European Commission

EOSC Interoperability Framework

Report from the EOSC Executive Board Working Groups (WG) FAIR and Architecture

Independent Expert Report

Research and Innovation

EOSC Executive Board
WGs FAIR and Architecture
February 2021

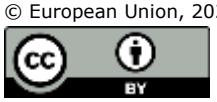
EOSC Interoperability Framework

European Commission
Directorate-General for Research and Innovation
Directorate G — Research and Innovation Outreach
Unit G.4 — Open Science
Contact Corina Pascu
Email Corina.PASCU@ec.europa.eu
RTD-EOSC@ec.europa.eu
RTD-PUBLICATIONS@ec.europa.eu
European Commission
B-1049 Brussels

Manuscript completed in January 2021.

PDF ISBN 978-92-76-28949-4 doi: 10.2777/620649 KI-02-21-055-EN-N

Luxembourg: Publications Office of the European Union, 2021



EUROPEAN COMMISSION

EOSC Interoperability Framework

Report from the EOSC Executive Board Working Groups FAIR and Architecture

Edited by: the EOSC Executive Board

February 2021

Authors

Oscar Corcho, Universidad Politécnica de Madrid, 0000-0002-9260-0753

Magnus Eriksson, Swedish Research Council, 0000-0003-1877-6168

Krzysztof Kurowski, Poznań Supercomputing and Networking Center IBCH PAS, 0000-0002-4478-6119

Milan Ojsteršek, University of Maribor, 0000-0003-1743-8300

Christine Choirat, Swiss Data Science Center, ETH Zürich and EPFL, 0000-0002-3745-9718

Mark van de Sanden, SURF, 0000-0002-2718-8918

Frederik Coppens, VIB-UGent Center for Plant Systems Biology, 0000-0001-6565-5145

With contributions from the EOSC FAIR WG chairs (Sarah Jones, Françoise Genova) and on legal interoperability from: Ohad Graber-Soudry, Timo Minssen, Daniel Nilsson, Marcelo Corrales, Jakob Wested, Bénédicte Illien

2021 Directorate-General for Research and Innovation

Contents

EXECUTIVE SUMMARY 3
1 INTRODUCTION 6
1.1 Context and definitions 6
1.1.1 The European Open Science Cloud (EOSC) 6
1.1.2 FAIR principles and the role of Interoperability 6
1.1.3 The European Interoperability Framework as a Starting Point 7
1.1.4 Definitions of relevant terms used in this document 7
1.2 Purpose and scope 9
1.3 How to read this document 9
2 INTEROPERABILITY LAYERS 11
2.1 Technical interoperability 11
2.2 Semantic interoperability 11
2.3 Organisational interoperability 12
2.4 Legal interoperability 12
3 MINIMUM REQUIREMENTS AND RECOMMENDATIONS FOR THE EOSC IF 14
3.1 Technical interoperability 14
3.1.1 Problems and needs 14
3.1.2 Recommendations 15
3.2 Semantic interoperability 16
3.2.1 Problems and needs 16
3.2.2 Recommendations 17
3.3 Organisational interoperability 18
3.3.1 Problems and needs 18
3.3.2 Recommendations 19
3.4 Legal interoperability 19
3.4.1 Problems and needs 19
3.4.2 Recommendations 23
3.5 Some general recommendations from the EIF 26
3.6 Summary of recommendations 27
4 TOWARDS THE EOSC IF: MODEL AND COMPONENTS 29
4.1 Model overview 29
4.2 Basic components 32
4.2.1 Generic and community-specific semantic artefacts 32
4.2.2 Generic metadata frameworks and data type registries 33
5 TOWARDS THE EOSC IF: REFERENCE ARCHITECTURE 35
5.1 EOSC-IF high-level Architecture viewpoint 36
5.1.1 Legal view 36
5.1.2 Organisational view 36
5.1.3 Semantic view 37
5.1.4 Technical view 38
5.2 EOSC-IF Reference Architecture – View details 39
5.2.1 EOSC-IF High-level Semantic view 39
5.2.2 EOSC-IF High-level technical view 42
5.3 Recommendations and next steps 45
APPENDIX I. ANALYSIS OF MINIMAL METADATA MODELS AND CROSSWALKS AMONG THEM 46
APPENDIX II. INTERVIEWS WITH STAKEHOLDERS 57

EXECUTIVE SUMMARY

This document has been developed by the Interoperability Task Force of the EOSC Executive Board FAIR Working Group, with participation from the Architecture WG.

Achieving interoperability within EOSC is essential in order for the federation of services that will compose EOSC to provide added value for service users. In the context of the FAIR principles[1], interoperability is discussed in relation to the fact that "research data usually need to be integrated with other data; in addition, the data need to interoperate with applications or workflows for analysis, storage, and processing". Our view on interoperability does not only consider data but also the many other research artefacts that may be used in the context of research activity, such as software code, scientific workflows, laboratory protocols, open hardware designs, etc. It also considers the need to make services and e-infrastructures as interoperable as possible.

This document identifies the general principles that should drive the creation of the EOSC Interoperability Framework (EOSC IF), and organises them into the four layers that are commonly considered in other interoperability frameworks (e.g., the European Interoperability Framework[2] - EIF): technical, semantic, organisational and legal interoperability.

For each of these layers, a catalogue of problems and needs, as well as challenges and high-level recommendations have been proposed, which should be considered in the further development and implementation of the EOSC IF components. Such requirements and recommendations have been developed after an extensive review of related literature as well as by running interviews with stakeholders from ERICs (European Research Infrastructure Consortia), ESFRI (European Strategy Forum on Research Infrastructures) projects, service providers and research communities. Some examples of such requirements are: "every semantic artefact[3] that is being maintained in EOSC must have sufficient associated documentation, with clear examples of usage and conceptual diagrams", or "Coarse-grained and fine-grained dataset (and other research object) search tools need to be made available", etc.

The document finally contains a proposal for the management of FAIR Digital Objects in the context of EOSC and a reference architecture for the EOSC Interoperability Framework that is inspired by and extends the European Interoperability Reference Architecture (EIRA), identifying the main building blocks required.

Two appendixes are provided for this document:

- Since semantic interoperability was highlighted as a challenging area in our interviews, we provide an analysis and a more detailed documentation, in appendix I, over the "Minimal Metadata" architectural building block in the reference architecture. An analysis of existing metadata models and an initial set of crosswalks among them are included. This initial work may set the initial steps for a future proposal for an EOSC Minimal Metadata Application profile, which should be widely discussed and agreed by a large palette of disciplinary communities.

- Appendix II provides the interview protocol followed with stakeholders.

1 Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016)

2 The New European Interoperability Framework. Available on https://ec.europa.eu/isa2/eif_en (last accessed: 31/Dec/2020)

3 Semantic artefact is defined in this document as a machine-actionable and -readable formalisation of a conceptualisation enabling sharing and reuse by humans and machines. These artefacts may have a broad range of formalisation, from loose set of terms, taxonomies, thesauri to higher-order logics.

The following table summarises the taskforce recommendations, organised by layers.

Layer	Recommendation
Technical	• Open Specifications for EOSC Services. • A common security and privacy framework (including Authorisation and Authentication Infrastructure). • Easy-to-understand Service-Level Agreements for all EOSC resource providers. • Easy access to data sources available in different formats. • Coarse-grained and fine-grained dataset (and other research object) search tools. • A clear EOSC PID policy.
Semantic	• Clear and precise, publicly-available definitions for all concepts, metadata and data schemas. • Semantic artefacts preferably with open licenses. • Associated documentation for semantic artefacts. • Repositories of semantic artefacts, rules with a clear governance framework. • A minimum metadata model (and crosswalks) to ease discovery over existing federated research data and metadata. • Extensibility options to allow for disciplinary metadata. • Clear protocols and building blocks for the federation/harvesting of semantic artefacts catalogues.
Organisational	• Interoperability-focused rules of participation recommendations. • Usage recommendations of standardised data formats and/or vocabularies, and with their corresponding metadata. • A clear management of permanent organisation names and functions.
Legal	• Standardised human and machine-readable licenses, with a centralised source of knowledge and support on copyright and licenses. • Permissive licenses for metadata (and preferably for data, whenever possible). And CC0 preferred over CC BY 4.0. • Identification of different parts of a dataset with different licenses. • Clearly marked instances of expired or inexistent copyright, as well as for orphan data. • A clear list of EOSC-recommended licenses and their compatibility with Member States' recommended licenses. • Tracking of license evolution over time for datasets. • Harmonised policy and guidance to dealing with cases where patent filing or trade secrets may be compromised by disclosure. • GDPR-compliance for personal data. • Additional restrictions on access and use of data only applied in cases of applicable legislation or legitimate reasons. • Harmonised terms of use across repositories • Alignment between Member States national legislations and EOSC.

The EOSC Interoperability task force further recommends continuing the work on the EOSC Interoperability Framework with:

- Detailed specification of Architectural building blocks, hand in hand with the communities, many of which have already their interoperability practices in place.

- Establishing governance structure and maintenance of the framework, to guide, organise and keep the work together.

Additionally, an accompanying document with more details on legal interoperability is available[4].

4 Graber-Soudry, Ohad, Minssen, Timo, Nilsson, Daniel, Corrales, Marcelo, Wested, Jakob, & Illien, Bénédicte. (2021, January 27). Legal Interoperability and the FAIR Data Principles (Version 1.0). Zenodo. http://doi.org/10.5281/zenodo.4471312

1 INTRODUCTION

1.1 Context and definitions

This section provides some context and general definitions related to this document.

1.1.1 The European Open Science Cloud (EOSC)

The European Open Science Cloud (EOSC)[5] is a European Commission initiative aiming at developing a federated infrastructure providing its users with services promoting Open Science practices.

EOSC aims to support three objectives: (1) to increase the value of scientific data assets by making them easily available to a larger number of researchers, across disciplines (interdisciplinarity) and borders (EU added value) and (2) to reduce the costs of scientific data management, while (3) ensuring adequate protection of information/personal data according to applicable EU rules.

1.1.2 FAIR principles and the role of Interoperability

In the context of the FAIR principles[6], interoperability is discussed in relation to the fact that "research data usually need to be integrated with other data [...] in addition, the data need to interoperate with applications or workflows for analysis, storage, and processing". The following principles are proposed:

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles.
- I3. (Meta)data include qualified references to other (meta)data.

As discussed in the "Turning FAIR into Reality" report[7], the role of interoperability frameworks is to "define community practices for data sharing, data formats, metadata standards, tools and infrastructure, recognising the objectives and cultures of different research communities". The report also stresses the fact that such frameworks need to support FAIR across traditional discipline boundaries and in the context of high priority interdisciplinary research areas.

Achieving interoperability within EOSC is essential in order for the federation of services that will compose EOSC to provide added value for service users, no matter which scientific disciplines they work on. The services within the EOSC will provide value by provisioning digital objects (which refer to the aforementioned research artefacts and whose definition is provided in Section 1.1.4). In order to realise the value of the services, the digital objects exchanged need to be efficiently consumed by other EOSC services and user systems.

In order for the user systems to consume the digital objects provisioned by the EOSC services they must understand how to read and interpret them, what restrictions there are to use the object and what processes are involved in their production, provisioning and

5 What the European Open Science Cloud is. https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud (last accessed: 31/Dec/2020)

6 Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016)

7 Turning FAIR into Reality. Final Report and Action Plan from the European Commission Expert Group on FAIR Data. 2018. https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf (last accessed: 31/Dec/2020)

consumption. All these processes should be independent from the specific scientific discipline where the digital objects were created or are being consumed.

Therefore, software/machines should be able to deduce or obtain these characteristics from the information provided by the digital object itself through its metadata. The EOSC interoperability framework aims to provide a set of recommendations on the components that need to be provided in the ecosystem and on the principles guiding digital object producers and/or consumers on their use, in order for the framework to set a foundation for an efficient machine-enabled exchange of digital objects within EOSC and between EOSC and the outside world. A final aspect to consider in this context is that there will be different degrees of interoperability that will be achievable, especially in interdisciplinary settings.

1.1.3 The European Interoperability Framework as a Starting Point

The structure of the EOSC IF is inspired by earlier work done for the European Interoperability Framework (EIF)[8], as well as in the context of other domain-specific interoperability frameworks (e.g., the Shift2Rail Interoperability Framework[9]).

The EIF, promoted and maintained by the ISA[2] programme, targets public administrations in Europe, so that they can design and deliver public services in an interoperable manner, contributing to the development of a single digital market by fostering cross-border and cross-sectoral interoperability for the delivery of such European public services.

The core targets of EIF are public administrations at all levels, including the national interoperability frameworks, and interactions between administrations - A2A -, administrations and citizens - A2C - and administrations and businesses - A2B -. They are thus somewhat different to the target of the EOSC IF, which is mostly focused on individual researchers, research performing organisations, research funding organisations and research infrastructures. However, they share many common underlying principles and core objectives. Indeed, using the EIF terminology, the EOSC IF may be seen as an example of a Domain-specific Interoperability Framework, which in turn focuses on multiple scientific domains.

For that reason, the EOSC IF is structured in a similar manner to EIF. More specifically, the EIF identifies four layers of interoperability (technical, semantic, organisational and legal), which have been also considered in the development of the EOSC IF.

1.1.4 Definitions of relevant terms used in this document

In this document, we use the term **Digital Object** to refer to the kind of objects that allow binding all critical information about any entity. The information that we are interested in in the context of the EOSC IF includes research data, software, scientific workflows, hardware designs, protocols, provenance logs, publications, presentations, etc., as well as all their metadata (for the complete object and for its constituents). The act of defining a Digital Object is the act of defining a boundary around a set of data points. From an interoperability point of view, not all actors will define the same boundaries in the same place (a simple example might be the choice to bundle data and metadata together vs handling data/metadata as two related objects). The RDA Data Foundation & Terminology (DFT) Core Terms and Model[10] states that "a Digital Object is represented by a bitstream,

8 New European interoperability framework. Promoting seamless services and data flows for European public administrations. Directorate-General for Informatics (European Commission). 2017. DOI: 10.2799/78681

9 Shift2Rail Interoperability Framework. Available on https://shift2rail.org/research-development/ip4/ (last accessed: 31/Dec/2020)

10 Gary Berg-Cross, Raphael Ritz, Peter Wittenburg (2016) RDA DFT Core Terms and Model. http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318

is referenced and identified by a persistent identifier and has properties that are described by metadata".

Examples of Digital Objects proposed in the past are Research Objects[11] and some of their implementations (e.g., RO-Crate[12], the BagIt specification[13]).

We also use the term **metadata** widely. For this, we have decided to choose the ISO11179 definition of metadata, which defines it as "descriptive data about an object". That is, metadata is a kind of data: data may act as metadata when the descriptive relationship is revealed between the data (now metadata) and the target object(s). And metadata that is the same for more than one object is metadata for a class of objects…" (ISO/IEC CD 11179-1)[14]. This definition also aligns well with the definition used in the paper on the FAIR Principles[15], which states that the term "data" is used to refer to all types of digital resources (not just data in the restricted sense, but also, for example, software, workflows, hardware designs, etc.) and metadata is any description of a resource that can serve the purpose of enabling findability and/or reusability and/or interpretation and/or assessment of that resource. In this context, data and metadata may be published together or as different inter-related entities (with their own identifiers), and different blocks of metadata may be associated to the same digital object (as described further in Section 4).

The term **semantic artefact** is used throughout the document, and more specifically in those sections describing semantic interoperability. There is no commonly agreed definition for semantic artefact, although a working definition is provided by Coen[16] as "the tools which allow humans and machines to locate, access and understand (meta)data, [..] including ontologies, knowledge organisation systems, data vocabularies, code lists, etc.". And FAIRsFAIR provides another definition: "a machine-actionable and -readable formalisation of a conceptualisation enabling sharing and reuse by humans and machines. These artefacts may have a broad range of formalisation, from loose set of terms, taxonomies, thesauri to higher-order logics. Moreover, semantic artefacts are serialised using a variety of digital representation formats, e.g., RDF Turtle, OWL-RDF, XML, JSON-LD"[17]. These definitions are in line with earlier representations of the continuum between lightweight and heavyweight ontologies from Lassila and McGuiness (2001)[18].

Finally, different definitions around **interoperability** are available in the state of the art. We summarise some of those that we are taking in the context of this document here:

- **Interoperability.** The European Interoperability Framework (EIF) defines interoperability as the "ability of organisations to interact towards mutually beneficial goals, involving the sharing of information and knowledge between these organisations, through the business processes they support, by means of the exchange of data between their ICT systems"[19].

11 Khalid Belhajjame, Jun Zhao, Daniel Garijo, Matthew Gamble, Kristina Hettne, Raul Palma, Eleni Mina, Oscar Corcho, José Manuel Gómez-Pérez, Sean Bechhofer, Graham Klyne, Carole Goble (2015) Using a suite of ontologies for preserving workflow-centric research objects, Web Semantics: Science, Services and Agents on the World Wide Web, https://doi.org/10.1016/j.websem.2015.01.003

12Research Object Crate (RO-Crate). Available on https://researchobject.github.io/ro-crate/ (last accessed: 31/Dec/2020)

13 The BagIt File Packaging Format (V1.0). Available on https://tools.ietf.org/html/draft-kunze-bagit-17 (last accessed: 31/Dec/2020)

14 ISO/IEC CD 11179-1. Available on https://www.iso.org/standard/78914.html (last accessed: 31/Dec/2020)

15 Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016)

16 Gerard Coen (2019) Introduction to Semantic Artefacts. FAIRsFAIR. http://doi.org/10.5281/zenodo.3549375

17 FAIRsFAIR - D2.2 FAIR Semantics: First recommendations. https://zenodo.org/record/3707985

18 Lassila O., & McGuiness D. The role of frame-based representation on the Semantic Web (Technical Report KSL-01-02. 2001). Knowledge Systems Laboratory, Stanford University.

19 EIF European interoperability framework - Introduction. Available on https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/1-introduction (last accessed: 31/Dec/2020)

- **Technical interoperability**. A characteristic of an Information Technology (IT) system, whose interfaces are completely understood, to work with other IT systems, at present or in the future, in either implementation or access, without any restrictions or with a controlled access (source: Interoperability - Wikipedia).

- **Syntactic interoperability.** If two or more systems use common data formats and communication protocols and are capable of communicating with each other using open standards (source: Interoperability - Wikipedia)

- **Semantic Interoperability.** It ensures that the precise format and meaning of exchanged data and information is preserved and understood throughout exchanges between parties, in other words 'what is sent is what is understood'.[20]

- **Organisational interoperability** refers to the way in which organisations align their business processes, responsibilities and expectations to achieve commonly agreed and mutually beneficial goals (source: European Interoperability Framework).

- **Legal interoperability**[21] is about ensuring that organisations operating under different legal frameworks, policies and strategies are able to work together. This might require that legislation does not block the establishment of European public services within and between Member States and that there are clear agreements about how to deal with differences in legislation across borders, including the option of putting in place new legislation.

1.2 Purpose and scope

The EOSC IF is meant to be a generic framework that can be used by all the entities participating in the development and deployment of EOSC, providing a common understanding of the requirements, challenges and recommendations that they should take into account, as well as a general set of principles on how these recommendations may be addressed. The EOSC IF does not propose any specific recommendation on how these recommendations should be actually implemented, although it provides a non-exhaustive list of illustrative examples of how some of them are being addressed in different contexts.

The different providers of EOSC-related services are also a relevant target for this document, since it provides some general recommendations for achieving interoperability across these services (e.g., interoperability in authentication and authorisation, interoperability in the exchange of data, interoperability for ensuring the findability of resources), enabling multidisciplinary and multi organisational collaborations.

1.3 How to read this document

This document is organised in three main sections:

- Section 2 provides a general overview of the four interoperability layers considered in the EOSC IF, and the types of challenges that are being addressed in each of them.

- Section 3 provides a summary of the main problems, needs, challenges, and recommendations at each layer, based on the analysis done on existing literature, plus the results of an extensive set of interviews run with researchers from different research communities, some of them involved in ESFRI projects and ERICs, as well as service providers.

20 Semantic interoperability. https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/glossary/term/semantic-interoperability (last accessed: 31/Dec/2020)

21 Legal interoperability definition. Available on https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/glossary/term/legal-interoperability (last accessed: 31/Dec/2020)

- Section 4 describes how interoperability may be addressed by adopting FAIR digital objects.

