DOI: 10.5281/zenodo.14501296

Deliverable D7.2: Data management plan (DMP)

Due Date of Deliverable	30.06.2023
Actual Submission Date	30.06.2023
Work Package	WP7
Tasks	T7.3
Type	R - document, report
Approval Status	Accepted
Version	v1.0
Number of Pages	81

Abstract

TThe deliverable elaborates on the data management plan for a set of active output Datasets for the SciLake project.

The DMP is implemented as a live document managed using the Argos DMP tool and important updates related to the evolution of the installations over the project lifetime will be uploaded on Zenodo (https://doi.org/10.5281/zenodo.8091822).

Revision history

Version	Date	Reason	Revised by
0.0		First Draft	Mary Melekopoglou
0.1		Peer Review	Stefania Amodeo
1.0		Final Version after proofreading	Mary Melekopoglou

Author List

Organisation	Name	Contact Information
ARC	Mary Melekopoglou	marmel@athenarc.gr
ARC	Thanasis Vergoulis	vergoulis@athenarc.gr
ARC	Serafeim Chatzopoulos	schatz@athenarc.gr

Contributor List

Organisation	Name	Contact Information
OpenAIRE	Stefania Amodeo	stefania.amodeo@openaire.eu

Table of Contents

Abbreviation List

- DMP   Data Management Plan
- FAIR   Findable Accessible Interoperable Reusable

1. Executive Summary

This document is a Data Management Plan (DMP) for SciLake project, developed in line with the guidelines in the Horizon Europe Data Management Plan template https://www.openaire.eu/images/Guides/HORIZON_EUROPE_Data-Management-Plan-Template.pdf. The policies define data archiving, preservation, and sharing (i.e. access rights) practices to be adopted.

2. Data Management Plan

The Data Management Plan was prepared using the Argos DMP tool (https://argos.openaire.eu) for each installation that requires such a plan. Argos allows researchers to create machine-actionable DMPs and ensure compliance with FAIR data handling. It enables collaboration on the definition of research data management plans in the context of projects or grants and subsequently the persistence, publication and exchange of those through a variety of mechanisms.

SciLake Data Management Plan, along with the dataset definitions for each installation, is publicly available at: https://doi.org/10.5281/zenodo.8091822

and is also available as an appendix below. The DMP along with the dataset descriptions defined for each installation will be kept up-to-date over the duration of the project to reflect possible policy changes due to ongoing evolution of project requirements.

1. Appendix-
Data Management Plan Information

1.2 SciLake Data Management Plan

Description

SciLake's mission is to build upon the OpenAIRE ecosystem and EOSC services to (a) facilitate and empower the creation, interlinking and maintenance of SKGs and the execution of data science and graph mining queries on top of them, (b) contribute to the democratisation of scholarly content and the related added value services implementing a community-driven management approach, and (c) offer advanced, AI-assisted services that exploit customised perspectives of scientific merit to assist the navigation of the vast scientific knowledge space. In brief, SciLake will develop, support, and offer customisable services to the research community following a two-tier service architecture. First, it will offer a comprehensive, open, transparent, and customisable scientific data-lake-as-a-service (service tier 1), empowering and facilitating the creation, interlinking, and maintenance of SKGs both across and within different scientific disciplines. On top of that, it will build and offer a tier of customisable, AI-assisted services that facilitate the navigation of scholarly content following a scientific merit-driven approach (tier 2), focusing on two merit aspects which are crucial for the research community at large: impact and reproducibility. The services in both tiers will leverage advanced AI techniques (text and graph mining) that are going to exploit and extend existing technologies provided by SciLake's technology partners. Finally, to showcase the value of the provided services and their capability to address current and anticipated needs of different research communities, four scientific domains (neuroscience, cancer research, transportation, and energy) have been selected to serve as pilots. For each, the developed services will be customised, to accommodate differences in research procedures, practices, impact measures and types of research objects, and will be validated and evaluated through real-world use cases.

Funder

European Commission||EC

Grant

Democratising and making sense out of heterogeneous scholarly content/ No 101058573

Researchers

Serafeim Chatzopoulos (orcid:0000-0003-1714-5225),
César Parra Rojas (orcid:0000-0003-3625-9412),
Fotis Psomopoulos (orcid:0000-0002-0222-4273),
Miriam Baglioni (orcid:0000-0002-2273-9004),
Daniel Hägerstrand (orcid:0000-0001-7270-0776),
Afroditi Anagnostopoulou (orcid:0000-0001-8292-2663),
Ingrid Reiten (orcid:0000-0002-9275-6475),
Vasileios Vasileiou (orcid:0000-0003-0145-033X),
Jakob Rager (orcid:0000-0003-4856-8140),
Sotiris Kotitsas (orcid:0000-0002-8114-6225),
Thanasis Vergoulis (orcid:0000-0003-0555-4128),
Trygve B. Leergaard (orcid:0000-0001-5965-8470),
Marek Horst (orcid:0000-0002-9038-9333),
Dimitris Pappas (orcid:0000-0001-5784-0658),
Claudio Atzori (orcid:0000-0001-9613-6639),
Xenophon Kitsios (orcid:0009-0000-7711-4960),
Leily Rabbani (orcid:0009-0007-0755-3643),
Nicolau Duran-Silva (orcid:0000-0001-5170-4129),
Haris Papageorgiou (orcid:0000-0002-7352-2403),
George Gavriilidis (orcid:0000-0003-2575-4354),
Pablo Accuosto (orcid:0000-0002-3493-6013),
Salmi Andrea (orcid:0000-0002-0305-7710)

Organisations

ATHENA Research Center,

OPIX IKE,

Centre for Research and Technology Hellas,

OPENAIRE AMKE,

University of Warsaw / ICM,

Karolinska Institutet, Stockholm, Sweden,

Consiglio Nazionale delle Ricerche - Istituto di Informatica e Telematica,

Ερευνητικό Πανεπιστημιακό Ινστιτούτο Συστημάτων Επικοινωνιών και Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου,

SIRIS Academic SL,

University of Oslo,

HES-SO Valais/Wallis

Dataset Title: OpenAIRE Full Texts

Template: Horizon Europe

Dataset includes all the openly available full texts, provided by the repositories registered in the OpenAIRE PROVIDE, which are linked to the OpenAIRE Research Graph entities bundled in a separate OpenAIRE Graph Dump dataset available at https://zenodo.org/record/7488618.

Dataset Description

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Other

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

Re-used

Full texts are going to be reusable in a way it will be possible to process their raw text format e.g. with the text data mining algorithms.

1.1.4 What is the type of the described dataset?

Other

OpenAIRE plaintexts are full texts (text files) extracted from the PDF files and shared in an unchanged form. Texts are associated with the OpenAIRE Graph publications.

1.1.5 What is its format?

JSON Data Interchange Format

1.1.6 What is its expected size?

Hundreds of gigabytes, tens of gigabytes when compressed.

1.1.7 Why are you collecting/generating or re-using it?

• To obtain information

• To combine with other data

• Other

To run the text data mining algorithms in order to infer additional data including references to other research objects.

1.1.8 What is its origin / provenance?

Datasources registered in the OpenAIRE aggregation system.

1.1.9 To whom might it be useful ('data utility')?

• Researchers

• Research communities

• Education

• The public

2.1 Publications

2.1.1 Does the described output support any scientific publication?

No

2.1.2 Is there a data availability statement provided along with the publication?

No

2.2 Datasets

2.2.1 Does the described output use or support any published dataset?

No

2.3 Software

2.3.1 Does the described output use or support any software?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

Other

DOI

OpenAIREId

3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

https://zenodo.org/record/7492151

3.1.1.3 What type(s) of metadata?

Other

OpenAIRE Research Graph Dump:

https://zenodo.org/record/7488618

3.1.1.4 Do the metadata use standardised vocabularies?

No

3.1.1.6 Are the metadata searchable?

Yes

3.1.1.7 How are searchable metadata provided?

• Registry/Catalogue

• Metadata repository

Metadata about the OpenAIRE Full Texts is searchable via Zenodo and the OpenAIRE via Explore portal and dashboards.

