FAIRSFAIR
Fostering Fair Data Practices in Europe

Project Title	Fostering FAIR Data Practices in Europe
Project Acronym	FAIRsFAIR
Grant Agreement No	831558
Instrument	H2020-INFRAEOSC-2018-4
Topic	INFRAEOSC-05-2018-2019 Support to the EOSC Governance
Start Date of Project	1st March 2019
Duration of Project	36 months
Project Website	www.fairsfair.eu

D3.6 Proposal on integration of metadata catalogues to support cross-disciplinary FAIR uptake

Work Package	WP3
Lead Author (Org)	Eva Méndez (UC3M)
Contributing Author(s) (Org)	Joy Davison (DCC), Angus White (DCC), Tony Hernández (UC3M),
Due Date	31.10.2020
Date	27.10.2020
Version	1.0
DOI	https://doi.org/10.5281/zenodo.4134787

Dissemination Level

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

Abstract

This deliverable provides an analysis of the (*meta)data catalogues* concept in different domain-specific research data infrastructures and research data repositories. It discusses the importance of metadata standards and vocabularies for improving the FAIRness of these research data collections, as well as the diversity of specific domain dependent metadata standards and vocabularies (a.k.a. semantic artifacts).

We discuss the problem with five domains (Life Sciences; Photon and Neutron; Social Sciences and Humanities; Environmental research, and with the corresponding five funded 'ESFRI cluster projects' (EOSC-Life, PaNOSC, SSHOC, ENVRIFAIR, and ESCAPE), and we describe a pilot proposal of metadata catalogue integration to improve cross-disciplinary FAIR uptake with the metadata catalogues of the clusters.

Versioning and contribution history

Version	Date	Authors	Notes
0.1	13.08.2020	Joy Davidson	TOC and first outline
0.2	05.10.2020	Eva Méndez	2nd TOC and First draft
0.3	14.10.2020	Eva Méndez, Tony Hernández, Angus Whyte, Joy Davidson	Draft to be shared for internal review
1.0	25.10.2020	Eva Méndez and Joy Davidson	Final reviewed version

Disclaimer

FAIRsFAIR has received funding from the European Commission's Horizon 2020 research and innovation programme under the Grant Agreement no. 831558 The content of this document does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of such content.

Abbreviations and Acronyms

a.k.a	As known as
BARTOC	Basel Register of Thesauri, Ontologies & Classifications
BPMN	Business Process Model and Notation
CERIF	Common European Research Information Format
CDI	Collaborative Data Infrastructure (EUDAT) Cross Domain Integration (DDI)
CLARIN	European Research Infrastructure for Language Resources and Technology
CMDI	Component MetaData Infrastructure
CODATA	COmmittee on Data (International Science Council)
CSMM	Core Scientific Metadata Model
DATS	DAta Tag Suite
DC	Dublin Core
DCAP	Dublin Core Application Profile
DCAT	Data CATalogue Vocabulary
DCAT-AP	DCAT Application Profile
DCATv2	Data CATalogue Vocabulary. Version 2
DCMI	Dublin Core Metadata Initiative
DCMI-Terms	Dublin Core Metadata Terms
DDI	Data Documentation Initiative
DDI-CDI	DDI- Cross Domain Integration
DOI	Digital Object Identifier

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

ECRIN	European Clinical Research Infrastructure Network
EDMI	EOSC Datasets Minimum Information
EIF	EOSC Interoperability Framework
ENVRI FAIR	Environmental Research Infrastructure (ENVRI) FAIR
ERIC	European Research Infrastructure Consortium
EOSC	European Open Science Cloud
EPOS	European Plate Observing System
ESCAPE	European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures
ESFRI	European Strategy Forum on Research Infrastructures
ExPaNDS	European national Photon and Neutron research infrastructures
FAIR	Findable, Accessible, Interoperable, Reusable
FNS-Cloud	Food Nutrition Security Cloud
GLAM	Galleries, Libraries, Archives and Museums
IIIIF	International Image Interoperability Framework
ISO	International Organisation for Standardisation
IVOA	International Virtual Observatory Alliance
JSON	JavaScript Object Notation
MES	Metadata Element Set
MDC	(Meta)Data Catalogue
MDCWS	(Meta)Data Catalogues Workshop(s)
MIG	Metadata Interest Group (RDA)

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

NGR	Next Generation Repositories
OAI-PMH	Open Access Initiative-Protocol for Metadata Harvesting
PaNOSC	The Photon and Neutron Open Science Cloud
PID(s)	Persistent Identifier(s)
PROV-O	PROV (Provenance) Ontology
PSI	Public Sector Information
RDA	Research Data Alliance Research Data Australia
RDF	Resource Description Framework
RDFa	Resource Description Framework in attributes
re3data	Registry of Research Data Repositories
RI (s)	Research Infrastructure (s)
RIF-CS	Research Interchange Format-Collections and Services
RM	Resource Metadata
ROR	Research Organization Registry
RPO	Research Performing Organization
SAM	Science And Metadata Community (DCMI)
SDMX	Statistical Data and Metadata eXchange
SPASE	Space Physics Archive Search and Extract
SRIA	Scientific Research and Innovation Agenda Registry
SSK	Standardization Survival Kit
SSN /SOSA	Semantic Sensor Network Ontology / Sensor Observation Sampling Actuator

UCD	Unified Content Descriptors
UKRDDS	UK Research Data Discovery Service
VO	Virtual Observatory
VRE	Virtual Research Environment
W3C	World Wide Web Consortium
W3C-REC	W3C Recommendation
WG	Working Group
YAMS	Yet Another Metadata Standard

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Executive Summary

FAIRsFAIR is a Coordination and Support Action in the context of two of the main challenges of the European Open Science agenda: EOSC and FAIR data. Therefore, FAIRsFAIR aims at accompanying other projects in the EOSC-FAIR ecosystem with different measures such as standardisation, dissemination, awareness-raising, networking, coordination or support services, policy dialogues and mutual learning exercises. EOSC-FAIR involves different relevant partners in the growing scenario of EOSC, including the ESFRI clusters as associated partners. This deliverable addresses the role of FAIRsFAIR in the EOSC ecosystem, particularly to foster the creation and interconnection of metadata catalogues in order to facilitate and incentivise sharing and finding interdisciplinary data for a common scientific performance among disciplines. This deliverable includes a proposal on the integration of metadata catalogues to support cross-disciplinary FAIR uptake and sets out a pilot which will then be trialed with invited repositories/Research Infrastructures (RIs) from within and outside the project.

Research data management infrastructures, services, and data repositories have their own metadata catalogues and rich descriptions often based on generic and domain-specific metadata and vocabulary standards. Seamless access to, and re-use of FAIR scientific data by researchers at large, cross-disciplines, requires common data models and appropriate domain-agnostic metadata schemas, and vocabularies. The Scientific Research and Innovation Agenda (SRIA) for EOSC reflected the transition of the European science system. It seeks to establish a multi-stakeholder European partnership to enhance the circulation of research data and knowledge in digital form across borders and disciplines, and to allow scientists and machines to collaborate in creating, storing, processing, finding, accessing, and reusing scientific data. This new system must therefore "be the sharing and reuse of data and metadata across all scientific disciplines".

(Meta)data catalogues, as we state in this deliverable, are not specifically mentioned in the FAIR principles, but they have become the key element for enabling research data findability (or even better "discoverability"). Principle F4 states that "(meta)data are registered or indexed in a searchable resource" and this searchable resource happens to be the (meta) data catalogue, as we define it here: the resource/database where you can seek for the datasets that you know are in that resource (find) or the datasets that might be useful for you but you did not know were there (discover). To have an internal metadata catalogue within a repository or set of repositories or thematic infrastructures might guarantee the findability of the research data within the domain or the particular system, but not necessarily the discoverability and interoperability among disciplinary facilities, research infrastructures, portals or other data collections alike.

In this deliverable we describe the complex challenge of facilitating cross-disciplinary data discovery, and the plethora of approaches and metadata standards in use. We define a proposal to test the (meta)data catalogue integration in the five disciplines represented by the ESFRI cluster projects, funded under INFRAEOSC-04-2018, using two domain-agnostic metadata standards: DCATv2-DCAT-AP and DDI-CDI, through B2FIND as a service provider.

Table of contents

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

1. Introduction and context

The European Open Science Cloud (EOSC) is almost a reality due to technical and policy advances. But if research data cannot be found, accessed, integrated and re-used, the goals of the EOSC will remain aspirations beyond the practical reach of many. Data management infrastructures, services, and openly published data, are necessary but not sufficient in themselves to achieve the aims of providing researchers with access to high quality data that supports reproducibility and enables wide reuse of scientific results to achieve both broader and faster innovation. Seamless access to, and re-use of scientific data by researchers at large, cross-disciplines, requires common data models and appropriate domain-agnostic metadata schemas, and vocabularies, that make the datasets not only FAIR (Findable, Accessible, Interoperable and Reusable) but "Discoverable". Shared use of data goes beyond one discipline, expanding the scope of research and diversifying perspectives (Fischer & Zigmond, 2010), motivating new knowledge by discovery, and discovery by serendipity.

The recent Scientific Research and Innovation Agenda (SRIA[1]) for EOSC, presented in July 2020, reflected the transition of the European science system and stressed the need for a multi-stakeholder European partnership to enhance the circulation of research data and knowledge in digital form across borders and disciplines, and to allow scientists and machines to collaborate in creating, storing, processing, finding, accessing and reusing scientific data. This new system must therefore facilitate **"…the sharing and reuse of data and metadata across all scientific disciplines"**. FAIRsFAIR is working to support the realization of an integrated, coherent and reliable research data approach for EOSC to deal with the ideal performance of data-intensive, cross-disciplinary, and global collaborative research.

Besides EOSC and all European Data produced in the context of ESFRI/ERICs or other Research Infrastructures (RIs), there are thousands of data repositories, data catalogues, data portals or other data collections on the Web, providing access to millions of research outputs and datasets. In recent years, a large number of research-data-management systems, "platforms", repositories and other kinds of data/metadata systems and solutions have been developed to gather and preserve research data. The registry of research data repositories (re3data[2]) as of October 2020 registered 2572 repositories/platforms from different domains and disciplines.

Datasets are often hard to discover and difficult to reuse, which causes harm both to quality and efficiency in research and hinders interdisciplinary research. However, the integration of disparate datasets from different domains offers a great potential for new discoveries. Unlike open access publications, research data are normally tightly constrained by the scientific discipline where it was created. The subject matter and the nature of the investigation and the attached e-Infrastructure determine how the data are described, and therefore how they can be retrieved, shared, and the extent to which they can be effectively re-used. Access to this data is crucial for many reasons, including: facilitating reproducibility of research, enabling scientists to build on others' work by re-

[1] https://www.eoscsecretariat.eu/sites/default/files/open_consultation_booklet_sria-eosc_20-july-2020.pdf

[2] http://www.re3data.org

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

using their data, or providing researchers, Research Performing Organizations (RPO), and funders, easy access to information, data, digital objects, and related provenance information.

The current landscape for FAIR data is complex and diverse. (Research) data are:

- **Big.** Many datasets are **massive**. At the end of August 2020, the complete Google Dataset Search corpus contained more than 31 million datasets from more than 4,600 internet domains. About half of these datasets come from .com domains, but .org and governmental domains are also well represented (Noy, 2020; Noy et al., 2019).
- **Growing rapidly.** Research data are getting larger in volume, but the variety of types of data is also increasing and they are being generated at ever greater velocity.
- **Heterogeneous.** Heterogeneity of data refers to both its type and subject. Google Data Search (that only gathers openly available datasets, described using schema.org or DCAT metadata standards) reflects that almost 50 % of the research data available in the Web are from Social and Geosciences, more than 15% from Biology, 9,3% Agriculture, 6% Medicine and between 4-6% Chemistry, Mechanical Engineering, Chemistry, Humanities, and Computer Sciences, along with other disciplines (Benjelloun et al., 2020).