- Finally, Section 5 proposes a reference architecture for the EOSC IF that extends the European Interoperability Reference Architecture, identifying the main building blocks to be considered. To promote reuse and further development, the resources used in the context of this reference architecture are made openly available[22]. Since semantic interoperability was highlighted as a challenging area in our interviews, we provide an analysis, and a more detailed documentation, over the "Minimal Metadata" architectural building block. This analysis also provides suggestions on crosswalks for the "Mapping Repository" and corresponding materials related to the "Semantic Artefact" architectural building block.

Appendix I provides documentation over the analysis of existing metadata models and an initial set of crosswalks among them, which could lead in the future to a proposal for an EOSC Minimal Metadata Application profile, if considered appropriate. The corresponding materials are also made publicly available[23]. Appendix II contains further information related to the interviews that have been performed as a first step towards the creation of this document.

22 Eriksson, van de Sanden, Kurowski, Coppens, Corcho, & Ojsteršek. (2021, January 5). EOSC Interoperability Framework Reference Architecture (Version 1.0). Zenodo. http://doi.org/10.5281/zenodo.4420096

23 Ojsteršek. (2021). Crosswalk of most used metadata schemes and guidelines for metadata interoperability (Version 1.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.4420116

2 INTEROPERABILITY LAYERS

As already discussed in the introduction, the EOSC IF is structured according to the four interoperability layers already identified by the European Interoperability Framework: technical, semantic, organisational and legal. Each of these will be described in its own subsection below.

2.1 Technical interoperability

Technical interoperability is commonly defined as the "ability of different information technology systems and software applications to communicate and exchange data". This definition may be also completed by adding the "ability to accept data from each other and perform a given task in an appropriate and satisfactory manner without the need for extra operator intervention", that is, with an aspect focused on the complete automation of such data exchange.

In the context of this document, we refer not only to the exchange of data (across scientific experiments, organisations or even communities), but also of other research artefacts that are commonly used in research (software, services, workflows, protocols, hardware designs, etc.). According to the EIF, technical interoperability covers "the applications and infrastructures linking systems and services, including interface specifications, interconnection services, data integration services, data presentation and exchange, and secure communication protocols". Indeed, in some cases the technical interoperability layer may be split into two parts: infrastructure and applications. This is further described in Section 4.

In the context of our interviews the aspects related to technical interoperability have arisen in many occasions, not only across communities, but also in the context of a given scientific community, where for example different systems that are used for the generation of data or for its consumption are not compatible with each other, or where different user identification methods exist for researchers that need to make use of different types of systems. Best practices have also been identified in this context, as a result of our interviews. For example, in the context of astronomy, many efforts have been done along the years on the creation of the Virtual Observatory (http://www.ivoa.net/), not only as a technical platform for sharing and exchanging data, but also as a set of specifications and standards for the definition of data sources that can be used by researchers, with a clear governance model. All of these aspects will be addressed in section 3.1.

2.2 Semantic interoperability

Semantic interoperability can be defined as "the ability of computer systems to transmit data with unambiguous, shared meaning. Semantic interoperability is a requirement to enable machine computable logic, inferencing, knowledge discovery, and data federation between information systems".[24]

That is, semantic interoperability is achieved when the information transferred has, in its communicated form, all of the meaning required for the receiving system to interpret it correctly, even when the algorithms used by the receiving system are unknown to the sending system. Syntactic interoperability (which is commonly associated with technical interoperability) is sometimes identified as a prerequisite to semantic interoperability. It ensures that the precise format and meaning of exchanged data and information is preserved and understood throughout exchanges between parties, in other words "what is sent is what is understood".

Besides this machine-based view of semantic interoperability, this layer also requires humans being sufficiently aligned to have satisfactory communication. For instance,

24 FAIRsFAIR deliverable D2.1 Report on FAIR requirements for persistence and interoperability 2019. https://zenodo.org/record/3557381

Service-Level Agreements (SLAs) must in themselves be based on common semantic artefacts, which shows how intertwined the different levels of interoperability are.

In the context of our interviews, aspects related to semantic interoperability have also arisen in many occasions, mainly related to the need to have a minimal set of common metadata formats inside and across communities and services so that the interpretation of the data is made easier, as well as shared semantic artefacts (ontologies, thesauri) inside and across the communities, which allow homogenising the interpretation and treatment of the exchanged data and all of its associated resources. A clear and well-established governance of all these artefacts is also a relevant requirement. In this sense, best practices have been identified, for instance in the case of CESSDA (Consortium of European Social Science Data Archives) or in many cases in Life Sciences (e.g., Genomics), where community-based repositories of semantic artefacts are being maintained, with a clear governance process.

2.3 Organisational interoperability

According to the EIF, organisational interoperability refers to the way in which organisations align their business processes, responsibilities and expectations to achieve commonly agreed and mutually beneficial goals. This type of interoperability is also focused on meeting the requirements of the user community by making services available, easily identifiable, accessible and user-focused.

Considering the overall agreed goal of Open Science that underlies all the activities at EOSC, this level of interoperability should be focused on the documentation, integration or alignment of the processes of different organisations providing services in EOSC, so as to ensure that researchers can reach their Open Science goals. It should also be clear who is responsible for providing (as well as developing, maintaining and curating) common interoperability services like service catalogues, registers and common PID services, among others.

In the context of our interviews, this is the aspect that has been less discussed, possibly because most research communities are already accounting for the need to align to the overall goals for Open Science that EOSC is looking for. It seemed that most of the interviewees understood the current impediments in their communities (additional work required to register their artefacts as Open Science-enabled ones and provide sufficient metadata, lack of recognition for this additional work, both from institutions and colleagues, lack of commonly agreed principles across funding agencies and organisations with respect to the Open Science approach, etc.).

2.4 Legal interoperability

Within the context of EOSC and the FAIR principles, legal interoperability requires, in particular, that data should be reusable. It concerns the ability to combine datasets from multiple sources without conflicts among restrictions imposed by the license of each dataset.[25] For example, assume that Anna wishes to embed two resources X and Y in a new content, Z, that she is working on. Both X and Y carry a Creative Commons (CC) open license but resource X carries an Attribution-NonCommercial license (CC-BY-NC) while resource Y carries a Creative Commons Attribution-ShareAlike license (CC-BY-SA). If Anna assigns a non-commercial open license to Z (for example the CC-BY-SA-NC), she will breach the terms of the license carried by Y. If Anna chooses a license that allows commercial use of Z (for example by using the CC-BY-SA license) she will breach the terms of the license carried by X. In other words, the licenses carried by X and Y separately, cannot be combined and reused in Anna's derivative work Z.[26] In summary, the fewest

25 Doldirina, Catherine & Eisenstadt, Anita & Onsrud, Harlan & Uhlir, Paul. (2018). Legal Approaches for Open Access to Research Data (p.8). DOI: 10.31228/osf.io/n7gfa.

26 Unless specific permission was granted for commercial use by the right holders of X.

restrictions contained in the source datasets will result in the fewest restrictions contained in the combined or derivative datasets. This implies that only where the data is free of any restrictions and in the public domain (e.g., CC0 or PDDL dedication), legal interoperability will be maximised.

Legal interoperability also concerns situations where regulatory or policy measures restrict the disclosure of data, or that datasets may be made available only in certain jurisdictions or under certain conditions. Examples include legal restrictions based on intellectual property law, national security, the protection of endangered species or privacy regulations, such as GDPR. A number of mechanisms are used in practice to restrict access to data where such regulatory or policy measures exist e.g., embargo, data redaction, data generalisation or simply restricting any access to the data.

FAIR data does not necessarily mean Open. FAIR principles do not restrict recognition of legitimate reasons for shielding data. However, in cases where access to data is restricted or subject to conditions, legal interoperability requires that metadata is FAIR, enabling data discovery, that the conditions for access and use are clearly and readily determinable through automated means and that they do not conflict with each other.

There are a number of 'enabling' legal instruments (EU directives and regulations, national laws, EU and national policies, international agreements, contractual agreements, individual or institutional policies and other forms of practice that may incorporate broader policy considerations) that support legal interoperability and the implementation of FAIR data. For example, the Open Data Directive requires that data generated by public sector bodies follow the principle of 'open by default'. However, there is a need to examine whether obligations or recommendations to use certain licenses, in particular at the national level, are coherent with specific recommendations that are or may be adopted by the EOSC interoperability framework.

Legal interoperability therefore covers the broader environment of laws, policies, procedures and cooperation agreements needed to allow the seamless exchange of information and reusability of data between different individuals, organisations and across jurisdictions. It occurs among multiple datasets from different sources when:[27]

- the legal use conditions are clearly and readily determinable for each of the datasets typically through automated means;
- the legal use conditions imposed on each dataset allow creation and use of combined or derivative products;
- users may legally access and use each dataset without seeking authorisation from data generators on a case-by-case basis assuming that the accumulated conditions of use for each and all of the datasets are met; and
- when access to the data is restricted, metadata is FAIR, i.e., using accepted standards to describe the data and thereby enabling their discovery.

27 *Ibid*. See also White Paper: Mechanisms to Share Data as Part of GEOSS Data-CORE. https://www.earthobservations.org/documents/dswg/Annex%20VI%20-%20%20Mechanisms%20to%20share%20data%20as%20part%20of%20GEOSS%20Data_CORE.pdf

3 MINIMUM REQUIREMENTS AND RECOMMENDATIONS FOR THE EOSC IF

This section presents usual problems and needs that are being faced by the user communities targeted by EOSC, as well as by those aiming at providing services for EOSC. These problems and needs are structured according to the four interoperability layers described in the previous sections (technical, semantic, organisational and legal), and can be understood as requirements for the EOSC IF. They affect the whole range of stakeholders involved in EOSC, from individual users to service providers.

Our problems, needs and recommendations have been compiled through a literature review of common types of requirements reported (including key documents such as the RDA FAIR data maturity guidelines[28] or the aforementioned FAIRsFAIR report on FAIR requirements for persistence and interoperability[29]), as well as through the series of interviews that we have run. A summary of these recommendations is provided in Section 3.6.

3.1 Technical interoperability

3.1.1 Problems and needs

At the level of technical interoperability, some of the problems typically identified by the communities that have been consulted and by ongoing work on other working groups are the following:

- When trying to work with infrastructures or services across communities, **authentication and authorisation often needs to be performed separately for each community/service**. Even though there are technical means and industry-based standards (e.g., SAML2.0, OAuth2.0) to overcome this, authentication often involves transferring personal information between identity provider and service provider, and authorisation is hard to harmonise based on centrally-maintained user attributes.

- Research data may be made **available in multiple general-purpose formats** (CSV, Excel, database dumps, JSON, XML, shapefiles, coding, etc.) **or community-based models** (Darwin Core, VOTable and VOResource, FITS, NetCDF), which are usually hard to align when reusing datasets across communities. In the case of general-purpose formats, semantic interoperability problems also appear because of the lack of agreement in attributes or column headers, the absence of headers or adequate documentation, etc.

- Coarse-grained or fine-grained research data from other communities may be difficult to find, given the **lack of knowledge about how to query their repositories**.

- Multiple service providers for different types of PIDs exist (e.g., IUPAC International Chemical Identifier[30], DOI[31], PURL[32], W3ID[33], Life Science Identifiers[34], handle[35], IVOA[36],

28 RDA FAIR Data Maturity Model Working Group (2020). FAIR Data Maturity Model: specification and guidelines. Research Data Alliance. DOI: 10.15497/RDA00050

29 Lehväslaiho, Heikki, Parland-von Essen, Jessica, Behnke, Claudia, Laine, Heidi, Riungu-Kalliosaari, Leah, Le Franc, Yann, & Staiger, Christine. (2019). D2.1 Report on FAIR requirements for persistence and interoperability 2019. FAIRsFAIR. https://zenodo.org/record/3557381

30 https://www.inchi-trust.org/ (last accessed: 31/Dec/2020)

31 https://www.doi.org/ (last accessed: 31/Dec/2020)

32 https://sites.google.com/site/persistenturls/ (last accessed: 31/Dec/2020)

33 https://w3id.org/ (last accessed: 31/Dec/2020)

34 https://fairsharing.org/bsg-s001184/ (last accessed: 31/Dec/2020)

35 http://handle.net/ (last accessed: 31/Dec/2020)

36 http://www.ivoa.net/documents/IVOAIdentifiers/index.html (last accessed: 31/Dec/2020)

RRID[37]). As a result, different sets of policies are enforced to varying degrees, and sometimes the identifiers are not resolvable (e.g., IUPAC InChi-KEY is a reverse identifier: given the chemical, the identifier can be generated, but not in the opposite direction).

As a result of this analysis, these are some of the needs that can be identified at the level of technical interoperability:

- There is a need for **support for the process of authenticating to and obtaining the rights to use** the services offered by EOSC in a way that is as unobtrusive as possible[38] [Reference: Architecture WG Authentication and Authorization Infrastructure (AAI) principles] and that is independent of any single community.

- There is a need for EOSC to provide a **trusted (and sustainability) framework** across scientific communities, collaborations and infrastructures. For the user this means that what works today will work tomorrow, only better, as referred to in the same document referred to above.

- There is a need for a minimum metadata application profile for the EOSC context to allow users to discover and deal seamlessly with data **available in multiple generic or community-based formats**.

- When searching for research data (or other research objects) that may be reusable across communities, such **data may need to be discovered at different levels of granularity**: high level / coarse-grained (e.g., look for data about DNA sequences or land-use) or low level / fine-grained (inside data collections, e.g., look for a specific DNA sequence or land-use in Hamburg).

- There is a need to have a **common and well-understood PID policy** across communities[39]

3.1.2 Recommendations

Some of the recommendations that can be made to service and data providers in this respect are:

- **Use open specifications**, where available, to ensure technical interoperability when establishing EOSC services.

- **Define a common security and privacy framework** and establish processes for EOSC services to ensure secure and trustworthy data exchange between all involved parties. For instance, there should be an AAI process for EOSC that is common across communities, easy to implement by resource providers and easy to understand by users.

- The **Service-Level Agreements for all EOSC resource providers should be easy to understand** by users from different communities.

- EOSC must enable **easy access to data sources available in different formats**, either generic or community-based, to facilitate overcoming their heterogeneity and

37 https://scicrunch.org/resources (last accessed: 31/Dec/2020)

38 EOSC Authentication and Authorization Infrastructure (AAI). DOI: 10.2777/8702. https://op.europa.eu/en/publication-detail/-/publication/d1bc3702-61e5-11eb-aeb5-01aa75ed71a1/language-en/format-PDF/source-189451671

39 Valle, Mrio; Heughebaert, André; Kotarski, Rachael; Weigel, Tobias; Ritz, Raphael; Matthews, Brian; Manghi, Paolo; Sparre Conrad, Anders; Hellström, Maggie; Wittenburg, Peter (2020) A Persistent Identifier (PID) policy for the European Open Science Cloud (EOSC) DOI: 10.2777/926037

allow integrating data across communities, and to tools enabling the usage of these data.

- **Coarse-grained and fine-grained dataset (and other research object) search tools** need to be made available. There will be a range of general-purpose and domain-specific/specialised search tools, exploiting general-purpose and domain-specific metadata.

- There should be a **clear EOSC PID policy**, accommodating any appropriate PID usage, recognising that established practises are at different levels of maturity for different resources and new PID types may emerge.

3.2 Semantic interoperability

3.2.1 Problems and needs

At the level of semantic interoperability, some of the usual problems that are identified by the communities that have been consulted are the following:

- There is a generalised **lack of common explicit definitions** about the terms that are used by user communities. This is especially a problem in the case of trying to share resources across communities.

- Not only term definitions are usually lacking, but also **common semantic artefacts across communities** (e.g., general ontologies that can be shared). And in case that they exist, these artefacts may not be sufficiently well documented.

- The previous problem is exacerbated by the fact that there is a generalised **lack of common reference repositories** or registries of semantic artefacts (e.g., ontology catalogues). Only some communities are actively maintaining such resources (e.g., Schema.org[40], BioPortal[41], Agroportal[42], CESSDA's Thesaurus Manager System, Linked Open Vocabularies[43]).

- Data collections are usually poorly documented, in terms of the metadata that is made available for them. Besides, there is **no common metadata schema across communities**, what results in different ones being used in different communities (e.g., DCAT, DDI4, DataCite, DarwinCore, RDA Metadata Directory[44], FAIRSharing[45])

- Depending on the discipline, there is a **lack or over-abundance of metadata models** that allow the description, functional preservation and ultimately re-use of the data stored.

- In some communities, there is **lack of expertise and skills related to semantics**, which negatively influences the availability and use of common definitions, semantic artefacts, reference repositories, etc. This aspect is sometimes known as the "human interoperability" problem.

---

40 Three communities are relevant in this context: **Libraries** (https://bib.schema.org/) – they have produced several classes and properties from library and information science; **Archives** (https://www.w3.org/community/architypes/) - their proposal for additional classes can be found on https://www.w3.org/community/architypes/wiki/Alternative_1_model_proposal; **Health and medicine** (https://bioschemas.org/ - Bioschemas aims to improve the findability of data in the life sciences, some types and properties are available on https://bioschemas.org/types/ and another link is https://www.w3.org/community/schemed/

41 Bioportal - https://bioportal.bioontology.org/ (last accessed: 31/Dec/2020)

42 Agroportal - http://agroportal.lirmm.fr/ (last accessed: 31/Dec/2020)

43 Linked Open Vocabularies - https://lov.linkeddata.es/ (last accessed: 31/Dec/2020)

44 RDA Metadata standard directory - http://rd-alliance.github.io/metadata-directory/standards/ (last accessed: 31/Dec/2020)

45 Fairsharing.org - https://fairsharing.org/standards/ (last accessed: 31/Dec/2020)

As a result of this analysis, these are some of the needs that can be identified at the level of semantic interoperability:

- Need for **principled approaches and tools for ontology and metadata schema** creation, maintenance, governance and use. Different communities are using different tools and representation models for their semantic artefacts. It is not uncommon to see UML models being used as standardised models for such representation, lacking sometimes the needed formality to describe terms and their relationships.

- Need for **harmonisation across disciplines**. It should be possible for a user of one community to add metadata to existing items (data and semantic artefacts) according to their own research discipline practices (e.g., a social scientist can add DDI-based metadata for a dataset coming from an environmental scientist). Allow a researcher from a discipline to transform metadata (or data) from one discipline's format/annotations to another.

- Need to **harmonise data of the same type** (e.g., observational data in environmental sciences as being done in the I-ADOPT RDA WG, a consistent coding for geographical locations where a sample was obtained, etc.).

- Need for **federated access over existing research data repositories** (both inside a discipline and across disciplines). How to support discovery of data on the basis of a high-level description, and possibly also on more details like concepts related to observations and variables?

3.2.2 Recommendations

Some of the recommendations that can be done in this respect are:

- All communities should generate **clear and precise definitions for the concepts** that they use, as well as their metadata and data schemas. These definitions should be **publicly available**, referenced by a persistent identifier and shared in EOSC. Furthermore, a classification for research disciplines (e.g., The German Research Foundation's subject area classification[46] or Frascati manual[47]) should be also explicitly chosen and implemented within the EOSC context.

- Semantic artefacts should be available **preferably using open licenses** (e.g., like in W3C).

- Every semantic artefact that is being maintained in EOSC must have **sufficient associated documentation**, with clear examples of usage and conceptual diagrams. Furthermore, any semantic artefact should also be FAIR.

- EOSC should provide support for the **maintenance of a repository of semantic artefacts, and a governance framework** for such a repository. For example, SKOS thesauri may be maintained using services similar to the CESSDA Vocabulary Service.

- A minimum metadata model should be proposed in the future to ease **discovery over existing federated research data and metadata** (based on the reuse of existing standards like DCAT-AP, DDI 4 Core, DataCite core schema, etc.). This metadata model is not meant to replace existing standards, but to facilitate findability and to support interoperability not only within domain-specific services or repositories, but also across domains and communities. A set of alignments (also known as crosswalks) among existing metadata models should be maintained (initial work with some of the most common metadata models is presented in Appendix I), and corresponding building

46 https://www.dfg.de/en/dfg_profile/statutory_bodies/review_boards/subject_areas/index.jsp (last accessed: 31/Dec/2020)

47 https://en.wikipedia.org/wiki/Frascati_Manual (last accessed: 31/Dec/2020)

blocks for metadata exchange should be made available for establishing crosswalks among services and data repositories. This is a priority for the next phase of the EOSC portal and is critical to enable the federation of research data. In defining a Minimum Metadata Framework, close collaboration and consultation with research communities is essential to ensure it is fit-for-purpose and can be adopted.

- There should be **extensibility options** to allow for disciplinary metadata that is typical for some research communities, allowing users/researchers to add annotations according to the established practices of their communities, if relevant (e.g., a social scientist adding DDI-based metadata on a GIS dataset that has only geographical-oriented metadata), with sufficient provenance information on the annotations, and with versioning support.

