D6.3: 1st Report on Data Interoperability: Findability and Interoperability

Author(s)	Ari Asmi (ICOS-ERIC), Bas Cordewener (JISC), Carole Goble (ELIXIR - UMAN), Donatella Castelli (CNR), Eileen Kühn (KIT), Fabio Pasian (INAF), Franco Niccolucci (UFlorence), Helen Glaves (BGS/NERC), Keith Jeffery (BGS/NERC), Massimiliano Assante (CNR), Matthew Dovey (JISC), Natalia Manola (Athena), Nick Juty (ELIXIR - UMAN), Niklas Blomberg (ELIXIR - EMBL), Rafael Jimenez (ELIXIR - EMBL), Volker Beckmann (CNRS)
Status	Final
Version	v1.1
Date	31/12/2017

Abstract:

The EOSCpilot project is the first phase in the development of the European Open Science Cloud (EOSC). It aims to improve interoperability between European data infrastructures, enabling efficient means to discover, access and share data in a sustainable manner, which is also easy to implement. This document reports on the development of a set of guidelines to facilitate findability and reuse through a simple set of metadata properties that can be easily adopted, and that are applicable across scientific domains. We describe the process used in the identification of these properties, the properties themselves, and analyse how they relate to those properties proposed for use by other, comparable, international efforts. In addition, we describe the means by which these guidelines can be adopted, and list the demonstrators that are under consideration to validate them.

Dissemination Level

- [X] PU: Public
- [ ] PP: Restricted to other programme participants (including the Commission)
- [ ] RE: Restricted to a group specified by the consortium (including the Commission)
- [ ] CO: Confidential, only for members of the consortium (including the Commission)

The European Open Science Cloud for Research pilot project (EOSCpilot) is funded by the European Commission, DG Research & Innovation under contract no. 739563

Document identifier: EOSCpilot -6-D6.3	
Deliverable lead	**ELIXIR**
Related work package	**WP6**
Author(s)	Ari Asmi (ICOS-ERIC), Bas Cordewener (JISC), Carole Goble (ELIXIR - UMAN), Donatella Castelli (CNR), Eileen Kühn (KIT), Fabio Pasian (INAF), Franco Niccolucci (UFlorence), Helen Glaves (BGS/NERC), Keith Jeffery (BGS/NERC), Massimiliano Assante (CNR), Matthew Dovey (JISC), Natalia Manola (Athena), Nick Juty (ELIXIR - UMAN), Niklas Blomberg (ELIXIR - EMBL), Rafael Jimenez (ELIXIR - EMBL), Volker Beckmann (CNRS)
Contributor(s)	Damien Lecarpentier (EUDAT), Steven J. Newhouse (EMBL-EBI) and Brian Matthews (STFC)
Due date	**31/12/2017**
Actual submission date	**03/01/2018**
Reviewed by	**Steven J. Newhouse (EMBL-EBI), Damien Lecarpentier (EUDAT) and Brian Matthews (STFC)**
Approved by	**Juan Bicarregui (STFC)**
Start date of Project	**01/01/2017**
Duration	**24 months**

Versioning and contribution history

Version	Date	Authors	Notes
0.1	18/12/2017	Ari Asmi (ICOS-ERIC), Bas Cordewener (JISC), Brian Matthews (STFC), Carole Goble (ELIXIR - UMAN), Donatella Castelli (CNR), Eileen Kühn (KIT), Fabio Pasian (INAF), Franco Niccolucci (UFlorence), Helen Glaves (BGS/NERC), Keith Jeffery (BGS/NERC), Massimiliano Assante (CNR), Matthew Dovey (JISC), Natalia Manola (Athena), Nick Juty (ELIXIR - UMAN), Niklas Blomberg (ELIXIR - EMBL), Rafael C Jimenez (ELIXIR - EMBL) and Volker Beckmann (CNRS)	Versioning and collaborative writing have been tracked in a google doc. All partners from 6.2 contributed providing feedback to shape this document via monthly calls, workshops and meetings.

1.1	31/12/2017	Damien Lecarpentier (EUDAT), Steven J. Newhouse (EMBL-EBI), Brian Matthews (STFC), Carole Goble (ELIXIR - UMAN), Nick Juty (ELIXIR - UMAN), Niklas Blomberg (ELIXIR - EMBL), Keith Jeffery (BGS/NERC) and Rafael C Jimenez (ELIXIR - EMBL)	This version was improved with the feedback provided by internal reviewers and 6.2 partners.

TABLE OF CONTENT

LIST OF FIGURES

1. EXECUTIVE SUMMARY

The EOSCpilot project is a crucial first step that will lay the foundations for the development of European Open Science Cloud (EOSC), being driven by the needs of the scientific community, and through alignment with multiple global partners.

In Work Package 6 (WP6) we aim to establish principles and develop mechanisms that enable the EOSC to provide research and data interoperability across the diversity of existing (and potential future) research communities, Research Infrastructures (RIs) and other research organisations. The objective is to develop principles that integrate and sustain community requirements for long term data stewardship in the EOSC that enables curation, provenance and quality using accepted community standards, conventions and established services.

Ultimately, we need mechanisms for representing common entities such as people, organisations, and resources in the EOSC that preserve established community standards and follow well established policies such as G8 GSO and the FAIR recommendations.

In the first phase the focus is on supporting (i) the finding and accessing of datasets across several scientific disciplines by exposing FAIR data to EOSC services and users and (ii) the interoperability of metadata catalogues.

In the execution of this work, we have explored existing processes and mechanisms facilitating interoperability, using a set of guiding principles. These principles allowed us to define the scope of the task, as well as focusing our efforts. These principles include:

- Reuse and repurpose existing metadata
- Identify and recommend improvements to RIs on their metadata
- Identify the minimal metadata for maximal benefit for users and services
- Propose strategies to expose metadata
- Recommend solutions that are simple to implement and easy to sustain
- Leverage upon existing work
- Align with international initiatives

A comparison of data models, vocabularies and standards, used across multiple disciplines to express the properties of scientific data, was performed with the aim to coalesce upon a core (minimal) set of properties. These metadata properties, termed EDMI (EOSC Dataset Minimum Information) is deemed, to be sufficient to enable data to be findable by users and suitably ‘aware’ programmatic services. Work has also been instigated to enable an additional layer of properties (‘recommended’ and ‘optional’) to be exposed, allowing more domain-specific data to become findable.

We also describe a strategy for exposing EDMI metadata, and recommend the process by which metadata should be registered and exchanged between data catalogues. Crucially, and guided by our established principles, our strategy draws upon pre-existing work (standards and data models), and can be easily adopted (for example, leveraging schema.org).

Working in conjunction with stakeholder data resource owners and catalogue providers, we have also begun the process of identifying suitable scientific demonstrators. These demonstrators will enable the testing and validation of the EDMI metadata, as well as of the comprehensiveness and adoptability of the guidelines generated by this work. It is our intention that the work reported here be refined in response to feedback and learnings from these demonstrators, as well as through continued discussions, calls, and targeted workshops.

2. INTRODUCTION

The objective of this task, within the EOSCpilot, is to demonstrate how to ensure availability of scientific data to users and services through an open cloud infrastructure. The purpose of this document is to provide a first draft of the strategy and recommendations to help with finding and accessing datasets across several scientific disciplines.

This document starts with an introduction including information about the EOSC, the EOSCpilot project, the data interoperability work package and the principles defining the scope of the data interoperability task, as well as the process to achieve its goals. The next chapter clarifies concepts relevant to understand the recommendations and strategy proposed in the following chapters. This is complemented with a short summary exposing the results of a survey, the outcomes of several workshops and the analysis of methods to expose metadata. The strategy and recommendations are followed by a description of EOSCpilot data interoperability demonstrators proposed for 2018. The demonstrators aim to test and evaluate the feasibility of the work proposed in this document. The last chapter describes other topics that need to be considered in this work.

2.1. EOSC and EOSCpilot

The European Open Science Cloud (EOSC)[1] has been proposed over the last few years, and is now under active development within European Commission funded programmes. The concept has arisen in recognition of a fragmented approach to the development of data and computing infrastructure to support science, which has meant that barriers have arisen between disciplines, regions and organizational structures.

The vision of the EOSC is as a common distributed environment which supports the publication of data and the publication of the services which store, manage, search, access, analyse, share and recombine and reuse that data. Thus, all researchers in Europe and beyond can then have open and seamless access, using open interfaces accessible to people and machines. The EOSC will federate within and across thematic research data infrastructures, and across horizontal e-infrastructures, providing common service across disciplines. Thus, the EOSC can: enable the more efficient use of tools and services, as best practice arising from earlier investments is more widely propagated; support data discovery and integration across cross disciplinary boundaries encouraging new science; and scale up the infrastructure, as new data providers and services have lower barriers to entry.

Key to overcoming these barriers, and promoting the development of a data infrastructure which supports open science, are standard formats, protocols and procedures which provide an interoperability layer for the sharing of data. A core framework to support data interoperability needs to be established which can be used to publish data into the EOSC.

The EOSCpilot project[2] has been funded to support the first phase in the development of the European Open Science Cloud (EOSC). It aims to set the baseline for the EOSC by: facilitating access of researchers across scientific disciplines to data, via science demonstrators; establish a governance and business model that sets the rules for the use of EOSC; and make technical recommendations and testbeds on creating a

[1] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
[2] https://eoscpilot.eu

cross-border and multi-disciplinary open innovation environment for research data, knowledge and services and establish global standards for interoperability for scientific data. Key activities include recommendations on interoperability, as discussed in the next section below.

2.2. Interoperability in EOSCpilot

The interoperability WP develops and demonstrates the interoperability requirements between e-Infrastructures, domain RIs and other service providers needed in the European Open Science Cloud. We map interoperability in two tracks: "Research and Data Interoperability" and "Infrastructure interoperability"

2.2.1. Research and Data Interoperability

The Research and Data Interoperability track provides the infrastructure and domain expert view in the work programme with focus on data interoperability. We base the definition of a Data Interoperability framework in the EOSC on the FAIR principles - data and services need to be Findable, Accessible, Interoperable and Reusable ("FAIR Principles for Data Stewardship" 2016). Based on the G8 GSO[3] recommendations and the FAIR principles this minimally encompasses machine actionable metadata as well as human consumable metadata. It also needs metadata standards and services that ensure and enable standards as the default - in the EOSC producing "FAIR" data should be routine, not an add-on requirement. This requires specific interfaces, standards and integration services between RI data-types.

2.2.2. Infrastructure interoperability

The track involves Cloud, Grid, HTC[4] and HPC[5] infrastructures, including large datastores, through high speed networks and performant data transfer protocols and tools. The high-level objective is to provide the most adequate infrastructures for the treatment of extensive amounts of data, generated by new generations of instruments, observatories, satellites, sensors, sequencers, imaging facilities and numerical simulations, and produced by well-known data intensive communities but also by the long tail of science. In the Infrastructure Interoperability track the provider view is in the centre of the work programme. Through the partners of this work package and the resources provided by the selected Science Demonstrators, federated infrastructure pilots will be set up. Those pilots will enable us to analyse the existing interoperability mechanisms for software components, services, workflows, users and resource access within existing Research Infrastructure[6] (RI) systems. Based on that analysis, this work package will develop a common mechanism utilising the 'best of breed' of existing mechanisms with a roadmap to evolution and convergence to a common framework.

The above objectives are envisioned to be implemented through the instantiation of multi-infrastructure, multi-community pilots. Services and the Science Demonstrators defined in work package 3 (WP3) and work package 4 (WP4) will then be deployed and validated in these pilots from the standpoint of maturity, scalability, and usability for a future EOSC.

WP6 is organised as three tasks. Task 6.1 aims to identify infrastructure interoperability gaps and propose an interoperability architecture. 6.2 is focused on recommendations for research and data interoperability and it is the task leading the work presented in this report. 6.3 aims to support interoperability use cases.

[3] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=gso
[4] High-throughput computing
[5] High-performance computing
[6] https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=about

The work of this work package is aligned with the expected impact of the INFRADEV-4-2016 call[7], requiring the project to "facilitate access of researchers across all scientific disciplines to the broadest possible set of data and to other resources needed for data driven science to flourish". In order to run this infrastructure and to enable usability, a rich set of interoperable infrastructure services (IaaS) ranging from AAI[8], reliable storage endpoints, cloud management frameworks, SDN[9] endpoints, to infrastructure monitoring, billing and accounting is required. This work package will gather infrastructures, operating computational facilities, and user communities from "the large" to "the long tail of science", avoiding unsustainable fragmentation.

2.3. Data interoperability

The vision of the EOSCpilot data interoperability task (task 6.2) is to establish principles and develop mechanisms that enable the EOSC to provide research and data interoperability across the diversity of existing (and potential future) research communities, RIs, and other research assets. To fulfil this vision, we must first explore the existing interoperability mechanisms/processes used for data, software components, and services (including workflows). In addition, we must investigate the mechanisms for user and resource access within existing RI systems, especially with respect to the syntactic (structure) and semantic (meaning) representation, and to the use of standards. Guidelines and recommendations will need to be created for the implementation and use of uniform descriptors, standard terms, persistent identifiers (PIDs) and encoding in the generation and storage of data, as well as defining a knowledge management architecture that provides services for users of these standards. An in-depth investigation into all these components/topics would require significant investment; while there are analogous efforts working in the same space, it would not be feasible and it is not our role in this project to engage with all the data interoperability efforts in the scientific community. **To balance the impact and effort of this group we decided to define the data interoperability goal focusing on one and probably the most important requirement described by the EOSCpilot project: demonstrate how to facilitate the availability of scientific data in EOSC.**

The objective of this task is to define and demonstrate the data interoperability architecture that would expose FAIR data to EOSC services and users. At the outset, it was agreed that it is not within scope to define how to make data itself FAIR, since that responsibility must lie with the individual RIs, e-Infrastructures and research communities, which must take into consideration their individual participant data resources. Moreover, there is already a working group funded by the European Commission, running in parallel to this group, tasked with defining a roadmap to make data FAIR across data repositories[10] [11].

After discussion within the group, and based on feedback collected from several EOSCpilot workshops including BlueBridge[12] workshop (Annex A) and the Open Science Fair[13] (Annex B), we generated a set of principles to define the scope and the direction of our activities in this project.

2.4. Guiding principles

These principles will drive the work of this task and help to refine the recommendations to be proposed to EOSC[14]. These are grouped into three categories: Reuse, Least and Practical.

[7] https://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/calls/h2020-infradev-2016-2017.html
[8] Authentication and Authorization Infrastructure
[9] Software Defined Networking
[10] https://github.com/FAIR-Data-EG/consultation
[11] http://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupDetail&groupID=3464
[12] https://www.bluebridge-vres.eu
[13] http://www.opensciencefair.eu/
[14] Recommendations summarised in section 4 and 5 and represented in Figure F, G and H

2.4.1. Reuse: Leverage the rich legacy of RIs

We must rely on metadata from RI metadata catalogues. A plethora of data repositories should exist to serve a particular scientific domain, with domain specific RIs maintaining metadata catalogues which collect, integrate, harmonise and enrich *metadata* from many dispersed and diverse data repositories to facilitate data discovery. We should leverage upon these existing metadata catalogues as the primary providers of scientific metadata for the EOSC. Thus, we should expect that domain specific metadata catalogues within EOSC will collect metadata from relevant data repositories.

We must support an ecosystem of catalogues and metadata flow. We envisage that an efficient system of metadata collection/sharing must contend with an ecosystem of coordinated metadata catalogues; domain specific metadata catalogues to collect domain-specific metadata from individual data repositories, while generic metadata catalogues collect a subset of (domain-agnostic) metadata from domain specific metadata catalogues. Ideally the generic EOSC metadata catalogues should pull metadata from domain specific metadata catalogues and recommend metadata submission to domain specific catalogues.

We should provide quality recommendations to feedback to RIs. With the analysis of metadata catalogues, metadata models and standards we should aim to provide recommendations to RIs about how to improve the quality of the metadata provided and collected by metadata catalogues (aggregators of metadata) and individual data repositories (metadata generators).

Making data FAIR is the responsibility of the RIs and their data repositories. The role of the EOSCpilot data interoperability working group should not be to define how to make data FAIR but to define and demonstrate a simple data interoperability architecture to expose FAIR data to EOSC services and EOSC users. We believe the responsibility of defining how to make data FAIR lies with RIs (and e-infrastructures), especially on their participant data repositories. Moreover, there is already a working group funded by the European Commission[15] which started in parallel to define a roadmap to make data FAIR ("FAIR Principles for Data Stewardship" 2016) across data repositories.

2.4.2. Least: Minimal metadata for maximal benefit

Findability first. Findability is naturally the first step to make data FAIR, being a precondition to subsequent data access and reuse. In considering findability, we should recognise the two main players, EOSC services and EOSC users, and determine their practices with respect to data access, and requirements for interoperability and reusability.

Common and minimum metadata. We should not aim to create a new data model to describe datasets or data repositories, but rather to coalesce upon a recommendation of the minimum metadata properties that are common across metadata catalogues. These properties should help EOSC services and EOSC users to find data repositories and datasets and should facilitate data access, interoperability and reusability. We should evaluate existing metadata models and recommend how to expose scientific data reusing one or several data models.

Focus on common data types: datasets and data repositories. Initially, we should focus our work on a few data types, such as datasets and data repositories, which are common across different scientific disciplines.

Flexible metadata models to embrace domain specifics. Each scientific domain should work in accordance with their own domain-specific standards and vocabularies, defining their specific entities which may also require the use of a standard format. Our objective should be to assist RIs and scientific communities in defining how better to describe their data, whilst respecting their existing formats and descriptions. Hence, we should aim to distil from this pool a set of minimum properties among metadata models, while retaining the flexibility to allow custom extensions for domain specific properties.

[15] http://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupDetail&groupID=3464

Service requirements and operational metadata as first-class citizens. Though there might be a considerable overlap it must be recognised that the metadata required in this task is of two equally important types. Scientific metadata is crucial for users to understand the details of the scientific records that are being served, while operational metadata is essential for (programmatic) services to be able to identify appropriate scientific data, and subsequently to access and (re)use it.

2.4.3. Practical: Sustainable and pragmatic delivery

Engage existing data repositories from EOSC science demonstrators. We should actively involve prominent data repositories from EOSC science demonstrators to demonstrate how their datasets can be made more discoverable and accessible through EOSC, via metadata catalogues.

Reuse methods to expose dataset metadata through metadata catalogues. Our strategy to expose dataset metadata (minimum and common properties) through metadata catalogues should relying on existing methods and guidelines to expose metadata. This work should be aligned with that done by initiatives such as RDA[16] and GO-FAIR[17] , as well as being informed by the expertise of our metadata catalogue partners.

