



D3.6 Data Marketplaces with Interoperability Solutions III

Authors: **Michael Boch, Stefan Gindl, Lorenzo Gugliotta, Victor Mireles**

Additional Information: **Follow-up of D3.5**

October 2022



TRUSTS Trusted Secure Data Sharing Space

D3.6 Data Marketplaces with Interoperability Solutions III

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	D3.6 Data Marketplaces with Interoperability Solutions III		
Work Package	WP3		
Contractual due date	30/06/2022	Actual submission date	20/10/2022
Nature	Report	Dissemination Level	Public
Lead Beneficiary	RSA		
Responsible Author	Stefan Gindl (RSA)		
Contributions from	Alan Barnett (DELL), Michael Boch (RSA), George Margetis (FORTH), Victor Mireles (SWC)		

Revision history (including peer reviewing & quality control)

Version	Issue Date	% Complete	Changes	Contributor(s)
v0.1	Feb. 22, 2022	5	Initial Deliverable Structure	Stefan Gindl (RSA)
v0.2	May 12, 2022	30	Technical content	Stefan Gindl (RSA)
v0.3	July 15, 2022	40	Legal aspects	Lorenzo Gugliotta (KUL)
v0.4	Sep. 23, 2022	60	1st version	Stefan Gindl (RSA)
V0.5	Sep. 28, 2022	90	Review-ready version	Stefan Gindl (RSA)
v0.6	Sep. 30, 2022	92	TRUSTS-internal review	Kim Fidomski, Ahmad Hemid (FhG)
v0.7	Oct. 11, 2022	94	TRUSTS-internal review	Andreas Truegler (KNOW)
v0.8	Oct. 19, 2022	96	Content update	Victor Mirelez (SWC)
v1.0	Oct. 20, 2022	100	Final version	Stefan Gindl (RSA)

Disclaimer

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

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

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

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

Copyright message

© TRUSTS, 2020-2022. This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the

work of others has been made through appropriate citation, quotation or both. Reproduction is authorised provided the source is acknowledged.

Table of Contents

1	Executive Summary	8
2	Introduction	10
2.1	Mapping Projects' Outputs	11
2.2	Deliverable Overview and Report Structure	12
3	Technical Interoperability	14
3.1	Interoperability for Interested Third Parties	14
3.1.1	TRUSTS Platform Client	14
3.1.2	The CKAN Harvesting Extension	16
3.2	EOSC	18
3.2.1	Connector Architecture	18
3.2.2	OpenAIRE	19
3.2.3	Europeana	20
4	Semantic Interoperability	24
4.1	Metadata Mapping	25
5	Organisational Interoperability	28
5.1	Interoperability Experiment	28
5.2	Smart Contracts	29
6	Legal Interoperability	30
6.1	Introduction	30
6.2	Potential pain points	30
6.2.1	Legal aspects related to intellectual property law	30
6.2.2	Legal aspects related to privacy and data protection law	31
6.3	Research avenues for solutions	31
6.3.1	Intellectual property law	31
6.3.2	Privacy and data protection law	33
7	Conclusions and Next Actions	34

List of Figures

Figure 1: The interoperability model defined in the EIF (p. 22) .	10
Figure 2: Installation of the TRUSTS platform client using Pip.	15
Figure 3: Description of the usage of the TRUSTS platform client.	16
Figure 4: The dataset browsing page of the Energi Data Service.	17
Figure 5: An overview of the ETL process as realised in the connectors.	19
Figure 6: Europeana’s search functionality.	21
Figure 7: Connection to Europeana’s FTP server using FileZilla.	21
Figure 8: Europeana metadata files in zipped form.	22
Figure 9: A dataset from Europeana published in the TRUSTS platform.	23
Figure 10: An extract of the properties of the EDM1.	25
Figure 11: Metadata properties for an aggregation object, i.e., an individual cultural heritage object.	27
Figure 12: Setup of the interoperability experiment.	28

List of Tables

Table 1: Adherence to TRUSTS GA Deliverable & Tasks Descriptions.	11
Table 2: TRUSTS properties and their equivalents in OpenAIRE and Europeana.	26
Table 3: Examples of metadata mapping between the TRUSTS database and the smart contract component.	29

Glossary of terms and abbreviations used

Abbreviation / Term	Description
API	Application Programming Interface
CKAN	Comprehensive Knowledge Archive Network
DCAT	Data Catalogue Vocabulary
EDM	Europeana Data Model
EDMI	EOSC Datasets Minimum Information
EOSC	European Open Science Cloud
ETL	Extract Transform Load
FAIR	Findable, Accessible, Interoperable, Reusable
FTP	File Transfer Protocol
GA	Grant Agreement
GZ	GNU-Zip, a file compression technology
HTTP	Hypertext Transfer Protocol
JSON	JavaScript Object Notation
PETs	Privacy-Enhancing Technologies
REST	Representational State Transfer
TAR	Tape ARchive, a technology to archive files and folders

1 Executive Summary

This report summarises the work accomplished in Task 3.3 “Data marketplaces interoperability solutions” and continues from Deliverable D3.5 “Data Marketplaces with Interoperability Solutions II”. It is the last deliverable of this task.

The subject of the deliverable: Task 3.3 covers research and development of the TRUSTS interoperability solution, as stated in the description of work: “... *the interoperability solution for TRUSTS will be designed in this task*”¹. The interoperability solution provides an interface for third-party datamarkets and EOSC initiatives and allows them to map their metadata catalogues into the TRUSTS platform, consequently making their offerings visible for organisations and partners involved in TRUSTS.

Summary of the work carried out: The work carried out in the last phase of this task can be divided into technical, semantic, organisational, and legal aspects of interoperability, which goes in line with the “EOSC Interoperability Framework”². From a technical perspective, the task has delivered a software solution usable for external, third-party datamarkets to interoperate with TRUSTS, i.e. to show their offerings from within the TRUSTS platform. The developed tool is a solution for the programmatic exchange of metadata of datasets, which enables a mirroring of the metadata catalogue of third-party datamarkets. From a semantic perspective, we researched the requirements for schema compatibility of the TRUSTS metadata schema with heterogeneous sources. This deliverable describes the outcomes and changes to the TRUSTS metadata schema, which is based on the work accomplished in the previous deliverable D3.5. The organisational perspective was an interoperability simulation. We set up an additional deployment of a TRUSTS platform. In other words, we worked with two individual deployments of TRUSTS: the main deployment is the one that has been used throughout the project lifetime for partners to upload their datasets and experiment with the platform. It is also the deployment that was used to carry out most of the use case trials. The second deployment, a TRUSTS “clone”, was used to mimic the behaviour of a hypothetical third-party datamarket. The goal was to demonstrate how the operators of such a datamarket could use the so-called **TRUSTS platform client** to transfer their offerings to the main TRUSTS platform and make it visible from there. Lastly, we also looked into legal aspects of interoperability, such as intellectual property law as well as privacy and data protection law.

The main conclusion(s): The learnings from the work in this task is that interoperability has to be viewed along different perspectives. Previous research (see deliverable D3.5) has shown that the “New European Interoperability Framework” (EIF)³ provides a solid basis for the purposes of TRUSTS. The framework defines the four layers “technical”, “semantic”, “organisational”, and “legal” interoperability. Task 3.3 adopts this strategy, and the outcomes are presented for each of these four layers in this report.

The purpose of the deliverable: This deliverable describes the work accomplished since the delivery of the previous deliverable D3.5⁴. D3.5 summarises our work in research about data management platforms, our efforts researching the EOSC ecosystem with a focus on interoperability, as well as the TRUSTS built-in methods for programmatic data transfer, i.e. via the Dataspace connector and the CKAN harvester. Furthermore, it describes the four different ways interoperability was addressed in the task, i.e. **technical**,

¹ Task 3.3 Definition of work, TRUSTS Trusted Secure Data Sharing Space grant agreement.

² EOSC Interoperability Framework: <https://op.europa.eu/en/publication-detail/-/publication/d787ea54-6a87-11eb-aeb5-01aa75ed71a1/language-en/format-PDF/source-190308283>, Sep. 21, 2022.

³ The New European Interoperability Framework: https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf, accessed Sep. 23, 2022.

⁴ TRUSTS Deliverable D3.6: https://www.trusts-data.eu/wp-content/uploads/2022/01/D3.5-Data-Marketplaces-with-Interoperability-Solutions-II_Dec2021.pdf, accessed Oct 11, 2022.

semantic, organisational, and legal interoperability. Technical interoperability covers the description and documentation of the software artefacts created to achieve interoperability. Semantic interoperability describes the adaptations required for the TRUSTS metadata model to be compatible with heterogeneous sources. Organisational interoperability shows the details of the interoperability simulation that was carried out to demonstrate interoperability with third-party datamarkets to the use case partners. Lastly, legal aspects of interoperability are covered in the chapter on legal interoperability.

2 Introduction

T3.3 deals with the requirements and development of an interoperability solution for the TRUSTS platform. This is aligned with TRUSTS fostering a data-driven economy, which is one of the goals of TRUSTS. A data-driven economy, which is also the goal of GAIA-X⁵, provides the means required for companies and organisations to share their data and acquire data from other participating members. The interoperability solution envisioned in TRUSTS provides the facilities for existing third-party data markets and for initiatives of the EOSC⁶ ecosystem to connect to TRUSTS. This allows them to upload their offerings to the TRUSTS platform and benefit from the TRUSTS ecosystem. It gives them access to the TRUSTS partners, as well as the involved commercial and non-commercial organisations.

The work on interoperability in this task is based on the recommendations by the European Union for interoperability of public administrations in Europe, summarised in the “New European Interoperability Framework” (EIF)⁷. This framework has been created for public administration. However, its lessons learned and recommendations are applicable beyond public administration and can reach out to the private sector as well. Specifically, the EIF is recognized and used by FREYA⁸, one of the early projects by the EOSC, aiming to provide an infrastructure for the management of persistent identifiers. Given that task 3.3 has a strong focus on EOSC, the selection of EIF as an established interoperability framework comes natural. Figure 1 gives an overview of the interoperability model defined by the EIF.

