D3.5 Data Marketplaces with Interoperability Solutions II

Authors: **Stefan Gindl (RSA), Michael Boch (RSA), Alan Barnett (EMC), George Margetis (FORTH), Victor Mireles (SWC)**

Additional Information: **Follow-up of D3.4**

December 2021

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 871481

TRUSTS Trusted Secure Data Sharing Space

D3.5 Data Marketplaces with Interoperability Solutions II

Document Summary Information

Grant Agreement No	871481	Acronym	TRUSTS
Full Title	TRUSTS Trusted Secure Data Sharing Space		
Start Date	01/01/2020	Duration	36 months
Project URL	https://trusts-data.eu/		
Deliverable	D3.5 Data Marketplaces with Interoperability Solutions II		
Work Package	WP3		
Contractual due date	31/12/2021	Actual submission date	17/12/2021
Nature	Report	Dissemination Level	Public
Lead Beneficiary	RSA		
Responsible Author	Stefan Gindl (RSA)		
Contributions from	Victor Mireles (SWC), George Margetis (FORTH)		

Revision history (including peer reviewing & quality control)

Version	Issue Date	% Complete	Changes	Contributor(s)
v0.1	Nov 19, 2021	10	Initial deliverable structure	Stefan Gindl (RSA)
v0.2	Dec 03, 2021	60	1st version	Stefan Gindl (RSA)
v0.3	Dec 03, 2021	70	Content update	Victor Mireles (SWC)
v0.4	Dec 07, 2021	75	2nd version	George Margetis (FORTH)
v0.5	Dec 14, 2021	92	Task-internal review	George Margetis (FORTH)
v0.6	Dec 15, 2021	94	TRUSTS-internal review	Gert Breitfuss (KNOW)
v0.7	Dec 15, 2021	96	TRUSTS-internal review	Benjamin Heitmann, Kim Fidomski (FHG)
v1.0	Dec 16, 2021	100	Final version	Stefan Gindl (RSA)

Disclaimer

Copyright message

Table of Contents

List of Figures

List of Tables

Glossary of terms and abbreviations used

Abbreviation / Term	Description
AAI	Authentication and Authorization Infrastructure
CKAN	Comprehensive Knowledge Archive Network
DMP	Data management platform
DSC	Dataspace Connector
EDMI	EOSC Datasets Minimum Information
EIF	European Interoperability Framework
EOSC	European Open Science Cloud
FAIR	Findability, Accessibility, Interoperability, and Reusability
IDS	International Data Spaces
IDS-IM	International Data Spaces - Information Model
PID	Persistent Identifier

1 Executive Summary

This report summarizes the efforts taken with regards to TRUSTS interoperability in Task 3.3 "Data marketplaces interoperability solutions". It continues from the previous deliverable "D3.4 Data Marketplaces with Interoperability Solutions I".

The subject of the deliverable: The goal of Task 3.3 is *"analysing existing interfaces and standards, and even developing new relevant standards"*[1] to design the interoperability solution of TRUSTS[2]. This covers interoperability with existing industrial data markets and the EOSC[3]. The interoperability solution has to be either flexible enough to cater for the needs of both, i.e., data markets and EOSC, or there are two (or more) interoperability solutions, one for each of the requirements.

Summary of the work carried out: We continue the work accomplished for and described in the previous deliverable, "D3.4 Data Marketplaces with Interoperability Solutions I". The work accomplished for this deliverable included experimentation with interfaces, i.e., APIs of existing data markets, research of components of the IDSA Reference Architecture Model[4], which is one of the backbones of TRUSTS, as well as conducting a survey with data market operators to learn more about the technical background of data markets and better understand how to design and conceptualize the interoperability solution. Given the low response rate of the survey, we decided to adopt a new strategy, which we describe in the present deliverable. We conducted a systematic review of existing data management platforms, i.e. software tools usable to set up a data repository within an organization. Such software tools can also be used to create data markets. TRUSTS, for instance, uses CKAN, a widely used open-source data management platform at its core.[5] The logic behind our review is that existing data markets might adopt a similar strategy, i.e., use an existing, mature, and openly available software tool as the core for its services. Our systematic review culminated in the creation of a generic criteria catalogue, applicable to assess the feature-richness of data management platforms, and an assessment of seven selected data management platforms using exactly this criteria catalogue.

Furthermore, we also researched that vast ecosystem of EOSC[6]. EOSC is not a singular platform or project. In contrast, it is an initiative spanning over several research projects and many years of research and development. We examined prominent EOSC projects and focused especially on their efforts regarding interoperability. This, as is the goal of Task 3.3, helps to inspire the design and conceptualization of a solution for interoperability with EOSC.

The main conclusion(s): Interoperability, even if applied to a seemingly narrow field of "only" data markets and the EOSC, turns out to be a task exhibiting many details, differences, and changing requirements. It is hard to conceptualize a universal interoperability solution catering for the needs of all the diverse platforms, as this would require a highly flexible and versatile architecture, providing a plethora of differing technical solutions. This would ultimately be infeasible in terms of available

[1] TRUSTS Trusted Secure Data Sharing Space grant agreement.

[2] *ibid.*

[3] *ibid.*

[4] IDSA Reference Architecture Model https://internationaldataspaces.org/wp-content/uploads/IDS-Reference-Architecture-Model-3.0-2019.pdf, accessed Dec 13, 2021.

[5] CKAN: https://ckan.org/, Dec 13, 2021.

[6] EOSC: https://eosc.eu/, accessed Dec 13, 2021.

resources and further become a threat for future maintenance of the platform, especially beyond project lifetime. Thus, it seems more beneficial to provide an outwards-looking interface for TRUSTS, that allows external initiatives to connect and exchange data assets with TRUSTS under the conditions of the technical requirements of TRUSTS, e.g., the used metadata model, which is based on the IDS-IM[7]. Such a solution is easier to maintain, as it requires external stakeholders to set up an API according to the needs of TRUSTS.

Furthermore, TRUSTS can learn from interoperability approaches on an organisational and legal level and adopt similar strategies. This knowledge is valuable for other WPs and Tasks within TRUSTS, which are not necessarily purely technical, as is the case for Task 3.3.

The purpose of the deliverable: This deliverable describes and legitimates the systematic review of data management platforms, which is crucial insight for conceptualizing an interoperability solution. It also describes our research related to the EOSC. The insights gained through this research path inspires the last phase of Task 3.3, which is the design of TRUSTS interoperability solution[8].

[7] IDS Information Model: https://internationaldataspaces.org/wp-content/uploads/IDS-Reference-Architecture-Model-3.0-2019.pdf, accessed Dec 13, 2021.

[8] TRUSTS Trusted Secure Data Sharing Space grant agreement.

2 Introduction

One major aspect and goal of TRUSTS is to establish interoperability with third-party initiatives, which is the objective of TRUSTS Task 3.3 “Data marketplaces interoperability solutions”. The third-party initiatives include existing industrial data markets as well as the European Open Science Cloud (EOSC). Interoperability is an increasingly relevant topic in the context of a European data-driven, such as GAIA-X[9]. It has received significant attention from the European Commission, which provides the “New European Interoperability Framework (EIF)”[10]. The EIF defines four layers of interoperability, i.e. the technical, semantic, organisational, and legal layer. Adherence to these four layers is crucial to ensure interoperability. In the context of the EU, interoperability is especially important for public services and how external sources and their information sources and services get integrated into the catalogues of public services (see Figure 1 for an overview).

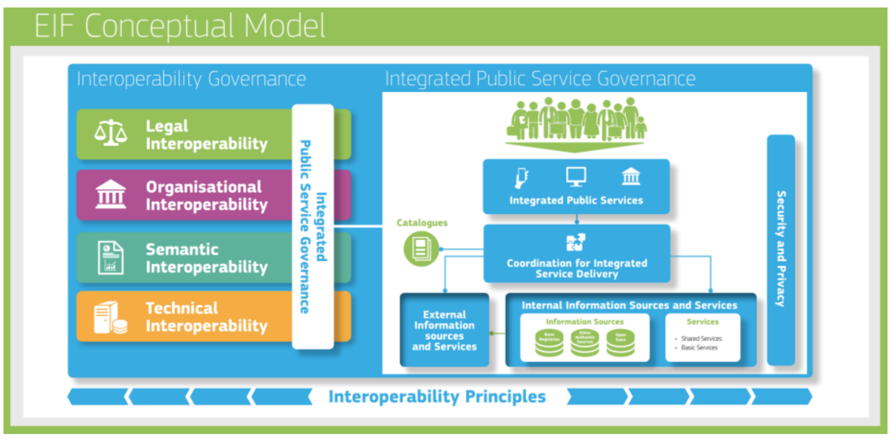

Figure 1: The EIF model (The New European Interoperability Framework[11], p. 44).

From the perspective of business and industry, a data-driven economy with a wide variety and diversity of stakeholders participating in it facilitates innovation and the invention of new and innovative business models. From the perspective of science and research, especially open science, such efforts facilitate collaboration and interactivity between research efforts, which otherwise run into the risk of becoming single, independent, and unconnected islands. This increases the chances of duplication of existing research efforts, since existing work might be unknown or hard to access. Combatting this risk by creating a lively and widely accepted ecosystem, where research institutions, academia, and organization, can effortlessly exchange their data, datasets, insights, service, etc., will increase cross-pollination of ideas and the reuse of existing results, consequently accelerating research efforts. This is especially important in an era where a global pandemic calls for strong and effortless collaboration.

In our work, we looked at both the sector of the data markets and the EOSC. We systematically examined software tools for data management to learn and understand the technical requirements

[9] GAIA-X: https://www.gaia-x.eu/, Dec 13, 2021.
[10] The New European Interoperability Framework: https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf, accessed Dec 14, 2021.
[11] *ibid.*

relevant to such platforms. Furthermore, we examined the material of relevant EOSC-related projects, to understand the technical requirements, but also organisational processes that are important to establish interoperability. In the deliverable, we describe and summarize the examined material. For each major topic, we provide lessons learned, where we summarize the major insights. From the lessons learned, we derive a set of next steps. The next steps will be translated into a strategy of actions to be followed in the remaining time of Task 3.3.

2.1 Mapping Projects' Outputs

The purpose of this section is to map TRUSTS Grand Agreement commitments, both within the formal Deliverable and Task description, against the project's respective outputs and work performed.

Table 1: Adherence to TRUSTS GA Deliverable & Tasks Descriptions

TRUSTS Task		**Respective Document Chapter(s)**	**Justification**
T3.3 Data marketplaces interoperability solutions	Based on the findings of D2.1: Definition and analysis of the EU and worldwide data market trends and industrial needs for growth, and by analysing existing interfaces and standards, and even developing new relevant standards (see T7.4 Standardisation), the interoperability solution for TRUSTS will be designed in this task. This means the definition of interfaces to ensure interoperability with other industrial data marketplaces. In addition interoperability solutions with the European Open Science Cloud (EOSC) will be evaluated and implemented where possible. Thereby this task has strong interdependencies with T3.2 Smart Contracts, T3.4 Data Governance & Metadata and the overall Work Packages: WP7 Business Plan and WP6 Legal Framework, to ensure interoperability solutions are reflected technically, legally and business wise.	Section 3 Section 4	Understand the technical requirements for interoperability with existing data markets. Understand the requirements to conceptualize an interoperability solution that can connect to the EOSC.
TRUSTS Deliverable			
D3.5 Data Marketplaces with Interoperability Solutions II The second version of a series of three deliverables (D3.4, D3.5, D3.6) will summarize the integration requirements as well as guidelines for both the TRUSTS Platform to interact with existing platforms, the EOSC, and for future platforms to integrate with TRUSTS.			