Simple to implement, easy to sustain. Any proposed solution should be looking at a high impact, low effort strategy especially in the short term, thereby lowering the barriers to adoption This strategy is therefore more likely to succeed, and give immediately tangible benefit; it should be simple to implement and easy to maintain providing just enough functionality to facilitate discovery, access and use of data in the EOSC.

Deliver guidelines and demonstrators. Besides a final report on making FAIR data findable, accessible and reusable in EOSC, the outcomes of our work should also:

- generate a set of guidelines for using metadata;
- propose an architecture which facilitates the flow of metadata;
- support demonstrators which apply our recommendations to show the feasibility of our proposed strategy, and generate tangible results.

2.5. Out of scope

We also need to specify what is and is not within the scope of this work. We describe below considerations that are currently 'out of scope'.

2.5.1. Make data FAIR

It is not within the scope of this project to make existing data FAIR, but to make existing FAIR data findable, accessible, reusable and as interoperable as possible within the EOSC.

2.5.2. Secure data access and authentication & authorization infrastructure (AAI)

There are many initiatives like AARC[18] working on a strategy for secure data access. Though security and AAI is important in this context it is not within the scope of this task. Moreover, it is premature to discuss data interoperability efforts across secure access infrastructure, especially given that EOSC is still debating the precise strategy to follow. These discussions also relate to interoperability, described in task 6.1.

2.5.3. Domain specific data entities

A data repository contains datasets, which may themselves be composed of more datasets, domain specific entities or specific types of experimental or observational data. The description of this information matters but to be able to demonstrate discovery and accessibility in EOSC this project will focus just on dataset and data repository types.

[16] Research Data Alliance, https://www.rd-alliance.org
[17] https://www.dtls.nl/go-fair
[18] Authentication and Authorisation for Research and Collaboration, https://aarc-project.eu

2.5.4. Software, Workflows, Containers, People and other content types

The discovery and access of software, workflows, containers, research objects, people, etc. is important, especially in its relationship with data and provenance. However, this project will focus on data and more specifically on the metadata of datasets and data repositories.

2.5.5. Searching technologies

This work will not consider the technologies or tools that are used to search for data, but rather the data models, metadata and technologies describing how to expose pertinent metadata, and hence how it may be found. Examples of search oriented approaches to data discovery include Elsevier's DataSearch[19].

2.5.6. Specific data models

The data model to expose dataset and data resources metadata is important, however the focus of this work is about the common properties among different data models important to facilitate the discovery and access of data.

2.5.7. How to collect data from data repositories

This work it is not about how metadata catalogues collect data from data repositories but about how metadata catalogues expose metadata to EOSC users and services.

2.5.8. Data managed by data repositories

This work is not about the data hosted in data repositories or how it is managed. However, for this task it is important how data repositories expose metadata about their datasets.

2.5.9. Data sustainability

Though sustainability is an important topic is not part of the scope of this work.

2.6. Process

This project aims to produce a set of guidelines and a data interoperability architecture proposal, with the goal of exposing minimum metadata to EOSC services and EOSC users. This will facilitate the efficient finding, access and use of public, scientific data. These guidelines are not intended to be a static document, but rather should evolve over time, following testing and evaluation by data providers and data consumers. The process to reach this goal is described by the following tasks:

2.6.1. Review

Undertake a critical review of existing metadata catalogues, from multiple scientific domains, evaluating their dataset descriptions/vocabularies and strategies employed to expose metadata.

2.6.2. Finding use cases

Domain specific data catalogues are interested in the requirements of their own scientific community, to understand which metadata are important for their users to find data. These requirements are embedded in, and reflected by, their choice of data models. We intend to look at requirements driven by use cases, from individual demonstrators, to find the necessary metadata to complement the expertise already captured by metadata catalogues. We will also examine the technical requirements for services (specially EOSC services) to be able to find, access and use data.

2.6.3. Mapping of important properties

The use case requirements, with respect to crucial metadata, will be mapped to the metadata models used by metadata catalogues, and to the standards used to describe datasets and data repositories. The goal of this mapping is to highlight which properties are important for finding data, which metadata models are better suited to expose metadata, and which important properties are missing in existing metadata catalogues.

[19] https://datasearch.elsevier.com

2.6.4. Guidelines

We will provide guidelines to recommend minimum properties, suitable data models and appropriate technologies to describe and expose dataset and data repository metadata. These guidelines will include an architecture proposal and will demonstrate solutions to make FAIR data findable, accessible and reusable in EOSC.

2.6.5. Evaluation

We will evaluate technologies to expose dataset and data repository metadata following the proposed guidelines.

2.6.6. Demonstrators

We will demonstrate our proposed strategy and recommendations, using at least one metadata catalogue and one e-infrastructure's services, working with individual datasets. The feedback from our demonstrators will help us to evaluate the feasibility and adoptability of our guidelines, allowing us to refine them accordingly.

2.7. Events

To drive and support the guidelines and the strategy proposal, we have engaged partners and external stakeholders directly through workshops and surveys. For the duration of the data interoperability task within the EOSCpilot project, we have planned four thematic events. These events are tasked with bringing our partners together, engaging with the community, continuous re-evaluation of the scope and the work done so far, working collaboratively on planned tasks, and the dissemination of plans and outcomes of the group. These events will be complemented by events organised by 6.3, to support the EOSCpilot data interoperability demonstrators (see section 6: Demonstrators).

2.7.1. BlueBRIDGE workshop: "FAIR friendly research data catalogues: How far are we?"

The workshop took place in April 3 2017 at the 9th RDA Plenary Meeting in Barcelona (Spain). This workshop brought together over 40 representatives from H2020 projects, e-infrastructures, European and global initiatives, and data users currently dealing with research metadata catalogues. Discussions focused on how these initiatives are approaching the FAIR principles, their current status, and how they plan to move forward. The audience and the speakers were asked to reflect on this topic and provide input on three questions. More information including a summary of the workshop, recommendations and the answers collected from the contributors are reported in the "Annex A"

2.7.2. How FAIR Friendly is your data catalogue? Exposing FAIR data in EOSC

The workshop took place in September 8 2017 at the Open Science Fair 2017 in Athens (Greece). The workshop brought together more than 40 representatives from EOSCpilot and many other related efforts. The workshop was tasked to provide an update of the activities of the EOSCpilot data interoperability working group, and to engage the diverse stakeholders to shape the work of this group. The workshop was structured into two sessions. In session 1, scene setting presentations on EOSC were followed by short presentations by representatives from eight metadata catalogues, each representing subject-specific and generic (subject-agnostic) systems, as well as a review of a previous meeting. Prior to this workshop, the organisers conducted a survey of 11 metadata catalogues and an early analysis of this information was presented during session 2. This presentation was followed by extensive breakout discussions of 13 principles of catalogue metadata exposure and interoperability. More information about the workshop including a summary, principles and recommendations are reported in "Annex B".

2.7.3. EOSCpilot data interoperability technical workshop: Data catalogues and datasets in the European Open Science Cloud

The workshop took place in October 4-5, 2017 at the European Bioinformatics Institute in Hinxton (UK). The goal of this workshop was to evaluate existing approaches to describe, expose and integrate dataset metadata. During this workshop, we also provided an update of some metadata catalogues, and data models used to describe datasets. We also evaluated requirements from researchers and infrastructure services. The workshop was attended by 47 people including service providers, representatives from metadata catalogues, RIs, e-infrastructures and partners from the EOSCpilot project. During this workshop, we presented a set of recommendations and actions to drive our work. A summary of the workshop is included in “Annex C”.

2.8. Survey

Between August and the beginning of September 2017 we conducted a survey with the aim of identifying the characteristics of the metadata catalogues operated by the initiatives that might adhere and become part of EOSC. The aim of the survey was to understand the possible starting point for making these catalogues interoperable and to what extent they already facilitated the FAIR data management principles.

A questionnaire, devised for the purposes of the survey, composed of 10 questions was submitted to initiatives, institutions, research infrastructures and repositories representatives of various scientific domains, ranging from cultural heritage to marine, environmental and High Energy Physics. The preliminary result of the survey was presented at the “How Friendly is Your Data Catalogue?” workshop co-located within the Open Science FAIR. This preliminary result revealed responses collected from eighteen participants. During the EOSCpilot 6.2 face to face meeting the questionnaire was further disseminated to different initiative representatives of additional scientific domains, such as astronomical, archaeology and biomedical. A description of the metadata catalogues surveyed is presented in the “Annex D”. Some highlights after the analysis of this survey are presented in sections 3 and 4: Metadata Catalogues.

3. METADATA CATALOGUES, DATA REPOSITORIES AND DATASETS

In this work, we found that different people use the same terminology to mean different things, or else different terminology to mean the same thing. This section aims to clarify terminology used around metadata catalogues, data repositories and datasets to help understanding the recommendations and strategy proposed in this document.

3.1. Main stakeholders involved in the process of data sharing

Data sharing in science involves a producer who is the source of the data to be shared (often its creator), and a consumer (also known as user or recipient) (Lord et al. 2005). In some cases, one or more data resources storing and making the data available may lie between the producer and the consumer. One or more data resources, therefore, may act as an intermediary, facilitating the process of data sharing by integrating data from different producers and/or data resources (see Figure A). Aiming to facilitate the process of data sharing, as well as increasing the availability of research data, many journals and funders encourage producers to submit data and metadata to specified data resources. Producers, however, might initially manage their data within a local content management system (CMS) or laboratory management system (LIMS). Sharing data among different such stakeholders, from a centralised source, can happen via a push or pull method. Whilst both push and pull models are technically feasible, it is more transparent and open to employ a pull model where many such stakeholders can participate in collecting the data.

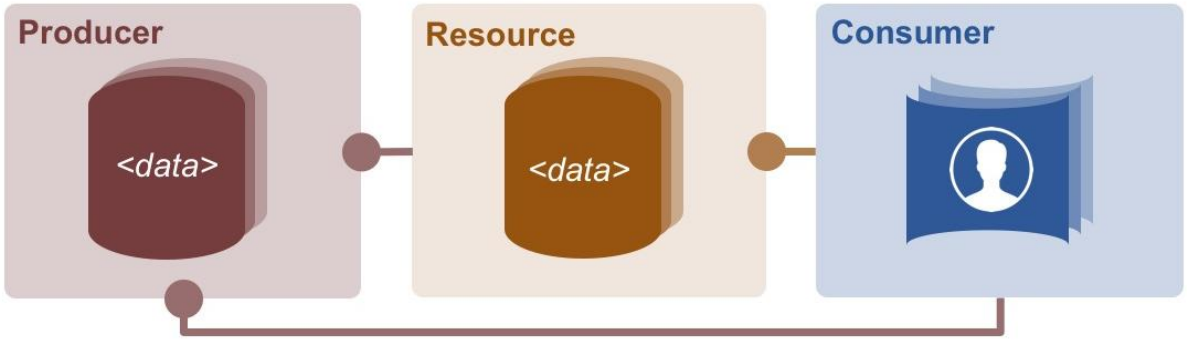


Figure A. Main stakeholders involved in the process of data sharing.

3.2. Types of data resources

3.2.1. Repositories, Knowledge-bases and Catalogues

We consider three types of data resources based on their content and functionality. A data repository (also known as data archive) is a deposition database collecting primary data from data producers (scientists, machines, etc.) arisen from different sources (experiments, observations, simulations, etc.). A knowledge-base is a database collecting information and accumulating experimental evidence, which is processed somehow to create knowledge. A metadata catalogue (also referred as registry) is a database collecting and integrating metadata from several resources to facilitate the discovery of third party data (see Figure B). Some data resources may traverse these classification boundaries, existing simultaneously as a repository, knowledge-base and/or metadata catalogue.

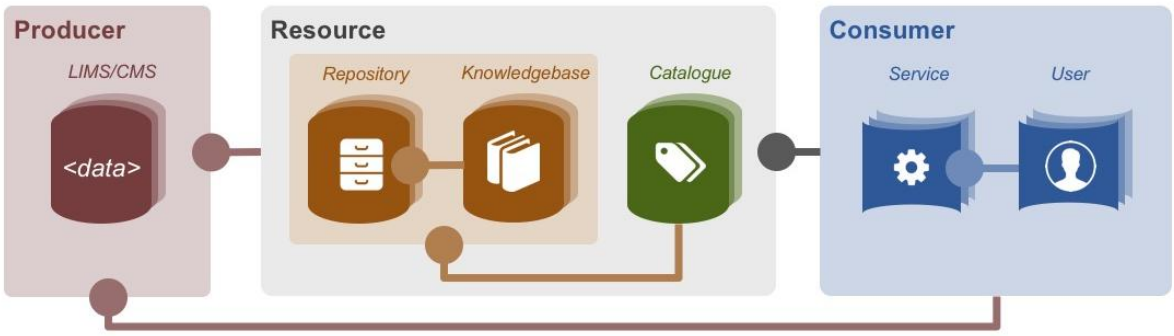


Figure B. Stakeholders involved in the process of data sharing and types of data resources.

There are some differences between data repositories and knowledge-bases which are worth considering especially when looking at the origin of the data. The content of an entry in a data repository normally remains static, while the entry in a knowledge-base is subject to change. Knowledge-base records can be enriched or changed based on emerging facts from new primary data. Data repositories are normally bigger in size since they contain the primary data (also known as raw data) produced during, for example, an experiment, while information produced as a result of processing data is much smaller. Thus knowledge-bases tend be smaller but richer in facts and information. Knowledge-base records require more human intervention since information usually requires interpretation and contextualisation, a job normally undertaken by domain experts called 'scientific curators'.

3.2.2. Generic vs specific

Data resources can also be classified by the type of scientific data they collect. At the highest level, resources can be classified for instance as life science, earth science and humanities. Within each domain we can find data resources with different levels of specificity. In life sciences for example there are protein resources, and within these protein resources lie several specific types like protein structure or protein sequence resources. Some other data resources are generic covering a broad spectrum of scientific data types. OpenAire[20] (Rettberg and Schmidt 2015) is an example of a generic data resource covering a wide range of scientific types.

3.2.3. Unstructured vs structured

This terminology is normally used for data repositories but it can also be applied to any type of data resource. Structured repositories are data resources with specific rules about how to represent and deposit data. These data resources have specific guidelines about how to annotate data and metadata and tend to follow specific formats and standards. Very often these types of resources are intended to collect a specific type of data. Unstructured repositories are more relaxed with rules since it is quite hard to come up with guidelines that satisfy different types of scientific data. Unstructured repositories are meant for those data that do not conform to existing standards and cannot be submitted to a structured repository. Normally there is a correlation between structured data repositories and domain specific repositories, as well as unstructured data repositories and generic data repositories. An example of a structured data repository is PRIDE[21] (Rettberg and Schmidt 2015; Jarnuczak and Vizcaíno 2017), a repository for proteomics experimental data. Figshare[22] (Singh 2011) is an example of an unstructured data repository.

3.3. Types of consumers

Most of the metadata catalogues and data resources considered in this work have excelled in collecting user requirements from researchers and other stakeholders. However, many of them have not considered

[20] https://www.openaire.eu/intro-data-providers
[21] https://www.ebi.ac.uk/pride/archive/
[22] https://figshare.com/

service requirements to be a priority, or had not considered them at all. In the EOSC, users will be able to access services, which to be successful, will need instructions on how to access the data that the user has requested. Therefore, it is important to take into account the different ways that data will be accessed, directly by users, as well as through services.

3.4. Datasets in data resources

Independently of the type of data resource (data repository, knowledge-base or metadata catalogue) we could consider all the data resources are organised in datasets. And datasets can be composed of other datasets or specific data records (Figure C). Across different scientific domains similar metadata properties are used to describe data resources and datasets. However, data records are described with very specific metadata related to the nature of the entity described. To illustrate these differences examples are provided below.

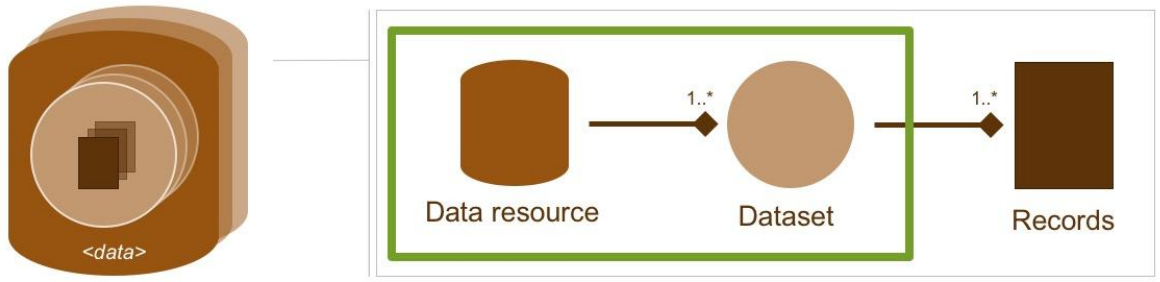


Figure C. Datasets and records in data resources. On the left a representation of a data resource including datasets including records. On the right box a decomposed and simplified representation showing a data resources can be composed of 1 or more datasets and a dataset can be composed of more data records.

This pattern appears in many well-known resources including those in the USA such as: dataONE[23], the DOE Data Explorer[24] and Dryad[25], as well as major international collections such as GBIF[26]. The RDA Data Foundation and Terminology Core model[27] also has components of it.

3.4.1. UniProt datasets and records

The UniProt (Magrane, Magrane, and UniProt Consortium 2010) knowledge-base is composed of different types of datasets. For example, based upon how the information is processed, UniProt can be defined as being composed of two datasets: SwissProt and TrEMBL. The SwissProt entries represent protein records which have been manually verified and extensively curated by a scientific curator. The TrEMBL set, on the other hand, are computationally predicted protein records, based upon the automated translation of EMBL sequence information (Translated EMBL). UniProt also contains datasets classifying protein records on a per species basis, with the Human proteome, instance, being one of the largest datasets. In this respect, therefore, the metadata used to describe the data resource and the dataset can be considered quite generic compared to the specific metadata required to describe a protein record. UniPort records contain specific metadata annotations like the sequence, associated diseases and protein interactions.

[23] https://www.dataone.org/
[24] https://www.osti.gov/dataexplorer/
[25] http://datadryad.org/
[26] https://www.gbif.org/
[27] https://b2share.eudat.eu/records/458fecc670ca4d13bdd635493f3449d6

3.5. Metadata catalogues

A metadata catalogue (also referred to as registry) is a database collecting and integrating metadata from several resources to facilitate the discovery of third party data. It can be described as a list of items with pointers to where to find the items, like the index on a database table or the card catalogue for a library. A repository stores the actual items, like a database table itself or a library shelf of books. Registries hold references to things while repositories hold the things.