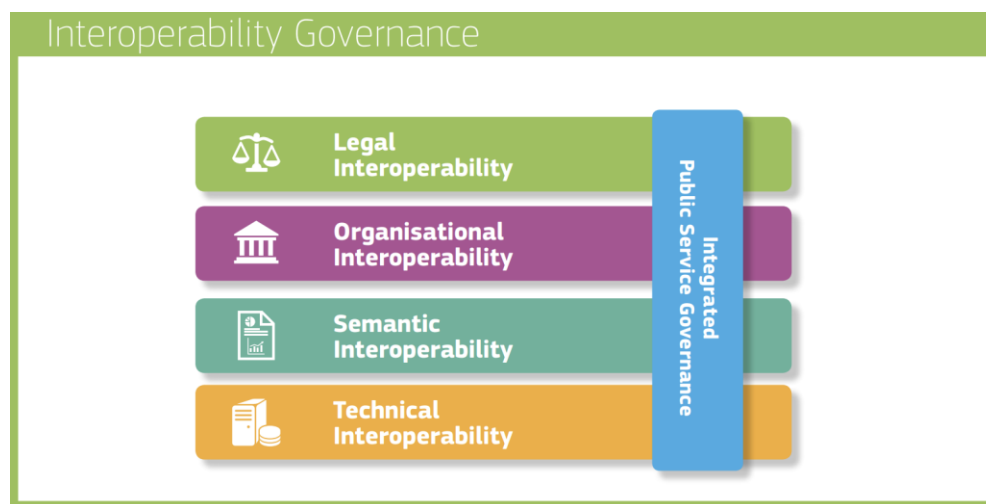


Figure 1: The interoperability model defined in the EIF (p. 22)⁹.

The four layers of interoperability of the interoperability model, as defined by the EIF, are as follows:

⁵ GAIA-X: <https://gaia-x.eu/>, accessed Sep. 21, 2022.

⁶ EOSC: <https://eosc.eu/>, accessed Sep. 21, 2022.

⁷ The New Interoperability Framework: https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf, accessed Sep. 23, 2022.

⁸ FREYA: <https://www.project-freya.eu/en/about/mission>, accessed Sep 23, 2022.

⁹ The New Interoperability Framework: https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf, accessed Sep. 23, 2022.

- **Technical interoperability:** “[...] covers the applications and infrastructures linking systems and services.” (EIF¹⁰, p. 30). Focus of this layer are technical applications and infrastructures used for setting up an interlinked network of partners. Reliable interfaces for the smooth exchange of data and services have strong importance.
- **Semantic interoperability:** “[...] ensures that the precise format and meaning of exchanged data and information is preserved and understood throughout exchanges between parties” (EIF¹¹, p. 29). This layer emphasises the value inherent to data and information, and recommends the usage of established taxonomies and vocabularies, to ensure that “*what is sent is what is understood*” (EIF¹², p. 29).
- **Organisational interoperability:** “[...] also aims to meet the requirements of the user community by making services available, easily identifiable, accessible and user-focused.” (EIF¹³, p. 28). This layer guides regarding the alignment of business processes and facilitates the exchange of information.
- **Legal interoperability:** “[...] is about ensuring that organisations operating under different legal frameworks, policies and strategies are able to work together.” (EIF¹⁴, p. 27). This layer ensures how to handle legal differences across borders and contradictions in legislation, as well as the diversity of licences and the avoidance of too strong legal restrictions.

This report is structured equivalently and describes the efforts of task 3.3 for each of the four EIF layers.

2.1 Mapping Projects’ Outputs

Purpose of this section, is to map TRUSTS Grant 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		These sections cover technical,

¹⁰ The New Interoperability Framework: https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf, accessed Sep. 23, 2022.

¹¹ *ibid.*

¹² *ibid.*

¹³ *ibid.*

¹⁴ *ibid*

	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, Section 5 Section 5.2 Section 6	semantic, and organisational interoperability. This section covers metadata mapping for smart contracts. This section covers legal interoperability.
TRUSTS Deliverable			
<p>D3.6 Data Marketplaces with Interoperability Solutions III</p> <p>The third 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.</p>			

2.2 Deliverable Overview and Report Structure

This report is divided into four main sections, covering technical, semantic, organizational, and legal interoperability. In the following, we give a brief description of each section:

- **Section 3 “Technical interoperability”:** This section documents the technical facilities of interoperability of the TRUSTS platform. It explains the design and usage of the TRUSTS platform client, which helps operators of external datamarkets to map their metadata catalogues into the TRUSTS metadata catalogue. It also explains how the CKAN Harvesting API, an extension of CKAN, can be used to harvest data from other platforms using CKAN (CKAN is the data management platform serving as the backbone of TRUSTS). Furthermore, it describes the two EOSC connectors, i.e. the OpenAIRE and Europeana connector, which are technical solutions to load data from them into TRUSTS. Interoperability with EOSC is one of the goals of T3.3, and it is demonstrated for these two EOSC initiatives.
- **Section 4 “Semantic interoperability”:** This chapter covers the adaptation of the TRUSTS metadata schema with regards to interoperability. We describe, how the metadata schema was extended with EDM¹⁵, the EOSC Datasets Minimum Information. The TRUSTS metadata schema does not only benefit from this extension with regards to EOSC, but EDM is flexible enough to be valuable with regards to interoperability with external datamarkets. The chapter also describes the metadata mapping, i.e. the conversion of external metadata schemas into a format that TRUSTS can handle.
- **Section 5 “organisational interoperability”:** This section describes the interoperability experiment conducted in Task 3.3, which was supposed to demonstrate the technical functionality of the TRUSTS platform client. The client is a software program for the programmatic exchange of a large set of metadata of datasets. It helps to map metadata catalogues into the TRUSTS platform, thus

¹⁵ EOSC Datasets Minimum Information: <https://eosc-edmi.github.io/properties>, accessed Sep. 22, 2022.

making them available from within the platform. The operators of external datamarkets can use this tool to make available their offerings from within TRUSTS. The experiment was designed as a simulation, where a second, individual deployment of a TRUSTS platform served as the external datamarket. This “clone” had a set of datasets in its catalogue, which were transferred to the TRUSTS main deployment using the TRUSTS platform client.

- **Section 6 “Legal interoperability”:** Legal aspects are crucial to establish interoperability across multiple individual players. Organisations might not only have their own understanding and opinions on legal aspects, but those might also differ across nations. Different national legal frameworks need to be taken into consideration, which creates considerable complexity. In this section we discuss several pain points that arise from interoperability. We discuss the challenges for intellectual property law, with potentially incompatible licences, which aggravate a smooth exchange and trade of data. We also discuss privacy and data protection challenges as well as GDPR. We conclude the section with an outline of potential research avenues to tackle the identified challenges.

3 Technical Interoperability

3.1 Interoperability for Interested Third Parties

The potential heterogeneity of existing datamarkets called for a flexible and adaptable solution. We decided to provide an interface for third parties willing to interact and exchange data with TRUSTS. The purpose of this interoperability solution was to provide a programmatic interface, allowing the exchange of assets (datasets, services and applications), along with their associated metadata, on a large scale. Simply speaking, it is an extension of the TRUSTS user interface. Organisations within the TRUSTS federation can manage their datasets from their own node. However, this is possible for single datasets at a time. Given that a third party, such as an external data market, aims to exchange datasets on a large scale and continuously, a solution is required that provides a programmatic interface for batch upload and continuous upload. In the following, we will describe the two programmatic ways how external parties can interoperate with the TRUSTS platform. The first way is the TRUSTS platform client, a wrapper around CKAN's action API specifically adapted to the requirements of TRUSTS, i.e. to access the TRUSTS central node and to generate, via the Dataspace Connector, asset descriptions that are conformant with the TRUSTS-IM and can be viewed and accessed by all participants of the TRUSTS platform. The second way is the CKAN harvesting extension, a generic mechanism for data exchange with other sources included in the CKAN extension ecosystem. The latter is available for all platforms built on top of CKAN and has not been developed within this project. However, it was one of the reasons why CKAN was chosen as the data management platform under the hood of TRUSTS, to provide exactly such a way of natural interoperability for other platforms built on top of CKAN.

3.1.1 TRUSTS Platform Client

The TRUSTS platform client is a solution for programmatic upload of metadata of datasets to TRUSTS. In other words, it allows batch-upload of metadata, in contrast to the user interface, which allows manual upload of metadata of single datasets. It is a standalone Python library, exposing an API, which can be used to implement custom solutions for batch upload and continuous upload of datasets. This "TRUSTS platform client" is a wrapper around the CKAN Action API¹⁶. In the following, we describe the installation and usage of the TRUSTS platform client.

The client was designed for ease of use. Installing it requires a simple call of the Python package manager Pip including the URL to the repository of the client (see Figure 2).

¹⁶ CKAN Action API: <https://docs.ckan.org/en/2.9/api/>, accessed Oct. 3, 2022.

Installation

1. **Virtual environment (optional but recommended):** create a python virtual environment and activate it:

```
python3 -m venv venv  
source venv/bin/activate
```

2. **Installation:**

```
pip install git+https://gitlab.com/trusts-platform/trusts-platform-client.git
```

Figure 2: Installation of the TRUSTS platform client using Pip.

The platform client can be used to access any corporate node in the TRUSTS platform. The client accesses the CKAN API of the TRUSTS platform and routes the uploaded metadata of datasets through the API and the dataspace connector to the central node of CKAN. The user of the client needs an API token as credentials, which can be retrieved via the TRUSTS's admin area, a sub-page in the TRUSTS user interface. Furthermore, the user needs to specify the URL of the deployment of TRUSTS, i.e. the IP address or domain where it is hosted. Here, both local and remote names are allowed. Using the CKAN action `status_show()` shows if the connection to TRUSTS was successful. Subsequently, datasets can be uploaded to TRUSTS. Each dataset requires contract data, which is sent along with the actual metadata to TRUSTS. Figure 3 is a screenshot of the README file of the platform client, as it is visible from TRUSTS Gitlab platform.