2.2 Deliverable Overview and Report Structure

The present deliverable is divided into four sections. The first two represent the two major aspects of Task 3.3 “Data marketplaces interoperability solutions”, i.e. the research and conceptualization of means of interoperability with (i) existing industrial data markets and (ii) the EOSC. In the following, we describe the deliverable structure in more detail:

- **Section 3 “Review of Data Management Platforms”** summarizes our efforts related to the former aspect, i.e. interoperability with existing industrial data markets. We describe how we proceeded since the last deliverable (“D3.4 Data Marketplaces with Interoperability Solutions I”). The section summarizes the systematic review of data management platforms. The seven assessed platforms are mature, feature-rich, free-to-use platforms. They could be used by data markets as the basic infrastructure for indexing and provision of data resources. Knowing how interoperability is possible with those platforms is beneficial to understand potential interoperability requirements of existing data markets. This research is a follow up of our previous efforts, which included the creation of technical prototypes to connect to interfaces (APIs) of online data markets, and the conduction of an online survey for data market operators, which was unsuccessful[12]. It closes the gaps to a better understanding of the technical basis of data markets, which is useful for the conceptualization of the interoperability platforms envisaged in Task 3.3
- **Section 4 “EOSC”** describes our research of EOSC, the European Open Science Cloud[13]. The EOSC is a considerable, EU-wide effort to make science more open and facilitate cross-pollination and innovation between research projects. We examine relevant EOSC projects regarding interoperability. The goal is to derive opportunities for interoperability with TRUSTS, i.e., to make accessible the offers hosted on EOSC within TRUSTS, or, vice versa, to make available TRUSTS’ data assets from within the EOSC. The results and insight gathered at this stage of our research will help us to better align the interoperability solution of Task 3.3 with the actual requirements of data markets and the EOSC. From there, we plan to conceptualize and implement interfaces for interoperability. The subsequent and last deliverable “D3.6 Data Marketplaces with Interoperability Solution III” of Task 3.3, upcoming in June 2022, will report about these efforts.
- **Section 5 “Exchanging Data with TRUSTS”** covers suggestions for the exchange of data with TRUSTS. We suggest an interface that is universal and flexible enough to be adopted by external initiatives for data exchange. It is tailored to the technical requirements of TRUSTS, i.e., it leverages the IDS-IM, and foresees functionality to connect using TRUSTS technical features, e.g. the Dataspace Connector and the CKAN harvester.
- **Section 6 “Conclusions & Next Actions”** summarizes this report, gives conclusions and outlines the future path of this task.

[12] TRUSTS D3.4 Data Marketplaces with Interoperability Solutions I: https://www.trusts-data.eu/wp-content/uploads/2021/01/D3.4-Data-Marketplaces-with-Interoperability-Solutions-I.pdf, accessed Dec 12, 2021.
[13] EOSC: https://eosc.eu/, accessed Dec 12, 2021.

3 Review of Data Management Platforms

The first major target of the TRUSTS interoperability solution is existing industrial data markets. TRUSTS T3.3 "Data marketplaces interoperability solutions" states, the task comprises

> *"[...] the definition of interfaces to ensure interoperability with other industrial data marketplaces.[14]"*

As further mentioned in the task description, this also comprises *"analysing existing interfaces and standards, and even developing new relevant standards"[15]*. To our understanding, an appropriate analysis as required in the task thus includes obtaining an understanding of the technical basis and requirements. This helps to lay the basis for technical prototypes and mature functionalities to be included in TRUSTS, which allow interoperation with those data markets. In TRUSTS "D3.4 Data Marketplaces with Interoperability Solutions I"[16] we describe our efforts in this respect.

We analyzed available interfaces and created software prototypes to retrieve data from a selected set of data markets (see D3.4[17], Section 3.1.2). This analysis showed the significant diversity among data markets in their interfaces. Also, the tackled data markets often do not deliver datasets. Instead, they provide an API to interact with their services. For example, HERE[18], lets users send queries with geo-coordinates and keywords to retrieve points of interest close to the location specified by the geo-coordinates. In the example in D3.4, Section 3.1.2, we show a request with longitude and latitude as well as the keyword restaurant, to retrieve names and addresses of restaurants close to the given location.

To learn more about the data markets, we decided to pursue another research path and attempted to do a survey among data market operators. We set up an online survey, consisting of eight general questions, 14 technical questions, and six business-related questions. The survey was accomplished in conjunction with partners from TRUSTS WP5. Despite continuous reminders to data market operators, we did not achieve a satisfying response rate. Thus, we decided to study already existing platforms, which could be used as the basis to build data markets. For example, the architecture of TRUSTS is built on top of CKAN[19] (for a detailed description see TRUSTS deliverable D2.6[20]), a widely used open-source data management platform. The rationale behind using CKAN is to use a ready-built, mature solution, providing functionality for the essential features of TRUSTS and thus solve more "mundane", less research-driven requirements. This includes, for example, user management, or the core infrastructure to manage datasets, i.e. to upload and store them or make them searchable. CKAN has such functionality built-in. For instance, dataset management is accomplished via a combination of a

[14] TRUSTS Trusted Secure Data Sharing Space grant agreement.

[15] *ibid.*

[16] TRUSTS "D3.4 Data Marketplaces with Interoperability Solutions I": https://www.trusts-data.eu/deliverables/, accessed Dec 12, 2021.

[17] *ibid.*

[18] HERE Technologies: https://www.here.com/, accessed Dec 12, 2021.

[19] CKAN: https://ckan.org/, accessed Dec 12, 2021.

[20] TRUSTS D2.6 Architecture design and technical specifications document I: https://www.trusts-data.eu/wp-content/uploads/2021/09/D2.6_Architecture-design-and-technical-specifications-document-I.pdf, accessed Dec 14, 2021.

PostgreSQL database[21], the search engine Apache SolR[22], and the key-value store Redis[23], served through a web application based.

In the following subsections, we describe these efforts. We submitted the conference article *"A systematic review of data management platforms"* for this work to WorldCist'22, the "*10th World Conference on Information Systems and Technologies"*[24], authored in conjunction by the TRUSTS project partners EMC, FORTH, RSA, and SWC. At the time of the writing of this deliverable, the acceptance of the paper has not been announced yet.

An existing work by Amorim et al. [1, 2] served as the starting point for our own research. The authors of the article defined a criteria catalogue to assess data management platforms and applied it to a set of selected platforms. We extended this criteria catalogue with additional criteria on the one hand, and applied it to a different set of platforms on the other hand.

In the following subsection, we describe the scientific framework we used for our work (design science research), the methodology, and the outcome. We give a shortened description of the article submitted to WorldCIST'22. This work was substantially driven by the desire to learn more about opportunities for interoperability with existing data markets, and tackles the problem from the side of potential core technologies, i.e., the data management platforms.

3.1 Approach

We followed a design science research approach, depicted in Figure 2, as the methodological framework of our research. This approach divides the study into the components "Problem identification & motivation", "Objectives of a solution", "Design & development", "Demonstration", "Evaluation", and "Communication", which we translated into the following five consecutive steps:

1. **Literature research:** the study of existing literature helped to identify existing, similar or related, artefacts, used for the classification of data management platforms. This resulted in the creation of a first draft of the criteria catalogue, assembled from existing catalogues or criteria collections.
2. **Initial expert round:** a sub-group of the article authors manually assessed the criteria catalogue and extended it by new criteria, adapted existing, or removed unnecessary.
3. **First manual round:** one author worked as an assessor and applied the catalogue to CKAN. This step helped to assess the validity of the selected criteria and identify criteria, that are too hard to assess or unnecessary.
4. **Extended expert round:** this step was like step 2, the "Initial expert round", but was accomplished with a larger group of experts (i.e., the article authors except for one author). This helped to further clean the catalogue.
5. **Final manual round:** the assessors, i.e., article authors, applied the criteria catalogue to all data management platforms selected for assessment. Each platform was assessed by at least two assessors to strengthen the validity and unveil criteria causing different results.

[21] PostgreSQL: https://www.postgresql.org/, accessed Dec 12, 2021.

[22] Apache Solr: https://solr.apache.org/, accessed Dec 12, 2021.

[23] Redis: https://redis.io/, accessed Dec 12, 2021.

[24] WorldCIST'22: http://worldcist.org/#:~:text=Welcome%20to%20WorldCist'22%20%2D%2010th,%2C%2012%20%2D%2014%20April%202022., accessed Dec 12, 2021.

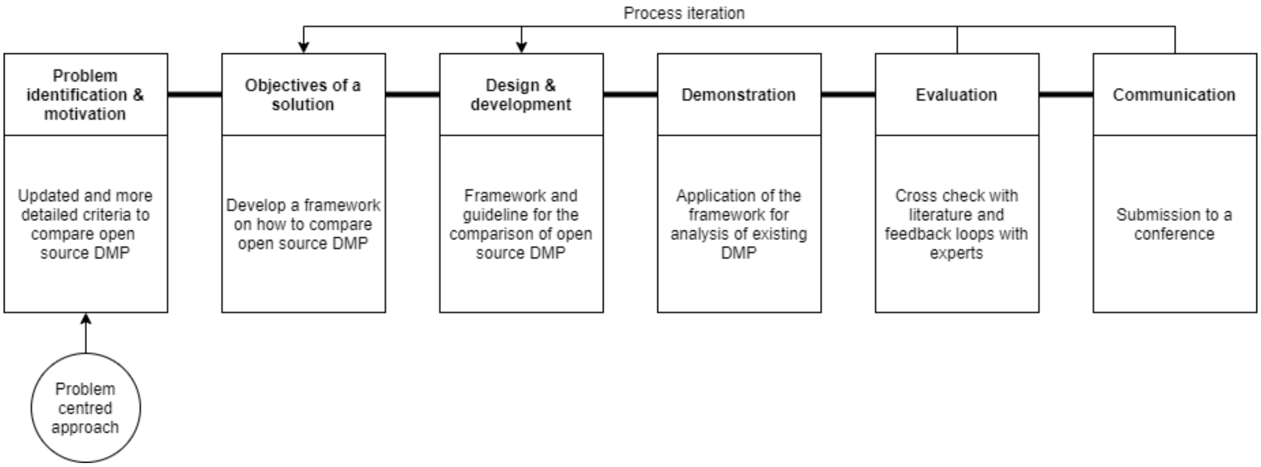

Figure 2: Overview of the design science research approach.

We applied the criteria catalogue to seven data management platforms, which were selected because of their open license and their wide utilization. Table 1 lists the platforms and gives a short description of them.

Table 1: The list of data management platforms.