- Not only data should be considered in this context, but also the recommendations should be extensible to other types of resources used in Science, such as **software, methods, scientific workflows, laboratory protocols**, **hardware designs**, etc.

- There should be clear protocols and building blocks for the **federation/harvesting of semantic artefacts catalogues**. These components are discussed in the architectural work presented in Section 4.

3.3 Organisational interoperability

3.3.1 Problems and needs

At the level of organisational interoperability, some of the usual problems that are identified by the communities that have been consulted, as well as by taking into account the Rules of Participation working group output at the time of making this analysis[48], are the following:

- There is not yet (although it is expected soon) a clearly-defined governance structure for EOSC that includes the **governance framework that will deal with interoperability across organisations and disciplines**, among many other aspects.

- There is not yet a **clear description of the "terms and conditions" and "acceptable use policies" that will rule the services provisioned by EOSC**, and most specifically in what respects to the management of interoperability aspects (e.g., how will metadata services be ruled, the governance of metadata schemas and other semantic resources, etc.).

- The current draft of the Rules of Participation does not enter into the details of how interoperability will be achieved across organisations and user communities in the context of EOSC.

- It is not always clear for users whether the infrastructures or services that they can use from other communities will be still running in the medium or long-term, because of **lack of knowledge about their sustainability policies** or robust long-term funding plans for the services.

As a result of this analysis, these are some of the needs that can be identified at the level of organisational interoperability:

- Need for a **clear governance framework** that includes clear instructions on how the other levels of interoperability will be handled across organisations and user

48 EOSC Rules of Participation (v0.5). https://www.eoscsecretariat.eu/sites/default/files/draft_eosc_rop_version_0.5_20-10-2020.pdf (last accessed: 31/Dec/2020)

communities (data formats, AAI services, metadata schemas, ontologies, etc.). This should also include the management of permanent organisation names and functions.

- Need for **documents explaining terms and conditions and acceptable use policies for services providing interoperability**. For instance, providing clear descriptions of the service-level agreements of those providing catalogues and registries of semantic artefacts, or providing systems to overcome semantic differences between different data sources, or alignments between models.

- Need for **interoperability certification mechanisms for service providers**, so that service users can set their own expectations about the support for interoperability of those services.

3.3.2 Recommendations

Some of the recommendations that can be done in this respect are:

- The **current set of rules of participation recommendations should be completed with aspects related to interoperability**. For instance, for data providers this may include asking explicitly that data is published according to specific data formats and/or vocabularies for a specific community.

- The same is applicable to **services**, which may be recommended to ingest or output data according to such standardised data formats and/or vocabularies, and with their corresponding metadata, with some level of quality.

- A clear management of **permanent organisation names and functions** needs to be provided.

3.4 Legal interoperability

The analysis of problem and needs, and the proposal of recommendations for this layer, are organised according to different groups of topics, related to: copyrights and licenses, other forms of Intellectual Property Rights (IPR), General Data Protection Regulation (GDPR), sensitive data, private law, enabling legal instruments.

3.4.1 Problems and needs

Copyright and licenses

- Lack or deficiencies in the provision (and awareness by users) of **clear statement of rights or information about legal conditions for reuse** of data hinders legal interoperability.

- Datasets may be subject to **different licenses which may not be compatible with each other**. This could limit reusability and combination of data.[49]

- Some historic copyrightable datasets and metadata may have **no license or unclear licenses arrangements** ('orphan data').[50] Without permission, a waiver, or specific

49 Depending on the type of license used for each dataset, this may mean that: (1) the conditions of use of one dataset negate the conditions of use of another data set, so the two (or more) datasets cannot be combined and carried forward; or, (2) the conditions of use of one dataset do not negate the conditions of use of another dataset, but the accumulated restrictions carried forward under the combined datasets are more restrictive than the initial conditions of use for one (or more) of the original datasets ('lowest common denominator' effect).

50 We use the term 'orphan data' in contrast to 'orphan works' which is used in the Orphan Works Directive 2012/28/EU. We consider that the Orphan Works Directive does not address the issue from a FAIR perspective completely because of its scope (the definition of 'orphan works' is relatively narrow) and because it has been criticised on the basis that it presumes that the reuse of orphan works should be

exemption, no one may use or sublicense the dataset and consequently, orphan data are often left unused due to the impossibility or the disproportionate cost to trace the copyright holder.[51]

- **National copyright protection may vary across jurisdictions**. In the absence of a permissive license (e.g., CC-BY) or a no-rights-reserved statement (e.g., CC0), data may be protected and limit re-use in a jurisdiction, but not in the other. Without permission, a waiver or an exemption (e.g., 'fair dealing' in the UK), it will be important to understand the scope of copyright conditions that regulate the use of the data and to obtain permission from the originator. Otherwise, data users may inadvertently breach the originator's copyright.

- Each dataset may, in practice, **include different copyrightable assets** ('embedded data'). For example, if a photo is embedded in a dataset, the photo may be subject to a separate license.

- Users within the EU need to obtain permission for (re-)using the whole compilation of a database, while those outside the EU are not required to do so, due to the territorial nature of DB rights (they only apply within the EU/EEA).

- Stakeholders raised certain data-sovereignty related concerns such as:[52] (1) the risk that open database, datasets and related software will be copied by an external entity and offered as a service on a commercial basis; or (2) control of downstream use, such as information on who has been granted access to data and for what purposes the data was used for; or, (3) wish of some stakeholders to include restrictions on duration and territorial use of data; or, (4) liabilities concerning issues such as inaccuracies, misuse, breach of privacy laws.

- User rights, restrictions and conditions of use may change over time and right-statements made in the past may no longer reflect the current rights-holder ownership claim regarding the data.

Other forms of IPR

- Most data as such is not patentable. However, patent requirements such as 'novelty' and 'inventive step/non-obviousness' may limit data generators' incentives to make data available,[53] in particular, if they believe that their data contributes to the description of a possible invention, or if the data fills a gap in the general knowledge, so that potential follow-up inventions are rendered "obvious" to a person skilled in the art.[54]

restricted. This situation may occur, for example, when the author is unknown, or deceased, leaving no locatable heirs, or when the holder of the copyright was a legal person but it has ceased to exist with no legal successor e.g., due to liquidation, or where any records about copyright ownership have been lost.

51 "*The British library estimates that 40% of works in their collections are orphan and over 1 million hours of TV programmes from BBC archives are not used due to the impossibility or the disproportionate cost to trace rightholders – and the risk of a subsequent legal action is simply too great for this material to be made available online*". Neelie Kroes, former Vice-President of the European Commission responsible for the Digital Agenda, addressing the challenge in the context of the Orphan Works Directive: https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_11_163.

52 See Proposed Re-Charter, RDA-CODATA Interest Group on Legal Interoperability of Research Data (Revision Sept. 2019), https://www.rd-alliance.org/sites/default/files/2019-09-20_IG-Charter-2019-Post-PHIL_0.pdf

53 In some cases, they may also be prevented from doing so due to legal requirements or organisational policies.

54 See Carroll MW (2015) Sharing Research Data and Intellectual Property Law: A Primer. PLOS Biology 13(8): e1002235. https://doi.org/10.1371/journal.pbio.1002235, also relating to a situation whereby a patentable process could claim a series of steps that would be practiced in connection with certain forms of data reuse.

- In some cases, software and functional data with a technical effect may form part of patent claims. If such functional data is made accessible, then specific reuses of the data may infringe patent rights.

General Data Protection Regulation (GDPR)

- The GDPR introduces constraints on handling personal and sensitive data[55] which may result in (1) legal impediments to making certain data open;[56] or, (2) Data producers, service providers and users inadvertently breaching the GDPR.

- The GDPR restricts the "processing"[57] of personal data – unless legal grounds exist such as consent by the data subject. Obtaining informed consent for each and every dataset is not practical (e.g., impossibility or disproportionate cost).

- Stricter rules as a result of the GDPR mean that transferring the personal data of EU nationals to third parties outside of the EU requires additional safeguards. Changing practices regarding the implementation of these rules such as the recent invalidation of the EU-US Privacy Shield Framework by the Court of Justice of the EU (CJEU)[58] have exacerbated this problem.

- A Data Privacy Impact Assessment (DPIA) is required when the data controller[59] begins to process personal data in a way that is likely to involve a "high risk" (e.g., when special categories of personal sensitive data are processed). Each EU Member State has implemented different guidelines and processes for completing a DPIA.

Sensitive data

- Certain laws and conventions may restrict the disclosure of, access to, or use of specific data ('sensitive data'). This may be in connection with, for example, the protection of endangered species, traditional cultural resources, national security, sovereign genetic resources or traditional knowledge.

- Measures used to restrict access to sensitive data include the generalisation of data, redaction of specific information (such as location of an endangered species), specific contractual arrangements, embargo periods, etc. However, even when specific data is redacted or generalised, sensitive information may nevertheless be deducted.

- In some cases, specific laws that restrict access to sensitive data may only be applicable in one jurisdiction but not in others. Equally, certain data, such as traditional knowledge[60] may be afforded protection by applicable intellectual property law in one

55 In the context of the GDPR 'personal sensitive data' relates to racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation. See Article 4 (13) (14) (15); Article 9; Recitals 51 and 56 of the GDPR.

56 For example, the use of PIDs may need to take into account constraints introduced by the GDPR, such as the right to erasure of personal data.

57 "Processing" has a broad meaning and means almost any operation in connection with personal data such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, (re-)use, disclosure by transmission, dissemination or otherwise making available of personal data.

58 Case C--311/18 (Schrems II case) of 16 July 2020.

59 The "data controller" is the entity who "defines the means and purposes of the processing" (Art. 4(7) GDPR).

60 There is no universal international definition of traditional knowledge. The World Intellectual Property Organisation (WIPO) considered it to include "knowledge, know-how, skills and practices that are developed, sustained and passed on from generation to generation within a community, often forming part of its cultural or spiritual identity". See: https://www.wipo.int/tk/en/tk/

jurisdiction but will not be afforded protection under intellectual property law in other jurisdictions.

Private law

- Repositories may use different terms and conditions of use, which may not be compatible with each other when data is combined from different repositories.
- Over regulated EOSC environment, onerous requirements and risk of liability to data generators or users may compromise the free flow of data.

'Enabling' legal instruments

- There are a number of enabling legal instruments that provide incentives or oblige public bodies, research funders, institutions and individuals to open data and apply open and FAIR principles. However, it is important to secure coherency between the requirements of such legal instruments and the general recommendations for the EOSC. For example, the Open Data Directive recommends the use of standard licenses for reuse of research data but leaves it open to the Member State to decide on the type of open licenses to be used. There is no guarantee that licenses recommended in the different Member States will be compatible with each other or with those recommended by EOSC.

As a result of this analysis, these are some of the **needs** that can be identified at the level of legal interoperability:

Copyright and licenses

- Prior to making data available, repositories or disseminators of data need to ascertain who holds the rights to the data, including any embedded data.
- Metadata needs to be available without restrictions on (re-)use in order to facilitate the FAIRness of the data it describes.
- User rights and use-conditions need to be clearly provided for each set of data. Users should be clearly informed of their rights and obligations.
- Licensing requirements should be both human and machine readable and allow data providers and users to understand license compatibility.
- There is a need to clarify the status of historic copyrightable datasets and metadata which have no license or unclear licenses arrangements and where the copyright holder is unknown or not reachable ('orphan data').
- There is a need to address concerns of stakeholders in relation to misuse and exploitation of data, liabilities and responsibilities connected with data reuse.
- Automatic database (sui generis) rights should not impose unintended restrictions on re-use for EU-based users (compared to non-EU users).
- User rights, restrictions and conditions of use may need to be updated from time to time and repositories of data must allow for an easy mechanism of doing so, including an audit trail for any licence changes.

Other forms of IPR

- Data generators may want to keep certain data secret or redact part of the data either (1) because of patent strategies; or (2) in order to protect data and knowhow as trade secrets where patents are not available.

- Users must be made aware of potential reuse restrictions, such as specific technical applications of functional data that form part of a patent filing or any other IP protected material which is included with the data.

GDPR

- There is a need to ensure adequate protection of personal data and general compliance with the GDPR and EU Member States' domestic law and guidelines (such as Data Protection Impact Assessment - DPIA - guidelines), where applicable.

- Need for communicating restrictions and limitations in human and a machine-readable form, for example, in the metadata.

Sensitive data

- Open data and FAIR principles need to be balanced against legal restrictions and legitimate interests, such as the protection of national security, endangered species, cultural resources, protection of sovereign genetic resources and traditional knowledge. At the same time, there is a need to ensure that restrictions imposed have a legitimate basis and do not go beyond what is necessary and required by law or by ethical and legitimate requirements.

- Mechanisms are required to ensure that combined or derivative data will not inadvertently generate information which is considered to be sensitive under the terms of use of one or more of the parent datasets.

'Enabling' legal instruments

- Ensure that open licensing requirements and recommendations provided by enabling legal instruments, particularly on Member States level, are coherent with the requirements and general recommendations provided by the EOSC.

3.4.2 Recommendations

Some of the recommendations that can be raised in this respect are:

Copyright and licenses

- All copyrightable **data and metadata should include a standardised human and machine-readable license** to downstream users, including a standardised statement of user rights, legal restrictions, applicable licenses and additional conditions of use (including applicable jurisdictions). EOSC should consider developing **a centralised source of knowledge and support on copyright and licenses** to users and data generators and to address common Q&A (e.g., something similar to http://licenses.openscience.si), and develop and implement **minimum standardised, human and machine readable, expressions of right statements and use conditions**.

- Copyrightable **metadata should always be assigned a permissive license** with no, or only legally necessary restrictions (e.g., CC0, PDDL). And copyrightable **data should be preferably assigned a permissive license**, unless legal or legitimate reasons apply. That is, open and permissive licenses, and the use of restricted data access collections[61] are preferred over the use of ad-hoc contracts entered into between a right

61 See RDA-CODATA Legal Interoperability Interest Group. (2016, October 20). Legal Interoperability of Research Data: Principles and Implementation Guidelines. Zenodo. http://doi.org/10.5281/zenodo.162241, on p. 21.

holder and a data user. Furthermore, from a license compatibility perspective, attribution should preferably be pursued by means of moral, ethical or other obligations (e.g., the European Code of Conduct for Research Integrity or the development of Persistent Identifiers, etc.) rather than the use of the CC-BY 4.0 license. **The CC0 is, in general, preferred over the CC BY 4.0**.

- Provide a mechanism to facilitate the inclusion of relevant information in the metadata to **identify different components in the dataset that are subject to different copyright** or under what license each component is provided.

- Instances of **expired or non-existent copyright**, or where data is already in the public domain, **should be clearly marked** (e.g., using CC PDM or equivalent). Besides, adopt a uniform set of recommendations or guidance on how to handle copyrightable dataset where the owner is unknown or not reachable and the data has no license assigned to it (e.g., an 'orphan data' standardised notice and related legal implications).

- The use of Creative Commons licenses is generally not recommended for licensing source code for software. An open and permissive license such as the MIT license or equivalent could be used instead.

- **A list of EOSC-recommended licenses and their compatibility with Member States' recommended licenses should be provided**, so as to avoid an inadvertent breach of copyright and with a view to harmonise and reduce the overall number of recommended licenses.

- Assess stakeholders' data-sovereignty related concerns and consider whether the authorisation and authentication processes (or similar mechanisms) should allow for additional control of downstream use.

- Data repositories should incorporate harmonised mechanisms to validate and allow for the **update of restrictions, right statements and conditions of use on data as these may change over time**. Data licences should only become more permissive, not more restrictive after first being shared within EOSC.

Other forms of IPR

- While remaining as open as possible, EOSC should balance the various legal interests and allow for the **seeking of IP protection of certain data in justified cases** where the disclosure of the data may compromise the ability to file for patents or protect trade secrets.

- Develop a **harmonised policy and guidance to dealing with instances where patent filing or trade secrets may be compromised by disclosure**.

- Metadata should indicate **reusability restrictions on software or data due to pending or existing patent claims** or when data had been redacted due to commercially confidential information.[62]

GDPR

- EOSC data and service providers (and any data controllers and data processors[63]) should implement appropriate mechanisms to **ensure compliance with the GDPR for all**

62 Or in other cases, such as regulatory exclusivities common in clinical trials.

63 The "data processor" is the entity which processes personal data on behalf of the data controller if the controller did not process personal data directly themselves but outsourced the task (Art. 4(8) GDPR).

personal data. In particular, the following requirements and principles shall be taken into consideration:[64]

- "lawfulness, fairness and transparency" (Art. 5, 6 and 9 GDPR), "purpose limitation" (Art. 5, 6 and 26 GDPR), "data minimisation" (Art. 5 and 26 GDPR), "accuracy" (Art. 5 and 16 GDPR), "storage limitation" (Art. 5 GDPR), "integrity and confidentiality" (Art. 5, 24 and 32 GDPR).

- Protocols for the implementation of data portability (Art. 20 GDPR), right to be forgotten (data erasure) (Art. 17 GDPR) as well as notification in the event of data breach (Art. 33 GDPR).

- Legal requirements should be technically implemented following the "privacy by design and by default" (Art. 25 GDPR) approach. In particular, making clear reference to data minimisation. The data controller must implement appropriate technical and organisational measures – such as pseudonymisation – for ensuring that only personal data which are necessary for each specific purpose of the processing are processed e.g., the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility.

- Legal safeguards including data security measures where international data transfers take place (where possible, in a machine-actionable manner such as in Service Level Agreements (SLAs) and the data management system).

- Data protection impact assessment to ensure that appropriate protections are in place (where possible, in the architecture design of software and SLAs).

- Extreme caution and necessary measures must be taken to protect personal data, in particular personal sensitive data. Data anonymisation may be the most appropriate approach to protect personal data since anonymised data can be shared for secondary purposes – such as scientific research – without placing individual privacy at risk.

Sensitive data

- Additional restrictions on access and use of data should only be applied in cases of applicable legislation or legitimate reasons. EOSC should consider preparing a list of 'legitimate reasons' that go beyond existing legislation (e.g., protection of transitional knowledge) and which could justify the introduction of additional restrictions on access to and reuse of data.

- Adopt a procedure for monitoring or reporting violations of use conditions and leakage of sensitive data.

Private law

- Repositories' terms of use should enable the enforcement of applicable rights so that users are made aware of, and can abide by, the specific rights applicable to re-use of data in relevant jurisdictions.

- Repositories' **terms of use should be harmonised to the extent possible** so as to avoid conflicting terms of use where data is combined from different disciplines and repositories.

'Enabling' legal instruments

64 The list shown here is not exhaustive but enunciative.

- Follow development and implementation of enabling legal instruments and coordinate directly with relevant entities and Member States to ensure that recommendations and adoption of open licensing requirements are coherent with the general recommendations provided by the EOSC.

3.5 Some general recommendations from the EIF

We also include here some general recommendations extracted from the EIF, which are applicable to the EOSC IF with some adaptations. These have been included as part of the more specific recommendations in the previous sections, and are maintained in this separate section to facilitate tracing back to the original EIF proposals:

- Ensure that national interoperability frameworks and interoperability strategies are aligned with the EOSC IF and, if needed, tailor and extend them to address the national context and needs.
- Publish research outputs openly unless certain restrictions apply (*"as open as possible, as closed as necessary"*).
- Give preference to open specifications, taking due account of the coverage of functional needs, maturity and market support and innovation.
- Use open source software. And if software needs to be implemented for data generation, presentation or analysis, it should be well developed, documented and published as open source.
- Reuse and share solutions (e.g., software components, Application Programming Interfaces, standards), and cooperate in the development of joint solutions when implementing EOSC services.
- Reuse and share information and data when implementing EOSC services, unless certain privacy or confidentiality restrictions apply.
- Secure the right to the protection of personal data, by respecting the applicable legal framework.
- Ensure that all EOSC services are accessible to all research organisations, researchers, citizens, including persons with disabilities, the elderly and other disadvantaged groups. EOSC services should comply, as much as possible, with e-accessibility specifications widely recognised at EU or international level.
- Ensure data portability: the data should be easily transferable between systems and applications supporting the implementation and evolution of EOSC services without unjustified restrictions.
- Use multiple channels (physical and digital) to provide EOSC services, to ensure that users can select the channel that best suits their needs.
- Put in place mechanisms to involve users in analysis, design, assessment and further development of EOSC services.
- As far as possible under current legislation (especially GDPR), ask users of EOSC services once-only and relevant-only information.
- Use information systems and technical architectures that cater for multilingualism when establishing an EOSC service.