Usage

1. Log into your TRUSTS node as an admin user and get an API token in the [admin area](#).

2. Import the class `TRUSTSCKAN`, the main class for exchanging data with TRUSTS:

```
>>> from trusts_platform_client import TRUSTSCKAN
```

3. Connect to a running TRUSTS instance:

```
>>> trusts_url = 'http://127.0.0.1:5000/' # Replace this with your actual URL
>>> CKAN_TOKEN = <YOUR_API_TOKEN>
>>> _trustsckan = TRUSTSCKAN(trusts_url, apikey=CKAN_TOKEN)
```

4. Check that you can access it:

```
>>> _trustsckan.action.status_show()
```

5. Import helper functions to create exemplary data for testing purposes:

```
>>> from trusts_platform_client.trustsckan import helper_load_europeana_dataset, helper_create_contract_data
```

6. Actually create the exemplary data:

```
>>> dataset = helper_load_europeana_data()
>>> contract_data = helper_create_contract_data()
```

7. Transfer the exemplary data into TRUSTS:

```
>>> _trustsckan.post_dataset(dataset, contract_data)
```

Figure 3: Description of the usage of the TRUSTS platform client.

3.1.2 The CKAN Harvesting Extension

CKAN is the data management platform under the hood of TRUSTS, which provides rich functionality in terms of data cataloguing and archiving, as well as user management. It was selected as a widely used and mature solution, especially with interoperability in mind. As reported in the previous deliverable D3.5 “Data Marketplaces with Interoperability Solutions II”, we conducted an extensive survey to compare the features of a variety of data management platforms. At the time of writing D3.5, the article was under review, but has been accepted in the meantime¹⁷.

Commercial and governmental platforms built on top of CKAN are automatically interoperable with TRUSTS. This is possible via [CKAN’s harvesting extension](#). The harvesting extension gives a natural and easy-to-use way for platforms built on top of CKAN to interoperate. CKAN is widely used by data providers,

¹⁷ Boch, M., Gindl, S., Barnett, A., Margetis, G., Mireles, V., Adamakis, E., Knoth, P. (2022). A systematic review of data management platforms. WorldCIST'22.

especially from governments, such as the Australian government¹⁸, the government of Denmark¹⁹, or the government of the United States²⁰. Enterprises also use the technology, e.g. the portal “Energi data service”²¹, which provides data about the Danish energy system. The website of CKAN provides a more extensive list of existing portals²².

The portals come in their own, unique design (see Figure 4 for an example). However, the metadata about the managed data, is accessible via the CKAN REST API. This mechanism allows data portals to “to create a federated network of data portals which share data between each other.”²³

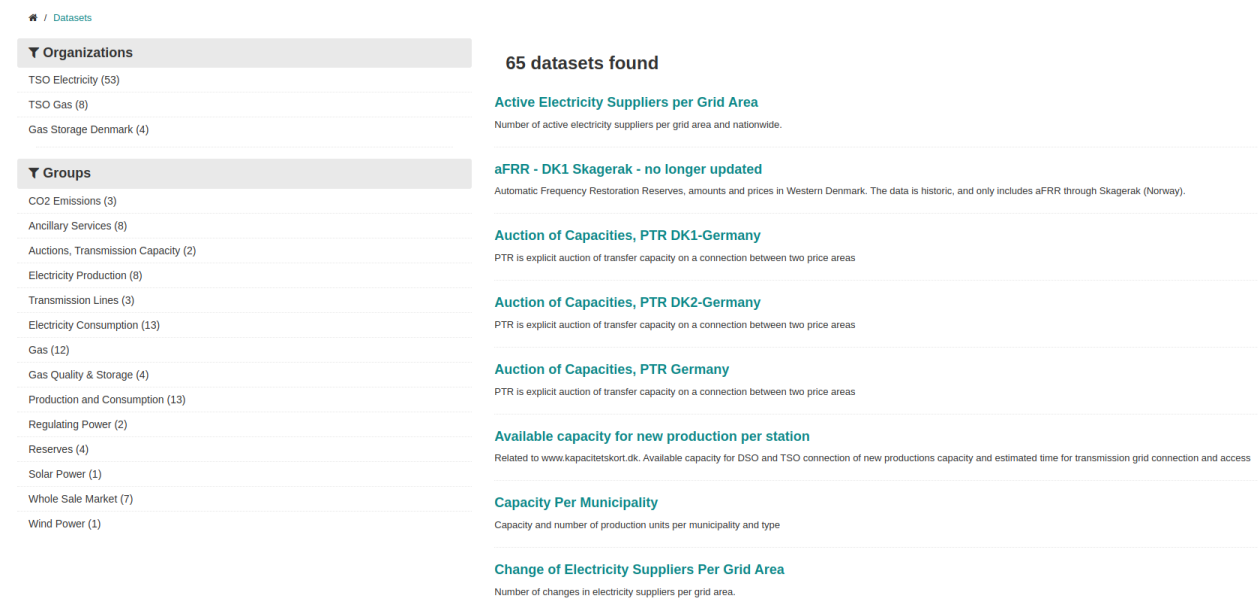


Figure 4: The dataset browsing page of the Energi Data Service²⁴.

On top of that, the CKAN extension mechanism is compatible with the DCAT standard via another extension mechanism²⁵. DCAT, the Data Catalog Vocabulary²⁶, aims “to facilitate interoperability between data catalogs published on the Web”²⁷ and has been defined by the W3C²⁸. Using DCAT, CKAN portals can be “federated from other non-CKAN catalogues.”²⁹

¹⁸ Data portal of the Australian government: <https://data.gov.au/>, accessed July 29, 2022.

¹⁹ Data portal of Denmark: <https://www.opendata.dk/>, accessed July 29, 2022.

²⁰ Data portal of the United States: <https://data.gov/>, accessed July 29, 2022.

²¹ Data portal “Energi data service” by Denmark: <https://www.energidataservice.dk/>, accessed July 29, 2022.

²² Open data portals powered by CKAN: <https://ckan.org/showcase>, accessed July 29, 2022.

²³ CKAN federation: <https://ckan.org/features/federate>, accessed July 29, 2022.

²⁴ Energi Data Service: <https://www.energidataservice.dk/>, accessed July 29, 2022.

²⁵ CKAN DCAT extension mechanism: <https://github.com/ckan/ckanext-dcat>, accessed July 29, 2022.

²⁶ DCAT: <https://www.w3.org/TR/vocab-dcat/>, accessed July 29, 2022.

²⁷ *ibid.*

²⁸ W3C: <https://www.w3.org/>, accessed July 29, 2022.

²⁹ CKAN federation: <https://ckan.org/features/federate>, accessed July 29, 2022.

3.2 EOSC

Next to interoperability with external data markets, task 3.3 is also responsible for interoperability with EOSC:

“In addition, interoperability solutions with the European Open Science Cloud (EOSC) will be evaluated and implemented where possible.”³⁰

EOSC is an effort to connect European research initiatives with each other. It aims to establish itself as a system of systems, where the diverse systems these initiatives are connected and made available to each other. It is supposed to foster innovation and increase mutual benefit from sharing research data and datasets within Europe and globally (see deliverable D3.5 for more details about EOSC). EOSC has a marketplace³¹, where participating initiatives and their details are listed.

3.2.1 Connector Architecture

The connectors follow the principles of the ETL process (“Extract, transform, load”) used for data acquisition in data warehouses:

- **Extract:** The first step in the ETL process is the extraction, or acquisition, of data from the data source. Data suppliers use a variety of methods to provide the data, such as FTP servers (e.g. Europeana³²) or dumps downloadable via a website (e.g. in the case of OpenAIRE³³). In the extraction step, the system first downloads the data to a so-called staging area, where the raw, unprocessed data persists and awaits further steps. Extraction also includes unpacking, e.g. unzipping, the raw files into the staging area in the case of the Europeana and OpenAIRE connectors.
- **Transform:** The transformation step in the ETL process adapts the incoming data into a format compatible with the TRUSTS environment. In other words, it maps its metadata into the TRUSTS metadata schema (see Section 4.1 Metadata Mapping). The transformation can be comparatively straight-forward or complicated, depending on the source format of the data. OpenAIRE delivers its data in .json format, which facilitates a conversion to the TRUSTS metadata schema. In contrast, Europeana delivers its data in .xml format, which requires the usage of further preprocessing tools to convert the .xml data into a .json format.
- **Load:** The final step is the transfer of the converted data items into the data archives of the TRUSTS platform, i.e. its relational database, search engine, key-value store, and SPARQL server. SWC and RSA developed the so-called TRUSTS platform client (available in the Gitlab of TRUSTS), a tool developed for that purpose. CKAN, the data management system under the hood of TRUSTS, provides a powerful REST API (the CKAN Action API³⁴), to manage data. The TRUSTS platform client is a Python wrapper to programmatically access this API. The wrapper itself is an extension of the matura Python library ckanapi³⁵.

³⁰ Task 3.3 Definition of work, TRUSTS Trusted Secure Data Sharing Space grant agreement.

³¹ EOSC Marketplace: <https://marketplace.eosc-portal.eu/>, accessed Sep. 22, 2022.

³² Europeana FTP Download: <https://pro.europeana.eu/page/harvesting-and-downloads>, accessed July 27, 2022.

³³ OpenAIRE Download page: <https://graph.openaire.eu/develop/graph-dumps.html>, accessed July 27, 2022.

³⁴ CKAN Action API: <https://docs.ckan.org/en/2.9/api/>, accessed July 27, 2022.

³⁵ ckanapi: <https://github.com/ckan/ckanapi>, accessed July 27, 2022.

The connector combines the three steps of the ETL process in a pipeline (see Figure 5).

