Data Management Plan

JUNE 2021

Scholarly Communication Services for EOSC users

D1.1 – Data Management Plan

Version 1.0 – Final

PUBLIC

The deliverable consists of a set of active DMPs in the DMP tool Argos, for the installations of OpenAIRE Research Graph, OpenCitations, Zenodo.org, Argos, episciences.org, OpenAPC, and ScholExplorer. DMPs are on-going and will be kept up-to-date based on the evolution of the installations over the project lifetime.

H2020-INFRAEOSC-2020-2

Grant Agreement 101017452

Document Description

D1.1 – Data Management Plan	
WP1 - Coordination and management	
WP participating organizations: **OPENAIRE AMKE**, CNR, UNIBI, CERN, ARC, UNIWARSAW, UNIBO, CNRS, CITE	
Contractual Delivery Date: 06/2021	Actual Delivery Date: 07/2021
Nature: Report	Version: 1.0 (Final)
Public	

Preparation Slip

	Name	Organization	Date
From	Marek Horst	UNIWARSAW	14/06/2021
Edited by	Marek Horst	UNIWARSAW	02/07/2021
Reviewed by	Elli Papadopoulou	ATHENA	30/06/2021
Approved by	Paolo Manghi	OpenAIRE AMKE	05/07/2021
For delivery	Mike Chatzopoulos	OpenAIRE AMKE	19/07/2021

Revision History

Issue	Item	Reason for Change	Author	Organization
V0.1	Draft version	The first version of deliverable.	Marek Horst	UNIWARSAW
V0.2	First Delivery	Updating DMP location in Argos and Zenodo, supplementing deliverable document with DMP dump attachment.	Marek Horst	UNIWARSAW
V1.0	Final version	Minor updates related to dates and versions.	Marek Horst	UNIWARSAW

Contents

Disclaimer

This document contains description of the OpenAIRE-Advance 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 for its content to be 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 OpenAIRE-Advance consortium and can in no way be taken to reflect the views of the European Union.

OpenAIRE-Nexus is a project funded by the European Union (Grant Agreement No 101017452).

Acronyms

DMP	Data Management Plan
FAIR	Findable Accessible Interoperable Reusable

Publishable Summary

This document is a Data Management Plan (DMP) for OpenAIRE-Nexus, developed in line with the Guidelines on FAIR Data Management in Horizon 2020, as requested by the EC Data Pilot. The policies define data archiving, preservation, and sharing (i.e. access rights) practices to be adopted.

1. DATA MANAGEMENT PLAN

The Data Management Plan was prepared using the DMP tool Argos[1] 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 under a variety of mechanisms.

OpenAIRE-Nexus Data Management Plan, along with the dataset definitions for each installation[2], is publicly available at:

https://argos.openaire.eu/explore-plans/publicOverview/09f0ea90-26e4-432b-b13d-c082868e2548

Its static snapshot taken on 06.07.2021 was published in Zenodo:

https://zenodo.org/record/5075490

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] https://argos.openaire.eu/

[2] There is no dataset defined for Zenodo because it does not produce any research data in OpenAIRE-Nexus project scope.

1| APPENDIX

DATA MANAGEMENT PLAN INFORMATION

OpenAIRE-Nexus

OpenAIRE-Nexus brings in Europe, EOSC and the world a set of services to implement and accelerate Open Science. The Data Management Plan for OpenAIRE-Nexus was developed in line with the Guidelines on FAIR Data Management in Horizon 2020, as requested by the EC Data Pilot. It was prepared for each installation that requires such a plan: OpenAIRE Research Graph (CNR, ICM), OpenCitations (UNIBO), Argos (CITE), episciences.org (CNRS), OpenAPC (UNIBI), ScholExplorer (CNR, ICM). The policies define data archiving, preservation, and sharing (i.e. access rights) practices to be adopted.

Funder

European Commission||EC

Grant

OpenAIRE NEXUS

Organisations

University of Warsaw, Interdisciplinary Centre for Mathematical and Computational Modelling, Communication & Information Technologies Experts S.A., University of Bielefeld, University of Bologna, CERN, French National Centre for Scientific Research, Consiglio Nazionale delle Ricerche (CNR), Athena Research and Innovation Center In Information Communication & Knowledge Technologies