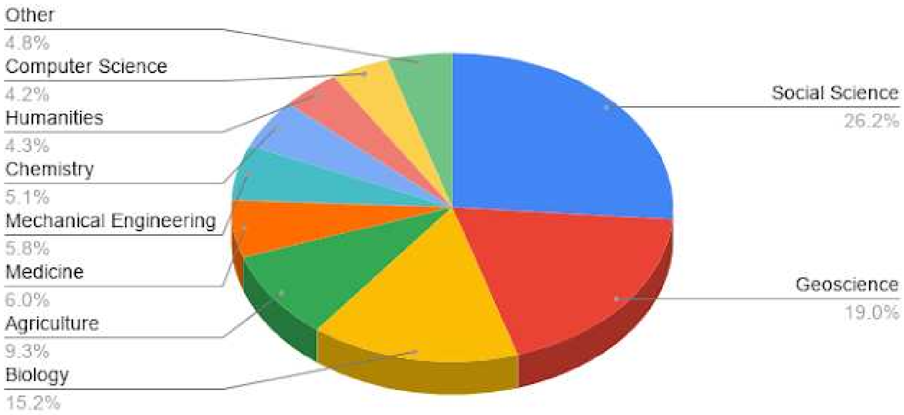


Figure 1. Datasets by topic in Google Dataset Search (Benjelloun et al., 2020)

- **Unique**. Research data are not like books, where a librarian can catalogue one copy and that catalogue information can be applied to every other copy of the book. There is often only one copy, or very few repeated datasets from the work, so it needs to be catalogued independently.
- **Difficult to discover.** Despite the capability of Google Data Search to search among a (huge but) limited number of datasets, most research data consists of non-text data, which are described using their domain specific standards and therefore cannot be found. While open, they are effectively hidden in the deep web of data, only visible and discoverable in Google Data Search when they are described with schema.org or DCAT generic metadata standards.

This complex and diverse landscape for data and research data, affects the decisions to be taken on the right metadata to make the data FAIR. Sometimes, the FAIRness of the data is only measured at domain, infrastructure, or service level, but the lack of interoperability among different disciplines

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

jeopardizes data discoverability from multiple domains. The EOSC Interoperability Framework[3] (EIF) highlights the need of **metadata frameworks** and elements, appropriate at a generic cross-disciplinary level as well as specific community-based interoperability frameworks. It also sets out the need to implement a semantic mapping mechanism and linking to common concepts, to support progress towards higher levels of interoperability.

2. Challenge, objectives and scope

This deliverable is a first attempt at the challenge of EOSC: to facilitate the sharing and reuse of data and metadata across all scientific disciplines. It proposes the integration of catalogues for a distributed set of metadata catalogues and infrastructures. In addressing this, the following five reflections and statements are considered:

1. Solving global problems (e.g. the COVID-19 crisis) requires multidisciplinary, cross-disciplinary, and interdisciplinary approaches by different disciplines or domains concurrently engaged in science and research. Multidisciplinary research implies people from different disciplines working together, each drawing on their disciplinary knowledge. Cross-disciplinary means the view of one discipline from the perspective of another; and interdisciplinary implies the integration of knowledge, methods and data from different disciplines, using a real synthesis of approaches. Metadata **catalogue integration** should foster research on the **concurrence of different disciplines** or research domains.

2. When publishing scholarly digital objects (data and other research outcomes), data generators have to use **appropriate domain-oriented metadata standards to make them FAIR**, as well as semantic artifacts (vocabularies) to add meaning and to specify relationships between them. This allows data consumers (humans or machines), to find, aggregate, and analyse data which would otherwise be invisible, building upon existing standards to push the state of the art in scientific data dissemination. The work described in this deliverable addresses only the metadata interoperability at syntactic and an element set level. Further work on semantics and vocabulary alignment is needed but is beyond the scope of this deliverable[4].

3. Open Science requires seamless integration of research infrastructure resources and it is going to be, from a technical point of view, built upon **next generation things**: next generation Research Infrastructures, next generation metrics (Wilsdon et al., 2017), but also **next generation metadata**, and next generation repositories (NGR[5]). New metadata approaches are needed in the open scholarly communication scenario (Metadata 2020[6] ; Smith-Yoshimura, 2020) but also for the European Open Science Cloud (EOSC), aiming at creating a unique infrastructure for FAIR research data.

4. FAIRsFAIR is a Coordination and Support Action (CSA) in the context of two of the main challenges of the European Open Science agenda: Realising the EOSC vision and increasing the production

[3] https://www.eoscsecretariat.eu/sites/default/files/eosc-interoperability-framework-v1.0.pdf (p.2)

[4] See also, D2.2 FAIR Semantics: First recommendations: https://doi.org/10.5281/zenodo.3707985

[5] https://www.coar-repositories.org/news-updates/what-we-do/next-generation-repositories

[6] http://www.metadata2020.org

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

and reuse of FAIR data. So, **FAIRsFAIR aims at accompanying other projects in the EOSC-FAIR ecosystem** with different measures such as standardisation, dissemination, awareness-raising, networking, coordination or support services, policy dialogues and mutual learning exercises. Inside the FAIRsFAIR project, D3.6 builds upon the work done around metadata and semantics in WP2.

5. This deliverable proposes **a pilot to trial the integration** of metadata catalogues to support cross/interdisciplinary research among the selected domains in the EOSC. The challenges facing the communities along with overviews of current activity were discussed during two workshops with the stakeholders. Building on the findings of these workshops, the small-scale pilot proposed in this deliverable should be seen as a first step and the results will be used to seed for future work coordinated by FAIRsFAIR. D3.7 (Report on integration of metadata catalogues) will report on the results of this piloting, the benefits for FAIR data discovery and reuse, and include recommendations for actions to see greater uptake.

This deliverable contributes to the **general objective** of the project on developing and implementing measures on FAIR data policy addressing, among others, data stewardship and curation and the creation and interconnection of metadata catalogues. But the **particular objectives** are to:

- Discuss approaches to the concept of data/metadata catalogues for disciplinary research data infrastructures/repositories defining types and a common denomination in the context of FAIR (meta)data catalogues.
- Recruit a small group of disciplinary domains to pilot the integration of metadata catalogues and to develop an approach for metadata alignment among disciplines: standards and workflows.
- Highlight interdisciplinary (cross-disciplinary) research needs in the context of EOSC (e.g. use case COVID-19), by discussing and looking for consensus of the “metadata layer” in the European Open Science Cloud (Fig. 2).
- Look further at metadata catalogue integration, bringing to light the complexity of the topic, to further develop other initiatives, inside and outside FAIRsFAIR/EOSC.

FAIRsFAIR does not want to re-invent the wheel or create YAMS (Yet Another Metadata Standard). We start from the EIF (EOSC Interoperability Framework) but also from other previous work, discussions and attempts to address cross/inter disciplinary data discovery, further at Google Dataset Search[7] and focusing on the (meta)data catalogue issue.

For this first proposal of (meta)data catalogue integration, we have chosen the main areas already targeted by the INFRAEOSC-04-2018[8] call, aiming at connecting ESFRI infrastructures through the ESFRI cluster projects, and we have limited the pilot to the final funded projects: Life Sciences (EOSC-Life); Photon and Neutron (PaNOSC); Social Sciences and Humanities (SSHOC); Environmental research (ENVRIFAIR); Astrophysics, astroparticle physics, and accelerator particle physics (ESCAPE).

[7] https://datasetsearch.research.google.com

[8] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/infraeosc-04-2018 The initial areas in the call were defined as: Biomedical Science, Environment and Earth Sciences, Physics and Analytical Facilities, Social Science and Humanities, Astronomy, Energy.

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

In the future, we might extend the (meta)data catalogue integration to other RIs, repositories, data portals/platforms, or other domains (See 5.3).

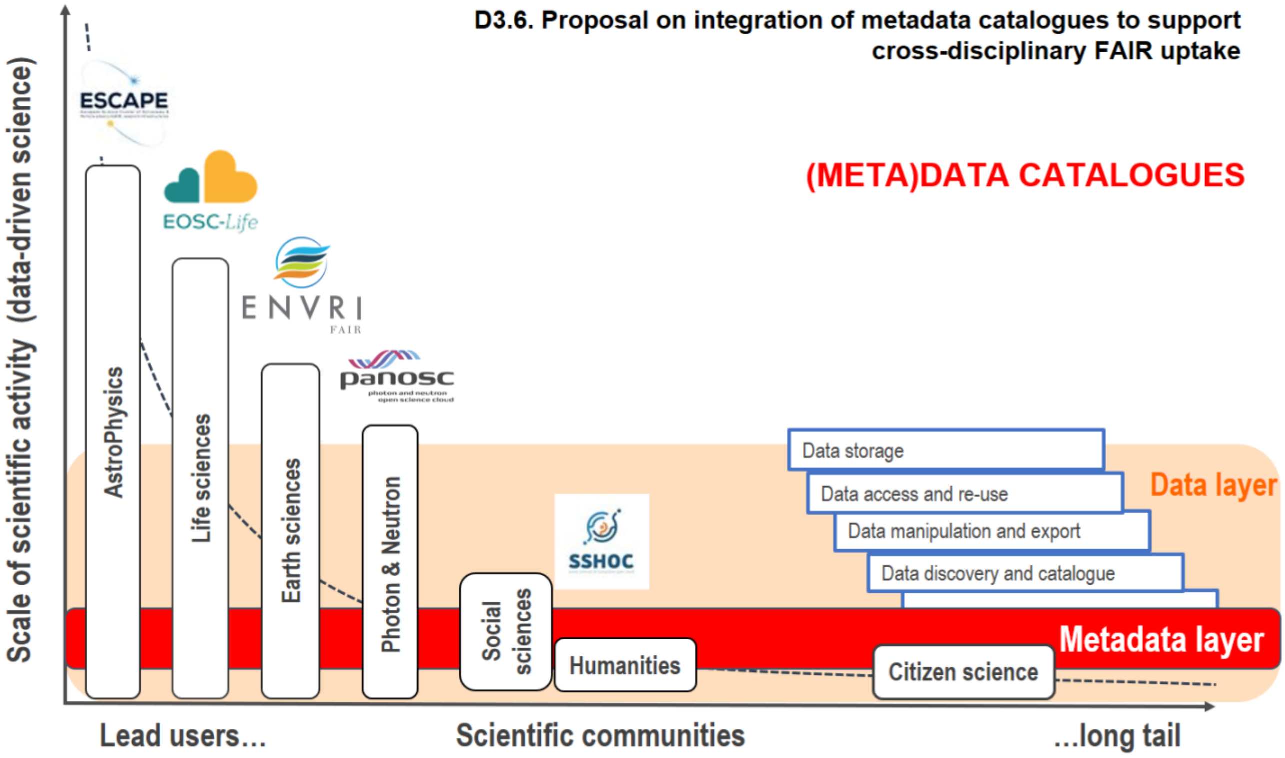


Figure 2. Metadata layer for EOSC. Presentation at the virtual workshop Metadata catalogues integration for interdisciplinary research (by Eva Méndez, 11/09/2020)[9]

As noted above, the aim of this work is not to create yet another metadata standard but rather to build upon existing approaches. Some of the initial questions that we pose to cover the objectives of this deliverable are:

- Which are the domain-agnostic metadata standards that we should consider?
- Which metadata format should we implement for the pilot on integration metadata catalogues?
- How is the metadata catalogue approach of each ESFRI cluster? Are they compatible/complementary?
- How rich can the metadata of a common interdisciplinary metadata catalogue be?
- Which would be the best workflow?
- How can we scale this approach? Could we think of integrating other domains and/or the long tail of science repositories in this new scenario of metadata catalogue integration?