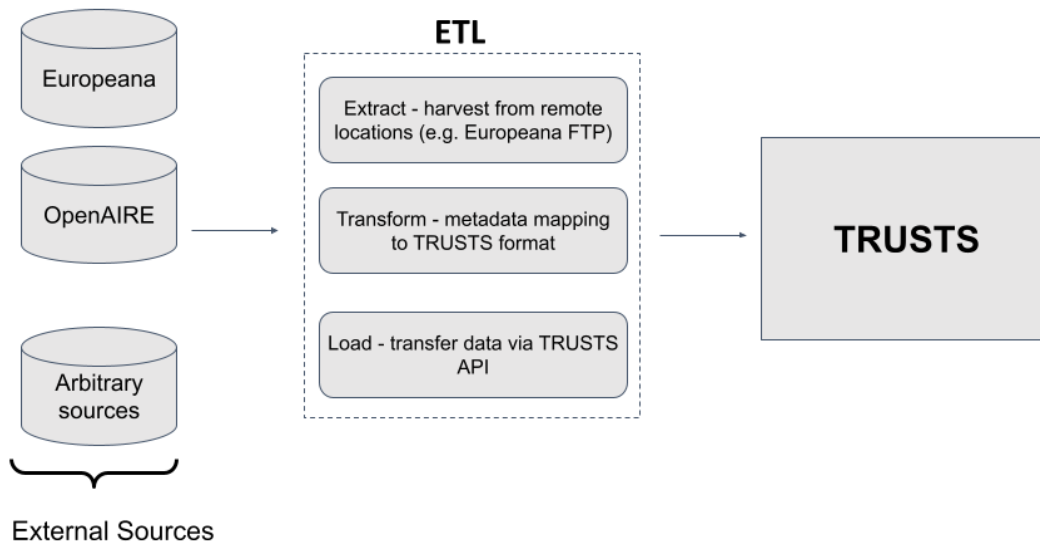


Figure 5: An overview of the ETL process as realised in the connectors.

3.2.2 OpenAIRE

We selected OpenAIRE to demonstrate interoperability with TRUSTS because the initiative was among the first ones in the EOSC universe, launched in 2018, i.e. one year after the EOSCPilot³⁶. Another obvious reason is the large amount of data available from OpenAIRE.

OpenAIRE aims to make research outcomes and data from publicly funded projects available to a broader audience³⁷. The amount of data has grown since the writing of deliverable D3.5, and currently, OpenAIRE hosts 144,379,330 publications, 17,097,102 items of research data, 297,805 software libraries, and 6,838,420 other research products³⁸. The data from heterogeneous resources is aggregated in six so-called OpenAIRE research graphs³⁹:

- **The whole OpenAIRE Research Graph Dump:** the entire research graph compressed as a dump.
- **The OpenAIRE COVID-19 dump:** contains the outcome of COVID-19-related research.
- **The dump of funded products:** research products including funding data.
- **The dumps about research communities, initiatives and infrastructures:** the partners collaborating with OpenAIRE.
- **The dump of ScholeXplorer:** a GZ-compressed dump of the ScholeXplorer⁴⁰, the explorer for the Scholix, the interoperability initiative to link scientific literature and research data⁴¹.

³⁶ EOSC Pilot: <https://eoscpiot.eu/>, accessed Sep. 26, 2022.

³⁷ OpenAIRE Mission and Vision: <https://www.openaire.eu/mission-and-vision>, accessed Sep. 26, 2022.

³⁸ OpenAIRE Research Graph: <https://graph.openaire.eu/>, accessed Sep. 26, 2022.

³⁹ OpenAIRE Research Graph Dumps: <https://graph.openaire.eu/develop/graph-dumps.html>, accessed Feb 24, 2022.

⁴⁰ ScholeXplorer: <https://scholexplorer.openaire.eu/#/>, accessed Sep. 26, 2022.

⁴¹ Scholix: <http://www.scholix.org/>, accessed Sep. 26, 2022.

- **The dump of DOIBoost:** this is an extension of Crossref, an initiative for easy access to research objects⁴².

We selected the whole research graph dump to create a connector to OpenAIRE. This dump is further subdivided into the 8 TAR dumps, covering publications, datasets, software, other research products, organizations, projects, and relations. The dump “dataset.tar” was the most interesting for us, since TRUSTS aims to become a digital marketplace for the exchange of datasets.

A software module serving as OpenAIRE connector was created in task 3.3. The module applies the ETL process described in Section 3.2.1 Connector Architecture. In the first step, the **extract** step, it acquires the dump of the dataset, which is a 10.2 GB tar file available from Zenodo page of the OpenAIRE research graph dump⁴³, which unpacks to a set of 169 GZ-compressed archives. Each archive contains several hundred MBs of data, available in JSON format. The fact that the data is available in JSON format facilitates further processing steps, since JSON can be conveniently transferred into Python dictionaries, which makes the implementation of a software component easier. In the next **transform** step, this data is transformed into the TRUSTS metadata schema using a customised mapping. The details of this metadata mapping are explained in detail in section 4.3 “Metadata Mapping”. After the completion of mapping, the metadata of datasets is transferred into the TRUSTS platform. This **load** step consists of loading the data into the local TRUSTS deployment and the subsequent push of the data to the central node of the platform. This last step makes the information about OpenAIRE datasets and the offerings of OpenAIRE available to all partners involved in TRUSTS.

3.2.3 Europeana

We selected Europeana as the second EOSC initiative for interoperability because of its data richness and the straightforward access to its raw content. Europeana is listed in the EOSC marketplace⁴⁴. Europeana is an effort to make digital cultural heritage material of Europe accessible⁴⁵. It connects over 4000 cultural institutions via a network of aggregating partners. The goal is to share Europe’s rich culture, history, and heritage, and to foster innovation around cultural heritage. For example, educators can use the material to create teaching material. In its own words,

“Europeana empowers the cultural heritage sector in its digital transformation. We develop expertise, tools and policies to embrace digital change and encourage partnerships that foster innovation.”⁴⁶

The digital data provided is quite substantial and covers over 50 million items of images, audio, text, videos, and 3D material⁴⁷. Europeana gives access to 30,122,687 images, 22,670,770 documents, 770,676 sound items, 338,492 videos, and 5,412 3D items from a diverse set of categories, such as archaeology, fashion, or science⁴⁸. The material is provided as “collections” and “stories”, and also searchable via Europeana’s search engine (see Figure 6).

⁴² Crossref: <https://www.crossref.org/>, accessed Sep. 26, 2022.

⁴³ OpenAIRE Research Graph dump on Zenodod: https://zenodo.org/record/6616871#.YzGa_NJBzLo, accessed Sep. 26, 2022.

⁴⁴ EOSC Marketplace: <https://marketplace.eosc-portal.eu/>, accessed Sep. 26, 2022.

⁴⁵ Europeana “About Us”: <https://www.europeana.eu/en/about-us>, accessed Sep. 26, 2022.

⁴⁶ Europeana, “Our Mission”: <https://www.europeana.eu/en>, accessed Sep. 26, 2022.

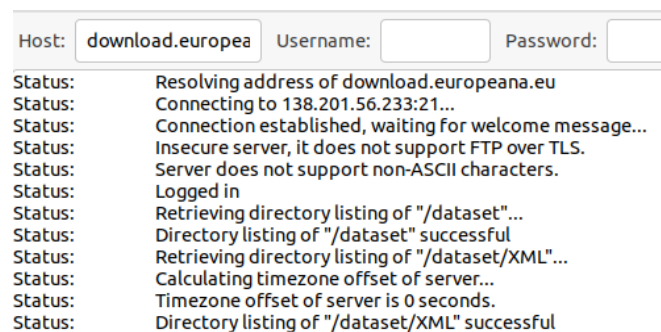
⁴⁷ *ibid.*

⁴⁸ Europeana “About Us”: <https://www.europeana.eu/en/about-us>, accessed Sep. 26, 2022.

Figure 6: Europeana's search functionality⁴⁹.

In addition to the web interface, Europeana also features multiple APIs to acquire the raw data: (i) the Europeana REST API, (ii) the Search API, (iii) the Record API, (iv) the Entity API, (v) Annotations API, (vi) IIIF APIs, (vii) SPARQL, (viii) Linked Open Data, (ix) Harvesting and Downloads⁵⁰.

We selected the latter option to load the metadata of Europeana's datasets into TRUSTS. This option allows bulk download of the entire content⁵¹. The data is available from an FTP server with open access. Any FTP software such as FileZilla⁵² can be used to download the data (see Figure 7).

Figure 7: Connection to Europeana's FTP server⁵³ using FileZilla.

⁴⁹ Europeana: <https://www.europeana.eu/en>, accessed Sep. 26, 2022.

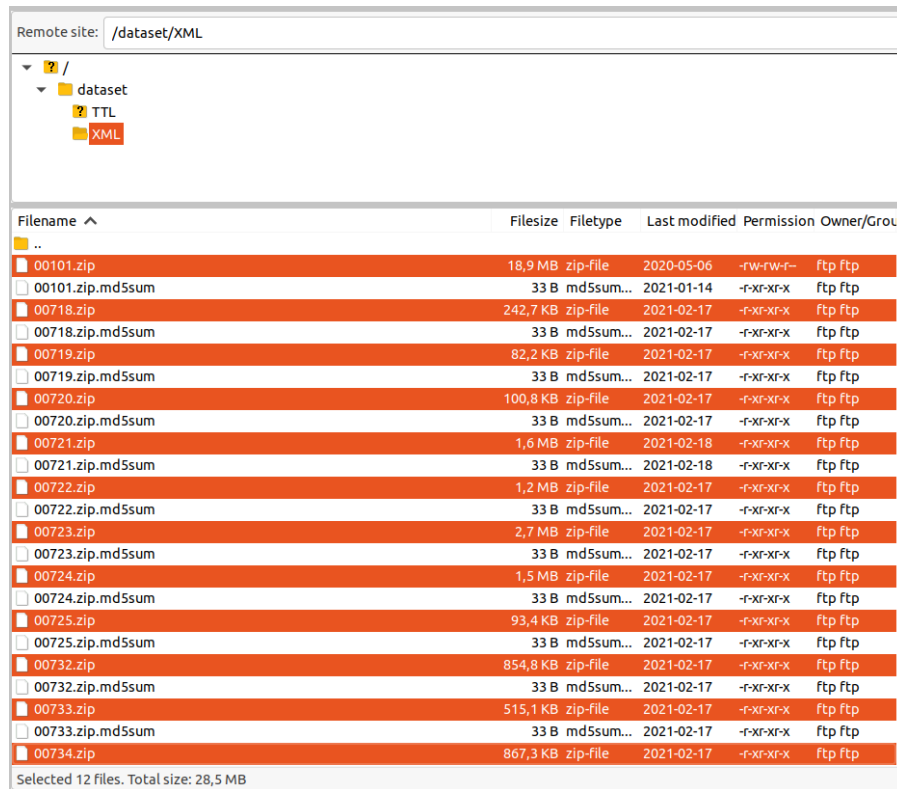
⁵⁰ Europeana APIs Overview: <https://pro.europeana.eu/page/apis>, accessed Sep. 26, 2022.