- Formulate a long-term preservation policy for information on EOSC services. To guarantee the long-term preservation of digital records and other kinds of information, formats should be chosen to ensure long-term accessibility.

Finally, when conducting interviews, we learned that "human interoperability" sometimes was an overlooked perspective that related to the common use of resources for interoperability such as metadata standards, terminologies/ontologies, licenses among others. Although the services provide machine readable representations of the different artefacts the people setting up mappings to, or using, metadata standards, concepts, licenses etc. often have different grounds for interpreting them and how the work should be done. We would therefore like to lift the perspective of human interoperability and the common FAIR resources needed to build the skills and competence needed to set a common ground for shared FAIR resource usage. The EIF also points to the lack of skills/competence needed to enable interoperability as "a barrier to implementing interoperability policies"[65], common FAIR resources to build skills and competence can contribute to remedy this.

3.6 Summary of recommendations

The following table summarises all the recommendations provided in this section, organised by layers.

Layer	Recommendation
Technical	• Open Specifications for EOSC Services. • A common security and privacy framework (including Authorisation and Authentication Infrastructure). • Easy-to-understand Service-Level Agreements for all EOSC resource providers. • Easy access to data sources available in different formats. • Coarse-grained and fine-grained dataset (and other research object) search tools. • A clear EOSC PID policy.
Semantic	• Clear and precise, publicly-available definitions for all concepts, metadata and data schemas. • Semantic artefacts preferably with open licenses. • Associated documentation for semantic artefacts. • Repositories of semantic artefacts, rules with a clear governance framework. • A minimum metadata model (and crosswalks) to ease discovery over existing federated research data and metadata. • Extensibility options to allow for disciplinary metadata. • Clear protocols and building blocks for the federation/harvesting of semantic artefacts catalogues.
Organisational	• Interoperability-focused rules of participation recommendations. • Usage recommendations of standardised data formats and/or vocabularies, and with their corresponding metadata. • A clear management of permanent organisation names and functions.
Legal	• Standardised human and machine-readable licenses, with a centralised source of knowledge and support on copyright and licenses.

65 http://doi.org/10.2799/78681, p23

	● Permissive licenses for metadata (and preferably for data, whenever possible). And CC0 preferred over CC BY 4.0. ● Identification of different parts of a dataset with different licenses. ● Clearly marked instances of expired or inexistent copyright, as well as for orphan data. ● A clear list of EOSC-recommended licenses and their compatibility with Member States' recommended licenses. ● Tracking of license evolution over time for datasets. ● Harmonised policy and guidance to dealing with cases where patent filing or trade secrets may be compromised by disclosure. ● GDPR-compliance for personal data. ● Additional restrictions on access and use of data only applied in cases of applicable legislation or legitimate reasons. ● Harmonised terms of use across repositories ● Alignment between Member States national legislations and EOSC.

4 TOWARDS THE EOSC IF: MODEL AND COMPONENTS

This section describes our proposal for the design and implementation of the EOSC IF. It discusses first the proposed model for the description of digital objects to be maintained and shared in EOSC, and then proceeds with a further description of the basic components of such a digital object model.

It is important to mention that this document does not provide a concrete recommendation on how digital objects should be implemented nor on how the basic building blocks of the reference architecture presented in Section 5 should be implemented and delivered, as this is out of the scope of this document.

4.1 Model overview

The Strategic Research and Innovation Agenda (SRIA) builds upon the outputs of the EOSC Executive Board Working Groups and the results from the Open Consultation processes, including key guiding principles and recommendations for EOSC IF. Additionally, the SRIA has already proposed a schematic representation of EOSC key elements: EOSC-Core, Federated Data and EOSC-Exchange to enable the federation of existing and planned research data infrastructures for the benefit of publicly funded researchers, to access openly available FAIR data and services. As the EOSC-Core aims to provide frameworks to discover, share, access and reuse resources together with core capabilities to transfer, store, process or preserve research data, EOSC IF is located around this layer respectively with corresponding elements representing key players: service and data providers as well as FAIR Digital Objects and users involved in EOSC ecosystems (Figure 1).

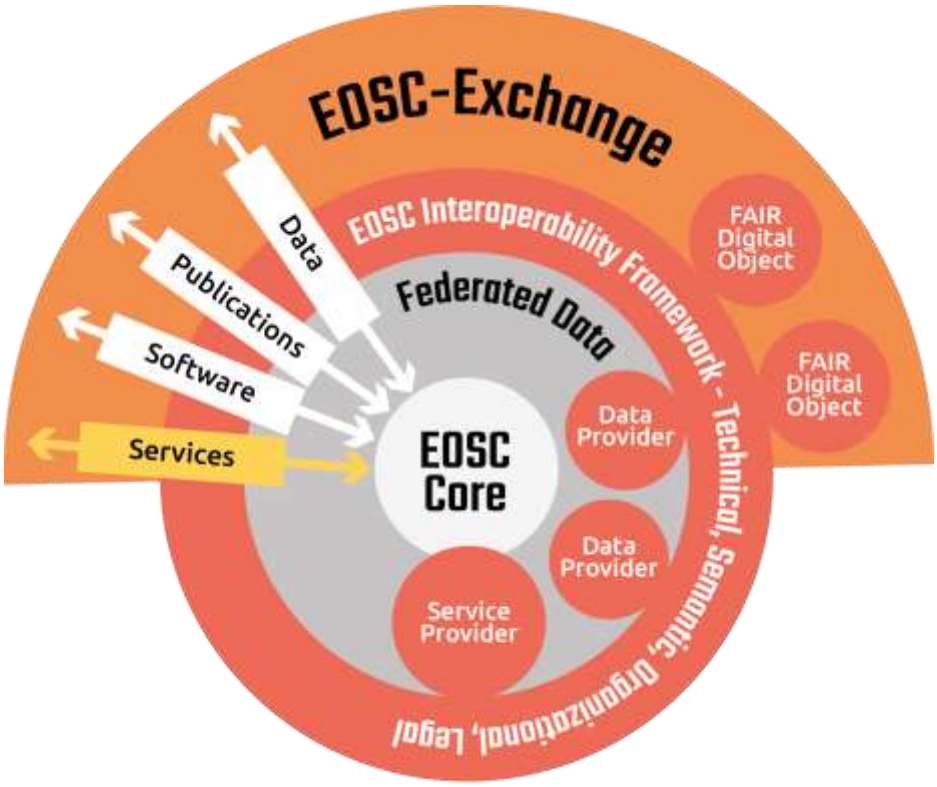


Figure 1. Schematic representation of key elements of EOSC Interoperability Framework.

In order for EOSC service providers, data providers and service consumers software to form a consensus on how digital objects are to be read, interpreted and used, they need to be able to use an agreed upon set of references to common resources describing these different aspects.

At the core of the EOSC IF we find the concept of FAIR Digital Object, which is described in the EC report "Turning FAIR into reality" as "the atomic entity for a FAIR ecosystem".[66] The FAIR digital object metadata layer will be essential in providing the elements needed to achieve different degrees of interoperability within the FAIR ecosystem. In figure 2 below we visualise the links between the metadata needed for interoperability according to the four layers that have been discussed in this document.

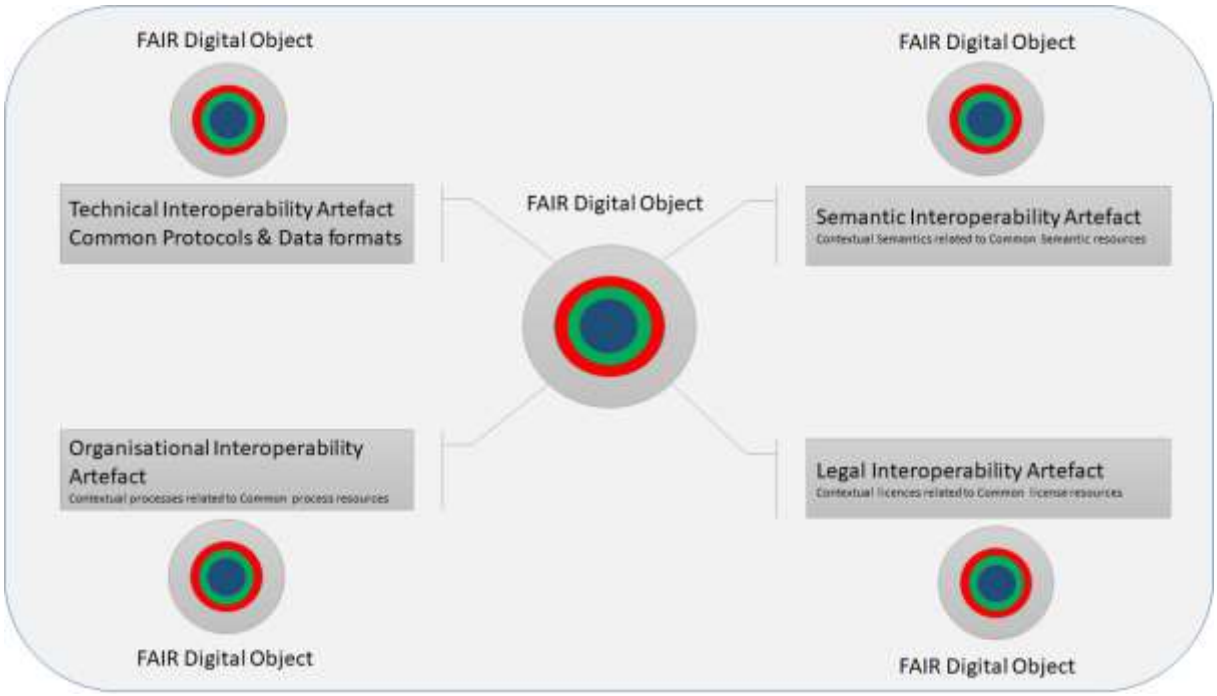


Figure 2. FAIR Digital Objects and their associated metadata, which may be themselves other FAIR Digital Objects.

That is, the metadata associated to a Digital Object should not be seen as a unique block of metadata, but rather as a set of layers that will point to different metadata items that will allow dealing with the problems and needs identified at each interoperability layer (technical, semantic, organisational, legal). Indeed, many different metadata items may exist for aspects related to the same layer.

The links available in the different metadata sections of these digital objects will normally resolve into FAIR digital objects themselves, possibly using a PID infrastructure supporting it, as described in the paper "Digital Objects as Drivers towards Convergence in Data Infrastructures".[67] Such a framework based on FAIR digital objects with PID links to common artefacts addresses problems expressed by the community during the interviews, as discussed in Section 3. For instance, the general lack of common explicit definitions that describe the data to be exchanged in a machine-readable way can then be met by linking, with a persistent identifier, to a common semantic artefact that is shared within the EOSC.

Furthermore, the FAIR digital object that acts as "the atomic entity for a FAIR ecosystem"[68] can exist on several levels of granularity, this is visualised in figure 3 below. One example can be a variable acting as a FAIR digital object in its own right, with references to the FAIR resources needed for interoperability, but also being a part of a dataset that is a FAIR digital object on a less granular level. If an EOSC service provides a digital object, the interoperability PID links will thus be provisioned with the object regardless of its granularity. The following figure provides an overview of how a FAIR Digital Object may be

66 https://doi.org/10.2777/1524, p39

67 http://doi.org/10.23728/b2share.b605d85809ca45679b110719b6c6cb11

68 http://doi.org/102777/1524, p39

decomposed into other Digital Objects as we are interested in obtaining more granularity in the access to its components.

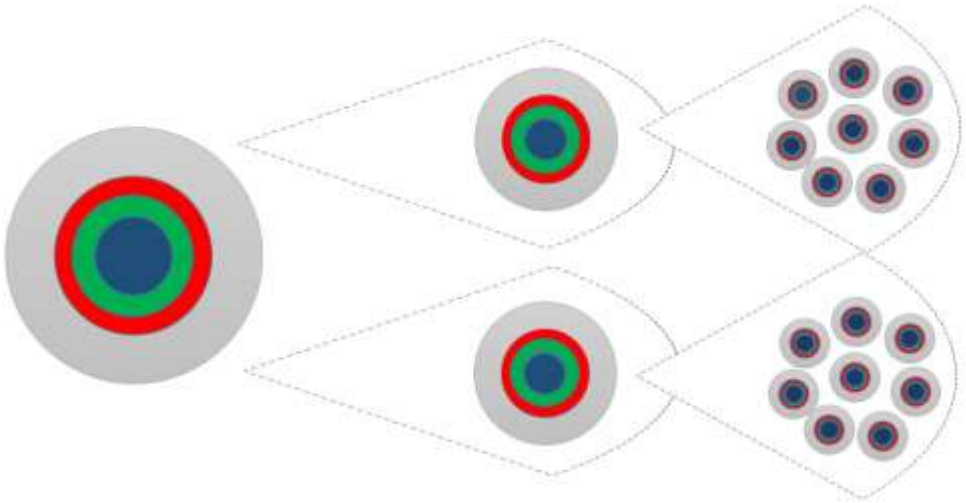


Figure 3. A FAIR Digital Object may be decomposed in other more fine-grained ones.

Figure 4 provides an additional overview of the role that FAIR Digital Objects will play in the context of the exchange of data and metadata, and the need to ensure that applications and services can interoperate. This is especially relevant in the context of providing applications as a service, rather than purely focusing on providing infrastructure or platform as a service. All the different layers of metadata will be relevant for this purpose.

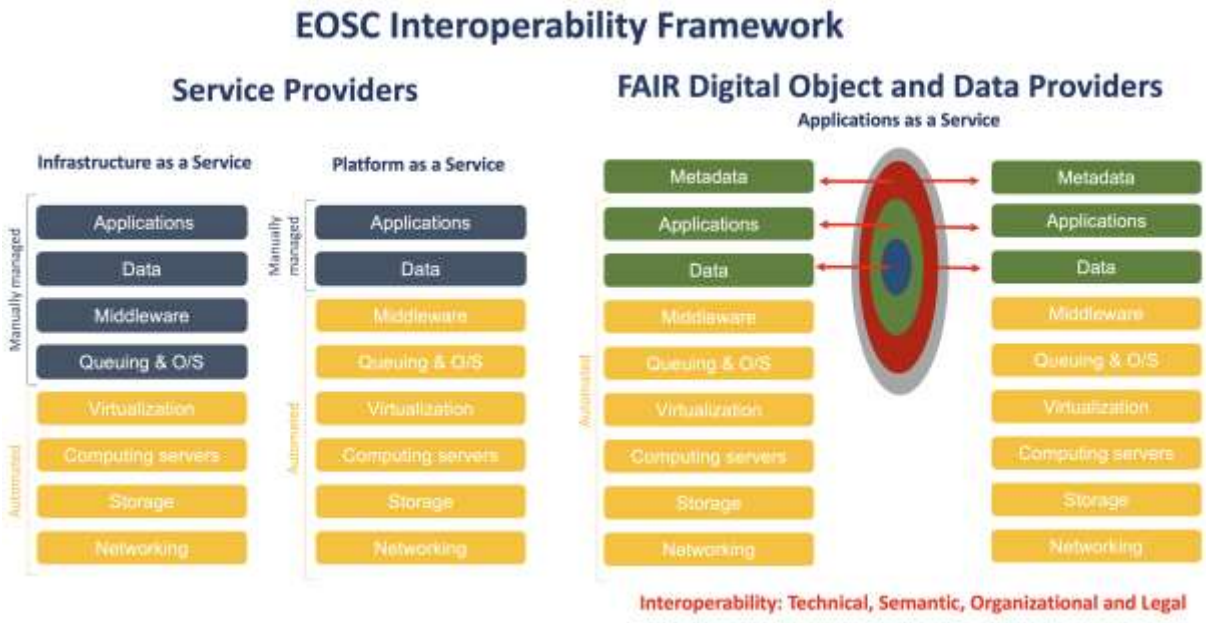


Figure 4. The role of FAIR Digital Objects in achieving interoperability of data, metadata and applications.

Figure 5 provides some high-level views on some examples of how data and metadata have been described in previous efforts to address technical and semantic interoperability (e.g., using Linked Data, microservice architectures, scientific workflows):

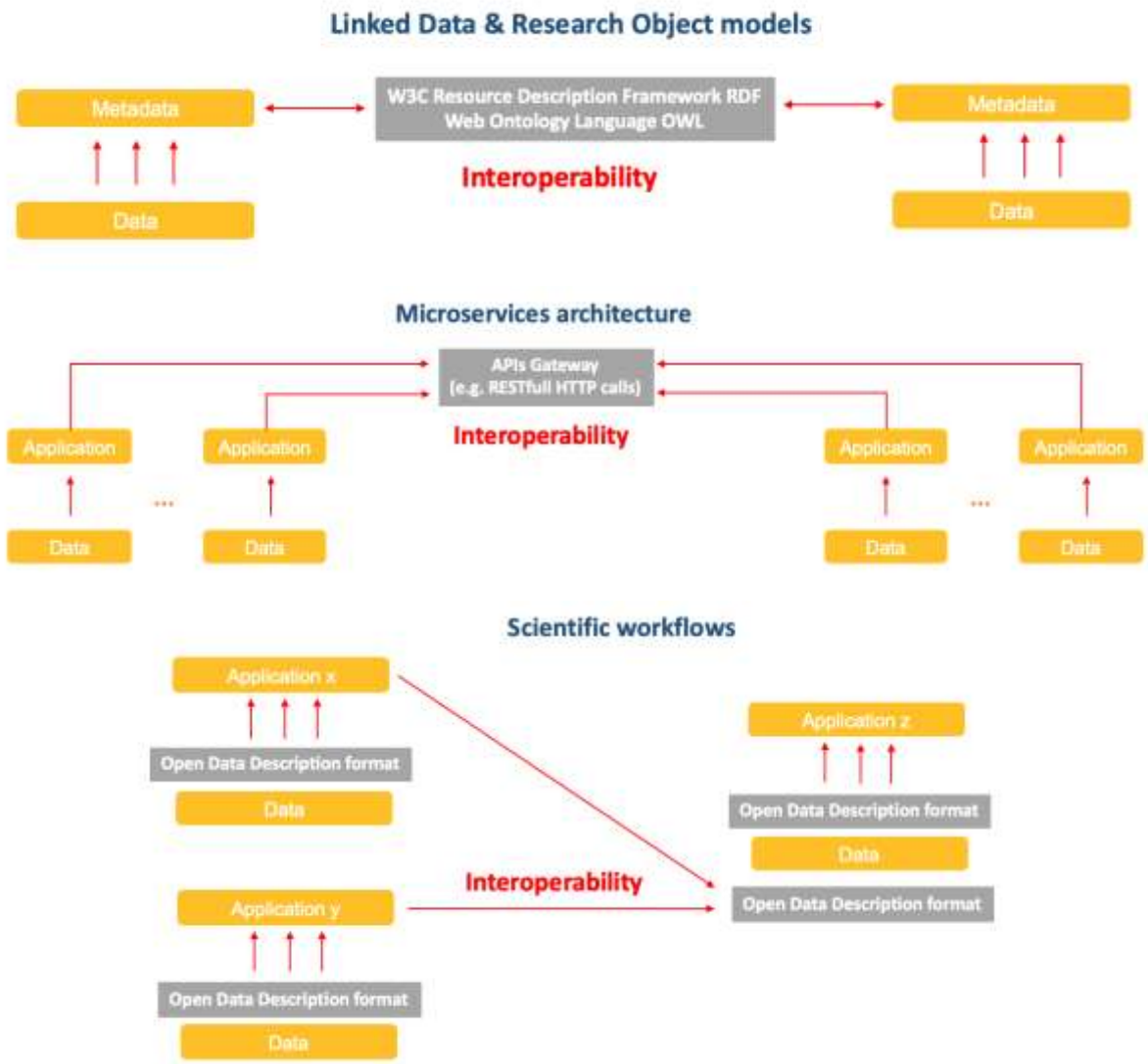


Figure 5. Sample approaches for interoperability (Linked Data, microservice architectures, scientific workflows).

4.2 Basic components

This section provides an overview of some of the basic components needed for the EOSC IF, considering our proposal for FAIR Digital Objects.

4.2.1 Generic and community-specific semantic artefacts

In order to reach semantic interoperability, the FAIR Digital Objects and the metadata elements describing them need to describe and interpret their content in a clear and machine-readable way. Links to concepts and properties within generic and community-specific semantic artefacts is a way of enabling this.

The different types of semantic artefacts can be specified according to the semantic spectrum already discussed in section 2, with two main groups of semantic artefacts: (i) those that are formally represented (e.g., heavyweight, like ontologies or more lightweight, like thesauri), and (ii) those that are less formal (e.g., UML models, database models, XML schemas). Independently of the type of semantic artefacts that we need to deal with, these

may be made available as well using FAIR principles (e.g., as discussed in the paper "Coming to Terms with FAIR Ontologies"[69]).

