EOSCpilot
The European Open Science Cloud for Research Pilot Project

D6.9: Final report on Data 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), Ricardo Arcila (ELIXIR - EMBL), Rafael Jimenez (ELIXIR - EMBL), Volker Beckmann (CNRS), Giorgos Papanikos (Athena), Leyla J Garcia (ELIXIR-EMBL)
Status	Submitted
Version	v1.3.1
Date	20/05/2019

Dissemination Level

☒ 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)

Abstract:
The objective of the EOSCpilot data interoperability task (6.2) is to demonstrate how to ensure availability of scientific data to services and users through an open cloud infrastructure. To do so, this task has produced a set of recommendations driving a coherent strategy, as well as a set of technical solutions to help users and programmatic services to find and access datasets across several scientific disciplines, enabling the EOSC to be built around a more coordinated and aligned data ecosystem.

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 -WP6-D6.9	
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), Ricardo Arcila (ELIXIR - EMBL), Rafael C Jimenez (ELIXIR - EMBL), Volker Beckmann (CNRS), Giorgos Papanikos (Athena), Leyla J Garcia (ELIXIR-EMBL)
Contributor(s)	Yin Chen (EGI), Brian Matthews (STFC) Nuno Ferreira (SURFsara), Juan A Vizcaino (EMBL-EBI), Henning Hermjakob (EMBL-EBI), Heinrich Widmann (DKRZ), Vicky Schneider (Amazon), Susanna A Sansone (University of Oxford), Peter McQuilton (University of Oxford), Sarala Wimalaratne (EMBL-EBI), Cristina Duma (IFN), Valentino Cavalli (LIBER)
Due date	28/02/2019
Actual submission date	08/03/2019
Reviewed by	All authors and contributors
Approved by	Brian Matthews (STFC)
Start date of Project	01/01/2017
Duration	28 months

Versioning and contribution history

Version	**Date**	**Authors**	**Notes**
0.1	02/01/2019	Rafael C Jimenez (ELIXIR - EMBL)	First draft taking into account the results of the data interoperability demonstrators and the feedback from the 2nd EOSC stakeholder forum
0.2	01/02/2019	Rafael C Jimenez (ELIXIR - EMBL)	Added modifications to the manuscript taking into account the feedback from the EOSC EAB (External Advisory Board)

1.0	12/02/2019	Rafael C Jimenez (ELIXIR - EMBL) and Nick Juty (ELIXIR - UMAN)	Draft ready to be shared with partners and reviewers
1.1	25/02/2019	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), Ricardo Arcila (ELIXIR - EMBL), Rafael C Jimenez (ELIXIR - EMBL), Volker Beckmann (CNRS), Giorgos Papanikos (Athena), Yin Chen (EGI), Brian Matthews (STFC) Nuno Ferreira (SURFsara), Juan A Vizcaino (EMBL-EBI), Henning Hermjakob (EMBL-EBI), Heinrich Widmann (DKRZ), Vicky Schneider (Amazon), Susanna A Sansone (University of Oxford), Peter McQuilton (University of Oxford), Sarala Wimalaratne (EMBL-EBI), Cristina Duma (INFN), Valentino Cavalli (LIBER), Leyla J Garcia (ELIXIR-EMBL)	Second draft including feedback from the partners, collaborators and reviewers.
1.2	28/02/2019	Rafael C Jimenez (ELIXIR - EMBL) and Nick Juty (ELIXIR - UMAN)	Final version addressing suggestions from reviewers and partners.
1.3	08/03/2019	Brian Matthews (STFC)	Formatting and editorial changes.
1.3.1	20/05/2019	Mark Thorley (UKRI)	Minor typographic edits.

Disclaimer*:* The content of the document herein is the sole responsibility of the publishers and it does not necessarily represent the views expressed by the European Commission or its services.

While the information contained in the document is believed to be accurate, the author(s) or any other participant in the EOSCpilot Consortium make no warranty of any kind with regard to this material including, but not limited to the implied warranties of merchantability and fitness for a particular purpose.

Neither the EOSCpilot Consortium nor any of its members, their officers, employees or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein.

Without derogating from the generality of the foregoing neither the EOSCpilot Consortium nor any of its members, their officers, employees or agents shall be liable for any direct or indirect or consequential loss or damage caused by or arising from any information advice or inaccuracy or omission herein.

TABLE OF CONTENT

EXECUTIVE SUMMARY

The objective of the EOSCpilot data interoperability task (6.2) is to demonstrate how to ensure availability of research data to services and users through an open cloud infrastructure. To do so, this task has produced a set of recommendations driving a coherent strategy, as well as a set of technical solutions to help users and programmatic services to find and access datasets across several disciplines, enabling the EOSC to be built around a more coordinated and aligned data ecosystem.

