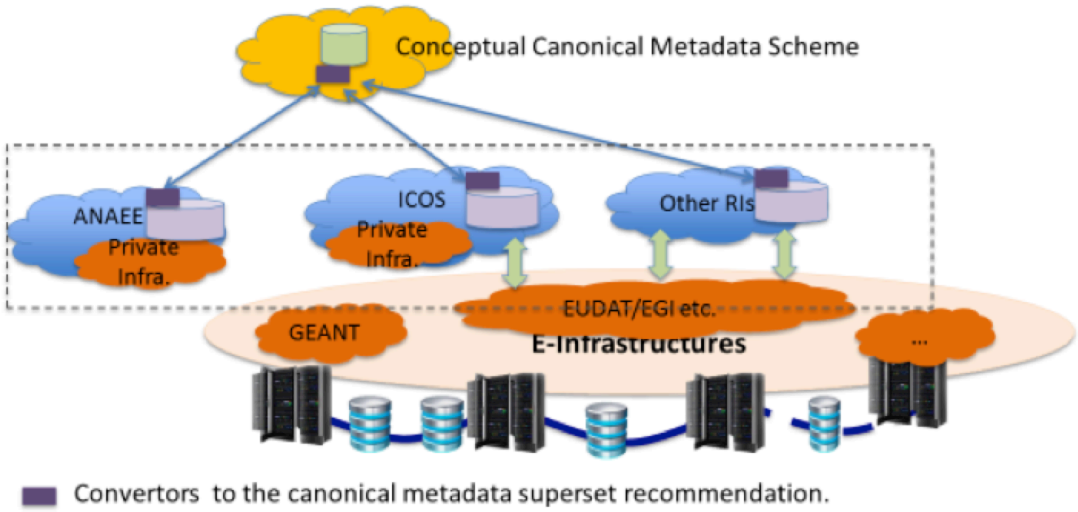

FIGURE 11: ARCHITECTURE PROPOSED IN DELIVERABLE 5.4

The catalog system is itself actually distributed in all the components of this overall targeted architecture:

- RI maintain their dedicated catalogs on their own private infrastructure. Especially, the descriptions (metadata) of the data and acquisition services provided by the RI are maintained where the expertise is. The catalog update operation, defined by the ENVRI Reference Model, is done by operators on these catalogs the structure of which best fits the RI discipline's specificities.
- Reference information can be managed on shared e-infrastructure such as documents registered on OpenAIRE, or persons on ORCID (where GEANT is shown as example on Figure 1Figure 11).
- Finally, to consolidate the information (here metadata) elaborated in RI, superset-catalogs replicate and gather information from the RIs (represented as conceptual canonical metadata scheme on Figure 11). However, It is recommended to allow flexibility and redundancy in the superset catalogs to adapt the maturing cross-disciplinary solutions developed in WP9 use cases and heterogeneous as-yet-undefined innovating applications (e.g. virtual research environments for specific communities) produced in the future.

Regarding the catalogs, specifically, the proposed architecture is as follow:

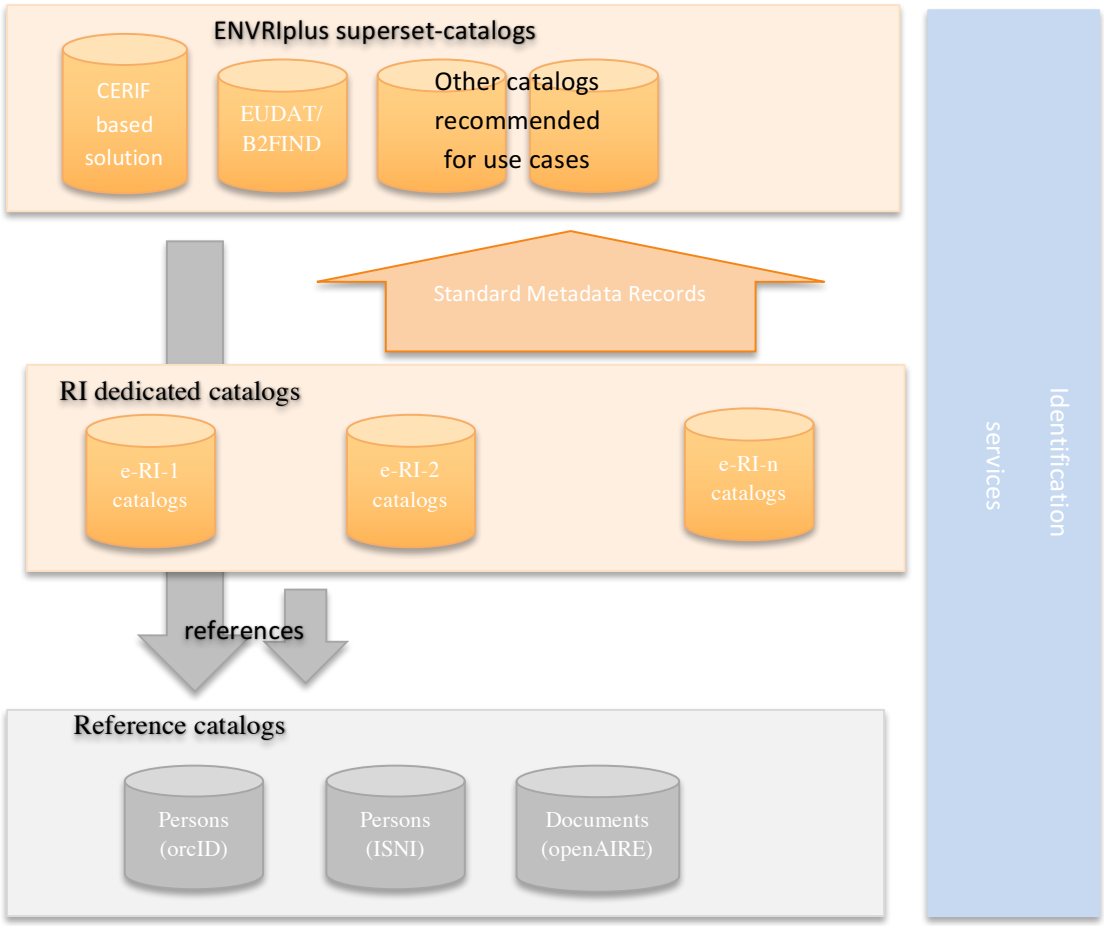

FIGURE 12: CATALOGS ARCHITECTURE

ENVRIplus superset-catalogs contain the standardized metadata records of the RI resources. Standard discovery functions are provided on top of these superset catalogs. Multiple instances of catalogs can be implemented, each have their strength and specific focus.

RI dedicated catalogs contains the ad-hoc metadata representation of RI resources, in their field of expertise. Updates and advanced discovery or reporting functions are provided on top of these catalogs. They are used by RI operators and their content is partially harvested in the superset-catalog from standard interfaces.

Reference catalogs contains information not specific to environmental RI but implemented by dedicated European or Global Infrastructures. The reference catalogs should be used under certain constraints (see 2.2.3.4).

Identification schemes or services following WP6 recommendations should be used to identify records in catalogs.

The reference, RI and superset catalogs should be discoverable in RDF through SPARQL end-points.

Specifically for the flagship data and acquisition services catalog, additional implementations details are given here after. For the other catalogs, the details on possible implementation will be given and demonstrated in the use case development, in WP9.

3.2 Flagship data service catalog architecture

CERIF and EUDAT/B2FIND are both candidates for implementing superset-catalogs for the ENVRIplus flagship data services catalog. They both have their specific strengths:

- CERIF: implements a data model the scope of which encompasses different aspects of science (funding, infrastructures, datasets….). This is for ENVRIplus an opportunity to collect in a single repository the information related to RIs and therefore ease discovery and reporting on those. In addition CERIF is interoperable with a range of metadata standards (ISO19XXX, Dublin-Core, DCAT ….) and can be made interoperable with the OIL-E semantic linking framework (https://confluence.egi.eu/display/EC/OIL-E).
- EUDAT/B2FIND: is a CKAN server dedicated to dataset discovery. It is up and running and provides a low threshold solution to initiate the flagship data services catalog. It can harvest ISO19XXX and Dublin-Core records.