Metadata catalogues can be classified by the type of metadata items they index. For instance, in the scientific domain we can find metadata catalogues of data resources (databases), datasets, publications, software tools, ontologies, standards, samples, training materials and scientific events. The majority of catalogues index metadata and build relationships for more than one type. Though all of these catalogues are important, in this project we are primarily interested in catalogues indexing data resources and dataset metadata. Metadata catalogues of data resources collect and integrate metadata from multiple individual data resources. This metadata can include information like the license of the data resource, the datasets available in the resource and the contact details of the maintainer of the resource. Examples of metadata catalogues of data resources are FAIRsharing (also a Force11 and RDA WG activity)[28], re3data[29], VizierR[30] and the Metadata Standards Directory (also a RDA WG activity)[31]. The other type of catalogue of interest in this work is the dataset metadata catalogue. The major role of dataset metadata catalogues is to index the dataset metadata of distributed data resources and facilitate the discovery of datasets. This metadata can include information like the date of the publication, the author of the dataset and the its identifier. Examples of metadata catalogues of datasets are OmicsDI[32] (Perez-Riverol et al. 2017), DataMed[33] (Ohno-Machado et al. 2017), OpenAIRE[34] and EUDAT-B2Find[35].

Metadata catalogues can also be classified by the role they play in the research workflow and their emphasis on experimental context. Those above, for example, are primarily targeted at the final deposition stage for holding results and put the dataset at the centre. Others, such as FAIRDOMHub[36] (Wolstencroft et al. 2017) and the US-based Open Science Framework[37] aim to support self-managed projects throughout the lifecycle and put the project at the centre, heavily referencing other metadata catalogues and repositories where the projects' data (and other types) are held. They also act in part as stores and are thus hybrids of repositories and metadata catalogues. A related catalogue type, the BioStudies database[38] (McEntyre, Sarkans, and Brazma 2015), holds descriptions of biological studies, linked to data in other databases at EMBL-EBI or outside, as well as data that do not fit in the structured archives at EMBL-EBI. These metadata catalogues promote the context of the datasets being catalogued.

Metadata catalogues can be generic or domain specific. Domain specific catalogues tend to collect more metadata details and have more restrictive guidelines to describe data. For instance, ProteomeXchange[39] (Jarnuczak and Vizcaíno 2017), a domain specific data catalogue, indexes proteomics datasets and uses

[28] https://fairsharing.org
[29] https://www.re3data.org
[30] http://vizier.u-strasbg.fr/viz-bin/VizieR
[31] https://rd-alliance.org/groups/metadata-standards-directory-working-group.html
[32] https://www.omicsdi.org
[33] https://www.nature.com/articles/ng.3864
[34] https://www.openaire.eu
[35] http://b2find.eudat.eu
[36] http://fairdomhub.org
[37] https://osf.io/
[38] https://www.ebi.ac.uk/biostudies/
[39] http://www.proteomexchange.org

several specific controlled vocabularies to describe many metadata properties like experimental methods or proteomics data types (see Figure D).

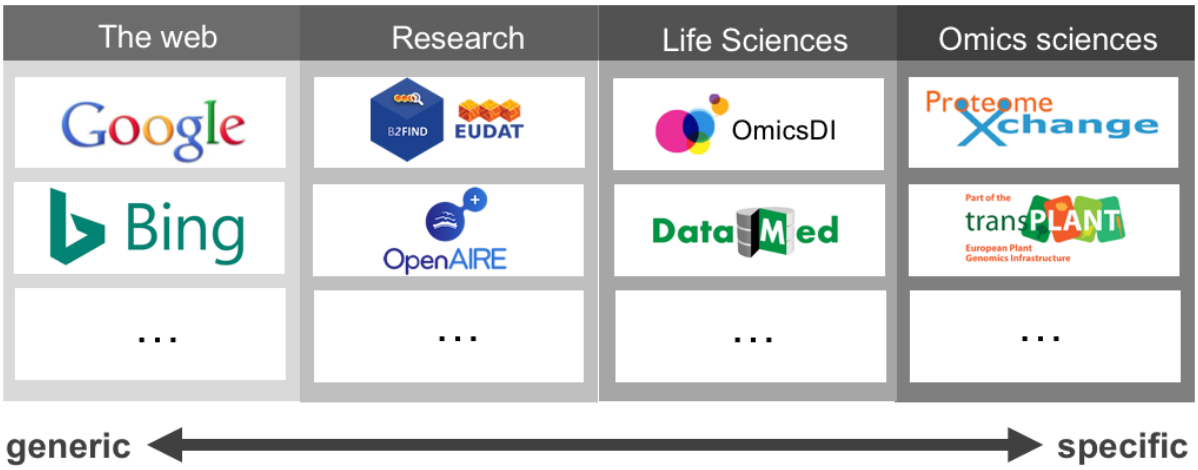


Figure D. Examples of metadata catalogues classified by specificity. Catalogues are listed from more generic on the left to more specific on the right. The examples (individual cells) include generic catalogues like EUDAT-B2Find and OpenAIRE, and specific catalogues in the life sciences like OmicsDI, DataMed. Even more specific within their respective domains are catalogues such as ProteomeXchange and transPLANT[40].

[40] http://www.transplantdb.eu/

4. METADATA CATALOGUES AND DATASETS IN EOSC

4.1. Metadata catalogues

The survey results from the questionnaire manifests a plethora of products types listed by research metadata catalogues (Annex D). It is not a surprise that the most common product types are Dataset and Publication, although these can be very domain specific, depending on the discipline addressed. A large majority of the catalogues contains metadata collected from third party data providers, the remaining ones are instead equipped with an internal infrastructure for maintaining the data.

A common metadata format (additional specific formats may coexist) is often used as a data model, though finding an agreement on common schemas is reported as one of the major difficulty encountered in the development of a metadata catalogue.

While all catalogues provide (or plan to) be able to be harvested from other catalogues or systems, their export methods may vary. Among the most common export formats, protocols and interfaces we can find the follow ones:

- DCAT (an RDF vocabulary designed to facilitate interoperability between metadata catalogues published on the Web)
- OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)
- RDF (Resource Description Framework); a general-purpose format for representing metadata on the web,
- Open Geospatial Consortium (OGC) protocols commonly based on ISO19115/139[41]
- RESTful API

For metadata catalogue implementations, custom frameworks based on Data Bases either Relational, NoSQL, RDF based and Search platforms (e.g. Apache Solr, ElasticSearch) are often used, though a significant percentage (about 25% of the sample) have chosen an already available Open Source digital repository platform such as CKAN[42] or EPrints[43]. Faceted search functions are recommended.

The survey revealed not all the metadata catalogues and data repositories have a programmatic interface or an open programmatic interface to expose metadata about themselves or their datasets. However, all of them provide a web GUI including search functionality.

According to the collected answers, development challenges are to be found in the heterogeneity of metadata records (model and agreement on common schemas), scalability and access control management. In operating and maintaining catalogues instead some of the common challenges are content growth at high quality (metadata collection and curation) and managing redundant entries (data de-duplication).

4.2. EOSCpilot data resources and datasets

The aim of the EOSCpilot Science Demonstrators[44] is to show the relevance and usefulness of the EOSC Services and their enabling of data reuse, to drive EOSC development. Each EOSCpilot demonstrator has specific goals. For instance, the "*Leveraging EOSC to offload updating and standardising life sciences datasets and to improve studies reproducibility, reusability and interoperability*" demonstrator aims to work

[41] https://www.iso.org/standard/26020.html
[42] https://ckan.org
[43] http://www.eprints.org
[44] https://eoscpilot.eu/science-demonstrators

on re-analysing of data and portability of tools and workflows. Independently of their specific goals, most of the demonstrators come with a set of third party datasets or data resources that need to be made available to a cloud infrastructure. In this case datasets from the EGA database[45] (Lappalainen et al. 2015) need to be accessed by tools and workflows in the cloud. It is within the remit of our work to propose recommendations to make the demonstrators' datasets more findable and accessible.

4.3. Datasets

All the research data resources we have explored organise their data records in datasets. The description of records tends to be specific for each discipline however the description of the dataset is quite similar. We therefore believe that datasets are a good starting point for improvements, and would allow EOSC users and services to find and access data, since it is common type, found across diverse scientific disciplines.

4.4. EDMI metadata guideline: EOSC Datasets Minimum Information

One of the outcomes of this work is a simple metadata guideline to help users to find and access datasets. We named this metadata guideline EDMI (EOSC Datasets Minimum Information). The EDMI metadata guidelines do not aim to be a new data model to describe datasets, but rather to complement existing data models (see Figure E). These guidelines will define the minimum metadata properties that should be present across existing data models, and which should be exposed by data resources, facilitating both users and programmatic services to locate and access data. The EDMI metadata guidelines thus aim to establish and encourage the adoption of a common and minimum set of metadata properties across different scientific domains, leveraging existing data models and access interfaces.

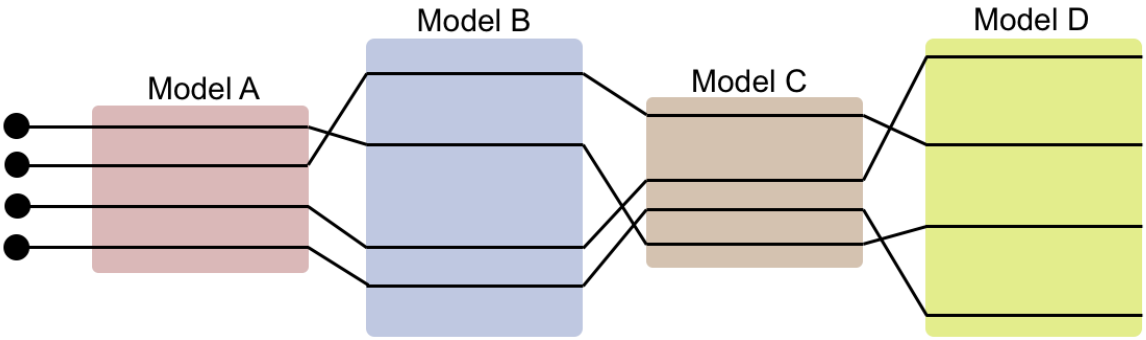


Figure E. Representation of a minimum set of metadata properties used in different data models. On the left, circles represent metadata properties (minimum properties) to be exposed to help EOSC services and users to find metadata. The coloured boxes represent different data models containing the minimum set of metadata properties (solid line).

One of the major goals of this work is to identify the common and minimum metadata properties important for consumers to find and access datasets. The main consumers identified in this work are users and services. In recognition of this distinction, we classify these metadata properties into two groups: Functional metadata and operational metadata. Functional metadata is the metadata of importance to users, especially researchers, and particularly for finding and accessing, for example, individual data records. This metadata includes properties such as the description and the methods used to generate the data. Operational metadata is of importance for services, particularly in locating and accessing, for example, a large dataset. This metadata includes properties such as the "type of interface" or the "URL to download the dataset" (see Table B).

[45] https://www.ebi.ac.uk/ega/home

Properties across different data models frequently have different names or descriptions, but with the intention to represent the same or a near identical concepts. For instance, Datacite[46] (Brase 2009) and the UKRDDS[47] use "title" while OpenAIRE and Schema.org[48] use "name" to describe the title or name of the dataset. Hence, it is clear that these concept names are being used interchangeably. The EDMI mapping activity helps users to know which properties should be used when working within the constraints of different metadata models, whilst remaining compliant with EDMI guidelines.

4.4.1. Selection process for minimal metadata properties (EDMI)

Two working groups were formed to identify minimum metadata with respect to: functional metadata and operational metadata. Both groups followed the same process to converge upon a preliminary selection and proposal for minimum metadata properties. These properties were identified in consideration of specific Use Cases, and the requirements deemed important by those stakeholders (Annex F). The focus was on the minimum properties necessary to find and access datasets. These parallel groups collected these properties in a table, and mapped those properties to dataset metadata properties in existing data models and resources (Annex E). Subsequently, the use cases requirements were mapped to the table of properties. The assignment of a property as being a 'minimal property' was through discussion evaluating the mapping exercise. This concerted effort attempted to strike a balance between users requirements, and what the data resources are currently provide, or able to provide in the future. In total six minimum functional metadata properties and six minimum operational metadata properties have been proposed as the core of the EDMI metadata guidelines (see table A and table B). An example of how to expose functional and operational metadata in JSON-LD[49] is available in "Annex I". The full list of metadata properties in available in "Annex H".

Functional metadata

The functional metadata mapping revealed good coverage of the EDMI proposed minimum metadata properties (Table A) across the evaluated catalogues and data models (see Annex E: Mapping of metadata properties). For those data models providing guidelines of minimum information like Bioschemas, the UKRDDS[50] metadata profile and Datacite, though not a complete overlap we found good alignment with the set of minimum properties.

Property	Description
name	A descriptive name for the dataset
description	A short summary describing a dataset
identifier	The identifier property represents any kind of identifier for any kind of dataset
creator	The creator/author of this dataset
dateCreated	The date on which the dataset was created
url	The location of a page describing the dataset

Table A. List of EDMI functional and operational minimum metadata properties.

Operational metadata

Though data resources have been very good at collecting requirements from users, many data resources do not include in their models the minimal operational metadata properties required for services to find and access datasets. For instance, something as trivial as the URL to download the dataset is missing in many cases and where it is present, it is not considered minimum or even recommended within the metadata

[46] https://www.datacite.org/
[47] UK Research Data Discovery Service, https://www.jisc.ac.uk/rd/projects/uk-research-data-discovery
[48] http://schema.org/
[49] https://json-ld.org/
[50] UK Research Data Discovery Service, https://www.jisc.ac.uk/rd/projects/uk-research-data-discovery

model employed by the data resource. One of the goals of this work is to test how important these properties are, and to increase awareness among existing data models and data resources.

Property	Description
license	A license under which the dataset is distributed
dateModified	The date on which the dataset was most recently modified
structure	The description of the structure of the dataset
dataStandard	The standard in which the content of the dataset is represented
accessUrl	The link to download the dataset
accessInterface	The type of interface to present the dataset

Table B. List of EDMI operational minimum metadata properties.

4.4.2. EDMI metadata guideline test and adoption

Our aim is to provide feedback to the communities and data resources maintaining dataset models and interfaces, recommending adoption of the EDMI metadata guidelines. However, we currently consider EMDI an early draft that needs to be tested and refined, ideally through feedback from data resources and consumers (users and services). In addition, before promoting this guideline, we would like to evaluate it; during the course of 2018, we intend to work with service providers and data resource volunteers on demonstrators to test the EDMI guideline (see demos section) with respect to being comprehensive and adoptable.

4.5. Interfaces to expose EDMI metadata

We conducted a survey, including 18 responses from metadata catalogues[51], which revealed around 30% of metadata catalogues do not provide a programmatic interface for services, while 100% of them provide a website to all users to browse the catalogue (Annex J). This shows many catalogues prioritise user requirements over requirements from services that need to access data programmatically. Though we highly recommend data resources to expose appropriate data through a programmatic interface, we acknowledge that not all of them have the resources or capacity to develop one. Thus, as an interim measure, we suggest that metadata catalogues and data repositories should at least expose structured metadata through a HTML mark-up vocabulary such as schema.org. This would allow services to access metadata programmatically, without placing undue burden upon catalogue or repository owners.

It is difficult and maybe counterproductive to recommend adoption of just one specific type of programmatic interface for all the metadata catalogues; we note many catalogues providing programmatic interfaces tailored to the needs of their community and we believe this is right. We also note that many metadata catalogues provide both an interface tailored to their domain needs, and a more generic and standard interface. For instance, OmicsDI[52] (Perez-Riverol et al. 2017) is a metadata catalogue which exposes a custom API[53], but also exposes the dataset metadata in a more standard way via schema.org.

[51] https://tinyurl.com/eosc-cat-survey-results
[52] https://www.omicsdi.org/
[53] https://github.com/OmicsDI/specifications

Our recommendation is that data resources provide at least one programmatic interface, and that their interface or underlying data model behind complies with the EDMI metadata guidelines. In those cases where this interface is a specific interface, we suggest adoption of an additional and more widespread method to expose metadata, for instance using a OAI-PMH[54] (Ward 2004) or schema.org.

4.6. Data models and standards

Minimum and common metadata is useful for data discovery and data access. However, it is important to highlight that data should be described beyond the minimum metadata, through the use of rich and domain specific metadata formats and guidelines. Rich metadata formats can be complex to adopt, but have the advantage of making data more "usable" by both humans and machines, that through a detailed and rich metadata description can filter, select, process, or even visualise data and data products in an appropriate way. Each scientific domain is working with standards to define their specific scientific entities[55]. We want to respect the existing formats and let RIs and scientific communities decide on how best to describe their data. In this work, we are looking for a set of minimum properties among existing models. This does not mean the models or providers have to stick just to the minimum. Custom or domain specific properties can be expressed on top of the minimum using existing formats. With this approach, we aim to satisfy requirements from EOSC to make data available to users and services and reduce the barriers to contribution from existing data providers.

We encourage the use of existing recommendations (control vocabularies and minimum information guidelines) described by existing data models and established communities. For each property proposed in the EDMI metadata guideline we aim to map to and reuse existing guidelines (generic and domain specific). Whenever possible, we will make suggestions on how to use controlled vocabularies, and provide guidelines to help with minimum metadata harmonisation for datasets across different scientific domains. This is work in progress which will be refined through collaboration with international partners and stakeholders, such as the RDA Metadata Interest Group (MIG)[56], and through demonstrators we aim to carry out in 2018 (see section 6: Demonstrators).

[54] https://www.openarchives.org/pmh/

[55] Over thousands standards registered in FAIRsharing

[56] https://www.rd-alliance.org/groups/metadata-ig.html

5. STRATEGY

The strategy proposed in this document is driven by the EOSCpilot data interoperability principles and the recommendations discussed and proposed in workshops by EOSCpilot partners and external stakeholders. This is not a final strategy but a first draft to be tested and reshaped as necessary through feedback from the EOSCpilot data interoperability demonstrators. This strategy is sustained by three main ideas:

1. The use of metadata catalogues to find and access data from 3rd party data resources.
2. The evolving EDMI metadata guidelines.
3. The development of a coordination strategy between existing dataset metadata catalogues.

5.1. Metadata catalogues and the EDMI metadata guideline

The “EDMI metadata guidelines” and the “dataset metadata catalogues” are two of the main components in this strategy to facilitate the findability and accessibility of data. Every scientific domain and many of the e-infrastructures represented in the EOSCpilot operates at least one metadata catalogue for datasets. Collectively, these catalogues play a fundamental role in facilitating the discoverability of scientific data. They integrate and harmonise metadata from different data resources helping user to search and find data. Metadata catalogues can be used by EOSC services and users not just to find the data but access the data from third party resources. To do so in the EOSC context, metadata catalogues need to provide sufficient metadata, namely, that complying with minimum requirements from EOSC users and services. These minimal requirements are described by the EDMI metadata guideline (see figure F).