The six recommendations driving our proposed strategy are:

- The data guidelines and technical solutions proposed by the EOSC should be specific, common, simple, lightweight and collaborative.
- All the data resources contributing to the EOSC should expose structured metadata.
- The EOSC should reuse or build upon existing standards and formats, and promote the use of common best practices across scientific domains.
- The EOSC should propose minimum information guidelines across scientific domains especially targeting key operational metadata which is important for services consuming data.
- The EOSC should support an interconnected ecosystem of metadata catalogues as a fundamental service component to facilitate data discovery.
- The EOSC should implement a monitoring service to validate standards and recommendations proposed by the EOSC.

The four technical solutions that comprise the data strategy are:

- A common minimum information metadata guideline for datasets in the EOSC.
- Recommendations to establish a coordinated and interconnected ecosystem of dataset metadata catalogues.
- Adoption of ResearchSchemas as a means to drive research data discoverability and accessibility.
- A set of recommendations for common properties across multiple data types.

1 INTRODUCTION

The FAIR Data Principles[1] are a set of high-level guidelines aimed to make data findable, accessible, interoperable and reusable. These principles provide guidance for research data management and stewardship. The adoption of the principles is advocated to data producers and data publishers, to promote data sharing and maximise the use (and reuse) of research data. However, the interpretation and implementation of the FAIR principles varies across different domains and by data resources. The objective of this task is to complement the FAIR principles by providing a strategy and a set of recommendations for the EOSC to improve the availability of research data to users and services through an open cloud infrastructure.

The work reported in this document has been driven by the feedback provided in EOSCpilot workshops and surveys, the outcomes of the EOSCpilot data interoperability demonstrators and specific recommendations proposed by partners and external stakeholders. More details about these activities are provided in the 1st and 2nd report of the EOSCpilot data interoperability deliverables[2].

The strategy proposed in this document is driven by a set of recommendations that can be applied to prioritise and define specific solutions needed to build a coordinated FAIR data ecosystem. The recommendations and the strategy take into account three major components (Data resources, Standards and metadata catalogues) and two types of consumers (Users and Services). During this project, we have proposed and piloted four solutions aiming to contribute to the strategy.

The EOSCpilot task 6.2, following its own recommendations, has produced a strategy prioritising solutions and guidelines contributing to the findability and accessibility of datasets across several scientific disciplines, targeting users as well as services. This task was planned in 3 phases aligned with the three deliverables. During the first phase, we developed a draft strategy based on the feedback collected from partners, and from a series of open community meetings. We also defined a set of principles to clarify and guide the scope of our work. These are available in the 1st Report on Data Interoperability[3]. During the second phase, we tested and evaluated the draft strategy. To do so we have proposed four internal demonstrators to test components of the strategy. The 2nd Report on data Interoperability provided a status update, and reviews aspects of the strategy based on the work done in these demonstrators. In the last phase (reported here, we propose a final EOSCpilot data interoperability strategy and a set of recommendations), taking into account the previous work and its context with other EOSC projects.

This document starts with an introductory chapter describing the data interoperability stakeholders involved in the data interoperability strategy. It continues describing general recommendations to propose guidelines and technical solutions to help build a cohesive FAIR data ecosystem in the EOSC. The final sections introduce a selection of technical solutions and a tailored strategy to improve the availability of research datasets to users and services in the EOSC.

[1] https://www.force11.org/group/fairgroup/fairprinciples

[2] https://eoscpilot.eu/themes/wp6-interoperability

[3] https://eoscpilot.eu/themes/wp6-interoperability/1st-report-on-data-report-findability-interoperability

2 DATA INTEROPERABILITY STAKEHOLDERS

The strategy and recommendations proposed in this document require an understanding of the different stakeholders involved, and of their relationships (Figure 1). We identify two major service providers (data resources and metadata catalogues) and two types of consumers (users and services consuming data). Data resources host data and information produced in research. Data resources are used by users as well as services. To facilitate the use of their hosted data, data resources often expose their data using access interfaces (e.g. web interface) and comply with data standards to increase data interoperability and integration. Metadata catalogues are specialised resources that index individual data resources, aggregating that information, and exposing it for consumption by users. Metadata catalogues play an important role as a broker to facilitate the discovery and accessibility of data by indexing metadata from third party data resources.