Researchers

Paolo Manghi (orcid:0000-0001-7291-3210), Antonis Lempesis (orcid:0000-0003-4483-1976), Argiro Kokogiannaki (orcid:0000-0002-3880-0244), Claudio Atzori (orcid:0000-0001-9613-6639), Marek Horst (orcid:0000-0002-9038-9333), Raphaël Tournoy (orcid:0000-0003-1244-0823), Sandro La Bruzzo (orcid:0000-0003-2855-1245), Andrea Mannocci (orcid:0000-0002-5193-7851), Pedro Príncipe (orcid:0000-0002-8588-4196), Michele De Bonis (orcid:0000-0003-2347-6012), Yannis Foufoulas (orcid:0000-0002-2785-946X), Andreas Czerniak (orcid:0000-0003-3883-4169), MICHELE ARTINI (orcid:0000-0002-4406-428X), Elli Papadopoulou, Jochen Schirrwagen (orcid:0000-0002-0458-1004), Christoph Broschinski (orcid:0000-0003-1972-7587), Harry Dimitropoulos (orcid:0000-0001-9791-587X), Enrico Ottonello (orcid:0000-0003-4606-7422), Natalia Manola (orcid:0000-0002-3477-3082), Alessia Bardi (orcid:0000-0002-1112-1292), Miriam Baglioni (orcid:0000-0002-2273-9004), Amelie Bäcker (orcid:0000-0001-6015-2063), Silvio Peroni (orcid:0000-0003-0530-4305)

DATASETS

Title: OpenAPC

Template: Horizon 2020

This dataset describes the OpenAPC research data.

Dataset Description

1.1 DATA SUMMARY

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

- ["To share information", "To make informed decisions", "To combine with other data"]
  - *Comment: Collecting and disseminating datasets on fees paid for open access publishing*

1.1.2 What types of data will the project generate/collect?

- ["sensor data", "text mining", "observational (e.g., sensor data, data from surveys)", "derived or compiled (e.g., text mining, 3D models)"]

1.1.3 What formats of data will the project generate/collect?

- [".txt files", "PDF", "RTF", "Text files - MS Word docs, .txt files, PDF, RTF, XML (Extensible Markup Language)"]
- CSV

1.1.4 What is the origin of the data?

- ["Secondary data"]

1.1.5 What is the expected size of the data?

- MB (megabyte)

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

- ["Researchers", "Research communities", "Decision makers", "Other"]

2.1 REUSED DATA

2.1.1 Will you re-use any existing data and how?

- Yes
- ["To reproduce and validate findings", "To compare and combine with other data"]

2.1.2 Where do the data reside?

- GitHub

2.1.3 Which data will be re-used?

- Tables in csv format
- CSVs send by contributed organizations and instiutions

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

- Yes
- OpenAPC
- Couldn't find it? Insert it manually

3.1.1.2 Please provide URL/Location describing the used metadata schema

- https://github.com/OpenAPC/openapc-de/wiki/Data-Submission-Handout
- GitHub

3.1.1.3 Will your metadata use standardised vocabularies?

- No

3.1.1.5 Will you make the metadata available free-of-charge?

- Yes
- *Comment: Open Data Commons license*

3.1.1.6 Will your metadata be harvestable?

- Yes
- *Comment: done at OpenAIRE to enrich the Research Graph*

3.1.1.7 Will you use naming conventions for your data?

- Yes

3.1.1.8 Please provide more details and examples on used naming conversions

- https://github.com/OpenAPC/openapc-de/wiki/schema#openapc-data-set

3.1.1.9 Will you provide clear version numbers for your data?

- Yes

3.1.1.10 Will you provide persistent identifiers for your data?

- Yes

3.1.1.11 Persistent identifiers

- URI

3.1.1.12 Will you provide searchable metadata for your data?

- No

3.1.1.15 Will you use standardised formats for your data?

- Yes
- Comma Separated Values

3.1.1.18 Are the file formats you will use open?

- Yes
- *Comment: csv*

3.1.1.20 Do supported open-source tools exist for accessing the data?

- Yes

3.1.1.21 Please describe if data require proprietary tools to access the data?

- NO

3.1.1.22 Will you provide metadata describing the quality of the data?

- No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

- No

3.1.2.2 Will your data be openly accessible?

- all

3.1.2.4 How will the data be made available?

- ["Other"]

3.1.2.7 Are there any methods or tools required to access the data?

- No

3.1.2.10 Will you also make auxiliary data that may be of interest to researchers available?

- after publication
- *Comment: Contribute instiutions and amount*

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

- Yes

3.1.4.1 When do you plan to make your data available for reuse?

- immediately

3.1.4.4 What internationally recognised licence will you use for your data?

- Open Data Commons Open Database License v1.0

3.1.4.5 Do you have documented procedures for quality assurance of your data?

- Yes

3.1.4.7 Describe the data quality assurance processes

- ["Use of tools for automatic checks", "Data conform to format specification"]

3.1.4.8 Will you provide any support for data reuse?

- Yes
- *Comment: openapc@uni-bielefeld.de*

4.1 Allocation of resources

4.1.1 How will the cost of making your data findable, accessible, interoperable and reusable be covered?

- ["Use of institution infrastructure"]

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

- Yes

4.1.3 Identify the people or roles that will be responsible for the management of the project data

- Christoph Broschinski (orcid:0000-0003-1972-7587)

4.1.4 How do you intend to ensure data reuse after your project finishes?

- ["Other"]

5.1 DATA SECURITY

5.1.1 What do you plan to do with research data of limited use?

- Delete at end of project

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

- No

7.1 OTHER

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

- No

Title: OpenAIRE Research Graph

Template: Horizon 2020

OpenAIRE Research Graph is an open resource that aggregates a collection of research data properties (metadata, links) available within the OpenAIRE Open Science infrastructure for funders, organizations, researchers, research communities and publishers to interlink information by using a semantic graph database approach.