Platform	Description
CKAN[25]	A data management platform used extensively by governments for publishing open data. It is the basis of many national data portals, and it is supported by an ecosystem of extensions that build upon its modular architecture. It provides a UI, a Database and document indexer, along with models abstracting data into python objects which form the basis of the extension mechanisms.
Dataverse[26]	A system used by many academic institutions for depositing documents and data resulting from research activities. Issues permanent identifiers to assets, as it is envisioned for preservation of openly-accessible assets. Offers limited support for metadata schemas and vocabularies.
DSpace[27]	A system for data and document institutional repositories with thousands of installations world-wide. In its latest iteration, it is one of the most technologically modern repository software. Complies with industry standards, and is customizable through extensions.
ePrints[28]	ePrints calls itself *"the world-leading open-source digital repository platform"*[29]. ePrints is available as a hosted service[30] and can be installed on-premises[31]. The

[25] https://ckan.org/, accessed Dec 10, 2021.
[26] https://dataverse.org/, accessed Dec 10, 2021.
[27] https://www.dspace.com/de/gmb/home.cfm, accessed Dec 10, 2021.
[28] https://www.eprints.org/uk/, accessed Dec 10, 2021.
[29] https://www.eprints.org/uk/index.php/about/, accessed Dec 10, 2021.
[30] https://www.eprints.org/uk/index.php/services/repository-hosting-0/, accessed Dec 10, 2021.

	services are available for a variety of purposes, e.g. for research data, for dataset showcases, social media, etc.[32]
Fedora[33]	Fedora (Flexible Extensible Digital Object Repository Architecture) is a *"robust, modular, open source repository system for the management and dissemination of digital content"*[34]. Founded in 1997, it is used by research organizations and academic institutions to host collections of research data for a variety of purposes, e.g. historical or cultural data[35].
InvenioRDM[36]	InvenioRDM is a "turn-key" solution to build "research data management repositories"[37] and based on the modular Invenio Framework[38] and Zenodo[39]. InvenioRDM aims to be scalable, flexible to changing business needs[40].
Omeka[41]	Omeka is a data publishing platform targeting mainly institutions interested in connecting digital cultural heritage collections with other resources online. It provides a full fledged dataset management system, supporting amongst others publishing items with linked open data, open vocabularies and item descriptions based on DPLA-ready resource templates, etc. Furthermore, it supports different fully-responsive themes to fit any screen size. Omeka is based on a modular architecture, employing a diversity of plugins for mapping, collecting, importing and connecting with external resources.

3.2 Results

We produced two results: (i) a criteria catalogue containing criteria useful to assess data management platforms, and (ii) an assessment of the seven mentioned data management platforms using this criteria catalogue. The criteria catalogue itself is useful for further research in that area, and can be used to continuously (re-)evaluate data management platforms, or even be extended.

Table 2 shows the assessment of the seven data management platforms using the criteria from the catalogue. The table does not show the full assessment, only an interoperability-relevant excerpt. The reason is that the full assessment is included in the aforementioned article, which has not been accepted for publication yet. The full assessment includes criteria such as the licence, the programming language, visualization, etc.

[31] http://files.eprints.org/, accessed Dec 10, 2021.
[32] https://www.eprints.org/uk/, accessed Dec 10, 2021.
[33] https://duraspace.org/fedora/, accessed Dec 10, 2021.
[34] https://duraspace.org/fedora/about/, accessed Dec 10, 2021.
[35] https://duraspace.org/fedora/about/, accessed Dec 10, 2021.
[36] https://inveniosoftware.org/products/rdm/, accessed Dec 10, 2021.
[37] *ibid.*
[38] https://inveniosoftware.org/products/framework/, accessed Dec 10, 2021.
[39] https://www.zenodo.org/, accessed Dec 10, 2021.
[40] https://inveniosoftware.org/products/rdm/, accessed Dec 10, 2021.
[41] https://omeka.org/, accessed Dec 10, 2021.

Table 2: An excerpt of the assessment of seven data management platforms using the criteria catalogue.

	Feature	CKAN	Dataverse	DSpace	ePrints	Fedora	InvenioRDM	Omeka
1	Ecosystem of extensions	✓	✓	✖	✓	✖	✖	✓
2	Internationalization support	✓	✓	✓	?	✖	✖	✓
3	Multi factor authentication	✓	✖	✓	✖	✓	✖	✖
4	Authorization (access control)	✓	✓	✓	✓	✓	✓	✓
5	Federated identity/Single sign-on	?	✓	✓	✓	✓	✓	✖
6	Review system	✖	✖	✖	✖	✖	✖	✖
7	User management and community features	✓	✖	✖	~	✖	✖	✓
8	Exporting schemas	✖	✓	✓	✓	✓	✓	✖
9	Established ontologies	?	✓	✓	✓	✓	✓	✓
10	Customized ontologies	~	✖	✓	✖	✖	✖	?
11	Domain-specific vocabularies	✓	?	✓	✖	✓	✓	✓
12	Customized vocabularies	~	?	✓	✖	✖	✓	✓
13	Data provenance management	?	✖	✓	✖	✖	✖	✖
14	Diverse data types	✓	✓	✓	✓	✓	✓	✓
15	Exporting and serialisation	✓	?	✖	✓	✓	✖	✓
16	Data services	✖	✖	✖	✖	✓	✓	✖
17	Unstructured data	✓	✓	✓	✓	✓	✓	✓
18	Harvesting / Crawling of datasets	✓		✓	✓	✖	✖	?
19	Batch access API	?	✓	✓	✖	✓	✖	✖
20	Transactional access API	✓	✓	✓	✓	✓	✓	✓
21	Analytical access API	?	?	✓	✖	✓	✓	✖

Lessons learned:

- **DMP Feature Diversity:** The data management platform strongly diverges in its features. Some features are common, such as the availability of transactional APIs, faceted search, or the usage of diverse data types.
- **DMP Ontologies:** Data management platforms use their own ontologies. A flexible metadata model will be beneficial for TRUSTS in case of future attempts to connect to data markets based on these platforms.

4 EOSC

The second of the two core goals of TRUSTS T3.3 "Data marketplaces interoperability solutions" is the investigation of potential ways to interoperate with the EOSC. The task description has the following statement in that respect:

> *"[...] interoperability solutions with the European Open Science Cloud (EOSC) will be evaluated and implemented where possible."*[42]

In the following, we describe our approach to assess ways of interoperation with the EOSC. We start with an introduction of EOSC and subsequently describe how interoperability was envisaged by the EOSC and how TRUSTS can leverage the existing approaches. Studying the EOSC with respect to interoperability will help to develop interfaces for TRUSTS to connect with EOSC and ideally become part of the EOSC ecosystem. Mutual exchange of data assets will be beneficial for both initiatives. We also describe procedures, e.g. for onboarding services, which are potentially relevant in the context of TRUSTS.

4.1 Introduction

It is not uncommon that research institutions and initiatives create their own procedural, organisational, managerial, legal, and technical stack of processes, tailored to their specific needs. This adaptation to individual requirements and goals can create a landscape of islands, where everyone works on their tasks in isolation, unaware of their surrounding vicinity. Cross-fertilization of ideas and innovations, sharing of insights, knowledge, expertise, and tangible assets such as datasets or services, is exacerbated, and sometimes even becomes impossible. The velocity of progress and innovation gets reduced. This fact calls for a solution allowing the flexible interchange of scientific stakeholders, giving them the infrastructure, tools, and procedures to facilitate active exchange. The EOSC has the goal to tackle exactly this challenge.

The EOSC (European Open Science Cloud) is an effort by the European Commission to create a functioning and lively ecosystem for researchers and scientists to exchange resources, research outcomes, data, demonstrators, etc. EOSC strives to become a system of systems, where the manifold and diverse systems of participants are connected, and become accessible and usable.

Purpose of EOSC[43]

- *"Seamless access to content and services via common AAI (EOSC Authentication and Authorization Infrastructure),"*
- *"Access to data from various sources which is FAIR and ideally open,"*
- *"Access to services for storage, computation, analysis, preservation and more,"*
- *"Adoption of standards so data and services can be combined,"*
- *"Helpdesk, training and support to improve the use of EOSC."*

[42] TRUSTS Trusted Secure Data Sharing Space grant agreement.

[43] EOSC - What the European Open Science Cloud is: https://www.eosc.eu/about-eosc, accessed Dec 6, 2021.

The European Commission initiated the EOSC to improve collaboration and exchange of scientists within Europe and globally. It finances the EOSC via a sequence of publicly funded, inter-related projects (see Figure 3 for a timeline). To facilitate the achievement and strengthen this vision, it included EOSC into the **"new European Research Area"**[44] and the **"European strategy for data"**[45], aiming to create a European single data market. The **"European strategy for data"** is one of the ten *actions* in **"A Europe fit for the digital age"**, which is in turn

one of the six *priorities* of the **European Commission for the years 2019 - 2024**. EOSC is also part of the

EU open science initiative[46], a set of eight priorities of the EU related to its open science policy[47] ("Open Data", "EOSC", "New generation metrics", "Future of scholarly communication", "Rewards", "Research integrity & reproducibility of scientific results", "Education and skills", "Citizen science"). On July 29, 2020, the EOSC Association[48] was founded as the legal entity to direct the future of the EOSC and currently has 13 task forces as well as 200 members and observers.

> *"Definition: the European Open Science Cloud promoted by the European Commission to provide all researchers, innovators, companies and citizens with seamless access to an **open-by-default**, efficient and cross-disciplinary environment for storing, accessing, reusing data, tools, publications and any EOSC Resource for research, innovation and educational purposes. EOSC is implemented by the EOSC System and governed by the EOSC Governance."* **EOSC Glossary V1.0 (Feb. '19)[49].**

> *"The ambition of the European Open Science Cloud (EOSC) is to develop "Web of **FAIR Data** and services' for science in Europe. EOSC will be a multi-disciplinary environment where researchers can publish, find and re-use data, tools and services, enabling them to better conduct their work."* **EOSC Association[50].**

> *"The ambition of the European Open Science Cloud (EOSC) is to provide European researchers, innovators, companies and citizens with a federated and **open** multi-disciplinary environment where they can publish, find and re-use data, tools and services for research, innovation and educational purposes."* **European Commission[51].**

As mentioned in the quotes above, data FAIRness is a special focus of the EOSC. The FAIR principles[52] (FAIR - Findability, Accessibility, Interoperability, and Reusability) for exchange of

---

[44] The New European Research Area: https://data.consilium.europa.eu/doc/document/ST-13567-2020-INIT/en/pdf, accessed Dec 6, 2021.

[45] A European strategy for data: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1593073685620&uri=CELEX%3A52020DC0066, accessed Dec 6, 2021.

[46] The EU's open science policy: https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science_en, accessed Dec 6, 2021.

[47] Open Science: https://ec.europa.eu/info/sites/default/files/research_and_innovation/knowledge_publications_tools_and_data/documents/ec_rtd_factsheet-open-science_2019.pdf, last accessed Dec 6, 2021.

[48] https://www.eosc.eu/, accessed Dec 6, 2021.

[49] EOSC Glossary: https://eoscpilot.eu/eosc-glossary#overlay-context=eosc-glossary, accessed Dec 6, 2021.