[9] https://drive.google.com/drive/u/1/folders/1md-lMPP2_sJLGTQE1TbGjjOhelYOomId

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

- How can we frame this proposal in a global implementation [political (Open Science) and technical (knowledge graph)]?

3. Methodology

To address the challenge of bringing interdisciplinary interoperability to domain-dependent research data FAIRsFAIR adopted a twofold methodology, combining top-down and bottom-up approaches:

- ***Top-down*** landscape analysis of the metadata standards currently used by the particular domains considered in this proposal, as well as those domain-agnostic metadata standards and their common properties that may be applied/adopted for metadata common catalogue.
- ***Bottom-up*** approach: validating with the ESFRI cluster projects (INFRAEOSC-04-2018 call funded projects[10]) their needs and complicity to test metadata catalogue integration approaches, mapping their domain-oriented metadata schemas to domain-agnostic metadata schemas through a generic metadata catalogue ingestion tool (B2FIND) of EUDAT CDI[11] (Collaborative Data Infrastructure) and EOSC-Hub service. Based on this, we have organized two workshops: one on September 11th 2020, with interested/interesting people to discuss in general about the challenge of "metadata catalogue integration for interdisciplinary research"; and one with the selected clusters, selected metadata standards and B2FIND, to validate the proposal and look forward to the pilot implementation.

This approach was shaped in the following **steps and decisions**:

1. **Selection of the disciplines** to be addressed and the sample of research data repositories or data RIs to implement the proposal. In this phase, we decided to target the ESFRI cluster projects and their represented disciplines for several reasons:
    - The five disciplines (Astrophysics, Life Sciences, Environmental research, Photon and Neutron; Social Sciences and Humanities) are different enough among them (in terms of the datasets they gather, and the topic) to challenge cross-disciplinary data discovery and they are the most data-intensive disciplines to test.
    - The ESFRI cluster projects have addressed themselves the problem of metadata interoperability and (meta)data catalogues integration inside their own domain, among different ESFRIs, and related RIs.
    - We (FAIRsFAIR and the ESFRI cluster projects) are significantly involved in the development of the European Open Science Cloud (EOSC). Furthermore, "working with the research clusters to document and propagate good examples and approaches to embed FAIR data practice in research culture" is one of the specific impacts envisaged in the FAIRsFAIR proposal.

[10] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/infraeosc-04-2018

[11] https://eudat.eu/eudat-cdi

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

2. **Analysis of the previous work** inside and outside FAIRsFAIR and EOSC projects, related with metadata interoperability, metadata catalogues integration and cross-disciplinary research. According to this, we have specifically considered the following resources.

- **FAIRsFAIR related deliverables:**
  - D2.2 FAIR Semantics: First recommendations
  - D2.3 Set of FAIR data repositories features
- **Previous workshop in the EOSCHub context (2018) about data catalogues**: How FAIR Friendly is your data catalogue? Exposing FAIR data in EOSC:
  - http://tinyurl.com/osf-eosc-datacat
  - http://tinyurl.com/osf-eosc-datacat-materials
- EOSCpilot Metadata catalogues strategy: https://eoscpilot.eu/metadata-catalogues-strategy
- EOSC Interoperability Framework: https://www.eoscsecretariat.eu/sites/default/files/eosc-interoperability-framework-v1.0.pdf
- B2FIND Metadata cross-discipline service: https://www.dkrz.de/pdfs/poster/eudat-b2find.pdf?lang=de
- **ESFRI cluster project approaches to metadata catalogue integration issues**:
  - ENVRI design out the service catalogue: https://envri.eu/wp-content/uploads/2020/07/MS18-WP5-Design-of-the-service-catalogue.pdf
  - Report on SSHOC (meta)data interoperability problems: https://zenodo.org/record/3569868#.X1D7YHkzY2z
  - PANOSC Data Policy framework: https://www.panosc.eu/wp-content/uploads/2020/05/PaNOSC-D2.1-PaNOSC-DataPolicyFramework.pdf
  - PANOSC-ExPaNDS: Data catalogue services: https://confluence.panosc.eu/display/wp3
- **Other interesting resources**:
  - OpenAire Research Graph: https://www.openaire.eu/blogs/the-openaire-research-graph
  - Position papers of the ESFRI cluster projects on expectations of planned contributions to EOSC: https://www.fairsfair.eu/sites/default/files/ESFRI_clusters_position_on_EOSC_jan_2020_v1.pdf
  - FAIR Data Maturity Model: specification and guidelines: https://doi.org/10.15497/RDA00050
  - GO-FAIR Discovery Implementation Network https://www.go-fair.org/implementation-networks/overview/discovery

3. **State of the art / Literature review** - this step involved the study of, but not limited to, the following topics:
   - Metadata catalogue approaches in the research data management environment. Literature review and analysis of initiatives, coming up with the concept of (meta)data catalogue.
   - Domain-specific metadata standards in the five chosen domains.
   - Domain-agnostic metadata standards for research data and open data.
   - Interdisciplinary research and cross disciplinary approaches.
   - Metadata catalogue integration and interoperability.
   - ESFRI clusters and their approaches to (meta)data catalogue integration.

1. **Discussion with the stakeholders -** this implied two workshops on *"(Meta)data catalogues integration for Interdisciplinary Research"* (MDCWS[12]):

[12] All the information about the two workshops (presentations, collective notes, chat and lists of invited and attendant people) can be found in: https://bit.ly/MDCWS

- The first workshop was held on September 11, 2020 and brought together more than 50 representatives of different domains, projects, standards and initiatives to discuss current activity and key challenges. The method to invite people was the snowball approach, targeting a small number of experts, the ESFRI cluster projects and the main domain-agnostic metadata standard. This first workshop was conceived as a collective mutual learning exercise that wanted to engage interested and interesting parties in the creation of "metadata catalogues" to grant access to cross-disciplinary datasets belonging to different domains within EOSC. Prior to the workshop some of the invitees contributed ranked the importance of the questions to be addressed in the workshop[13], and provided their definition/understanding of "metadata / data catalogues" (This was achieved using a participation tool called 'wooclap'). Some of the discussions from the workshop are reflected in this deliverable, as well as the vision of the ESFRI cluster projects on (meta)data catalogues.
- The second workshop was held on October 9, 2020 with a smaller subset of the first group (19) including partners of the ESFRI cluster projects engaged with metadata work and catalogue integration; metadata schemas representatives (DCAT, DDC-CDI) and B2FIND to review the draft proposal and agree the participants of the pilot described in section 5 of this document.

4. State of the art: (Meta)Data catalogues for FAIR (meta)data

4.1. Metadata and FAIR data

It does not seem necessary to define **metadata** when its definition is as simple and self-contained as "data about data" that in the context of (research) data management, should also be machine readable and actionable. Metadata is "an unglamorous corner of science" but they are a crucial component or infrastructure often holding the key for data-driven research discoveries (Schriml et al., 2020). However, we agree with (Habermann, 2020) on using a more informative definition coming from the domain of geospatial information ''metadata are the information necessary to understand and effectively use data, including documentation of the dataset contents, context, quality, structure, and accessibility''. Another simpler but good definition in our context is the one of the ISO 11179, chosen in the EIF, which defines it as "descriptive data about an object".

Much of the metadata understanding: types, classifications and uses, come from the previous work on digital libraries and collections, addressing mainly digital GLAM (Galleries, Libraries, Archives and Museums) as well as other information gateways and repositories. But metadata principles, types and uses (Méndez & van Hooland, 2014) are still applicable to the higher heterogeneity of datasets among different and complex scientific domains.

The most traditional classifications of **metadata types** are based on metadata elements, assuming that every metadata schema has elements of similar types, describing similar features of the digital object. Almost every metadata handbook distinguishes the following types of metadata: descriptive,

[13] The most voted questions are included in this presentation https://drive.google.com/drive/u/1/folders/1md-lMPP2_sJLGTQE1TbGjjOhelYOomId

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

structural, and administrative. Thinking of the digital object to be described in our case (research data), this implies:

- Descriptive metadata allow the identification and the retrieval of a digital object. Descriptive elements might include: creator, the title or the subject. They are the metadata used for information/data findability when you know the data to be found, or data discoverability when you do not know them.
- Structural metadata elements facilitate storage, navigation and/or presentation of digital objects, in our case, within the dataset. They provide information about the internal structure of the dataset and may also describe relationships among objects/data.
- Administrative metadata elements help to structure information regarding the management and conservation of a digital object/dataset, such for example the date. These metadata elements are used for managing and preserving objects in a RI or repository, and they can also incorporate 'meta-metadata', information regarding the metadata themselves.

The need to describe data with metadata is not in question. The problem is how best to do it within a domain but also to facilitate understandability outside that domain. There are as many standards as disciplines to make data FAIR. Even at the same domain level, different infrastructures or repositories might use different metadata standards/schemas (Gómez et al., 2016) and of course different schemes or semantic artifacts. Therefore, there are several vocabularies (semantic artefacts) and metadata standards, but a comprehensive vocabulary service does not exist, nor a common metadata standard for data that fits all highly divergent disciplines.

FAIR is a fortunate acronym that everybody repeats and cites when speaking about Open Science and data-driven research, but to make data FAIR is more complex and implies a lot of technical work on the side of semantics and persistent identification. But FAIR are principles, not standards and are "agnostic": technology-agnostic, as suggested in D2.2, but also domain-agnostic. They are a set of clear but wide principles, to be interpreted by any domain or discipline, with different technologies, standards, terminologies, and approaches.

The FAIR Guiding Principles for scientific data management (Wilkinson et al., 2016) are very explicit in the value of the importance of metadata to make research data, (and any other kind of data) Findable, Accessible, Interoperable and Reusable. FAIRsFAIR has already underlined the role of metadata in the context of data repositories[14], and here we are highlighting, one more time, its role in the context of thematic research infrastructures (RIs).

Findable	**Accessible**
The data and **metadata** can be found by the community after its publication, using search tools.	**(Meta)data** are accessible and can therefore be downloaded by other researchers using their identifiers.

[14] See Workshop organized under FAIRsFAIR WP2 *Digital Workshop on common metadata interfaces for "FAIR repositories":* https://drive.google.com/drive/folders/1Fosi-8OVOXGSeJiCACMyyBmosCFvKGPj

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

F1. Assign the **(meta)data** a globally unique and persistent identifier F2. Describe the data with **rich metadata** F3. Register/index the **(meta)data** in a searchable resource F4. The **metadata** should clearly and explicitly include the identifier of the data described.	A1 **(Meta)data** are retrievable by their identifiers using a standardized communications protocol A1.1 The protocols have to be open, free and universally implementable A1.2 The protocol must allow for an authentication and authorization procedure (where necessary) A2 The **metadata** must be accessible, even when the data are no longer available.
Interoperable: Both the data and the **metadata** should be described following the rules of the community, using open standards, in order to allow for their exchange and reuse.	**Reusable:** **(Meta)data** can be reused by other researchers, since their origin and conditions of reuse are clear.
I1. **(Meta)data** must 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**.	R1. **(Meta)data** have a plurality of accurate and relevant attributes R1.1. **(Meta)data** are released with a clear and accessible data usage license R1.2. **(Meta)da**ta are associated with information on their provenance R1.3. **(Meta)data** meet domain-relevant community standards.