Indeed, less formal semantic artefacts can provide links to more formal semantic artefacts to increase their level of interoperability. A UML model containing a class with the name "Person" can in this scenario provide a machine-readable link to a concept defining the meaning of "Person" in an existing ontology, hence more clearly and precisely than what is possible to deduct from the class references in the UML model. Domain agnostic metadata standards are also often used to map one standard to another as a way to establish commonalities and create interoperability.

4.2.2 Generic metadata frameworks and data type registries

The core mechanism provided by generic metadata frameworks (aka conceptual metadata standards) and the data type registry model set the foundation for semantic interoperability.

Generic metadata frameworks (e.g., GSIM[70], ISO11179[71]) determine how metadata should be described and what metadata concepts should be used. They do not describe the specific concepts and properties used in a domain, since this is reserved to the domain-specific metadata frameworks and to specific semantic resources.

In the Generic Statistical Information Model (GSIM) case, a variable (e.g., diagnosis) may inherits its meaning from a Concept (e.g., a cancer diagnosis) with two representations (e.g., two different code lists for categorising different kinds of cancer diseases). These representations in turn also inherit their meaning from a concept and this also applies to the provided diagnosis codes. The (meta)data elements designation can then differ but the semantic interoperability can be evaluated by comparing the referenced concepts on each level of granularity, that is on the variable level, the representation level and the level of the code used in the code list.

In the ISO11179 standard a similar mechanism is constructed using the data element and the data element concept, where the relation between them provides a mechanism for semantic mapping. Analogous to the example above the data element concept (e.g., cancer diagnosis) may be related to two data elements (e.g., different cancer variables) using two different code lists for representation. Different designations and representations can, in a similar way as in the earlier example, be semantically compared by comparing the linked concepts on each level of granularity.

Links between domain-specific metadata frameworks and generic ones can be used as a way to establish commonalities for increased interoperability, and a common language for data semantics, acknowledging differences between domains. Examples of this are:

- Data documentation initiative[72], used within CESSDA[73], which maps to GSIM.
- HL7FHIR[74], used for exchanging health data, which maps to ISO11179.

69 Poveda-Villalón M., Espinoza-Arias P., Garijo D., Corcho O. (2020) Coming to Terms with FAIR Ontologies. In: Keet C.M., Dumontier M. (eds) Knowledge Engineering and Knowledge Management. EKAW 2020. Lecture Notes in Computer Science, vol 12387. Springer, Cham. https://doi.org/10.1007/978-3-030-61244-3_18

70 Generic Statistical Information model (GSIM): Specification, United Nations Economic Commission for Europe (UNECE)

71 ISO/IEC JTC1 SC32 WG2, ISO/IEC 11179

72 https://ddialliance.org/

73 https://www.cessda.eu/

74 https://www.hl7.org/fhir/elementdefinition.html

This mechanism is also enabled by the **RDA Data Type Registries Model**[75]. In the work continued by the RDA WG more detailed models are further describing this in the context of data type registries see "Documentation of Data Representations - A proposed scheme for documenting data structures and vocabularies for machine applications"[76]. This work maps well to ISO11179 and GSIM.

Domain specific and community driven metadata standards that are not mapping to a framework/conceptual metadata standard/data type registry model today can progress towards improved interoperability by mapping to one. Implementation of a semantic mapping mechanism and linking to common concepts will support progress towards higher levels of interoperability.

75 http://doi.org/10.15497/A5BCD108-ECC4-41BE-91A7-20112FF77458

76 https://github.com/usgin/usginspecs/raw/gh-pages/DataTypeModelDraft.pdf (last accessed: 31/Dec/2020)

5 TOWARDS THE EOSC IF: REFERENCE ARCHITECTURE

This section presents a Reference Architecture Framework in which the EOSC Interoperability Framework can be described and modelled. The target audience of the EOSC Interoperability Reference Architecture Framework is providers developing and making resources available through EOSC as well as users consuming EOSC resources and services as part of a solution. The EOSC Interoperability Reference Architecture Framework will also set guidelines for connecting resources to the EOSC Core.

The Reference Architecture contains framework definitions and uses abstract Building Blocks as a tool to group functionality that will be needed to meet the requirements for the EOSC Interoperability Framework (EOSC-IF). It does not provide the Solution Building Blocks (SBB)[77] themselves, since it is foreseen that there will be a range of different implementations from different domains delivering the needed functionality of the EOSC-IF.

The base of the EOSC-IF Reference Architecture has been derived from the European Interoperability Reference Architecture (EIRA)[78] developed by ISA2. The frameworks have somewhat different, but sometimes overlapping, target groups but share many common underlying principles and core objectives. Despite the differences, there is a value of aligning both frameworks and to develop the EOSC-IF as an extension of the EIRA for the research domain.

The EIRA is modelled in Archimate and thus the extension that the EOSC Interoperability task force created is an Archimate representation of the interoperability framework reference architecture. The Archimate file is available[79] but in order to present the interoperability framework to a broader audience, this section uses a less formal notation.

The section presents an overview of the different viewpoints for the frameworks as used throughout the EOSC IF document and in the European Interoperability Framework: Legal, Organisational, Semantic and Technical views. These views encompass the main components that need to be specified to align the different solutions needed to implement the EOSC Interoperability Framework.

- Definition of the Legal Interoperability building blocks rests on the work done by the Legal team of the Interoperability task force.
- Definition of the Organisational Interoperability building blocks rests on the work done by the EOSC Rules of Participation WG and the EOSC Sustainability WG.
- The architecture Building Blocks and the structure needed to enable the EOSC Interoperability Framework Semantic and Technical view have been the focus of the EOSC Interoperability taskforce.

The proposed EOSC Interoperability framework is to be considered as a Reference Architecture and common structures that are designed with extension and evolution, through community input, in mind.

77 "A candidate solution which conforms to the specification of an Architecture Building Block (ABB)", https://pubs.opengroup.org/architecture/togaf9-doc/arch/chap03.html.

78 https://joinup.ec.europa.eu/collection/european-interoperability-reference-architecture-eira/solution/eira

79 Eriksson, van de Sanden, Kurowski, Coppens, Corcho, & Ojsteršek. (2021, January 5). EOSC Interoperability Framework Reference Architecture (Version 1.0). Zenodo. http://doi.org/10.5281/zenodo.4420096

5.1 EOSC-IF high-level Architecture viewpoint

The high-level viewpoint model below aims to provide a general description of the EOSC-IF Reference Architecture and Building Blocks that enable interoperability for the four views/layers above and also acts as an extension of the EIRA.

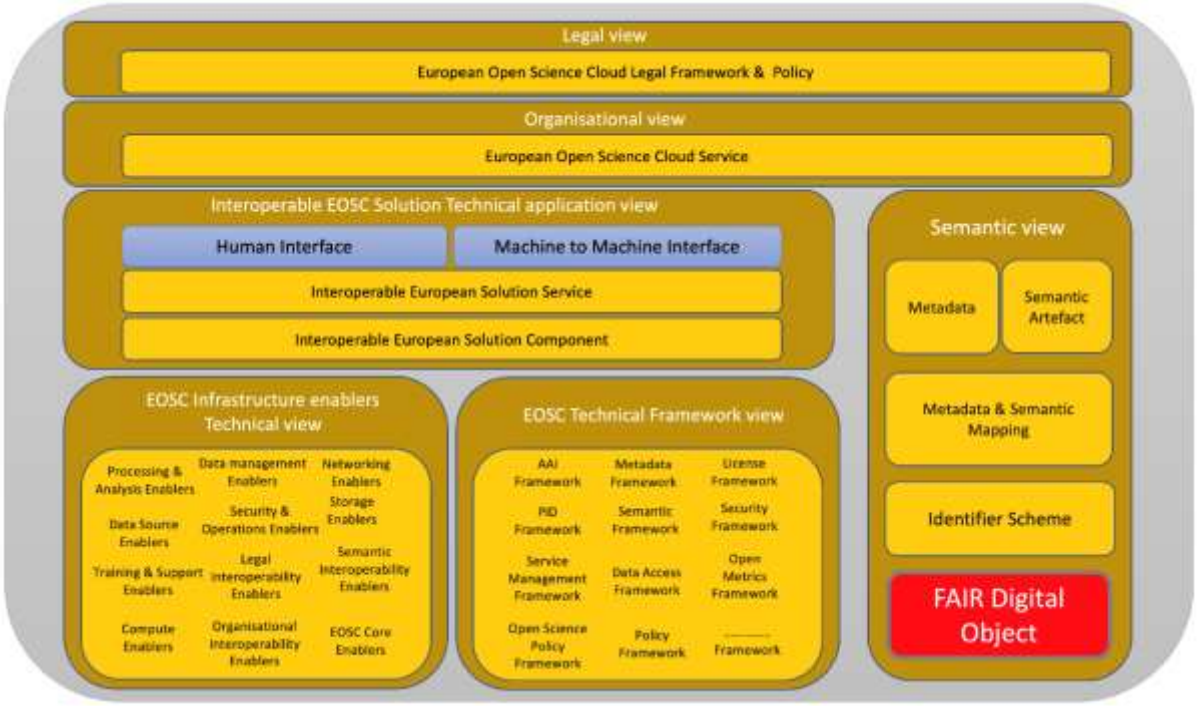


Figure 6. EOSC-IF High Level viewpoint.

The next sections go into detail describing a structure supporting the evolution of the EOSC interoperability framework, exemplified with a number of example frameworks and architectural building blocks for each interoperability view/layer in the subsequent sections.

5.1.1 Legal view

The different aspects of EOSC Policy that affect interoperability are shaped within a separate governance structure and a Shared Legal Framework.

At the time of writing, the EOSC Association has been established, although the details of the policies are still under development. This will need to be taken into account for a future update of this document and the implementation of the reference architecture as described here. AARC-BPA 2019 and the EOSC AAI

AARC-BPA 2019 distinguishes between two types of AAI services: One focuses on infrastructure management, while the other focuses on community management. Both types of AAI services may comprise the same interfaces (e.g., a proxy), but their functionality and their organisational purposes differ.

5.1.2 Organisational view

Both public and private service providers and consumers are actors within EOSC and can provide and/or consume services offered through EOSC. This affects the organisational view by including building blocks for providers offering resources and services through EOSC and to provision the interoperability needs between providers/consumers and across organisations in a common way. This will be based on work done by the EOSC Rules of Participation and the Sustainability WGs.

5.1.3 Semantic view

In the interviews and consultations done we found that Semantic Interoperability is one of the greatest interoperability challenges that the community meets.

FAIR digital objects are at the core of the EOSC interoperability frameworks Semantic View, as described in Section 4. Based on the FAIR Digital Object concept, the semantic view has been composed of architecture building blocks to describe and model the FAIR Digital Object and functional content on which it is depending.

5.1.3.1 Semantic Interoperability Concepts

The semantic view (Figure 7) represents the conceptual framework used to provide clear definitions, which enable interpretation of the digital objects' meaning, as required during exchange. In this high-level viewpoint these building blocks are summarised under the Semantic Artefact concept.

The EOSC Interoperability taskforce also wants to highlight the need to support different scientific domains at different levels of maturity. The concepts aimed to represent different types of Semantic artefacts at different levels of formality and machine readability embrace this.

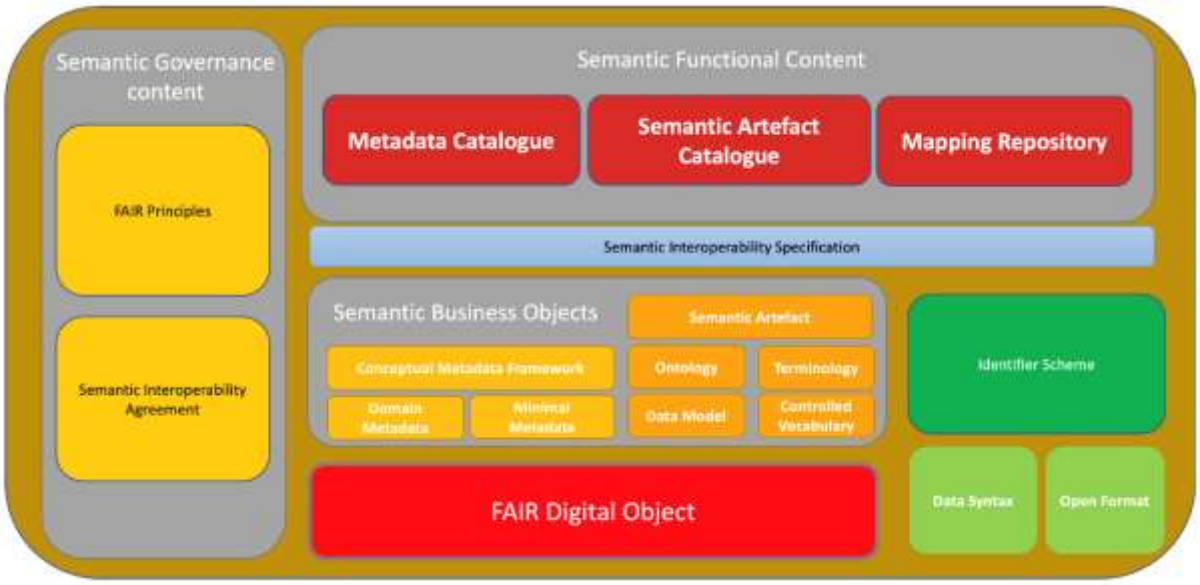


Figure 7. Semantic view.

5.1.3.2 Semantic Functional Content

The concepts described are utilised for provisioning the functional content needed to enable semantic interoperability. The functional content can be summarised as different types of knowledge bases/repositories provisioning metadata, semantic artefacts and crosswalks/mappings that enable translation between different metadata standards and semantic artefacts to enable an effective exchange.

Since the EIRA High level viewpoint section on Semantic Interoperability is quite small, and the semantic interoperability challenge within EOSC is large, this view is extended with several Architectural building blocks.

This also reflects the need to go beyond data and open up for other types of digital objects to be exchanged. The Architectural building blocks described in the Semantic view should from an EOSC perspective be interpreted as conceptual and can be delivered by one or many Solution building blocks and by one or several services in a federated way.

5.1.4 Technical view

The starting point for the EOSC interoperability framework emanates from the work done by the EOSC Sustainability WG that is presented in the FAIR Lady report[80]. This report presents the EOSC Core as central to implement a Minimal Viable EOSC and that this rests on ”an instantiation of the EOSC Interoperability Framework…”.

The EOSC Interoperability Framework is composed of several frameworks that encompass a broad portfolio of interoperability alignment and support functions. The report highlights a subset of these as important for the realisation of the EOSC Core and a Minimal Viable EOSC:

- Authentication and Authorization Interoperability (AAI) framework
- PID Framework
- Metadata framework
- Data access framework
- Service management and access framework
- Open metrics framework
- Security framework
- Support framework

The frameworks presented is a core set that the Sustainability WG views as essential to instantiate in order to support the implementation of EOSC Core.

Since these frameworks will continue to evolve with community input and the EOSC Interoperability Framework also will continue to evolve, encompassing new frameworks the EOSC Interoperability taskforce sees the need to propose a reference Architecture and Interoperability Framework that is designed with extension in mind to evolve with community input.

The different components within the EOSC Interoperability Framework also need to support the implementation of EOSC Digital Infrastructure and other EOSC Solutions in a flexible way.

80 Solutions for a sustainable EOSC (2020). DOI: 10.2777/870770

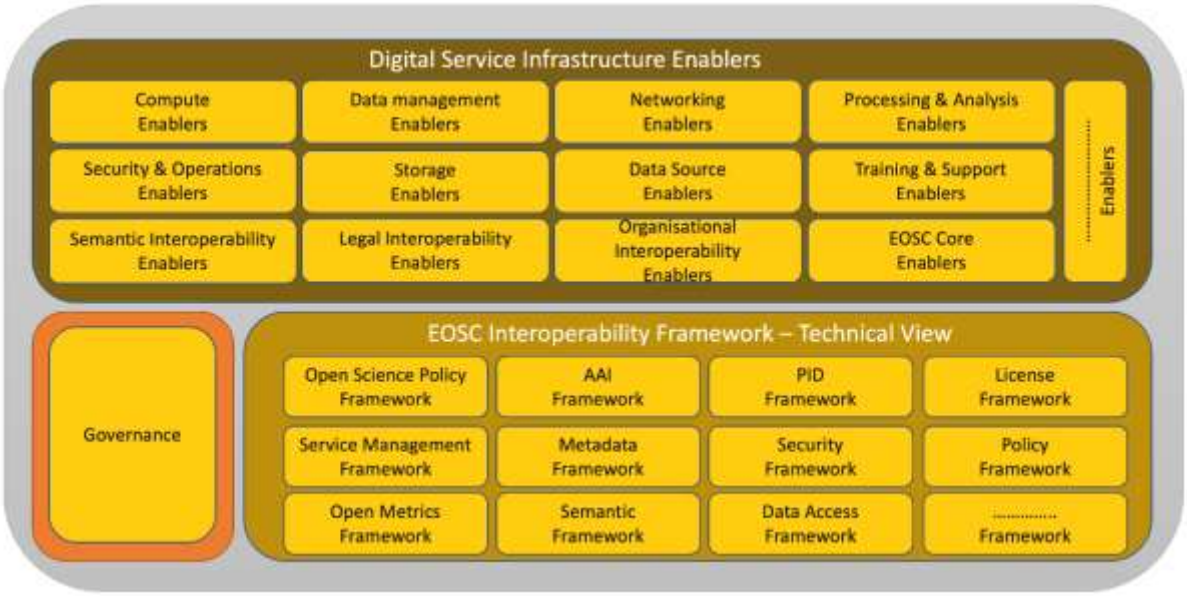


Figure 8. Technical view.

In order to meet the Legal, Organisational and Semantic interoperability requirements enabling digital service infrastructure that implements the EOSC Interoperability Framework, and the components supporting these, are of course essential.

The common structure for this is described in the reference Architecture where the components within the EOSC Interoperability Framework are instantiated when developing digital service infrastructure enablers that provision services to meet community needs.

5.2 EOSC-IF Reference Architecture – View details

This section presents more detailed views of the EOSC Interoperability Framework reference architecture and the different views it consists of.

5.2.1 EOSC-IF High-level Semantic view

The EOSC-IF Semantic view extends the EIRA Semantic view while also removing some of the EIRA building blocks that are of less focus for the EOSC-IF.

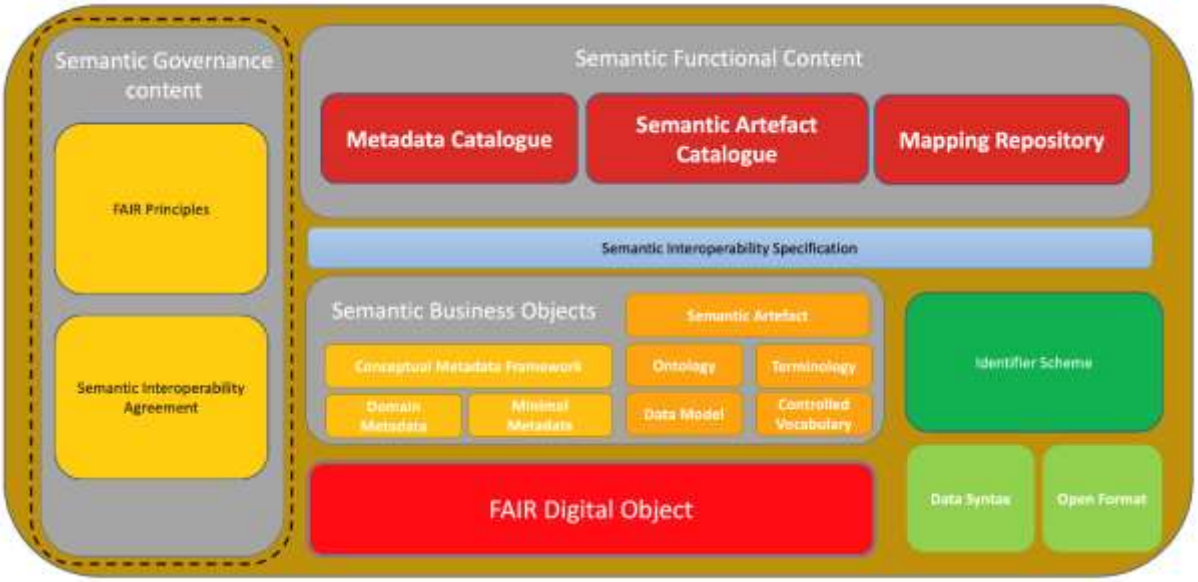


Figure 9. EOSC-IF Semantic view - Governance.