Dataset Description

1.1 DATA SUMMARY

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

- ["To obtain information", "To share information", "To develop a product", "To improve a product", "To combine with other data"]
  - *Comment: OpenAIRE aggregates metadata about entities of the research life-cycle, combines them and enriches them to generate the OpenAIRE Research Graph. The graph is then made publicly available and used by many OpenAIRE products targeting differnet stakeholders, such as researchers, funders, content providers, organisations, research infrastructures and communities.*

1.1.2 What types of data will the project generate/collect?

- ["reference or canonical (e.g.", "static", "peer-reviewed data sets", "likely published or curated", "such as gene sequence databanks or chemical structures)", "Other"]

- Resource that aggregates a collection of research data properties (metadata, links) available within the OpenAIRE Open Science infrastructure which were harvested, resolved, harmonized and deduplicated

1.1.3 What formats of data will the project generate/collect?

- [".txt files", "PDF", "RTF", "Text files - MS Word docs, .txt files, PDF, RTF, XML (Extensible Markup Language)"]
- formats: JSON, XML, PDF; Data Models: Datacite, Crossref, OpenAiRE, MEDLINE/PubMed

1.1.4 What is the origin of the data?

- ["Secondary data"]

1.1.5 What is the expected size of the data?

- GB (gigabyte)
  - *Comment: over 120GB of data in tar format encapsulating files in JSON format*

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

- ["Researchers", "Research communities", "Decision makers", "Education", "The public"]

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

- Yes
- ["To compare and combine with other data", "To develop new products/ services"]

2.1.2 Where do the data reside?

- Re-used data are hosted in all sources aggregated by OpenAIRE

2.1.3 Which data will be re-used?

- Bibliographic metadata about research products and descriptive metadata about research entities available via scholarly communication sources and authoritative registries

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

- Yes
- OpenAIRE Metadata Format
- Couldn't find it? Insert it manually

3.1.1.2 Please provide URL/Location describing the used metadata schema

- https://zenodo.org/record/4723403
- Zenodo
- https://doi.org/10.5281/zenodo.3974225
- Zenodo

3.1.1.3 Will your metadata use standardised vocabularies?

- Yes
- MeSH (Medical Subject Headings)

- COAR vocabularies, Datacite vocabularies, other vocabularies as indicated in the OpenAIRE guidelines
- Couldn't find it? Insert it manually

3.1.1.4 Please provide URL/Description of used vocabularies

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

3.1.1.5 Will you make the metadata available free-of-charge?

- Yes

3.1.1.6 Will your metadata be harvestable?

- Yes
- *Comment: Metadata harvestable via Zenodo*

3.1.1.7 Will you use naming conventions for your data?

- Yes

3.1.1.8 Please provide more details and examples on used naming conversions

- https://zenodo.org/record/4723403

3.1.1.9 Will you provide clear version numbers for your data?

- Yes

3.1.1.10 Will you provide persistent identifiers for your data?

- Yes
- *Comment: DOI released by Zenodo*

3.1.1.11 Persistent identifiers

- DOI

3.1.1.12 Will you provide searchable metadata for your data?

- Yes
- *Comment: Metadata about the OpenAIRE Research Graph is searchable via Zenodo and OpenAIRE itself*

3.1.1.13 What services will you use to provide searchable metadata?

- Metadata repository
- OpenAIRE

3.1.1.15 Will you use standardised formats for your data?

- Yes
- JSON Data Interchange Format
- Couldn't find it? Insert it manually

3.1.1.16 Provide information about used standardised formats

- https://zenodo.org/record/4723403
- JSON

3.1.1.18 Are the file formats you will use open?

- Yes

- *Comment: JSON*

3.1.1.20 Do supported open-source tools exist for accessing the data?

- Yes
- OpenAIRE Explore Portal

3.1.1.21 Please describe if data require proprietary tools to access the data?

- NO

3.1.1.22 Will you provide metadata describing the quality of the data?

- Yes

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

- No

3.1.2.2 Will your data be openly accessible?

- all

3.1.2.4 How will the data be made available?

- ["Repository of Archive"]
- Zenodo

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

- secure with backup and recovery
- *Comment: All files uploaded to Zenodo are stored in CERN's EOS service where each file copy has two replicas located on different disk servers.*