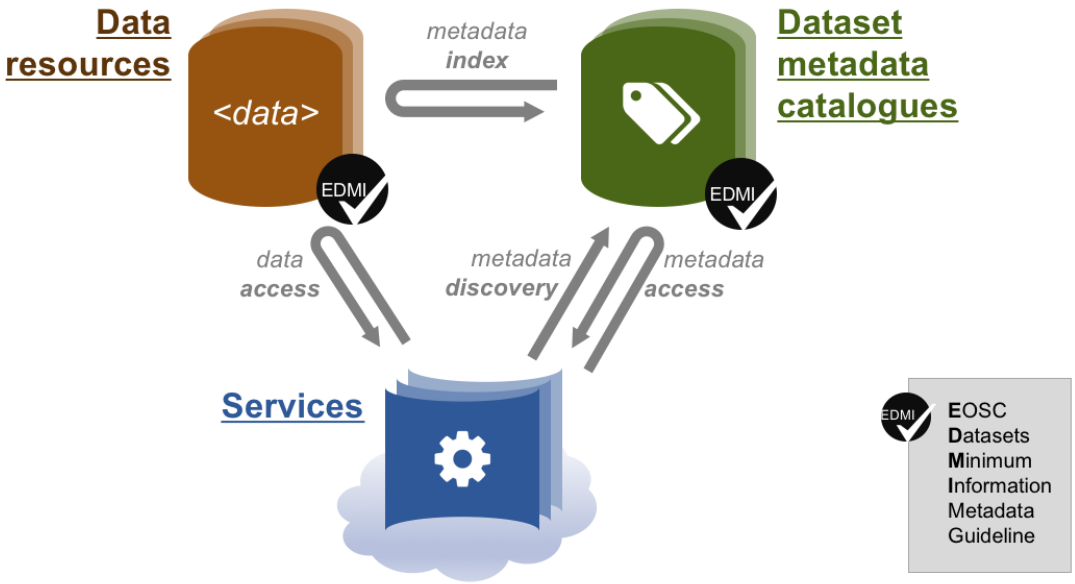


Figure F. Dataset metadata catalogues and the EDMI metadata guideline as major components of the data interoperability architecture. The metadata catalogues index metadata from 3rd party data resources. Each metadata catalogue recommends the best way for each resource to provide metadata to the catalogue, and can chose to make these recommendations compliant with EDMI. Services to find datasets will be able to use the programmatic interfaces exposed by the catalogues. If the metadata provided by the catalogue is compliant with the EDMI guidelines, the services will know they will have enough information to find and access datasets from third party resources.

5.2. Better coordination among existing dataset metadata catalogues

We believe in an ecosystem of coordinated metadata catalogues. There are many different catalogues for different purposes, covering different user needs, and collecting metadata at different levels. For instance,

some catalogues are specialised in specific content types like datasets, software and computer resources; and some others provide different coverage and granularity for a scientific domain e.g. science, life sciences or proteomics. We believe **there should not be just one EOSC metadata catalogue but an ecosystem supported by a sustainable and coordinated strategy** to provide users with a better service to find and access data.

When looking for a hotel, many users might use a search engine like Google[57] as a first entry point to find information. Though the information provided by a search engine might not be good enough to make a decision, it is good enough to direct the user to a more specialised site. This specialised site might be a hotel website or a hotel metadata catalogue like Booking.com[58] where the user counts upon sufficient functionality and additional information with which to make an informed choice. It is in the best interest of the specialised hotel catalogue to be indexed by the generic search engine which directs users to find it. At the same time, it is not the goal of the generic search engine to provide the customised functionality and information provided by the hotel catalogues. We believe our metadata catalogues should have a similar relationship, where generic metadata catalogues should leverage domain specific metadata catalogue.

5.2.1. Metadata registration

As explained in this report, metadata tends to be richer in specialised metadata catalogues than in generic catalogues. Thus, we believe the entry point of dataset metadata registration should be the metadata catalogues which are more aligned to the scientific scope of the dataset. For instance, a proteomics dataset should be registered in a catalogue like ProteomeXchange (catalogue of proteomics datasets) rather than in a generic catalogue like EUDAT-B2Find[59] or Figshare[60]. Generic catalogues should not encourage metadata registration in their own catalogues unless there is no domain specific catalogue where the dataset can be registered.

5.2.2. Metadata exchange

To facilitate the indexing and sharing of metadata into more generic metadata catalogues the metadata from domain specific catalogues should be available for harvesting via programmatic interfaces. This will help generic catalogues aggregating dataset metadata like EUDAT-B2Find. It will also be easier for generic catalogues to import integrated and harmonised metadata from domain specific catalogues, rather than directly from the source. To assure a minimum level of completeness and fulfil EOSC user and service requirements the metadata should follow the EDMI metadata guidelines (see Figure G). To facilitate the discovery of domain specific catalogues the generic metadata catalogues should acknowledge where the metadata came from.

[57] https://www.google.com
[58] https://www.booking.com
[59] http://b2find.eudat.eu
[60] https://figshare.com

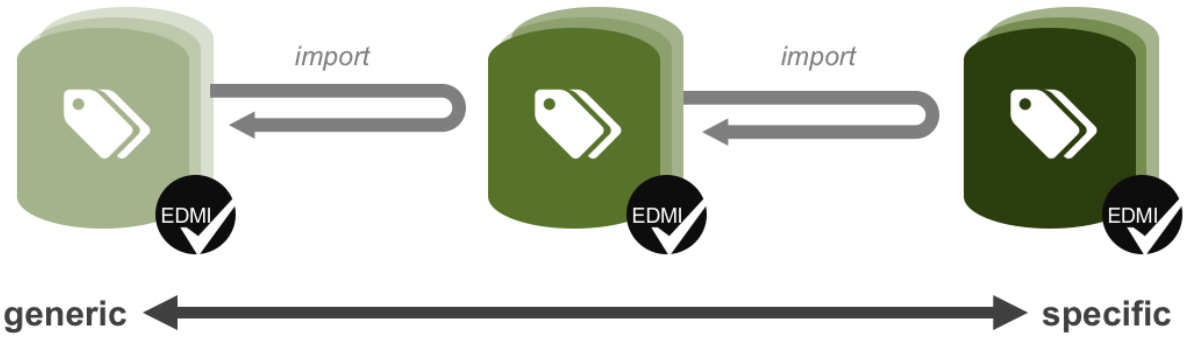


Figure G. Metadata catalogues and metadata exchange. Generic catalogues importing metadata from specialised catalogues.

5.2.3. Discovery of metadata catalogues and data resources

Defining the minimal metadata required to facilitate the finding and use of data across diverse resources is the key first step, but some key questions remain: How will services and users know which metadata catalogues are available to find and access research datasets? How will services and users know which data resources are indexed by a specific metadata catalogue? How will services and users know if catalogues, indexed data resources and datasets are compliant with the EDMI guidelines? How will generic metadata catalogues know which domain specific metadata catalogues they can import metadata from?

To help users, services, data resources and metadata catalogues to find metadata catalogues to use, import or submit dataset metadata we recommend the use of a catalogue of data resources and catalogues like FAIRsharing[61] (see figure H). FAIRsharing it is participating in a demonstrator to help find data resources and catalogues compliant with the EDMI guidelines. FAIRsharing interrelates standards with data resources (and with data policies), so as part of the demonstrator, will also be able to suggest which catalogues are more suited for the registration of dataset metadata, and will provide provenance relationship between catalogues and data resources, and between generic metadata catalogues and domain specific catalogues. For more information see section 6 - demonstrators activities planned for 2018.

Data resources and metadata catalogues, besides exposing metadata at the dataset level, will also need to expose metadata about their own resource. A resource can be considered as one integrated dataset and thus the properties provided at the dataset level are also valid to describe data resources and metadata catalogues. This is something that we will evaluate during 2018. This metadata will be especially helpful for other catalogues and services, enabling them to be aware of what interfaces, data models and updates are made in metadata catalogues and data resources

[61] https://fairsharing.org

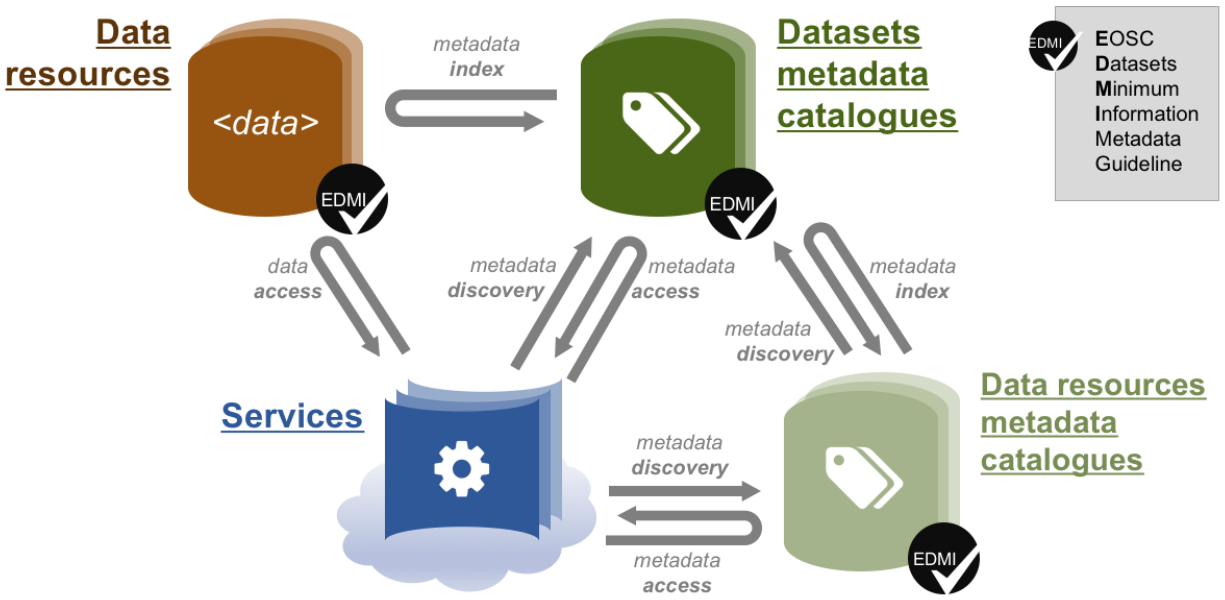


Figure H. Data resource metadata catalogue(s) to facilitate the discovery of the ecosystem of datasets metadata catalogues and data resources. FAIRsharing will be one of the data resource metadata catalogues and will participate in a demonstrator to test this strategy.

5.3. The EOSC metadata catalogues

Even when we think about a generic metadata catalogue for EOSC we should be open to the possibility to have more than one. Generic metadata catalogues might have been created for different purposes and might have different value to different users. At the moment in EOSC, we note partners supporting and maintaining dataset metadata catalogues with different functionalities, and with different objectives or target users. Some examples are eInfraCentral[62], OpenAIRE[63] and EUDAT-B2Find[64]. We believe there should be some coordination among EOSC catalogues at this level starting with the cross-referencing of their common entries. The generic metadata catalogues will be useful for EOSC services since they will provide an entry point to find dataset metadata, domain specific catalogue metadata and the individual data resources that host the data.

[62] http://einfracentral.eu
[63] https://www.openaire.eu
[64] https://www.eudat.eu/services/b2find

6. DEMONSTRATORS

These demonstrators are proposed as a result of the discussions and feedback collected during the open workshops hosted by the EOSCpilot data interoperability working group. The demonstrators aim to test and evaluate the feasibility of the recommendations proposed in this project. The feedback from these demonstrators will be used to make the necessary changes and improvements in order to have a practical data interoperability strategy by the end of the EOSCpilot project.

6.1. Findability and accessibility of datasets via functional and operational metadata

6.1.1. Introduction

Many services in science rely on data maintained in third party data resources. For these services, it is not easy to find, access, transfer and keep updated copies of the data hosted by those data resources. Neither is it currently possible to have a simple solution, since data resources will employ different data models, a diversity of interfaces, together with the highly distributed nature of the data itself. These are the main challenges for current services in finding the right operational metadata that can help them to manage data. The EOSCpilot data interoperability project focused on this problem, providing metadata recommendations and a strategy to make data from third party resources more findable and accessible for services.

6.1.2. Goal

This demonstrator aims to test how the functional and operational metadata proposed by this project, EDMI, will help services to find, access, transfer and replicate data available in third party data resources.

6.1.3. Objectives

- Involve at least two data repositories to adopt the recommendations of this project to expose dataset functional and operational metadata.
- Involve one catalogue of datasets to index and expose minimum functional and operational metadata from at least one data repository
- Involve at least one service to test the benefits of using the metadata proposed in this project.

6.1.4. Proposed participants

Several stakeholders volunteered to participate in this demonstrator, including:

- The PRIDE database, which hosts Proteomics datasets
- The OMICsDI catalogue of datasets, which hosts metadata about omics datasets
- The EUDAT-B2Find metadata catalogue, which indexes the metadata of scientific records

6.2. Discovery of compliant data resources and metadata catalogues

6.2.1. Introduction

Catalogues of datasets index and integrate metadata from data resources making it easier for users and services to have an overview of what data is available from data resources, and where it can be found. However, it is more difficult to ascertain which data resources have been indexed by a particular metadata catalogue and which resources are compliant with our recommendations.

6.2.2. Goal

This demonstrator aims to provide users and services a better overview of existing catalogues and data resources indexed by these catalogues. It also aims to recognise which catalogues and data resources

comply with the recommendations of this project.

6.2.3. Objectives

- Involve at least an existing catalogue indexing data resources to help users and services to find dataset catalogues and data resources compliant with the project recommendations
    - Create a collection of data resources per dataset catalogue
    - Associate the EOSCpilot recommendations to those catalogues and data resources complaint with the recommendations
- Involve a catalogue of datasets to register in a catalogue of data resources the list of data resources indexed and their compliance with the recommendations

6.2.4. Proposed participants

Several stakeholders volunteered to participate in this demonstrator, including:

- The FAIRsharing catalogue of databases, repositories, standards and data policies
- The OMICsDI catalogue of datasets, which hosts metadata about omics datasets

6.3. Research schemas for exposing dataset metadata

6.3.1. Introduction

As highlighted in our survey findings, around 30% of the metadata catalogues do not provide a programmatic interface that could help services to find and access data. Furthermore, they do not provide all the properties which are considered minimum in our EDMI recommendations. This demonstrator focuses on the need to provide a simple and quick way to implement a solution which allows metadata catalogues to expose this structured metadata. Schema.org (Mika 2015) provides a simple mechanism to expose structured metadata using the existing web interfaces of metadata catalogues and data resources.

6.3.2. Goal

We would like to explore how to use Schema.org in a manner akin to that used by Bioschemas[65] to facilitate exposing minimum scientific metadata. We call this “Research Schemas”.

Objectives

- Community
    - Start and support the Research Schemas community effort
    - Organise the first community meeting to engage the community and plan future activities
- Technical
    - Use Research Schemas as a vehicle to expose the minimum metadata properties proposed in the recommendations.
    - Recycle Bioschemas ideas to come up with a prototype of how to expose dataset metadata based on the recommendations.
    - Start to define profiles (metadata specification on top of existing schema.org types) based on the recommendations for scientific dataset and data catalogue
    - Come up with several examples to facilitate adoption
    - Test Research Schemas with one data resource to expose metadata
    - Test with one catalogue of datasets to expose metadata
    - Test with one catalogue of datasets to index schema.org metadata from a data resource.

[65] http://bioschemas.org

6.3.3. Proposed participants

Several stakeholders volunteered to participate in this demonstrator, including:

- The PRIDE database, which hosts Proteomics datasets
- The OMICsDI catalogue of datasets, which hosts metadata about omics datasets
- The EUDAT-B2Find metadata catalogue, which indexes the metadata of scientific records
- The Bioschemas.org community

6.4. Description and guidelines per metadata property

6.4.1. Introduction

The RDA Metadata Interest Group (MIG)[66] is working on providing detailed descriptions and recommendations for dataset metadata properties. The EOSCpilot wants to contribute to and reuse these descriptions and recommendations, with particular focus on those properties identified by the EOSCpilot as being part of the recommended set of minimum properties (EDMI).

6.4.2. Goal

Collaborate with the RDA MIG group to describe dataset metadata properties

6.4.3. Objectives

- Select a set of dataset properties of interest to start with (e.g. identifiers)
- Propose new properties if any EDMI property is missing in the RDA MIG dataset proposal
- Propose structure and template for properties to capture and harmonise feedback from the community.
- Contribute to the definition of the property, linking with existing guidelines (especially domain specific) and summarise recommendations

6.4.4. Proposed participants

Several stakeholders volunteered to participate in this demonstrator, including:

- RDA MIG group
- Identifier.org
- Nick Juty (ELIXIR/CORBEL UMAN)

[66] https://www.rd-alliance.org/groups/metadata-ig.html

7. FUTURE WORK

Partners from several disciplines have been involved in this work and feedback has been collected from several scientific domains. Still the current engagement is dominated by the life sciences. To make sure the recommendations and outcomes of this work are as interdisciplinary as possible we aim to engage resources and representatives from different scientific domains.

The timeline, plan and specific tasks proposed for the data interoperability task is outlined in "Annex K". The first year was specially dedicated to work on the Findability and Accessibility ('F' and 'A' aspects of the FAIR principles, respectively). Within the same scope next year, we will work on improving our guidelines regarding Interoperability and Reusability ('I' and 'R', respectively). We will cover ways to improve interoperability and reusability by looking at specific guidelines for each minimum metadata property. Some of these guidelines will be generic and domain specific and will cover topics like identifiers, and controlled vocabularies, licenses and provenance. Other important topics that require more discussion are: The metadata used to describe metadata catalogues, how to measure quality, specific profiles for domain specific datasets and the validation of EDMI guidelines.

7.1. Metadata about the catalogues and data resources

The EDMI guideline can be applied to catalogues and data resources, however it would be worthwhile to explore whether the EDMI properties are good enough at this level looking at user and service requirements.

7.2. Quality

The EDMI guideline could partly contribute to the evaluation of FAIRness and quality across several data providers of scientific datasets. Though it is still early to do this evaluation, we will need to consider how to measure EDMI compliance, and how this compliance correlates with FAIRness. There is a great deal of activity in FAIR Metrics (Wilkinson et al. 2017), including FAIRmetrics.org (Wilkinson et al. 2017) and recent investments by the NIH FAIR Data Commons[67].

7.3. Profiles

The properties defined by EDMI are meant to be minimal across datasets and between different scientific disciplines. However, each disciple has its own additional requirements beyond this minimal set, and would therefore need to define an extended set of minimum properties. For instance, for the geospatial domain, it is important to add properties capturing longitude and latitude, which might not be required minimum properties for other domains. These domain specific minimum properties could be an extension of the EDMI properties. It might be worth exploring whether EDMI should help to define domain specific properties, as domain profiles, or should just rely upon, and point to, existing minimum information guidelines.

7.4. Validation

Once the EDMI guidelines reach maturity, we will need a way to evaluate adoption within and across disciplines and infrastructures, as well as a way to ascertain specific levels of compliance on a per resource or dataset level. This will require a lightweight solution to perform validation.

