Pathos

Open Science Impact Pathways

Deliverable 3.5

Data Management Plan Update

Deliverable Number and Name	D3.5 Data Management Plan Update
Due Date	29/02/2024
Delivery Date	11/04/2024
Work Package	WP3
Type	DMP – Data Management Plan
Author	Petros Stavropoulos, Ioanna Grypari, Haris Papageorgiou, Erika Balsyte, Nicki Lisa Cole, Pedro Principe, Vincent Traag, Corinne Martin
Reviewers	Andrew Hoffman, Natalia Manola
Approved by	Ioanna Grypari
Dissemination Level	Public
Version	1.0
Number of Pages	69

The information in this document reflects only the author's views and the European Commission is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability.

This project has received funding from the European Union's Horizon Europe framework programme under grant agreement No. 101058728. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the European Research Executive Agency can be held responsible for them.

Revision History

VERSION	DATE	REASON	REVISED BY
0.5	15/02/2024	First Draft	All authors
0.7	21/03/2024	Peer Reviewed	Andrew Hoffman, Natalia Manola
0.8	28/03/2024	Second Draft	All authors
0.9	08/04/2024	Project Coordinator QA	Ioanna Grypari
1.0	11/04/2024	Final Draft	Petros Stavropoulos

Table 1: Document Revision History

Table of Contents

Disclaimer

This document contains description of the PathOS project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) rules so, prior to using its content please contact the consortium head for approval.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order to ensure that its content is accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the PathOS consortium and can in no way be taken as a reflection of the views of the European Union.

PathOS is a project funded by the European Union (Grant Agreement No 101058728).

Abbreviations

ANR	Agence National de la Research
API	Application Programming Interface
CC	Creative Commons[1]
CERIF	Common European Research Information Format
DMP	Data Management Plan
DOAJ	Directory of Open Access Journals
DOI	Digital Object Identifier
EOSC	European Open Science Cloud
FAIR	Findability, Accessibility, Interoperability, Reusability
FCT	Foundation for Science and Technology
FoS	Fields of Science
FWF	Austrian Science Fund
HE	Horizon Europe
HRZZ	Croatian Science Foundation
IPCR	International Patent Classification
M	Month
NIH	National Institutes of Health
NSF	National Science Foundation
NWO	Netherlands Organisation for Scientific Research
PID	Persistent Identifier
RCAAP	Repositório Científico de Acesso Aberto de Portugal
SDG	Sustainable Development Goal(s)
SNSF	Swiss National Science Foundation
UKRI	UK Research and Innovation
WT	Wellcome Trust

[1] https://creativecommons.org/

Executive Summary

This document contains the Data Management Plan (DMP) of PathOS on M20 of the project. It is a living document that will formally be updated on M34. It has been created using the ARGOS service (https://argos.openaire.eu) an online tool for creating, managing and sharing DMPs and linking them with the research artifacts they correspond to. In the DMP, we describe the input (re-used) and output (created) datasets that will be used in the technical work of PathOS, following the principles of Open (whenever possible) and FAIR as outlined in the Grant Agreement.[2]

[2] The numbering of the questions skips in order to allow the updating of the DMP in the future (i.e. filling out questions for which the information is currently missing) via the Argos service.

1. Data Management Plan

1.1. Input Datasets

1.1.1. OpenAIRE Research Graph Dump

“The OpenAIRE Graph (formerly known as the OpenAIRE Research Graph) is one of the largest open scholarly record collections worldwide, key in fostering Open Science and establishing its practices in the daily research activities. Conceived as a public and transparent good, populated out of data sources trusted by scientists, the Graph aims at bringing discovery, monitoring, and assessment of science back in the hands of the scientific community.

Imagine a vast collection of research products all linked together, contextualised and openly available. For the past years OpenAIRE has been working to gather this valuable record. It is a massive collection of metadata and links between scientific products such as articles, datasets, software, and other research products, entities like organisations, funders, funding streams, projects, communities, and data sources.

As of today, the OpenAIRE Graph aggregates around 450Mi metadata records with links collecting from 2K data sources trusted by scientists, including:

- Open Access journals registered in DOAJ
- Crossref
- Unpaywall
- ORCID
- Microsoft Academic Graph
- Datacite

and repositories registered in OpenDOAR, re3data.org, FAIRSharing.org, and the EOSC Service Catalogue. Among these, prominent repositories such as:

- UKPubMed
- ArXiv
- HAL
- Zenodo

- Figshare
- Dryad
- Repec

After cleaning, deduplication, enrichment and full-text mining processes, the graph is analysed to produce statistics for the OpenAIRE MONITOR, the Open Science Observatory, made discoverable via the OpenAIRE EXPLORE and programmatically accessible via OpenAIRE Public APIs. Last but not least, the Graph data are openly available and can be used by third-parties to create added value services. "[3]

In the context of PathOS, this dataset will be used in the following case studies:

- EASY (ULEI)
- RCAAP (UMINHO)
- EMERGING TOPICS (ATHENA RC)
- COVID-19 (ATHENA RC)

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

1.1.4 What is the type of the described dataset?

Observational

1.1.5 What is its format?

JSON-LD

1.1.6 What is its expected size?

[3] https://graph.openaire.eu/

237.3 GB packed

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

- To obtain information
- To combine with other data
- Other

The OpenAIRE Graph will be used to estimate open science impact indicators for case studies.

1.1.8 What is its origin / provenance?

OpenAIRE

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

- Researchers
- Research communities
- Decision makers
- The public
- Industry
- Other

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

DOI for dataset: 10.5281/zenodo.4201546

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

Yes

Please refer to https://zenodo.org/record/7492313#.Y_319HZBxPY for full details.

3.1.1.9 Are metadata harvestable?

Yes

Please refer to https://zenodo.org/record/7492313#.Y_319HZBxPY for full details.