3.1.2.7 Are there any methods or tools required to access the data?

- No

3.1.2.10 Will you also make auxiliary data that may be of interest to researchers available?

- no auxiliary data

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

- Yes
- *Comment: http://api.openaire.eu/vocabularies*

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

- immediately

3.1.4.4 What internationally recognised licence will you use for your data?

- Creative Commons Attribution 4.0 International

3.1.4.5 Do you have documented procedures for quality assurance of your data?

- Yes

3.1.4.6 Please provide URL with the documented procedures

- https://graph.openaire.eu/about#architecture

3.1.4.7 Describe the data quality assurance processes

- ["Use of tools for automatic checks", "Data conform to format specification", "Consistency verified with data models and standards"]

3.1.4.8 Will you provide any support for data reuse?

- Yes

3.1.4.9 How long do you intend to support data reuse?

- More than 10 years

4.1 ALLOCATION OF RESOURCES

4.1.1 How will the cost of making your data findable, accessible, interoperable and reusable be covered?

- ["Infrastructure Grant", "Collaboration with other Projects"]

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

- Yes

4.1.3 Identify the people or roles that will be responsible for the management of the project data

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

4.1.4 How do you intend to ensure data reuse after your project finishes?

- ["Other"]

5.1 DATA SECURITY

5.1.1 What do you plan to do with research data of limited use?

- Kept on secure, managed storage for limited time

6.1 ETHICAL ASPECTS

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

- No

6.1.2 What are the methods used for processing sensitive/personal data?

- ["Not available"]

7.1 OTHER

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

- No

Title: OpenCitations

Template: Horizon 2020

This dataset describes the bibliographic and citation data that are made available by OpenCitations. The data mainly concerns one main collection, i.e. the OpenCitations Indexes, which comprises several subcollections including COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

- ["To obtain information", "To share information", "To combine with other data"]
  - *Comment: The data gathered by OpenCitations from several sources will be reorganised according to the OpenCitations Data Model, combined with other existing data available, and finally shared with the community*

1.1.2 What types of data will the project generate/collect?

- ["reference or canonical (e.g.", "static", "peer-reviewed data sets", "likely published or curated", "such as gene sequence databanks or chemical structures)"]
- OpenCitations gathers and provides bibliographic metadata and citation data.

1.1.3 What formats of data will the project generate/collect?

- ["Text files - MS Word docs", ".txt files", "PDF", "RTF", "XML (Extensible Markup Language)"]
- The main formats used are CSV, RDF (N-quads) and JSON (Scholix).

1.1.4 What is the origin of the data?

- ["Secondary data"]

1.1.5 What is the expected size of the data?

- TB (terabyte)
  - *Comment: 5 to 10 terabytes of data are acpected to be produced and shared.*

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

- ["Researchers", "Research communities", "Decision makers", "Education", "The public"]
- The data made available for OpenCitations can be used fo research studies in bibliometrics, to support stakeholders community via university libraries, to assess scientific works, to build new applications and visualisations.

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

- Yes
- The data are provided by Crossref (https://crossref.org) and, in future iterations, by DataCite (https://datacite.org).
- ["To compare and combine with other data"]

2.1.2 Where do the data reside?

- https://api.crossref.org, https://api.datacite.org

2.1.3 Which data will be re-used?

- Bibliographic and citation data.

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

- Yes
- OpenCitations Data Model
- Couldn't find it? Insert it manually

3.1.1.2 Please provide URL/Location describing the used metadata schema

- https://doi.org/10.6084/m9.figshare.3443876.v7
- Figshare

3.1.1.3 Will your metadata use standardised vocabularies?

- Yes
- Couldn't find it? Insert it manually

3.1.1.4 Please provide URL/Description of used vocabularies

- http://www.sparontologies.net
- The Semantic Publishing and Referencing Ontologies, a.k.a. SPAR Ontologies, form a suite of orthogonal and complementary OWL 2 DL ontology modules for the creation of comprehensive machine-readable RDF metadata for every aspect of semantic publishing and referencing: document description, bibliographic resource identifiers, types of citations and related contexts, bibliographic references, document parts and status, agents' roles and contributions, bibliometric data and workflow processes.

3.1.1.5 Will you make the metadata available free-of-charge?

- Yes
- *Comment: All data and metadata are licensed in CC0.*

3.1.1.6 Will your metadata be harvestable?

- Yes
- *Comment: All the metadata are uploaded in the same repository used for the data.*

3.1.1.7 Will you use naming conventions for your data?

- Yes

3.1.1.8 Please provide more details and examples on used naming conversions

- https://doi.org/10.6084/m9.figshare.3443876.v7
- The naming convention adopted is defined in the OpenCitations Data Model, according to the following schema: '[base URL]/[dataset name]/][entity dataset identifier]'.

3.1.1.9 Will you provide clear version numbers for your data?

- Yes

- *Comment: Different versions of the data/metadata are defined according to the principles implemented in the repository used (which specifies different DOIs for each new version), while the versioning of each single entity included in the data is provided at metadata level (via provenance information).*

3.1.1.10 Will you provide persistent identifiers for your data?

- Yes
- *Comment: DOIs are used for datasets, while w3id.org Web PURL are used to refer to each single entity included in the data and metadata.*

3.1.1.11 Persistent identifiers

- DOI
- PURL

3.1.1.12 Will you provide searchable metadata for your data?

- Yes

3.1.1.13 What services will you use to provide searchable metadata?

- Linked Open Data
- Couldn't find it? Insert it manually

3.1.1.14 Please provide URL/Name for the used searchable metadata

- http://opencitations.net/index/sparql
- OpenCitations Indexes SPARQL endpoint.