[67] https://commonfund.nih.gov/bd2k/commons

8. REFERENCES

Brase, Jan. 2009. "DataCite - A Global Registration Agency for Research Data." In *2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology*. https://doi.org/10.1109/coinfo.2009.66.

"FAIR Principles for Data Stewardship." 2016. *Nature Genetics* 48 (4):343–343.

Jarnuczak, Andrew F., and Juan Antonio Vizcaíno. 2017. "Using the PRIDE Database and ProteomeXchange for Submitting and Accessing Public Proteomics Datasets." In *Current Protocols in Bioinformatics*, 13.31.1–13.31.12.

Lappalainen, Ilkka, Jeff Almeida-King, Vasudev Kumanduri, Alexander Senf, John Dylan Spalding, Saif ur-Rehman, Gary Saunders, et al. 2015. "The European Genome-Phenome Archive of Human Data Consented for Biomedical Research." *Nature Genetics* 47 (7). Nature Publishing Group:692.

Lord, P. W., A. MacDonald, R. O. Sinnott, D. Ecklund, M. Westhead, and A. Jones. 2005. "Large-Scale Data Sharing in the Life Sciences: Data Standards, Incentives, Barriers and Funding Models (The 'Joint Data Standards Study')." National e-Science Centre, 193.

Magrane, Michele, Michele Magrane, and UniProt Consortium. 2010. "UniProt Knowledgebase: A Hub of Integrated Data." *Nature Precedings*. https://doi.org/10.1038/npre.2010.5092.1.

McEntyre, Jo, Ugis Sarkans, and Alvis Brazma. 2015. "The BioStudies Database." *Molecular Systems Biology* 11 (12):847.

Mika, P. 2015. "On Schema.org and Why It Matters for the Web." *IEEE Internet Computing* 19 (4):52–55.

Ohno-Machado, Lucila, Susanna-Assunta Sansone, George Alter, Ian Fore, Jeffrey Grethe, Hua Xu, Alejandra Gonzalez-Beltran, et al. 2017. "Finding Useful Data across Multiple Biomedical Data Repositories Using DataMed." *Nature Genetics* 49 (6). Nature Publishing Group:816.

Perez-Riverol, Yasset, Mingze Bai, Felipe da Veiga Leprevost, Silvano Squizzato, Young Mi Park, Kenneth Haug, Adam J. Carroll, et al. 2017. "Discovering and Linking Public Omics Data Sets Using the Omics Discovery Index." *Nature Biotechnology* 35 (5). Nature Publishing Group:406.

Rettberg, Najla, and Birgit Schmidt. 2015. "OpenAIRE: Supporting a European Open Access Mandate." *College & Research Libraries News* 76 (6):306–10.

Singh, J. 2011. "FigShare." *Journal of Pharmacology & Pharmacotherapeutics* 2 (2). Wolters Kluwer -- Medknow Publications:138–39.

Ward, Jewel. 2004. "Unqualified Dublin Core Usage in OAI-PMH Data Providers." *OCLC Systems & Services: International Digital Library Perspectives* 20 (1):40–47.

Wilkinson, Mark D., Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos, and Michel Dumontier. 2017. "A Design Framework and Exemplar Metrics for FAIRness." *bioRxiv*. https://doi.org/10.1101/225490.

Wolstencroft, Katherine, Olga Krebs, Jacky L. Snoep, Natalie J. Stanford, Finn Bacall, Martin Golebiewski, Rostyk Kuzyakiv, et al. 2017. "FAIRDOMHub: A Repository and Collaboration Environment for Sharing Systems Biology Research." *Nucleic Acids Research* 45 (D1):D404–7.

9. ANNEXES

9.1. Workshop reports

- Annex A - BlueBRIDGE workshop: “FAIR friendly research data catalogues: How far are we?” - Workshop recommendations
- Annex B - How FAIR Friendly is your data catalogue? Exposing FAIR data in EOSC - Summary report
- Annex C - EOSCpilot data interoperability technical workshop: Data catalogues and datasets in the European Open Science Cloud - Summary report

9.2. Metadata catalogues

- Annex D - Description of the metadata catalogues surveyed
- Annex J - Survey analysis: Matrix comparing metadata catalogues

9.3. Metadata

- Annex E - Dataset metadata properties mapping
- Annex F - EDMI Metadata properties, use cases and mappings.
    - Functional metadata properties: use cases
    - Functional metadata properties: mappings
    - Operational metadata properties: use cases
    - Operational metadata properties: mappings
- Annex H - List of minimum, recommended and optional metadata properties
- Annex I - Example of how to expose functional and operational metadata

9.4. Other

- Annex K - Proposed timeline, plan and specific tasks for the EOSCpilot data interoperability task
- Annex L - Glossary

ANNEX L - GLOSSARY

Term	Explanation
AAI	Authentication and Authorization Infrastructure
CMS	Content Management System
EDMI	EOSC Dataset Minimum Information
EOSC	The European Open Science Cloud
FAIR	Findable, Accessible, Interoperable and Reusable
GSO	Group of Senior Officials
HPC	High-performance computing
HTC	High-throughput computing
IaaS	Infrastructure as a Service
JSON-LD	JavaScript Object Notation for Linked Data
LIMS	Laboratory Information Management System
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
RDA	Research Data Alliance
RIs	Research infrastructures
SDN	Software Defined Networking
UKRDDS	UK Research Data Discovery Service

ANNEX A - BLUEBRIDGE WORKSHOP: "FAIR FRIENDLY RESEARCH DATA CATALOGUES: HOW FAR ARE WE?"

FAIR friendly research data catalogues: How far are we?
April 3, 2017. 9th RDA Plenary Meeting. Barcelona, Spain

Workshop recommendations

The agenda in the BlueBridge website[68] provide more details about the event including access to the presentations and the summary outcome[69] of the workshop. This annex aims to highlight the recommendations proposed during the workshop.

Collective recommendations

This section reports the individual contributions of the participants who contributed to the report driven by three questions.

Question1: How could the EOSC catalogue facility be implemented? Through a single global data catalogue that gathers the metadata of all the published "data"? By harvesting metadata from the participating infrastructure data catalogues? Or what other model do you envisage as the most appropriate?

E3.1 - Catalogue of catalogues

All the contributors agree on the fact that EOSC should offer a data catalogue to its users and that it has necessarily be built as a "Catalogue of catalogues" where existing catalogues can be national, institutional, discipline or project specific ones;

E3.2 - Rely on work done by existing initiatives

There are several infrastructures and initiatives that are already making an effort to integrate data and metadata from multiple catalogues. They are adopting different solutions more or less based on shared protocols and standards. The EOSC Catalogue can rely on the work already done by these initiatives.

E3.3 - Low effort pragmatic solutions in the short term

It is certainly unrealistic at the moment to assume that the EOSC Catalogue can be built by asking to all the existing component catalogues to adopt common metadata standards and interfaces. Reaching this harmonization requires many changes and years of works from participating actors. This common solution may possibly be reached in the long term when the return of investment of sharing will be well understood. In the meantime, more pragmatic solutions, based on ad-hoc transformations and mediators that do not necessarily require considerable changes in existing catalogues should be put in place. In parallel, actions can be done to progressively introduce shared guidelines starting from very simple ones. Catalogues presented in the workshop showed few commonalities. Initial guidelines might leverage them.

E3.4 - Reuse and no create a new catalogue

Creating a new "catalogue of catalogues" is a solution that presents some risks: a) maintenance of the catalogue, b) creation and update of entries, c) choice of a model to store metadata in an appropriate way d) technical issues: is it really feasible? e) scalability and granularity: How to group and structure the metadata from the various sources?

[68] http://www.bluebridge-vres.eu/FAIR-workshop/agenda

[69] https://goo.gl/xBgZxu

E3.5 - Flexible metadata model

The model selected to store metadata should take into account both the capacity of representing "rich" metadata (i.e. metadata related to different concepts, as for instance users, datasets, catalogues, projects, equipment etc.), and the possibility of dealing with semantics in a smart way, (i.e. use multi-domain ontologies, or support the capability of representing mappings among different ontologies).

E3.6 - Common taxonomy

It would also be useful to define a common taxonomy to classify data (public and open data, private and big data, sensitive data, anonymized personal data).

E3.7 - Referring to domain specific catalogues

One potential disadvantage for a single catalogue of catalogues may be that community specific fields and 'search interfaces' could not be offered. It would be important to identify solutions that enable to refer to the more specific catalogue information when needed.

Question 2: Should EOSC aim at introducing a single, even if minimal, common metadata format that is used by each infrastructure to publish data outside its boundaries or should we introduce mediators between metadata formats as basic components of the EOSC architecture? The inputs collected from the contributors are quite diverse and reported below:

E3.8 - Minimum metadata should be extendable

The adoption of a minimal common metadata format with associated protocols can be useful in the case where it is of interest the discovery of the available resources. However, this must be 'extendable' by templates or something similar.

E3.9 - Encourage rich metadata and semantic metadata descriptions

As findable data is depending on rich use of metadata, a minimal format will not make it easier to find data. Working towards use of semantic metadata description will facilitate easier exchange of metadata between different formats. Richest metadata formats can be complex to adopt, but have the advantage to make the data more "usable" by both humans and machines, that through a detailed and rich metadata description can filter, select, process, or even visualise data and data products in an appropriate way.

With both the solutions proposed some common actions need to be performed:

E3.10 - Promote the adoption of existing metadata standards

To promote the adoption by Research Infrastructures and e-Infrastructures of already existing metadata standards (e.g. INSPIRE, OGC etc.), protocols and practices.

E3.11 - Promote the best practice of publishing metadata in multiple formats

To promote the best practice of publishing metadata in multiple formats thus to match different needs. Such formats include both community specific standards (e.g. Darwin Core), data type specific standards (e.g. ISO 19115), as well as community agnostic / generic Standards (e.g. Dublin Core, Schema.org)

E3.12 - Provide metadata in the format that work best

It is important to reduce the barriers to contribution to such catalogue as far as possible: a model where people provide metadata in the format that work best for them is the solution. Forcing people to provide data in a way decided by externals will reduce adoption.

Question 3: Currently each infrastructure has its own publication policies. Should EOSC impose a set of common policies on what, when and under which conditions data can be published in the catalogue?

• Common policies, particularly if reinforced by funding policies, can be very helpful. A clear set of guidelines and recommendations for the data providers should be envisaged with regards to the provided metadata and to the underlying data collections.

• The recommendation is that the common policies focus initially on ensuring that the data is FAIR. Disseminating and pushing to create a FAIR culture is the way to go if shared principles and publication policies must be adopted.

• Looking for greater consistency on data licenses would be the next thing to tackle.

• Already quite a lot of work has been carried out on publication policies, so it is recommended to maximise the re-usage of existing results. Indeed, only a balance between top down and bottom up approach, (i.e. the co-development approach) can ensure that solutions are agreed and finally adopted.

• EU-funded research could impose a common policy, but any other research should be able to have its own policy. However, it must be considered that imposing a set of rules might turn out in an action without results.

• Relying on clustering initiatives (e.g. ESFRIs) is probably a good opportunity to be sure that communities are involved and adopting policies.

Individual recommendations

This section reports the individual contributions of the participants who contributed to the report driven by three questions.

Question 1: How could the EOSC catalogue facility be implemented? Through a single global data catalogue that gathers the metadata of all the published "data"? By harvesting metadata from the participating infrastructure data catalogues? Or what other model do you envisage as the most appropriate?

• Massimiliano Assante, CNR: Assuming that each participating infrastructure provides its own data catalogue, the EOSC catalogue facility might be implemented as a Catalogue of Catalogues. It should offer to its users the possibility to transparently querying it also using specific metadata formats (or set of metadata formats) even if not natively supported by the underlying original catalogues. To enable this behavior in a so heterogeneous context as the one delineated by the existing catalogues a mix of technical solutions will have to be supported. These will have to combine harvesting into a central catalogue with distributed search and access facilities according to characteristics and policies of the interfaced catalogues.

• Daniele Bailo, INGV: In the framework of the EOSC is of course fundamental to have access to heterogeneous resources in a simple way. With this objective in mind, creating a new catalogue is one of the viable options. However, a serious discussion should be undertaken about the effectivity of this solution in order to match the objective (i.e. facilitate access to heterogeneous EOSC resources). Creating a new "catalogue of catalogues" is a solution that presents some risks or issues: a) maintenance of the catalogue, b) creation and update of entries, c) choice of a model to store metadata in an appropriate way. In this sense, previous to the question "should we create a new catalogue", a harmonisation activity that promotes the adoption of common metadata standards and interfaces, that in turn will enable existing catalogues to expose their metadata in a machine-understandable way, should be carried on. Such an initiative will improve interoperability of system. With this premise, also the adoption and creation of a new catalogue, becomes an action whose risks are mitigated: a) maintenance and b) updates of entries can be done in an automated way by harvesting metadata from participating infrastructure catalogues; c) one of the models now used to store metadata can then be adopted. Such a model should take into account both the capacity of representing "rich" metadata (i.e. metadata related to different concepts, as for instance users, datasets, catalogues, projects, equipment etc.), and the possibility of dealing with semantics in a smart way, (i.e. use multi-domain ontologies, or support the capability of representing mappings among different ontologies).

• Ramon Codina, Communications Maritime Hub: It is important first to define a common criteria science

taxonomy and classification data (public and open data, private and big data, sensitive data, anonymized personal data). I propose to include in EOSC a metadata about ecological footprint and biocapacity (see http://data.footprintnetwork.org). Into the EEZ (Economic Exclusive Zone) 200NM we propose to use our international initiative IaaS (Communication Maritime Hub) Connecting the Smart Sea (Oceanography observatories, Oceanographic buoys of EMSO ERIC, and marine rescue), Smart Port (Cruises and Sustainable Shipping and Port Logistics) and People in a Smart Maritime Hub with a very low latency and a coverage mobile broadband into the EEZ, see our propose in https://eu- smartcities.eu/commitment/2621.

• Harry Lankreijer, ICOS: Could one single global data catalogue be technically feasible and make data easy accessible. Harvesting metadata from other catalogues seems to be a better solution. Either way, rich metadata is essential.

• Andrew Treloar, ANDS25: The catalogue facility should bootstrap on existing endeavours. These might be national (such as the Dutch NARCIS), institutional, discipline or project. The catalogue should aim to harvest from these into a single catalogue, and remove any duplications along the way. A model for how to do this at a national scale that could be generalised is http://researchdata.ands.org.au/, run by the Australian National Data Service. The approach that we use and all the source code are freely available for adoption by EOSC if that is useful.

• Heinrich Widmann, EUDAT: One single global data catalogue would be the approach as implemented by EUDAT-B2FIND. The advantage is that users must access only one single entry point (interface) to search in a comprehensive and common search space. The disadvantage may be that you cannot offer community specific fields and 'search interfaces' and that you have to homogenize to one common schema. Other issues to be considered with this global approach are scalability and granularity: How to group and structure the metadata from the various sources?

Question 2. Should EOSC aim at introducing a single, even if minimal, common metadata format that is used by each infrastructure to publish data outside its boundaries or should we introduce mediators between metadata formats as basic components of the EOSC architecture?

• Massimiliano Assante, CNR: The EOSC catalogue facility cannot rely on a single, even if minimal, common metadata format so as to fall under the "Agreement-based" approaches for interoperability. There is a need to guarantee a high level of autonomy among the partaking infrastructures. Thus it is required to use approaches able to isolate the interoperability machinery and implement it in mediators between metadata formats.

• Daniele Bailo, INGV: When planning and promoting the adoption of European wide models, rules and standards, it is of great importance to take into account technical and social issues and also to focus on the objectives. The two options proposed in the questions are both interesting according to the goal they want to pursue. The adoption of a minimal common metadata format with associated protocols (for instance Dublin Core and OAI-PMH) can be useful in the case where it is of interest the discovery of the available resources. Then a manual refinement is required if a user wants to access the actual data (or - in general - resources). Richest metadata formats can be complex to adopt, but have the advantage to make the data more "usable" both by humans and machines, that through a detailed an rich metadata description can filter, select, process, or even visualise data and data products in an appropriate way. In order to match both objectives and maximise impact, some key principles might be outlined, for instance: promote the adoption by Research Infrastructures and e-Infrastructure of already existing metadata standards (e.g. INSPIRE, OGC etc.) o promote the best practice of publishing metadata both in rich metadata standards (sometimes very community specific) and in generic standards (Dublin core). With this premise, the creation of a mediator, which is a task to which much resources should be dedicated, can become simple and - with the adoption of appropriate metadata models - feasible. In any case, building on the experience of Research Infrastructure like EPOS, I think that the EOSC should be used as an opportunity to harmonise data, metadata, best practices and tools. Questions like "should we build a catalogue" or "should we build a mediator" are out of the scope at the moment. I think we should FIRST start from a common basis where all

RIs and e-Infrastructures adopt common standards, protocols and practices. With this premise, new scenarios will open up, where anybody (even skilled IT users) could harvest metadata, build their own mediators or applications (even mobile). Likewise, with the above premise, access to resources and building of catalogues will be simpler. The creation of a mediator would be facilitated and even several mediators could be built, according to the needs of specific communities and domains.

• Ramon Codina, Communications Maritime Hub: A common metadata format, or common criteria science taxonomy is basic. EOSC should aim at introducing this common criteria science taxonomy to publish data or open data. The EOSC architecture must define an API, and the controller rule must be mandatory in big data and open data. In case to use a sensitive data Binding European Research Council Rules must be mandatory to all European Research Council members26. See an example list of BCR at http://ec.europa.eu/justice/data-protection/international- transfers/binding-corporate-rules/bcr_cooperation/index_en.htm

• Harry Lankreijer, ICOS: Today the work done on metadata standards is going towards a certain common minimum. However, findable data is depending on rich use of metadata and thus a minimal format will not make it easier to find data. Working towards use of semantic metadata description will facilitate easier exchange of metadata between different formats.

• Andrew Treloar27, ANDS: The challenge in producing such a catalogue is that the catalogue producers care more about getting the data instead of caring about providing data. In other words, it's important to reduce the barriers to contribution as far as possible. In the early days of ANDS we were able to provide funding to data producers to do things "our way". Once we were no longer providing such funding, many of the feeds dried up. We needed to move to a model where people provided metadata in the format that worked best for them and we took on the work of converting this into what we needed. So, I would argue for the mediator approach.

• Heinrich Widmann, EUDAT: I would suggest a minimal, common metadata schema. However, this must be ‘extendable’ by templates or something similar. “Mediators between metadata formats” sound nice as well, but I have no idea how they should be implemented.

Question 3. Currently each infrastructure has its own publication policies. Should EOSC impose a set of common policies on what, when and under which conditions data can be published in the catalogue?