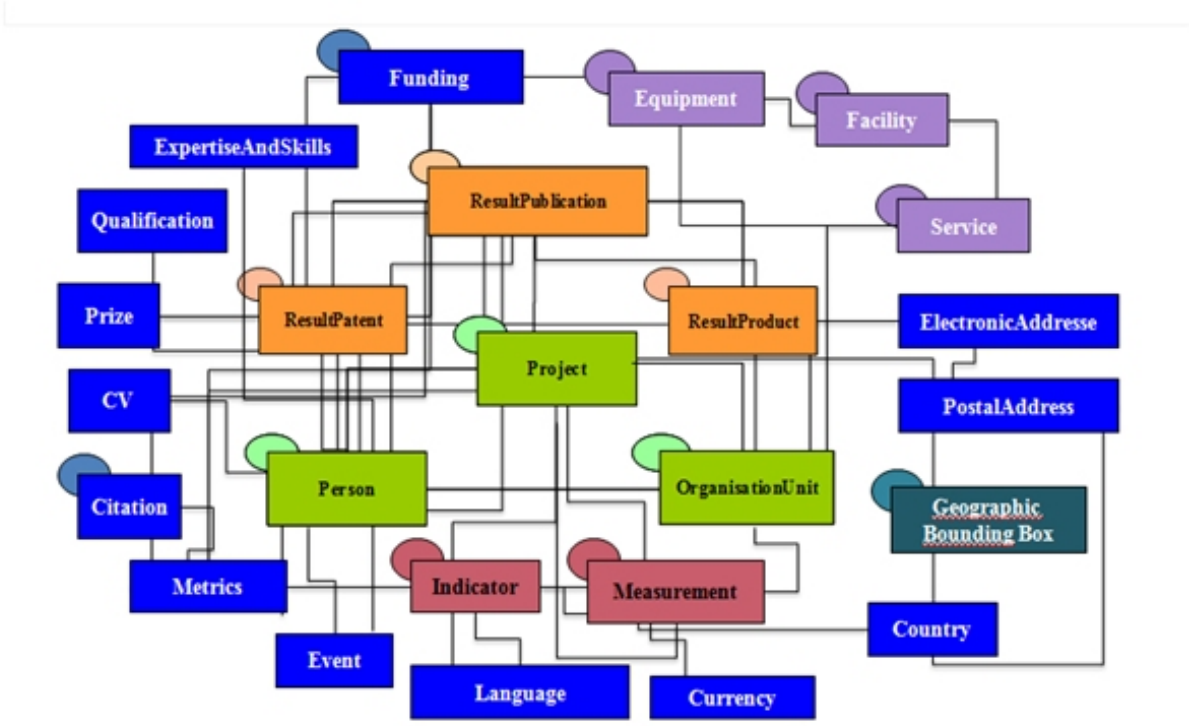

FIGURE 13: CERIF DATA MODEL

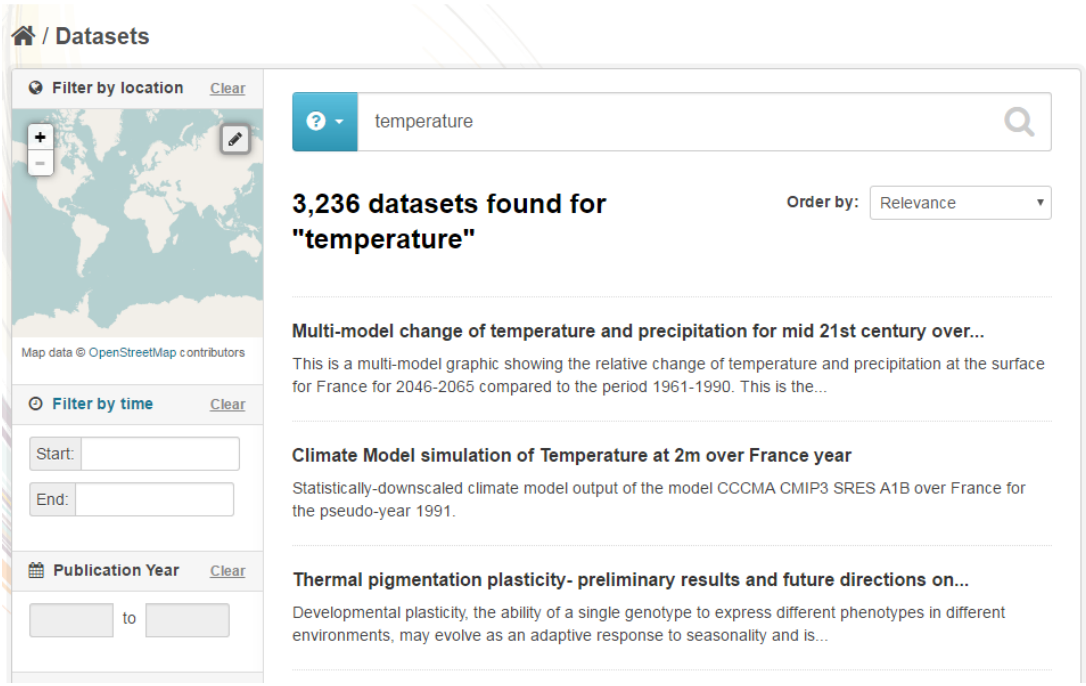

FIGURE 14: EUDAT/B2FIND DISCOVERY INTERFACE

For a first version, these superset-catalogs will collect metadata records from the following first circle of RI catalogs: ANAEE, Euro-ARGO, EMBRC, EPOS, IAGOS, ICOS, LTER and SeaDataNet. The record collection will be implemented by a so-called "harvesting" process: the RI catalog content will be exposed in standard interfaces implementing canonical metadata profiles defined in 5.1 (ISO19139, DC, DCAT). The superset catalog will periodically extract latest version of metadata from these interfaces to consolidate them in their repository.

The description evaluation function (2.2.3.3.13) is meant to be simple and will be implemented by both catalogs.

The CKAN implementation is configured and operated by DKRZ as B2FIND function of EUDAT.

The ENVRIplus/CERIF implementation will be set up and operated by EPOS. As already proposed there, the mapping from CERIF to CKAN will expose the richer metadata content possible in CERIF, and also will make the ENVRIplus catalog compatible with VRE4EIC (a Virtual Research Environment EU project) allowing appropriate coupling for multidisciplinary research.

4 Summary

The current document provides foundational guidelines for RIs to set up catalogs (see 2.2.2) for any of the concepts which need to be documented (datasets, sensors, physical samples…). The detailed recommendations for the implementation of each of these catalogs are given in WP9 use cases as done for example for TC_4, sensor registry. They will propose standards and application profiles for the interoperability of the catalogs and propose tools for the implementation of the

catalog in the RI and federation of the RI's catalog in a single portal or superset catalogs dedicated to specific items (datasets, sensors, physical samples, processes…)

In order to promote the trans-disciplinary nature of the ENVRIplus cluster of Research Infrastructures, a federated flagship data services and acquisition services catalogs is designed and detailed in this document. The detailed functional requirements (see 2.2), standards, application profiles and tools (see architecture 3.2) are described in this document.

As a first step, for first half of 2017, the flagship data services of RI's represented as partners in task 8.2 (ANAEE, Euro-ARGO, EMBRC, EPOS, IAGOS, ICOS, LTER and SeaDataNet) will be described in a EUDAT/B2FIND community instance. The superset catalog will be fed from ISO19139 descriptions of these services provided and harvested from the RI's catalogs. This catalog will be completed and improved then during the project with data service descriptions from other RI.