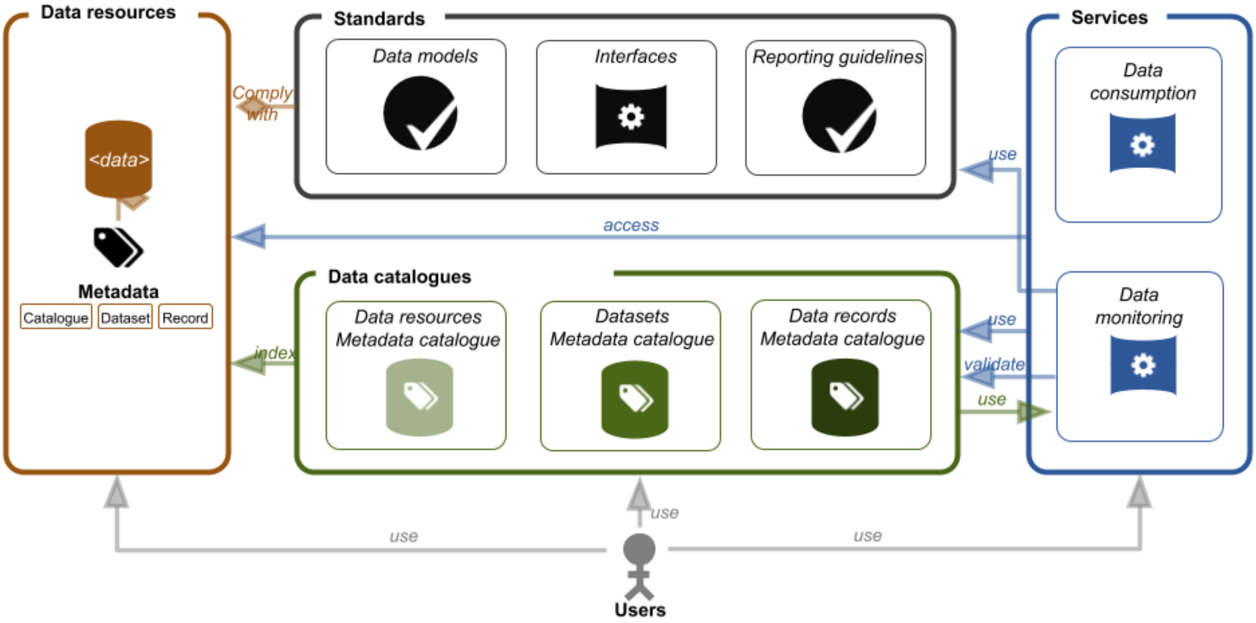

Figure 1. Stakeholders and relationships among stakeholders.

2.1 Data Resources

There are many types of research data resources distributed across the Internet. Each data resource tends to be structured and developed following its own user requirements. Though this is a good thing, it creates a great diversity in the way the data is modelled and exposed, presenting a challenge for users that want to use and integrate data from multiple such sources. The FAIR principles are helping data resource providers to think about how to improve their findability, accessibility, interoperability and reusability. The principles promote data sharing and quality, however they are not enough alone to build an integrated FAIR data ecosystem. The FAIR principles are high-level recommendations that can and have been interpreted in many different ways. They are aspirational, in that they do not strictly define how to achieve a state of "FAIRness", but rather they describe a continuum of features, attributes, and behaviours that will move a digital resource closer to that goal (Wilkinson et al. 2017). If the EOSC wants to build an integrated FAIR data ecosystem, it needs to agree on a concrete set of recommendations and technical solutions that complement the FAIR principles which are shared and adopted across all its participating data resources.

2.2 Data standards

Data standards are published documents that establish specifications and procedures designed to ensure that data is expressed or represented in a community prescribed and approved manner, thereby ensuring reliability of the data that researchers and services will use. Standards address a range of issues, including but not limited to various protocols to facilitate interoperability and product functionality and compatibility. Data standards make it easier to create, share, and integrate data by making sure that there is a clear understanding of how the data are represented.

We consider three types of data standard according to their focus: reporting guidelines (content), data models (syntax) and terminology artefacts (semantics). Reporting guidelines detail the information elements that need to be expressed in order to create a common core set of descriptors for different data types. For example, the MIAPTE[4] (Minimum Information About Particle Tracking Experiments) guidelines, which consist of a checklist and recommendations for authors for reporting from intracellular multiple particle tracking (MPT) experiments. Data models cover exchange formats that are used in data sharing. For example, DATS[5] and DCAT[6] are data models used to describe datasets. Terminology artefacts, also known as Controlled Vocabularies (CVs) or ontologies, are semantic representations of a topic or field used to catalogue and organise data with coded relationships between them. DOLCE[7], for example, is a terminology artefact used for Linguistic and Cognitive Engineering with related ontology terms like “process” and “state”. These types of data standards can be registered in for example FAIRsharing[8,9] and interlinked to the repositories that implement them.