• Massimiliano Assante, CNR: Both the EOSC High Level Expert Group report and the successive GO-FAIR reports suggest choosing a lightweight integration at the level of EOSC. This is also confirmed by our experience in dealing with the federation of heterogeneous providers. The engagement rules may progressively become more prescriptive once the EOSC emerges as a useful and operational reality. For example, initial light rules might be limited to imposing the specification for any item in the catalogue of its terms of use and of a persistent identifier while all the other description fields might be optional both in the term of the format and in the used vocabulary.

• Daniele Bailo, INGV: Already quite a lot of work has been carried out on publication policies, so we should maximise the re-usage of existing results. Indeed, only a balance between top down and bottom up approach, (i.e. the co-development approach) can ensure that solutions are agreed and finally adopted. Imposing another set of rules might turn out in an action without results. In Europe, we already have INSPIRE regulations and indications. Creative Commons licenses are often adopted by many RIs. Additionally, for data publication in a commons we have OpenAIRE. Disseminating and pushing to create a FAIR culture is the way to go if shared principles and publication policies must be adopted. Finally, relying on clustering initiatives (e.g. ESFRIs) is probably a good opportunity to be sure that communities are involved and adopting policies.

• Ramon Codina, Communications Maritime Hub: It’s important the rule of the Data Privacy Officer (DPO) in all European Research Council members if EOSC were to use a model of DaaS (Data as a Service) and compliance with the F.A.I.R. data principles. The rule of the DPO is to ensure to the digital society the best

practice about controllers and processors, in a model of DaaS (Data as a Service) and compliance with the F.A.I.R. data principles.

• Harry Lankreijer, ICOS: If the aim is to obtain as many data as possible for re-use in the data catalogue, the researcher should be motivated to publish the data. EU- funded research could impose a common policy, but any other research should be able to have its own policy. However, researchers should be motivated to publish by seeing the benefits of it: increased chances for funding. However, this will need also a good system for citation to published and downloaded data.

• Andrew Treloar, ANDS: Common policies, particularly if reinforced by funding policies, can be very helpful. I would recommend that these policies focus initially on ensuring that the data is FAIR. Looking for greater consistency on data licenses would be the next thing to tackle.

• Heinrich Widmann, EUDAT: Yes, at least regarding the provided metadata there should be a clear set of guidelines and recommendations for the data providers. Another thing are the policies of the underlying data collections, e.g. the access permissions of the data resources may differ between the infrastructures - and that's ok, as long this is clearly specified in the metadata (e.g. in a field 'Licences' or 'Rights').

ANNEX B - HOW FAIR FRIENDLY IS YOUR DATA CATALOGUE? EXPOSING FAIR DATA IN EOSC

How FAIR Friendly is your data catalogue? Exposing FAIR data in EOSC
September 8, 2017. Open Science Fair 2017. Athens, Greece

Summary report

Introduction

Research communities and specially research infrastructures are making a concerted effort to homogenize, collect their (meta)data and publish them in the open through community specific data catalogues. This is a good start towards making data FAIR, but how can we ensure availability of domain specific FAIR data and data-analysis services through a common virtual research environment like the European Open Science Cloud (EOSC)? From vertical domains (e.g., research infrastructures) to horizontal approaches (e.g., OpenAIRE, DataCite) which cover national settings and libraries/repositories, we see different content, data models, interfaces, frameworks, architectures and vocabularies being used.

The EOSCpilot data interoperability task aims to establish principles, propose recommendations and demonstrate how FAIR data hosted by domain specific data repositories and catalogues can be exposed to EOSC to be used and reused by EOSC services, repositories and users.

Goal, objectives and structure

The workshop had the goal to provide an update of the activities of the EOSCpilot data interoperability working group and engage different stakeholders to shape the work of this group. This workshop was a follow-up of the BlueBridge workshop held on April 3 at the RDA meeting. The workshop was structured into two sessions. In session 1 scene setting presentations on EOSC were followed by short presentations by eight data catalogues representing subject-specific and generic systems and a review of a previous meeting. Prior to this workshop the organisers conducted a survey of 11 data catalogues and an early analysis was presented in session 2 followed by extensive breakout discussions of 13 principles of Data Catalogue metadata exposure and interoperability.

Outcomes

The participants got an overview of EOSC, the EOSCpilot project and the direction and scope of the EOSCpilot data interoperability working group. The presentations, the survey and the discussions contributed with feedback to the group about how data catalogues can contribute to make FAIR data available into EOSC. The overall workshop contributed to define a set of principles that will drive the work of the EOSCpilot data interoperability working group and the recommendations we will proposed for EOSC.

Guiding principles

We have defined the scope of the EOSCpilot data interoperability task following the guiding principles mentioned below. These principles will drive the work of this task and the recommendations we will propose to EOSC. These principles have been created based on the feedback collected from EOSCpilot data interoperability partners, surveys and EOSCpilot workshops like the BlueBridge workshop and the Open Science FAIR workshop.

Reuse: Leverage the rich legacy of Research Infrastructures

P1 - Making data FAIR is the responsibility of the Research Infrastructures and their data repositories

The role of the EOSCpilot data interoperability working group is not to define how to make data FAIR but to define and demonstrate a simple data interoperability architecture to expose FAIR data to EOSC services and EOSC users. We believe the responsibility of defining how to make data FAIR lies on Research Infrastructures (and e-infrastructures), especially on their participant data repositories. Moreover there is already a working group funded by the European Commision which started in parallel to define a roadmap to make data FAIR across data repositories.

P2 - We must rely on research infrastructure data catalogues

Many data repositories exist per scientific domain. Domain specific research infrastructures maintain data catalogues which collect, integrate, harmonise and enrich metadata from many dispersed and diverse data repositories to facilitate data discovery. We plan to rely on existing metadata catalogues as a main providers of scientific metadata for EOSC. We expect domain specific data catalogues will collect metadata from relevant data repositories.

P3 - We must support an ecosystem of catalogues

We believe in an ecosystem of coordinated data catalogues where domain specific data catalogues collect specific metadata from data repositories and generic data catalogues collect a subset of metadata from domain specific data catalogues. Ideally the generic data catalogues should pull information from specific data catalogues and recommend metadata submission to domain specific catalogues.

P4 - We should provide quality recommendations to feedback to RIs

With the analysis of data catalogues, metadata models and standards we aim to provide recommendations about how to improve the quality of the metadata provided by data catalogues and data repositories.

Least: The least possible metadata for the most benefit

P5 - Findability should come first

Findability is the first step to make data FAIR and the main condition to access and reuse data. We will focus on how EOSC services and EOSC users can find data taking into account their access, interoperability and reusability requirements.

P6 - Common and minimum metadata

We do not aim to create a new data model to describe datasets or data repositories but create a recommendation of minimum metadata properties common across data catalogues. Properties that will help EOSC services and users to find data repositories and datasets and will facilitate data access, interoperability and reusability. We aim to evaluate existing data models and recommend how to expose data reusing one or several data models.

P7 - Focus on common data types: datasets and data repositories

We will focus our work on few data types which are common across different scientific disciplines to start with. These data types are datasets and data repositories.

P8 - Flexible metadata models to embrace domain specifics

Each scientific domain work with standards to define their specific scientific entities which might or not be described with a standard format. We want to respect the existing formats and let research infrastructures and scientific communities decide on how better describe their data. We are looking for a set of minimum properties among models but we should be flexible enough to allow space for custom or domain specific properties.

P9 - Service requirements and operational metadata first class citizen

It is about the scientific metadata but also importantly about the operational metadata required by services to be able to find, access and use the data.

Practical: Sustainable and pragmatic delivery

P10 - Engage EOSC demonstrator data repositories

Most of the EOSC demonstrators involved datasets at least from one data repository. We will engage these data repositories to demo how their datasets can be discovered and accessed in EOSC via data catalogues.

P11 - Propose methods to expose metadata

We will evaluate existing methods and guidelines to expose metadata and propose one or more technologies to expose in data catalogues minimum and common properties. We will rely on work done by initiatives like RDA and GO-FAIR as well as the expertise of our data catalogues.

P12 - Simple to implement, easy to sustain

Any proposed solution should be looking at a high impact low effort strategy specially in the short term. It should be simple to implement and easy to maintain providing just enough functionality to facilitate discovery, access and use of data in the EOSC.

P13 - Deliver guidelines and demonstrators

The outcome of our work will be a report but also a set of guidelines, an architecture proposal and demonstrators applying our recommendations and showing the feasibility of our proposed strategy to make FAIR data findable, accessible and reusable in EOSC.

Recommendations

Several recommendations were made during the workshop which should be considered by the EOSCpilot data interoperability task.

E4.1 - Clarity on terminology

Need to clarify the distinctions between data catalogues, metadata catalogues, registries, etc.

E4.2 - Define relationship

Need to define relationship between generic and domain specific data catalogues and data repositories

E4.3 - Validation

We will need validation on top of existing data resources and data catalogues to evaluate the adoption of minimum information. This would help data providers and metadata catalogues.

E4.4 - Balance between mandating and being flexible

Mandating minimum information but being flexible with the data models and interfaces that people might want to use to expose and share metadata.

E4.5 - EOSC incentives

We should highlight what are the benefits of adoption this recommendations for different stakeholders.

E4.6 - Beyond minimum

Minimum is important and should be our priority however we should also allow data providers expose metadata beyond minimum. The MoSCoW method (Must have, Should have, Could have Won't have) could be used to prioritise metadata properties.

E4.7 Push and pull models

We should probably support both models to collect metadata, however we should prioritise the pull model where data is exposed and open to any service.

E4.8 - Finding use cases

We should have several use cases for finding and accessing data.

E4.9 - De-duplication

Aggregating data from different data catalogues will lead to uncontrolled duplication. We should have a de-

duplication strategy and reuse existing de-duplication functionality like the open provided by OPENAire.

E4.10 - Multiple entries to search datasets

We should not just consider a catalogues of catalogues but multiple entries to search datasets via different catalogues.

E4.11 - Engage with RDA groups with similar interest

The Metadata Interest Group https://www.rd-alliance.org/groups/metadata-ig.html and the Data Discovery Paradigms Interest Group https://www.rd-alliance.org/groups/data-discovery-paradigms-ig

E4.12 - Start with data provided by RIs

Not all the data is provided or collected by research infrastructure data repositories and not all the data is indexed by research data catalogues. Though it is a good to start with the data hosted by RIs we should not forget about the rest.

E4.13 - Simple to implement easy to sustain

Generate the metadata as much as automatic as possible, validate them as much as automatic as possible. In this case deep learning algorithms may be of help.

E4.14 - Pointers and training to understand data

It's not straightforward to work with data from specialised domains. We need pointers and training on how to understand the data rather than specialist domain data.

References

The agenda including links to presentations, raw notes and more specific points raised during the workshop discussion are available in http://tinyurl.com/osf-eosc-datacat

ANNEX C - EOSCPILOT DATA INTEROPERABILITY TECHNICAL WORKSHOP: DATA CATALOGUES AND DATASETS IN THE EUROPEAN OPEN SCIENCE CLOUD

Summary report

Introduction

The European Open Science Cloud (EOSC) aims to provide by 2020 a virtual environment bringing together services and data from publicly funded research. The EOSCpilot is a two year project to support the first phase of development of EOSC. The EOSCpilot data interoperability task aims to establish recommendations and demonstrate how FAIR data can be exposed to EOSC to be used and reused by EOSC services and EOSC users. This workshop was organised with the goal to evaluate existing approaches to describe, expose and integrate dataset metadata. During this workshop we also provided an update of some data catalogues, data models used to describe datasets. We did also evaluate requirements from researchers and infrastructure services. The workshop was participated by 47 people including service providers, representatives from data catalogues, research infrastructures, e-infrastructures and partners from the EOSCpilot project.

The first day was dedicated to updates and it was dominated by presentations and discussions. It was organised in 7 topics: 1.- EOSC and EOSCpilot, 2.- Progress of the EOSCpilot data interoperability task, 3.- Data catalogues, 4.- Datasets in EOSCpilot demonstrators, 5.- Dataset models, 6.- Requirements from e-infrastructure services, and 7.- Methods to expose dataset metadata.

The updates were complemented with a hand-on session splitted in 3 working groups to collect feedback from participants. The data catalogue and data repositories group did focus on mapping dataset models and collecting more information from data catalogues. During this session, this group managed to include new 10 dataset mappings and add 5 new data catalogues to our list. The researchers group did focus on reviewing requirements and use cases from researchers identifying minimum functional metadata properties. The services group did look into the dataset requirements from infrastructure services. This group did come up with a set of service use cases to evaluate minimum operational metadata properties.

Our second day was dedicated to learn different technologies to help exposing dataset metadata. Four hands-on tutorials were organised to expand on some of the technologies presented the day before. RDA MIG Metadata Element Set, Bioschemas, EUDAT-B2FIND, GO-FAIR. This session helped to understand how to better align, support and push forward existing efforts and ideas discussed during the meeting.

Outcomes

During this workshop we came with a set of set of recommendations and actions to drive our work.

Recommendations

Minimum metadata model

E5.1 - Rely on existing dataset metadata standards

EOSCpilot do not aim to create another dataset standard but recommend user, services and data providers to reuse existing dataset metadata standards. Still we want to identify the minimum properties important to describe datasets in EOSC. We also want to contribute with recommendations of how to expose minimum properties. To start with and for testing purposes we will work with metadata models we have expertise with and those willing to participate in this project. These metadata models are: DCAT, the metadata profile for the UK Research Data Discovery Service, DATS, the RDA MIG dataset model and the schema.org dataset model.

E5.2 - Minimum at different levels

We should focus first on minimum common properties to cover any type of scientific datasets. On top we should work on profiles defining minimum properties for domain specific datasets. eg. for the geospatial domain it is important to define longitude and latitude properties which might not be minimum for other domains.

E5.3 - Recommendations and descriptions at the level of property

The RDA Metadata Interest Group wants to provide detailed descriptions and recommendations for dataset metadata properties. The EOSCpilot should contribute and reuse these descriptions and recommendations per property specially focusing on the recommended set of minimum properties agreed by the EOSCpilot.

E5.4 - Research schemas

Bioschemas is an effort built on top of schema.org aimed at facilitating the discovery of content of life sciences data. It proposes a simple generic model reusing schema.org types and a set of domain specific profiles. We would like to explore how to use this approach for other scientific domains where the generic model could be shared among different scientific domains and specific profiles could be developed tailored to each scientific domain. The generic model could be maintained as part of Research schemas and the profiles could be maintained by each specific domain.

Methods to access structured metadata

E5.5 - Programmatic access of structured metadata

Not all the data catalogues and data repositories have an API or Webservice to expose metadata about themselves or their datasets however they all have a web GUI. Though we really recommend data resources to expose their data through a programmatic interface, we acknowledge no all of them have the resources or capacity to develop one. Thus we suggest at least data catalogues and data repositories expose structure metadata via a HTML markup vocabulary like schema.org which still allows services to access structured metadata programmatically.

E5.6 - Rely on existing existing technologies to expose dataset metadata

EOSCpilot data interoperability do not aim to create an API or web service but recommend user, services and data providers to reuse existing technologies to expose dataset metadata. Still we would like to influence existing technologies to expose the minimum properties important to describe datasets in EOSC. To start with and for testing purposes we will work with metadata models we have expertise with and those willing to participate in this project. These technologies are OAI-PMH and schema.org markup.

E5.7 - Support an ecosystem of catalogues

We did reinforce the idea we need to have a better coordinated strategy between domain specific and generic metadata catalogues. We should work on recommendations for collecting and sharing metadata among data catalogues.

Collaboration

E5.8 - Working with the W3C Data Exchange Working Group

Though the purpose of the groups are different the EOSCpilot working group should provide feedback to the W3C Data Exchange Working Group (DXWG) that is currently revising DCAT. The recommendations from DXWG should also be considered by the EOSCpilot data interoperability working group.

Not within the scope

E5.9 - Measuring FAIRness

This group does not aim to work or evaluate how FAIR a dataset is. There already several groups working on this topic. Still we hope the minimum information defined by this project will contribute to evaluate how FAIR a dataset and a data resource is.

E5.10 - Making data FAIR

As agreed previously making data FAIR is not within the scope. Research Infrastructures and other working groups are already working on this topic.

Demonstrators

E5.11 - Steps within the demonstrator

Our demonstrator will need to prove dataset metadata can be used by services to be able to effectively discover and access data. This demonstrator will include several steps. 1.- Registration of EOSCpilot dataset metadata in a domain specific data catalogue. 2.- Metadata sharing with other data catalogues and 3.- Use of metadata by a service to find and access the data.

E5.12 - Indexing datasets from data catalogues

We will use EUDAT-B2FIND as a test case to access and index metadata provided by research infrastructure data catalogues. EUDAT-B2FIND will start a demonstrator focusing on minimum metadata prvideed by OAI-PMH interfaces and schema.org markup.

E5.13 - Registries of data resources

Registries of data resources can help finding data catalogues, data resources and data resources indexed by data catalogues. They could also highlight catalogues and resources compliant with a minimum metadata recommendation provided by EOSC. We plan to pilot with FAIRsharing the discovery of catalogues and resources compliant with EOSCpilot data interoperability recommendations. FAIRsharing will also be able to provide useful information and links to standards and data policies adopted by data resources.

References

The agenda including links to presentations, raw notes and more specific points raised during the workshop discussion are available in http://tinyurl.com/cats-dats

ANNEX D - DESCRIPTION OF THE METADATA CATALOGUES SURVEYED

Summary of the metadata catalogues involved in the survey[70] [71] performed by EOSCpilot data interoperability. Eighteen catalogues participated, listed in alphabetical order:

1. **ACTRIS data portal: http://actris.nilu.no**: the catalogue lists atmospheric variables related to aerosols, clouds and trace gases in situ, columnar and profiling observations. Metadata standards are established for each topical datacenter. Aerosol and cloud profiles adopted NetCDF data with embedded metadata. For in situ data NASA-AMES 10'00 standard is used. These standards were selected for accomplishing CF standard and integration with companion datasets at global scale, so in compliance with WMO GAW and GALION typically used formats. All these aspects facilitate the application of FAIR principle.

2. **ARIADNE** (http://www.ariadne-infrastructure.eu): the catalogue lists description of various items pertaining to archaeological research, such datasets, images, maps, collections, structured databases. The catalogue data model in use is the ARIADNE Catalogue Data Model (ACDM), an extension of the Data Catalogue Vocabulary (DCAT), a recommendation of the W3C Consortium (http://www.w3.org/TR/vocab-dcat/). The content-providing ARIADNE partners mapped descriptions of their content to the ACDM in a number of ways, through this mappings a specific service (MORe aggregator) imported and published the metadata into the catalogue;

3. **Astronomical Virtual Observatory [(A)VO] Registry of Resources** (http://rofr.ivoa.net): the catalogue contains astronomical and astrophysical domain collections of resources: images, spectra, multidimensional cubes, listings of astronomical sources with annotations, numerical simulations and models, as well as services deployed on top of them for access purposes; VOResource is used as metadata standard, it takes its start out of the OAI-DC set of metadata and specializes on it. VOResource elements map easily into the DataCite metadata format. The metadata integration is done by means of OAI-PMH protocol.