Table 1: FAIR principles highlighting the importance of metadata.

Metadata are at the core of FAIR principles. Metadata, rich metadata records, metadata standards and also other vocabularies (semantic artefacts) are crucial to meet FAIR principles in the research data management practice. The metadata are key for findability (F1, F2, F3, F4) and interoperability (I1, I2, I3) but there is also a need for other controlled vocabularies. Accessibility relies on Persistent identifiers but also on the metadata - especially when the data are not openly available (A2) - and to support reusability, in terms of providing provenance and license information. However, using only relevant standards acknowledged by the disciplinary community (R1.3) limits the re-use potential outside the domain.

(Meta)data catalogues are not specifically mentioned in the FAIR principles, but they have become the key element for supporting research data findability (or even better "discoverability"). Principle F3 states that *"(meta)data are registered or indexed in a searchable resource"* and this searchable resource happens to be the "metadata catalogue or the data catalogue" that might (in the case of repositories) or might not include direct access to the data itself. To have a metadata catalogue within the infrastructure, repository or set of repositories or thematic infrastructures might guarantee the findability of the research data within the domain, but never the interoperability among disciplinary facilities, e-Infrastructures or data collections alike.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Metadata elements that might support **findability** are those properties (title, author, keywords, abstract, temporal, and spatial extent) common across many domains, required in many metadata schemas and therefore easily mappable. (Meta)data **accessibility** or availability relies more on the Persistent Identifier (both the PID of the data and the PID of the metadata). The element sets in a metadata schema that support **interoperability** and **reusability** are more specific from the domain and are found also in community vocabularies (schemes or semantic artifacts).

4.2. Data catalogues and Metadata catalogues

Data catalogues, metadata catalogues (MDC) or even metadata platforms (Shaw et al., 2020) or metadata portals (Martin et al., 2019) are different expressions to name the reality of different databases or other computational systems and services that enable researchers to describe their data or other research outcomes (raw or processed data, code, methods, images, etc.) in a standardised and consistent way, using metadata element sets (schemas) and vocabularies (schemes), in general, validated within their community in a specific scientific domain.

There is not much difference between *metadata catalogues* and *data catalogues*, only preferences of denomination. Actually, all data catalogues are eventually metadata catalogues. Similarly, to the traditional library catalogues where all catalogues are metadata catalogues describing the books, whether they have the book (digital library) or not (they only point to a location either in a self or in a server). Similarly, some data catalogues are in fact, metadata catalogues, because they contain the descriptions but they might not have the datasets themselves (they are just registries). However, when it comes to research data repositories, they have both, the datasets deposited and the metadata describing them. But in both cases, there is no real difference between them apart from putting the stress on the data or the metadata. Both denominations can even be used in the same sentence, for example**:**

> ***Data catalogues** have been used in data management for a long time. Under the impetus of European regulations, the number of **metadata catalogues** has been growing steadily over the last decade…* (Quimbert et al., 2020).

The expression "metadata catalogue" might be also understood as a collection of metadata vocabularies: registries or directories of metadata schemas[15] and/or vocabularies or semantic artifacts/assets (vocabularies, ontologies, thesauri[16]), perhaps in that case we might call it "catalogue of metadata".

Almost half of the participants of the first workshop held online September 11th 2020, answered the question "What is for you a metadata catalogue? and a data catalogue?" in the poll prior to the workshop[17], and in most of the cases they agreed upon the approach that we reflect here. In the case of four participants understanding it as a collection of metadata standards ("...a repository where I can search and browse for ontologies, vocabularies and shared data models"). Most of them also

[15] Ex. RDA metadata standards catalog (WG): https://rdamsc.bath.ac.uk

[16] Ex. BARTOC: https://bartoc.org

[17] All the answers can be accessed here: https://docs.google.com/document/d/17nzHL-V6naKKq9zWR3tLUgNmZ6E6bX_pyGYmRw0eQZY/edit

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

agree that “they are the same thing. Catalogues allow me to find datasets based on information about the data in them”, but with particular nuances and understandings, sometimes similarly to what we state here:

> “A metadata catalogue contains information about a dataset, with a persistent identifier, that enables the placement of that dataset into context. A data catalogue can be the same as a metadata catalogue, with the exception that it also contains the data that the metadata describes”.

For the purpose of this deliverable and the proposal of integration, we will use the expression **(Meta) Data catalogue** being conscious that *strictu sensu* such catalog is a pure metadata catalogue but it would give access to the data (if possible) through the PIDs. We use on purpose “(meta)data'' in accordance with the FAIR data principles explanation as well as in coherence what we have discussed here.

4.3. Standards and specifications: Domain-specific metadata *vs* domain-agnostic metadata schemas

> “The nice thing about standards is that there are so many of them to choose from”
> – Andrew S. Tanenbaum

Different RIs, repositories and other digital portals or platforms use different metadata formats and each has its own roadmap or evolution path improving metadata as required by their community. Unfortunately, there are many metadata standards, some general (and usually too wide for scientific use) and some detailed and domain-specific (but not easily mapped against other formats). On the other hand, there are at least seven criteria to typify metadata and metadata standards but one of the most important is the purpose or the focus of the metadata schema, so we can distinguish “metadata for general purposes”, like the Dublin Core, and “metadata for specific purposes”, like ISO 19115 to describe geospatial information, (Méndez & van Hooland, 2014) or the IIIF to describe images. This criterion is applicable at schema level, but if we apply it to the element level, we can also distinguish:

- Domain independent metadata: those elements reflecting general properties of information objects, enabling the abstraction of representational details, such as, the file format, the type of document, the title, etc. When this kind of elements conform a particular schema of generic metadata properties or elements are called ***domain-agnostic metadata*** schemas/standards.
- Domain dependent metadata: the elements enabling a particular representation of domain information and knowledge, describing the information domain to which the underlying digital data or digital object belongs. When these kinds of properties are built up in a metadata schema, we call them ***domain-specific metadata*** standards, and along with the descriptive specific element sets or properties, are also crucial the specific semantic artifacts (vocabularies, content schemes, ontologies, thesaurus, etc.). However, as we mentioned in section 2, addressing adequately domain semantic artifacts exceed the scope of this deliverable.

In any case or classification, metadata standards are crucial to describe, retrieve, access and reuse all research outputs, including publications and data. In the scholarly communication process, and from the Open Science perspective, metadata is the information describing research outputs (publications, data, methods, software and others). Common to most scholarly research outputs are metadata elements such as author, date, title, subject, language, and the persistent identifier. In the case of research data, metadata describes specialized aspects such as the geographic location where the data was collected, the name and identifier of the research funder, the institutional affiliation of the researchers, contributors such as editors and data curators, or the number of the grant awarded to fund the research (Gregg et al., 2019). The minimum element sets or properties to be described are those that make the research data citable and discoverable (data identifier, creator(s), title, publisher, and publication or release date). But to make data FAIR we need richer metadata, and at some point (re-usability and reproducibility) domain specific metadata schemas and aligned vocabularies.

There are different federated (meta)data catalogues that have developed a specific metadata standard/schema format to aggregate research data from different collections, repositories, virtual research facilities, or other research infrastructures. For example, Research Data Australia (RDA) uses RIF-CS (Registry Interchange Format - Collections and Services[18]) that supports the electronic exchange of collection and service descriptions. It is the metadata format required by the RDA Registry to enable descriptions to be harvested automatically for display and discovery in Research Data Australia[19]. However, in the case of EOSC there is an initial tacit consensus that we do not need a "YAMS" (Yet Another Metadata Standard") or a new data model to create a metadata catalogue, however in the context of EOSCpilot project there was an approach to create the EOSC Dataset Minimum Information (EDMI).

There are as many metadata schemas, standards, and application profiles based on them, as domains, projects, collections or digital information services on the Web. Research data repositories and Research Infrastructures are not an exception to creating, using or adapting metadata schemas and/or standards, and they might be found in resources like:

- FAIRSharing[20] collects, as of October 2020, 1458 "standards" in general (including semantic artifacts, taxonomies, etc.) to be applicable to RIs, repositories, portals, platforms or other research data systems. 69 of these standards are "metadata standards" applicable to more than 50 identified subjects or uses; 42 are specific metadata models/formats, focusing on element set/property vocabularies, and 13 semantic or terminology artifacts).
- Digital Curation Centre (DCC) Disciplinary Metadata[21], where metadata standards are classified in 4 domains: Social Sciences and Humanities, Physical Science, Earth Science, Biology and an extra category called "general research data".
- RDA Metadata Directory/Metadata Standards Catalog[22], created by the Research Data Alliance Metadata Standards Catalog WG, is a community managed version of the DCC guide on disciplinary metadata. The RDA Metadata directory, also known as "metadata catalogue"

[18] https://documentation.ardc.edu.au/display/DOC/About+RIF-CS

[19] https://researchdata.edu.au

[20] https://fairsharing.org

[21] https://www.dcc.ac.uk/guidance/standards/metadata

[22] Metadata Standards Directory https://rd-alliance.github.io/metadata-directory / Metadata Standards Catalog: https://rdamsc.bath.ac.uk

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

includes, like that of DCC, metadata standards and applications profiles (extensions) by discipline that might be potentially used to describe research datasets of those identified domains.

It is important to underline that even though all metadata schemas and specifications are probably born with the goal of becoming a metadata standard, at least within the domain or the group defining or creating them, not all of them finally become standards. It is needed to assess the level of standardisation that a metadata model or schema reaches. That level ranges from *de jure* ISO standards, like the Dublin Core Element Set (DCMITerms), to different levels of *de facto* standards like the Recommendations of the W3C, the GIS (Geographic Information Systems) standards[23] or other open standards or public access specifications. "Meet domain-relevant community standards", as we have mentioned, is one of the FAIR principles (R1.3) to guarantee data re-usability, however the relevance for a community of a metadata "standard" (schema or specification) is not necessary related with its level of formal standardisation, but is maintenance and credibility.

The most important aspect in choosing a metadata schema, once it fulfills our descriptive needs and functional requirements, is that it is generally adopted within a community evidenced by its massive use and easy implementation. We analyse here below the most important **domain-agnostic metadata standards-schemas to be considered** for our proposal.

▪ DCMI (Dublin Core Metadata Initiative)

DCMI is probably one of the most adopted and adapted metadata standards. Born in 1995, its collection of 15 basic elements (DCMITerms[24]) became an ISO standard (ISO 15836[25]) in 2003. DCMI is a domain-agnostic and type-agnostic metadata element set (MES) adopted by all the scholarly publications repositories as well as many digital libraries and digital information services. Its adaptability and extensibility to a specific domain, through the creation of application profiles, has made DCMI both an extensible metadata schema and a simple element set.

The Dublin Core (DC) metadata schema/format/standard was not developed to describe research datasets. It was created for the-web-of-documents not the web-of-data. Nevertheless, several disciplines, like the Social Sciences and Humanities (SSH) use DCMI as a generic metadata format to describe their research datasets, both in the EOSC Cluster project approach (SSHOC[26]) and in the thematic repositories of these disciplines (Gómez et al., 2016), like Datorium (Andias, 2014). Data in SSH are in many cases very close to digital GLAM collections where the Dublin Core has been used for years.