[50] EOSC - What the European Open Science Cloud is: https://www.eosc.eu/about-eosc, accessed Dec 6, 2021.

[51] European Open Science Cloud (EOSC) - What the cloud is, how it was developed and being implemented: https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science/european-open-science-cloud-eosc_en, accessed Dec 6, 2021.

[52] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

scholarly data aim to foster open science and data exchange. They are highly relevant in the context of EOSC, where open data exchange is a vital focus.

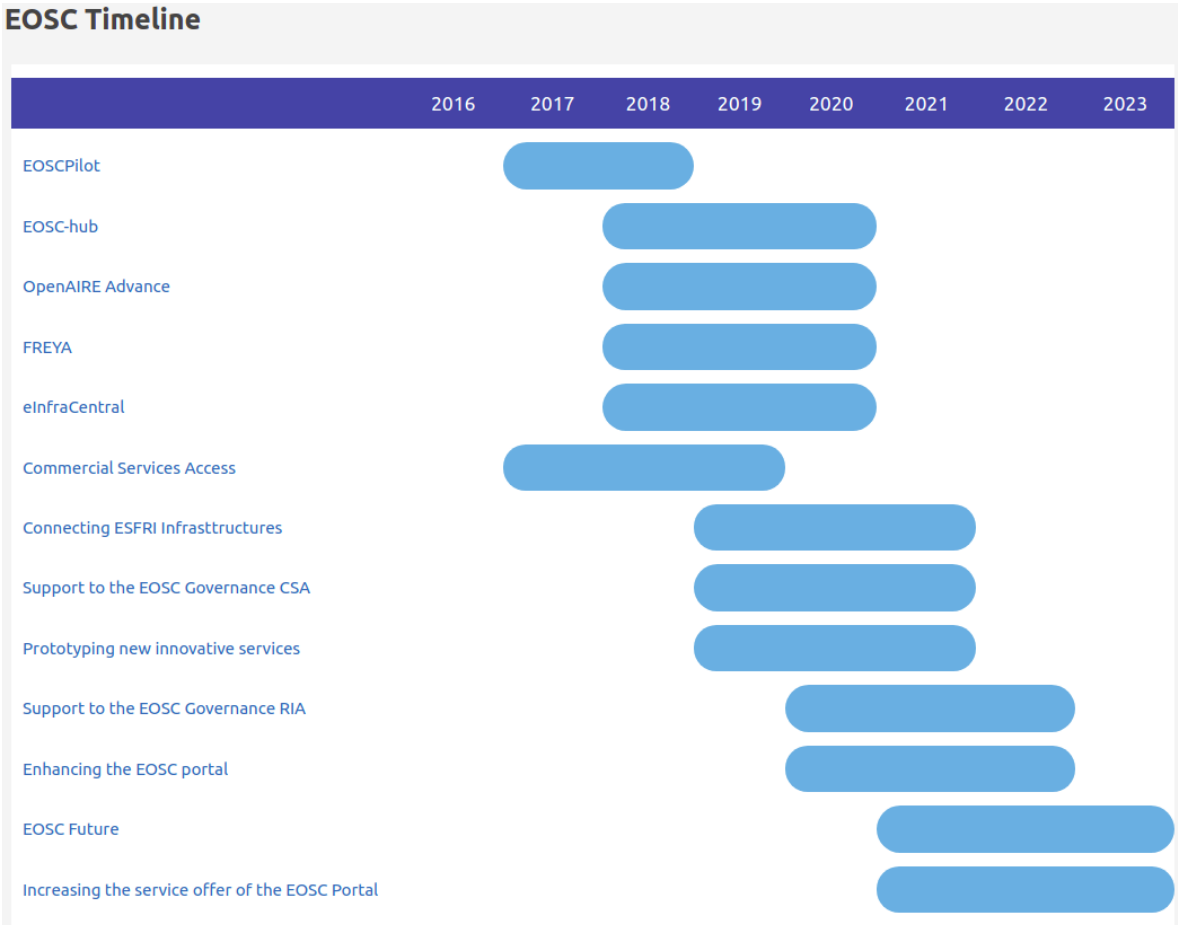

Figure 3: Timeline of past and ongoing EOSC-related projects[53].

In the following, we give an overview and description of highly notable projects and initiatives related to the EOSC:

- **EOSCpilot[54]:** the initial project, researching technical, scientific, and cultural challenges related to the implementation of such an endeavour.
- **EOSC-hub[55]:** this H2020 project aimed to create a central location, the hub, where researchers can search and find data resources and services relevant to their research needs. Notable outcomes of this project are the EOSC portal/marketplace, a search engine for accessing EOSC resources[56], as well as guidelines for the interoperability with EOSC (the rules of participation[57], the inclusion criteria[58], and the provider documentation[59]).

[53] EOSC portal - About: https://eosc-portal.eu/about/eosc, accessed Dec 3, 2021.
[54] EOSCpilot: https://eoscpilot.eu/, accessed Dec 1, 2021.
[55] EOSC-hub: https://eosc-hub.eu/, accessed Dec 1, 2021.
[56] EOSC Portal and Marketplace: https://eosc-hub.eu/key-exploitable-results/eosc-portal-and-marketplace, accessed Dec 1, 2021.

- **OpenAIRE ADVANCE[60]:** the vision of this H2020 project is to provide unimpeded access to the outputs of publicly funded projects in the EU. The consortium consists of 65 partners (universities and research organizations). The project produced a platform listing EOSC services, defined open science policies, and provided training to familiarize with open science. OpenAIRe provides easy access to its data resources, both via API[61] and as a downloadable dump[62].
- **FREYA[63]:** its main objective is the development of infrastructure for persistent identifiers (PIDs) for research. The focus was on the cross-linking between different PID services, and the creation of sustainable PID e-infrastructure for the European research community. One outcome of this effort is the FREYA PID graph, which links scientific resources with each other and helps a researcher to understand underlying networks[64].
- **eInfraCentral:** This project aimed to find a consensus about technical requirements for European e-infrastructures and also created a portal, where existing offers became searchable. Unfortunately, the project website[65] is not accessible anymore, which makes it impossible to examine the project outcomes with regards to interoperability with TRUSTS. The information given here comes from the project's CORDIS website[66].
- **EOSC executive board working groups[67]:** The EOSC governance structure established six working groups to tackle EOSC challenges, i.e. the working groups (i) "Landscape", (ii) "FAIR", (iii) "Architecture", (iv) "Rules of Participation", (v) "Skills and Training", and (vi) "Sustainability". Among these, the outputs of the "Architecture" and "Rules of Participation" working groups are specifically relevant for TRUSTS. We describe these in Section 4.2.5.

The next section focuses particularly on the interoperability of the EOSC.

4.2 Interoperability in EOSC

In the following, we summarize the output of initiatives and projects previously introduced with regards to resources and approaches related to interoperability. The goal is to identify connection points, be it technical or procedural, that are also relevant for TRUSTS and help to implement the interoperability solution in a way that it becomes compatible with the EOSC.

---

[57] EOSC rules of participation: https://op.europa.eu/en/publication-detail/-/publication/a96d6233-554e-11eb-b59f-01aa75ed71a1/language-en/format-PDF/source-184432576, accessed Dec 1, 2021.

[58] EOSC Provider Portal - Inclusion Criteria: https://eosc-portal.eu/providers-documentation/eosc-provider-portal-inclusion-criteria, accessed Dec 1, 2021.

[59] EOSC Providers Documentation: https://eosc-portal.eu/providers-documentation, accessed Dec 1, 2021.

[60] OpenAIRE Advance: https://www.openaire.eu/advance/, accessed Dec 1, 2021.

[61] The OpenAIRE APIs: https://graph.openaire.eu/develop/, accessed Dec 3, 2021.

[62] OpenAIRE Research Graph Dumps: https://graph.openaire.eu/develop/graph-dumps.html, accessed Dec 3, 2021.

[63] FREYA: https://www.project-freya.eu/en/about/mission, accessed Dec 1, 2021.

[64] FREYA - The PID Graph: https://www.project-freya.eu/en/pid-graph/the-pid-graph, accessed Dec 10, 2021.

[65] eInfraCentral: http://einfracentral.eu/, website discontinued; access attempt Dec 7, 2021.

[66] European E-Infrastructure Services Gateway: https://cordis.europa.eu/project/id/731049, accessed Dec 7, 2021.

[67] EOSC Working Groups: https://www.eoscsecretariat.eu/eosc-working-groups, accessed Dec 9, 2021.

4.2.1 EOSCpilot

Interoperability and adherence to the FAIR principles are at the core of EOSC[68]. EOSCpilot conduct a gap analysis and derived recommendations from the identified gaps[69]. These recommendations are supposed to guide the EOSC and create an interoperable infrastructure[70].

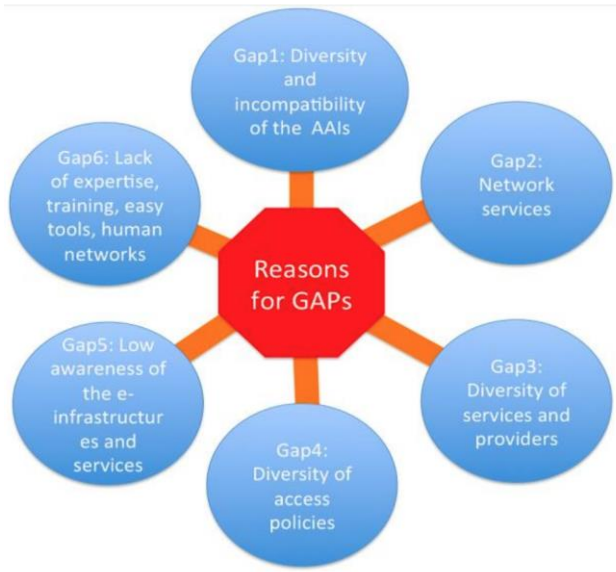

Figure 4: The gaps for EOSC interoperability as identified in the EOSCpilot (EOSCpilot "D6.8: Final EOSC Architecture" [71], p. 6).

Figure 4 depicts the gaps identified in the EOSCpilot deliverable "D6.8: Final EOSC Architecture"[72]. This involves technical aspects, (e.g. "Gap 2: Network services"), organisational aspects (e.g. "Gap 6: Lack of expertise, training, easy tools, human networks"), but also a lack of visibility (e.g. "Gap 5: Low awareness of the e-infrastructures and services"). From there, a set of recommendations was derived (see Figure 5) to guide the development of the EOSC infrastructure to ensure interoperability. For each of the gaps, one or more recommendations to bridge the gap were identified. For instance, for "GAP 3 Diversity of services and providers", EOSCpilot formulated the recommendations that services should have standardized descriptions (Recommendation 3A), should define how interoperability with them can be achieved (Recommendation 3B), how to ensure user privacy and data traceability (Recommendation 3C), and how to ensure security (Recommendation 3D)[73]. Next to the recommendation to bridge the gap, EOSCpilot also provides six recommendations on the application of FAIR principles from within EOSC[74].