5.2.1.1 Semantic governance content

The Semantic governance content, visualised in figure 9 above, rests on the FAIR principles. They are implemented using the Semantic Interoperability Agreement.

The Organisational view connects to the Semantic view by an association from the EOSC Service provider/consumer to the Semantic Interoperability Agreement building block where the rules governing semantic interoperability requirements are agreed.

5.2.1.2 Semantic Business objects

5.2.1.2.1 Semantic artefact

The EOSC-IF recognises that different domains use more or less formal ways to provide clear definitions of their digital objects. In order to support different domain starting points, and progress from less formal ways of providing clear definitions to more formal and machine-readable ways, we extend the EIRA with the concept of semantic artefacts in the business object section (Figure 10).

The EOSC service providers/consumers agreement around the Semantic Interoperability Specification can thus differ between domains when it comes to how formal and machine-readable semantic artefacts will be used during exchange of the FAIR Digital objects.

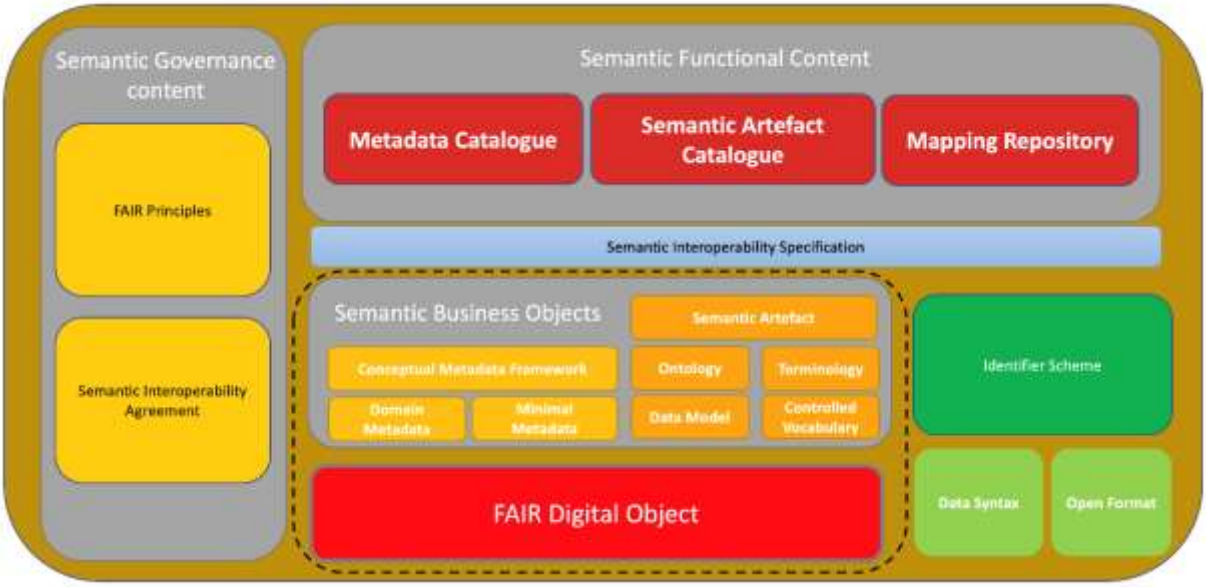
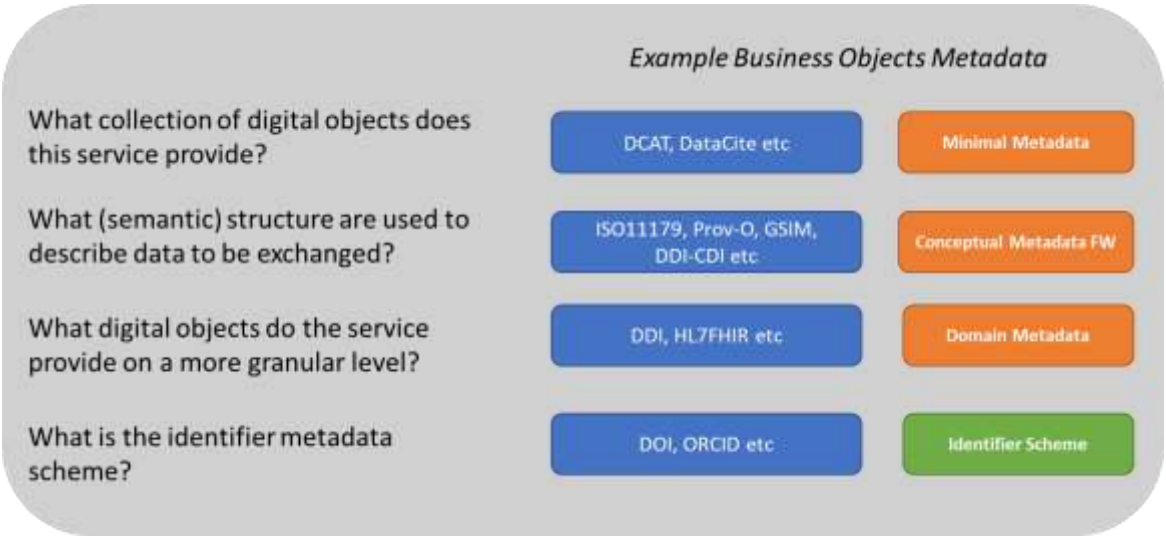


Figure 10. EOSC-IF Semantic view - Objects.

5.2.1.2.2 Metadata

Three types of metadata are also introduced in the business object section, visualised in the figure above (Figure 10).

The Conceptual Metadata framework - that provides a common terminology of how data is described and how information objects are put together. Examples of these are ISO11179, the RDA Data Type Registry, the UNECE Generic Statistical Information Model and the W3C Prov standard.



Minimal Metadata – High granularity metadata describing a small selection of properties that describe resources. Examples of these are Dublin Core, DCAT-AP, Datacite, etc (see Appendix I for a review of crosswalks among them).

The Domain Metadata standard – that specifies **what** metadata attributes should be provided. By mapping to a conceptual metadata framework, the Domain Metadata increases the possibility to find commonalities between digital objects, increasing interoperability.

5.2.1.2.3 Semantic functional content

In the Semantic functional content section, visualised in figure 11 below, we describe Architectural building blocks that are needed in order to realise the business objects supporting semantic interoperability and providing the needed functionality. The functionality will be provided by many different services delivered within the EOSC federation of services.

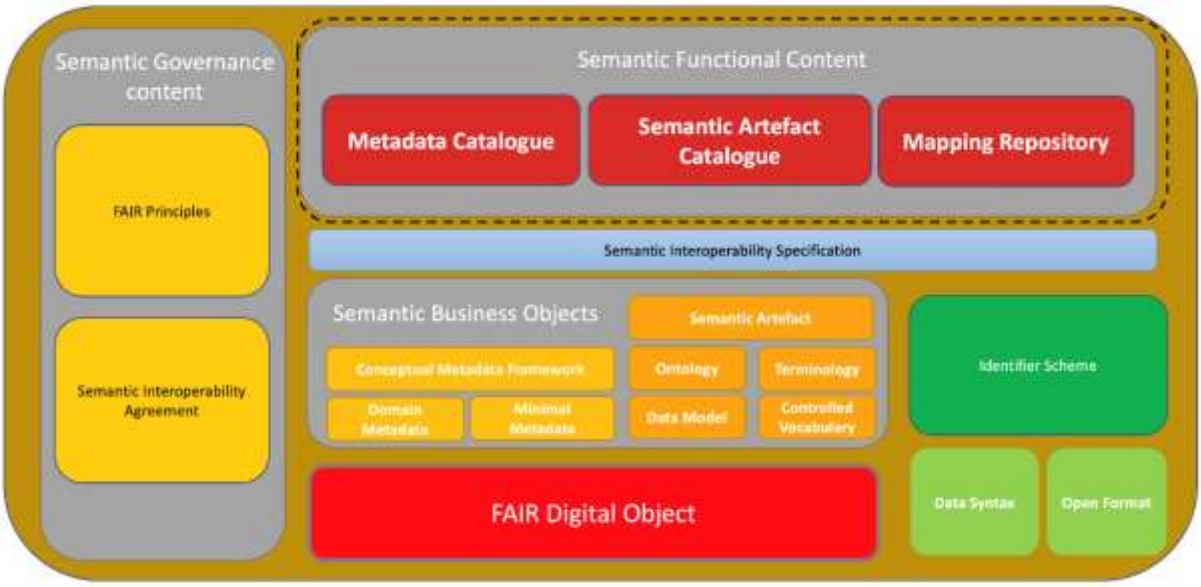


Figure 11. EOSC-IF Semantic view - Functional Content.

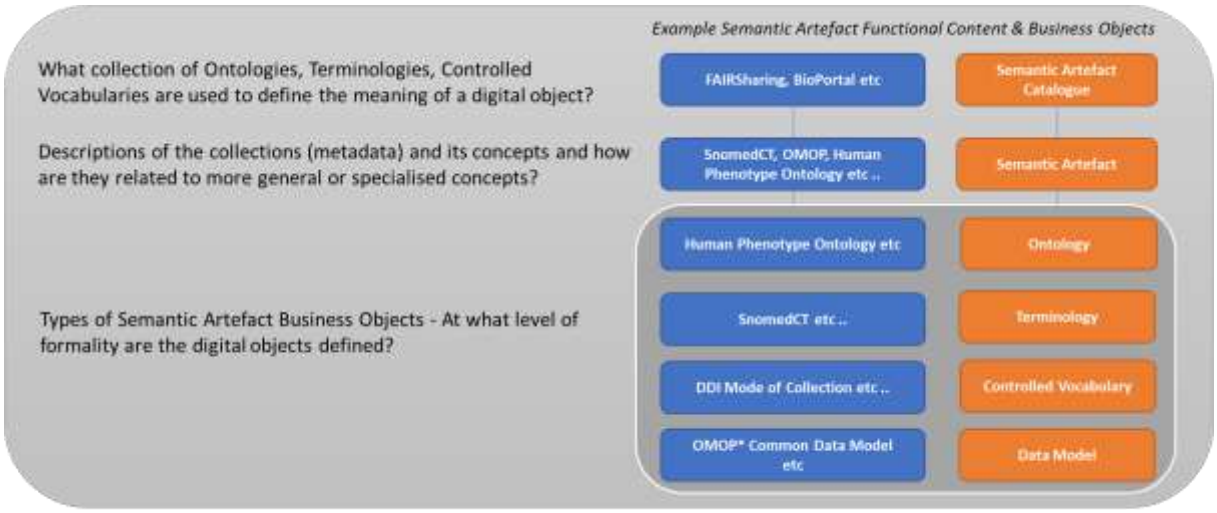
FAIR Digital Object – This is a central building block for increased interoperability within EOSC. This extension to the EIRA, aggregates Digital object, Metadata, Identifier Scheme and Open format building blocks and takes meaning from the Concept represented within a Semantic Artefact.

Metadata catalogue

Metadata and Metadata catalogues are core building blocks when supporting interoperability and implements several of the Business objects that are included in the Semantic Interoperability Specification.

Semantic Artefact Catalogue

This also applies to the Semantic artefact building block implementing the different Semantic artefact Business objects that provides support to different domains using different ways to provide clear definitions of concepts defining their digital objects. Domains can thus progress from less formal to more formal machine-readable semantic artifacts and provide these within EOSC as agreed upon by the service providers and consumers in the Semantic interoperability specification.



Mapping repository

An architectural building block that represents different services providing different types of mappings between FAIR digital objects in order to enhance the interoperability. Examples of useful mappings can be between FAIR digital objects providing (meta)data elements and those providing meaning through concepts within semantic artefacts.

Crosswalk repository - One special type of building block that enhances interoperability is the crosswalk repository providing relations between attributes in different metadata standards to be used then translating from one standard to another when exchanging metadata and/or data.

5.2.2 EOSC-IF High-level technical view

The EOSC-IF Technical view extends the EIRA Application and Infrastructure view but since there are some differences between the public and research data domain and choice of architectural support the views presented will differ.

The EOSC Interoperability framework builds on a subset of frameworks aligning and supporting digital infrastructure solutions that act as enablers within EOSC and EOSC core capabilities. The different components within the frameworks also align and support development of other services and solutions. With this architectural choice the EOSC reference architecture differs from the EIRA that has not adopted the framework approach in the reference architecture model.

5.2.2.1 Starting point

The FAIR Lady report highlights a set of frameworks that is viewed as important for the realisation of the EOSC Core and a Minimal Viable EOSC (MVE). The FAIR Lady report has been developed by the Sustainability WG exploring possible means for sustaining the EOSC beyond its initial phase.

Together with the EOSC Architecture WG vision for a Minimal Viable EOSC this provides the starting point for the technical view of the EOSC Interoperability Framework.

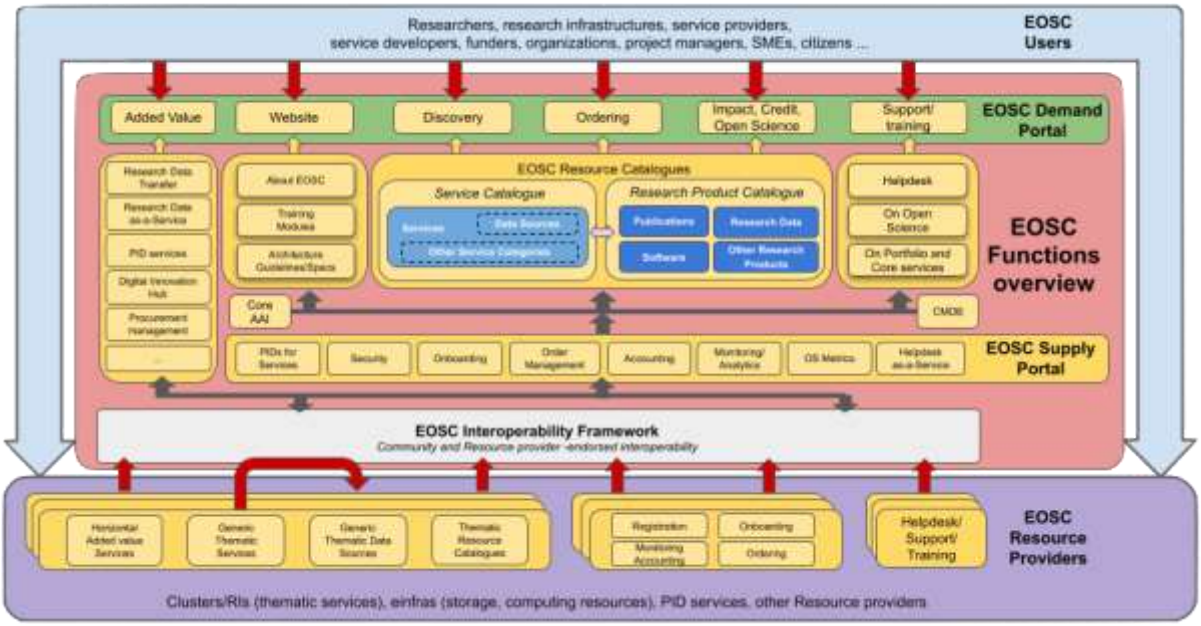


Figure 12. Architecture WG - Envisioned EOSC MVE.

5.2.2.2 Frameworks

The identified frameworks are aimed to support interoperability across EOSC, to lower the barrier for communities to build solutions for their researchers based on services and resources offered by RI and e-infrastructure within EOSC. These frameworks provide guidelines for service engineers to develop interoperable infrastructure components and to connect these services to the EOSC Core.

Each framework included within the EOSC Interoperability Framework is composed of a set of components targeting a specific topic or element within the framework. The EOSC Interoperability Framework and the individual frameworks included must be flexible in nature to meet the evolving needs of EOSC in a way that gives the highest benefit to the research community.

In the figure below an example view of a top-level organisation of the EOSC Interoperability Framework is provided including some of the frameworks identified in the FAIR Lady report. The schematic diagram also provides an indicative view of components comprising the frameworks. The arrow on the right and the components with dashed lines indicate possible extensions in frameworks to be included within the EOSC Interoperability Framework and components within the frameworks.

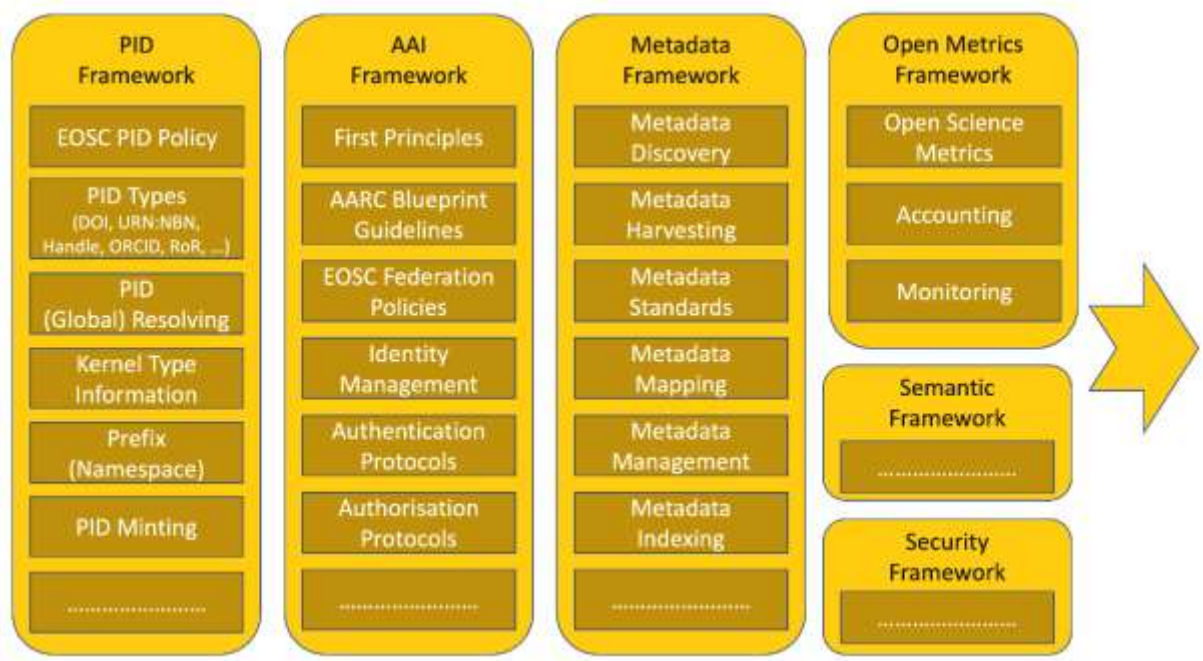


Figure 13. Framework structure.

5.2.2.3 Infrastructure Perspective

To support service developers developing infrastructure components an infrastructure perfective is also provided. The infrastructure view comprises Infrastructure enablers which are organised in enabler categories such as Compute Enablers, Semantic Interoperability Enablers, Data Management Enablers etc. The enabler categories include different infrastructure building blocks. For example, within the data management enablers category, building blocks such as digital repository, data discovery, and data archive are included, while within the compute category building blocks for HTC, HPC, Cloud, and Containers are included.

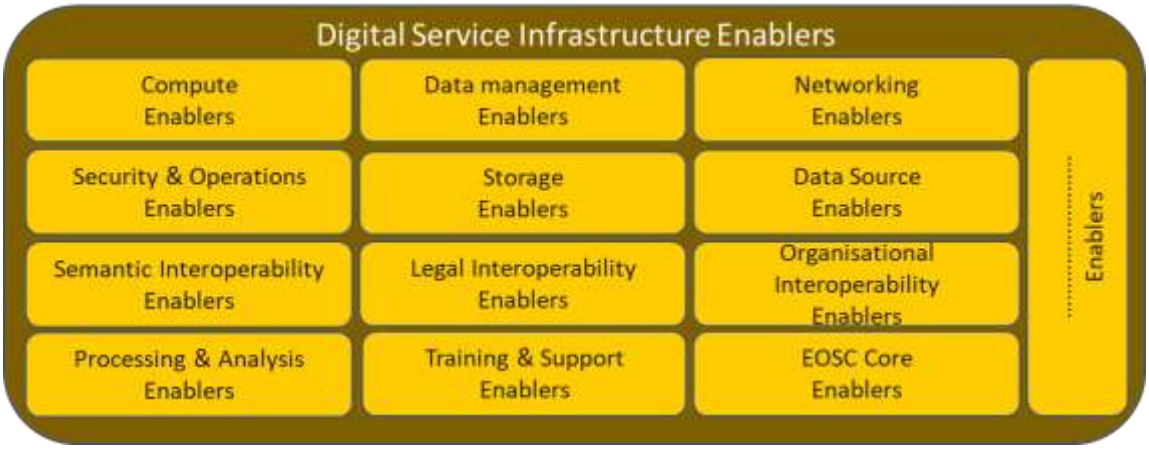


Figure 14. Enablers.