4. **Bio.tools** (https://bio.tools)**:** the catalogue includes all types of bioinformatics tools - application software with well-defined data processing functions. This ranges from simple tools with a single primary function, to complex, multimodal tools with many distinct functions. The catalogue data model in use id the biotoolsSchema, a description model for bioinformatics software defining 50 important scientific, technical and administrative attributes.

5. **BlueBRIDGE** (http://www.bluebridge-vres.eu): the catalogue comprises biodiversity environmental data such as geo-referenced chemical and physical variables, marine species distribution maps, observations, methods/algorithms, stock and fishery records, but also to Virtual Research Environments and research objects generated through the use of D4Science infrastructure. The catalogue data model in use is the BlueBRIDGE Core Metadata Model (BCMM), an extensible model which allows to support multiple (specialised) Metadata models. The metadata integration from external repositories is done by means of the Open Geospatial Consortium (OGC) CSW protocol. Users can instead publish their products by means of specialised GUIs or RESTful APIs;

[70] http://tinyurl.com/eosc-cat-survey

[71] https://tinyurl.com/eosc-cat-survey-results

6. **CLARIN** (https://www.clarin.eu): the catalogue contains language resources and language studies in general, It aggregates the metadata about the resources from over 60 data providers, containing more than 900.000 records. The catalogue data model in use is the Component Metadata Infrastructure (CMDI), a modular meta model allowing to defined custom schemas. Products are weekly harvested from all CLARIN centres via OAI-PMH;

7. **CMS Collaboration:** (https://cms.cern/collaboration)**:** the catalogue lists description of High Energy Physics events. Descriptions include the origin of data from either Monte Carlo simulation or collected by detectors, configurations used to extract derived data, as well as settings used to create simulated events, where applicable. Products are not associated with a common metadata format. Instead, the common data format JSON is utilised to expose the associated descriptions of products. The products are stored distributed in the Worldwide LHC Computing Grid (WLCG) on infrastructure pledged to the CMS collaboration by computing centres;

8. **DataMed** (https://datamed.org): the catalogue lists metadata records of biological datasets. The data repositories covered in this initial release have been selected by the bioCADDIE team and represent only a small sample of biomedical data. The catalogue data model in use is Biocaddie DATS;

9. **EPOS** (https://www.epos-ip.org): the catalogue provides access to an heterogeneity of environmental data, data products and software, ranging raw data such as seismograms and accelerograms to strain maps, earthquake source models and integrated data products coming from complex analyses or community shared products (hazards maps, catalogue of active faults, etc). The catalogue data model in use is CERIF (Common European Research Information Format) model.

10. **EUDAT-B2FIND** (https://www.eudat.eu/services/b2find): this catalogue is a catalogue of catalogues "federating" interdisciplinary research data, coming from different initiatives and communities such as DARIAH, CLARIN, GBIF, IVOA and others. The The catalogue data mode in use is the common B2FIND schema based on DataCite Metadata schema. Products are harvested via OAI-PMH;

11. **FAIRDOMHub** (http://www.fair-dom.org): the catalogue lists description of datafiles (including omics, images, specimens, strains and samples) models of different kinds (SBML, MATLAB, software etc), Standard Operating Procedures, publications, presentations, genetic parts, devices and plasmids. The products in the catalogue are associated with metadata that adhere to standards of their type: for example, SBML for Systems Biology Models; SBOL for Synthetic Biology; the Minimum Information models (e.g. MIAPE) omics standards for data etc. Across these we have the Just Enough Results Model (http://jermontology.org), a common metadata to describe the interrelations between assets in the catalogue and the metadata fields required to describe them. DCAT is being reintroduced as a mechanism for other FAIR resources to harvest content. The content is imported and exported using the Research Object model;

12. **FAIRSharing** (https://www.fairsharing.org): the catalogue describes and links standards, databases/data repositories and data policies across disciplines. It allows users to see which resources are maintained, used and endorsed by the community, be they researchers, data managers, journal publishers, funders or the developers/curators of the resources themselves. The catalogue data model in use is the BioDBcore standard. All FAIRSharing data is available in JSON

and can be accessed via a number of different metadata formats.

13. **ICAT STFC (https://www.isis.stfc.ac.uk/Pages/ICAT.aspx)**: the catalogue contains Experimental Data Generated at Large-scale Analytic Facilities (e.g. ISIS, DLS, ESRF). The data files produced at the aforementioned facilities are captured and catalogued along with the metadata about the sample conditions for that experimental run, and the metadata from the proposal. The catalogue data model in use is the Core Scientific Metadata Model (CSMD).

14. **OMICsDI** (http://www.omicsdi.org): the catalogue is an open-source platform that can be used to access, discover and disseminate omics datasets. OmicsDI integrates proteomics, genomics, metabolomics and transcriptomics dataset metadata from several experimental data repositories. Datasets in OmicsDI are described using the OmicsDI XML format following the OmicsDI XML Schema. The catalogue exposes it content through a RESTful API. It is an open source project distributed under the Apache License Version 2. It is accessible at http://www.omicsdi.org and the source code at https://github.com/OmicsDI

15. **OpenAIRE** (https://www.openaire.eu): the catalogue lists metadata records of publications and datasets. Publications are included only if Open Access or linked to a research grant. Datasets are included only if linked to a publication in OpenAIRE ("in-context" datasets). The catalogue data model in use is the OpenAIRE data model, a common metadata format as provided by the OpenAIRE guidelines (guidelines.openaire.eu). In addition to the common metadata format, a number of specific metadata are associated with the products including Dublin Core and DataCite. OpenAIRE aggregates more than 25 millions of metadata records from more than 2,700 data sources featuring 3 workflows for metadata aggregation. Harvesting frequency may vary from repository to repository, weekly on average;

16. **OpenMinTeD** (http://openminted.eu): the catalogue lists language resources that can be used for Text and Data Mining purposes distinguished in the following resource types: scholarly publications, corpora, tools and web services, lexical/conceptual resources and language descriptions. The catalogue data model in use is the OMTD-SHARE metadata schema, links to other popular metadata schemas are also included in the schema.

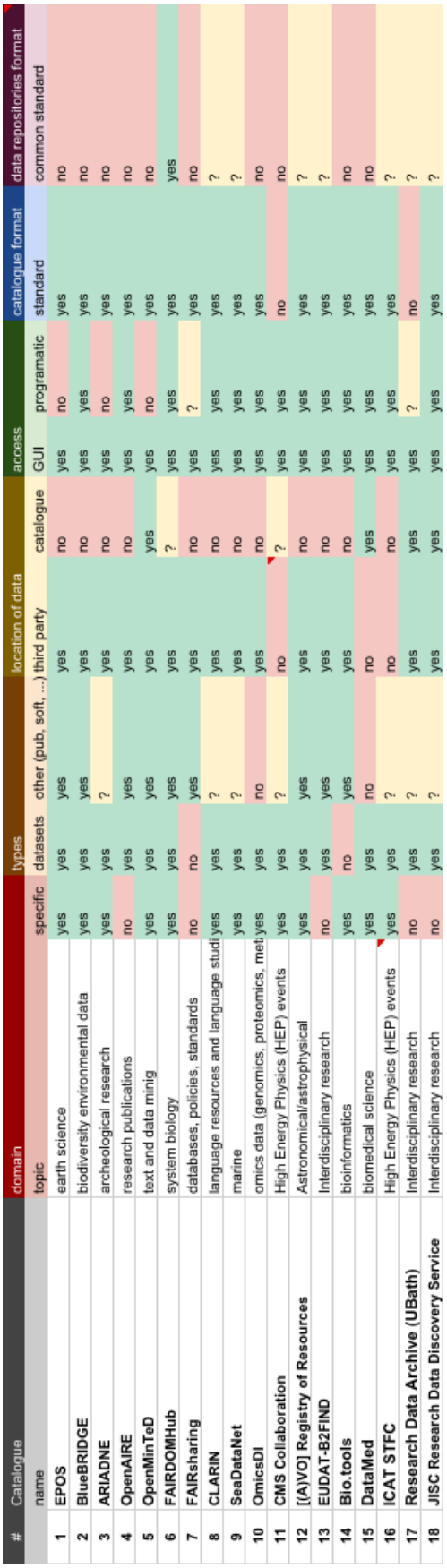
17. **Research Data Archive** (https://researchdata.bath.ac.uk): the catalogue contains high-quality research Datasets produced by members of the University of Bath. Datasets published in the catalogue have a minimum standard of metadata that ensure that data are described and discoverable. Dataset records are checked by the University of Bath Library for presence of appropriate metadata. The catalogue data model is a custom metadata scheme adapted from EPrints ReCollect, designed for compatibility with DataCite.

18. **JISC Research Data Discovery Service** (http://researchdiscoveryservice.jisc.ac.uk): the catalogue (still in beta version) lists multidisciplinary datasets and repositories. Metadata is mapped to an internal core metadata format from a number of supported schema (DataCite, EPrints, MODS, OAI-PMH, Figshare, GEMINI2) named UKRDDS.

19. **SeaDataNet (**www.seadatanet.org): the catalogue lists descriptions of datasets (including geographical coverage and other metadata features), collections, data observing, associated organisations connecting more the 100 data centres. The core directory service, the Common Data

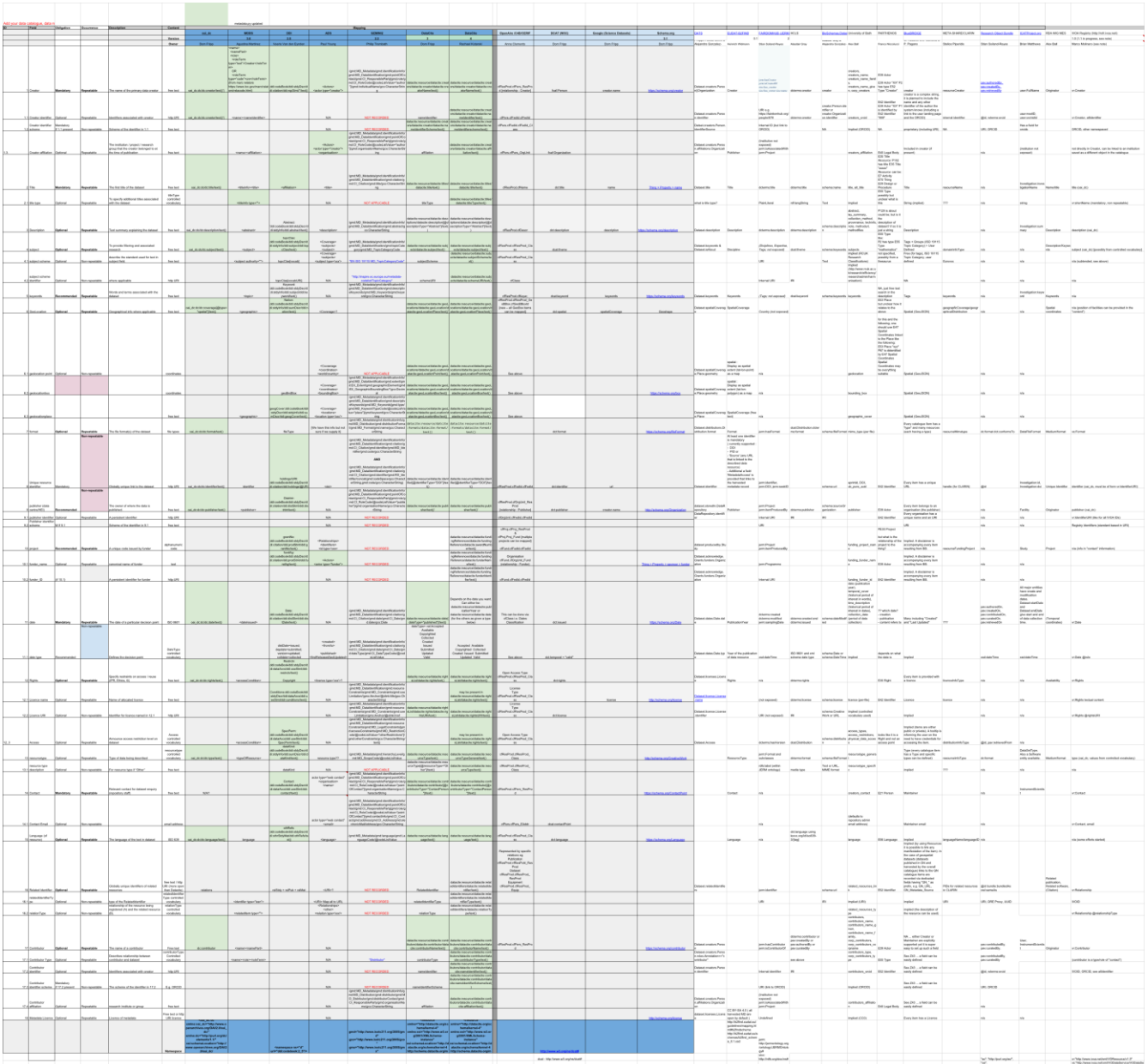
Index (CDI), provides users with discovery and access to the vast resources (metadata and access to more than 1.98 Million data sets, originating from more than 600 organisations in Europe) of marine and ocean datasets, managed by the distributed data centres (covering physical, geological, chemical, biological and geophysical data, and acquired in European waters and global oceans). Several data centers, in particular those concerning data, are a profile of the ISO19115 - ISO19139 standard for geographical information.

ANNEX J - SURVEY ANALYSIS: MATRIX COMPARING METADATA CATALOGUES



#	Catalogue	domain		types		location of data		access		catalogue format	data repositories format
	name	topic	specific	datasets	other (pub, soft, ...)	third party	catalogue	GUI	programatic	standard	common standard
1	EPOS	earth science	yes	yes	yes	yes	no	yes	no	yes	no
2	BlueBRIDGE	biodiversity environmental data	yes	yes	yes	yes	no	yes	yes	yes	no
3	ARIADNE	archeological research	yes	yes	?	yes	no	yes	no	yes	no
4	OpenAIRE	research publications	no	yes	yes	yes	no	yes	yes	yes	no
5	OpenMinTeD	text and data minig	yes	yes	yes	yes	yes	yes	no	yes	no
6	FAIRDOMHub	system biology	yes	yes	yes	yes	?	yes	yes	yes	yes
7	FAIRsharing	databases, policies, standards	no	no	yes	yes	no	yes	?	yes	no
8	CLARIN	language resources and language studi	yes	yes	?	yes	no	yes	yes	yes	?
9	SeaDataNet	marine	yes	yes	?	yes	no	yes	yes	yes	?
10	OmicsDI	omics data (genomics, proteomics, met	yes	yes	no	yes	no	yes	yes	yes	no
11	CMS Collaboration	High Energy Physics (HEP) events	yes	yes	?	no	?	yes	yes	no	no
12	[(A)VO] Registry of Resources	Astronomical/astrophysical	yes	yes	yes	yes	no	yes	yes	yes	?
13	EUDAT-B2FIND	Interdisciplinary research	no	yes	yes	yes	no	yes	yes	yes	?
14	Bio.tools	bioinformatics	yes	no	yes	yes	no	yes	yes	yes	no
15	DataMed	biomedical science	yes	yes	no	no	yes	yes	yes	yes	no
16	ICAT STFC	High Energy Physics (HEP) events	yes	yes	?	no	no	yes	yes	yes	?
17	Research Data Archive (UBath)	Interdisciplinary research	no	yes	?	yes	yes	yes	?	no	?
18	JISC Research Data Discovery Service	Interdisciplinary research	no	yes	?	yes	yes	yes	yes	yes	?

ANNEX E - DATASET METADATA PROPERTIES MAPPING



ANNEX F - EDMI METADATA PROPERTIES, USE CASES AND MAPPINGS

Functional metadata properties: use cases

Topic	Use Case	Minimum Level				
Entity	As a researcher I want to find datasets in a specific format.		R: distribution.fileFormat			
Entity	As a researcher I want to search on a title of a dataset.		M: name			
Entity	As a researcher I want to discover datasets based on free text search.		O: description	R: keywords		
Provenance	As a researcher I want to discover datasets based on their method of generation.		R: measurementTechnique			
Citation	As a researcher I want to enable proper credits/attribution.		O: citation	M: creator	M: name	M: dateCreated
Access	As a researcher I want to find a specific dataset regardless of its location.		M: identifier			
Entity	As a researcher I want to find datasets containing specific dimensions/variables.		R: variablesMeasured	R: keywords		
Access	As a researcher I want to find datasets included in a specific repository or catalogue.		R: includedInDataCatalog			
Context	As a researcher I want to discover related datasets (e.g. for a series of data takings).		O: sameAs	O: metrics		
Provenance	As a researcher I want to find raw data a given dataset was derived from.					
Metric	As a researcher I want to find datasets that are commonly used in my field.		O: metrics			
Provenance	As a researcher I want to find information on how a dataset was derived from raw data.		R: measurementTechnique	?		
Entity	As a researcher I want to find datasets based on specific values for a dimension/variable.		O: spatialCoverage	O: temporalCoverage		
Entity	As a researcher I want to find variables filtered by their units.					
Entity	As a researcher I want to find datasets that provide a proper license.		R: license			
Entity	As a researcher I want to find data across disciplines based on a location.		O: spatialCoverage			
Entity	As a researcher I want to find exemplary data to inspire my research.		M: name	M: description	R: variablesMeasured	
Entity	As a researcher I want to find datasets based on their quality of metadata.		O: metrics			

Functional metadata properties: mappings

Property Name	Requirement	Schema.org Datasets	Google Datasets	DataCite	Project Open Data	Description
name	Minimum	name	name	Title	title	The name of the da
description	Minimum	description	description	Description	description	A description of the
url	Recommended	url	url	valueURI	landingPage	Location of a page
identifier	Minimum	identifier		Identifier	identifier	The property repres
sameAs	Optional	sameAs	sameAs	RelatedIdentifier	references	Other URLs that ca
version	Optional	version	version	Version		The version of the c
keywords	Recommended	keywords	keywords		keyword	Keywords or tags u
variablesMeasured	Recommended	variablesMeasured (pending)	variableMeasured			The property indica
creator	Minimum	creator	creator.name	Creator	contactPoint	The creator/author
includedInDataCatalog	Recommended	includedInDataCatalog	includedInDataCatalog	Publisher	publisher	A data catalog whic
distribution	Recommended	distribution	distribution		distribution	A downloadable for
distribution.fileFormat	Recommended	distribution.fileFormat	distribution.fileFormat	Format	distribution.mediaType	Media type of the c
distribution.contentUrl	Minimum (derived from distribution)	distribution.contentUrl	distribution.contentUrl		distribution.accessURL	URL of the dataset
citation	Optional	citation	citation			A citation or referer
license	Recommended	license	license	Rights	license	A license documen
measurementTechnique	Recommended	measurementTechnique (pending)				A technique or tech
dateCreated	Minimum	dateCreated		Date		The date on which
spatialCoverage	Optional	contentLocation / spatialCoverage		GeoLocation	spatial	The location depict
temporalCoverage	Optional	temporalCoverage		Date	temporal	The property indica
metrics	Optional					