3.1.1.15 Will you use standardised formats for your data?

- Yes
- Couldn't find it? Insert it manually

3.1.1.16 Provide information about used standardised formats

- http://opencitations.net/download#index
- CSV, RDF, Scholix

3.1.1.18 Are the file formats you will use open?

- Yes

3.1.1.20 Do supported open-source tools exist for accessing the data?

- Yes

3.1.1.21 Please describe if data require proprietary tools to access the data?

- NO

3.1.1.22 Will you provide metadata describing the quality of the data?

- No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

- No
- *Comment: Published data does not include personal or confidential data.*

3.1.2.2 Will your data be openly accessible?

- all

3.1.2.4 How will the data be made available?

- ["Project website", "Domain-specific database", "Repository of Archive"]
- Couldn't find it? Insert it manually

3.1.2.5 Please provide URL/Name of used data repositories

- https://figshare.com
- Figshare
- https://zenodo.org/
- Zenodo
- https://archive.org/
- Internet Archive

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

- secure with backup and recovery

3.1.2.7 Are there any methods or tools required to access the data?

- Yes

3.1.2.8 Please provide information about the method(s) needed to access the data

- Data can be accessed programmatically using SPARQL endpoints, REST APIs and visual interfaces.
- http://opencitations.net/querying

3.1.2.10 Will you also make auxiliary data that may be of interest to researchers available?

- no auxiliary data

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

- Yes
- *Comment: SPAR Ontologies (http://www.sparontologies.net).*

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

- immediately

3.1.4.4 What internationally recognised licence will you use for your data?

- Creative Commons Zero v1.0 Universal

3.1.4.5 Do you have documented procedures for quality assurance of your data?

- Yes
- *Comment: Currently, the actual procedures we used are described in research papers.*

3.1.4.6 Please provide URL with the documented procedures

- http://opencitations.net/publications

3.1.4.7 Describe the data quality assurance processes

- ["Use of tools for automatic checks", "Data conform to format specification"]

3.1.4.8 Will you provide any support for data reuse?

- Yes
- *Comment: Indirectly, by using open standards and formats for defining the data.*

3.1.4.9 How long do you intend to support data reuse?

- More than 10 years

4.1 Allocation of resources

4.1.1 How will the cost of making your data findable, accessible, interoperable and reusable be covered?

- ["Infrastructure Grant", "Collaboration with other Projects"]

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

- Yes
  - *Comment: Silvio Peroni, Director of OpenCitations, is currently in charge for the management of the data.*

4.1.3 Identify the people or roles that will be responsible for the management of the project data

- Silvio Peroni (orcid:0000-0003-0530-4305)
  - *Comment: Data manager*

4.1.4 How do you intend to ensure data reuse after your project finishes?

- ["Institutional archive", "Data Center Archive Storage"]

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

- Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

- No

6.1.2 What are the methods used for processing sensitive/personal data?

- ["Anonymising data where necessary"]

7.1 Other

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

- No

Title: Argos-data

Template: Horizon 2020

This dataset describes the Argos research data.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

- ["To make informed decisions", "To develop a product", "To combine with other data"]
  - *Comment: Argos generates Data Management Plans (DMPs) and connects with OpenAIRE services to strengthen their FAIRness and exploitation in the Open Science and Scholarly Communication ecosystem, where OpenAIRE NEXUS significantly contributes. Furthermore, Argos exposes DMPs to Zenodo, thus closing the DMPs publication lifecycle and allowing for outputs to be described according to best practices (DOIs, licenses, etc). The Research Graph is enriched with machine actionable DMPs from Argos and data are combined with other types of inferred re-sources to show relationships/semantics with each other. Information exchange of Argos defined metadata sets between PROVIDE and MONITOR facilitates better organization of researchers' data deposits as well as measurements of their uptake and evolution through time in support of informed policy and decision making.*

1.1.2 What types of data will the project generate/collect?

- ["Other"]
- Argos generates and publishes DMP documents, , which consists of dataset descriptions. Additionally, the project tracks researchers, projects, and organisations.

1.1.3 What formats of data will the project generate/collect?

- [".txt files", "PDF", "RTF", "Text files - MS Word docs, .txt files, PDF, RTF, XML (Extensible Markup Language)", "Other"]
- Argos generates DMP outputs and offers them both as plain text documents (.txt, .pdf) and as machine readable/actionable files (.xml, .json).