We also consider standard interfaces, which play a major role in accessing and sharing data and metadata in a uniform manner. Data resources might provide access to their data via interfaces like GUIs, APIs, Web Services or FTP end-points. A minimum level of agreement and standardisation at the level of the access interfaces can make a difference helping consumers to query and integrate data from different data resources. For example, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[10] is used as a common programmatic interface by many data resources to facilitate computational access of metadata descriptions of data resources and their data records. The adoption of ready to use Data Management Systems (DMS) like Dataverse[11], the Open Science Framework[12] or SEEK[13] also provide a simple way to create a distributed and federated network of data resources exposing common programmatic interfaces.

[4] http://big.umassmed.edu/omegaweb/resources/miapte/

[5] https://github.com/datatagsuite

[6] https://www.w3.org/TR/vocab-dcat/

[7] http://www.loa.istc.cnr.it/old/DOLCE.html

[8] https://fairsharing.org

[9] https://doi.org/10.1101/245183

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

[11] https://dataverse.org/

[12] https://osf.io/

[13] https://seek4science.org/

2.3 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. Hence, metadata catalogues hold references to things while repositories hold the things.

Metadata catalogues can be classified by the type of metadata items they index. In the research domain, we can find a wide variety of metadata catalogues indexing different data types such as 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 metadata of data resources and datasets. 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 Identifiers.org[14], FAIRsharing (now also a RDA WG[15]), re3data[16], VizierR[17] and the Metadata Standards Directory (now also a RDA WG activity)[18].

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 its identifier. Examples of metadata catalogues of datasets are OmicsDI[19] (Perez-Riverol et al. 2017), DataMed (Ohno-Machado et al. 2017), OpenAIRE[20] and EUDAT-B2Find[21].

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[22] (Jarnuczak and Vizcaíno 2017), a domain specific data catalogue, indexes proteomics datasets and uses several specific controlled vocabularies to describe many metadata properties like experimental methods or proteomics data types.

[14] https://identifiers.org

[15] https://www.rd-alliance.org/group/fairsharing-registry-connecting-data-policies-standards-databases.html

[16] https://www.re3data.org

[17] http://vizier.u-strasbg.fr/viz-bin/VizieR

[18] https://rd-alliance.org/groups/metadata-standards-directory-working-group.html

[19] https://www.omicsdi.org

[20] https://www.openaire.eu

[21] http://b2find.eudat.eu

[22] http://www.proteomexchange.org

2.4 Consumers

When we talk about users of data we usually think about people. But in many cases the data is accessed, shaped or interpreted directly via services. Services may directly need to access the data, or might be used by people or by other services. Many data resource owners or maintainers are focussed on collecting user requirements from researchers and other people that might want to use the data, however, many of them do not consider service level requirements to be a priority, or may not have considered them at all. In the EOSC, services need to have the necessary information from data resources on how to access the hosted data. Therefore, it is important to take into account the different ways that data will be accessed, directly by users, as well as through services. One of the services that plays a key role in making sure the minimum requirements will be met to be part of the EOSC ecosystem will be a data monitoring service. Such a service will check data resources provided enough metadata for programmatic services to understand how to access the data. More information about how the EOSC should implement a monitoring service applicable to open science but also data is available in the report from EOSCpilot, D3.2: EOSC Open Science Monitor specifications[23]

[23] https://eoscpilot.eu/sites/default/files/eoscpilot-d3.2.pdf

3 RECOMMENDATIONS TO BUILD AN EOSC DATA ECOSYSTEM

The following set of six recommendations aim to prioritise and define specific solutions needed to build a coordinated FAIR data ecosystem. The first is a generic recommendation about how EOSC should develop guidelines and technical solutions. The other five recommendations focus on interfaces, data models, reporting guidelines, metadata catalogues and monitoring services.

R1. The data guidelines and technical solutions proposed by EOSC should be specific, common, simple, lightweight and collaborative.

Research data is generated across numerous and specialised research areas, thus it is important domain specific communities and specialised data resources collaborate to define their own standards and mechanisms to make data FAIR. EOSC should support the use of standards defined by communities but should also propose guidelines and technical solutions that it makes sense to have across all the research domains, helping build a more cohesive FAIR data ecosystem. The EOSC should focus and prioritise the guidelines and technical solutions that could bring more value to their participant stakeholders and to the ecosystem as a whole. The data guidelines and technical solutions proposed by the EOSC should be:

- Specific:– the EOSC should strive to provide specific guidelines and concrete solutions which let data service providers know how to implement them. Specific recommendation should accompany and complement agreed high-level recommendations like the FAIR principles.
- Common: the EOSC should work on guidelines and technical solutions that are common across different research disciplines. Guidelines and solutions should cover common and basic data types such as datasets and data repositories.
- Simple: the EOSC should deliver simple guidelines and technical solutions presenting a low barrier for adoption, implementation and maintainability.
- Lightweight: the EOSC should work on lightweight guidelines providing just enough functionality to facilitate FAIR data to be used in the EOSC. At the same time, the EOSC should support and promote more comprehensive guidelines agreed and used by specific domain communities.
- Collaborative: the EOSC should develop guidelines and technical solutions relying and complementing existing and well-adopted guidelines and technical solutions. Agreements should happen in collaboration with domain-specific initiatives and established data providers.