3.1.1.8 Are keywords provided in the metadata?

No

3.1.1.9 Are metadata harvestable?

Yes

Metadata are harvestable via Zenodo.

3.2.1 Repository

3.2.1.1 In which repository will the dataset / output be deposited?

Zenodo

https://zenodo.org/

Zenodo is a catch-all repository hosted in CERN.

3.2.1.2 Is the selected repository a trusted source?

Yes

• Follows repository standards

• Has an open access content policy

• Supports back up

• Provides Open Access content (free at the point of use)

• Assigns PIDs

• Follows metadata standards

• Supports mid- and long-term preservation

• Supports authentication and authorization of users

• Has data security mechanisms in place

• Other

3.2.1.5 Does the repository(ies) assign datasets / outputs with persistent identifiers?

Yes

3.2.1.6 Does the repository(ies) resolve the identifiers to a digital object?

Zenodo assigns and resolves Digital Object Identifiers.

3.2.1.7 Does the repository support versioning?

Yes

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

OpenAIRE Full Texts Dump

3.2.2.2 How is the dataset / output shared?

Open

3.2.2.5 Are there any methods or tools required to access the dataset / output?

No

3.2.2.8 Is the described dataset / output supported by a data access committee?

No

3.2.2.9 Please specify how the dataset / output will be accessed during and after the project ends

It will remain available in Zenodo.

3.2.2.10 Please specify how long after the project has ended the dataset / output will be made accessible for

At least 10 years.

3.2.3 Metadata

3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

Yes

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.2.3.3 Do metadata provide information about how to access the described dataset / output?

No

3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

Yes

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

No

3.3.2 If you created the vocabulary, where can it be found?

https://api.openaire.eu/vocabularies/

3.3.3 Have you applied a standard schema for your (meta)data?

No

3.3.4 Will you provide a mapping to more commonly used ontologies?

No

3.3.7 Does the described dataset / output provide qualified references with other outputs?

No

3.4 Increasing data and other outputs reuse

3.4.1 What internationally recognised licence will you use for your dataset / output?

License Not Specified

Each individual plaintext document may be licensed differently according to the original document license.

3.4.2 What reusability and / or reproducibility methods are followed?

Other

It is possible to identify the publication the text was extracted from by DOI or OpenAIRE id.

3.4.3 Will you provide the described dataset / output in the public domain?

Yes

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

Dataset persisted in Zenodo archive.

3.4.5 Is provenance well documented?

Yes

Data provenance is tracked and available in the OpenAIRE Research Graph records. Each full text record is linked to the graph record with the OpenAIRE identifier.

3.4.6 What documented procedures for quality assurance do you have in place?

• Use of tools for automatic checks

• Data conform to format specification

• Consistency verified with data models and standards

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

N/A

Euro

• Storage

• Archiving

• Re-use

Direct cost

4.1.2 How will this cost be covered?

Other

4.1.3 Identify the people who will be responsible and their role(s) in the management of the described output

Marek Horst (orcid:0000-0002-9038-9333)

text extraction, data packaging

5.1 Data Security

5.1.1 What security measures are followed?

Other

Kept on secure, managed storage for a limited time.

5.1.2 What conditions do the security measures meet?

• Data storage

• Data recovery

• Data sharing

5.1.3 How will you preserve the described dataset / output in the long term?

Persisted in Zenodo archive.

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on sharing the described dataset / output?

no

6.1.2 Does the described dataset / output contain sensitive information?

No

6.1.3 Does the described dataset / output contain personal data?

No

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Dataset Title: BIP! DB

Template: Horizon Europe

BIP! DB is an open dataset that contains a variety of impact measures calculated for a large collection of scientific publications from various disciplines.

Dataset Description

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Research Data

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

Re-used

This dataset is based on the integration of three major datasets, which provide citation data: OpenCitation's COCI dataset, Microsoft Academic Graph, and Crossref.

Based on these data, we perform citation network analysis to produce five useful impact measures, which capture distinctly different aspects of scientific impact.

Note that this dataset is already available in Zenodo (https://doi.org/10.5281/zenodo.4386934); in the context of the SciLake project, we plan to extend it, ensuring regular updates.

1.1.4 What is the type of the described dataset?

Other

BIP! DB is an open dataset that contains a variety of impact measures calculated for a large collection of scientific publications from various disciplines.

1.1.5 What is its format?

Comma Separated Values

1.1.6 What is its expected size?

More than 25GB when compressed

1.1.7 Why are you collecting/generating or re-using it?

• To share information

• To combine with other data

The impact measures that BIP! DB offers, can be invaluable, since they can effectively quantify the scientific impact of research works.

These impact measures are calculated at the DOI and OpenAIRE id level, thus they can be easily integrated, and combined with other datasets.

1.1.8 What is its origin / provenance?

BIP! DB gathers citation data and metadata from three data sources: OpenCitations' COCI dataset, Microsoft's Academic Graph (MAG), and Crossref.

The dataset production workflow collects, cleans, and integrates data from these sources to produce a citation graph based on the distinct DOI-to-DOI relationships found.

Since the publication year is required for some of the measures to be calculated, publications lacking this information were excluded from the final network.

1.1.9 To whom might it be useful ('data utility')?

• Researchers

• Research communities

• Decision makers

• The public

Identifying the most valuable publications for any given research topic has become extremely tedious and time consuming.

Therefore, quantifying the impact of scientific works could facilitate the related tasks, which make up the daily routine of researchers and other professionals of the broader scientific and

academic community.

Note that all impact indicators can be aggregated at the level of the researcher, organisation (RPO or RFO) or country to facilitate research monitoring.

2.1 Publications

2.1.1 Does the described output support any scientific publication?

Yes

BIP! DB: A Dataset of Impact Measures for Scientific Publications

2.1.2 Is there a data availability statement provided along with the publication?

Yes

All files published are freely available in Zenodo (https://doi.org/10.5281/zenodo.4386934) under the Creative Commons Attribution 4.0 International license.

2.3 Software

2.3.1 Does the described output use or support any software?

Yes

The described output uses BIP! Ranker (https://github.com/athenarc/Bip-Ranker) to calculate impact indicators. Additionally, BIP! DB is used as the underlying data of the BIP! Services (https://bip.imsi.athenarc.gr/).

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

Data identifiers

DOI

3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

For each calculation level (DOI / OpenAIRE-id) we provide five (5) compressed CSV files (one for each measure/score provided) where each line follows the format "identifier <tab> score <tab> class". The parameter setting of each measure is encoded in the corresponding filename.

3.1.1.3 What type(s) of metadata?

• Descriptive

• Reference

3.1.1.4 Do the metadata use standardised vocabularies?

No

3.1.1.6 Are the metadata searchable?

Yes

3.1.1.7 How are searchable metadata provided?

Metadata repository

Through Zenodo API and portal.

3.1.1.8 Are keywords provided in the metadata?

Yes

3.1.1.9 Are metadata harvestable?

Yes

Through Zenodo API.

3.2.1 Repository

3.2.1.1 In which repository will the dataset / output be deposited?

Zenodo

https://zenodo.org/

Zenodo is a general-purpose open repository developed under the European OpenAIRE program and operated by CERN.

3.2.1.2 Is the selected repository a trusted source?

Yes

3.2.1.5 Does the repository(ies) assign datasets / outputs with persistent identifiers?

Yes

3.2.1.6 Does the repository(ies) resolve the identifiers to a digital object?

Yes, it assigns depositions with a DOI.

3.2.1.7 Does the repository support versioning?

Yes

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

BIP! DB: A Dataset of Impact Measures for Scientific Publications

3.2.2.2 How is the dataset / output shared?

Open

3.2.2.5 Are there any methods or tools required to access the dataset / output?

No

3.2.2.8 Is the described dataset / output supported by a data access committee?

No

3.2.3 Metadata

3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

Yes

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

Yes

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

No

3.4 Increasing data and other outputs reuse

3.4.1 What internationally recognised licence will you use for your dataset / output?

Creative Commons Attribution 4.0

3.4.3 Will you provide the described dataset / output in the public domain?

Yes

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

Yes, the dump will be uploaded in Zenodo, so it will be accessible after the end of the project.