⁵¹ Europeana Harvesting and Downloads: <https://pro.europeana.eu/page/harvesting-and-downloads>, accessed Sep. 26, 2022.

⁵² FileZilla: <https://filezilla-project.org/>, accessed Sep. 26, 2022.

⁵³ Europeana FTP server: <ftp://download.europeana.eu/dataset/>, accessed April 19, 2022.

The access to the FTP server⁵⁴ is open, using the user name “anonymous” and a blank password⁵⁵. After logging in the data can be accessed either in XML or TTL format. The data in XML format is substantially more condensed with 450.8 GB vs. 504.9 GB in TTL format. The data is packed and compressed into roughly 2.000 ZIP archives. Each archive is provided with an MD5 checksum file to detect corrupted archives (see Figure 8).



Filename	Filesize	Filetype	Last modified	Permission	Owner/Group
00101.zip	18,9 MB	zip-file	2020-05-06	-rw-rw-r--	ftp ftp
00101.zip.md5sum	33 B	md5sum...	2021-01-14	-r-xr-xr-x	ftp ftp
00718.zip	242,7 KB	zip-file	2021-02-17	-r-xr-xr-x	ftp ftp
00718.zip.md5sum	33 B	md5sum...	2021-02-17	-r-xr-xr-x	ftp ftp
00719.zip	82,2 KB	zip-file	2021-02-17	-r-xr-xr-x	ftp ftp
00719.zip.md5sum	33 B	md5sum...	2021-02-17	-r-xr-xr-x	ftp ftp
00720.zip	100,8 KB	zip-file	2021-02-17	-r-xr-xr-x	ftp ftp
00720.zip.md5sum	33 B	md5sum...	2021-02-17	-r-xr-xr-x	ftp ftp
00721.zip	1,6 MB	zip-file	2021-02-18	-r-xr-xr-x	ftp ftp
00721.zip.md5sum	33 B	md5sum...	2021-02-18	-r-xr-xr-x	ftp ftp
00722.zip	1,2 MB	zip-file	2021-02-17	-r-xr-xr-x	ftp ftp
00722.zip.md5sum	33 B	md5sum...	2021-02-17	-r-xr-xr-x	ftp ftp
00723.zip	2,7 MB	zip-file	2021-02-17	-r-xr-xr-x	ftp ftp
00723.zip.md5sum	33 B	md5sum...	2021-02-17	-r-xr-xr-x	ftp ftp
00724.zip	1,5 MB	zip-file	2021-02-17	-r-xr-xr-x	ftp ftp
00724.zip.md5sum	33 B	md5sum...	2021-02-17	-r-xr-xr-x	ftp ftp
00725.zip	93,4 KB	zip-file	2021-02-17	-r-xr-xr-x	ftp ftp
00725.zip.md5sum	33 B	md5sum...	2021-02-17	-r-xr-xr-x	ftp ftp
00732.zip	854,8 KB	zip-file	2021-02-17	-r-xr-xr-x	ftp ftp
00732.zip.md5sum	33 B	md5sum...	2021-02-17	-r-xr-xr-x	ftp ftp
00733.zip	515,1 KB	zip-file	2021-02-17	-r-xr-xr-x	ftp ftp
00733.zip.md5sum	33 B	md5sum...	2021-02-17	-r-xr-xr-x	ftp ftp
00734.zip	867,3 KB	zip-file	2021-02-17	-r-xr-xr-x	ftp ftp

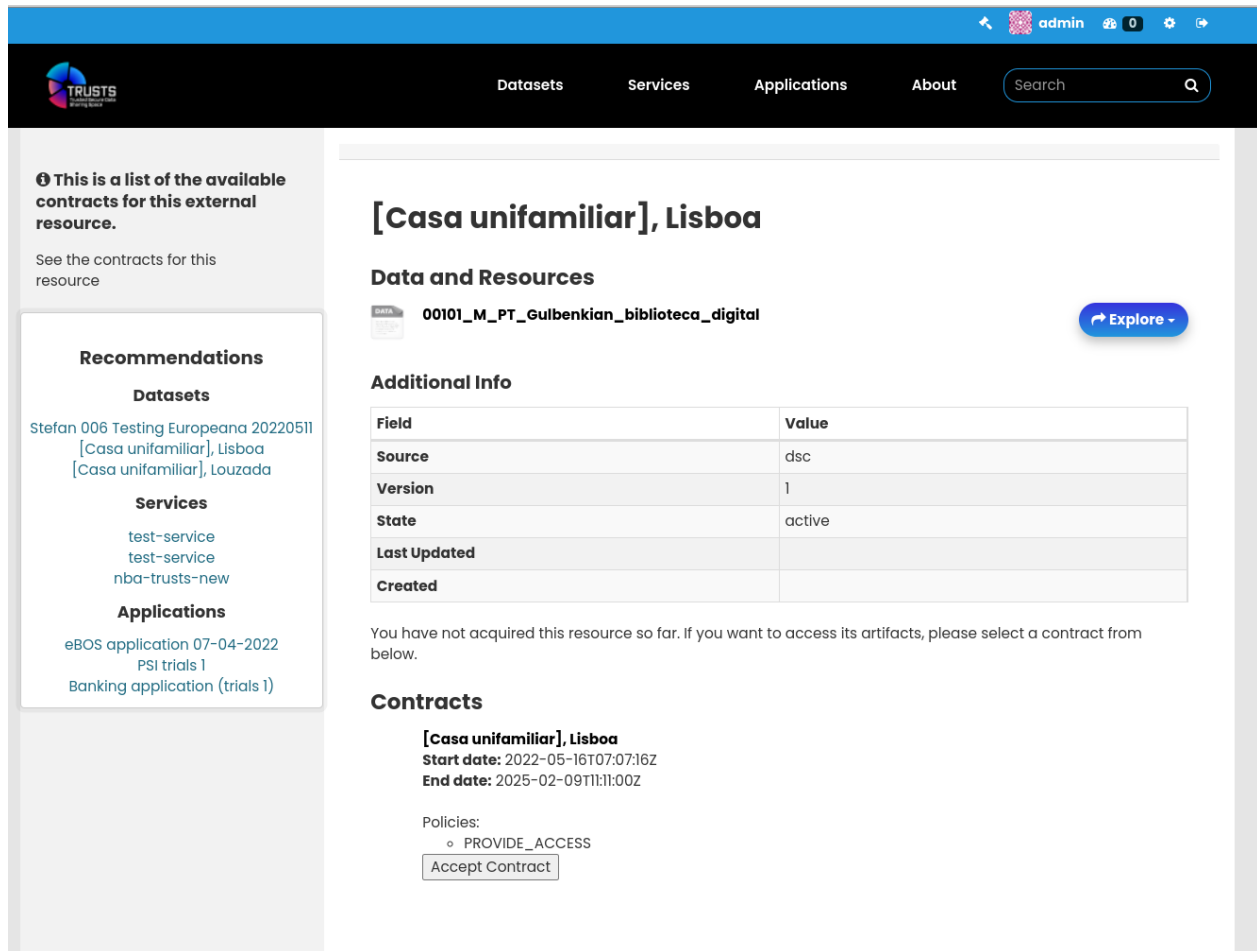
Selected 12 files. Total size: 28,5 MB

Figure 8: Europeana metadata files in zipped form.

We developed a software module that serves as a connector to Europeana. This Europeana connector applies the same ETL process as described in Section 3.2.1 Connector Architecture. In the **extract** step the data is loaded from Europeana, using a Python implementation for connecting to FTP servers. The **transform** step converts the Europeana metadata into the TRUSTS schema. Details for this metadata mapping follows in section 4.3 “Metadata Mapping”. After metadata conversion the data is loaded into the TRUSTS platform, using the TRUSTS platform client. Once the metadata of the dataset has been loaded into the TRUSTS platform and pushed to the central node, it is visible for everyone hosting a TRUSTS node. We created an organization “Europeana” in the TRUSTS platform, where all datasets from Europeana are available (Figure 9 shows the screenshot of an example).

⁵⁴ *ibid.*

⁵⁵ Europeana FTP Credentials: <https://pro.europeana.eu/page/harvesting-and-downloads>, accessed Sep. 26, 2022.



The screenshot shows the TRUSTS platform interface. The top navigation bar includes 'Datasets', 'Services', 'Applications', and 'About', along with a search bar and a user profile 'admin'. The main content area displays the dataset '[Casa unifamiliar], Lisboa' with the identifier '00101_M_PT_Gulbenkian_biblioteca_digital'. A table under 'Additional Info' lists fields like Source, Version, State, Last Updated, and Created. A 'Contracts' section shows the start and end dates and a policy 'PROVIDE_ACCESS' with an 'Accept Contract' button. A sidebar on the left provides recommendations for datasets, services, and applications.

[Casa unifamiliar], Lisboa

Data and Resources

00101_M_PT_Gulbenkian_biblioteca_digital [Explore](#)

Additional Info

Field	Value
Source	dsc
Version	1
State	active
Last Updated	
Created	

You have not acquired this resource so far. If you want to access its artifacts, please select a contract from below.

Contracts

[Casa unifamiliar], Lisboa
Start date: 2022-05-16T07:07:16Z
End date: 2025-02-09T11:11:00Z

Policies:

- PROVIDE_ACCESS

[Accept Contract](#)

Recommendations

Datasets

Stefan 006 Testing Europeana 20220511
 [Casa unifamiliar], Lisboa
 [Casa unifamiliar], Louzada

Services

test-service
 test-service
 nba-trusts-new

Applications

eBOS application 07-04-2022
 PSI trials 1
 Banking application (trials 1)

Figure 9: A dataset from Europeana published in the TRUSTS platform.