3.2.1 Repository

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

https://zenodo.org

3.2.1.2 Is the selected repository a trusted source?

Yes

- Follows repository standards
- Details terms of use
- Has an open access content policy
- Supports retention
- Supports withdrawal
- 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

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

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?

OpenAIRE Research Graph 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

Dataset will remain available from OpenAIRE.

All data is openly available, no identity will be ascertained

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

Dataset will remain available from OpenAIRE.

3.2.3 Metadata

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

Yes

Metadata are already available from OpenAIRE.

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

Metadata are already available from OpenAIRE.

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

Yes

Metadata are already available from OpenAIRE.

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?

OpenAIRE documentation: https://graph.openaire.eu/docs/

Sustainable Development Goals (SDG): https://explore.openaire.eu/sdgs

Fields of Science (FoS): https://explore.openaire.eu/fields-of-science

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

Yes

DataCite Metadata Schema

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?

No

This is provided by OpenAIRE.

3.4.5 Is provenance well documented?

Yes

documentation: https://graph.openaire.eu/docs/

7.1 Other

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

No

1.1.2. OpenAIRE Research Graph: Dump of funded products

"This dataset contains the metadata records about research products (research literature, data, software, other types of research products) with funding information available in the OpenAIRE Research Graph produced on December 2022. [...] You can also search and browse this dataset (and more) in the the OpenAIRE EXPLORE portal and via the OpenAIRE API."[4]

In the context of PathOS, this dataset will be used in the following case studies:

- EMERGING TOPICS (ATHENA RC)
- COVID-19 (ATHENA RC)
- RCAAP (UMINHO)

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

1.1.4 WHAT IS THE TYPE OF THE DESCRIBED DATASET?

Observational

1.1.5 WHAT IS ITS FORMAT?

JSON-LD

1.1.6 WHAT IS ITS EXPECTED SIZE?

[4] https://graph.openaire.eu/

3.6 GB packed

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

- To obtain information
- To combine with other data
- Other

The OpenAIRE research graph will be used to estimate open science impact indicators for case studies.

1.1.8 What is its origin / provenance?

OpenAIRE

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

- Researchers
- Research communities
- Decision makers
- The public
- Industry
- Other

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

Please refer to https://zenodo.org/record/7492313#.Y_319HZBxPY for full details.

3.1.1.9 Are metadata harvestable?

Yes

Please refer to https://zenodo.org/record/7492313#.Y_319HZBxPY for full details.

3.2.1 Repository

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

https://zenodo.org

3.2.1.2 Is the selected repository a trusted source?

Yes

- Follows repository standards
- Details terms of use
- Has an open access content policy
- Supports retention
- Supports withdrawal
- 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

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

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?

OpenAIRE Research Graph: Dump of funded products

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

Dataset will remain available from OpenAIRE.

All data is openly available, no identity will be ascertained

3.2.2.10 PLEASE SPECIFY HOW LONG AFTER THE PROJECT HAS ENDED THE DATASET / OUTPUT WILL BE MADE ACCESSIBLE FOR

Dataset will remain available from OpenAIRE.

3.2.3 METADATA

3.2.3.1 WILL YOU PROVIDE METADATA EVEN IF THE DESCRIBED DATASET / OUTPUT CANNOT BE OPENLY SHARED?

Yes

Metadata are already available from OpenAIRE.

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

Metadata are already available from OpenAIRE.

3.2.3.4 WILL METADATA REMAIN AVAILABLE AFTER THE DATASET / OUTPUT IS NO LONGER AVAILABLE?

Yes

Metadata are already available from OpenAIRE.

3.3 MAKING DATA AND OTHER OUTPUTS INTEROPERABLE

3.3.3 HAVE YOU APPLIED A STANDARD SCHEMA FOR YOUR (META)DATA?

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?

No

This is provided by OpenAIRE.

3.4.5 Is provenance well documented?

Yes

7.1 Other

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

No

1.1.3. OpenAIRE Research Graph Dump: new collected projects

“The dataset includes metadata about projects grants collected by OpenAIRE since September 2022. This dump involves:

- 640 new SNSF projects
- 1088 new HE projects
- 38462 NWO projects. Some of them are old projects collected with different OpenAIRE identifier
- 2008 new ANR projects
- 284571 new NIH projects
- 152 FWF projects
- 1 HRZZ new project
- 47089 NSF new projects
- 1 WT new project
- 8443 UKRI new projects “[5]

In the context of PathOS, this dataset will be used in the following case studies:

- EMERGING TOPICS (ATHENA RC)

[5] https://graph.openaire.eu/

- COVID-19 (ATHENA RC)

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

1.1.4 What is the type of the described dataset?

Observational

1.1.5 What is its format?

JSON-LD

1.1.6 What is its expected size?

1.7 MB packed

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

- To obtain information

- To combine with other data

- Other

The OpenAIRE research graph will be used to estimate open science impact indicators for case studies.

1.1.8 What is its origin / provenance?

OpenAIRE

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

- Researchers

- Research communities

- Decision makers
- The public
- Industry
- Other

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?

Projects identifiers

DOI

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

Yes

Please refer to https://zenodo.org/record/6685030#.Y_35kXZBxPZ for full details.

3.1.1.9 Are metadata harvestable?

Yes

Please refer to https://zenodo.org/record/6685030#.Y_35kXZBxPZ for full details.

3.2.1 Repository

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

https://zenodo.org

3.2.1.2 Is the selected repository a trusted source?

Yes

- Follows repository standards
- Details terms of use
- Has an open access content policy
- Supports retention
- Supports withdrawal
- 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

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

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?

OpenAIRE Research Graph Dump: new collected projects

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

Dataset will remain available from OpenAIRE.