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

N/A

Euro

• Storage

• Archiving

4.1.2 How will this cost be covered?

Other

The fairness of additional resources produced in the context of this project have been already taken into account in the funding.

In addition, Open Science tools (e.g., Zenodo) are utilized whenever possible, so no additional cost will incur.

4.1.3 Identify the people who will be responsible and their role(s) in the management of the described output

a. Thanasis Vergoulis (orcid:0000-0003-0555-4128)

b. Serafeim Chatzopoulos (orcid:0000-0003-1714-5225)

Couldn't find it? Insert it manually

5.1 Data Security

5.1.3 How will you preserve the described dataset / output in the long term?

Since it is uploaded on Zenodo, it is responsible for the long term preservation of the dataset.

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on sharing the described dataset / output?

no

6.1.2 Does the described dataset / output contain sensitive information?

No

6.1.3 Does the described dataset / output contain personal data?

No

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Dataset Title: FOS (SciNoBo)

Template: Horizon Europe

Fields of Study Classification System (FoS) is a suite of tools automatically assigning publications to scientific fields of study. The service will be publicly available through OpenAIRE.

Dataset Description

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Research Data

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

Re-used

This dataset is based on the integration of Crossref, Microsoft Academic Graph (MAG) and OpenAIRE Graph. We utilize the names of the published venues of the publications as well as their citations, references and additional metadata (e.g. titles & abstracts) to assign to each publication a Field of Science label.

1.1.4 What is the type of the described dataset?

Derived or compiled

We collect scientific publications from Crossref, MAG and OpenAIRE Graph. We use the SciNoBo FoS classifier to assign FoS labels to each of the publications.

1.1.5 What is its format?

• JSON Data Interchange Format

• Comma Separated Values

1.1.6 What is its expected size?

More than 20GB

1.1.7 Why are you collecting/generating or re-using it?

• To share information

• To make informed decisions

Classifying scientific publications according to Field-of-Science taxonomies is of crucial importance, powering a wealth of relevant applications including Search Engines, Tools for Scientific Literature, Recommendation Systems, and Science Monitoring. Furthermore, it allows funders, publishers, scholars, companies, and other stakeholders to organize scientific literature more effectively, calculate impact indicators along Science Impact pathways and identify emerging topics that can also facilitate Science, Technology, and Innovation policy-making.

1.1.8 What is its origin / provenance?

We process scientific publications provided by OpenAIRE Graph, Crossref and MAG.

The pipeline of the algorithm utilizes the metadata of each publication to produce FoS labels for each one of them. The metadata required are the published venue, the citation venues, the reference venues, the title and the abstract.

Publications lacking the abovementioned metadata are excluded from the analysis.

1.1.9 To whom might it be useful ('data utility')?

• Researchers

• Research communities

• Decision makers

• Education

• Economy

Since the dataset contains scientific publications classified to scientific FoS fields a lot of interest parties can be benefited. For example, researchers and research communities can use the data to perform different analyses (train new classifiers, provide insights in the research communities of AI and Energy etc.). Furthermore, decision makers can analyze the classified data and identify emerging technologies and make informed decisions (the same goes for Economy and Industry). Finally, from an educational and public aspect, users can easily explore scientific publications related to these domains.

2.1 Publications

2.1.1 Does the described output support any scientific publication?

Yes

SciNoBo : A Hierarchical Multi-Label Classifier of Scientific Publications

World Wide Web

2.1.2 Is there a data availability statement provided along with the publication?

No

2.3 Software

2.3.1 Does the described output use or support any software?

Yes

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

Data identifiers

DOI

3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

```
 1    [
 2        {
 3            "pub_id": "the publication id",
 4            "fos_predictions": [
 5                {
 6                    "level_1": "level 1 fos",
 7                    "level_2": "level 2 fos",
 8                    "level_3": "level 3 fos",
 9                    "level_4": "level 4 fos",
10                    "level_5": "level 5 fos",
11                    "level_6": "level 6 fos"
12                }
13            ],
14            "fos_scores": {
15                "score_for_l3": 0.0,
16                "score_for_l4": 0.0,
17                "score_for_l5": 0.0
18            }
19        }
20    ]
```

Image 1

We provided an image containing one record of a JSON file, showcasing the schema used.

3.1.1.3 What type(s) of metadata?

Descriptive

3.1.1.4 Do the metadata use standardised vocabularies?

No

3.1.1.6 Are the metadata searchable?

Yes

3.1.1.7 How are searchable metadata provided?

The metadata (FoS predictions) are searchable and available through OpenAIRE (https://explore.openaire.eu/fields-of-science)

3.1.1.8 Are keywords provided in the metadata?

No

3.1.1.9 Are metadata harvestable?

No

3.2.1 Repository

3.2.1.1 In which repository will the dataset / output be deposited?

OpenAIRE Graph

3.2.1.2 Is the selected repository a trusted source?

Yes

3.2.1.7 Does the repository support versioning?

Yes

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

Scinobo: a hierarchical multi-label classifier of scientific publications / SCINOBO: a novel system classifying scholarly communication in a dynamically constructed hierarchical Field-of-Science taxonomy

3.2.2.2 How is the dataset / output shared?

Shared

3.2.2.5 Are there any methods or tools required to access the dataset / output?

No

3.2.2.8 Is the described dataset / output supported by a data access committee?

No

3.2.3 Metadata

3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

No

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

No

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

No

3.4 Increasing data and other outputs reuse

3.4.2 What reusability and / or reproducibility methods are followed?

Readme files

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

The FoS assignments will also be available through the OpenAIRE Explore portal even after the project finishes.

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

N/A

Euro

• Storage

• Archiving

• Other

Infrastructure

4.1.2 How will this cost be covered?

Other

4.1.3 Identify the people who will be responsible and their role(s) in the management of the described output

a. Sotiris Kotitsas (orcid:0000-0002-8114-6225)

b. Haris Papageorgiou (orcid:0000-0002-7352-2403)

Couldn't find it? Insert it manually

5.1 Data Security

5.1.3 How will you preserve the described dataset / output in the long term?

Since the predictions are uploaded and searchable through OpenAIRE, it is responsible for the long term preservation of the dataset. However, we will continue to provide FoS assignments for publications and updating the dataset.

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on sharing the described dataset / output?

no

6.1.2 Does the described dataset / output contain sensitive information?

No

6.1.3 Does the described dataset / output contain personal data?

No

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Dataset Title: VocTagger

Template: Horizon Europe

VocTagger is a python library that indexes textual corpora in accordance with given controlled vocabularies, in a flexible and scalable manner.

Dataset Description

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Research Data

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

Re-used

The VocTagger application can be used with various controlled vocabularies, both new and existing.
It has been used by SIRIS Academic in multiple domains and scientific disciplines.

The VocTagger dataset, currently available on Zenodo (https://zenodo.org/record/4118028#.ZHXrrxlBwkI), includes a set of controlled terms that define the scope and breadth of Sustainable Development Goals (SDGs) as defined by the United Nations. These terms may be used to tag and index textual records in accordance with SDGs.The original vocabularies will be improved and expanded.

1.1.4 What is the type of the described dataset?

Derived or compiled

In order to expand the coverage of the current dataset, additional sources (provided by the project pilots in the area of cancer, neuroscience, energy, and transportation) will be considered.

1.1.5 What is its format?

Comma Separated Values

1.1.6 What is its expected size?

Aprox. 1MB

1.1.7 Why are you collecting/generating or re-using it?

• To obtain information

• To make informed decisions

• To improve a product

Information retrieval from large textual corpora involves the identification of some key terms within the collection of text at hand. These terms are typically compiled in "controlled vocabularies".