The abstract building blocks are constructed from elements defined within the EOSC Interoperability Framework. In the figure below an example building block of a digital repository is provided. As described in the diagram, a digital repository consists of many elements. For the implementation of the building block elements guidelines need to be followed related to different frameworks components defined within the Framework view, as shown in the diagram below. The abstract building blocks descriptions guides service developers in the development of new services and in identifying important components within services.

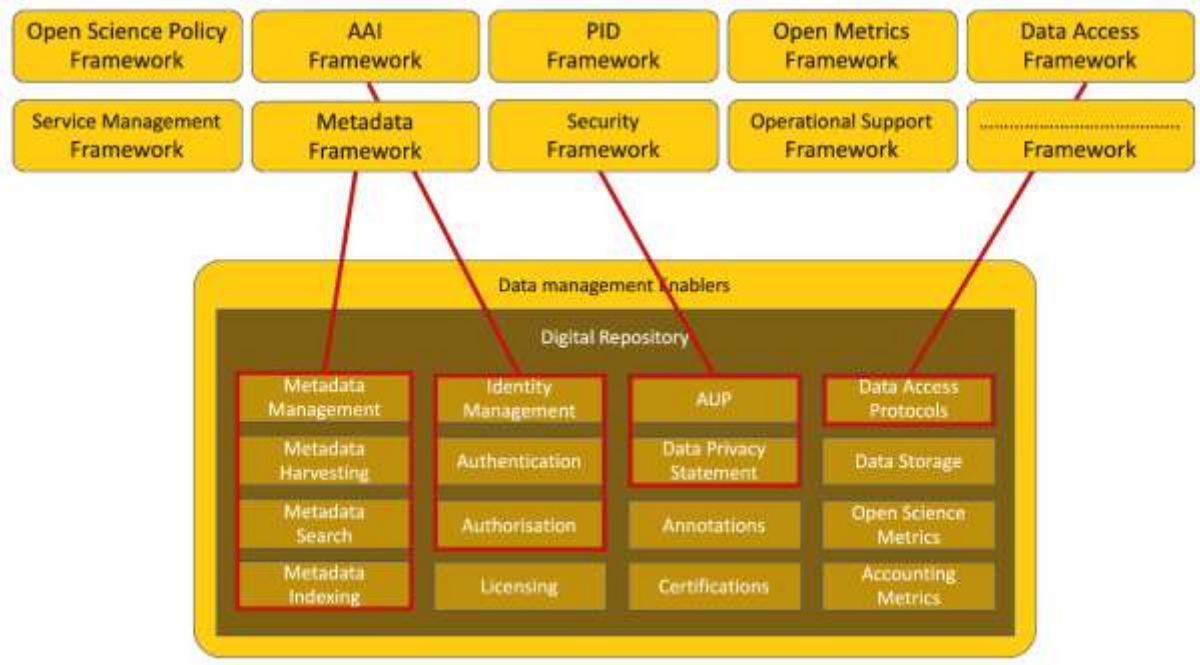


Figure 15. EOSC Infrastructure building blocks.

The strength of the Reference Architecture is in the reuse of interoperability components defined within the framework view in the infrastructure view in defining the infrastructure building blocks.

This architectural structure further enables extension and supports evolution on the infrastructure level. Extension of the capabilities delivered by infrastructure will then be aligned and supported by the EOSC Interoperability Framework that, as described above, is designed with extension and reuse in mind.

5.3 Recommendations and next steps

The EOSC Interoperability task force recommends continuing the work on the EOSC Interoperability Framework with the areas listed below:

- **Detailed specification of Architectural building blocks**. The architectural building blocks that compose the EOSC Interoperability Framework need to be further detailed. This should be done hand in hand with the communities, many of which have already their interoperability practices in place.
    - The EOSC Interoperability taskforce is not a group that has resources to deliver this output, so it is handed over to the EOSC Association.
    - Extensions will be concerning all types of building blocks, including frameworks, framework components and infrastructure enablers
- **Establishing governance structure and maintenance of the framework**. Since the EOSC Interoperability Framework is designed with extension and evolution in mind it is of utmost importance to establish a governance structure and maintenance organisation to guide, organise and keep the work together. This is especially important when implementing the core framework that will set the foundation for the future.

APPENDIX I. ANALYSIS OF MINIMAL METADATA MODELS AND CROSSWALKS AMONG THEM

Repositories and data archives are an incredible treasure trove of open knowledge. They store scientific production that researchers need in their research. Most of these resources are openly available. Others are accessible in restricted or closed access. The OpenDOAR database contains 5508 repositories. Re3data has 2994 registered repositories and data archives. Information about 15641 scientific journals is available in DOAJ. Europeana provides access to more than 50 million digital objects located in more than 3,500 digital libraries.

Metadata standards, profiles and schemes

Metadata records of digital objects in repositories and data archives are in various metadata standards, profiles and schemes[81,82,83,84]. According to re3data.org[85] the most common used metadata formats are Dublin Core[86], Datacite metadata schema[87], DDI - Data Documentation Initiative[88], ISO 19115[89], FGDC / CSDGM[90] , and DIF - Directory Interchange Format[91]. Metadata in public administration data archives are mostly in the DCAT - Data Catalogue Vocabulary[92]. MODS[93] and MARC[94] metadata schemas are still widely used in the world of digital libraries. Data archives use many domain-specific metadata standards (CF - (Climate and Forecast) Metadata Conventions[95], Bioschemas[96], DarwinCore[97], CIF[98], ABCD[99], FITS[100], IVOA[101], SPASE[102], CIDOC-CRM[103],CMDI[104] …).

Some repositories have defined customised metadata schemas. They transform their metadata elements into the metadata application profile required by the search engines (Google, Google Scholar, Google Dataset Search, Microsoft Academics), metadata aggregators (OpenAire, B2Find, Europeana, DART Europe, national aggregators (e.g., Narcis, RCAAP, Recolecta, Openscience.si), bibliographic cataloguing systems, commercial discovery systems (EBSCO, ProQuest Exlibris Primo, Proquest Exlibris Summon, OCLC WorldCat Discovery…) or DOI providers (Datacite, Crossref)). Metadata aggregators have

81 Digital Curation Centre's metadata directory. Available on https://www.dcc.ac.uk/guidance/standards/metadata [19.12.2020]

82 RDA Metadata Standard Catalogue. Available on https://rdamsc.bath.ac.uk/ [19.12.2020]

83 Library of Congress Standards. Available on https://www.loc.gov/librarians/standards [19.12.2020]

84 Jenn Rilley and Devin Becker. Seeing Standards: A Visualization of the Metadata Universe. Available on http://jennriley.com/metadatamap/ [19.12.2020]

85 Re3Data usage of metadata standards. Available on https://www.re3data.org/metrics/metadataStandards [19.12.2020]

86 DCMI Metadata Terms. Available on https://dublincore.org/specifications/dublin-core/dcmi-terms/ [19.12.2020]

87 DataCite Metadata Schema 4.3. Available on https://schema.datacite.org/meta/kernel-4.3/ [19.12.2020]

88 DDI Products. https://ddialliance.org/products/overview-of-current-products [19.12.2020]

89 ISO 19115:2014. https://www.iso.org/standard/53798.html [19.12.2020]

90 FGDC/CSDGM - Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata.c Available on https://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf [19.12.2020]

91 DIF - Directory Interchange Format. Available on https://earthdata.nasa.gov/esdis/eso/standards-and-references/directory-interchange-format-dif-standard [19.12.2020]

92 DCAT - Data Catalogue Vocabulary. Available on https://www.w3.org/TR/vocab-dcat/ [19.12.2020]

93 MODS – Metadata object description schema. Available on http://www.loc.gov/standards/mods/ [19.12.2020]

94 MARC 21 Format for Bibliographic Data. Available on https://www.loc.gov/marc/bibliographic/ [19.12.2020]

95 CF - (Climate and Forecast) Metadata Conventions. Available on https://cfconventions.org/ [19.12.2020]

96 Bioschemas. Available on https://bioschemas.org/ [19.12.2020]

97 DarwinCore. Available on https://dwc.tdwg.org/ [19.12.2020]

98 CIF specifications. Available on https://www.iucr.org/resources/cif/spec [19.12.2020]

99 Access to Biological Collection Data (ABCD). Available on https://github.com/tdwg/abcd [19.12.2020]

100 FITS - Flexible Image Transport System. Available on https://fits.gsfc.nasa.gov/fits_standard.html [19.12.2020]

101 International Virtual Observatory Alliance. Documents and standards. Available on https://www.ivoa.net/documents/index.html [19.12.2020]

102 SPASE ontology 2.3.2. Available on https://spase-group.org/data/schema/ [19.12.2020]

103 CIDOC-CRM – Conceptual reference model. Available on http://www.cidoc-crm.org/Version/version-7.0.1 [28.12.2020]

104 CMDI – Component Metadata Infrastructure. Available on https://www.clarin.eu/content/component-metadata [28.12.2020]

developed guidelines for metadata records that they can aggregate. Most of the guidelines use a Dublin Core or Datacite schema with some additional features[105,106,107]. Crossref[108] has its metadata schema for scientific publications. For the aggregation of cultural heritage digital objects, Europeana EDM metadata schema[109] and OAI-ORE metadata schema[110] are used. Schema.org[111] is gaining ground as a de facto standard on the discoverability of digital objects in search engines.

Shortcomings of metadata records in different repositories and data archives

The CORE collection[112] collects metadata and scholarly texts. In 2018 it contained 123 million metadata records, 85.6 million records with abstract, and 9.8 million records with full text. Metadata records and full-text are available in different languages. When we reviewed the metadata records from this dataset, we found that the metadata are of deficient quality. The problem is that repository and data archives platforms are still using technologies and protocols designed almost twenty years ago. Repositories and data archives use different software platforms in different versions[113,114]. They are not well prepared for search engines, social networks, semantic web, and "machine and human-understandable" FAIR digital objects.

After reviewing the metadata records in the CORE dataset, we found the following shortcomings:

- When system administrators upgrade repository software, its metadata application profile usually changes as well. Direct mapping of metadata elements between old and new metadata profiles are not established in many older repository software platforms. Metadata records that use an older version of the application profile are incompatible with metadata records that use a newer version of the metadata application profile.

- Metadata records resulting from the conversion of repository metadata records to metadata records required by metadata aggregators and search engines are also low quality. Cause for the poor quality of metadata records is in the processes for managing them. Researchers, data stewards, and librarians typically insert the minimum required set of metadata elements needed by repository software.

- Files that are part of digital objects stored in repositories are in various file formats. Some files are available in formats that can only be read by commercially available software. Some files cannot be opened in newer software versions.

- Many metadata elements in metadata schemas should have data types defined. Due to the evolution of repository software platforms, metadata schemas allow the entering of character strings for most metadata elements. Inserters of metadata records insert character strings in different languages without specifying the language in these metadata elements. The main problems are when researchers, librarians, or data stewards insert character strings instead of dates, numbers, subject codes (keywords,

105 Google. Inclusion guidelines for webmasters. Available on https://scholar.google.com/intl/en/scholar/inclusion.html#indexing [19.12.2020]

106 OpenAire guidelines. Available on https://guidelines.openaire.eu/en/latest/ [19.12.2020]

107 B2Find guidelines. Available on http://b2find.eudat.eu/guidelines/mapping.html [19.12.2020]

108 CROSSREF Schema. Available on https://data.crossref.org/reports/help/schema_doc/4.4.2/index.html [19.12.2020]

109 EDM Mapping Guidelines. Available on https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines_v2.4_102017.pdf [19.12.2020]

110 OAI ORE Specifications and user guides. Available on http://www.openarchives.org/ore/1.0/toc [19.12.2020]

111 Schema.org. Available on https://schema.org/ [25.12.2020]

112 CORE dataset. Available on https://core.ac.uk/documentation/dataset/ [28.12.2020]

113 OpenDOAR Software platforms overview. Available on https://v2.sherpa.ac.uk/view/repository_visualisations/1.html [26.12.2020]

114 Re3data Software statistics. Available on https://www.re3data.org/metrics/software [26.12.2020]

resource types, subject headings, temporal and spatial coverage), citations, and relations between digital objects.

- In many metadata schemas, very few metadata elements use controlled vocabularies (e.g., DDI vocabularies[115], COAR vocabularies[116], Datacite vocabularies[117], Bioportal ontologies, and controlled vocabularies[118]…). Some repositories have their vocabularies that cannot be easily mapped to established controlled vocabularies.

- Controlled vocabularies and other semantic artefacts used by metadata application profiles have changed over time. Repository software platforms do not have supported vocabulary versioning and mapping between different versions of the same vocabulary.

- Repository platforms do not use authority control for persons, places, corporate names, projects, and research groups.

- In many cases, non-functioning web and persistent links are found in metadata records.

- Most metadata records also do not have defined access rules, digital objects' provenance, and licenses for their use. Consequently, these digital objects are challenging to reuse.

- The quality of metadata is also affected by different versions of aggregators' application metadata profiles. Aggregators' older versions of metadata application profiles were adapted to older versions of the application metadata profiles used by repository software platforms. In most cases, they allow uncontrolled inputs of metadata elements.

- Another problem of aggregators is the duplication of digital objects. The reason for a high number of duplicated digital objects is again in the missing or the uncontrolled inputs of metadata elements. In most duplicate cases of digital objects, the publisher's version of the metadata records differs from the metadata records published in the repository.

- Metadata records of digital objects stored in repositories do not have defined persistent identifiers or have a changed order of authors, or do not list all authors. Older versions of repository software platforms do not have defined fields for metadata elements that are important for citation (e.g., ISSN, journal name, volume number, issue number, start and end pages, proceedings name, ISBN etc…). Often digital objects from repositories do not have a defined year of publication and resource type of digital object or other subject headings. Due to publishers' requirements, repositories often contain different content files (e.g., preprint, accepted manuscript, postprint, author's version of the article). The same situation is also in the case of published research data or software.

Metadata records of digital objects in repositories can be improved by supplementing them with the missing metadata elements in metadata records. To improve metadata elements in repository metadata records are possible to use the CROSSREF[119] and Datacite API[120] (if the digital objects have a defined DOI or other persistent identifiers) or OpenAIRE ResearchGraph[121] or Microsoft Academic Graph[122] APIs.

115 DDI controlled vocabularies. Available on https://ddialliance.org/controlled-vocabularies [28.12.2020]

116 COAR vocabularies. Available on https://www.coar-repositories.org/news-updates/what-we-do/controlled-vocabularies/ [28.12.2020]

117 Datacite Metadata schema 4.3. Available on https://schema.datacite.org/ [28.12.2020]

118 Bioportal. Available on https://bioportal.bioontology.org/ [31.12.2020]

119 CROSSREF API. Available on https://github.com/CrossRef/rest-api-doc/blob/master/api_format.md [3.1.2021]

120 Datacite API. Available on https://support.datacite.org/docs/api [3.1.2021]

121 OpenAIRE ResearchGraph. Available on https://graph.openaire.eu/ [3.1.2021]

122 Microsoft Academic Graph. Available on https://docs.microsoft.com/en-us/academic-services/graph/ [3.1.2021]

Crosswalks among the most commonly used metadata schemes and aggregators' guidelines

Due to the shortcomings presented above, this section presents crosswalks among the most commonly used metadata schemes and aggregators' guidelines to describe digital objects in open science. In comparison are included[123]:

- RDA Metadata Interest Group recommendation of the metadata element set[124],
- EOSC Pilot - EDMI metadata set[125],
- Dublin CORE Metadata Terms[126],
- Datacite 4.3 metadata schema[127],
- DCAT 2.0 metadata schema and DCAT 2.0 application profile[128],
- EUDAT B2Find metadata recommendation[129],
- OpenAIRE Guidelines for Data Archives[130],
- OpenAire Guidelines for literature repositories 4.0[131],
- OpenAIRE Guidelines for Other Research Products[132],
- OpenAIRE Guidelines for Software Repository Managers[133],
- OpenAIRE Guidelines for CRIS Managers[134],
- Crossref 4.4.2 metadata XML schema[135],
- Harvard Dataverse metadata schema[136],

123 Ojsteršek. (2021). Crosswalk of most used metadata schemes and guidelines for metadata interoperability (Version 1.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.4420116 [5.1.2021]

124 RDA metadata IG recommendation of the metadata element set. Available on https://www.rd-alliance.org/groups/metadata-ig.html [3.1.2021]

125 EOSC Pilot - EDMI metadata set. Available on https://docs.google.com/spreadsheets/d/1dtHpbp5cVaooVdqhvDjLHKM5Y8IfC-iRSU6OA6BLSUg/edit#gid=1110916251 [3.1.2021]

126 Dublin CORE Metadata Terms. Available on https://dublincore.org/specifications/dublin-core/dcmi-terms/ [3.1.2021]

127 Datacite 4.3 metadata schema. Available on https://schema.datacite.org/meta/kernel-4.3/ [3.1.2021]

128 DCAT 2.0 metadata schema and DCAT 2.0 application profile. Available on https://www.w3.org/TR/vocab-dcat-2/ and https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe [3.1.2021]

129 Mapping onto EUDAT-B2FIND Metadata Schema. Available on http://b2find.eudat.eu/guidelines/mapping.html [3.1.2021]

130 OpenAIRE Guidelines for Data Archives. Available on https://guidelines.openaire.eu/en/latest/data/index.html [3.1.2021]

131 OpenAire Guidelines for literature repositories 4.0. Available on https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/v4.0.0/index.html [3.1.2021]

132 OpenAIRE Guidelines for Other Research Products. Available on https://guidelines-other-products.readthedocs.io/en/latest/ [3.1.2021]

133 OpenAIRE Guidelines for Software Repository Managers. Available on https://software-guidelines.readthedocs.io/en/latest/ [3.1.2021]

134 OpenAIRE Guidelines for CRIS Managers. Available on https://openaire-guidelines-for-cris-managers.readthedocs.io/en/latest/ [3.1.2021]

135 Crossref 4.4.2 XML metadata schema. Available on https://data.crossref.org/reports/help/schema_doc/4.4.2/schema_4_4_2.html [3.1.2021]

136 Dataverse Metadata Crosswalk. Available on https://goo.gl/yN2f9V [3.1.2021]

- DDI Codebook 2.5 metadata XML schema[137],.
- Europeana EDM metadata schema[138],
- Schema.org[139],
- Bioschemas[140] and
- The PROV Ontology[141].

We also present the most commonly used controlled vocabularies in the crosswalk. We describe Datacite, Crossref, OpenAIRE, and other vocabularies (from COAR[142], MARC, and Dublin Core).

Controlled vocabularies should[143]:

- be published under an open license.
- Be operated and/or maintained by a recognised standards organisation or another trusted organisation.
- Be properly documented.
- Have labels in multiple languages, ideally in all official languages of the European Union.
- Contain a relatively small number of terms that are general enough to enable a wide range of resources to be classified.
- Have terms that are identified by URIs, with each URI resolving to documentation about the term.
- Have associated persistence and versioning policies.

After examining mappings between metadata schemas, it is possible to find:

- that metadata elements in different schemas are in different granularity levels.
- Some metadata elements from the same group use for values character strings or values from vocabularies or metadata objects (e.g., persons, organisations, geolocation objects, temporal coverage objects). The problem is how to harmonise metadata schemes to the level that we do not lose semantics when we map one metadata element to a metadata element in a target scheme. If all schemes will use recommended data types for specific metadata elements and if it is possible to align used vocabularies in

137 DDI Codebook 2.5 XML metadata schema. Available on https://ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/field_level_documentation.html [3.1.2021]

138 EDM Mapping Guidelines. Available on https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines_v2.4_102017.pdf [3.1.2021]

139 Schema.org and Crosswalks from schemas to schema.org. Available on https://schema.org/docs/schemas.html and https://docs.google.com/spreadsheets/d/1P6WH8h4OnIVR9UJj3FcOebNUpLnKNBCuvEp3NsLRho4/edit#gid=1789151191 [3.1.2021]

140 Bioschemas and Schema.org Dataset Mapping. Available on https://bioschemas.org/ and https://docs.google.com/spreadsheets/d/16HNJVKUdueVIPEdcp3x2HXI0RJ4zrlpQWrTIkAf-lB4/edit#gid=0 [3.1.2021]

141 The PROV Ontology. Available on https://www.w3.org/TR/prov-o/ [3.1.2021].

142 COAR vocabularies. Available on https://www.coar-repositories.org/news-updates/what-we-do/controlled-vocabularies/ [3.1.2021].