All data is openly available, no identity will be ascertained

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

Dataset will remain available from OpenAIRE.

3.2.3 Metadata

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

Yes

Metadata are already available from OpenAIRE.

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

Metadata are already available from OpenAIRE.

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

Yes

Metadata are already available from OpenAIRE.

3.3 Making data and other outputs interoperable

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

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

Output and data used for analysis will be deposited.

3.4.5 Is provenance well documented?

Yes

7.1 Other

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

No

1.1.4. AI EU Collection of Projects and Publications enriched with FoS labels powered by SciNoBo

Collection of metadata from project documents and publications that are classified into the "artificial intelligence (AI)" field of science (FoS) by the SciNoBo tool. For more information regarding the algorithm of SciNoBo, please refer to:

https://dl.acm.org/doi/abs/10.1145/3487553.3524677

In the context of PathOS, this dataset will be used in the following case studies:

- EMERGING TOPICS (ATHENA RC)

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

1.1.4 WHAT IS THE TYPE OF THE DESCRIBED DATASET?

Observational

1.1.5 WHAT IS ITS FORMAT?

JSON-LD

1.1.9 TO WHOM MIGHT IT BE USEFUL ('DATA UTILITY')?

• Researchers

- Research communities
- Decision makers
- The public
- Industry
- Other

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?

No

3.2.3 Metadata

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

1.1.5. Covid-19 EU Collection of Projects and Publications enriched with FoS labels powered by SciNoBo

Collection of metadata from project documents and publications that are classified into the "Covid-19" field of science (FoS) by the SciNoBo tool. For more information regarding the algorithm of SciNoBo, please refer to:

https://dl.acm.org/doi/abs/10.1145/3487553.3524677

In the context, of PathOS, this dataset will be used in the following case studies:

- COVID-19 (ATHENA RC)

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

1.1.4 What is the type of the described dataset?

Observational

1.1.5 What is its format?

JSON-LD

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

- Researchers
- Research communities
- Decision makers
- The public
- Industry
- Other

3.2.3 Metadata

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

1.1.6. Climate EU Collection of Projects and Publications enriched with FoS labels powered by SciNoBo

Collection of metadata from project documents and publications that are classified into the "climate change" field of science (FoS) by the SciNoBo tool. For more information regarding the algorithm of SciNoBo, please refer to:

https://dl.acm.org/doi/abs/10.1145/3487553.3524677

In the context, of PathOS, this dataset will be used in the following case studies:

- EMERGING TOPICS (ATHENA RC)

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

1.1.4 WHAT IS THE TYPE OF THE DESCRIBED DATASET?

Observational

1.1.5 WHAT IS ITS FORMAT?

JSON-LD

1.1.9 TO WHOM MIGHT IT BE USEFUL ('DATA UTILITY')?

- Researchers

- Research communities
- Decision makers
- The public
- Industry
- Other

3.2.3 Metadata

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

1.1.7. The Lens (lens.org)

The Lens, formerly called Patent Lens, is an online patent and scholarly literature search facility, provided by Cambia, an Australia-based non-profit organization. The Lens serves all the patents and scholarly work in the world as a free, open and secure digital public good.

In the context of PathOS, this dataset will be used in the following case studies:

- BIOINFORMATICS (ELIXIR)

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

The Lens will be used as an input dataset

1.1.4 What is the type of the described dataset?

Derived or compiled

The Lens is an agglomeration database, that takes bibliometric data from other databases (such as PubMed and Crossref) and combines them into one, deduplicated and with unified search syntax.

1.1.5 What is its format?

JSON Data Interchange Format

With over 20 years of development, supported by prominent philanthropic organizations, The Lens ingests, cleans, aggregates, normalizes and serves over 225+ million scholarly works, 127+ million global patent records, and more than 370+ million patent sequences, with rich metadata including the people and institutions that generate this knowledge and the linkages between them, drawn from diverse data sources. The Lens allows data exporting in JSON format.

1.1.6 What is its expected size?

225+ million scholarly works, 127+ million global patent records, and more than 370+ million patent sequences

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

To obtain information

The Lens will be used to text-mine names of ELIXIR bioinformatics resources (encompassing databases, software, tools, workflows, standards, ontologies, cloud computing, etc), as proxy for their usefulness in work leading to patent applications.

1.1.8 What is its origin / provenance?

The Lens is the flagship project of the social enterprise Cambia (https://cambia.org/).

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

• Researchers

• Research communities

• Decision makers

• Industry

Nature Biotechnology (https://www.nature.com/articles/nbt0506-474a) called the Patent Lens "a giant leap in the right direction" for providing researchers, technology transfer offices and company executives a facile means of establishing the novelty of their offerings and the nature of their competitors' inventions.

2.1 Publications

2.1.1 Does the described output support any scientific 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?

Other

URL

Guidance on acknowledging The Lens is at https://support.lens.org/knowledge-base/attribution-badge/

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

No

Not applicable as The Lens will be used an input dataset

3.2.1 Repository

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

Not applicable as The Lens will be used an input dataset

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)

• Supports mid- and long-term preservation

• Follows curation processes

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

Not applicable as The Lens will be used an input dataset

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?

The Lens

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

Not applicable as The Lens will be used an input dataset

3.2.3 Metadata

3.2.3.2 Under which license will metadata be provided?

Other

https://about.lens.org/policies/

3.3 Making data and other outputs interoperable

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

Yes

7.1 Other

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

No

1.1.8. FCT funding streams and projects information - SciPROJ

The FCT funding projects information - SciPROJ - is the national database that aggregates funding records (national and international) that support the science and technology activities developed in Portugal. Is based on the CERIF data model, SciPROJ provides information on 4 interconnected entities: funding, people, projects and organizations, in accordance with the adoption of the OpenAIRE data profile.