To successfully carry out this task, a series of pattern matching rules must be defined to capture possible variants of the same concept, such as permutations of words within the concept and/or the presence of null words to be skipped. For this reason, we have carefully crafted matching rules that take into account permutations of words and that allow words within concept to be within a certain distance. Some relatively ambiguous keywords (which may match unwanted pieces of text), have a set of associated "extra" terms.

These "extra" terms are defined as further terms that must co-appear, in the same sentence, together with their associated ambiguous keywords.

1.1.8 What is its origin / provenance?

The vocabulary is constructed by means of the following steps:

1.   An initial set of terms per SDG target is built by extracting key terms from the UN official list of Goals, Targets and Indicators

2.   The list is manually enriched by performing a review of the literature produced around SDGs and by compiling lists of pertinent words per Target mentioned by the reviewed documents

3.   A reference textual corpus is downloaded by searching for the initial set terms defined at step 1. and 2. The corpus is used to train a Word2Vec[1] word embedding model (a machine learning model based on neural networks).

4.   The terms' list is then enriched by means of automatic methods, which are run in parallel:

The trained Word2Vec model is used to select, among the indexed keywords of the reference corpus, all terms "semantically close" to the initial set of words. This step is carried out to select terms that might not appear in the texts themselves, but that were deemed pertinent to label the textual records.

Further terms that are mentioned in the texts of the reference corpus and that are valued by the trained Word2Vec model as "semantically close" to the initial set of words are also retained. This step is performed to include in the controlled vocabulary a series of terms that are related to the focus of the SDGs and which are used by practitioners.

An automated algorithm is used to retrieve, from the APIs of WikiPedia a series of terms that have some categorical relationships (i.e. those that are indexed as "a broader concept of", or "equivalent to" in DBpedia) with the initial set of words.

5.   The final list produced by steps 1-4 is manually revised

[1] Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. Retrieved from http://arxiv.org/abs/1301.3781

1.1.9 To whom might it be useful ('data utility')?

• Researchers

• Research communities

- Decision makers

- Education

- Industry

2.1 Publications

2.1.1 Does the described output support any scientific publication?

Yes

Mapping STI Ecosystems via Open Data: Overcoming the Limitations of Conflicting Taxonomies. A Case Study for Climate Change Research in Denmark

SpringerOpen

2.1.2 Is there a data availability statement provided along with the publication?

Yes

Duran-Silva, N., Fuster, E., Massucci, F.A., Quinquillà, A.: A controlled vocabulary defining the semantic perimeter of sustainable development goals. Dataset, Zenodo (2019). https://doi.org/10.5281/zenodo.4118028

2.3 Software

2.3.1 Does the described output use or support any software?

Yes

https://github.com/sirisacademic/VocTagger

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

Data identifiers

DOI

10.5281/zenodo.4118028

3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

Metadata is available through Zenodo:

https://zenodo.org/record/4118028/export/json#.ZHXv8hlBwkI

3.1.1.3 What type(s) of metadata?

• Descriptive

• Reference

• Legal

3.1.1.4 Do the metadata use standardised vocabularies?

No

3.1.1.6 Are the metadata searchable?

Yes

3.1.1.7 How are searchable metadata provided?

Metadata repository

The metadata is available through Zenodo.

3.1.1.8 Are keywords provided in the metadata?

Yes

sdg, nlp

3.1.1.9 Are metadata harvestable?

Yes

Yes. Through Zenodo API.

3.2.1 Repository

3.2.1.1 In which repository will the dataset / output be deposited?

Zenodo

https://zenodo.org/

Zenodo is a general-purpose open repository developed under the European OpenAIRE program and operated by CERN.

3.2.1.2 Is the selected repository a trusted source?

Yes

• Follows repository standards

• Details terms of use

• Provides Open Access content (free at the point of use)

• Follows metadata standards

• Uses non-proprietary formats

3.2.1.5 Does the repository(ies) assign datasets / outputs with persistent identifiers?

Yes

3.2.1.6 Does the repository(ies) resolve the identifiers to a digital object?

Yes, it assigns DOIs.

3.2.1.7 Does the repository support versioning?

Yes

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

VocTagger

3.2.2.2 How is the dataset / output shared?

Open

3.2.2.5 Are there any methods or tools required to access the dataset / output?

Yes

Couldn't find it? Insert it manually

Download

3.2.2.6 Please provide information about the method(s) needed to access the dataset / output.

Login via Zenodo, GitHub or ORCID.

3.2.2.8 Is the described dataset / output supported by a data access committee?

No

3.2.3 Metadata

3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

Yes

The dataset will be openly shared.

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.2.3.3 Do metadata provide information about how to access the described dataset / output?

Yes

3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

Yes

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

Yes

Couldn't find it? Insert it manually

Generated SDG vocabulary plus vocabularies to be created in the context of the project based on the pilots proposed datasets.

3.3.2 If you created the vocabulary, where can it be found?

https://zenodo.org/record/4118028

3.3.3 Have you applied a standard schema for your (meta)data?

No

3.3.4 Will you provide a mapping to more commonly used ontologies?

Yes

3.3.7 Does the described dataset / output provide qualified references with other outputs?

No

3.4 Increasing data and other outputs reuse

3.4.1 What internationally recognised licence will you use for your dataset / output?

Creative Commons Attribution Share-Alike 4.0

3.4.2 What reusability and / or reproducibility methods are followed?

Readme files

3.4.3 Will you provide the described dataset / output in the public domain?

Yes

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

N/A

Euro

• Storage

• Archiving

Direct cost

4.1.2 How will this cost be covered?

Other

SciLake project (Horizon Europe framework programme under grant agreement No. 101058573)

4.1.3 Identify the people who will be responsible and their role(s) in the management of the described output

Couldn't find it? Insert it manually

a. César Parra Rojas (orcid:0000-0003-3625-9412)

b. Nicolau Duran-Silva (orcid:0000-0001-5170-4129)

c. Pablo Accuosto (orcid:0000-0002-3493-6013)

5.1 Data Security

5.1.3 How will you preserve the described dataset / output in the long term?

By means of Zenodo preserving mechanisms.

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on sharing the described dataset / output?

no

6.1.2 Does the described dataset / output contain sensitive information?

No

6.1.3 Does the described dataset / output contain personal data?

No

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Dataset Title: SciTagIT

Template: Horizon Europe

A suite of Domain-specific Classifiers automatically assigning publications to categories leveraging domain ontologies and classification schemes (e.g. International Classification of Diseases 11th Revision (ICD-11), the Glossary for Transport Statistics, United Nations Sustainable Development Goals (SDGs)). The toolkit will be made available under the CC-BY-NC-ND license.

Dataset Description

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Research Data

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

New

1.1.4 What is the type of the described dataset?

Derived or compiled

Three sets of annotations will be generated automatically by means of classification algorithms that assign terms from the 11th Revision of the International Classification of Diseases, the Glossary for Transport Statistics, and UN SDGs.

1.1.5 What is its format?

CSV Schema

1.1.6 What is its expected size?

more than 10GB

1.1.7 Why are you collecting/generating or re-using it?

• To obtain information

• To share information

- To make informed decisions

1.1.9 To whom might it be useful ('data utility')?

- Researchers

- Research communities

- Decision makers

- Education

- Economy

- The public

- Industry

- Other

2.1 Publications

2.1.1 Does the described output support any scientific publication?

No

2.1.2 Is there a data availability statement provided along with the publication?

No

2.3 Software

2.3.1 Does the described output use or support any software?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

Data identifiers

DOI

3.1.1.2 Will you provide metadata for the described dataset / output?

No

3.2.2 Data

3.2.2.2 How is the dataset / output shared?

Shared

3.2.3 Metadata

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.2.3.3 Do metadata provide information about how to access the described dataset / output?