[23] https://www.gistandards.eu/gis-standards

[24] https://www.dublincore.org/specifications/dublin-core/dcmi-terms

[25] https://www.iso.org/standard/71341.html

[26] Mari Kleemola (SSHOC), during the discussion of the 1st workshop: "SSHOC experience is that Dublin Core is general enough to be suitable for different domains and for data discovery (findability) in aggregated metadata catalogues. However, conversion from domain specific metadata standard to it leads almost always to loss of information. But of course careful conversion can lead to meaningful and informative documentation with Dublin Core" See: https://drive.google.com/drive/u/1/folders/1EQmA9q-pm6ieEDAyWCfMNKAcwRWx0SUb).

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Even though it is a very generic standard, its simplicity, generic use and level of adoption in the scholarly communication environment, as well as its role as a substantial mapping standard for OAI-PMH harvesting, makes DCMI "a must" in the consideration of domain-agnostic metadata standards for cross-disciplinary uptake.

DataCite metadata schema

Data Cite metadata schema is the metadata standard associated with the DOI (Digital Object Identifier) assigned by DataCite to the research data (typically a dataset) that makes the datasets persistently identified. It is intended to be generic to the broadest range of research datasets, rather than customized to the needs of any particular discipline, and it does not replace the discipline or community specific metadata. However, it is a global standard that primarily supports citation and discovery of data.

DataCite's Metadata Schema has been expanded in each new version, but it is intended to be generic to the broadest range of research datasets, rather than customized to the needs of any particular discipline. DataCite metadata primarily supports citation and discovery of data; it is not intended to supplant or replace the discipline or community specific metadata that fully describes the data, and that is vital for understanding and reuse.

It includes a list of properties chosen for an accurate and consistent identification of a resource for citation and retrieval purposes, along with recommended use instructions. The metadata properties are presented in three categories: Mandatory (M), Recommended (R), and Optional (O) (DataCite Metadata Working Group, 2019). This metadata schema contains support of organizational identifiers, like ROR IDs (Research Organization Registry[27] Identifiers) as well as other content schemes or vocabularies. DataCite metadata schema had collaborated with the Dublin Core Metadata Initiative (DCMI) Science and Metadata Community (SAM[28]) to maintain a Dublin Core Application Profile (DCAP) for the schema.

DCAT (Data CATalog Vocabulary)

The Data Catalog Vocabulary is a "new-popular" and widely adopted standard for describing datasets and establishing interoperability between data catalogues or in the context of Public Sector Information, data portals. DCAT became first a W3C Recommendation (W3C-REC) in January 2014 and in February 2020 the second version of the W3C-REC was published (DCATAv2). Founded in Dublin Core principles, DCAT captures many essential features of a description of a dataset: the abstract concepts of the catalogue and datasets, the realizable distributions of the datasets, keywords, landing pages, links to licenses, publishers etc.

DCAT-AP[29] built up as a Linked Data extension of DCAT which adds metadata fields and mandatory ranges for specific properties. This application profile is the customization of the data catalogue

[27] https://ror.org

[28] https://www.dublincore.org/groups/sam

[29] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/about

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

vocabulary for data portals in Europe, particularly the European Data Portal[30]. It is a specification for describing public sector information (PSI), particularly datasets, and enables the exchange of descriptions of datasets among different data portals. DCAT-AP is a specification for metadata records to meet the specific application needs of data portals in Europe while providing semantic interoperability with other applications on the basis of reuse of established controlled vocabularies and mappings to existing metadata schemas (DCMI, SDMX (Statistical Data and Metadata eXchange), INSPIRE metadata, etc.). The popularity of DCAT-AP has been increasing in the dataGOV domain and country-specific extensions have been published and used in their governmental data portals.

The first EOSC Declaration[31] highlighted the need of interoperable research data standards. So, in May 2018 a report entitled *Research Data Analysis*[32] was published to analyse the suitability of DCAT-AP to represent and exchange research (meta)data both between research data repositories themselves, and also between research data repositories and general purpose open data portals. The study concluded that DCAT-AP can act as a common language between domain-agnostic research metadata models and can thus help the exchange of metadata between research data catalogues.

Furthermore, the new European Directive of Open Data and PSI (Directive 2019/1024), published in June 2019, specifically includes “research data” as Public Sector Information[33] acknowledging the potential re-use of research data beyond the scientific community, and therefore the need of being FAIR, interoperable with other PSI infrastructures, and Discoverable beyond research context.

▪ DDI-CDI (Data Documentation Initiative- Cross Domain Integration)

DDI-CDI is a specification aimed at helping implementers integrate data across domain and institutional boundaries. It stems from the idea that modern research increasingly involves large amounts of data, much of which comes from non-traditional sources (sensors, big data, social media, etc.) and often from other domains. DDI-CDI focuses on a uniform approach to describing a range of needed data formats which allows them to be connected and understood to support transformation and processing for integrated use. DDI-CDI is aligned with other DDI specifications (DDI-Codebook[34], DDI-Lifecycle[35]) to support integration of external data in systems which use DDI.

DDI-CDI offers an extension to the suite of DDI work products which helps those in the Social, Behavioral and Economic (SBE) domains (and outside of them) to integrate the expanding range of data required by today's research. It is explicitly designed to work with many popular generic technology standards used in that domain, such as PROV-O, BPMN (Business Process Model and Notation), DCAT (Data Catalog Vocabulary), SDMX, DataCube, SSN/SOSA and Schema.org to allow for easy integration into systems which support them.

[30] https://www.europeandataportal.eu/en

[31] EOSC Declaration (26 October 2017): https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf

[32] https://joinup.ec.europa.eu/sites/default/files/document/2018-05/Research%20Data%20Analysis_v1.00.pdf
Previously (November 2016) Andrea Perego et al. published also a document on this issue, in the context of a W3C workshop See: Using DCAT-AP for Research Data: https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_27

[33] https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1561563110433&uri=CELEX:32019L1024

[34] https://ddialliance.org/Specification/DDI-Codebook

[35] https://ddialliance.org/Specification/DDI-Lifecycle

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

The Cross-Domain Integration approach of DDI (Data Documentation Initiative) is currently undergoing review so it is an opportune moment to test the potential value for integrating metadata catalogues and to feed our findings into the wider review process.

▪ Schema.org

Schema.org was created approximately ten years ago, by the main Search Engines of the web (Google, Bing, Yahoo and Yandex) to develop a metadata vocabulary and description mechanism for any kind of information in the web. It is basically a machine-readable semantic annotation mechanism for web content, including data (as a type of content) published on the web. Schema.org has two components: 1) an agreed simple hierarchy of resource types and a vocabulary for naming the characteristics of resources, their relationships, and constraints on how to describe these characteristics and relationships; and 2) a simple syntax to express that information in machine readable formats such as microdata and RDFa.

Since the creation of schema.org different domains have tried to customise its generic nature to domain or type specific information through extensions, and so called Schema.org Community Groups, hosted by the W3C but not necessarily supported by it. The Schema.org extension is like DCMI or DCAT application profiles, a mechanism to become more specific or to adapt a domain-agnostic metadata standard to a specific domain or discipline (e.g. Bioschemas for life sciences[36], Schema Bib Extend and Bibframe2schema for bibliographic data[37]; or Open Educational Resources[38], etc.)

As they describe themselves, Schema.org as a project is a collection of terms, and is entirely devoted to data. It always describes or encodes some form of data. However, for the interest here, we focus on its domain-agnostic standard to describe data and datasets. Its wide use to add structured metadata have driven the creation of a specific RDA WG based in the Data Discovery Paradigms Interest Group[39] task force ("Using schema.org for research data discovery") to discuss how the research community could come together to embrace the advantages of discovering data via search engines as well as to identify issues, gaps and deficiencies.

Schema.org (along with DCAT) is the crucial metadata vocabulary behind Google Dataset Search to discover datasets publicly available on the web. It would be naïve to not consider this metadata standard: first for its great level of implementation and simplicity but also for its potential role of making visible the research data on an all-the-web approach provided by Google. However, we must bear in mind that not all the FAIR data published in the ESFRI/ERIC landscape are Open, as well as consider the envisaged added value of EOSC services.

[36] https://www.w3.org/community/bioschemas; https://bioschemas.org

[37] https://www.w3.org/community/schemabibex/; https://www.w3.org/community/bibframe2schema

[38] https://www.w3.org/community/oerschema

[39] https://www.rd-alliance.org/groups/data-discovery-paradigms-ig

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

- Other metadata schemas/initiatives to describe any scientific dataset

From the previous sections, it is clear that there are several domain-agnostic metadata schemas/standards that could potentially describe datasets:

- Specific standards specially created to describe any dataset (Datacite schema, DCAT)
- Specific standards looking at the interdisciplinary nature of datasets (DCATv2/DCAT-AP and DDI-DCI)
- Standards not created specifically to describe datasets nor even research datasets (DCMI, and Schema.org) but very flexible to adapt themselves, through application profiles or extensions, to describe datasets.

Also, we can add, those initiatives that are not yet standards but are developing and discussing the possible minimum set of metadata elements (schema level) that might be used to describe research data or scientific data:

○ EDMI (EOSC Dataset Minimum Information), an initiative developed under the EOSCpilot project that includes an extensible interoperable set of FAIR catalogues for metadata, tools, workflows and data standards adhering to a minimal metadata standard, extending for community specific needs. It includes a set of properties[40] (metadata elements) at three levels: minimum (11 properties) recommended, and optional as well as metadata properties/element mapping among other independent domain agnostic schemas (DataCite, Schema.org and B2FIND) and others, domain agnostic, but specific to a project or an infrastructure like UKRDDS[41], developed for the UK Research Data Service, or DATS[42], Data Tag Suite developed to support the DataMed data discovery index.

○ List of terms of RDA Metadata Interest Group (MIG): RDA-MIG has been working since the Plenary 9 on a list of elements to address the interdisciplinary issues and the discoverability among disciplines. It is also a simple list of a few elements, not single-valued attributes (17 in this case)[43]. It is a domain-agnostic Metadata Element Set (MES) which assumes that subject domains and particular disciplines might have much greater metadata element lists. This list is intended to be the recommended list of elements that should be provided by all within RDA to: permit discovery, support contextualisation (assessment of relevance and value) and facilitate action (interoperation including query and integration).

5. Proposal for cross-disciplinary (meta)data catalogue integration

Two main immediate options to foster interdisciplinary research through a common discoverability framework can be considered: a) To develop and apply common standards across domains and

[40] https://eosc-edmi.github.io/properties

[41] https://rdds.jiscinvolve.org/wp/2016/03/11/core_metadata_profile ; https://drive.google.com/file/d/0B3v6Fm7XStdBWUpvc3FWQjhoMTA/view

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

[43] https://drive.google.com/drive/folders/0B8FnM3PsoL2dd2RnYVBmcjRMYXc

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

institutions and b) To create a (meta)data catalogue for a distributed set of metadata catalogues and infrastructures. Option (a) is almost rejected for three fundamental reasons:

- The community of stakeholders participating in our workshop agreed that a **single metadata standard** is a **"no-go"**
- The implementation of web-based information services during the last 25 years has demonstrated that it is better to **build upon already existing standards** than create a complete new one from scratch
- It is **impossible** (or discouraged) to create a new metadata **(one-size-fits-all)** standard that satisfies all research communities, data facilities and data-driven research procedures
- Creating a brand new "ideal" metadata schema implies a **standardisation process,** very costly and too long

Therefore, the FAIRsFAIR proposal is to agree upon a **"high level metadata element-set"** that will support data integration across diverse types of research data and different domains. Understanding "high-level" as low specificity and high interoperability potential. As the EIF states:

> *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.* (EOSC Interoperability Framework, p.27[44]).