In the context of PathOS, this dataset will be used in the following case studies:

- ACAI (UMINHO)

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

Reuse the content of the SciPROJ (Portuguese database with lists of funded projects by the major Portuguese funder - FCT).

1.1.4 What is the type of the described dataset?

Other

Metadata records with funding information.

1.1.5 What is its format?

- CSV Schema
- JSON-LD

1.1.6 What is its expected size?

500 MB

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

- To obtain information
- To combine with other data

1.1.8 What is its origin / provenance?

Public API available in the context of the PTCRIS initiative.

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

- Researchers
- Research communities
- Decision makers

Research on the context of a case study analysing the compliance with the national mandate mandates the publications resulting from FCT funding to be deposited in a repository belonging to the RCAAP infrastructure.

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?

The described dataset has included in the data PIDs from Researchers identifiers, Organizations identifiers, Projects identifiers and others. The processed dataset will be deposited in DataRepositóriUM, so a DOI will be registered.

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

Yes

3.1.1.3 What type(s) of metadata?

Sciproj data have mainly administrative metadata, but the processed dataset will be deposited in DatRepositoriUM and via that mean have structured metadata following the standards from Dataverse software.

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

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?

DataRepositoriUM

https://datarepositorium.uminho.pt/

Institutional Data Repository to share, publish and manage research data generated and collected by the researchers' activity and in the research units of the University of Minho.

3.2.1.2 Is the selected repository a trusted source?

Yes

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

Institutional repository from the University of Minho managed by the unit involved in the project.

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

Yes, Digital Object Identifier.

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

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?

National database that aggregates funding records from FCT funder that support the science and technology activities developed in Portugal.

3.2.2.2 How is the dataset / output shared?

Shared

Public API.

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 cannot be openly shared?

Yes, based on Dataverse standards.

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?

No

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

Yes, we will apply descriptive metadata based on Dataverse standards, exposing the metadata compliant with Dublin core, DDI codebook, OpenAIRE, Datacite and Shecma.org.

3.3.5 What is the methodology followed?

CERIF and OpenAIRE

3.3.6 What community-endorsed interoperability best practices are followed?

CERIF and OpenAIRE

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

Yes

3.4 Increasing data and other outputs reuse

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

CC0 1.0

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

3.4.5 Is provenance well documented?

Yes

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

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?

Research data will be collected for reuse with minimal effort from public API collection, without many payments involved for storage and deposit in repository, as an Institutional facility is used.

Euro

Storage

Indirect cost

4.1.2 How will this cost be covered?

Use of institution infrastructure

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

a. Antonia Correia (orcid:0000-0002-6610-8853)

Curate and deposit.

b. Pedro Príncipe (orcid:0000-0002-8588-4196)

Curate and storage/preserve.

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

1.1.9. CNRS – OpenEditions connexion logs

Server connexion logs to https://www.openedition.org/ cleaned and enriched by Ezpaarse (A tool developped by CNRS support unit INIST-UAR71) and made available in the COUNTER format on OpenEditions servers hosted by Huma-num (CNRS infrastructure) In the context of PathOS, this dataset will be used in the following case studies:

- FRANCE (CNRS)

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

Those are connection logs to OpenEdition's website

1.1.4 What is the type of the described dataset?

Observational data

Those are connection logs to OpenEdition's website

1.1.5 What is its format?

comma separated values

1.1.6 What is its expected size?

10 gigabytes

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

• To obtain information

• To combine with other data

The connection logs to OpenEdition's website will be aggregated with the connection logs to Hal's website and the connection logs to RechercheDataGouv's website, to create overall statistic on usage of those three platforms

1.1.8 What is its origin / provenance?

Those are connection logs to OpenEdition's website. They are generated by the server, and post-processed by an automatic tool that de-noises the raw file and converts it into a CSV.

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

- Researchers
- Research communities
- Decision makers
- Education
- Economy
- The public
- Industry
- Other

This data will not be made public. It stays on OpenEdition's servers and, apart from OpenEdition, is only accessible to CIS working team during the project.

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?

We will use and develop different software to collect, visualize, and process the data. Elasticsearch and Kibana as tools; SSH, Linux, Python and Node.js as protocols, OS and programming frameworks.

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?

this data will not be made public

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

this data will not be made public

3.2.1 Repository

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

this data will not be made public

this data will not be made public

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

this data will not be made public

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

OpenEdition connection logs

3.2.2.2 How is the dataset / output shared?

Closed

this data will not be made public

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

The dataset contains personal information as per §4 of GDPR. it will be aggregated in order not to contain any directly or indirectly identifying information.

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

Yes

Couldn't find it? Insert it manually

SSH connexion

OpenEdition's Elasticsearch API

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

We connect in SSH to OpenEdition's virtual machine. Everything is done through linux terminal. We also use OpenEdition's Elasticsearch API.

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

https://en.wikipedia.org/wiki/Secure_Shell

The Secure Shell Protocol (SSH) is a cryptographic network protocol for operating network services securely over an unsecured network.[1] Its most notable applications are remote login and command-line execution.

https://en.wikipedia.org/wiki/Elasticsearch

Elasticsearch is a search engine based on the Lucene library. It provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents.

https://en.wikipedia.org/wiki/API

An application programming interface (API) is a way for two or more computer programs or components to communicate with each other. It is a type of software interface, offering a service to other pieces of software.

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

Yes

This data will not be made public. In order to obtain SSH access to the data, we had to ask OpenEdition's sys admin to create individual accounts to access the data.

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

This data will not be made public.

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

This data will not be made public.

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

3.3.3 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

3.4.2 What reusability and / or reproducibility methods are followed?

Codebooks

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

3.4.5 Is provenance well documented?

Yes

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

Data conforms to COUNTER5 standards https://cop5.projectcounter.org/en/5.0.2/03-specifications/index.html

4.1 Allocation of resources

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

This data will not be made public.