R2. All the data resources comprising the EOSC should expose structured metadata.

Though we highly recommend that data resources expose appropriate data through a programmatic interface, we acknowledge that not all of them have the resources or capacity to develop one. In this case, as an interim measure, we suggest that data resources at least expose structured metadata through a low barrier method such as schema.org markup[24].

It is difficult and counterproductive to recommend the adoption of just one programmatic interface for all the data resources in the EOSC. Many catalogues provide programmatic interfaces tailored to the needs of their community and we believe this is the right approach. Data resources should still make an effort to provide structured metadata with a complementary and more generic standard interface allowing a more

[24] https://schema.org/

federated environment. For instance, OmicsDI (Perez-Riverol et al. 2017) is a metadata catalogue that exposes a customised API, but also exposes the dataset metadata in a more standard way via schema.org.

Structured metadata should be exposed by data resources even if they hold sensitive data. The sensitive components of the data can be kept private but at the same time can be made findable by making part of the metadata available to users. Data should be “as open as possible, and as closed as necessary”[25] as recommended by the guidelines on FAIR Data Management in Horizon 2020. This controlled exposure of metadata will allow users to find relevant datasets and to later request access to the data.

R3. The EOSC should reuse existing standard formats promoting the use of common best practices across scientific domains.

Each scientific domain data resource should work in accordance with their own domain-specific standards and vocabularies, defining their specific entities that may also require the use of a standard format. The EOSC should avoid creating new data formats and encourage reusing existing ones. The EOSC recommendations should promote best practices across existing standards.

R4. The EOSC should propose minimum information guidelines across scientific domains especially encouraging exposing key operational metadata important for services consuming data.

Many standard formats are accompanied by reporting guidelines helping to define what information elements are important to be presented. More than one reporting guideline can exist for the same format, expressing different needs, while a single reporting guideline can be applied across several data formats. The EOSC should work on common reporting guidelines that could be applied across scientific domains reusing existing standard formats and interfaces. Common reporting guidelines focusing on a minimum set of key information elements are important to build a data ecosystem and promote common best practices.

Data resources, especially domain specific data resources, tend to excel in collecting user requirements from their main users, the researchers. However, many of them have not considered service requirements to be a priority, or had not considered them at all. In a data ecosystem, data is expected to be consumed not just by users but also by third party services, services that might want to integrate, analyse, copy or evaluate the data in a seamless and automatic way. This is especially important in a federated data ecosystem like the EOSC. Therefore, the EOSC should make sure their participant data providers also take into account requirements from third party services.

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. Though there might be considerable overlap among them, the EOSC should recognise scientific and operational metadata are equally important.

R5. The EOSC should support an interconnected system of metadata catalogues as a fundamental service component to facilitate data discovery.

[25] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

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 compute resources; and some others provide different coverage and granularity for a research domain e.g. science, life sciences or proteomics. We believe there should not be just one EOSC metadata catalogue but a collaborative and interconnected system of metadata catalogues supported by a sustainable and coordinated strategy to provide data consumers with a better service to find and access data.

R6. The EOSC should implement a monitoring service to validate standards and recommendations proposed by the EOSC.

A data monitoring service will play a key role on advising services, users, metadata catalogues and data resources about how well data comply with the standards and recommendations proposed by the EOSC. An EOSC monitoring system should be implemented to validate how well data standards and recommendations are being met. Metadata catalogues could use such monitoring system to evaluate different requirements integrating the validation information and making it accessible to users and services.

4 TECHNICAL SOLUTIONS

Following our recommendations we propose and prioritise four technical solutions:

- A common dataset minimum information metadata guidelines across research domains relying on existing standards to promote operational metadata required by services.
- Recommendations to establish a coordinated effort on dataset metadata catalogues to empower data providers and data consumers.
- ResearchSchemas as an initiative to promote research data discoverability and accessibility exposing structured markup via schema.org
- A set of recommendations for common properties across multiple data types to promote best practices across scientific domains contributing to the “RDA Metadata Interest Group” and “RDA Metadata Element Set” efforts.

These four solutions aim to be specific, common, simple, lightweight and collaborative. They are ‘Specific’ since they aim to facilitate dataset findability and discoverability focusing on operational and functional metadata to improve the availability of research data to users and services in EOSC; ‘Common’ since they are applicable to all research domains; ‘Simple’ since they are easy to adopt, implement and sustain; ‘Lightweight’ since they focus on minimum requirements and enough functionality across data resources; and ’Collaborative’ since they encourage reusing existing standards and approaches across domain specific disciplines.