[68] EOSCpilot's Contributions to EOSC Interoperability: https://eoscpilot.eu/eoscpilot%E2%80%99s-contributions-eosc-interoperability, accessed Dec 13, 2021.
[69] *ibid.*
[70] *ibid.*
[71] EOSCpilot, "D6.8: Final EOSC Architecture": https://eoscpilot.eu/sites/default/files/eoscpilot-d6.8-v2.4_0.pdf, accessed Dec 7, 2021.
[72] *ibid.*
[73] EOSCpilot's Contributions to EOSC Interoperability: https://eoscpilot.eu/eoscpilot%E2%80%99s-contributions-eosc-interoperability, accessed Dec 13, 2021.
[74] *ibid.*

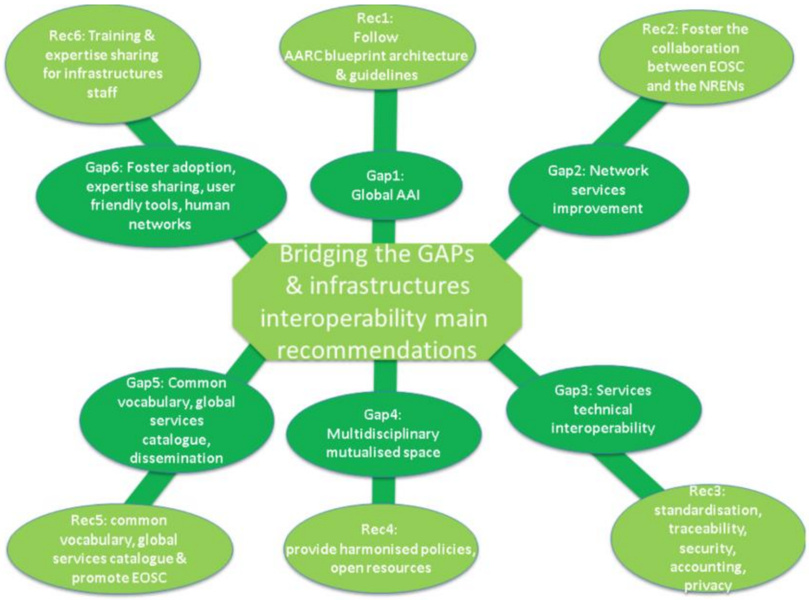

Figure 5: Recommendations to close interoperability gaps (EOSCpilot "D6.8: Final EOSC Architecture" [75], p. 9).

EOSC Datasets Minimum Information (EDMI)

Making data assets, i.e., datasets and data-related services, searchable, requires the availability and usage of a metadata model. TRUSTS builds upon the existing IDS-IM, the information model created and maintained by the IDS. The EDMI developed in EOSCpilot is a metadata guideline to make datasets searchable and has been developed as part of the solution to synchronize metadata between EOSC services and the metadata catalogue (see Figure 6 for the paths of synchronization within the EOSC architecture).

[75] EOSCpilot "D6.8: Final EOSC Architecture": https://eoscpilot.eu/sites/default/files/eoscpilot-d6.8-v2.4_0.pdf, accessed Dec 7, 2021.

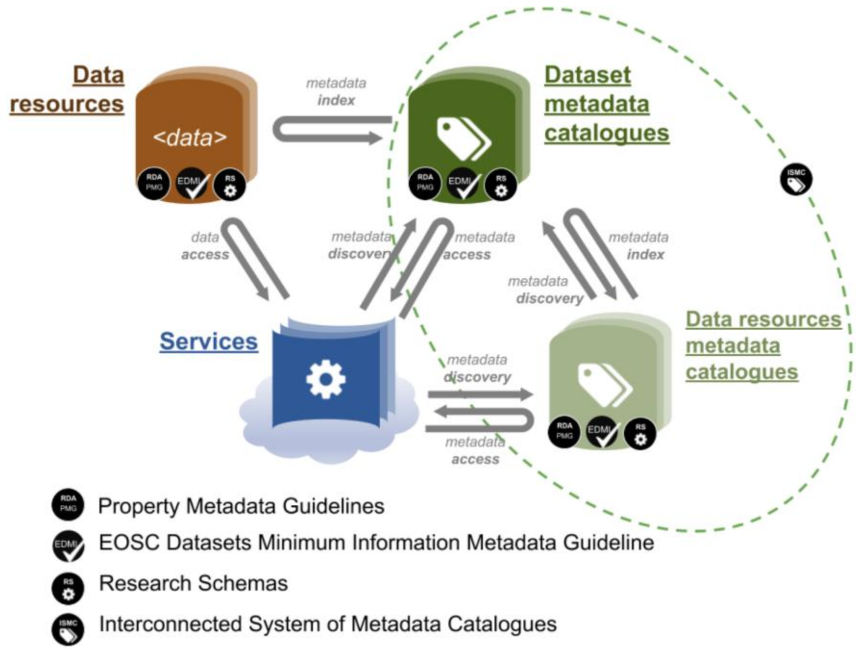

Figure 6: Synchronization of metadata between EOSC data resources/services with metadata catalogues (EOSC Pilot "D6.9: Final report on Data Interoperability" [76], p. 17).

EDMI is supposed to be a minimum set of universal metadata properties, making datasets searchable for both services and users. EDMI is general enough to be applicable across different scientific domains, i.e., it is a minimal set of metadata properties applicable for a wide variety of datasets (see Figure 7 for an illustration). EDMI has two main categories, which are functional and operational metadata properties. The former ones are relevant for human users, such as scientists, wanting to locate and access datasets. They cover aspects such as the dataset name, a description, or the license of the dataset. The latter ones are relevant for services, and cover properties such as "accessUrl" or "accessInterface". The full set of properties is available on the Github website of EDMI[77], the guidelines are available under an MIT license.

Lessons learned:

- **Bridging EOSC Interoperability Gaps:** Development of a detailed strategic plan for the adoption of the recommendation formulated to bridge the gaps. This should include a relevance check of recommendations (i.e., which recommendations of EOSCpilot are also relevant for TRUSTS), as well as an adaptation of the recommendations where necessary.
- **EDMI Integration:** EDMI is relevant for TRUSTS as a potential extension of the metadata model used in the project, IDS-IM. Extending IDS-IM by the properties used in EDMI would naturally make it compatible with the EOSC. We found out during our research that the IDS-IM is extensive enough that it already covers significant aspects of EDMI and does not require significant improvement.

[76] EOSCpilot "D6.9: Final report on Data Interoperability": https://eoscpilot.eu/sites/default/files/eoscpilot-d6.9-v1.3.1-submitted.pdf, accessed Dec 1, 2021.
[77] EDMI metadata properties: https://eosc-edmi.github.io/properties, accessed Dec 7, 2021.

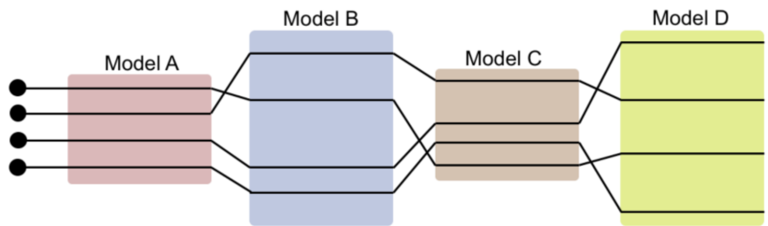

Figure 7: The minimum set of metadata properties required to describe universal models, or datasets (EOSCpilot D6.3: 1st Report on Data Interoperability: Findability and Interoperability[78], p. 23).

4.2.2 EOSC-hub

EOSC-hub interoperability guidelines differentiate between "common services" and "federation services"[79]:

- **Common services:**
  - Cloud Compute (inc. containerisation and orchestration)
  - HTC/HPC Specification
  - Metadata Management and Data Discovery
  - PaaS Solutions
  - Workflow management and user interfaces and Data analytics
- **Federation services:**
  - Accounting
  - Helpdesk
  - Monitoring
  - Security
  - Software Quality Assurance
  - AAI service

The "Metadata management and data discovery" is the relevant service for the research in our task, which itself comprises three services:

- **Annotation service:** *"comprises the ability for end-users to create, manage and search for annotations on data resources"*[80].
- **Data discovery and access:** *"comprises the ability for end-users to search for data resources and access the referenced data"*[81]. This includes the mapping of metadata catalogues from external resources into the EOSC-hub catalogue[82].

---

[78] EOSCpilot "D6.3: 1st Report on Data Interoperability: Findability and Interoperability": https://eosc-edmi.github.io/resources/eoscpilot-d6.3.pdf, last accessed Dec 7, 2021.
[79] EOSC Technical specifications and interoperability guidelines: https://www.eosc-hub.eu/technical-documentation#overlay-context=, accessed Dec 14, 2021.
[80] EOSC Technical Specification - Annotation Service: https://wiki.eosc-hub.eu/display/EOSCDOC/Metadata+Management+and+Data+Discovery?preview=/68223176/68223179/EOSC_Technical_specification_AnnotationService_Proposal_v1.pdf, accessed Dec 14, 2021.
[81] EOSC Technical Specification - Data Discovery and Access: https://wiki.eosc-hub.eu/display/EOSCDOC/Metadata+Management+and+Data+Discovery?preview=/68223176/68223180/EOSC_Technical_specification_DataDiscoveryAndAccess_Proposal_v1.pdf, accessed Dec 14, 2021.
[82] *ibid.*

- **Metadata cataloguing and indexing:** *"comprises the management of metadata in the whole life cycle from generation up to uploading and indexing metadata in a searchable catalogue"*[83].

Each of these services uses a set of standards, see Table 3.

Table 3: Standards and protocols used for metadata management in the EOSC-hub.

Annotation service[84]	**Data discovery and access**[85]	**Metadata cataloguing and indexing**[86]
• JSON-LD • W3C Web Annotation data model • OpenAPI specification • RDF • SPARQL • SolR • RESTful API • User interface integration (Iframe, web component, …) • SPARQL endpoint	• DataCite Metadata Schema 4.1 • ISO 639-1 codes • B2FIND classification for disciplines • ElasticSearch • CKAN-API • SolR	Standard • Community specific metadata schemas and standards • DataCite Metadata Schema 4.1 • Controlled vocabularies Protocol/API • OAI-PMH • ResourceSync • REST APIs • CSW/OGC

Lessons learned:

- **EOSC-hub Technical Stack:** The technical stack of technologies related to interoperability for metadata exchange consists of widely used technologies and tools such as JSON, RDF, SPARQL, SolR, etc.

4.2.3 OpenAIRE ADVANCE

OpenAIRE aims to make research outcomes of publicly funded EU projects openly and easily available. In its own words, its goal is *" [...] to provide unlimited, barrier-free, open access to research outputs financed by public funding in Europe. OpenAIRE fulfils the EOSC vision substantially [...]"*[87]. OpenAIRE hosts a substantial amount of research data, for example, 124,403,533 publications, 12,661,374

[83] EOSC Technical Specification - Metadata Cataloguing and Indexing: https://wiki.eosc-hub.eu/display/EOSCDOC/Metadata+Management+and+Data+Discovery?preview=/68223176/68223181/EOSC_Technical_specification_MetadataCatalogingIndexing_Proposal_v1.pdf, accessed Dec 14, 2021.