In addition, as described in 2.2.3.3.2, the flagship acquisition services descriptions will be added to this catalog. The flagship acquisition services will nicely complement the data services in this catalog exposing and promoting the outcome of the RI at trans-disciplinary level, beyond their traditional communities of users.

This catalog 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 project.

5 Appendices: Detailed components and interfaces description

5.1 Flagship Data Service Descriptions: Encoding guidelines

For the agreed attributes, canonical encoding is proposed in different standards:

- ISO19139 [http://www.iso.org/iso/catalog_detail.htm?csnumber=32557]
- Dublin-Core and dedicated extension profile [http://dublincore.org/]
- DCAT [https://www.w3.org/TR/vocab-dcat/]
- Schema.org [https://schema.org/Dataset]
- openAire and dataCite [https://guidelines.openaire.eu/en/latest/data/use_of_oai_pmh.html]

In addition, mapping with ENVRIplus pivot models is described:

- OIL-E
- CERIF

5.1.1 ISO19139

5.1.1.1 References

Namespaces:

xmlns:gmd="**http://www.isotc211.org/2005/gmd**"
xmlns:gts="**http://www.isotc211.org/2005/gts**"
xmlns:xsi="**http://www.w3.org/2001/XMLSchema-instance**"
xmlns:gml="**http://www.opengis.net/gml**"
xmlns:gco="**http://www.isotc211.org/2005/gco**"
xmlns:gmx="**http://www.isotc211.org/2005/gmx**"
xmlns:xlink="**http://www.w3.org/1999/xlink**"

Schema (xsd):

xsi:schemaLocation="**http://www.isotc211.org/2005/gmd http://schemas.opengis.net/iso/19139/20060504/gmx/gmx.xsd**">

5.1.1.2 Attribute encoding

Attribute	Xpath	Comment	Supported by
RI	/gmd:MD_Metadata/gmd:contact/gmd:CI_ResponsibleParty/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:pointOfContact[./gmd:CI_ResponsibleParty/gmd:role/gmd:CI_RoleCode/@codeListValu=”Publisher]/gmd:CI_ResponsibleParty/@uuid	Apply contact identification rules described in 2.2.3.4.2 Complete with available attributes	IAGOS, SeaData Net, ARGO
Identifier	Internal identifier of dataset, non unique /gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:identifier/gmd:MD_Identifier/gmd:code/gco:CharacterString URI (universally unique): /gmd:MD_Metadata/gmd:distributionInfo/gmd:MD_Distribution/gmd:transferOptions[./gmd:onLine/gmd:CI_OnlineResource/gmd:protocol/gco:CharacterString/text() = “WWW:LINK-1.0-http--metadata-URL”]/gmd:MD_DigitalTransferOptions/gmd:onLine/gmd:CI_OnlineResourc/gmd:linkag/gmd:URL	Internal identifiers and PID are managed.	SeaData Net, ARGO
Title	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:title/gco:CharacterString	Free text	IAGOS, SeaData Net, ARGO,

Description	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:abstract/gco:CharacterString	Free text	IAGOS, SeaDataNet, ARGO,
Quicklook	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:graphicOverview[contains(./gmd:MD_BrowseGraphic/gmd:fileDescription/gco:CharacterString/text(), 'thumbnail')]/gmd:MD_BrowseGraphic/gmd:fileName/gco:CharacterString	URL	IAGOS, SeaDataNet, ARGO,
Keywords	**Freetext keywords**: /gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords/gmd:keyword[*]/gco:CharacterString Or **for linked data references** (recommended) /gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords/gmd:keyword/gmx:Anchor	Freetext keyword or possibility to add enumeration reference (e.g. as SKOS) in /gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords/gmd:type/gmd:MD_KeywordTypeCode/@codeList	
Temporal window	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent/gmd:extent/gml:TimePeriod/gml:beginPosition …gml:endPosition	Start and end date time of the period of validity of the dataset. Date time are encoded in ISO8601	
Geo-spatial bounding box	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement/gmd:EX_GeographicBoundingBox/gmd:westBoundLongitude/gco:Decimal … gmd:eastBoundLongitude/gco:Decimal … gmd:southBoundLatitude/gco:Decimal … gmd:northBoundLatitude/gco:Decimal		