4 Semantic Interoperability

This section covers the efforts taken with regards to semantic interoperability in TRUSTS. Semantic interoperability ensures that the metadata schema of TRUSTS is flexible and diverse enough to cater for the needs of a multitude of technical solutions of participating partners. Each participating organisation potentially has its own technical solution, e.g., a platform, where they host and provide their data. They use their own taxonomies and naming conventions for data items and might even share datasets that are completely different to each other content-wise. For instance, a dataset in one domain, such as biology, might require a different metadata schema than a dataset in nuclear physics. Even datasets in one and the same domain might have different naming conventions, e.g., “name” might be used for the identifying term of a dataset, but also “title”.

This situation called for a solution that offers flexibility when exchanging metadata of datasets. The research documented in the previous deliverable D3.5 showed that the adoption of the EDM (EOSC Minimum Datasets Information) is beneficial for the TRUSTS metadata schema. TRUSTS uses a modified version of the IDS-IM (International Data Spaces - Information Model)⁵⁶ as the basis of its own information model, the TRUSTS-IM (see TRUSTS deliverables D3.7 and D3.8 for details). The IDS-IM was designed for data representation in “International Data Spaces” (IDS)⁵⁷, a concept for organisations to exchange data. The goal is to create mutual benefit from the shared data and encourage trading of data. The IDS is driven forward by the IDSA (International Data Spaces Association)⁵⁸. The IDSA has supported the creation of the IDS-IM, which was led by the two Fraunhofer Institutes FIT⁵⁹ and IAIS⁶⁰. The IDS-IM is meant for “... describing, publishing and detecting data products (Data Assets) and reusable data processing software (Data Apps) ...”⁶¹ and it “... is a generic model, with no commitment to any particular domain.”. This makes it a perfect fit for TRUSTS, which is a platform for the exchange of data assets, i.e., datasets or data services without any constraints with regards to the domain. The IDS-IM was used as a basis for the TRUSTS metadata schema “TRUSTS-IM”. To further increase the flexibility of the TRUSTS-IM, we decided to also include EDM, which has a special focus on the requirements of the EOSC, i.e., the scientific domain.

The EDM was developed in the EOSCpilot⁶², one of the first EOSC-related projects. It is a set of metadata properties to make datasets searchable. The properties of the EDM are specifically targeting the scientific area and cover aspects such as “measurementTechnique”, “scientificType”, “metric”, “citation”, or “referenceCitation”. Figure 10 shows an extract of the EDM.

⁵⁶ International Data Spaces - Information Model: <https://github.com/International-Data-Spaces-Association/InformationModel>, accessed Sep. 27, 2022.

⁵⁷ International Data Spaces Association: <https://internationaldataspaces.org/>, accessed Sep. 27, 2022.

⁵⁸ *ibid.*

⁵⁹ Fraunhofer Institute for Applied Information Technology FIT: <https://www.fit.fraunhofer.de/en.html>, accessed Sep. 27, 2022.

⁶⁰ Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS: <https://www.iais.fraunhofer.de/en.html>, accessed Sep. 27, 2022.

⁶¹ International Data Spaces Information Model Ontology: <https://international-data-spaces-association.github.io/InformationModel/docs/index.html#>, accessed Sep. 27, 2022.

⁶² EOSCpilot: <https://eoscipilot.eu/>, accessed Sep. 27, 2022.

measurementTechnique	A technique or technology used in a dataset corresponding to the method used for measuring the corresponding variables	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No guideline yet
keywords	Keywords or tags used to describe the dataset	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Keywords
variablesMeasured	The variables that are measured in the dataset	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No guideline yet
format	The format in which the content of the dataset is encoded to present the information, typically a MIME format	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Medium / Format
scientificType	Scientific domain or type of the information provided in the dataset	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No guideline yet
includes	A dataset or data catalog contained in the dataset	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No guideline yet
contentType	Type of content provided in the dataset based on its origin and type of processes (raw, processed, summarised)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No guideline yet
size	Size of the dataset using a digital information multiple unit byte symbol (MB, GB, PT, ...)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No guideline yet
authentications	Type of authentication required to access the dataset	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	No guideline yet
version	The version of the dataset	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	No guideline yet
metric	Metric to provide some quantitative or qualitative information about the dataset	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	No guideline yet

Figure 10: An extract of the properties of the EDM⁶³.

4.1 Metadata Mapping

Metadata mapping is the process of merging two metadata schemas to unify differences in the two schemas. Differences in metadata schemas are common, since they are usually defined by humans who understand different things under the same name or the same thing under different names. Even though the TRUSTS-IM has been designed with flexibility in mind, it cannot cater for all requirements of metadata schemas. When loading metadata into TRUSTS, it is therefore necessary that the external schema, i.e. the metadata schema of the third party interacting with TRUSTS, is transformed into the schema used by TRUSTS. This step is called metadata mapping. The **transform** step of the OpenAIRE and Europeana connector uses metadata mapping to transform those two external schemas into the TRUSTS schema.

For example, the TRUSTS property “name” is known as “maintitle” in OpenAIRE and as “datasetName” in Europeana. Furthermore, TRUSTS knows a “title” property, which is unknown to OpenAIRE and Europeana. Thus, and for future flexibility, we map OpenAIRE’s “maintitle” and Europeana’s “datasetName” to both TRUSTS’ “name” and “title” property. Another example is the TRUSTS property “resources:rights”: this property has the same name in Europeana, “edm:rights”, but is called “bestaccessright::code” in OpenAIRE. Consequently, equivalent tables have to be created for each third party aiming to interoperate with TRUSTS. Table 2 gives an overview of the metadata mapping accomplished by the OpenAIRE and Europeana connectors. Europeana required a special pre-processing step, since its data is not available as JSON, but as XML and TTL instead. We decided to extract the metadata from the XML, because it is easier

⁶³ EDM Metadata Properties: <https://eosc-edmi.github.io/properties>, accessed Sep. 27, 2022.

than extracting it from TTL. However, this requires an additional component that is apply to traverse the XML tree. We used the free Python XML toolkit lxml⁶⁴ for this purpose. By using the XPath notation, a notation to access nodes in XML documents, we are able to traverse the XML tree of the Europeana documents and access relevant nodes and the respective information in those nodes⁶⁵.

Table 2: TRUSTS properties and their equivalents in OpenAIRE and Europeana.

TRUSTS property	OpenAIRE equivalent	Europeana equivalent (XPath notation)
name	maintitle	string(edm:EuropeanaAggregation/edm:datasetName)
title	maintitle	string(edm:EuropeanaAggregation/edm:datasetName)
notes	description	string(ore:Proxy/dc:description)
owner_org	publisher	Pre-set: Europeana
resources::rights	bestaccessright::code	string(ore:Aggregation/edm:rights)
resources::url	url	string(edm:WebResource/@rdf:about)
resources::name	maintitle	string(edm:EuropeanaAggregation/edm:datasetName)
resources::dataProvider	publisher	string(ore:Aggregation/edm:rights)
resources::created	publicationdate	string(dqv:QualityAnnotation/dcterms:created)
resources::remoteId	id	string(ore:Proxy/dc:identifier)

Europeana uses EDM (Europeana Data Model)⁶⁶ as one of their metadata schemas, as shown in the example of the “rights” property of Europeana, which is “ore:Aggregation/edm:rights” in XPath notation. Europeana also uses other established namespaces such as the Dublin core⁶⁷ or the OAI-ORE terms vocabulary (Open Archives Initiative - Object Reuse and Exchange)⁶⁸, as seen in the example before in the stub “ore:Aggregation”. Europeana has an extensive set of classes and properties, e.g. “edm:EuropeanaAggregation” object, which represents Europeana items referring to the same cultural heritage object, or the “edm:WebResource”, pointing to the digital address of a cultural heritage object. Figure 11 shows an overview of properties for the aforementioned EuropeanaAggregation object.

⁶⁴ lxml: <https://lxml.de/>, accessed Sept. 26, 2022.

⁶⁵ XPath: https://www.w3schools.com/xml/xpath_syntax.asp, accessed Sep. 26, 2022.

⁶⁶ Europeana Data Model <https://pro.europeana.eu/page/intro#edm>, accessed Sep. 26, 2022.

⁶⁷ Dublin core: <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>, accessed Sep. 26, 2022.

⁶⁸ OAI-ORE: <http://www.openarchives.org/ore/1.0/vocabulary>, accessed Sep. 26, 2022.

EDM Property	Datatype	#	Description
dc:creator	Literal	0..1	A creator definitions. This field has always the value "Europeana".
edm:aggregatedCHO	Reference	1..1	The ID of the record corresponding to the CHO of this aggregation
edm:collectionName	Literal	1..1	<i>This property is deprecated and edm:datasetName should be used instead.</i>
edm:datasetName	Literal	0..1	This property holds the identifier given to the dataset in Europeana.
edm:country	Literal	1..1	This is the name of the country in which the Provider is based or "Europe" in the case of Europe-wide projects.
edm:hasView	Reference	0..*	This property relates a ORE aggregation about a CHO with a web resource providing a view of that CHO. Examples of view are: a thumbnail, a textual abstract and a table of contents.
edm:isShownBy	Reference	0..1	An unambiguous URL reference to the digital object on the provider's web site in the best available resolution/quality.
edm:landingPage	Reference	0..1	The URL of Europeana HTML object page. It captures the relation between an aggregation representing a cultural heritage object and the Web resource representing that object on the provider's web site.

Figure 11: Metadata properties for an aggregation object, i.e., an individual cultural heritage object⁶⁹.

⁶⁹ Europeana REST API: <https://pro.europeana.eu/page/intro#edm>, accessed Feb 22, 2022.

5 Organisational Interoperability

This chapter covers two aspects: (i) the interoperability experiment we conducted to see the feasibility of the interoperability solutions, and (ii) our work on the development of a smart contract client. The main work on smart contracts is covered in Task 3.2. In Task 3.3 we created a smart contract client including a metadata mapper, which allows to persist data from the smart contract blockchain in the TRUSTS database and, vice versa, that transactions accomplished in the platform are persisted in the blockchain. The metadata mapping resembles interoperability between two differing components, which is why we are reporting about this activity at this place.