US Dollar

• Storage

• Archiving

• Re-use

• Security

• Other

this data will not be made public

This data will not be made public.

All costs related to data management are supported by OpenEdition and Huma-Num, french international digital infrastructure for human and social sciences https://www.huma-num.fr/quest-ce-que-l-ir-huma-num/

4.1.2 How will this cost be covered?

• Use of national infrastructure

• Use of institution infrastructure

All costs related to data management are supported by OpenEdition and Huma-Num, french international digital infrastructure for human and social sciences https://www.huma-num.fr/quest-ce-que-l-ir-huma-num/

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

Tommaso Venturini (orcid:0000-0003-0004-5308)

5.1 Data Security

5.1.1 What security measures are followed?

• Encryption

• Firewall

• Passwords

5.1.2 What conditions do the security measures meet?

• Data access

• Data storage

• Data transmission

• Data sharing

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

This data will not be made public.

As they are connexion logs, RGPD compliance rules forces OpenEdition's system administrators to delete them after a legal period of time.

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?

Yes

7.1.2 Documentation of other procedures

We also use CNRS data treatment declaration tool called Revcil. All our data treatments are validated and supervised by CNRS's data protection officer, which also supervises OpenEdition's overall data RGPD compliance, as OpenEdition is also a CNRS unit.

CNRS – HAL connexion logs

Server connexion logs to https://hal.science/ cleaned and enriched by Ezpaarse (A tool developped by CNRS support unit INIST-UAR71) and made available in the COUNTER format on OpenEditions servers hosted by Huma-num (CNRS infrastructure)

In the context of PathOS, this dataset will be used in the following case studies:

- FRANCE (CNRS)

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

Those are connection logs to OpenEdition's website

1.1.4 What is the type of the described dataset?

Observational

Those are connection logs to HAL's website

1.1.6 What is its expected size?

10 go

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

• To obtain information

• To combine with other data

The connection logs to Hal's website will be aggregated with the connection logs to OpenEdition's website and the connection logs to RechercheDataGouv's website, to create overall statistics on usage of those three platforms.

1.1.8 What is its origin / provenance?

Those are connection logs to Hal's website. They are generated by the server, and post-processed by an automatic tool that de-noises the raw file and converts it into a CSV.

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

- Researchers
- Research communities
- Decision makers
- Education
- Economy
- The public
- Industry
- Other

This data will not be made public. It stays on CNRS's servers and, apart from HAL, is only accessible to CIS working team during the project.

2.1 Publications

2.1.1 Does the described output support any scientific publication?

a. No

b. 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?

None

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

Yes

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?

No

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?

Data will not be made public

3.2.1.2 Is the selected repository a trusted source?

No

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

Data will not be made public

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

Hal connexion logs

3.2.2.2 How is the dataset / output shared?

Closed

Data will not be made public

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

The dataset contains personal information as per §4 of GDPR. it will be aggregated in the output dataset in order not to contain any directly or indirectly identifying information.

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

Yes SSH connection.

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

We connect in SSH to CNRS' virtual machine. Everything is done through linux terminal. We also use an Elasticsearch API.

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

□ https://en.wikipedia.org/wiki/Secure_Shell

The Secure Shell Protocol (SSH) is a cryptographic network protocol for operating network services securely over an unsecured network.[1] Its most notable applications are remote login and command-line execution.

□ https://en.wikipedia.org/wiki/Elasticsearch

Elasticsearch is a search engine based on the Lucene library. It provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents.

□ https://en.wikipedia.org/wiki/API

An application programming interface (API) is a way for two or more computer programs or components to communicate with each other. It is a type of software interface, offering a service to other pieces of software.

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

Yes

This data will not be made public. In order to obtain SSH access to the data, we had to ask HAL' s sys admin to give us access to the data.

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

This data will not be made public.

This data will not be made public.

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

This data will not be made public.

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

This data will not be made public

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

3.3.3 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

3.4.3 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

3.4.5 Is provenance well documented?

https://casrai.org/term/provenance-metadata/ link is dead

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

Data conforms to COUNTER5 standards

https://cop5.projectcounter.org/en/5.0.2/03-specifications/index.html

4.1 Allocation of resources

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

This data will not be made public

Euro

- Storage

- Archiving

- Re-use

- Security
- Other

This data will not be made public

All costs related to data management are supported by HAL and Huma-Num, french international digital infrastructure for human and social sciences https://www.huma-num.fr/quest-ce-que-l-ir-huma-num/

4.1.2 How will this cost be covered?

- Use of national infrastructure
- Use of institution infrastructure

All costs related to data management are supported by OpenEdition and Huma-Num, french international digital infrastructure for human and social sciences https://www.huma-num.fr/quest-ce-que-l-ir-huma-num/

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

Tommaso Venturini (orcid:0000-0003-0004-5308)

5.1 Data Security

5.1.1 What security measures are followed?

- Encryption
- Firewall
- Passwords

5.1.2 What conditions do the security measures meet?

- Data access
- Data storage
- Data transmission
- Data sharing

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

This data will not be made public.

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

As they are connexion logs, GDPR compliance rules forces HAL's system administrators to delete them after a legal period of time.

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?

Yes

7.1.2 Documentation of other procedures

We also use CNRS data treatment declaration tool called Revcil. All our data treatments are validated and supervised by CNRS's data protection officer, which also supervises HAL's overall data GDPR compliance, as HAL is also a CNRS unit.