4.1 A common dataset minimum information metadata guidelines

For services it is not easy to find, access, transfer and keep updated copies of data hosted by third party data resources. It is challenging since there are many data resources, often highly distributed, and employing different data models and a diversity of access interfaces. Operational metadata, provided at the level of a dataset, would help services to efficiently access data. We propose EDMI (EOSC Datasets Minimum Information), a simple metadata guideline to help users to find and access datasets. The EDMI metadata guidelines do not aim to be a new data model to describe datasets, but rather to complement existing data models. These guidelines define the minimum metadata properties that should be present *across* existing data models, and which should be exposed by EOSC data resources to facilitate 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. They can also be measured and thus could be easily used as one parameter to evaluate the ‘FAIRness’ of datasets and data resources
More information about EDMI, the metadata properties proposed and cross walk across several standards can be found in https://eosc-edmi.github.io

4.2 Recommendations to establish a coordinated effort of dataset metadata catalogues

A coordinated strategy to support a collaborative and interconnected system of dataset metadata catalogues should start by defining how the metadata is registered, exchanged and discovered.

a. **Register data sets at the most specific catalogue available**

Metadata catalogues should make an effort to define a common strategy to register metadata. This should include how to share metadata content and how to automate the ingestion of metadata from existing third party resources. Metadata tends to be richer in specialised metadata catalogues than in generic catalogues. Thus we believe the entry point of metadata registration should be the metadata catalogues which are more aligned to the research scope of the data produced. For instance, a proteomics dataset should be registered in a catalogue like ProteomeXchange[26] (catalogue of proteomics datasets) rather than in a generic catalogue like EUDAT-B2Find[27]. Generic catalogues should not encourage metadata registration in their own catalogues unless there is no domain specific catalogue where the dataset can be registered.

b. Generic catalogues should harvest domain catalogues.

Domain specific metadata catalogues should facilitate the indexing of their data into more generic metadata catalogues in a way it is easy to be automatically harvested (e.g. via programmatic interfaces). Generic catalogues should partner with domain specific catalogues so generic catalogues can easily import integrated and harmonised metadata from domain specific catalogues, rather than directly from the source. To recognise and sustain the important work of domain specific catalogues and facilitate the discovery of their richer metadata, generic metadata catalogues should acknowledge where the metadata came from and recommend metadata submission to domain specific catalogues.

c. Use catalogues of data resources

We recommend the use of a catalogue of data resources to facilitate the discovery of metadata catalogues and indexed data resources. A catalogue of data resources could easily suggest which catalogues are more suited for the registration of metadata and provide provenance relationships between data resources and metadata catalogues. A catalogue like FAIRsharing, which also contains information about reporting guidelines and standard formats, could also facilitate to the evaluation of compliance to EOSC guidelines, such as EDMI, across data resources and data catalogues.

d. Promote the federation of metadata catalogues

Agreements across metadata catalogues on a common strategy for metadata registration, exchange and discovery are key to building an EOSC data ecosystem. But the EOSC should not forget about promoting other activities that are important to integrate and bring together metadata catalogues. Activities such as:

- Develop common software components that could be easily reused to facilitate the sustainability, usability, and the technical implementation of the metadata catalogues. For example, tools and functionality like visualisation components, an authentication system, interfaces, crawlers, validators, formats, etc.
- Make the catalogue software components open-source in a publicly accessible, version controlled repository to facilitate the adoption of best practices and community engagement following the 4OSS guidelines (Jiménez et al. 2017), and make the registries software easy to reuse, configure and deploy.
- Promote knowledge exchange, community engagement and capacity building across metadata catalogues to better support sustainability of the metadata catalogues and their activities.

[26] http://www.proteomexchange.org/

[27] http://b2find.eudat.eu/

- Support users and service providers with the development of training materials and training workshops.
- Explore mechanisms to better sustain and support metadata catalogues.
- Adopt common guidelines, regulations and recommendations improving the quality and security of the catalogues. For instance identifying which guidelines (e.g. 4OSS), regulations (e.g. GDPR[28]) and technology recommendations (e.g. ResearchSchemas) metadata catalogues should comply with and how to adopt them in a collective and consistent manner.
- Agree on common Key Performance Indicator (KPIs) for the assessment of maturity and quality of metadata catalogues.