Location	/gmd:MD_Metadata/gmd:distributionInfo/gmd:MD_Distribution/gmd:transferOptions[./gmd:onLine/gmd:CI_OnlineResource/gmd:protocol/gco:CharacterString/text() = “WWW:DOWNLOAD-1.0-link--download”]/gmd:MD_DigitalTransferOptions/gmd:onLine/gmd:CI_OnlineResourc/gmd:linkag/gmd:URL		SeaDataNet, ARGO
Format	/gmd:MD_Metadata/gmd:distributionInfo/gmd:MD_Distribution/gmd:distributionFormat/gmd:MD_Format/gmd:name/gco:CharacterString		IAGOS, SeaDataNet, ARGO,
Contact	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:pointOfContact/gmd:CI_ResponsibleParty/@uuid	Apply contact identification rules described in 2.2.3.4.2 Complete with available attributes (email, name, affiliation) Role of contact is among list in **Contacts** section.	IAGOS, SeaDataNet, ARGO,
Related documents or resources	/gmd:MD_Metadata/gmd:distributionInfo/gmd:MD_Distribution/gmd:transferOptions[./gmd:onLine/gmd:CI_OnlineResource/gmd:protocol/gco:CharacterString/text() = “WWW:LINK”]/gmd:MD_DigitalTransferOptions/gmd:onLine/gmd:CI_OnlineResourc/gmd:linkag/gmd:URL		SeaDataNet, ARGO
Availability, license	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:resourceConstraints/gmd:MD_Constraints/gmd:useLimitation/gco:CharacterString		SeaDataNet, ARGO
Update date	/gmd:MD_Metadata/gmd:dateStamp/gco:DateTime	Update date of the data encoded in ISO8601	IAGOS, SeaDataNet, ARGO

5.1.1.3 Multi-linguality management

The attribute of type free text should be multi-valued (cartinality n) with language and text in the language. However in order to reduce the cost for setting up the initial version, only English will be managed.

For attributes being references to external resources (vocabularies or semantic layer, contact directories…), the multi-linguality is managed in the catalog or system managing the resource description.

5.1.1.4 Examples

http://sextant.ifremer.fr/geonetwork/srv/eng/csw-MYOCEAN-CORE-PRODUCTS?service=CSW&request=GetRecordById&version=2.0.2&id=f91cd16e-c47f-4b2f-8c38-d19b633ba02b&outputSchema=http://www.isotc211.org/2005/gmd&outputFormat=application/xml&ElementSetName=full

https://data.lter-europe.net/deims/node/8709/iso19139

5.1.2 Dublin-Core

5.1.2.1 References

Namespaces:

```
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:xml="http://www.w3.org/XML/1998/namespace"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:dcmi="http://purl.org/dc/dcmitype"
xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
```

Schemas :

```
xsi:schemaLocation="
http://purl.org/dc/elements/1.1/ http://dublincore.org/schemas/xmls/qdc/dc.xsd
http://purl.org/dc/dcmitype/
http://dublincore.org/schemas/xmls/qdc/2003/04/02/dcmitype.xsd
http://purl.org/dc/terms/ http://dublincore.org/schemas/xmls/qdc/dcterms.xsd
http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
```