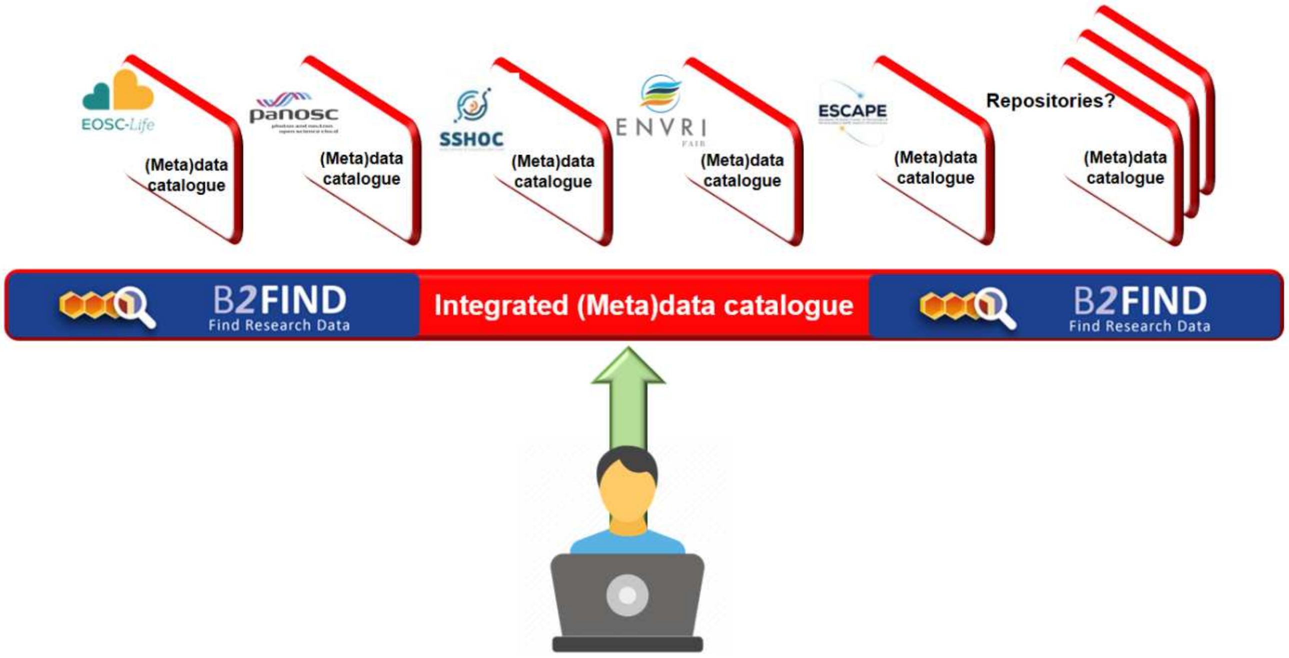


Figure 4: FAIRsFAIR conceptual proposal of (meta)data integration catalogue

[44] https://www.eoscsecretariat.eu/sites/default/files/eosc-interoperability-framework-v1.0.pdf

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

In FAIRsFAIR-Task 3.4, we have been analysing the requirements to enable data repositories, research infrastructures and other service providers to support FAIR data in terms of general practice and the specifics of metadata handling and integration. In the work described here (D3.6) we have agreed to test already existing metadata standards specifically adequate for metadata integration: DCATv2/DCAT-AP, and DDI-CDI. So, starting from domain specific metadata, each cluster will implement a semantic mapping and linking mechanism to common elements of those domain-agnostic models, to explore progress towards higher levels of interoperability.

5.1 EOSC thematic clusters. (Meta)data catalogue issues and current approaches

In this section we detail a brief first approach to the ESFRI cluster projects that we have targeted in our proposal, particularly paying attention to their individual approaches on (meta)data catalogues integration, first within their own domain, but also on the discussion of a potential cross-disciplinary discovery and future integration with EOSC[45].

▪ Life Sciences: EOSC-Life

EOSC-Life represents 13 RIs across the food and health domains. They are very interdisciplinary and, accordingly, there are numerous approaches to metadata descriptions making the use of a single metadata catalogue not feasible. Key challenges around metadata catalogues include:

- The variability in data resource maturity between different RI communities
- Complexity of data and relationships between datasets
- Wide range of cataloguing approaches, from highly centralised to highly distributed
- Varying requirements of different user groups data generators, data browsers, downloaders, data analysts, and data owners
- Open and controlled data projects
- Legacy: 2/3rd of the first wave projects have their own data cataloguing platforms

To support effective data discovery and reuse, existing and future metadata catalogues must interoperate and will require links in interfaces, common vocabularies, common minimum cataloguing terms and cross-walks, shared API standards.

A key issue for this community is handling sensitive data and access to some of the datasets require AAI authentication (provided by EOSC Life WP4) and may never be indexed more generally.

There are many deposition databases to share results - most use minimum metadata standard e.g., MIAPPE. For example, OmicsDI[46] which is an aggregator harvesting experimental and sample metadata from 23 open life sciences molecular databases and FAIRsharing which is a curated resource of data and metadata standards, inter-related to databases and data policies. There are also workflow registries which live in native repositories and are harvested into workflow hubs.

[45] The descriptions included in this section are based in the presentations given by the clusters in the first MDWS on September 11, available here: https://drive.google.com/drive/u/1/folders/1md-lMPP2_sJLGTQE1TbGjjOhelYOomId , and the information provided by the clusters on the web.

[46] https://github.com/OmicsDI

The metadata ecosystem for the Life Sciences involves pooling and maintaining metadata. EOSC Life looked at simple ways to integrate these and explored the tailoring of schema.org leading to Bioschema.

Within the Life Sciences, there are several levels of catalogues that are needed:

- Cataloguing level: about the catalogues
- Data Catalogue level: about the datasets, data records
- Data type level: about what is being catalogued
- Other, e.g. training materials, tools, workflows, protocols

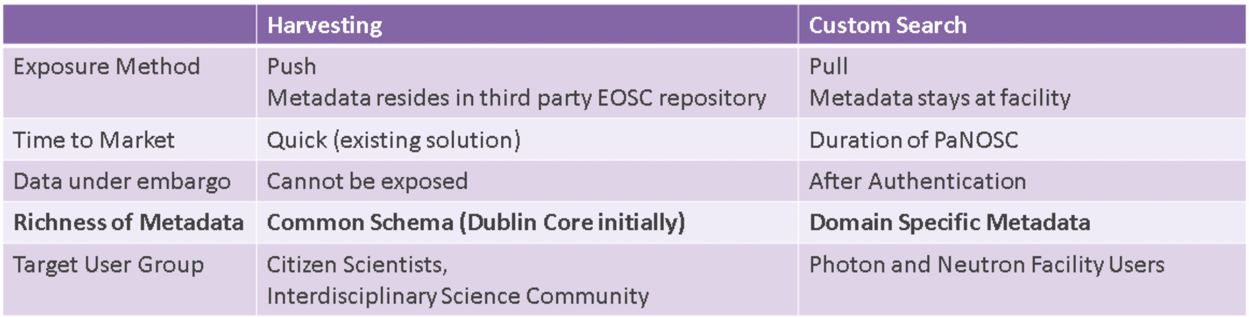
Due to the variation in catalogue levels, there is a need for an extended and interoperable set of catalogues.

▪ Photon and Neutron: PaNOSC (+ ExPaNDS)

PaNOSC and ExPaNDS are two EOSC EU funded projects bringing together research infrastructures in the Photon and Neutron domain. PaNOSC targets the Photon and Neutron ESFRIs, and ExPaNDS the national infrastructures in the domain, and they have the objective to adopt and implement data management, simulation and analysis services, and to make their open data available to the EOSC. Both projects are addressing the metadata catalogues integration together. Within this community, catalogue information typically includes:

- Proposal Information
- Sample Information
- Experimental Parameters
- Previews
- Provenance
- Relationship to derived datasets
- Relationship to people

There is often a two-tiered embargo period that must be respected within this community of researchers. In close cooperation with ExPaNDS, the PaNOSC project team have been working on pushing data into B2FIND and OpenAIRE with some success.



	Harvesting	Custom Search
Exposure Method	Push Metadata resides in third party EOSC repository	Pull Metadata stays at facility
Time to Market	Quick (existing solution)	Duration of PaNOSC
Data under embargo	Cannot be exposed	After Authentication
Richness of Metadata	**Common Schema (Dublin Core initially)**	**Domain Specific Metadata**
Target User Group	Citizen Scientists, Interdisciplinary Science Community	Photon and Neutron Facility Users

Figure 5: Two routes (push/pull) for Opening Datasets in PaNOSC. Presentation at the virtual workshop Metadata catalogues integration for interdisciplinary research (by Tobias Richter, 11/09/2020)

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

The projects are currently working on developing dictionaries and common vocabularies based on NeXus[47] (its domain-specific metadata schema, seen as a common data format for neutron, x-ray and muon science). Ongoing metadata catalogue issues are being explored by the projects include describing:

- Roles for persons involved with the data
- Experimental or Measurement Technique
- Sample & Parameters
- Experimental Equipment and Parameters
- Cross-over with RDA activity on PIDS for experimental equipment

▪ Social Sciences and Humanities: SSHOC

The SSHOC cluster brings together research infrastructures in the Humanities and Social Sciences domains. The research carried out in these domains is multidisciplinary in nature with very heterogeneous and diverse metadata approaches in use. In their landscaping activity, SSHOC carried out a series of interviews which highlighted no fewer than 19 metadata standards currently being used by the communities. SSHOC recommends the use of domain-specific standards for expressivity and common standards for low-level interoperability/discoverability. Schema.org and DDI-CDI were not covered in the SSHOC metadata report as they are not generally used by SSHOC communities at present however, interest in them is rising.

For the SSHOC Open Marketplace, the project:

- Collects metadata on tools, service, training material, workflows, datasets, publications
- Targets researchers from the humanities and social sciences
- Supports the usage of digital methods and allow to discover (new) tools/approaches/findings
- Relies on curated items: Curators generate high-quality metadata based on initial ingestions from sources (amongst others metadata catalogues)
- Relates items with each other
- Ingests list of tools
- Ingests publications and relates to tools (i.e., if a tool is mentioned in a publication create a relation
- Enriches items (by using different sources), e.g. identify research community where an item belongs to or is used

Currently, SSHOC's approach to the integration of metadata catalogues with respect to the SSHOC Marketplace is to:

- Decouple software architecture: API is the main connection point for ingestion/ enrichment/relation
- Ingest sources (e.g. metadata catalogues) with the help of dedicated tools
- Manual mapping of sources to our data model / vocabularies

[47] https://www.nexusformat.org

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

- History of changes: versioning and merging of items (identify the same item at different sources) requires a lot of effort

For the internal alpha release of the marketplace, they ingested data from different sources in respect of scope (list of tools, training materials) and methods (use of source API, GitHub API, spreadsheet document). The sources include:

- TAPoR[48]: list of tools, use source API
- Programming Historian[49]: list of training materials (including references to tools in text), use GitHub API
- SSK (Standardization Survival Kit[50]): list of workflows (including references to training material/publications in a Zotero library), use GitHub/Zotero API
- Curated items: mixture of items, use spreadsheet document

Challenges that SSHOC has encountered with its sources/metadata catalogues include:

- Gathering of data not always easy/well documented
    - Lack of machine readability, e.g. a machine-readable metadata schema
    - APIs often not so well documented in regard to semantics of data
    - FAIR data principles as a good checklist/recommendation
- Identify same items: works well for publications but not so well for tools
    - PIDs can help but we miss something like rdfs:sameAs
    - Disambiguation as an issue where curators are necessary
- Good metadata quality on SSHOC side relies on good metadata quality at the source side otherwise curators are necessary. For example, information on the research community that uses an item is often “hidden” for machines in the contexts of the source
- Relations of items from different and especially cross-disciplinary sources: Different data schemas, interpretations and vocabularies

▪ Environmental research: ENVRI-FAIR

ENVRI-FAIR brings together the ESFRI RIs and landmarks from the cluster of European Environmental research infrastructures creating policies and standards aligned with EOSC and the current European initiatives, like Inspire. ENVRIFAIR will implement the ENVRI-hub, a virtual, federated machine-to-machine interface to access environmental data and services provided by the contributing RIs. The complete set of thematic data services and tools will be incorporated into the EOSC service catalogue, with the following goals:

- Goal 1: Cataloguing all RIs in the environment domain
- Goal 2: Starting point for accessing RI datasets
- Goal 3: Interface to the European Open Science Cloud

[48] http://tapor.ca
[49] https://programminghistorian.org
[50] http://ssk.huma-num.fr

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

To reach its goals, ENVRI-FAIR has a Specific Task Force dedicated to metadata catalogues (Catalogue Task Force) and they are testing the EPOS solution[51]. To ensure its integration with EOSC ENVRI-FAIR will provide metadata extracted from the RI catalogues and transformed into an EOSC services catalogue.