4.3 ResearchSchemas for exposing dataset metadata

30% of the data resources evaluated in EOSCpilot do not provide a programmatic interface that could help services to find and access data. Furthermore, they do not provide all the properties that are considered minimum in our EDMI recommendations. However all the evaluated resources do provide their information via a web interface. ResearchSchemas is proposed as a solution to provide a simple and quick way for data resources to expose structured metadata. ResearchSchemas complements and builds upon Schema.org (Mika 2015) to expose structured metadata for datasets and other generic research types using the existing web interfaces of data resources. ResearchSchemas 'types' could be leveraged by EOSC guidelines like EDMI, in this case to define the minimum dataset properties to be exposed in schema.org.
More information about ResearchSchemas can be found in the RDA Data Discovery Paradigms Interest Group https://www.rd-alliance.org/groups/data-discovery-paradigms-ig

4.4 A set of recommendations for common properties across multiple data types

The goal of this recommendation is to promote best practice across research domains for specific metadata properties, particularly for common metadata properties like "license" and "identifiers" recommended through EOSC guidelines such as EDMI. To achieve this goal EOSCpilot has extended and contributed to the work of the RDA Metadata Interest Group (MIG)[29] and the "RDA Metadata Element Set". We have proposed a new template providing more structure to the recommendations including generic as well as domain specific recommendations starting with the "identifiers" and "license" properties.
More information about the set of recommendations for common properties can be found in the RDA Data Metadata Interest Group "https://rd-alliance.org/groups/metadata-ig.html".

[28] General Data Protection Regulation

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

5 DATA INTEROPERABILITY STRATEGY

The overall strategy focuses on improving the availability of research data to users and services in the EOSC. The strategy involves data resources, metadata catalogues as well as users and services consuming data and metadata (Figure 2). Users and services that need data can search, find and discover data hosted by data resources via metadata catalogues. Metadata catalogues play an important role in integrating and harmonising metadata from dispersed and disparate data resources. In this is strategy we focus on metadata catalogues of datasets, since datasets cover a wide spectrum of data resources and are connected with the records they contain and the data resources where they are hosted.

A catalogue of data resources like FAIRsharing, which is also recommended by the EOSC report on Turning FAIR into reality[30], is key to finding dataset metadata catalogues and the data resources they index. It is also a first entry point to identify which metadata catalogues and data resources are compliant with EOSC guidelines like EDMI. Dataset metadata catalogues can do this validation at the level of the dataset. This validation can be automated and facilitated by a common monitoring service that could be used by all metadata catalogues.

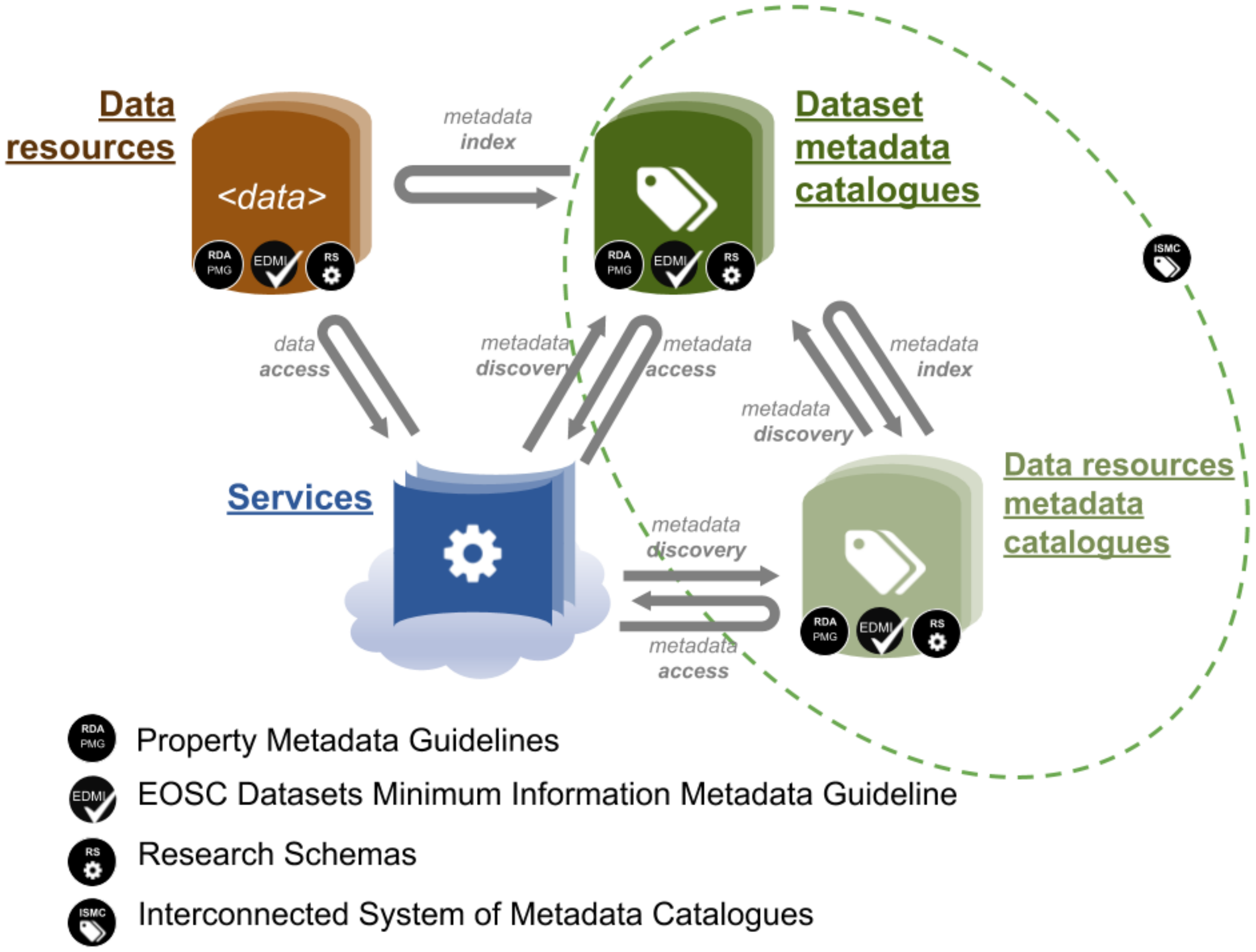
Data resources
<data>
metadata index
Dataset metadata catalogues
ISMC
data access
metadata discovery
metadata access
metadata index
metadata discovery
Services
Data resources metadata catalogues
metadata discovery
metadata access
RDA PMG
EDMI
RS
RDA PMG Property Metadata Guidelines
EDMI EOSC Datasets Minimum Information Metadata Guideline
RS Research Schemas
ISMC Interconnected System of Metadata Catalogues