No

3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

Yes

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

Yes

Couldn't find it? Insert it manually

International Classification of Diseases 11th Revision
Sustainable Development Goals vocabulary
The Glossary for Transport Statistics

3.3.3 Have you applied a standard schema for your (meta)data?

No

3.3.4 Will you provide a mapping to more commonly used ontologies?

No

3.3.5 What is the methodology followed?

Using the tree structure of the ICD11 classification we have constructed a mapping from disease terms to ICD11 categories at higher levels. These mappings have been extended to also include synonyms of disease terms originating from UMLS. The ICD11 classification is then performed by means of a text-processing pipeline that: i) parses the text, ii) identifies disease terms, and iii) computes statistics for each ICD11 category, iv) assigns the highest-ranked categories to the text.

For the task of SDG classification, an ensemble of deep learning models, topic models, and lexicographic rules are applied using Natural Language Processing to classify scientific literature into one or more SDG categories.

3.3.7 Does the described dataset / output provide qualified references with other outputs?

No

3.4 Increasing data and other outputs reuse

3.4.1 What internationally recognised licence will you use for your dataset / output?

Apache Software License 2.0

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

No

3.4.6 What documented procedures for quality assurance do you have in place?

Not available

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

N/A

Euro

• Storage

• Archiving

Direct cost

4.1.2 How will this cost be covered?

Infrastructure Grant

4.1.3 Identify the people who will be responsible and their role(s) in the management of the described output

Dimitris Pappas (orcid:0000-0001-5784-0658)

6.1 Ethical aspects

6.1.2 Does the described dataset / output contain sensitive information?

No

6.1.3 Does the described dataset / output contain personal data?

No

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Dataset Title: Cancer research data for Pilot 2 (CLL Knowledge Graph)

Template: Horizon Europe

In the context of this proposal, a combination of institutional and publicly available data will be used. Specifically, the services developed by the SciLake project will be interlinked with aggregated output that is continuously produced through analysis of omics data (such as RNA-seq, amplicon,- based sequencing and whole-genome sequencing), as these are generated by national efforts (including Precision Medicine Initiatives) and cancer research projects. In addition, publicly available domain-specific data from sources such as PubMed (literature, information retrieval) and ClinVar (aggregation of information about genomic variation and its relationship to human health) will also be taken into consideration.

Dataset Description

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Research Data

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

New

The overarching goal is to create a knowledge graph relating to Chronic Lymphocytic Leukemia (CLL). We will interlink various databases using established ontologies, schemas, and vocabularies that will enable semantic reasoning and inference of nascent information.

1.1.4 What is the type of the described dataset?

Derived or compiled

The Knowledge Graph will contain:

- **EnrichR-KG** is a knowledge graph database and a web-server application that combines selected gene set libraries from Enrichr for integrative enrichment analysis and visualization, containing plenty of categories of databases for transcription **(ARCHS4 TFs, ChEA3, FAN-TOM6,** and **TRRUST);** pathways **(KEGG, PFOCR, Reactome, the Kinase Library,** and **WikiPathways);** ontologies **(Gene Ontology, Human Phenotype Ontology, Jensen DISEASES,** and **MGI Mammalian Phenotypes);** diseases/drugs **(Project Achilles, LINCS L1000 perturbation signatures,** and **Drug Perturbation Proteome Atlas),** cell types **(CCLE, Descartes, Human Gene Atlas, Tabula Muris,** and **Tabula Sapiens);** and other **(Pfam).**

- **PubMed** comprises more than 35 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher websites.

- **ClinVar** is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence. ClinVar thus facilitates access to and communication about the relationships asserted between human variation and observed health status, and the history of that interpretation.

- The Single Nucleotide Polymorphism database (**dbSNP**) is a public-domain archive for a broad collection of simple genetic polymorphisms.

- **COSMIC**, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

- The Genome Aggregation Database (**gnomAD**) is a resource developed by an international coalition of investigators, with the goal of aggregating and harmonizing both exome and genome sequencing data from a wide variety of large-scale sequencing projects and making summary data available for the wider scientific community.

- The Genotype-Tissue Expression (**GTEx**) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 54 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. Remaining samples are available from the GTEx Biobank. The GTEx Portal provides open access to data including gene expression, QTLs, and histology images.

- **OMIM** is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. The full-text, referenced overviews in OMIM contain information on all known mendelian disorders and over 16,000 genes. OMIM focuses on the relationship between phenotype and genotype.

• The **BioGRID** Open Repository of CRISPR Screens (ORCS) is a publicly accessible database of CRISPR screens compiled through comprehensive curation of all genome-wide CRISPR screen data reported in the biomedical literature. ORCS is updated on a quarterly basis and is fully searchable by gene/protein, phenotype, cell line, authors, and other attributes.

• **Genomics of Drug Sensitivity in Cancer** is a database that comprises information about 1000 human cancer cell lines, screened with 100s of compounds.

1.1.5 What is its format?

RDF/XML

1.1.6 What is its expected size?

~500 GB

1.1.7 Why are you collecting/generating or re-using it?

• To obtain information

• To make informed decisions

• To develop a product

• To combine with other data

1.1.8 What is its origin / provenance?

Datasets containing bio-data and meta-data derived from various experiments and bioinformatic analysis.

1.1.9 To whom might it be useful ('data utility')?

• Researchers

• Research communities

• Decision makers

• Education

• Industry

2.1 Publications

2.1.1 Does the described output support any scientific publication?

No

2.3 Software

2.3.1 Does the described output use or support any software?

Yes

EnrichR-KG tool: https://maayanlab.cloud/enrichr-kg

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

• Data identifiers

• Researchers identifiers

• Organizations identifiers

• Projects identifiers

DOI

ORCID

Crossref

Cordis

3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

The potential paradigm for this section could be Bioschemas (**https://bioschemas.org/).**

3.1.1.3 What type(s) of metadata?

• Descriptive

• Administrative

• Structural

• Reference

• Statistical

• Legal

3.1.1.4 Do the metadata use standardised vocabularies?

Yes

Couldn't find it? Insert it manually

3.1.1.5 Please provide URL/Description of used vocabularies

UBERON: https://www.ebi.ac.uk/ols/ontologies/uberon

Cell Ontology: https://www.ebi.ac.uk/ols/ontologies/cl

Cellosaurus: https://www.cellosaurus.org/

PubChem: https://pubchem.ncbi.nlm.nih.gov/

Entrez Gene: https://www.ncbi.nlm.nih.gov/gene

MSigDB: https://www.gsea-msigdb.org/gsea/msigdb

3.1.1.6 Are the metadata searchable?

Yes

3.1.1.8 Are keywords provided in the metadata?

Yes

3.1.1.9 Are metadata harvestable?

Yes

3.2.2 Data

3.2.2.2 How is the dataset / output shared?

Open

3.2.2.5 Are there any methods or tools required to access the dataset / output?

Yes

3.2.2.7 Please provide information about the tools needed to access the dataset / output.

(System to perform SPARQL queries on the graph and leverage information)

3.2.2.9 Please specify how the dataset / output will be accessed during and after the project ends

A web-based application accessible to the scientific community.

3.2.3 Metadata

3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

Yes

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.2.3.3 Do metadata provide information about how to access the described dataset / output?

Yes

3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

Yes

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

Yes

Couldn't find it? Insert it manually

https://www.ebi.ac.uk/ols/ontologies/uberon, https://www.ebi.ac.uk/ols/ontologies/cl, https://pubchem.ncbi.nlm.nih.gov/, https://www.ncbi.nlm.nih.gov/gene/, https://www.gsea-msigdb.org/gsea/msigdb

3.4 Increasing data and other outputs reuse

3.4.2 What reusability and / or reproducibility methods are followed?

Readme files

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

TBC

Euro

• Storage

• Re-use

• Other

tbc

4.1.2 How will this cost be covered?

Other

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on sharing the described dataset / output?

unknown

6.1.2 Does the described dataset / output contain sensitive information?

Unknown

6.1.3 Does the described dataset / output contain personal data?

Unkown

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Dataset Title: Transportation Research Knowledge Graph | Pilot 3

Template: Horizon Europe

This domain-specific Science Knowledge Graph (SKG) builds upon the OpenAIRE ecosystem and EOSC services, utilizing the Open Research Gateway on Transport Research (https://beopen.openaire.eu) delivered under the EC H2020 BE OPEN project. It will offer services for integrated SKG smart browsing based on impact and reproducibility using AI by also serving several categories of stakeholders (i.e., researchers, developers, regulators). The data collection/generation involves datasets, experimental data, informational data, observational data, CCAM-related data, simulation, derived data, Research Papers, Academic Publications, and any other piece of work which will have implications for the transport research.