[84] EOSC Technical Specification - Annotation Service: https://wiki.eosc-hub.eu/display/EOSCDOC/Metadata+Management+and+Data+Discovery?preview=/68223176/68223179/EOSC_Technical_specification_AnnotationService_Proposal_v1.pdf, accessed Dec 14, 2021.

[85] EOSC Technical Specification - Data Discovery and Access: https://wiki.eosc-hub.eu/display/EOSCDOC/Metadata+Management+and+Data+Discovery?preview=/68223176/68223180/EOSC_Technical_specification_DataDiscoveryAndAccess_Proposal_v1.pdf, accessed Dec 14, 2021.

[86] EOSC Technical Specification - Metadata Cataloguing and Indexing: https://wiki.eosc-hub.eu/display/EOSCDOC/Metadata+Management+and+Data+Discovery?preview=/68223176/68223181/EOSC_Technical_specification_MetadataCatalogingIndexing_Proposal_v1.pdf, accessed Dec 14, 2021.

[87] OpenAIRE: https://www.openaire.eu/mission-and-vision, accessed Dec 12, 2021.

datasets, and 7,003,553 other research products[88]. The service is widely visible and had 162,367,716 views and downloads in 2021[89]. The service is available as the so-called OpenAIRE Research Graph[90].

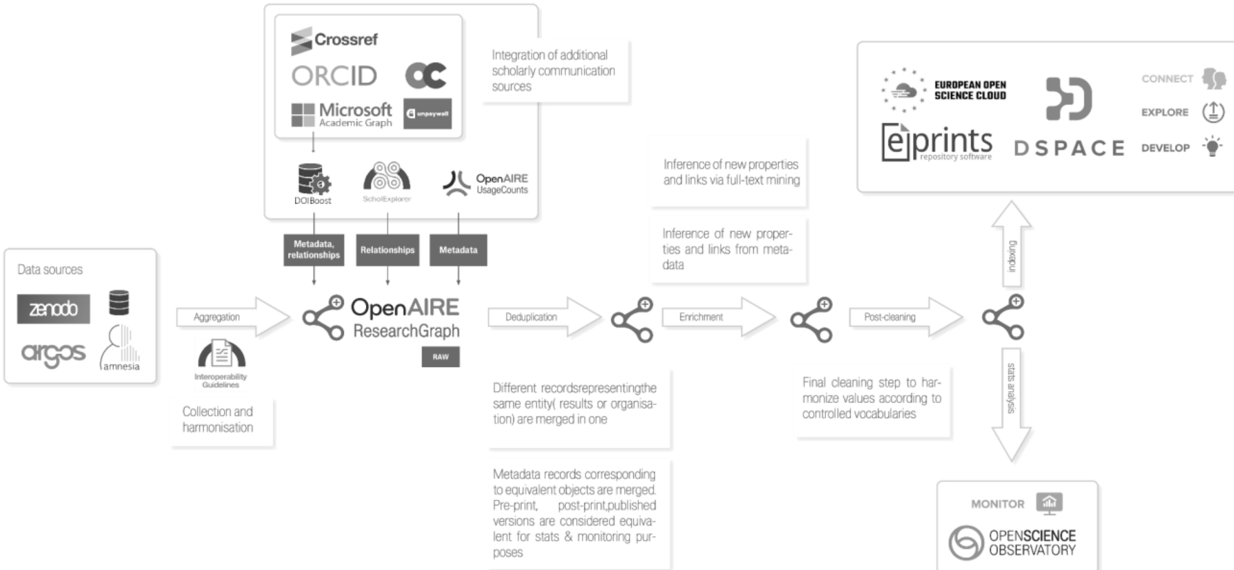

Figure 8: The architecture of the OpenAIRE Research Graph[91].

Figure 8 depicts the architecture of the OpenAIRE Research Graph. Data is acquired from a variety of sources, e.g., Zenodo. The data is cleaned, i.e., deduplicated, enriched, and harmonized. Subsequently, it is indexed and analyzed statistically. The tools used for indexing are the well-known data management platforms ePrints[92] and DSpace[93]. The two platforms are also covered by our systematic platform reviews (see Section 3).

The OpenAIRE Research Graph is accessible in two ways. There is an API available to registered users, free of charge[94]. For users preferring to access the data in bulk, there is a downloadable dump[95].

Lessons learned: there are three opportunities for TRUSTS to interoperate with OpenAIRE, and thus with EOSC:

- **OpenAIRE Bulk Inclusion:** Inclusion of the OpenAIRE dump to make all data resources available from within TRUSTS is possible.
- **OpenAIRE Continuous Update:** Connecting to the OpenAIRE API for continuous updates about changes to the data resources is possible, e.g., to maintain an own version of the OpenAIRE dump and keep it up to date.
- **OpenAIRE CKAN Harvesting:** The OpenAIRE architecture uses ePrints and DSpace for indexing.

[88] OpenAIRE Research Graph: https://graph.openaire.eu/, accessed Dec 12, 2021.
[89] OpenAIRE Provide: https://provide.openaire.eu/home, accessed Dec 12, 2021.
[90] OpenAIRE Research Graph: https://graph.openaire.eu/, accessed Dec 12, 2021.
[91] OpenAIRE Research Graph Architecture: https://graph.openaire.eu/about#tabs_card, accessed Dec 12, 2021.
[92] ePrints: https://www.eprints.org/uk/, accessed Dec 12, 2021.
[93] DSpace: https://www.dspace.com/de/gmb/home.cfm, accessed Dec 12, 2021.
[94] OpenAIRE Research Graph APIs: https://graph.openaire.eu/develop/overview.html, accessed Dec 12, 2021.
[95] *ibid.*

4.2.4 FREYA

FREYA connects and links existing PID services with each other[96], thus creating a form of interoperability among them. FREYA explicitly describes efforts of interoperability with the EOSC in FREYA *"D4.5 Integration of the PID Graph with the EOSC"*[97]. It adheres to the EIF and the four layers of interoperability defined therein[98], i.e., technical, organisational, semantic, and legal interoperability.

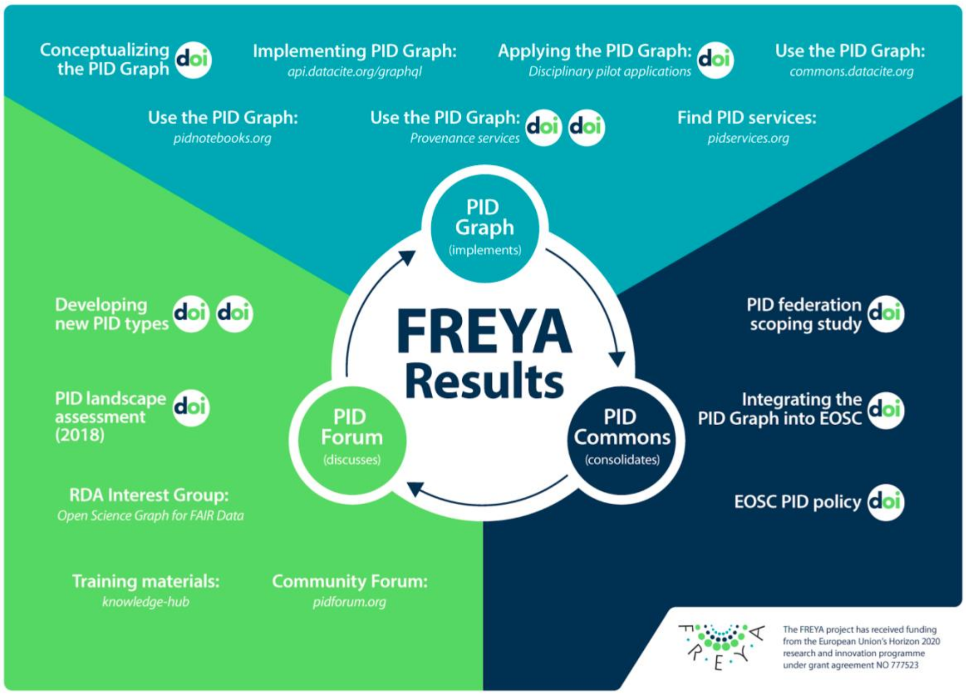

Figure 9: Outcomes of the FREYA project (FREYA[99], p. 17).

The outcomes of FREYA, depicted in , are incorporated, thus "interoperate", with EOSC in the following ways[100] (at the time of the creation of the deliverable):

- **Technical/semantic interoperability** is firstly achieved by registering FREYA's PID graph in the EOSC marketplace[101]. Secondly, since the PID graph is a fundamental service, it should also get integrated into the EOSC core services.
- **Organisational interoperability** is achieved via including multiple of its services into the EOSC marketplace.
- **Legal interoperability** by an adequate future path of the PID graph.

[96] https://www.project-freya.eu/en/about/mission, accessed Dec 10, 2021.
[97] FREYA "D4.5 Integration of the PID Graph with the EOSC": https://www.project-freya.eu/en/deliverables/freya_d4-5.pdf, accessed Dec 12, 2021.
[98] European Interoperability Framework: https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/european-interoperability-framework-detail, accessed Dec 12, 2021.
[99] FREYA "D4.5 Integration of the PID Graph with the EOSC": https://www.project-freya.eu/en/deliverables/freya_d4-5.pdf, accessed Dec 12, 2021.
[100] *ibid.*
[101] EOSC marketplace: https://marketplace.eosc-portal.eu/, accessed Dec 12, 2021.

FREYA aims to fulfill further aspects of interoperability with EOSC, which were mentioned in the deliverable but have not been put into reality at that time.

Lessons learned:

- **EOSC Marketplace Registration:** TRUSTS could follow a similar approach as FREYA, i.e., register for the EOSC marketplace.
- **FREYA PID Graph Connection:** if TRUSTS integrates a PID service for its data assets, adopting the FREYA PID graph could be an option to further strengthen interoperability.

4.2.5 Executive Board Working Groups

EOSC interoperability framework

EOSC provides a framework on interoperability, the "EOSC Interoperability Framework"[102]. In the following text, we describe aspects of this framework relevant for TRUSTS. In adherence to other common interoperability frameworks, such as the European Interoperability Framework[103], it splits interoperability into four different layers: the (i) technical, (ii) semantic, (iii) organisational, and (iv) legal layer. The working group created a catalogue of criteria for each layer. Adherence to the criteria is strongly recommended in order to establish interoperability with the EOSC. The working group provides the following recommendations based on identified problems and needs:

- **Technical interoperability:**
  - **Problems:** differing authentication/authorisation mechanisms, manifold, and diverse data exchange formats, differing PIDs, etc.
  - **Recommendations:** usage of common security frameworks, facilitate access to data sources in various formats, unified PID policy.
- **Semantic interoperability:**
  - **Problems:** no common definitions, no shared metadata schema.
  - **Recommendations:** clear definition of concepts, a minimum metadata model to facilitate search.
- **Organisational interoperability:**
  - **Problems:** unclear use policies, no clear rules for interoperability across organizations and domains.
  - **Recommendations:** inclusion of policies into the "Rules of Participation"[104] (the participation rules are summarized in the section below)
- **Legal interoperability:**
  - **Problems:** incompatibility of different licenses and varying copyrights (e.g., because of different national jurisdiction).
  - **Recommendations:** licences should be standardized and available in human- and machine-readable format, licenses should be as open as possible.