CNRS – RechercheDataGouv connexion logs

Server connexion logs to https://www.openedition.org/ cleaned and enriched by Ezpaarse (A tool developed by CNRS support unit INIST-UAR71) and made available in the COUNTER format on OpenEditions servers hosted by Huma-num (CNRS infrastructure)

In the context of PathOS, this dataset will be used in the following case studies:

- FRANCE (CNRS)

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

Those are connection logs to RechercheDataGouv's website

https://recherche.data.gouv.fr/fr

1.1.4 What is the type of the described dataset?

Sample or specimen data

Those are connection logs to RechercheDataGouv's website

https://recherche.data.gouv.fr/fr

1.1.5 What is its format?

comma separated values

1.1.6 What is its expected size?

10 GB

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

- To obtain information

- To combine with other data

1.1.8 What is its origin / provenance?

Those are connection logs to RechercheDataGouv's website. They are generated by the server, and post-processed by an automatic tool that de-noises the raw file and converts it into a CSV.

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

- Researchers

- Research communities

- Decision makers

- Education

- Economy

- The public

- Industry

- Other

This data will not be made public. It is deposited on OpenEdition's servers and, apart from OpenEdition, is only accessible to CIS working team during the project.

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?

None

this data will not be made public

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

No

this data will not be made public

3.2.1 Repository

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

this data will not be made public

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

No

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

RechercheDataGouv's connection logs

3.2.2.2 How is the dataset / output shared?

Closed

this data will not be made public

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

The dataset contains personal information as per §4 of GDPR. it will be aggregated in an output dataset in order not to contain any directly or indirectly identifying information.

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

Yes

SSH connexion OpenEdition's Elasticsearch API

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

We connect in SSH to OpenEdition's virtual machine. Everything is done through linux terminal. We also use OpenEdition's Elasticsearch API.

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

https://en.wikipedia.org/wiki/Secure_Shell

The Secure Shell Protocol (SSH) is a cryptographic network protocol for operating network services securely over an unsecured network.[1] Its most notable applications are remote login and command-line execution. https://en.wikipedia.org/wiki/Elasticsearch

Elasticsearch is a search engine based on the Lucene library. It provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents.

https://en.wikipedia.org/wiki/API

An application programming interface (API) is a way for two or more computer programs or components to communicate with each other. It is a type of software interface, offering a service to other pieces of software.

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

Yes

This data will not be made public. In order to obtain SSH access to the data, we had to ask OpenEdition's sys admin to create individual accounts to access their virtual machines and RechercheDataGouv's system administrators to deposit the log files on this VM.

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

This data will not be made public.

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

This data will not be made public.

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

3.3.3 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?

This data will not be made public.

3.4.2 What reusability and / or reproducibility methods are followed?

Codebooks

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

3.4.5 Is provenance well documented?

Yes

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 data is conform to COUNTER5 standards https://cop5.projectcounter.org/en/5.0.2/03-specifications/index.html

4.1 Allocation of resources

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

This data will not be made public

Euro

- Storage
- Archiving
- Re-use
- Security
- Other

This data will not be made public

All costs related to data management are supported by OpenEdition and Huma-Num, french international digital infrastructure for human and social sciences https://www.huma-num.fr/quest-ce-que-l-ir-huma-num/

4.1.2 How will this cost be covered?

- Use of national infrastructure
- Use of institution infrastructure

All costs related to data management are supported by OpenEdition and Huma-Num, french international digital infrastructure for human and social sciences https://www.huma-num.fr/quest-ce-que-l-ir-huma-num/

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

Tommaso Venturini (orcid:0000-0003-0004-5308)

5.1 Data Security

5.1.1 What security measures are followed?

- Encryption
- Firewall
- Passwords

5.1.2 What conditions do the security measures meet?

- Data access
- Data storage
- Data transmission
- Data recovery
- Data sharing

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

This data will not be made public. As they are connexion logs, RGPD compliance rules forces RechercheDataGouv's system administrators to delete them after a legal period of time.

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?

Yes

7.1.2 Documentation of other procedures

We also use CNRS data treatment declaration tool called Revcil. All our data treatments are validated and supervised by CNRS's data protection officer, which also supervises RechercheDataGouv's overall data RGPD compliance, asOpenEdition is also a CNRS unit.

1.1.10. EP full-text data 2024

A bulk data collection including the full text in machine-readable format of all patent applications and granted patent specifications published by the EPO since it was set up in 1978. It's possible to search the complete EP full-text collection using our own search interface, in-house database or search engine.

In the context of PathOS, this dataset will be used in the following case studies:

- EMERGING TOPICS (ARC)
- COVID-19 (ARC)
- ACAI (UMIHNO)

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

1.1.4 What is the type of the described dataset?

Derived or compiled

Database

1.1.5 What is its format?

XML ST36 and PDF/A

1.1.6 What is its expected size?

4 GB (zipped)

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

To obtain information

1.1.8 What is its origin / provenance?

European Patent Office

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

• Researchers

• Research communities

• Education

• Industry

2.1 Publications

2.1.1 Does the described output support any scientific publication?

No

3.2.3 Metadata

3.2.3.2 Under which license will metadata be provided?

Creative Commons Zero (CC0)

1.2. Output Datasets

All Output Datasets

This section of the Data Management Plan (DMP) outlines the procedures and metadata standards we will employ to ensure the datasets created during the project adhere to the principles of FAIRness—making them Findable, Accessible, Interoperable, and Reusable. While each project partner retains their data on their own premises, adhering to their internal data management protocols, we have established a unified approach for handling the datasets generated within the scope of the project. The rules and guidelines detailed in this entry are designed to harmonize our efforts, ensuring that all project-generated datasets meet the current standards of data stewardship. This approach not only facilitates compliance with FAIR principles but also enhances the datasets' utility and longevity, enabling broader access and collaboration within the research community.

In the context of PathOS, this dataset will be created for the following case studies:

- BIOINFORMATICS (ELIXIR)
- EASY (ULEI)
- ACAI (UMINHO)

- FRANCE (CNRS)
- EMERGING TOPICS (ATHENA RC)
- COVID-19 (ATHENA RC)

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?