Operational metadata properties: use cases

Use Case	mum Level					
As a service consumer I cannot use the data unless I am allowed by a license	M	License (break it down)				
As a services I need to know not just the url of the end point but as well the format, metadata model, protocol, and interface	M	End point (protocols, interfaces, format, ...)	Format (XML, JSON, ...)	Metadata model (SBML, TDD, XS	Protocol (FTP, HTTP, ...)	Interface (Webserv
As a service tryinng to syncronise with the data. Want to keep my data updated	R	Update frequency	Version	Last update time /timestamp		
As a service I would like to know the type of data I need access to. For search purposes	R	DataType (Functional)	Kind of data (Genomic, Geospatial, etc)			
	M	DataFormat (Operational)	FileFormat (ZIP, VCF, KML, etc)			
As a service manager I would like to know how to contact in case of legal issues	R	LegalContact	Email, webpage, etc			
As a service I would like to know who to contact in case of technical issues	R	TechnicalContact	Email, webpage, etc			
As a service I would like to know which external dataset is integrated in the data set. Is a dataset composed of other datasets, does it have recor	R	Structure of the dataset				
As a service I need to plan the data transfer (time) or storage (space) so knowing the sze of the dataset is important	R	Size				
As a service I would like to know the contextual information (situational information)	R	Contextual infomration, origine of data, related da	Meta-data again			
As a service I would like to know the different locations to access replicas of the dataset	R	DataDistribution	Replica / Mirror of C24 (IMHO) (Endpoints - internal & external)			
As a service I need the know if my dataset or data is part of other dataset or collections	R	context of the dataset		Structure of dataset		
As a service I would need to know whether specific security or privacy levels apply	M	Access rights, legal limitations	security/privacy model required	AAI		
As a service I need to know how to cite an ingested dataset	O	Citation	(DataCite, DOI, etc)			
As a service I need to know what the level of underlying evidence levels or quality ratings are	R	In vitro/in vivo star rating	metrics			
As a service I need to know how the data is compressed so I know which tools to use to decompress it (unzip a zip, etc)	M	Compression				

Operational metadata properties: mappings

EOSCpilot example (values)	EOSCpilot comments	New prop.	Minimum / Recom	EOSCpilot properties	EOSCpilot properties (decomposed)	DATS	schema.org
https://www.ebi.ac.uk/pride/archive/projects/PXD000004			R	url	url	dataset.distributions(DatasetDistribution).ac	dataset.url
CC0 1.0			M	**license**	license.name	dataset.licenses.name	dataset.license(CreativeWork).name
https://creativecommons.org/publicdomain/zero/1.0/			M		license.url	dataset.licenses.url	dataset.license(CreativeWork).url
1			O	version	version	dataset.version	dataset.version
2012-12-10			M	**dateModified**	dateModified	dataset.dates.date dataset.dates.date.type=	dataset.dateModified
2012-12-10			O	datePublished	datePublished	dataset.dates.date dataset.dates.date.type=	dataset.datePublished
"proteomics", "protein identification", "mass spectrometry"			R	scientificType	scientificType (Genomics, Earth Science,	dataset.type	dataset.category
Creator			R	contact	contact.type [legal, technical,...]	dataset.creator(Person).roles	dataset.creator.jobTitle
Matthew MacDonald			R		contact.name	dataset.creator(Person).fullName	dataset.creator.name
mlmacdonald@gmail.com			R		contact.email	dataset.creator(Person).email	dataset.creator.email
http://identifiers.org/px/PXD000004			O	referenceCitation	referenceCitation.url	-	dataset.citation.url
Identification and quantification of human postmortem frontal cortex proteome with a			O		referenceCitation.name	-	dataset.citation.name
Matthew MacDonald			O		referenceCitation.author	-	dataset.citation.author
mlmacdonald@gmail.com			O		referenceCitation.date	-	dataset.citation.dateCreated
quality		x	O	metric	metric.category	-	-
PRIDEQ			O		metric.name	-	-
high			O		metric.value	-	-
https://www.ebi.ac.uk/pride/archive/projects/PXD000004/assays/26881			R	includedIn(Dataset)	includedIn(Dataset/DataCatalogue).url	-	-
https://www.ebi.ac.uk/pride			R	includes(Dataset)	includes(Dataset).url	dataset.hasPart	-
Repository,Project,Assay,Entry	Repository,Project,Assay,E	x	M	**structure**	structure.levels	-	-
Project	Project		M		structure.level	-	-
application/xml	http://www.iana.org/assignments/media-ty		R	**format**	distribution.mimeFormat	dataset.distributions(DatasetDistribution).for	dataset.distribution(DataDownload).fileFormat
ftp://ftp.pride.ebi.ac.uk/pride/data/archive/2	accessURL		M	**accessUrl**	distribution.accessURL	dataset.distributions(DatasetDistribution).ac	dataset.distribution(DataDownload).contentUrl
FTP		x	M	**accessInterface**	distribution.accessInterface [HTML | API |	-	-
processed			R	contentType	distribution.contentType [raw | processed	dataset.distributions(DatasetDistribution).qu	-
https://fairsharing.org/bsg-s000693			M	**dataStandard**	distribution.dataStandards.id	dataset.distributions(DatasetDistribution).co	-
mzTab			M		distribution.dataStandards.name	dataset.distributions(DatasetDistribution).co	dataset.distribution(DataDownload).fileFormat
format			M		distribution.dataStandards.type	dataset.distributions(DatasetDistribution).co	-
18GB			R	size	distribution.size	dataset.distributions(DatasetDistribution).siz	dataset.distribution(DataDownload).contentSize
some			O	compression	distribution.compression [yes | no | some]	dataset.distributions(DatasetDistribution).qu	-
-	https://docs.google.com/document/d/15Q				~~distribution.accessTypes [download, remo~~	dataset.distributions(DatasetDistribution).ac	-
none	https://docs.google.com/document/d/15Q		O	authorizations	distribution.authorizations [none, clickLice	dataset.distributions(DatasetDistribution).ac	-
none	https://docs.google.com/document/d/15Q		R	**authentications**	distribution.authentications.type [none, sir	dataset.distributions(DatasetDistribution).ac	-
none			O		distribution.authentications.supportedIds (	-	-

ANNEX H - LIST OF MINIMUM, RECOMMENDED AND OPTIONAL METADATA PROPERTIES

List of EDMI metadata properties[72]. On the left column the name of the property, in the middle column the description of the properties and in the right columns the identification of functional and operational metadata and its classification into minimum, recommended and optional properties. M/F: Minimum functional metadata. M/O: Minimum operational metadata. R/F: Recommended functiona metadata. R/O: Recommended operational metadata. O/F: Optional functional metadata. O/O: Optional operational metadata.

Properties	Description	M/F	M/O	R/F	R/O	O/F	O/O
MINIMUM							
name	A descriptive name of the dataset	yes					
description	A short summary describing a dataset	yes					
identifier	The identifier property represents any kind of identifier for any kind of dataset	yes					
url	The location of a page describing the dataset	yes			yes		
creator	The creator/author of this dataset	yes			yes		
dateCreated	The date on which the dataset was created	yes					yes
license	A license under which the dataset is distributed		yes	yes			
dataStandard	The standard in which the content of the dataset is represented		yes	yes			
dateModified	The date on which the dataset was most recently modified		yes				
structure	The description of the structure of the dataset		yes				
accessUrl	The link to download the dataset		yes				
accessInterface	The type of interface to present the dataset		yes				
RECOMMENDED							
includedeIn	A dataset or data catalog which contains the dataset			yes	yes		
measurementTechnique	A technique or technology used in a dataset corresponding to the method used for measuring the corresponding variables			yes			
keywords	Keywords or tags used to describe the dataset			yes			
variablesMeasured	The variables that are measured in the dataset			yes			
format	The format in which the content of the dataset is encoded to present the information, typically a MIME format				yes		
scientificType	Scientific domain or type of the information provided in the datataset				yes		
includes	A dataset or data catalog contained in the dataset				yes		
contentType	Type of content provided in the dataset based on its origin and type of processes (raw, processed, summarised)				yes		

[72] https://tinyurl.com/dats-cats-edmi

size	Size of the dataset using a digital information multiple unit byte siymbol (MB, GB, PT, ...)				yes		
authentications	Type of authentication required to access the dataset				yes		
OPTIONAL							
version	The version of the dataset					yes	yes
metric	Metric to provide some quantitative or qualitative information about the dataset					yes	yes
sameAs	Other URLs that can be used to access the dataset page					yes	
spatialCoverage	The location depicted or described in the content					yes	
temporalCoverage	The property indicates the period that the content applies to					yes	
citation	A citation or reference to another work that describes the dataset					yes	
referenceCitation	A citation or reference to that describes the dataset						yes
compression	Type of compression used in the dadataset						yes
authorisations	Type of authorisation required to access the dataset						yes

ANNEX I - EXAMPLE OF HOW TO EXPOSE FUNCTIONAL AND OPERATIONAL METADATA

Example of how to expose EDMI operational metadata for a dataset from PRIDE[73] using the metadata properties from schema.org[74] and DATS[75] (Sansone et al. 2017).

```
{
        "@context":     ["http://schema.org",     "dats":     "https://github.com/biocaddie/WG3-
MetadataSpecifications/tree/v2.2", "epdi":"https://eoscpilot.eu"],
        "@type": "Dataset",
        "name": "Identification and quantification of human postmortem frontal cortex proteome
with a SILAM mouse brain standard",
        "sameAs": "http://identifiers.org/px/PXD000004"
        "url": "https://www.ebi.ac.uk/pride/archive/projects/PXD000004"
        "license": [
                {
                        "@type": "CreativeWork",
                        "name": "EMBL-EBI terms of use",
                        "url": "https://www.ebi.ac.uk/about/terms-of-use"
                },
                {
                        "@type": "CreativeWork",
                        "name": "CC0 1.0",
                        "url": "https://creativecommons.org/publicdomain/zero/1.0/"
                },
        ],
        "version":"1.0",
        "datePublished": "2012-12-10",
        "dateModified": "2012-12-10",
        "category": ["proteomics", "protein identification", "mass spectrometry"],
        "creator": [{
                "@type": "Person",
                "name": "Matthew MacDonald",
                "email": "mlmacdonald@gmail.com"
        }],
        "citation": {
                "@type": "CreativeWork",
                "name": "Identification and quantification of human postmortem frontal cortex
proteome with a SILAM mouse brain standard",
                "url": "http://identifiers.org/px/PXD000004",
                "sameAs":
                "author": {
                        "@type": "Person",
                        "name": "Matthew MacDonald",
                        "email": "mlmacdonald@gmail.com"
                },
                "publisher": {
                        "@type": "Organization",
                        "name": "PRIDE"
                }
        },
        "epdi:metric": {
                "@type": "Metric",
                "category": "quality",
                "name": "PRIDEQ",
                "value": "high"
        },
        "epdi:structure": {
                "levels": "Repository,Project,Assay,Entry"
                "level":"Project"
        },
        "epdi:includedIn":{
                @dataCatalogue,
                "url": "https://www.ebi.ac.uk/pride"
        }
```

[73] https://www.ebi.ac.uk/pride

[74] http://schema.org/

[75] https://github.com/biocaddie/DATS

```
    "epdi:includes":[
    {
            @dataset,
            "url": "https://www.ebi.ac.uk/pride/archive/projects/PXD000004/assays/26881"
    },
    {
            @dataset,
            "url": "https://www.ebi.ac.uk/pride/archive/projects/PXD000004/assays/26882"
    },
    {
            @dataset,
            "url": "https://www.ebi.ac.uk/pride/archive/projects/PXD000004/assays/26883"
    },
    {
            @dataset,
            "url": "https://www.ebi.ac.uk/pride/archive/projects/PXD000004/assays/26884"
    },
    {
            @dataset,
            "url": "https://www.ebi.ac.uk/pride/archive/projects/PXD000004/assays/26885"
    },
    ],
    "distribution": [{
            "@type": "DataDownload",
            "fileFormat": [
                    "application/xml",
                    "raw",
                    "application/gzip"
            ],
            "contentUrl": "ftp://ftp.pride.ebi.ac.uk/pride/data/archive/2012/12/PXD000004",
            "epdi:accessInterface": "FTP",
            "dats:conformsTo": [
             {
               "@type": "DataStandard",
               "identifier": "https://fairsharing.org/bsg-s000693",
               "name": "mzTab",
               "type": { "value": "format"}
             },
             {
               "@type": "DataStandard",
               "identifier": "https://www.biosharing.org/bsg-s000112",
               "name": "mzML",
               "type": { "value": "format"}
             }
    ],
            "epdi:contentType": [
                    "raw",
                    "processed",
                    "summarised"
            ],
            "contentSize": "18GB",
            "epdi:compression": "some",
            "authorizations":"none",
            "authentications":"none"
},
{
    "@type": "DataDownload",
            "fileFormat": [
                    "application/xml",
                    "raw",
                    "application/gzip"
            ],
            "contentUrl": "https://www.ebi.ac.uk/pride/archive/projects/PXD000004/files",
            "epdi:accessInterface": "HTML",
            "dats:conformsTo": [
             {
               "@type": "DataStandard",
               "identifier": "https://fairsharing.org/bsg-s000693",
               "name": "mzTab",
               "type": { "value": "format"}
             },
             {
               "@type": "DataStandard",
               "identifier": "https://www.biosharing.org/bsg-s000112",
               "name": "mzML",
```

```
                  "type": { "value": "format"}
                }
        ],
                "epdi:contentType": [
                        "raw",
                        "processed",
                        "summarised"
                ],
                "contentSize": "18GB",
                "epdi:compression": "some",
                "authorizations":"none",
                "authentications":"none"
    },
    {
        "@type": "DataDownload",
                "fileFormat": [
                        "application/xml",
                        "raw",
                        "application/gzip"
                ],
                "epdi:accessInterface": "Aspera",
                "dats:conformsTo": [
                {
                  "@type": "DataStandard",
                  "identifier": "https://fairsharing.org/bsg-s000693",
                  "name": "mzTab",
                  "type": { "value": "format"}
                },
                {
                  "@type": "DataStandard",
                  "identifier": "https://www.biosharing.org/bsg-s000112",
                  "name": "mzML",
                  "type": { "value": "format"}
                }
        ],
                "epdi:contentType": [
                        "raw",
                        "processed",
                        "summarised"
                ],
                "contentSize": "18GB",
                "epdi:compression": "some",
                "dats:authorizations":"none",
                "dats:authentications":"none"
    }]
}
```

ANNEX K - PROPOSED TIMELINE, PLAN AND SPECIFIC TASKS FOR THE EOSCPILOT DATA INTEROPERABILITY TASK

		Subtasks				Stakeholders											Deliverables			Events				Effort		Timeline	
ID	Tasks	F	A	I	R	ELIXIR-	INAF	JISC	UFlorer	KIT	ELIXIR-	BGS/NE	Athena	CNR	ICOS-E		M12	M18	M24	E3	E4	E5	E6	49 PMs	Status	Start	End
T1	**EOSC data review**										LEAD													5			
T1.1	Metadata catalogues and data repositories		B,C,D	A,D	C	R	C				R	C	C	A	C		x			x	x				Complete	2017-04-03	2017-09-08
T1.2	Scientific datasets		B,C,D	D	C	A	C				R	C	C	C	C		x				x				Complete	2017-06-06	2017-09-08
T1.3	e-infrastructure service requirements		B,C,D		C	C	C	A			C		R	C			x				x				Complete	2017-06-06	2017-09-08
T2	**Define findability use cases**						LEAD	LEAD																5			
T2.1	Technical use cases	A					A				C		R				x				x				Complete	2017-09-06	2017-09-08
T2.2	Scientific use cases	A				C	A			R		C	C	C	C		x				x				Complete	2017-09-06	2017-09-08
T3	**Map metadata and use cases**					LEAD																		10			
T3.1	Findability use cases (T2)	A,C				A	R	R		C	C		C				x					x			Complete	2017-09-08	2017-10-04
T3.2	Metadata catalogues	C				A	C				C			R			x					x			Complete	2017-09-08	2017-10-04
T3.3	Metadata standards	C				C			A			R			R		x					x			Complete	2017-09-08	2017-10-04
T4	**Propose metadata guidelines**											LEAD			LEAD									10			
T4.1	Metadata model(s)	D		A			C	C		C		A			R		x					x			Complete	2017-10-04	2017-11-03
T4.2	Minimum information	D		A		C	R	C		A		C	C	C	C		x					x			Complete	2017-10-04	2017-11-03
T4.3	Controlled vocabularies	D		A		C	R	C		A		C	C	C	C			x				x	x		Active	2017-10-04	2018-04-26
T4.4	Alignment with best practices			A,B			C	C	R	C		R			A			x				x	x		Active	2017-10-04	2018-04-26
T5	**Evaluate and prototype solutions**								LEAD					LEAD										10			
T5.1	Methods/technologies to expose metadata		D	C		C		C	R		C	A	C	C	R		x					x			Complete	2017-10-04	2017-12-01
T5.2	Architecture and interface to find datasets in EOSC	B		C				A	C		C		R					x					x		Active	2017-12-31	2018-04-26
T5.3	Quality of datasets				A	C			C			C	C	A	C				x				x				2018-04-25	2018-12-01
T5.4	Terms of use				B	C			C			C	C	A	C				x				x				2018-04-25	2018-12-01
T6	**Demo proof of concept**												LEAD											9			
T6.1	Metadata providers	B				C					A	R	R	R	R			x					x			2018-04-25	2018-06-01
T6.2	Metadata EOSC index	B				C		R			C	C	A	C	C			x					x			2018-04-25	2018-06-01
	Deliverables																										
M12	Report - discovery architechtures for finding																x			x	x	x			Complete	2017-12-01	2017-12-31
M18	Report - science demostrators																	x					x			2018-06-01	2018-06-30
M24	Report																		x							2018-12-01	2018-12-31
	Events																										
E1	EOSCpilot KickOff																								Complete	2017-01-17	2017-01-19
E2	WP6 kick off																								Complete	2017-02-20	2017-02-21
E3	BlueBridge data catalogues																			x					Complete	2017-04-03	2017-04-03
E4	Open Science FAIR																				x				Complete	2017-09-06	2017-09-08
E5	EOSCpilot data interoperability technical workshop																					x			Complete	2017-10-04	2017-10-05
E6	EOSCpilot data interoperability workshop																						x			2018-04-25	2018-04-26