Dataset Description

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Other

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

Re-used

https://beopen.openaire.eu/

1.1.4 What is the type of the described dataset?

Other

- Observational data in transport research i.e., sensor readings, survey results, human observations, etc.

- Experimental data i.e., data from equipment such as cameras, sensors, etc.

- Simulation data i.e., data generated from transport models simulating operations, systems, etc.

- Derived data i.e., results from data analysis

- Reference data i.e., peer-reviewed and often published and curated

1.1.5 What is its format?

Acrobat PDF 1.0 - Portable Document Format

TBC

• Text files

• Numerical

• Software

• Other

i.e.

- Publications

- Research data

- Research software

- Other research products

1.1.6 What is its expected size?

TBC

1.1.7 Why are you collecting/generating or re-using it?

• To obtain information

• To share information

• To keep on record

• To combine with other data

Data Collections, Research Papers, Academic Publications, and any other piece of work which will have implications for the transport research.

1.1.8 What is its origin / provenance?

OpenAIRE Graph

1.1.9 To whom might it be useful ('data utility')?

• Researchers

• Research communities

• Education

Promoting territorial and cross-border cooperation and contributing in the optimization of open science in transport research.

2.1 Publications

2.1.1 Does the described output support any scientific publication?

No

2.1.2 Is there a data availability statement provided along with the publication?

No

2.2 Datasets

2.2.1 Does the described output use or support any published dataset?

Yes

OpenAIRE Research Graph

2.3 Software

2.3.1 Does the described output use or support any software?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

• Data identifiers

• Researchers identifiers

• Projects identifiers

DOI

ORCID

Cordis

3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

OAI-ORE (Open Archives Initiative Object Reuse and Exchange)

3.1.1.3 What type(s) of metadata?

• Descriptive

• Administrative

• Structural

• Reference

• Statistical

• Legal

3.1.1.4 Do the metadata use standardised vocabularies?

No

3.1.1.6 Are the metadata searchable?

Yes

3.1.1.7 How are searchable metadata provided?

Other

3.1.1.8 Are keywords provided in the metadata?

Yes

3.1.1.9 Are metadata harvestable?

Yes

3.2.1 Repository

3.2.1.1 In which repository will the dataset / output be deposited?

OpenAIRE Research Graph

Will be uploaded to Zenodo through the OpenAIRE Graph

3.2.1.2 Is the selected repository a trusted source?

Yes

3.2.2 Data

3.2.2.2 How is the dataset / output shared?

Open

3.2.2.9 Please specify how the dataset / output will be accessed during and after the project ends

A web-based application accessible to the scientific community.

3.2.2.10 Please specify how long after the project has ended the dataset / output will be made accessible for

At least 2 years

3.2.3 Metadata

3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

Yes

We will curate the main information of the output in the form of metadata.

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.2.3.3 Do metadata provide information about how to access the described dataset / output?

No

3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

Yes

Metadata may be transcribed to Zenodo or other open repository.

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

No

3.4 Increasing data and other outputs reuse

3.4.3 Will you provide the described dataset / output in the public domain?

Yes

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

Zenondo may be a good alternative.

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

TBC

Euro

• Storage

• Archiving

• Re-use

• Other

TBC

4.1.2 How will this cost be covered?

Other

5.1 Data Security

5.1.1 What security measures are followed?

Other

Data security is based on the OpenAIRE ecosystem and uses integrated security rules

5.1.2 What conditions do the security measures meet?

Other

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on sharing the described dataset / output?

no

6.1.2 Does the described dataset / output contain sensitive information?

No

6.1.3 Does the described dataset / output contain personal data?

Unkown

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Dataset Title: Energy Pilot 4 - EnerMaps database

Template: Horizon Europe

Besides usual scientific repositories, energy research uses an extensive array of governmental data at local, national, or European levels. Different databases are also catering to particular data needs. The data based on the results of the H2020 EnerMaps project is described in the following way: - Resolution: data with a spatial resolution ranging from local to worldwide, temporal data with annual values down to measurements per second. - Domain coverage: Databases containing domain-specific information to cross-cutting databases with multiple domains. Energy thread: data is directly or indirectly linked to energy, such as climate, economic or social data

Dataset Description

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Other

1.1.2 Is it physical or digital?

Physical

1.1.3 Are you generating or re-using it?

Re-used

We are re-using the Enermaps platform and respective content, which is is also part of the Open AIRE.

1.1.4 What is the type of the described dataset?

Reference or canonical

Most of the data are collected and published by third organisations. An example of data type: the census or environmental data

1.1.5 What is its format?

8-bit ASCII Text

correct answer: stored data format unknown

1.1.6 What is its expected size?

164 GB , 50 databases, heterogeneous sizes and formats

1.1.7 Why are you collecting/generating or re-using it?

• To share information

• To make informed decisions

The main goal is to allow smart access to energy research data, to trigger informed decision-making, as well as reuse of data and comparative research.

1.1.8 What is its origin / provenance?

Enermaps: https://enermaps.eu/

1.1.9 To whom might it be useful ('data utility')?

• Researchers

• Research communities

• Decision makers

• Economy

Stakeholders that are active in the sector of energy are mainly related to research and governance

2.1 Publications

2.1.1 Does the described output support any scientific publication?

Yes

Sustainable Cities and Society

2.1.2 Is there a data availability statement provided along with the publication?

No

2.2 Datasets

2.2.1 Does the described output use or support any published dataset?

No

2.3 Software

2.3.1 Does the described output use or support any software?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

Data identifiers

DOI

nothing to add

3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

nothing to add

3.1.1.3 What type(s) of metadata?

Other

Depending on the database metadata my include multiple of the listed characteristics

3.1.1.4 Do the metadata use standardised vocabularies?

No

3.1.1.6 Are the metadata searchable?

No

3.1.1.8 Are keywords provided in the metadata?

Yes

energy

nothing to add. Keywords of enermaps full list available here: https://enermaps.openaire.eu/subjects

3.1.1.9 Are metadata harvestable?

Yes

nothing to add

3.2.1 Repository

3.2.1.1 In which repository will the dataset / output be deposited?

EnerMaps

https://enermaps.openaire.eu/

nothing to add

3.2.1.2 Is the selected repository a trusted source?

Yes

Follows repository standards

3.2.1.4 Add appropriate arrangements made with the repository(ies) where the described dataset will be deposited

cannot be specified

3.2.1.5 Does the repository(ies) assign datasets / outputs with persistent identifiers?

Yes

3.2.1.6 Does the repository(ies) resolve the identifiers to a digital object?

Unknown

3.2.1.7 Does the repository support versioning?

Unknown

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

Energy dataset

3.2.2.2 How is the dataset / output shared?

Shared

The data are already shared through enermaps

3.2.2.3 What is the reason of limiting access to the dataset / output?

there are no reasons nowadays

3.2.2.5 Are there any methods or tools required to access the dataset / output?

No

3.2.2.8 Is the described dataset / output supported by a data access committee?

No

3.2.2.9 Please specify how the dataset / output will be accessed during and after the project ends

Unknown

Unknown

3.2.2.10 Please specify how long after the project has ended the dataset / output will be made accessible for

Unknown

3.2.3 Metadata

3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

Yes

nothing to add

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.2.3.3 Do metadata provide information about how to access the described dataset / output?