5.1 Interoperability Experiment

In the following, we describe our experiment of interoperability between two TRUSTS markets. The goal of the experiment was to show the processes and demonstrate the software modules we have created for the case that an external data market wants to interoperate with TRUSTS, e.g., after the project lifetime. For that purpose, we decided to simulate such a case and deploy an additional instance of TRUSTS, which functioned as the external data marketplace. By using such a “clone” deployment, we were able to simulate the aspects required for asset exchange, while retaining full control over the entire system. The additional TRUSTS instance was completely separated from the main TRUSTS deployment and contained its own datasets. In the experiment, the operator of this simulated marketplace aimed to make available their offerings from within the main TRUSTS platform. They selected datasets and using a Python module developed for that purpose as well as the TRUSTS platform client, were able to map the metadata of those datasets into the main TRUSTS platform. Figure 12 shows a visual overview of the experimental setup.

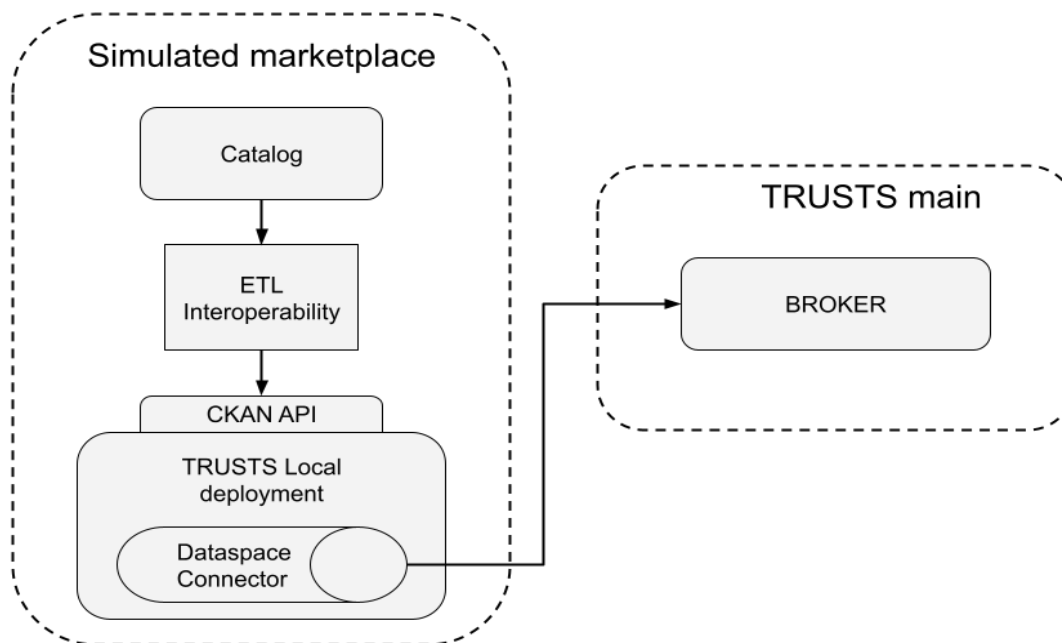


Figure 12: Setup of the interoperability experiment.

5.2 Smart Contracts

Smart contracts are used in TRUSTS to persist transactions taken in the platform, e.g., when sharing or accessing datasets. The core work is accomplished in Task 3.2 and reported in deliverable D3.3 “Smart Contracts”. The work accomplished for smart contracts in Task 3.3 was to make the blockchain component, i.e., the smart contract demonstrator, interoperable with TRUSTS. This required a metadata mapping between the schemas of both components.

The smart contract demonstrator uses Hyperledger Fabric, a software for the creation of distributed ledgers⁷⁰. It is used as the basis for the smart contract blockchain used in TRUSTS (for details see deliverable D3.3 “Smart Contracts”). In order to persist transactions that are accomplished in the TRUSTS platform in the blockchain, the platform has to communicate with the blockchain. Communication is done using the REST API endpoints the blockchain application offers. The TRUSTS platform needs to contact these endpoints when a transaction is happening, for which we developed a dedicated software. This software, the so-called smart contract client, is called when a transaction is accomplished and contacts the blockchain to inform it about the transaction. Furthermore, it passes all required data of the transaction to the blockchain. The blockchain component, in turn, returns an output signal, which is transferred back to the platform and stored in its database. This allows to show the smart contract information in the user interface of TRUSTS.

The metadata schema of the blockchain differs from the main TRUSTS metadata schema. This required the creation of a mapping, to make both schemas compatible with each other. Table 3 shows examples of this mapping. For example, the TRUSTS database knows an “author” of a dataset, which is called a “creator” in the blockchain component, or the “id” is called “assetid”.

Table 3: Examples of metadata mapping between the TRUSTS database and the smart contract component.

TRUSTS property	Smart contract equivalent
author	creator
author_email	contactPoint
id	assetid
license_title	license
maintainer	publisher
metadata_created	creationDate
name	title
notes	description

⁷⁰ Hyperledger Fabric: <https://www.hyperledger.org/use/fabric>, accessed Sep. 28, 2022.

6 Legal Interoperability

6.1 Introduction

Legal interoperability refers to the ability of entities subject to various legal frameworks – that may differ because of substantive and/or geographic scope – to achieve joint goals and work together smoothly. In the context of data marketplaces, legal interoperability is key to the extent that the data, as well as the entities processing and/or holding them, are regulated by a plethora of legal frameworks that vary in terms of subject matter and geography.

This document acknowledges the importance of legal interoperability for the endeavours of the TRUSTS project. Based on the principles elaborated by EOSC, the document discusses the main pain points concerning TRUSTS related to legal interoperability and proposes avenues for solutions, including solutions that have already been researched on and envisaged within the project.

6.2 Potential pain points

This section describes the aspects related to legal interoperability that have been investigated with regard to the databases envisaged to be shared and traded through the TRUSTS platform. The aspects have been broken down into the following two categories: intellectual property law; and privacy and data protection law.

6.2.1 Legal aspects related to intellectual property law

Intellectual property law, and especially copyright law, plays a crucial role in the possibilities to share databases that are protected by either EU law on copyright or EU law protecting databases. Intellectual property law protection may be one of the most important protection frameworks that databases shared on the TRUSTS platform may rely on, aside from e.g., trade secrecy law.

The intellectual property law protection is usually laid down in licences accompanying a given database or data set. Variability of data set licences are one significant hindrance to legal interoperability. As pointed out in the EOSC Interoperability Framework, the furthermore far apart the conditions of two data sets are, the more unlikely it is for a user to be able to combine the two data sets in a new data set under a licence compatible with both initial data sets. Variability may also be worsened by the differences in copyright law across Member States, which may have consequences on the degree of reusability of various databases.

The potential divergence or incompatibility between licences may arise in two different scenarios, which are based on a slightly different understanding of ‘legal interoperability’: a) the attempt to combine and/or improve on two or more data sets uploaded to the TRUSTS platform; and b) the attempt to combine and/or improve on the data sets uploaded to the TRUSTS platform and data sets from other data marketplaces.

Another interesting aspect pertains to the territorial scope of EU law protecting databases and to the claim that users from within the EU need to acquire the rights to reuse a database whereas users from outside the EU would not. However, in actuality, the Database Directive should not be given this interpretation; rather, the Directive confers protection to databases originating in the EU and/or created by companies

established or having a genuine link with economic activities in the EU. In such a scenario, therefore, a potential source of uncertainty may pertain to non-EU entities who would like to participate in the TRUSTS platform with their database but would not be afforded the same legal protection as EU databases. This might prevent business opportunities linked to the participation of non-EU entities.

6.2.2 Legal aspects related to privacy and data protection law

The EU and national data protection framework, grounded in EU constitutional texts, in the General Data Protection Regulation (GDPR), and in national law, may also constitute a hindrance to data reuse. This is because 'reuse' implies transferring and/or additional operations done on the data set, all of which constitute data processing pursuant to Article 4(2) GDPR. Since the GDPR encourages as little data processing operations as possible, in compliance with the purpose limitation and data minimisation principles, from a conceptual perspective it sits at odds with the goal of promoting data sharing and data reuse to the extent that the data sets concerned contain personal data.

Here the legal interoperability angle is visible from two distinct perspective: a) a data use perspective: if a user intended to combine and improve on two separate data sets, of which only one contains personal data, the compatibility may be low as the personal data set would have more constraints to sharing than the data set without personal data; b) a regulatory perspective: compliance with the terms of EU data protection law 'trumps' any reuse ambition, hence the intra-users legal interoperability needs to yield to the compatibility of the data marketplace and its practices with the law.

6.3 Research avenues for solutions

This chapter attempts to shed some light on research avenues that may supply solutions to prevent and solve issues of legal interoperability both between data sets from within TRUSTS as well as between the TRUSTS platform and other data marketplaces.

6.3.1 Intellectual property law

It can be argued that legal interoperability concerning data set licences can be enhanced in two steps: first, enhance clarity; second, promote harmonisation.

1. *Enhance clarity.* In the first step, all data holders uploading data sets to the TRUSTS platform are encouraged to be clear about the licences regulating the reuse of each data set. This also has an impact on the actual terms of reuse that are to be laid down in the smart contract regulating data set exchanges between two or more parties. This step is also likely to help a great deal cope with the intra-EU variability of copyright laws, i.e., an aspect that can only be dealt with substantially through legislative reforms, but on which clear and transparent licences can at least shed light to prevent the risk of misuse;
2. *Promote harmonisation.* In the second step, that goes forward compared to the first one, data holders are also encouraged to reduce the variability of their licences and align as much as possible to existing standards to increase the likelihood of reuse.

Both steps can in principle be useful in intra-TRUSTS scenarios as well as in scenarios where TRUSTS relates to other data marketplaces. The main difference that is likely to arise concerns the governance of the

platforms and the level of complexity required to push these changes through: in the case of TRUSTS, the governance being unique, clarity and harmonisation can be promoted by a sole actor by adopting specific policies; whereas the TRUSTS governance – aside from common fora and initiatives – has in principle no influence on the policies adopted by other marketplaces.