1.1.4 What is the origin of the data?

- ["Primary data"]

1.1.5 What is the expected size of the data?

- GB (gigabyte)
  - *Comment: Given the nature of the DMPs content, the step of the project lifecycle when DMPs are created and the lightweight format that DMPs are exposed to, a single DMP output in Argos is not expected to go beyond some MBs. The expected size of the total number of DMPs generated since the beginning of the OpenAIRE NEXUS project, in accordance to the KPIs set, is estimated at the range 5-10GBs. In case the project evolves to support binary elements accompanying dataset descriptions, the size may increase substantially by 1 or even 2 orders of magnitude.*

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

- ["Researchers", "Research communities", "Decision makers", "Economy", "Industry"]
- Argos DMPs data are useful (either directly or indirectly) to a number of stakeholders: Researchers can consult publicly available Argos DMPs and their associated dataset descriptions to inspire their own DMPs or use Argos data for their analytical studies Research communities can identify data trends and data profiles in their domains as well as strengths and weaknesses in specific Research Data Management (RDM) areas from Argos data exploitation. Policy and decision makers can exploit Argos data in support of their RDM strategies and plans. Industry can use Argos and enrich semantics in the Research Graph as well as exploit data and drive changes in RDM across sectors. Finally, Argos builds on data economy principles and consequently works towards accommodating them.

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

- No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

- Yes
- CERIF (Common European Research Information Format)
- DMPs are described according to three models: Datacite/ DC for published DMP records on Zenodo and CERIF in the context of the Research Graph. Internal model of Argos, utilized inside the application and in native xml exports of the service.

3.1.1.3 Will your metadata use standardised vocabularies?

- Yes
- Yes, metadata integrates with controlled vocabularies to achieve maximum information organization and retrieval. Vocabularies vary from ontologies, such as RDF, to ISO standards as used by the aforementioned services that handle Argos data sharing and processing. More specific, for Zenodo https://help.zenodo.org/guides/search/ and for CERIF https://github.com/EuroCRIS/CERIF-Vocabularies.

3.1.1.5 Will you make the metadata available free-of-charge?

- Yes
- *Comment: Metadata records are already available for free via Zenodo.*

3.1.1.6 Will your metadata be harvestable?

- Yes
- *Comment: Metadata comply with the Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH). Harvesting of metadata is supported only for DMPs that are published to Zenodo. Harvesting is supported from Zenodo.*

3.1.1.7 Will you use naming conventions for your data?

- Yes

3.1.1.8 Please provide more details and examples on used naming conversions

- Naming conversions target the exported file of the DMPs which are automatically created from user requests. They currently follow the format: DMP-nameofproject-version.

3.1.1.9 Will you provide clear version numbers for your data?

- Yes
- *Comment: Versioning is provided for non published and published DMPs to keep track of all changes, despite publication status, and treat DMPs as living documents while securing progress.*

3.1.1.10 Will you provide persistent identifiers for your data?

- Yes
- *Comment: Argos published DMP outputs are assigned Digital Object Identifiers (DOI) via Zenodo integrations. Private ones, do not acquire a PID.*

3.1.1.11 Persistent identifiers

- DOI

3.1.1.12 Will you provide searchable metadata for your data?

- Yes
- *Comment: Published DMP metadata are searchable via OpenAIRE catalogue and Zenodo which support web-based search. Complementary, Argos tags increases searchability within and outside of the platform. Furthermore all DMPs are searchable inside the Argos system utilizing its internal data model for search.*

3.1.1.13 What services will you use to provide searchable metadata?

- Metadata repository
- OpenAIRE
- Couldn't find it? Insert it manually

3.1.1.14 Please provide URL/Name for the used searchable metadata

- https://zenodo.org/
- Argos catalogue

3.1.1.15 Will you use standardised formats for your data?

- Yes
- JSON-LD

3.1.1.18 Are the file formats you will use open?

- Yes
- *Comment: Most of the formats that Argos data are exposed are standard: .xml and .json (maDMP), while the rest are well established and widely used: .txt and .pdf. OpenXML is also supported.*

3.1.1.20 Do supported open-source tools exist for accessing the data?

- Yes
- *Comment: All data exported can be opened by numerous opensource editors and Argos which is also FOSS.*

3.1.1.21 Please describe if data require proprietary tools to access the data?

- NO

3.1.1.22 Will you provide metadata describing the quality of the data?

- Yes
- *Comment: Metadata cover quality assurance issues as expressed in the RDA DMP Commons Standard: https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard#dataset_quality_assurance.*

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

- Yes
- *Comment: That depends on the types and nature of datasets described in Argos. Datasets with sensitive information can be isolated from the rest before DMP publication. Hence, exposed metadata won't record confidential or restricted information. Furthermore, users may opt out of publishing and sharing their DMPs that may expose confidential data.*

3.1.2.2 Will your data be openly accessible?

- some

3.1.2.4 How will the data be made available?

- ["Repository of Archive"]
- Zenodo holds an Argos community: https://zenodo.org/communities/argos/?page=1&size=20. Argos data are also expected to be made available via OpenAIRE API.