Various types of output datasets are expected.

3.1.1 Making data findable, including provisions for metadata

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

Yes

In PathOS, all produced datasets, whether open or closed, will be accompanied by detailed metadata to ensure they meet FAIR principles. Closed datasets will be catalogued in Zenodo as separate sources and our Data Management Plan (DMP), with metadata available under a CC-0 license. Open datasets will be deposited in Zenodo with a CC-BY license if they are small in size, and we are considering alternative sharing methods, like APIs, for larger datasets. We use the OpenAIRE Guidelines metadata schema, which is aligned with EOSC the EOSC interoperability framework, based on the Dublin Core and aligned with Zenodo as well (https://guidelines.openaire.eu/en/latest/data/index.html).

3.1.1.4 Do the metadata use standardised vocabularies?

Yes

3.1.1.5 Please provide URL/Description of used vocabularies

https://developers.zenodo.org/?python#representation.

3.1.1.6 Are the metadata searchable?

Yes

3.1.1.7 How are searchable metadata provided?

Registry/Catalogue

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?

Open datasets will be deposited in Zenodo

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

3.2.1.7 Does the repository support versioning?

Yes

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?

Yes

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

Yes

1.2.1. Collaborations extracted from RCAAP based on the citation and network analysis

Dataset with the results of the Citation and network analysis to derive collaborations in specific domains between academia and industry using the Portuguese network of repositories (RCAAP) content harvested in OpenAIRE graph and including the Open Citations data included in the graph.

In the context of PathOS, this dataset will be created for the following case studies:

- ACAI (UMINHO)

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

Generating a new dataset based on datasets consider in the PATHOS DMP.

1.1.4 WHAT IS THE TYPE OF THE DESCRIBED DATASET?

Derived or compiled

1.1.5 WHAT IS ITS FORMAT?

CSV Schema

3.2.3 METADATA

3.2.3.2 UNDER WHICH LICENSE WILL METADATA BE PROVIDED?

Creative Commons Zero (CC0)

1.2.2. Case study focus group data

This dataset is constituted by video recordings of online focus groups held with stakeholders to provide feedback on and insights for the development of the project's case studies, impact pathways, and indicators. This dataset will be used internally to facilitate this work and will not be shared openly to protect the privacy of participants.

In the context of PathOS, this dataset will be created for the following case studies:

- BIOINFORMATICS (ELIXIR)
- EASY (ULEI)
- ACAI (UMINHO)
- FRANCE (CNRS)
- EMERGING TOPICS (ATHENA RC)
- COVID-19 (ATHENA RC)

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?

Observational

These data are video recordings of online focus groups conducted throughout the PathOS project.

1.1.5 WHAT IS ITS FORMAT?

Digital Video

1.1.6 WHAT IS ITS EXPECTED SIZE?

Unknown at this time

1.1.7 WHY ARE YOU COLLECTING/GENERATING OR RE-USING IT?

• To obtain information

• To keep on record

- To make informed decisions

1.1.8 What is its origin / provenance?

The data will be generated through a series of focus group discussions, including a range of participants, across seven case studies and over the lifespan of the PathOS project.

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

Researchers

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?

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

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

No

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

PathOS case study focus group data

3.2.2.2 How is the dataset / output shared?

Closed

The dataset is closed to protect the privacy of those who participate in the focus groups.

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

Within the focus groups participants may critique aspects of products or processes that they are directly involved with through employment and therefore sharing the dataset openly would reduce their willingness to participate and/or limit what they are willing to share in the focus groups, thus hampering the aims of the research.

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

The data from each case will be available only to the case team in their secured cloud storage environment. Raw data will not be shared outside of each case team (e.g., not with the entire project consortium). Raw data will be preserved for up to three years after the conclusion of the project.

Case team members are required to enter private and confidential login details to access their cloud-based storage systems.

3.2.3 Metadata

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

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

3.4.4 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

Storage

4.1.2 How will this cost be covered?

Other

5.1 Data Security

5.1.1 What security measures are followed?

Passwords

Data are stored in password-protected, cloud-storage environments (e.g., Microsoft Teams).

5.1.2 What conditions do the security measures meet?

Data storage

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?

yes

Participants may share critiques of products or processes that they are involved with through their employment, and therefore could experience the risk of institutional repercussions if the data were openly shared.

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

Yes

Members of case study teams alone will be able to access raw data from their focus groups. Information created and shared based on these data, including notes, cross-case data synthesis, project meetings, documents and reports will not contain any personally identifiable information. Results of focus groups will be aggregated and fully anonymized when shared outside of the case team.

Personal information, including participants' first and last names, affiliations, positions, gender and email addresses are recorded in a project-internal spreadsheet that is secured in a password-protected Microsoft Teams environment that is only accessible to project consortium members.

7.1 Other

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

No

Mentions of ELIXIR resource names in patent applications (via lens.org)

Using The Lens (lens.org), a patent and scholarly literature search facility, we regularly text-mine names of ELIXIR resources in patent applications. We use this as proxy of their usefulness to bioinformaticians of all sectors (academic, industry), and across the globe. We are working to expand our searches to cover all ELIXIR resources (>400), as well as being able to visualise patent applications by main sectors.

In the context of PathOS, this dataset will be used in the following case studies:

- BIOINFORMATICS (ELIXIR)

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

Our approach is loosely inspired from earlier ELIXIR work (https://f1000research.com/articles/5-160/v1), but less computationally intensive, yet lightweight and repeatable so as to operationalise it as an indicator.