Even within TRUSTS, an important aspect that should be investigated is the downstream effect that a tightening of licensing policies might have on the platform's business opportunities. If TRUSTS were to tighten the licensing requirements and demand, e.g. licences on data sets to be significantly more relaxed, several organisations unwilling to take a step back on their IPRs might choose not to use the TRUSTS platform at all. This may be either because these organisations may have no interest in relaxing their licences; or because, even if such an interest existed, as these organisations sit outside the TRUSTS consortium, the TRUSTS governance structure would not have the means to influence their policy decisions.

Because of its ambition to make data trading possible, the TRUSTS platform might therefore be confronted with a trade-off: either narrow down the scope of admissible data set licences to come as close as possible to the principle of openness, potentially reducing compatibility and reusability issues but discouraging some organisations from joining; or limit its ambitions in terms of openness, thereby leveraging those business opportunities, but exposing itself to more divergences in licensing. It can be argued, however, that having divergences across licences – albeit not desirable in itself as it thwarts reuse – might still be preferable to foreclosing the possibility of reuse in the first place which would risk being the outcome of the former approach. In other words, to be sure, allowing typically restrictive licensors to join TRUSTS may bring data sets that are difficult to make interoperable with more open data sets, hence skew reuse and business opportunities stemming from data set combination. However, it may still be preferable to apply the latter approach which runs the danger that these opportunities stemming from combination, rather than difficulty, become non-existent because the more restrictive data sets held by the organisations discouraged from joining TRUSTS would never reach the platform in the first place.

It can be argued that this would be the case because the trade-off mentioned above is unlikely to be a 'black-or-white' type of choice. It is a matter of degree both in how 'narrow' the scope of the allowed licensing is defined by TRUSTS, and in how inclined other organisations are to still make business through the TRUSTS platform given the licensing scope decided at policy level by TRUSTS. A compromise should therefore be sought that balances the need to promote open licences with the need to make open data sets coexist with data sets composed of IPR-protected data. The concepts of 'fairness' and 'FAIR data' may help a great deal in this regard as guiding principles slightly different from the concept of 'open data'. Transparent licences and data sharing policies constructed to maximise reuse and accessibility are perfectly compatible with the idea behind FAIR data, without this data having to be 'open' per se.

Regarding the pain point mentioned above on the **protection of non-EU databases**, it goes without saying that legislation would be the main source for change. However, the legislative framework does not prevent, in itself, private actors from constructing a trust-based licensing system that treats foreign databases on equal terms as EU databases. This should be a platform-level policy decision to make, whereby the potential benefits of such a decision should be weighed against the potential cost that this equality may represent for some EU-based entities who would like to participate in the data marketplace under the privileged regime granted by the EU Database Directive. Moreover, the actual extent of the problem should be evaluated prior to making decisions: according to a 2020 study to support the

evaluation of the Database Directive,⁷¹ more than 80% of the stakeholders consulted did not see the territorial scope of the Directive as a source of legal uncertainty.⁷²

6.3.2 Privacy and data protection law

Taking into account the privacy and data protection challenges mentioned above, in the context of work packages 4 and 6 TRUSTS partners investigated privacy-enhancing methods and their application to the data sets envisaged to be shared via the TRUSTS platform.

Deliverables D4.1 and D4.2 dive into the technology and rationale behind the privacy-enhancing methods chosen for TRUSTS; Deliverable D6.3 assesses their capability of making the TRUSTS platform comply with relevant data protection law. The preliminary conclusion is that the privacy-enhancing technologies (PETs) envisaged in TRUSTS are likely to induce compliance with data protection legislation. This is going to be achieved by a) providing data anonymisation techniques to organisations willing to share their data sets containing personal data; b) applying a Federate Learning (FL)-based Machine Learning (ML) model to the data analytics provided by the TRUSTS platform in order to minimise the number of data processing operations and the need for data; and c) combining the FL model with modern cryptographic techniques (such as Multi-Party Computation and Homomorphic Encryption) to further shield the data processing from privacy leaking.

In this context, for TRUSTS to keep being compatible with relevant legislation it will be essential to monitor the performance of the approaches and PETs chosen to enhance privacy especially vis-à-vis the development of de-anonymisation technologies potentially capable of breaching advanced privacy-enhancing methods.

⁷¹ European Commission (DG CNECT), *Study in support of the evaluation of Directive 96/9/EC on the legal protection of databases*, 2020. Available at: [Study-in-Support-of-the-Evaluation-of-the-Database-Directive.pdf \(technopolis-group.com\)](#).

⁷² Ibid., p. 9.

7 Conclusions and Next Actions

This report describes the efforts taken in task 3.3 “Data marketplaces interoperability solutions” and is the last deliverable in a series of three deliverables (D3.4, D3.5, D3.6). TRUSTS aims at interoperability with third-party data marketplaces and the EOSC (the latter being specified in the Grant Agreement with the statement “...interoperability solutions with the European Science Cloud (EOSC) will be evaluated and implemented, where possible.”⁷³). The work in this task has delivered software artefacts for those two aspects, i.e., software for external market operators to load the metadata of their data into the TRUSTS platform using an ETL process. Furthermore, we created two connectors for two EOSC initiatives to pre-load TRUSTS with datasets from those initiatives.

The work in this deliverable is divided into four main sections, covering “technical”, “semantic”, “organisational”, and “legal” interoperability. This categorization is aligned with existing research in interoperability, it has been adopted from the EIF (“New European Interoperability Framework”⁷⁴), which uses the same four categories.

Technical interoperability (see Section 3 Technical Interoperability) are actual software artefacts that have been created in this task to implement interoperability, specifically the TRUSTS platform client and an interoperability solution for the mapping of metadata catalogues between TRUSTS and a potential third-party marketplace. Furthermore, we developed two connectors for the EOSC initiatives OpenAIRE and Europeana. OpenAIRE makes existing research work and output available and searchable. Europeana is a cultural heritage platform, aggregating digital information about cultural heritage objects from several thousand institutions within Europe. The relevance of those two initiatives, the large amount they provide, and their existing APIs have been the reasons for selecting them.

Semantic interoperability (see Section 4 Semantic Interoperability) is concerned with the requirements of the TRUSTS metadata schema to be flexible enough to interoperate with external platforms and markets. The TRUSTS-IM has the IDS-IM as a basis, which is a domain-agnostic schema. Furthermore, the TRUSTS-IM has been extended with EDM1 to further increase its flexibility. We also describe a metadata mapping functionality, which converts the properties in external metadata schemas into the format that TRUSTS understands and is compatible with.

Organisational interoperability (see Section 5 Organisational Interoperability) is a showcase on interoperability with a third-party datamarket. We simulated an external data marketplace by creating an own, individual, and separated TRUSTS instance. We demonstrate how the mapping of metadata between the catalogues can be accomplished using a dedicated software module.

Legal interoperability (see Section 6 Legal Interoperability) gives insights on the legal aspects concerning interoperability. Potentially, operators from countries with strongly differing legal frameworks could come together and work, exchange, and trade data via the TRUSTS platform. This diversity must be considered legally, for instance with regards to the selection of appropriate licences for data trade.

Business benefits. An interoperable data marketplace offers new commercial and business opportunities for its operators, partners, and participating organisations. A data-driven economy becomes lively when data sharing and trading is easy, and when all participating partners see new business opportunities and

⁷³ Task 3.3 Definition of work, TRUSTS Trusted Secure Data Sharing Space grant agreement.

⁷⁴ The New Interoperability Framework: https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf, accessed Sep. 23, 2022.

room for innovations. For instance, companies could trade data that they possess but do not currently actively work with. The reason might be a lack of resources to exploit the data, or merely the fact that the knowledge for an appropriate data exploitation is missing, e.g., when the company does not have the required expertise in data science, artificial intelligence, machine learning, statistical analysis, etc. By putting their data on the market for sale they can attract other professionals, who recognize the business value of the data and create new innovations using this data. This situation creates mutual benefit, where both parties derive value from the data, where prior to that no value was generated.

Interoperability, as envisioned in Task 3.3, contributes to this benefit by increasing the overall amount of data that is available on the TRUSTS platform. It does this by (i) giving the means for commercial operators, such as existing third-party data marketplaces, to share and trade their data in TRUSTS, and by (ii) making available research outputs from publicly funded projects, i.e., the EOSC.

A digital data marketplace, such as TRUSTS, is able to attract more participants and customers when more and diverse data is available. Thus, interoperability with parties offering such data is important. Especially the EOSC, with its diverse offerings in research data, contributes to availability of large and diverse amounts of data. The advantage of including the EOSC is that the data is already readily available and covers a broad range of topics. Furthermore, since it is from publicly funded projects, the data is often available under a convenient licence, aiming at the wide-spread usage by interested parties. In this way, TRUSTS also contributes to the spread of the data and raises the awareness for it and can potentially also attract new research initiatives. On the other hand, having cutting-edge research data available in a digital data marketplace fosters data-driven innovation. Organisations can use this data for completely new innovations and generate business value with, either by a direct exploitation of the data or a merging with their own data.

Overlaps with other tasks and work packages. As stated in the Grant Agreement, Task 3.3 overlaps with T3.2 Smart Contracts, documented in Section 5.2 “Smart Contracts”, and T3.4 “Data Governance”, documented in Section 4 “Semantic Interoperability”. Furthermore, it overlaps with WP6 “Legal & Ethical Framework”, documented in Section 6 “Legal Interoperability” and WP7 “Business Model, Exploitation & Innovation Impact Assurance”. Our work in collaboration with the latter work package is documented in D3.4, Section 5.1 “DM Survey”.

Next actions. The work in Task 3.3 is finished with this deliverable. However, interoperability in TRUSTS is still a lively topic. Changing technical and semantic requirements, new market operators, and emerging EOSC initiatives make it crucial to continue development and align the TRUSTS with the state of the art. Also, legal frameworks change over time and might make it necessary to deal with more constrained or more relaxed requirements and licences.