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

- secure with backup and recovery
- *Comment: For published DMPs: https://about.zenodo.org/infrastructure/. Regular backups are scheduled for data, under the policies of the data center that hosts the VMs. Furthermore data are accessible only by authorized system administrators for maintenance and troubleshooting.*

3.1.2.7 Are there any methods or tools required to access the data?

- No

3.1.2.10 Will you also make auxiliary data that may be of interest to researchers available?

- no auxiliary data

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

- Yes
- *Comment: The RDA DMP Common Standard is supported.*

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

- immediately

3.1.4.4 What internationally recognised licence will you use for your data?

- Creative Commons Attribution 4.0 International

3.1.4.5 Do you have documented procedures for quality assurance of your data?

- Yes

3.1.4.7 Describe the data quality assurance processes

- ["Data conform to format specification", "Consistency verified with data models and standards"]
- *Comment: Argos data conform to maDMP specification as exportable. Metadata conform to specifications mentioned for export to OpenAIRE graph and Zenodo.*

3.1.4.8 Will you provide any support for data reuse?

- Yes
- *Comment: Support is provided for Argos DMPs and data access and reuse.*

3.1.4.9 How long do you intend to support data reuse?

- Up to 5 years

4.1 Allocation of resources

4.1.1 How will the cost of making your data findable, accessible, interoperable and reusable be covered?

- ["Use of institution infrastructure", "Collaboration with other Projects"]

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

- Yes
  - *Comment: The Argos data manager, the Argos product manager and the OpenAIRE NEXUS communication officers are involved in Argos data management on different areas, eg from activities that aim to provide statistics that measure usage and impact to activities that relate to data ingestions and service integrations.*

4.1.3 Identify the people or roles that will be responsible for the management of the project data

- Diamantis Tziotzios (orcid:0000-0003-1670-4611)
  - *Comment: Data Manager*

4.1.4 How do you intend to ensure data reuse after your project finishes?

- ["Institutional archive", "Other"]

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

- Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

- No
  - *Comment: Researchers can select which data descriptions they would like to deposit in Zenodo. That way they comply with relevant regulations.*

6.1.2 What are the methods used for processing sensitive/personal data?

- ["Other"]
  - *Comment: The OpenAIRE methods are utilised during data contextualisation.*

7.1 OTHER

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

- No

Title: ScholExplorer Data Dump

Template: Horizon 2020

This dataset contains the GZ-compressed dump of the Scholix links (schema Version 3) exposed by the OpenAIRE ScholeXplorer service. The dataset consists of 445+Mi bi-directional links (i.e. 890+Mi directed links) between literature-dataset and dataset-dataset involving 17+ Mi literature objects and 50+ Mi datasets. Links were collected from publishers (CrossRef, EventData), data centers (DataCite and data centers), institutional/thematic repositories (OpenAIRE), and life-science databases (EMBL-EBI). The links are organized in 12 compressed files, each of at most 25.1Gb, for a total of ~300GB.

Dataset Description

1.1 DATA SUMMARY

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

- ["To share information", "To combine with other data"]
  - *Comment: Research data and scientific literature interlinking.*

1.1.2 What types of data will the project generate/collect?

- ["Other"]

1.1.3 What formats of data will the project generate/collect?

- [".txt files", "PDF", "RTF", "Text files - MS Word docs, .txt files, PDF, RTF, XML (Extensible Markup Language)"]
- formats: XML, JSON, FASTA - Data Models: Datacite, Scholix, Crossref, OpenAiRE, MEDLINE/PubMed

1.1.4 What is the origin of the data?

- ["Primary data", "Secondary data"]

1.1.5 What is the expected size of the data?

- GB (gigabyte)
  - *Comment: over 300GB of data in tar format encapsulating files with links in Scholix format (schema Version 3)*

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

- ["Researchers", "Research communities", "Education", "The public", "Industry"]
- Main target: Content providers (data archives, publishers, publication repositories), Research Infrastructures, Researchers

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

- No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

- Yes
- DataCite Metadata Schema

3.1.1.3 Will your metadata use standardised vocabularies?

- No
- Datacite relationships terms

3.1.1.5 Will you make the metadata available free-of-charge?

- Yes

3.1.1.6 Will your metadata be harvestable?

- Yes
- *Comment: Rest API: http://api.scholexplorer.openaire.eu*

3.1.1.7 Will you use naming conventions for your data?

- No

3.1.1.9 Will you provide clear version numbers for your data?

- Yes

3.1.1.10 Will you provide persistent identifiers for your data?

- Yes

3.1.1.11 Persistent identifiers

- DOI

3.1.1.12 Will you provide searchable metadata for your data?

- Yes
- *Comment: Datacite format, searchable from OpenAire, Datacite and Google Scholar*