5.1.2.2 Attributes encoding

Attribute	Xpath	Comment	Supported by
RI	/oai_dc:dc/dc:publisher	Additional attribute required to identify the RI among publishers.	
Identifier	/oai_dc:dc/dc:identifier	Recommended by openaire and datacite : DOI:10.17882/39475 Preferred for linked data: http://dx.doi.org/10.17882/39475	
Title	/oai_dc:dc/dc:title		
Description	/oai_dc:dc/dc:description		
Quicklook		Not available	
Keywords	/oai_dc:dc/dc:subject[*]	Data linked reference (href…) would be interesting to complement the free text keyword.	
Temporal window	/oai_dc:dc/dcterms:temporal	ISO8601 start, end	
Geo-spatial bounding box	/oai_dc:dc/dcterms:spatial[@xsi:type="@DCTERMS:Box"]		EUDAT
Location	/oai_dc:dc/dc:source		
Format	/oai_dc:dc/dcterms:format	Use mime type, see	
Contact	/oai_dc:dc/dc:creator /oai_dc:dc/dc:contributor /oa_dc:dc/dcterms:mediator (?) /oai_dc:dc/dc:publisher	Creator = author | originator Any other role Mediator = user Publisher = publisher Only contact name value fits in Dublin-core. Use it for person name or	

		affiliation name depending on the context. Further detail on how to transfer contact URI (e.g. ORCID) will defined with EUDAT during implementation phase.	
Related documents or resources	/oai_dc:dc/dc:relation	URI of the documentation	
Availability, license	/oai_dc:dc/dc:rights	URI of the license description	
Update date	/oai_dc:dc/dc:date	ISO8601	

5.1.2.3 Multi-linguality

@xml:lang, value among ISO 639-1

5.1.2.4 Examples

http://www.seanoe.org/oai/OAIHandler?verb=GetRecord&identifier=oai:seanoe.org:39475&metadataPrefix=oai_dc

http://sextant.ifremer.fr/geonetwork/srv/eng/csw-MYOCEAN-CORE-PRODUCTS?service=CSW&request=GetRecordById&version=2.0.2&id=f91cd16e-c47f-4b2f-8c38-d19b633ba02b&outputSchema=http://www.opengis.net/cat/csw/2.0.2&outputFormat=application/xml&ElementSetName=full

5.1.3 DCAT

No candidate encoding has been proposed yet. The model proposed by ICOS might be used as default proposal.

5.1.4 Schema.org

See https://schema.org/Dataset

5.1.5 Other recommended encoding, openAIRE, dataCite

See openaire recommendation

https://guidelines.openaire.eu/en/latest/data/use_of_oai_pmh.html

OpenAIRE utilizes CERIF as its storage format and exchanges information in XML format within the OAI-PMH harvesting protocol.

5.1.6 OIL-E and CERIF mapping

As for ISO, DC or DCAT standards, CERIF and OIL-E should be mapped with the profile proposed from the RDA interest group in section 2.2.3.3.12.

OIL-E is not refined at this level yet while CERIF has numerous built-in mappings which should comply with the encoding proposed above.

CERIF mappings exist to:

(a) INSPIRE/ISO19115: a mapping has been done between CERIF and INSPIRE and further work is ongoing in the EPOS project to deal with the varying dialects of INSPIRE.

(b) DC (Dublin Core). While CERIF to DC is quite straightforward (although lossy) the reverse is not so simple because of ambiguities in interpreting the DC metadata records which are commonly in textual or HTML form. If they are qualified DC (now deprecated) or encoded as RDF then resolution is more straightforward. However, there are various 'dialects' of DC so convertors may need revision.

(c) DCAT: a mapping has been produced between CERIF and DCAT. There is a workshop sponsored by the VRE4EIC project and W3C at the end of November 2016 to plan extensions to DCAT to make it appropriately expressive and more aligned with CERIF (which is used in VRE4EIC)

(d) CKAN metadata: the CKAN system has a standard metadata not unlike DC. However, it is usually encoded as RDF and so conversion is relatively straightforward. This conversion (CERIF-RDF) was done initially in the ENGAGE project and the software is now maintained by EKT in Athens as a general CERIF-RDF mapping and conversion.

(e) OIL-E: in the VRE4EIC project (where ENVRIplus through UvA is a partner) a mapping between CERIF and OIL-E has been done using tools supplied by FORTH in Heraklion. This is being extended in cooperative work between euroCRIS, UvA and FORTH within VRE4EIC.