ENVRI-FAIR does not simply combine the metadata from the RIs (meta)data catalogues. In order to harmonise them, enrich the metadata on a canonical metadata catalogue that represents descriptions of services from the different RIs. The catalogue will be filled with existing resources and assets, which will be represented by metadata records describing, datasets, services, workflows/tools, e-services and other assets such as equipment. In terms of metadata, this cluster had pointed out the inadequacy of the metadata schema proposed in the context of EOSC Hub, inherited from eInfracentral. The previous project ENVRIplus concluded with trials of two homogenizing technical frameworks, namely CKAN (as used by EUDAT) and CERIF (as used by EPOS). Environmental RIs were invited to try these solutions and propose what they wanted for any common catalogue. An existing, well-tested option is on the table, that is to say the CERIF metadata catalogue already used by the EPOS RI, that permits export to DCAT and DCMI.

▪ Astrophysics: ESCAPE

ESCAPE brings together ESFRI facilities of astronomy, astroparticle and particle physics into a single EU collaborative cluster. Plus, it will create a cross-border and multi-disciplinary environment that will benefit EOSC thanks to the management of extremely large data volumes at the multi-exabyte level. Focusing on semantic aspects of metadata, ESCAPE addresses the Open Science challenges shared by the astronomy/astroparticle/accelerator particle physics communities and aims at connecting the domain ESFRI to EOSC via the Virtual Observatory (VO) framework.

The ESCAPE approach to (meta)data catalogue integration has done several steps within a domain where building a metadata standard took 20 years in the context of the IVOA (International Virtual Observatory Alliance):

- (Meta)data catalogue integration involves challenges across infrastructures
- They have already connected the VO registry to B2FIND
- The architecture of the VO is FAIR

The domain standards used are:

- Metadata about data collections and data services: RM 1.12 (2017. Resource Metadata for the Virtual Observatory)
- Metadata for the distributed and harvestable VO registries are: DC with disciplinary extensions that are well aligned with DataCite metadata schema
- It also includes semantic standards:
    - IVOA standard for Unified Content Descriptors (UCD), that imply key concepts in astronomy
    - VO units to express physical units in all measurements tables

[51] European Plate Observing System: https://www.epos-ip.or

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

- Vocabularies and semantic artifacts like dictionaries of standardised labels used in the metadata profile of the Virtual Observatory, as well as other domain specific vocabularies encoded in RDF and JSON.

ESCAPE is working to include the existing interoperability framework of the Virtual Observatory in EOSC. They consider that interdisciplinary catalogues look very ambitious but they are happy to build on their experience within the domain to keep legacy vocabularies and find ways to build bridges.

5.2 Pilot on implementation of (meta)data catalogues integration in B2FIND

To support the improved discoverability of (meta)data catalogues by aggregators such as **B2FIND** and to increase interdisciplinary reuse, a small-scale pilot will be carried out between late 2020 and mid-2021. The pilot will trial the use of **DCAT/DCAT-AP** and **DDI-CDI** for (meta)data catalogues within domain e-Infrastructures (ESFRI clusters projects described in 5.1) and repositories (to be analysed as M3.7 'Identification of candidates for testing proposal on integration of metadata catalogues'). It is also important to underline that DCAT and DDI-CDI were among those recommended standards in FAIRsFAIR D2.4: 2nd Report on FAIR requirements for persistence and interoperability[52].

As shown in Fig. 3 above, our **(meta)data catalogue** integration will be a service, harvesting, aggregating, storing and accessing descriptive metadata from five different disciplines. It will allow users/researchers, from different domains, to search for research datasets based on a particular intended attribute (metadata element) of the information that the user is looking for, or discovering datasets that the user was not aware of (serendipitous discovery). In both cases, if the dataset is openly available, eventually the user might have access to the data.

The integration service (Fig. 4) will be B2FIND. EUDAT's B2FIND is a discovery service based on metadata steadily harvested from research data collections from EUDAT data centres and other repositories, included in the EOSC marketplace[53]. B2FIND aims to provide a simple and user- friendly discovery portal. It is moving away from a disciplinary focus to be a more generic service covering all domains and as such must harvest and support various metadata schemas and standards. These are mapped to the B2FIND schema which is similar to DataCite but goes a step further by supporting cross-disciplinary search through the use of facets (Fig. 6). As part of the Freya project[54] DataCite is currently developing a discovery platform 'DataCite Commons'[55] which enables users to query for PID metadata (DOIs for datasets and articles, ORCIDs for people and RORs for organisations) and their connections.

EUDAT is trying to raise awareness about the FAIR principles and act as an ambassador. EUDAT/B2FIND are keen to provide community support on how to deal with metadata and to encourage repositories to map to B2FIND schema. B2FIND does not currently support DCAT nor DDI-CDI but is interested in working on this approach to add to the already supported domain-agnostic

[52] D2.4 2nd Report on FAIR requirements for persistence and interoperability (Version v1.0 draft): https://zenodo.org/record/4001630

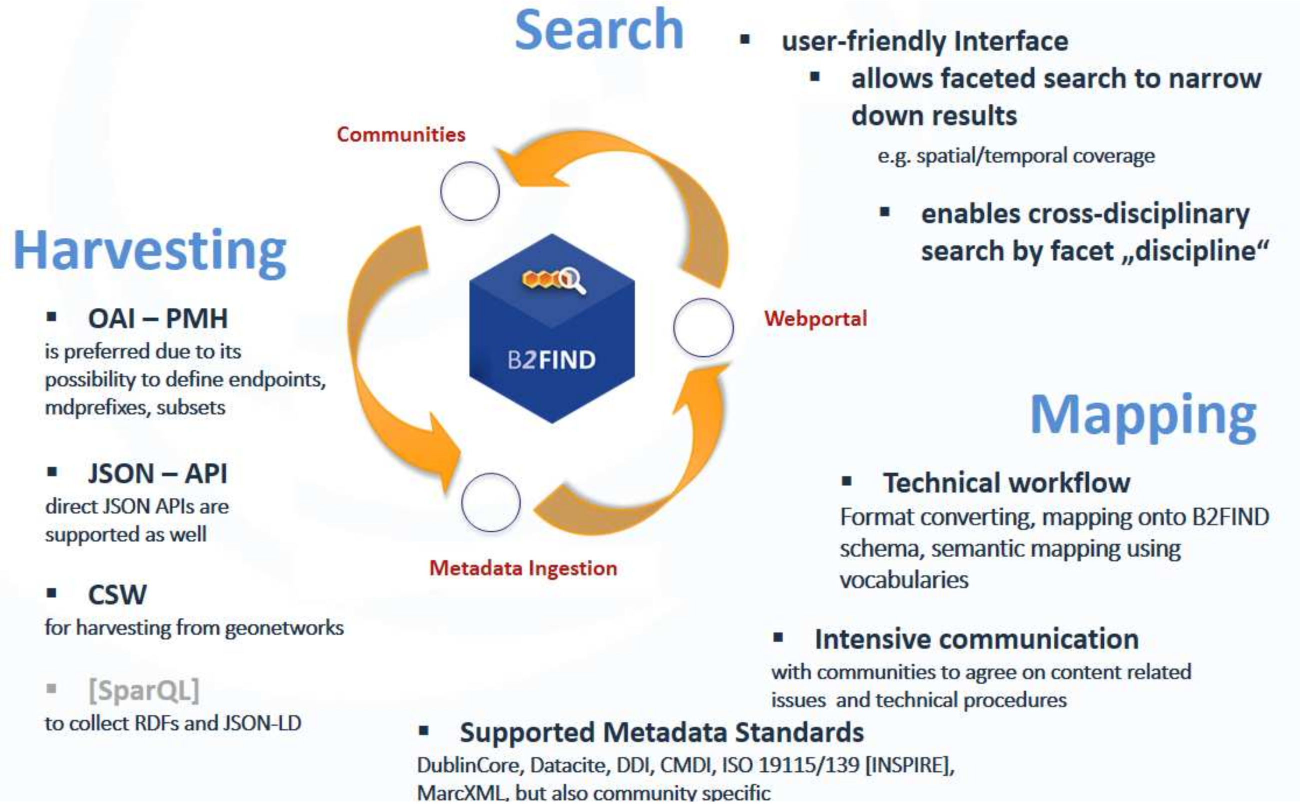
[53] https://marketplace.eosc-portal.eu/services/b2find

[54] https://www.project-freya.eu/en/about/mission

[55] https://commons.datacite.org/

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

metadata schemas (Fig. 6) (DCMI, DataCite schema, DDI, CMDI (used by CLARIN), ISO 19115 (INSPIRE), MarcXML, + some community specific standards).



Search

- user-friendly Interface
  - allows faceted search to narrow down results
    e.g. spatial/temporal coverage
  - enables cross-disciplinary search by facet „discipline“

Harvesting

- OAI – PMH
  is preferred due to its possibility to define endpoints, mdprefixes, subsets
- JSON – API
  direct JSON APIs are supported as well
- CSW
  for harvesting from geonetworks
- [SparQL]
  to collect RDFs and JSON-LD

Mapping

- Technical workflow
  Format converting, mapping onto B2FIND schema, semantic mapping using vocabularies
- Intensive communication
  with communities to agree on content related issues and technical procedures
- Supported Metadata Standards
  DublinCore, Datacite, DDI, CMDI, ISO 19115/139 [INSPIRE], MarcXML, but also community specific

Figure 6: Current B2FIND functionalities. Presentation at the virtual workshop Metadata catalogues integration for interdisciplinary research (by Anna-Lena Fügel, 11/09/2020)[56]

Along with the willingness of B2FIND to help and be involved in the implementation of this proposal, CODATA is starting an **EOSC co-creation project** entitled: “Applying DDI-CDI (Data Documentation Initiative-Cross Domain Integration) to the EOSC[57]”. The project will analyse the innovative capabilities of the new domain-agnostic DDI-DCI specification (currently in review, as we mentioned above) and apply them to needed search and data integration functions within the EOSC. The project will perform a structured consultation with European and international fora. The envisioned outputs of the project are: a refined profile of DDI-CDI to enable efficient reuse and discovery of data across disciplines; guidelines for the implementation of DDI-CDI in EOSC infrastructures and data sources; feedback to the DDI-CDI committee on any identified gaps. The project will be finished at the end of March 2021 and FAIRsFAIR will join its testing and implementation with the ESFRI cluster projects. DCATv2/DCAT-AP are also willing to be involved and one of the editors of the specification is involved in the metadata work of ExPANDs as well as involvement in FAIRsFAIR. Therefore, we think that a strong synergy can be guaranteed.