1.1.4 What is the type of the described dataset?

Derived or compiled

The dataset is a list of patent applications in which ELIXIR resources (https://elixir-europe.org/services) are mentioned by name. The Lens is text-mined once a week using a number of search terms relating to ELIXIR resource names. Patent applications containing one or more of these search terms are then added to a spreadsheet, with the following fields being extracted from The Lens:

- jurisdiction where the patent application was filed
- lens.org unique ID
- publication date
- application number and date
- title, abstract
- applicant names (institutions)
- URL in lens.org
- IPCR classifications

Manual curation of the returned patent applications is then carried out, so as to exclude false positives. The curated list feeds visualisations at https://elixir-europe.org/about-us/impact/patents.

1.1.5 What is its format?

OpenDocument Spreadsheet

The dataset is currently maintained as a Google spreadsheet - the format allows for collaborative manual curation to be carried out, as well as direct feeding of online visualisations. It is anticipated that this will remain the format for as long as the indicator will be maintained. A download option may be added to the visualisation URL once the exercise has become less qualitative.

1.1.6 What is its expected size?

20MB

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

- To obtain information
- To share information
- To make informed decisions

In the context of the PathOS project, the dataset will inform the work of the Focus Group convened as part of the bioinformatics case study, as well as the cost-benefit analysis. See https://pathos-project.eu/how-open-bioinformatics-resources-foster-innovation-in-industry-a-short-interview for a summary.

1.1.8 What is its origin / provenance?

The Lens (lens.org)

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

- Decision makers
- Other

This dataset is part of a broader body of work carried out by ELIXIR to demonstrate its public value to funders and decision-makers.

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

This feature will be implemented once the dataset is deemed suitable (i.e. sufficiently robust) for sharing.

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

Yes

Metadata will be produced once the dataset is deemed suitable (i.e. sufficiently robust) for sharing.

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?

No

3.2.1 Repository

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

https://elixir-europe.org/about-us/impact/patents, through a download functionality

3.2.2 Data

3.2.2.1 What is the described dataset / output title?

Mentions of ELIXIR resource names in patent applications (using lens.org)

3.2.2.2 How is the dataset / output shared?

Shared

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

The dataset will be openly accessible when deemed suitable (i.e., sufficiently robust).

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

https://elixir-europe.org/about-us/impact/patents, through a download functionality

3.2.3 METADATA

3.2.3.1 Will you provide metadata even if the described dataset / output cannot 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

https://elixir-europe.org/about-us/impact/patents, through a download functionality

3.3 Making data and other outputs interoperable

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

Yes

3.4 Increasing data and other outputs reuse

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

CC0 1.0

4.1 Allocation of resources

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

Corinne Martin (orcid:0000-0002-5428-2766)

Mentions of ELIXIR-supported publications in EuropePMC

Using EuropePMC (an ELIXIR Core Data Resource providing comprehensive access to life sciences literature from trusted sources), we use a range of search terms on funding linked to

ELIXIR, as well as on its achievements through its operation and projects, to identify ELIXIR-supported publications.

In the context of PathOS, this dataset will be used in the following case studies:

- BIOINFORMATICS (ELIXIR)

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

1.1.4 What is the type of the described dataset?

Derived or compiled

The dataset is a list of publications that have received financial support (e.g. grant funding, event sponsorship) or other support (e.g. in-kind contribution, use of the infrastructure and its many services) linked to ELIXIR (https://elixir-europe.org/about-us/publications/how-acknowledge). Manual curation of monthly searches is carried out, so as to exclude false positives. The curated list feeds visualisations at https://elixir-europe.org/about-us/impact/publications.

Note: the dataset excludes publications that make use of ELIXIR resources, as this is monitored through other processes.

ELIXIR partners collaborate to publish research articles (peer-reviewed and preprints) on the development and operation of bioinformatics resources encompassing databases, tools, cloud computing, standards and training. These publications highlight ELIXIR's scientific legacy as a research infrastructure, and their citations by others (in the open literature) demonstrate the extent of ELIXIR's contribution and appreciation by others.

1.1.5 What is its format?

The dataset is currently maintained as a Google spreadsheet - the format allows for collaborative manual curation to be carried out, as well as direct feeding of online visualisations. It is anticipated that this will remain the format for as long as the indicator will be maintained. A download option may be added to the visualisation URL once the exercise has become less qualitative.

1.1.6 What is its expected size?

20MB

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

- To obtain information
- To share information
- To make informed decisions

In the context of the PathOS project, the dataset will inform the work of the Focus Group convened as part of the bioinformatics case study, as well as the cost-benefit analysis. See https://pathos-project.eu/how-open-bioinformatics-resources-foster-innovation-in-industry-a-short-interview for a summary.

1.1.8 What is its origin / provenance?

EuropePMC

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

- Decision makers
- Other

This dataset is part of a broader body of work carried out by ELIXIR to demonstrate its public value to funders and decision-makers.

2.1 Publications

2.1.1 Does the described output support any scientific 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

URL

This feature will be implemented once the dataset is deemed suitable (i.e. sufficiently robust) for sharing.

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

Yes

3.1.1.3 What type(s) of metadata?

Metadata will be produced once the dataset is deemed suitable (i.e. sufficiently robust) for sharing.

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?

No

3.2.1 Repository

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

Not applicable as EuropePMC will be used an input dataset

3.2.1.2 Is the selected repository a trusted source?

Yes

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?

Mentions of ELIXIR-supported publications in EuropePMC

3.2.2.2 How is the dataset / output shared?

Shared

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

The dataset will be openly accessible when deemed suitable (i.e. sufficiently robust).

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

https://elixir-europe.org/about-us/impact/publications, through a download functionality

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?

https://elixir-europe.org/about-us/impact/publications, through a download functionality

3.3 Making data and other outputs interoperable

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

Yes

3.4 Increasing data and other outputs reuse

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

CC0 1.0

7.1 Other

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

No

Powered by

argos