Figure 2. Stakeholders and technical solution proposed in the data interoperability strategy.

[30] Commission Directorate General for Research and Innovation: Turning FAIR into reality European. https://doi.org/10.2777/1524 (2018).

Services that find relevant datasets through a metadata catalogue might want to take the next step and access the datasets. With metadata compliant to the minimum operational metadata as described by EDMI, there should be enough information provided to services to know how to access those datasets. Part of this information reveals information such as the location of the dataset, its format and the type of interface used to be able to access the dataset.

Many dataset metadata catalogues struggle to index dataset metadata from different data resources since each data resource tends to provide their data and metadata using different formats, vocabularies and interfaces. ResearchSchemas could help data resources to expose their structured metadata in a common, simple and complementary way making datasets and data records easy to be indexed by metadata catalogues and search engines. The RDA “Metadata Element Set” recommendations on metadata properties can also be applied to ResearchSchemas helping to harmonise metadata across different schema.org data types.

6 CONCLUSION AND NEXT STEPS

This strategy proposes common solutions compatible with existing domain specific solutions, thus having a low entry barrier for entry and adoption, whilst maximising benefits for data consumers and data providers. These solutions were proposed, tested and shaped during the EOSCpilot, however they need further development and support to be implemented and adopted. Some of the guidelines and technical solutions proposed by EOSCpilot have been picked up by other projects. ResearchSchemas is now an initiative driven by the RDA Data Discovery Paradigms and participated in by EOSCpilot members and a wider community. The recommendation for common properties across multiple data types are also part of RDA and are driven by the Metadata Interest Group. Some recommendations to establish a coordinated effort of dataset metadata catalogues are partially followed by EOSC projects like EOSCHub and EOSC-Life, however to develop and implement an interconnected system of metadata catalogues, the EOSC would need to bring together existing metadata catalogues from domain specific communities and e-infrastructures. EDMI has its own website and it is hosted in github so the community can provide feedback and develop EDMI further, but it needs to be owned and sustained by the EOSC. The strategy should not be limited to the technical solutions presented in this document, however it is important the EOSC makes a decision about what guidelines and technical solutions are important and make a decision about how to support them and implement them. Data guidelines and technical solutions proposed by the EOSC to build a data ecosystem should consider the proposed 6 recommendations, especially aiming to be specific, common, simple, lightweight and collaborative to facilitate their adoption and implementation.

REFERENCES

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.

Jiménez, Rafael C., Mateusz Kuzak, Monther Alhamdoosh, Michelle Barker, Bérénice Batut, Mikael Borg, Salvador Capella-Gutierrez, et al. 2017. “Four Simple Recommendations to Encourage Best Practices in Research Software.” *F1000Research* 6 (June). https://doi.org/10.12688/f1000research.11407.1.

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): 816. https://www.nature.com/articles/ng.3864

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): 406.

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.