[56] https://drive.google.com/drive/u/1/folders/1md-lMPP2_sJLGTQE1TbGjjOhelYOomId

[57] https://www.eoscsecretariat.eu/funding-opportunities/list-approved-co-creation-activities

During the pilot, FAIRsFAIR, representatives of the ESFRI cluster projects, representatives from the two metadata models/standards (DCAT and DDI-CDI), and B2FIND will develop an assessment framework to enable a comparison of the two approaches from the domain perspective and also from the aggregator perspective. At least two virtual workshops will be held during the pilot to bring together the participants to discuss their experiences and to provide insights into the feasibility of the approaches. The results of the pilot will be shared in D3.7 'Report on integration of metadata catalogues' in August 2021 along with recommendations for wider adoption.

Phases of the pilot include:

PHASE1 (Q4-2020)
- Metadata elements mapping and crosswalk development. This task will include the mapping among the domain-agnostic metadata to DDI-CDI and DCATv2-DCAT-AP, as well as those with B2FIND metadata model. But may also include the analysis of the domain specific metadata schemas to DDI-CDI and DCATv2-DCAT-AP. This work will also engage in and look at other metadata mapping initiatives, and prior work[58] in this area - Identification of other possible candidates (repositories) for testing the proposal on integration of (meta)data catalogues (M3.7 Identification of candidates for testing proposal on integration of metadata catalogues, due November, 2020)
PHASE 2 (Q1-2021)
- Identification and agreement of use cases for the different participants (see annex 1) - Definition of the assessment framework for metadata catalogue integration based on the use cases
PHASE 3 (Q1-3 2021)
- Carry out metadata catalogue description against DCAT and DDI-CDI (Q1-2 2021) - Workshop 1 to review progress and feasibility (Q1 2021) - Workshop 2 to review progress and feasibility (Q2 2021) - D3.7 report on pilot and recommendations (Q3 2021)

[58] The previous metadata mapping works to follow, include:

- Crosswalks from schemas to schema.org (work lead by the RDA schemas WG): https://docs.google.com/spreadsheets/d/1P6WH8h4OnIVR9UJj3FcOebNUpLnKNBCuvEp3NsLRho4/edit#gid=1673841184 This work includes mapping for DCAT-AP, DCATv2, ISO19115, EOSC-EDI, Dataverse, DATS, RIF-CS, BioSchema, B2FIND, DDI ECRIN, CodeMeta, SPASE
- Research Data Analysis (work done in the context of Joinup Initiative of the European Commission): https://joinup.ec.europa.eu/sites/default/files/document/2018-05/Research%20Data%20Analysis_v1.00.pdf This work includes 1:1 mappings with DCAT-AP with DataCite schema, CERIF, schema.org, CSMD, and re3data metadata model (mapping from re3data only can be done at the class dcat:Catalog).

5.3Future steps: other disciplines, other thematic repositories and the long tail of data

The findings of this pilot will be presented in D3.7 which is due in August 2021 and, based on the outcomes, FAIRsFAIR will provide guidance to assist other repositories to improve the FAIRness of their data collections through integration of their metadata catalogues. There will potentially be at least, two ways to extend the work on metadata catalogue integration: 1) extending the number of repositories and RIs in the same chosen domains, or 2) extending the number of disciplines.

- Metadata integration from other repositories

Besides the great RIs that collect, manage and curate big data (ESFRIs, ERICS, etc.) there are also small thematic repositories where data and metadata are stored together and the metadata catalogue becomes a data catalogue. FAIRsFAIR is also thinking in the long-term integration in EOSC of the long tail data (repositories at different levels, that include data from different domains, or multi-domain). For example, just taking into account the five disciplinary domains of the EOSC clusters, we found the following repositories in re3data:

- Life Sciences: 436 repositories[59]
- Photon and Neutron: 2 repositories[60]
- Social Sciences and Humanities: 874 repositories[61]
- Environmental research: 490 repositories by searching environment*[62]
- Astrophysics and Astronomy: 10 repositories, but 182 on general search by "astronomy"[63]

- Metadata integration from other domains/projects.

Another way to amplify this proposal in the future would be including more domains or disciplines, for example Engineering, Material sciences, Computer Science, or Agriculture, to mention just a few. For example, projects like FNS-Cloud (Food and Nutrition Security Cloud[64]) are also interested in piloting our approach.

In the future, the integration of specific datasets from the selected domains (the five considered here, or others) could include datasets stored in generic or multidisciplinary repositories (e.g. Figshare, Zenodo, etc.) or even institutional repositories. Along with the domain-specific repositories, there are other multi-thematic or multidisciplinary repositories that might have content from these disciplines, but they use generic metadata standards (like DataCite schema or DCMI) to describe them. This super-low level of specificity in terms of metadata property catalogues, might be also

[59] https://www.re3data.org/search?query=&subjects%5B%5D=1%20Humanities%20and%20Social%20Sciences&subjects%5B%5D=2%20Life%20Sciences

[60] https://www.re3data.org/search?query=Photon+and+Neutron

[61] https://www.re3data.org/search?query=&subjects%5B%5D=1%20Humanities%20and%20Social%20Sciences

[62] https://www.re3data.org/search?query=environment*

[63] https://www.re3data.org/search?query=&subjects%5B%5D=1%20Humanities%20and%20Social%20Sciences&subjects%5B%5D=2%20Life%20Sciences&subjects%5B%5D=311%20Astrophysics%20and%20Astronomy

[64] https://www.fns-cloud.eu

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

willing to map to DDI-CDI, DCAT for a common aggregation service like B2FIND. The integration of all these repositories of the long-tail data is not yet addressed in EOSC, but it deserves mention in the interdisciplinary research scenario that we envisioned here.

Bibliography

Andias, W. A. (2014). Dublin Core Metadata for Research Data—Lessons Learned in a Real-World Scenario with datorium. *DC-2014--The Austin Proceedings*. DCMI International Conference on Dublin Core and Metadata Applications. https://dcpapers.dublincore.org/pubs/article/view/3703

Benjelloun, O., Chen, S., & Noy, N. (2020). Google Dataset Search by the Numbers. *arXiv:2006.06894 [cs]*. http://arxiv.org/abs/2006.06894

DataCite Metadata Working Group. (2019). *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data v4.3* [Application/pdf]. 73 pages. https://doi.org/10.14454/7XQ3-ZF69

Fischer, B. A., & Zigmond, M. J. (2010). The Essential Nature of Sharing in Science. *Science and Engineering Ethics*, *16*(4), 783-799. https://doi.org/10.1007/s11948-010-9239-x

Gómez, N.-D., Méndez, E., & Hernández-Pérez, T. (2016). Data and metadata research in the social sciences and humanities: An approach from data repositories in these disciplines. *El Profesional de la Información*, *25*(4), 545. https://doi.org/10.3145/epi.2016.jul.04

Gregg, W., Erdmann, C., Paglione, L., Schneider, J., & Dean, C. (2019). A literature review of scholarly communications metadata. *Research Ideas and Outcomes*, *5*, e38698. https://doi.org/10.3897/rio.5.e38698

Habermann, T. (2020). Metadata and Reuse: Antidotes to Information Entropy. *Patterns*, *1*(1), 100004. https://doi.org/10.1016/j.patter.2020.100004

Martin, P., Remy, L., Theodoridou, M., Jeffery, K., & Zhao, Z. (2019). Mapping heterogeneous research infrastructure metadata into a unified catalogue for use in a generic virtual research environment. *Future Generation Computer Systems*, *101*, 1-13. https://doi.org/10.1016/j.future.2019.05.076

Méndez, E., & van Hooland, S. (2014). Metadata typology and Metadata uses. In: M.-A. Sicilia, *Handbook of Metadata, Semantics and Ontologies* (pp. 9-39). World Scientific. https://doi.org/10.1142/9789812836304_0002

Noy, N. (2020). *Google Dataset Search—Building an open ecosystem for dataset discovery*. https://science.unimelb.edu.au/mcds/events/mcds-seminar-series-aug2020

Noy, N., Burgess, M., & Brickley, D. (2019). Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. *28th Web Conference (WebConf 2019)*. https://ai.google/research/pubs/pub47845

Quimbert, E., Jeffery, K., Martens, C., Martin, P., & Zhao, Z. (2020). Data Cataloguing. En Z. Zhao & M. Hellström (Eds.), *Towards Interoperable Research Infrastructures for Environmental and Earth*

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Sciences (Vol. 12003, pp. 140-161). Springer International Publishing. https://doi.org/10.1007/978-3-030-52829-4_8

Schriml, L. M., Chuvochina, M., Davies, N., Eloe-Fadrosh, E. A., Finn, R. D., Hugenholtz, P., Hunter, C. I., Hurwitz, B. L., Kyrpides, N. C., Meyer, F., Mizrachi, I. K., Sansone, S.-A., Sutton, G., Tighe, S., & Walls, R. (2020). COVID-19 pandemic reveals the peril of ignoring metadata standards. *Scientific Data*, *7*(1), 188. https://doi.org/10.1038/s41597-020-0524-5

Shaw, F., Etuk, A., Minotto, A., Gonzalez-Beltran, A., Johnson, D., Rocca-Serra, P., Laporte, M.-A., Arnaud, E., Devare, M., Kersey, P., Sansone, S.-A., & Davey, R. P. (2020). COPO: A metadata platform for brokering FAIR data in the life sciences. *F1000Research*, *9*, 495. https://doi.org/10.12688/f1000research.23889.1

Smith-Yoshimura, K. (2020). *Transitioning to the Next Generation of Metadata.* https://doi.org/10.25333/RQGD-B343

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018.

Wilsdon, J., Bar-Ilan, J., Frodeman, R., Lex, E., Peters, I., & Wouters, P. (2017). *Next-generation metrics: Responsible metrics and evaluation for open science.* Publications Office. https://data.europa.eu/doi/10.2777/337729

Annex 1: Draft use cases for pilot metadata catalogue integration

These use cases were suggested during the 2nd workshop mentioned in the methodology, held online on October, 9 2020[65].

Project	Example domain metadata catalogue	Cross Domain Use Case
EOSCLife (Parkinson)	FAIRSharing, using schema.org standard OmicsDI - using schema.org standard	Example: In order to model Covid19 infections across a geographical region we need to connect clinical datasets indicating infection rates to datasets which contain geospatial and population information. The search query: 'Find clinical datasets for Covid19 aggregate datasets for <some defined region> and find geospatial and population information for <some defined region>
SSHOC		Example: In order to study social recovery, and to anticipate the "new normal", we would need to connect survey data with clinical datasets about infections and also with geospatial data, social media data or text corpora, and official statistics.
PaNOSC ExPaNDS	ICAT / SciCat	Potential example by Alejandra (to be checked with the consortium): integrating data from the SARS-COV2 protease structure and how the different strains are represented in the different populations across the world

[65] See: https://docs.google.com/document/d/1QT5wB7fQYebxBfGyjouu_bCRt0IYXxsSWTRgEBrIo90/edit