No

it is not established if metadata will provide this content yet

3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

Yes

nothing to add

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

No

3.3.2 If you created the vocabulary, where can it be found?

Unknown

3.3.3 Have you applied a standard schema for your (meta)data?

No

3.3.4 Will you provide a mapping to more commonly used ontologies?

No

nothing to add

3.3.5 What is the methodology followed?

Unknown

3.3.6 What community-endorsed interoperability best practices are followed?

Unknown

3.3.7 Does the described dataset / output provide qualified references with other outputs?

Yes

it follows FAIR principles

3.4 Increasing data and other outputs reuse

3.4.1 What internationally recognised licence will you use for your dataset / output?

Other (Open)

nothing to add

3.4.2 What reusability and / or reproducibility methods are followed?

Other

Unknown

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

nothing to add

3.4.5 Is provenance well documented?

Yes

nothing to add

3.4.6 What documented procedures for quality assurance do you have in place?

Not available

Unknown

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

0

Euro

Other

Direct cost

The content is already designed to be FAIR

4.1.2 How will this cost be covered?

Other

Cost is 0

4.1.3 Identify the people who will be responsible and their role(s) in the management of the described output

Jakob Rager (orcid:0000-0003-4856-8140)

Project manager

5.1 Data Security

5.1.1 What security measures are followed?

Other

Unknown

5.1.2 What conditions do the security measures meet?

Other

Unknown

5.1.3 How will you preserve the described dataset / output in the long term?

Unknown

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on sharing the described dataset / output?

no

nothing to add

6.1.2 Does the described dataset / output contain sensitive information?

No

6.1.3 Does the described dataset / output contain personal data?

No

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Dataset Title: EBRAINS Knowledge Graph

Template: Horizon Europe

The scientific knowledge graph EBRAINS KG is part of the data sharing services of the EBRAINS research infrastructure (http://www.ebrains.eu). Together with the openMINDS metadata framework, the EBRAINS KG is used by the EBRAINS curation team to publish neuroscientific datasets, models and software, provided by researchers worldwide. The KG is a powerful metadata management system, linking multimodal and multilevel datasets through metadata. It has an API and a GUI for search and navigation of metadata, and access to DOI citable data with data descriptors and licenses for data use.

Dataset Description

1.1 Brief description of the described research output

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

Re-used

The EBRAINS KG was first released in 2018. The EBRAINS KG, delivered by SciLake, will contain updates compatible with the relevant features the SciLake project can contribute with.

1.1.4 What is the type of the described dataset?

Other

The EBRAINS KG is a graph database, embedded in the EBRAINS research infrastructure. It contains scholarly data, i.e., metadata and links between research products (datasets, models, software, including workflows and metadata models), projects (collections of datasets), and contributors (persons).

1.1.5 What is its format?

JSON-LD

The KG is a graph database.

1.1.6 What is its expected size?

Data are stored in the CSCS supercomputing centre in Switzerland, currently containing >100 TB of data. Individual data sets vary considerably in size.

1.1.7 Why are you collecting/generating or re-using it?

• To obtain information

• To share information

• To keep on record

• To develop a product

• To combine with other data

• Other

1.1.8 What is its origin / provenance?

Neuroscience data are heterogeneous, collected at multiple spatial and temporal scales using a broad range of methods, resulting in a plethora of procedures and formats and accompanying challenges with reproducibility and replicability of findings. Data are shared through domain-specific repositories (e.g., EBRAINS, DANDI, SPARC, CONP, BRAIN/MINDS, INCF Knowledge Space, SciCrunch, brain-map.org, openNeuro, Human Connectome Project, CBRAIN) or general data sharing repositories (e.g., Zenodo, figshare, Data Dryad). Motivated by the general need for metadata harmonization and interoperability in the field, the Human Brain Project initiated the development of the EBRAINS data services, which includes a curation service targeting data producers, a scientific knowledge graph (EBRAINS KG) with an API and a GUI for search and navigation of metadata, and access to DOI citable data with data descriptors and licenses for data use. EBRAINS is the new digital research infrastructure for brain and brain-?inspired sciences, recently included in the ESFRI roadmap

1.1.9 To whom might it be useful ('data utility')?

• Researchers

• Research communities

• Decision makers

• Education

• The public

• Industry

Researchers can explore research outputs on specific topics. Research communities can have dedicated gateways, gathering their research outputs.

Decision makers can evaluate the outcomes of a funding stream. Data available in online viewers are very suitable for education, and there are many options for designing studies based on data available in the KG. The public and industry may also find neuroscience data of interest.

2.1 Publications

2.1.1 Does the described output support any scientific publication?

No

2.1.2 Is there a data availability statement provided along with the publication?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

• Data identifiers

• Other

URL

The KG can be found at https://search.kg.ebrains.eu/. Data shared via the EBRAINS RI are provided with a DOI and a Creative Commons license

3.1.1.2 Will you provide metadata for the described dataset / output?

No

No, but please see the documentation: https://docs.kg.ebrains.eu. Each data set shared via the EBRAINS RI are provided with OpenMinds structured metadata.

3.2.1 Repository

3.2.1.1 In which repository will the dataset / output be deposited?

EBRAINS

https://www.ebrains.eu/

3.2.1.2 Is the selected repository a trusted source?

No

3.2.1.5 Does the repository(ies) assign datasets / outputs with persistent identifiers?

Yes

3.2.1.6 Does the repository(ies) resolve the identifiers to a digital object?

Data sets are assigned Data Cite DOIs

3.2.1.7 Does the repository support versioning?

Yes

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

EBRAINS research infrastructure

3.2.2.2 How is the dataset / output shared?

Open

https://www.ebrains.eu/

3.2.2.5 Are there any methods or tools required to access the dataset / output?

No

3.2.2.8 Is the described dataset / output supported by a data access committee?

Yes

Data sharing requests are evaluated by an interim curation request evaluation group and an ethics compliance team, both of the EU Human Brain Project. Access requests for sensitive data are handled case by case.

3.2.2.9 Please specify how the dataset / output will be accessed during and after the project ends

Via the EBRAINS research infrastructure

3.2.3 Metadata

3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

Yes

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.2.3.3 Do metadata provide information about how to access the described dataset / output?

Yes

3.2.3.4 Will metadata remain available after the dataset / output is no longer available?

Yes

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

Yes

3.3.2 If you created the vocabulary, where can it be found?

https://github.com/HumanBrainProject/openMINDS

3.3.3 Have you applied a standard schema for your (meta)data?

Yes

Couldn't find it? Insert it manually

open Metadata Initiative for Neuroscience Data Structures (openMINDS)

3.4 Increasing data and other outputs reuse

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

3.4.5 Is provenance well documented?

Yes

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

0

Euro

Storage

4.1.2 How will this cost be covered?

Use of national infrastructure

6.1 Ethical aspects

6.1.2 Does the described dataset / output contain sensitive information?

Yes

6.1.3 Does the described dataset / output contain personal data?

Yes

6.1.4 What are the methods used for processing and accessing sensitive/personal information?

• Anonymising data where necessary

• Privacy constraints and applicable ethical norms

• Data accompanied by informed consent statements

• Privacy policies

• National laws

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Dataset Title: OpenAIRE Graph

Template: Horizon Europe

The OpenAIRE Graph is a service that populates and provides access to an SKG that includes metadata and links between scientific products (e.g. literature, datasets, software, "other research products"), organizations, funders, funding streams, projects, communities, and (provenance) data sources.

Dataset Description

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Research Data

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

Re-used

The OpenAIRE Graph available in the Scientific Lake will be a specific materialization compatible with the specification from the RDA Working Group for SKGs Interoperability.

1.1.4 What is the type of the described dataset?

Other

The OpenAIRE Graph includes scholarly data, i.e., metadata and links between research products (e.g., literature, datasets, software), organizations, funders, funding streams, projects, communities, and data sources.

1.1.5 What is its format?

JSON Data Interchange Format

1.1.6 What is its expected size?

More than 100GB of storage when compressed.

1.1.7 Why are you collecting/generating or re-using it?

• To share information

• To develop a product

We share information aggregated from different data sources, deduplicate, enrich them, and make them available in different formats, portals and APIs.

We provide services upon the data that can be used to develop a product.

1.1.8 What is its origin / provenance?

Metadata records are collected from different data sources, including Open Access institutional repositories, data archives, journals, etc.