[102] EOSC interoperability framework: https://op.europa.eu/en/publication-detail/-/publication/d787ea54-6a87-11eb-aeb5-01aa75ed71a1/language-en/format-PDF/source-190308283, accessed Dec 9, 2021.
[103] New European Interoperability Framework: https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf, accessed Dec 9, 2021.
[104] EOSC rules of participation: https://op.europa.eu/en/publication-detail/-/publication/a96d6233-554e-11eb-b59f-01aa75ed71a1/language-en/format-PDF/source-184432576, last accessed Dec 9, 2021.

A full list of recommendations is listed in the "EOSC Interoperability Framework", p. 27.

Rules of participation

The "EOSC Rules of Participation Working Group" has published a document with rules[105] for EOSC participants. The rules are policies and procedures to safe-guard trust and openness with the EOSC ecosystem. They are relatively universal rules, applicable to all EOSC participants (national, international, commercial, publicly funded) and all resources published within EOSC (workflows, publications, software, etc.). The working group made the deliberate decision to keep the total number of rules small, resulting in the following set of eight rules:

- **EOSC is based on the principle of openness:** all resources should be provided as open as possible and are accessible in a publicly available registry. Individual resources should also follow this principle of openness, but might be closed if justified (e.g. for commercial reasons).
- **EOSC resources align with FAIR principles:** resources should follow FAIR principles[106] (Findable, accessible, interoperable, reusable), even though the levels of FAIRness might differ.
- **EOSC services align with EOSC architecture & interoperability guidelines:** these rules target the creation of an architecture to reduce resource silos and foster integration, ideally beyond domains and geographical borders.
- **EOSC is based on principles of ethical behaviour and research integrity:** adherence to this rule targets diversity and integration, privacy and anonymity, and the promise to not intentionally provide wrong information.
- **EOSC users are expected to contribute to EOSC:** these rules encourage the inclusion of results achieved with EOSC resources back into the EOSC ecosystem, ideally with open licenses.
- **EOSC users adhere to terms and conditions associated with the resources they use:** the original terms of use of resources included in EOSC do not change. Consequently, users agree to adhere to the original terms of use.
- **EOSC users reference the resources they use in their work:** usage of resources shall follow scientific best practices and referenced accordingly, ideally using PIDs.
- **Participation in EOSC is subject to applicable policies and legislation:** adherence to all policies and legislations, e.g., in a global context to objectives such as the UN's Sustainable Development Goals[107].

Lessons learned:

- **Adoption of rules of participation:** These rules help to establish a common playground and quasi-"legal" framework for the participation of stakeholders in the EOSC.

4.3 Systematic Review of EOSC Platforms

In addition to our review of the documents and material produced by previous EOSC projects, we also systematically examined initiatives that are already part of the EOSC ecosystem. The goal of this examination was to identify opportunities to directly interact and exchange with

[105] *ibid.*
[106] https://www.go-fair.org/fair-principles/, last accessed Dec 3, 2021.
[107] https://sdgs.un.org/goals, last accessed Dec 3, 2021.

them. For this purpose, we used the EOSC portal[108] to find resources. A team of two reviewers assessed the resources of the two EOSC domains "Engineering & Technology" and "Medical & Health Sciences". In the spirit of open and FAIR data, we only selected **open access** initiatives for this examination. The goal of the examination was to get an overview of the technical capabilities and requirements of the individual initiatives. Furthermore, we wanted to find out if there are any existing interfaces ready to be used from within TRUSTS. Interfaces, in this context, are programming interfaces, i.e., APIs, where an automatized exchange of data/metadata is possible for the two respective parties.

Table 4 and Table 5 show the results of this examination. Unfortunately, there are multiple EOSC initiatives without a machine-readable API, which will make it hard to integrate them directly into TRUSTS. Also, the available interfaces are very diverse. Conceptualizing and implementing dedicated interoperability solutions for all of them is unfeasible.

Lessons learned: Individual EOSC initiatives do not have a shared interface. The initiatives strongly differ with respect to their technical background. For example, some of them are websites with an interactive toolkit included. Others provide infrastructure, such as in the case of Jaqpot, which is an e-infrastructure for sharing machine-learning models. The technical diversity of the individual platforms would require individual solutions for each platform, which is not feasible. Thus, for TRUSTS, instead of developing a large set of individual solutions, it is recommended to develop an interface that exposes the TRUSTS metadata model in a machine-readable way. Such an interface will allow interested initiatives to connect to this interface and share the metadata they have readily available in the format required by TRUSTS. On the one hand, this provides a universal interface usable for all external EOSC initiatives, and on the other hand, is a solution feasible without the requirement to develop many individual solutions.

Table 4: Assessment of the open-access resources included in the EOSC domain "Medical & Health Sciences" [109].

Initiative name	**Short description provided on EOSC Marketplace**[110]	**Type of resource**	**API?**
3DBIONOTES-WS	*"A web application designed to automatically annotate biochemical and biomedical information onto structural models."*	Web application	Y (http://3dbionotes.cnb.csic.es/ws/api)
AMBER-based Portal Server for NMR structures (AMPS-NMR)	*"Web portal for the refinement of Nuclear Magnetic Resonance (NMR) structures of macromolecules"*	Web portal	Website discontinued
CyVerse UK	*"Cyberinfrastructure for life science"*	Virtual machines	No
DisVis web portal	*"DisVis allows you to visualize and quantify the*	Web application,	

[108] EOSC Portal - A gateway to information and resources in EOSC: https://eosc-portal.eu/, accessed Dec 1, 2021.

[109] EOSC marketplace, domain "Medical & Health Sciences": https://marketplace.eosc-portal.eu/services?scientific domains=38, accessed Dec 7, 2021.

[110] *ibid.*

	information content of distance restraints between macromolecular complexes."	seems to be discontinued	
European Galaxy Server	*"Open, reproducible, web-based platform for data intensive research."*	Web application to access different data	No, but: ● data libraries are openly available and could be crawled ● data can be uploaded to their premises
HADDOCK2.4 web portal	*"Integrative modelling of biomolecular complexes"*	Has a "workspace", so probably a web app to analyze data. Registration required	No
Identifiers.org	*"Persistent Identifier (PID) services"*	Search engine to find scientific identifiers	No
MOLGENIS	*"Molgenis is a flexible interoperability platform for scientific data and it enables quick setup of local and central catalogues making finding, managing and sharing samples and data much easier."*	On-premises, open source, installable data management platform	No
MetalPDB: a database of metal-binding sites in 3D structures of biological macromolecules	*"Knowledge on metal sites in biological macromolecules, built from 3D structural information."*	Web application	Y (http://metalpdb.cerm.unifi.it/api_help)
Neuroinformatics OpenAIRE Community Gateway	*"Single entry point for discovery and sharing of scientific results in Neuroinformatics"*	Search engine	No
PDB-Tools web	*"Process and massage your PDB file using PDB-Tools"*	Web application for the Python pdb-tools library	No
PhenoMeNal	*"Large-Scale computing for medical metabolomics"*	Cloud-based research environment	No, but Jupyter notebooks are supported

		(Registration required)	
PowerFit web portal	*"Fit an atomic model with an electron density map."*	Web portal and Python library	No
SpotOn	*"Identify Hot-Spots at protein-protein interfaces."*	Web application	No
UBORA e-platform	*"Open Biomedical Engineering e-platform for innovation through education"*	e-platform (registration required)	No
Virtual Imaging Platform	*"Medical simulation and image data analysis as a service"*	Web application	No

Table 5: Assessment of the open-access resources included in the EOSC domain "Engineering & Technology" [111].

Initiative name	**Short description provided on EOSC Marketplace**[112]	**Type of resource**	**API?**
Europeana APIs	*"Large-Scale Data Discovery, Acquisition and Management of Digital Cultural Heritage in Research"*	Data Retrieval API Set (Free - registration required)	Y - several (https://pro.europeana.eu/page/apis#our-apis)
MetaCentrum Cloud	*"Czech national scientific cloud"*	IAAS for scientific users (Free - registration required)	Y (https://cloud.gitlab-pages.ics.muni.cz/documentation/register/?q=API)
Plant Classification	*"Plant classification algorithm"*	Deep learning Application (Other order type)	Y (https://github.com/IgnacioHeredia/plant_classification/blob/master/webpage/README.md#using-the-api)

[111] https://marketplace.eosc-portal.eu/services?scientific_domains=27, last accessed Dec 7, 2021.
[112] *ibid.*

4.4 TRUSTS as EOSC Provider

EOSC provides a documentation[113] on how external organizations can become EOSC providers. This section summarizes the available information. According to this documentation, *"An EOSC Provider is an EOSC System User responsible for the provisioning of one or more Resources to the EOSC. EOSC Providers are organisations, a part of an organisation or a federation that manages and delivers Resources to End-Users."*[114] Becoming an EOSC provider is possible for any legal entity[115]. At the time of research, providers could only onboard services, instead of datasets, documents, or even software[116]. However, the inclusion of such resources was foreseen for a later future[117]. This is confirmed in the "EOSC Resource Profile", where a resource is defined as *"Services, Data Sources, Research Products and any other asset"*[118].

The "EOSC Portal Onboarding Process" document divides the process of onboarding a new provider into the following phases[119]:

- *"Phase 1: An Authorised Representative of a Provider (ARP) registers into the EOSC Portal."*
- *"Phase 2: The Authorised and Authenticated Representative of a Provider (AARP) onboards the Provider (organisation)."*
- *"Phase 3: The AARP onboards the Resources offered by the Provider."*
- *"Phase 4: The AARP onboards the Options/Offerings of a Resource offered by the Provider."*
- *"Phase 5: The AARP and the EPOT[120] maintain the quality of the Profiles."*

The EPOT supervises the onboarding process: It reviews freshly registered providers, checks onboard resources, creates an onboarding report, and assists the providers with feedback or best practices[121]. This manual supervision of the onboarding process is in contrast to the registration on arbitrary online platforms, where interested users can create accounts without a manual review. Such a process can increase trustworthiness of included users, in this case providers, and makes sure that registration attempts are valid and accomplished by a serious entity.

113 EOSC Providers Documentation: https://eosc-portal.eu/providers-documentation, accessed Dec 9, 2021.
114 *ibid.*
115 EOSC Provider Portal - Inclusion Criteria: https://eosc-portal.eu/providers-documentation/eosc-provider-portal-inclusion-criteria, accessed Dec 9, 2021.
116 *ibid.*
117 *ibid.*
118 EOSC Resource Profile: https://eosc-portal.eu/providers-documentation/eosc-resource-profile, accessed Dec 9, 2021.
119 EOSC Portal Onboarding Process: https://eosc-portal.eu/providers-documentation/eosc-portal-onboarding-process, accessed Dec 9, 2021.
120 EPOT = EOSC Portal Onboarding Team.
121 EOSC Portal Onboarding Process: https://eosc-portal.eu/providers-documentation/eosc-portal-onboarding-process, accessed Dec 9, 2021.

5 Exchanging Data with TRUSTS

In the following, we describe two technical features of TRUSTS usable for the exchange of data, i.e. the Dataspace Connector[122] and the CKAN harvesting extension[123].