143 DCAT Application profile for data portals in Europe 2.0.1. Available on https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe [3.1.2021].

these metadata elements, it is possible to achieve interoperability of metadata records from different metadata schemes.

Recommendation of minimum metadata set to describe metadata records for FAIR digital objects

It is possible to use various metadata schemes to describe metadata records for FAIR digital objects. EOSC does not want to invent a new metadata scheme. It is possible to define minimal metadata requirements for FAIR digital objects using existing metadata schemes and aggregators' guidelines.

The following requirement levels for the metadata properties are used:

- **Mandatory (M)**: The property must always be present in the metadata. An empty value for the property is not allowed.
- **Mandatory if Applicable (MA)**: When the property value can be obtained it must be present in the metadata.
- **Recommended (R)**: The use of the property is recommended.
- **Optional (O)**: It is not important whether the property is used or not, but it may provide complementary information about the resource if used.

Mandatory fields use for all metadata elements, which are essential for reusability, accessibility, and digital object citation.

We will use for the occurrence (cardinality or quantity constraint) of metadata elements following notation:

- 0-n = optional and repeatable,
- 0-1 = optional, but not repeatable,
- 1-n = required and repeatable,
- 1 = required, but not repeatable.

We recommend the following metadata elements for the description of FAIR digital object:

- **Identifier (M) (1)**: The Identifier is a unique string that identifies a resource. Example formal identification systems include the Uniform Resource Identifier (URI), the Uniform Resource Locator (URL), the Digital Object Identifier (DOI), and the URN. This can also be a direct URL, or a persistent identifier, like PURL, HANDLE, ARXIV, ARK, IGSN, PMID, LSID, or other international resolution mechanisms. If it is possible, please use one of persistent identifiers.
- **Creator (MA) (0-n)**: The authors of the digital object in priority order. It may be a corporate/institutional or personal name. Use authority control databases for persons or institutions if available and persistent identifiers (e.g., ORCID, ISNI, VIAFID) or national authority database of personal and corporate names.
- **Title (M) (1-n)**: A name or title by which a digital object is known. It is possible to have different titles such as title, subtitle, abbreviated title, or translated title. Titles are multilingual.
- **Publisher (M) (1)**: The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource. This property will be used to formulate the citation, so consider the prominence of the role. The digital object must

have one publisher. Use authority control databases for persons or institutions if available and persistent identifiers (e.g., ORCID, ISNI, VIAFID) or national authority database of personal and corporate names.

- **Publication Year (M) (1)**: The year when the digital object was or will be made publicly available. In the case of digital objects such as software or dynamic data where there may be multiple releases in one year, please include the vocabulary of different types of date (e.g., issued, deposited, available) such as Datacite DateType vocabulary. Do not use string for publication year. Please use the date format.

- **Resource type (M) (1)**: A type of digital object (e.g., research paper, dataset, software, workflow). Please use one of the most used resource type vocabularies (e.g., COAR Resource type vocabulary, Crossref resource types, Datacite resourceTypeGeneral, MARC Genre/Terms Scheme). Resource type has multilingual values.

- **Rights and terms of access (M) (0-n)**: Any rights and terms of access information for this digital object. Typically, rights information includes a statement about various property rights associated with the digital object, including intellectual property rights. The recommended practice is to refer to a rights statement with a URI. If this is not possible or feasible, a literal value (name, label, or short text) may be provided. These metadata elements consist of the:

    - standardised version of the license name,
    - URI of the license,
    - copyrights holder (use authority control databases for persons or institutions if available and persistent identifiers (e.g., ORCID, ISNI, VIAFID) or national authority database of personal and corporate names),
    - link to license ODRL[144] (a machine-readable (RDF) representation of different licenses for data, software, or other digital objects),
    - start date,
    - access rights[145],
    - confidentiality declaration,
    - special permissions,
    - restrictions,
    - citation requirements,
    - conditions,
    - disclaimer,
    - level of access[146],
    - access type, authentications, authorisation, access method, and granularity[147].

These metadata elements are essential for the reusability of the digital object. License information and access rights are mandatory. Other metadata elements are recommended or optional.

- **File (M) (0-n)**: File metadata (e.g., file name, file title, file description, file persistent identifier, download URL, file format, file size, compression format, checksum, checksum

144 ODRL information model. Available on https://www.w3.org/TR/odrl-model/ [19.12.2020]

145 COAR controlled vocabulary for access rights (version 1.0). Available on http://vocabularies.coar-repositories.org/documentation/access_rights/ [19.12.2020]

146 Latanya Sweeney, Mercè Crosas, and Michael Bar-Sinai. Sharing Sensitive Data with Confidence: The Datatags System. Available on https://techscience.org/a/2015101601/ [19.12.2020]

147 George Alter, Alejandra Gonzalez-Beltran, Lucila Ohno-Machado, Philippe Rocca-Serra, The Data Tags Suite (DATS) model for discovering data access and use requirements, *GigaScience*, Volume 9, Issue 2, February 2020, giz165, https://doi.org/10.1093/gigascience/giz165

algorithm, rights, terms of access, processing URL, file version, update/modification date, API URL, update frequency).

- **Subject (R) (0-n)**: Subject, keyword, classification code, or key phrase describing the digital object. If it is possible, please use in subject metadata field well-known subject classification schemes (e.g., UDC, IMT, Eurovoc, Agrovoc, GEMET, ZBW, DDC, MeSH, LCC, LCSH, TGN) instead of character strings. Recommended is also to use subject schemes URIs if it is available. The digital object may have one or more one or more subjects. This metadata element is multilingual.

- **Description (R) (0-n)**: This element is used for a textual description of the content. Description may include but is not limited to: an abstract, methods, table of contents, technical information, reference to a graphical representation of content, or a free-text account of the content. This metadata element is multilingual.

- **Contributor (R) (0-n)**: The institution or person responsible for collecting, managing, distributing, or otherwise contributing to the digital object's development. Please use vocabulary for determining the different types of contributors (e.g., ContributorType vocabulary from Datacite). To supply multiple contributors, repeat this property. Use authority control databases for persons or institutions if available and persistent identifiers (e.g., ORCID, ISNI, VIAFID) or national authority database of personal and corporate names.

- **Date (R) (0-n)**: Different types of dates relevant to the digital object (e.g., accepted, available, collected, copyrighted, created, deposited, distributed, issued, modified, produced, published, submitted, version, valid, uploaded, withdrawn). Allowed values are Date (YYYY-MM-DD) and type of date.

- **Language (R) (1)**: The primary language of the digital object. Allowed values are taken from IETF BCP 47, ISO 639-1 language codes.

- **Alternate Identifier (R) (0-n)**: An identifier or identifiers other than the primary identifier applied to the digital object being registered. This may be any alphanumeric string which is unique within its domain of issue. May be used for local identifiers or other persistent identifiers. Alternate Identifier should be used for another identifier of the same instance (same location, same file). If it is possible, please use persistent identifiers.

- **Related Identifier (R) (0-n)**: Identifiers of related digital objects. These must be globally unique identifiers. Identifiers consist of the type of identifier, digital object type, description of the relationship of the digital object being registered and the related digital object. Please use for identifier type one of values from Datacite (relationType), Crossref intra and inter relation type or attributes from the PROV ontology[148]. If it is possible, please use persistent identifiers.

- **Version (R) (0-n)**: The version number of the digital object (the version number of a dataset or software, the status in the publication process of journal articles). Allowed values are string or COAR Version Types or Related IdentifierType (use only isVersionOf relation).

- **Coverage (O) (0-n)**: The spatial or temporal topic of the resource, spatial applicability of the resource, or jurisdiction under which the resource is relevant.

    - **Temporal coverage** describes the temporal characteristics of the digital object. What the digital object is about or depicts in terms of time (e.g., a period, date or date range (start and end date)). Allowed values are date (YYYY-MM-DD), text, time

148 PROV-O: The PROV Ontology. Available on https://www.w3.org/TR/prov-o/ [19.12.2020]

ontology elements or appropriate vocabularies (e.g., Geological Timescale vocabulary[149],Dublin Core Collection Description Frequency Vocabulary[150], DDI TimeMethod vocabulary[151]).

- **Spatial coverage** describes all aspects of geographic location: not only coordinates indicating where observations were made or what spatial region was observed (geolocation place, geolocation points, geolocation polygons, geolocation boxes, area, volume) but also the coordinate system used, accuracy, precision, resolution, and so on. Allowed values are text or one of Geolocation subject schemes (e.g., TGN[152], Geonames[153], OpenStreetMap[154]…) or geolocation point, or geolocation polygon, or geolocation box values.

- **Project (MA) (0-n)**: A project is an organised effort, either by the individual or collaborative enterprise, that is carefully planned and designed to achieve a particular aim in a specific time frame. Information about financial support (funding) for the digital object being registered. The group of metadata elements that describes a project is project title, project number, project URI, funding stream, funder (funder identifier type, funder identifier, funder name). A good example of a group of metadata elements for the description of projects is "Funding reference" in OpenAIRE guidelines for literature repositories[155]. Very important is the establishment of the distributed authority database of projects and funders on the level of EU.

- **Source (R) (1..n)**: A reference to a resource from which the present digital object is derived. Allowed values are string or Related Identifier metadata. Use only relations IsSourceOf or prov:wasDerivedFrom.

Additional metadata which is recommended for a dataset are:

- **Contact (R) (1..n)**: The contact(s) for this dataset.

- **Producer (R) (1..n):** Person or organisation with the financial or administrative responsibility over this dataset.

- **Production date (R) (1):** Date when the data collection or other materials were produced (not distributed, published or archived).

- **Production Place (R) (1-n):** The location where the data collection and any other related materials were produced.

- **Dataset distribution (R) (0-n):** This property links the dataset to an available distribution (for example CSV, RDF, XLS distribution of data).

- **Depositor (R) (1):** The person (family name, given name) or the name of the organisation that deposited this dataset to the repository.

- **Deposit Date(R) (1):** Date that the dataset was deposited into the repository.

149 Geological Timescale vocabulary. Available on https://vocabs.ardc.edu.au/viewById/196 [19.12.2020]

150 Dublin Core Collection Description Frequency Vocabulary. Available on https://www.dublincore.org/specifications/dublin-core/collection-description/frequency/ [19.12.2020]

151 DDI TimeMethod vocabulary. Available on https://ddialliance.org/Specification/DDI-CV/TimeMethod_1.2.html [19.12.2020]

152 Getty Thesaurus of Geographic Names. Available on https://www.getty.edu/research/tools/vocabularies/tgn/ [19.12.2020]

153 Geonames. Available on https://www.geonames.org/ [19.12.2020]

154 OpenStreetMap. Available on https://www.openstreetmap.org/about [19.12.2020]

155 Funding reference in OpenAIRE guidelines for literature repositories. Available on https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/v4.0.0/field_projectid.html [19.12.2020]

- **Data Collector (R) (0-n):** Individual, agency or organisation responsible for administering the questionnaire or interview or compiling the data.
- **Date of Collection (R) (1):** Contains the date(s) when the data were collected.
    - Start: Date when the data collection started.
    - End: Date when the data collection ended.
- **Kind of Data (R) (1-n):** Type of data included in the file: survey data, census/enumeration data, aggregate data, clinical data, event/transaction data, program source code, machine-readable text, administrative records data, experimental data, psychological test, textual data, coded textual, coded documents, time budget diaries, observation data/ratings, process-produced data, or other.
- **Series (R) (1):** Information about the dataset series
    - Series Name: Name of the dataset series to which the dataset belongs.
    - Series Information: History of the series and summary of those features that apply to the series as a whole.
    - Series identifier: Link to more information about series.
- **Software (R) (0-n):** Information about the software used to generate the dataset.
    - Software Name: Name of software used to generate the dataset.
    - Software Version: Version of the software used to generate the dataset.
    - Software identifier: Link to more information about software.
- **Data Sources (R) (0-n):** List of books, articles, serials, or machine-readable data files that served as the sources of the data collection.
- **Origin of Sources (R) (0-n):** For historical materials, information about the origin of the sources and the rules followed in establishing the sources should be specified.
- **Documentation and Access to Sources (R) (0-n):** Level of documentation of the original sources.
- **Characteristic of Sources Noted (O) (0-n):** Assessment of characteristics and source material.
- **Time Method (R) (0-n):** The time method or time dimension of the data collection, such as panel, cross-sectional, trend, time- series, or other.
- **Frequency (R) (0-n):** If the data collected includes more than one point in time, indicate the frequency with which the data was collected; that is, monthly, quarterly, or other.
- **Universe (R) (0-n):** Description of the population covered by the data in the file; the group of people or other elements that are the object of the study and to which the study results refer. Age, nationality, and residence commonly help to delineate a given universe, but any number of other factors may be used, such as age limits, sex, marital status, race, ethnic group, nationality, income, veteran status, criminal convictions, and more. The universe may consist of elements other than persons, such as housing units, court cases, deaths, countries, and so on. In general, it should be possible to tell from the description of the universe whether a given individual or element is a member of the population under study. Also known as the universe of interest, population of interest, and target population.
- **Unit of Analysis (R) (0-n):** Basic unit of analysis or observation that this Dataset describes, such as individuals, families/households, groups, institutions/organisations,

administrative units, and more. For information about the DDI's controlled vocabulary for this element, please refer to the DDI web page at http://www.ddialliance.org/Specification/DDI-CV/.

- **Standard (R) (1):** The standard in which the service is implemented.

For scholarly publications (journal, journal volume, journal issue, journal article, book set, book series, book, conference paper, content item, database, dissertation, proceedings, proceeding series, report paper, report paper series, series, set and standard) is recommended to use Crossref XML schema[156].

RDA Metadata Interest Group has defined four metadata principles[157]:

- The only difference between metadata and data is the mode of use.
- Metadata is not just for data; it is also for users, software services, computing resources.
- Metadata is not just for description and discovery; it is also for contextualisation (relevance, quality, restrictions (rights, costs)) and for coupling users, software, and computing resources to data (to provide a Virtual Research Environment).
- Metadata must be machine-understandable as well as human-understandable for autonomicity (formalism).
- Management (meta)data is also relevant (research proposal, funding, project information, research outputs, outcomes, impact…).

A researcher wants to find, access, and reuse digital objects in the shortest time possible. In reality, digital objects are often hard to discover (find) and difficult to reuse, hence causing harm to the quality and efficiency of research. Technology can solve many interoperability problems at a technical level - but this does not solve misunderstandings at the semantic level. Humans still need to communicate, agree on terms, and vocabularies. It is important to take advantage of existing frameworks to build cohesion.

156 Crossref XML schema 4.4.2. Available on https://data.crossref.org/reports/help/schema_doc/4.4.2/schema_4_4_2.html [19.12.2020]

157 Keith G Jeffery, Rebecca Koskela. RDA Metadata Principles and their Use. Available on https://rd-alliance.org/metadata-principles-and-their-use.html [19.12.2020]

APPENDIX II. INTERVIEWS WITH STAKEHOLDERS

During the process of creating this document we performed a set of interviews to different stakeholders (researchers from different disciplines) in order to gain a better understanding of their views related to interoperability. As a result, many of the examples used throughout this document are based on examples provided by the interviewees.

The interview process was done during Q4 2019. The following disciplines were covered: Astrophysics, Vulcanology, Marine Sciences, Social Science, Language Resources and Technologies, and Biobanks.

The interview template was as follows:

According to the definition provided in the FAIR data principles (https://doi.org/10.1038/sdata.2016.18**)**, Interoperability is focused on making sure that the data can be integrated with other data, and can be used with applications or workflows for analysis, storage, and processing. Furthermore, the following principles are identified (for data and its corresponding metadata):

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

Q0.1: Do you agree with this definition and principles?

Q0.2: What would you add, if any, based on your understanding of interoperability in your research area?

Q0.3: Is there any other type of resource that should be considered in the context of addressing interoperability (e.g., software, methods, protocols)?

There are classifications of interoperability that focus on different levels: technical, semantic, legal, organisational.

Q1.1: Do you understand and agree with these levels?

Q1.2: Do they happen in your research area?

Q1.3: Would you add any other level, or propose changes to this classification?

Interoperability

Q2.1 Do you or your organisation encounter issues using/integrating data/services from different sources? If so, describe the issues and how you tackle these interoperability issues?

Q2.2 Are there any best practices that you would recommend checking?

Q2.3 Do you have training or use external consulting services regarding any aspects of interoperability?

Technical interoperability

Q3.1 Is technical interoperability relevant for your research area/project/services?

Q3.2 If relevant, how do you address technical interoperability in your research area/project/service? Can you provide some examples?

Q3.3 Are the principles/techniques applied in your area/project applicable to other areas or projects/services? Which principles/techniques? To which types of areas/projects/services?

Q3.4 What are the next steps in technical interoperability that should be addressed in the short and medium-term in your area?

Semantic interoperability in metadata

Q4.1 What metadata standards are recommended in your community? (from FAIRsFAIR survey)

Q4.2 Are metadata standards published in a FAIR manner? Which ones?

Q4.3 Do metadata standards reuse other existing metadata standards (generic, such as Dublin Core, or domain specific)?

Q4.4 How do the metadata standards used utilise or relate to semantic resources/concept systems such as ontologies, terminologies, vocabularies?

Q4.5 Are researchers adding such metadata normally? Are they helped by librarians?

Q4.6 Do they normally fill in all the metadata items or only a subset of them (e.g., Dublin-Core like)?

Q4.7 In your experience, are the metadata standards available well suited for your community? If not, please elaborate (from FAIRsFAIR survey)

Q4.8 Do any of your metadata standards have the potential to be reused/used by another community?

Semantic interoperability in data

Q5.1 Are there any good practices in your community on how to best publish data in a usable/reusable manner?

Q5.2 Is data published using any standards (e.g., W3C standards such as RDF, or as Linked Data)?

Q5.3 Do you use semantic resources (ontologies/thesauri/terminologies/vocabularies) in your community to achieve semantic interoperability? If yes, which ones?

Q5.4 Are semantic resources published in a FAIR manner? Which ones?

Q5.5 Do such semantic resources reuse other existing resources (generic or domain-specific)?

Q5.6 Do most researchers know how to use (and effectively use) such semantic resources?

Q5.7 In your experience, are the semantic resources available well suited for your community? If not, please elaborate

Q5.8 Do any of your semantic resources have the potential to be reused/used by another community?

Legal interoperability

Q6.1 Are there any legal obstacles/barriers for the exchange of data in your community (e.g., data protection, copyright issues, etc.)?

Q6.2 Do researchers understand well those barriers and the actions needed to overcome/deal with them?

Q6.3 Is there any agent/mediator that provides legal support in a centralised or distributed manner, or is it done locally at each project/organisation?

Organisational interoperability

Q7.1 Do you have any policies or procedures defined in advance to encourage your community to work together and exchange information?

Q7.2 Do you have to obey any cooperation agreements with respect to interoperability?

Q7.3 Do you participate in training sessions or use external consulting services regarding interoperability?

Getting in touch with the EU

IN PERSON
All over the European Union there are hundreds of Europe Direct information centres.
You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

ON THE PHONE OR BY EMAIL
Europe Direct is a service that answers your questions about the European Union.
You can contact this service:
– by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
– at the following standard number: +32 22999696, or
– by email via: https://europa.eu/european-union/contact_en

Finding information about the EU

ONLINE
Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU PUBLICATIONS
You can download or order free and priced EU publications from:
https://op.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en)

EU LAW AND RELATED DOCUMENTS
For access to legal information from the EU, including all EU law since 1952 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

OPEN DATA FROM THE EU
The EU Open Data Portal (http://data.europa.eu/euodp/en) provides access to datasets from the EU. Data can be downloaded and reused for free, for both commercial and non-commercial purposes.

This document has been developed by the Interoperability Task Force of the EOSC Executive Board FAIR Working Group, with participation from the Architecture WG.

Achieving interoperability within EOSC is essential in order for the federation of services that will compose EOSC to provide added value for service users. In the context of the FAIR principles, interoperability is discussed in relation to the fact that "research data usually need to be integrated with other data; in addition, the data need to interoperate with applications or workflows for analysis, storage, and processing".

The WGs view on interoperability does not only consider data but also the many other research artefacts that may be used in the context of research activity, such as software code, scientific workflows, laboratory protocols, open hardware designs, etc. It also considers the need to make services and e-infrastructures as interoperable as possible.

Research and Innovation policy

Publications Office
of the European Union