They are further enriched with records from Crossref, Unpaywall, DataCite, ORCID, ROR, and information about projects provided by national and international funders.

Dedicated inference algorithms applied to metadata and full-texts of Open Access publications enrich the content with links between research results and projects, author affiliations, subject classification, links to entries from domain-specific databases.

Duplicated organizations, data sources and results are identified and merged together in a unified form.

1.1.9 To whom might it be useful ('data utility')?

• Researchers

• Research communities

• Decision makers

• The public

Researchers can explore research outputs on specific topics, since the OpenAIRE Graph collects data from the most common and large data sources.

Research communities can have dedicated gateways, gathering their research outputs.

Decision makers to evaluate the outcomes of a funding stream.

2.1 Publications

2.1.1 Does the described output support any scientific publication?

No

2.1.2 Is there a data availability statement provided along with the publication?

No

2.3 Software

2.3.1 Does the described output use or support any software?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

• Researchers identifiers

• Projects identifiers

DOI

ORCID

Cordis

The dataset will be uploaded on Zenodo, which assigns a DOI. All contributing authors will provide their ORCID, and eventually, the dataset will be linked with the related project. Furthermore, the metadata will have a description and be linked to the relevant resources describing the schema of the data in the dataset.

3.1.1.2 Will you provide metadata for the described dataset / output?

Yes

3.1.1.3 What type(s) of metadata?

• Descriptive

• Reference

3.1.1.4 Do the metadata use standardised vocabularies?

No

3.1.1.6 Are the metadata searchable?

Yes

3.1.1.7 How are searchable metadata provided?

Metadata repository

Through Zenodo API and portal.

3.1.1.8 Are keywords provided in the metadata?

Yes

3.1.1.9 Are metadata harvestable?

Yes

Through Zenodo APIs.

3.2.1 Repository

3.2.1.1 In which repository will the dataset / output be deposited?

Zenodo

https://zenodo.org/

Zenodo is a catch-all repository hosted in CERN.

3.2.1.2 Is the selected repository a trusted source?

Yes

3.2.1.5 Does the repository(ies) assign datasets / outputs with persistent identifiers?

Yes

3.2.1.6 Does the repository(ies) resolve the identifiers to a digital object?

Yes, it provides a DOI.

3.2.1.7 Does the repository support versioning?

Yes

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

The RDA Interoperable OpenAIRE Graph Dataset

3.2.2.2 How is the dataset / output shared?

Open

3.2.2.5 Are there any methods or tools required to access the dataset / output?

No

3.2.2.8 Is the described dataset / output supported by a data access committee?

No

3.2.3 Metadata

3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

Yes

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

No

3.3.3 Have you applied a standard schema for your (meta)data?

No

3.3.7 Does the described dataset / output provide qualified references with other outputs?

Yes

It provides reference to the schema of the data within the dataset.

3.4 Increasing data and other outputs reuse

3.4.1 What internationally recognised licence will you use for your dataset / output?

Creative Commons Attribution 4.0

3.4.3 Will you provide the described dataset / output in the public domain?

Yes

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

Yes

Yes, the dump will be uploaded in Zenodo, so it will be accessible after the end of the project.

3.4.5 Is provenance well documented?

Yes

Yes, every record has dedicated information for its provenance.

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

N/A

Euro

• Storage

• Archiving

• Re-use

Direct cost

4.1.2 How will this cost be covered?

Other

The SKG will be based on the OpenAIRE Graph, an open resource already FAIR.

Furthermore, the fairness of additional resources produced in this project's context has already been considered in the funding.

Last, Open Science tools (e.g., Zenodo) are utilized whenever possible so that no additional cost will incur.

4.1.3 Identify the people who will be responsible and their role(s) in the management of the described output

a. Claudio Atzori (orcid:0000-0001-9613-6639)

b. Miriam Baglioni (orcid:0000-0002-2273-9004)

Couldn't find it? Insert it manually

5.1 Data Security

5.1.3 How will you preserve the described dataset / output in the long term?

It is preserved via uploading in Zenodo.

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on sharing the described dataset / output?

no

6.1.2 Does the described dataset / output contain sensitive information?

No

6.1.3 Does the described dataset / output contain personal data?

Yes

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Dataset Title: Questionnaire Results

Template: Horizon Europe

This dataset will gather the answers to the Pilots Questionnaire, which aims to facilitate the collection of the requirements for the SciLake services from the pilot communities (in the context of T1.1 - Requirements & KPIs). The main objective is to ensure that the current and anticipated needs of the respective research communities are adequately documented and to deliver services to match them.

Dataset Description

1.1 Brief description of the described research output

1.1.1 What kind of research output are you describing?

Other

1.1.2 Is it physical or digital?

Digital

1.1.3 Are you generating or re-using it?

New

Indeed it is the collection of answers to the SciLake Pilots Questionnaire which is performed specifically for the project's needs.

1.1.4 What is the type of the described dataset?

Other

1.1.5 What is its format?

Google Document Link File

The answers will be collected in a Google Spreadsheet and stored on the SciLake GDrive.

1.1.6 What is its expected size?

< 5 MB

1.1.7 Why are you collecting/generating or re-using it?

• To obtain information

• To make informed decisions

• To develop a product

We developed the questionnaire to have a clear view of the communities' current and anticipated needs, so as to understand which services are relevant for each of them and how to personalize them in a way that will meet their needs.

1.1.8 What is its origin / provenance?

The dataset will be composed of the answers provided by each partner participating in the pilot cases to the aforementioned questionnaire. In the future, the set of answers can be extended by asking people outside of the consortium, but related to the communities involved in the pilots, to reply to the same set of questions.

1.1.9 To whom might it be useful ('data utility')?

Other

It will be used by SciLake management and interested partners in the context of the project to make informed decisions and drive the implementation process.

2.1 Publications

2.1.1 Does the described output support any scientific publication?

No

2.1.2 Is there a data availability statement provided along with the publication?

No

2.2 Datasets

2.2.1 Does the described output use or support any published dataset?

No

2.3 Software

2.3.1 Does the described output use or support any software?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 What type(s) of persistent identifier(s) are used for the described dataset / output?

None

We do not intend to make the answers publicly available. We will store them locally for requirement acquisition purposes and we will report our findings in the architecture document.

3.1.1.2 Will you provide metadata for the described dataset / output?

No

We will not deposit the dataset, so no metadata will be provided for it.

3.2.1 Repository

3.2.1.1 In which repository will the dataset / output be deposited?

We do not intend to make the answers publicly available. We will store them locally for requirement acquisition reasons and we will report our findings in the architecture document.

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

SciLake - Pilot Questionnaire (Responses)

3.2.2.2 How is the dataset / output shared?

Closed

We do not intend to make the answers publicly available. We will store them locally for requirement acquisition reasons and we will report our findings in the architecture document

3.2.2.3 What is the reason of limiting access to the dataset / output?

The questions have relevance in the context of the project only. Our intention is to use them to make informed decision and to drive the subsequent implementation process. We will report about the questionnaire output in the deliverable we have to provide as a milestone in the project.

3.2.3 Metadata

3.2.3.1 Will you provide metadata even if the described dataset / output can not be openly shared?

No

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

3.3 Making data and other outputs interoperable

3.3.1 Does your (meta)data use a controlled vocabulary?

No

3.4 Increasing data and other outputs reuse

3.4.3 Will you provide the described dataset / output in the public domain?

No

3.4.4 Do you intend to ensure (re)use by third parties after your project finishes?

No

4.1 Allocation of resources

4.1.1 What will be the cost of making the described output FAIR?

0

Euro

Other

No activity since the dataset will not be made public

4.1.2 How will this cost be covered?

Other

4.1.3 Identify the people who will be responsible and their role(s) in the management of the described output

Miriam Baglioni (orcid:0000-0002-2273-9004)

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on sharing the described dataset / output?

no

6.1.2 Does the described dataset / output contain sensitive information?

No

6.1.3 Does the described dataset / output contain personal data?

No

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Powered by