5.1 Pushing Directly to the Dataspace Connector

An alternative method for including metadata from external sources into the TRUSTS platform is to push it directly into the harvesting-node's Dataspace Connector (DSC). Once ingested into the local DSC, its metadata can then be propagated to the rest of the TRUSTS platform, and the external data marketplace or EOSC initiative will then act as a backend of the DSC, to which requests will be forwarded.

In brief, the DSC exposes a REST API for the creation and maintenance of assets. Out of the box, it can be used for datasets and services (as defined by TRUSTS), and the present project will develop the additional components necessary to make it suitable also for the trading of applications. The endpoints of this API roughly correspond to the different entity types as defined by the IDS-IM (of which the TRUSTS information model is a superset). Since the IDS-IM is a very descriptive and detailed metadata schema, listing an asset in the IDS or TRUSTS ecosystems requires a certain amount of metadata entries (expressed as RDF triples) to be created. For example, an asset has a description, a title, a set of keywords, and several distributions. Each distribution, in turn, can have several contracting-relating attributes along with a set of artifacts, each of which is accompanied by configuration metadata to aid in its access or configuration. For a full explanation of the TRUSTS-IM, the reader is referred to this project's deliverable 3.7. The API provided by the DSC makes the creation of these triples more intuitive for system developers, turning the task of creating a metadata graph for an asset into a series of HTTP calls with a simple client.

A node harvesting from an external data marketplace or EOSC initiative would then have to go over the different assets listed in the external source and issue the corresponding API calls to its DSC. There are several example scripts on how to do this[124], which can serve as a general guide for the node administrator. These need to be adapted to the metadata and file structure of the source which is being integrated.

Once this ingestion process is concluded, the DSC can push the corresponding metadata to the rest of the TRUSTS platform. Users in other TRUSTS nodes can then ask the central Metadata Broker for information about an asset, and upon discovering that it is the harvesting node's connector that is exposing this, can make further inquiries on its metadata directly to that node. When these users wish to access contained in the external data source, they will 1) make an HTTP request to their own DSC asking for that asset 2) that connector will forward the request to the harvesting node's DSC, which will check

[122] Github repository "Dataspace Connector": https://github.com/International-Data-Spaces-Association/DataspaceConnector/tree/21d54bef7bdfbc7b3d2ab37c2338fba2ec0ea137/scripts/tests accessed Dec 14, 2021.
[123] CKAN Harvesting extension: https://github.com/ckan/ckanext-harvest, accessed Dec 14, 2021.
[124] Github repository "Dataspace Connector": https://github.com/International-Data-Spaces-Association/DataspaceConnector/tree/21d54bef7bdfbc7b3d2ab37c2338fba2ec0ea137/scripts/tests accessed Dec 14, 2021.

for access control policy compliance. If the access control is successful, the request will be forwarded to the external data source in the form of an HTTP GET call.

This method of interoperability has the advantage that no assumption is made as of the organization and description of the external sources' assets, except that their description is transformable into the IDS-IM and that access to them can be achieved using HTTP methods. This flexibility allows for interoperation with, potentially, any pre-existing data source, albeit with a fair amount of programming and systems architecture required.

The disadvantages of this approach are several. First, a certain amount of programming will be required from the part of the harvesting node operator. It might be necessary to develop scripts and conversion pipelines to process the metadata. While the above-cited examples, as well as the Metadata Mapper and Metadata Mapping Builder components of the TRUSTS platform, can be of assistance in these efforts, it is still a laborious process requiring technical expertise and familiarity both with the DSC API and with the external organization of assets in the external data source. One further, and probably more crucial, a disadvantage is that authorization and access control are only provided in a very basic form by the DSC. That means that the monetary interoperability with the external data source must be implemented from scratch, making use of the logging mechanisms provided by the DSC. It is important to point out that the DSC offers great flexibility for integrating with other applications, for example any developed to handle transactions between TRUSTS and the external data source, by means of the Apache Camel implementation of the Enterprise Data Bus concept.

5.2 The CKAN Harvester

CKAN is a component in the TRUSTS architecture[125], which provides basic functionality such as user management, metadata cataloguing, faceted search, etc. CKAN also has a lively ecosystem of extensions to add new functionality to a pure installation of CKAN. The CKAN harvesting extension is interesting in the context of interoperability and this task. By installing it, operators of a CKAN instance have an easy-to-use way of exchanging metadata and data with other platforms using CKAN. After installation, it is possible to define harvesting jobs, which download external metadata from the specified source and serialize it into the CKAN database and search engine index. Harvesting jobs can be run periodically, e.g. to provide continuous updates of another CKAN instance's data resources.

[125] TRUSTS D2.6 Architecture design and technical specifications document I:https://www.trusts-data.eu/wp-content/uploads/2021/09/D2.6_Architecture-design-and-technical-specifications-document-I.pdf, accessed Dec 14, 2021.

6 Conclusions and Next Actions

This deliverable describes the efforts and results achieved so far within TRUSTS with respect to interoperability. It focuses on the two major aspects of the interoperability within the task, i.e., interoperability with existing data markets and interoperability with the EOSC. Both aspects contribute strongly to the efforts taken, on an EU-level, to create a data-driven economy, where stakeholders, i.e., businesses, research institutions, etc. participate to create new and innovative data-driven business models and scientific initiatives.

The deliverable describes the efforts that have been pursued as a continuation of previous work. We complete the research of our understanding of the technical requirements of existing data markets, which was previously accomplished by setting up software prototypes and conducting a survey with data market operators. The latter has, unfortunately, not delivered the desired results. Thus, we adapted our research strategy and accomplished a systematic review of existing data management platforms. The rationale of this is that data market operators might leverage existing software solutions and integrate them into the architecture of their systems. TRUSTS itself follows a similar path: TRUSTS uses the open-source data management platform CKAN[126]. CKAN provides functionality to build data repositories, which is also relevant for TRUSTS. Furthermore, CKAN is mature, has a broad community, and relieves TRUSTS from having to implement everything from scratch. By adopting CKAN, TRUSTS avoids the problem of having to reinvent the wheel. OpenAIRE, an EOSC-related project, adopts a similar rationale. Its architecture uses ePrints[127] and DSpace[128] for data asset indexing. Analysing existing data management platforms will thus give us crucial insights into how an interoperability solution has to be designed. It further sheds light on the types of standards and interfaces commonly used.

Next actions in this task include the dissemination of the relevant organizational and legal insight to the respective work packages and include insights gained from analysing the technical background of data markets and data management platforms back into the development of the interoperability solution. For example, our analysis of the data management platforms showed that they provide various ways of data transaction, such as transactional APIs or batch access. While transactional APIs are provided by all examined platforms, batch access is not available for ePrints, InvenioRDM, and Omeka. CKAN provides batch access via a plugin and not natively. The platforms use established ontologies, which is relevant for TRUSTS with regards to connecting the IDS-IM to existing metadata models. All platforms leverage existing ontologies except for CKAN, which again provides a plugin for this functionality.

Besides technical aspects, there are also organizational and legal aspects. An adherence to these and integration into the TRUSTS processes will help to comply with the policies, especially with the EOSC. We

[126] CKAN: https://ckan.org/, accessed Dec 13, 2021.

[127] ePrints: https://www.eprints.org/uk/, accessed Dec 13, 2021.

[128] DSpace: https://www.dspace.com/de/gmb/home.cfm, accessed Dec 13, 2021.

discuss the "Rules of participation" in Section 4.2.5, which is a set of guidelines for data asset providers who want to participate in EOSC and share data assets.

The analysis both of data management platforms, data markets, and the EOSC has revealed crucial and valuable insights that will help to design and conceptualize the interoperability solution. We summarize them as "Lessons learned" for each major topic throughout this deliverable. From the lessons learned we derive the next steps that will be tackled in the remaining phase of the task. We align the next steps with the four layers of interoperability as defined in the EIF[129], i.e., technical, organisational, semantic, and legal interoperability. Table 6 summarizes the lessons learned and translates them into concrete next steps.

Table 6: Next steps derived from lessons learned aligned with the layers of the EIF.

Topic	Lessons learned	Next steps
Technical interoperability	• DMP Feature Diversity • Unified Interface • EOSC-hub Technical Stack	Design a unified interface and deploy it within TRUSTS. Use technologies such as RDF, JSON, REST APIs (also part of the technical stack of EOSC-hub).
	• OpenAIRE Bulk Inclusion • OpenAIRE Continuous Update	Explore, if an inclusion of the OpenAIRE dump is possible and beneficial for TRUSTS. Explore, if the usage of the OpenAIRE API is beneficial to keep the metadata stored about OpenAIRE up to date.
	• OpenAIRE Bulk Inclusion	A TRUSTS database dump could be provided on a periodic basis to allow for bulk upload of TRUSTS metadata to other infrastructures.
	• OpenAIRE CKAN Harvesting	Find out if there are interfaces available to directly connect them to the CKAN, the data management platform at the core of TRUSTS. This method might be easier to implement, because of the usage of existing, mature interfaces ready to be used. The CKAN harvesting mechanism is a versatile and mature mechanism, which might be utilizable for this purpose.
Semantic interoperability	• DMP ontologies • EDMI Integration	A flexible metadata model is beneficial for TRUSTS, which is given by the usage of IDS-IM. A next step will include the integration of EDMI, or adaptation of the IDS-IM to become compatible with EDMI.
Organizational	• Bridging EOSC Interoperability Gaps	Address non-technical gaps, such as low awareness or lack of knowledge using outreach activities and

[129] New European Interoperability Framework: https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf, accessed Dec 13, 2021.

		the production of information material, ideally combined.
Legal	• Adoption of Rules of Participation	The EOSC rules of participation need to be translated into a set of rules compatible with the requirements of TRUSTS. TRUSTS is envisioned as a data market, i.e., a commercial aspect can/should be considered to make the rules compatible.

The next steps require the involvement of other WP3 Tasks and TRUSTS WPs, i.e., "Task 3.4 "WP6 Legal & Ethical Framework", and "WP7 Business Model, Exploitation & Innovation Impact Assurance". We will connect to these WPs/Tasks, inform about our gained insights, and assist in the implementation and integration into the procedures and processes of TRUSTS. The collaboration with these tasks fulfills the last requirement of Task 3.3, i.e., *"this task has strong interdependencies with T3.2 Smart Contracts, T3.4 Data Governance & Metadata and the overall Work Packages: WP7 Business Plan and WP6 Legal Framework, to ensure interoperability solutions are reflected technically, legally and business wise."*[130]

[130] TRUSTS Trusted Secure Data Sharing Space grant agreement.

7 References

[1] Aguiar, J., da Silva Jo˜ao Rocha, Carvalho, R.C.A.R., Castro: A comparative study of platforms for research data management: Interoperability, metadata capabilities and integration potential. pp. 101–111. Springer International Publishing (2015)

[2] Amorim, R.C., Castro, J.A., da Silva, J.R., Ribeiro, C.: A comparison of research data management platforms: architecture, flexible metadata and interoperability. Universal Access in the Information Society 16, 851–862 (2017). https://doi.org/10.1007/s10209-016-0475-y, https://doi.org/10.1007/s10209-016-0475-y