3.1.1.13 What services will you use to provide searchable metadata?

- Registry/Catalogue
- OpenAIRE

3.1.1.15 Will you use standardised formats for your data?

- Yes
- Tape Archive Format
- Couldn't find it? Insert it manually

3.1.1.16 Provide information about used standardised formats

- https://zenodo.org/record/1120275#.YNB5_3X7SfY
- Scholix Guidelines Version 3

3.1.1.18 Are the file formats you will use open?

- Yes

3.1.1.20 Do supported open-source tools exist for accessing the data?

- Yes
- OpenAIRE ScholeXplorer

3.1.1.21 Please describe if data require proprietary tools to access the data?

- NO

3.1.1.22 Will you provide metadata describing the quality of the data?

- No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

- No
- *Comment: Published data does not include personal or confidential data.*

3.1.2.2 Will your data be openly accessible?

- all

3.1.2.4 How will the data be made available?

- ["Repository of Archive"]
- Zenodo

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

- secure with backup and recovery
- *Comment: All files uploaded to Zenodo are stored in CERN's EOS service where each file copy has two replicas located on different disk servers.*

3.1.2.7 Are there any methods or tools required to access the data?

- No

3.1.2.10 Will you also make auxiliary data that may be of interest to researchers available?

- no auxiliary data

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

- Yes
- *Comment: Controlled Vocabularies described in the schema documentation: https://zenodo.org/record/1120275#.YNWlXxP7SDV*

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

- immediately

3.1.4.4 What internationally recognised licence will you use for your data?

- Creative Commons Attribution 4.0 International

3.1.4.5 Do you have documented procedures for quality assurance of your data?

- No

3.1.4.8 Will you provide any support for data reuse?

- Yes
- *Comment: Technical support*

3.1.4.9 How long do you intend to support data reuse?

- More than 10 years

4.1 Allocation of resources

4.1.1 How will the cost of making your data findable, accessible, interoperable and reusable be covered?

- ["Use of institution infrastructure", "Other", "Collaboration with other Projects"]

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

- Yes

4.1.3 Identify the people or roles that will be responsible for the management of the project data

- Sandro La Bruzzo (orcid:0000-0003-2855-1245)

4.1.4 How do you intend to ensure data reuse after your project finishes?

- ["Other"]

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

- Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

- No

6.1.2 What are the methods used for processing sensitive/personal data?

- ["Not available"]

7.1 Other

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

- No

Title: Episciences

Template: Horizon 2020

This dataset describes the Episciences research data. The data is related to the activity of publishing scientific open access articles as an overlay service, operating on top of open access repositories.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

- ["To obtain information", "To share information", "To keep on record", "To make informed decisions", "To improve a product", "To combine with other data"]
  - *Comment: We collect data to peer-review preprints and publish open access scientific articles or data*

1.1.2 What types of data will the project generate/collect?

- ["observational (e.g.", "sensor data", "data from surveys)", "Other"]
- The data we collect or generate is preprints, peer-reviews about those preprints and published scientific articles when the preprints have been endorsed and published by the journals

1.1.3 What formats of data will the project generate/collect?

- ["Text files - MS Word docs", ".txt files", "PDF", "RTF", "XML (Extensible Markup Language)", "Numerical - SPSS", "Stata", "Excel", "Multimedia - jpg / jpeg", "gif", "tiff", "png", "mpeg", "mp4", "QuickTime", "Models - 3D", "statistical", "Software - Java", "C", "Python"]
- Most of the content is text files but we are able to collect any other type of content for instance to publish data journals

1.1.4 What is the origin of the data?

- ["Primary data", "Secondary data"]

1.1.5 What is the expected size of the data?

- GB (gigabyte)
  - *Comment: The size depends on the needs and disciplines of the users, it is difficult to anticipate*

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

- ["Researchers", "Research communities", "Decision makers", "Education", "Economy", "The public", "Industry"]
- Preprints and scientific articles have a broad audience

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

- No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

- Yes
- TEI - Text Encoding Initiative
- Others formats: Dublin Core ; Datacite ; Crossref

3.1.1.3 Will your metadata use standardised vocabularies?

- No

3.1.1.5 Will you make the metadata available free-of-charge?

- Yes
- *Comment: Available with a CC0 licence*

3.1.1.6 Will your metadata be harvestable?

- Yes
- *Comment: OAI-PMH protocol*

3.1.1.7 Will you use naming conventions for your data?

- No

3.1.1.9 Will you provide clear version numbers for your data?

- Yes
- *Comment: We use unique identifiers that are linked to document versions*

3.1.1.10 Will you provide persistent identifiers for your data?

- Yes
- *Comment: A DOI is assigned to each published resource*

3.1.1.11 Persistent identifiers

- DOI

3.1.1.12 Will you provide searchable metadata for your data?

- Yes

3.1.1.13 What services will you use to provide searchable metadata?

- Metadata repository
- OpenAIRE

3.1.1.15 Will you use standardised formats for your data?

- Yes
- Acrobat PDF 1.3 - Portable Document Format

3.1.1.18 Are the file formats you will use open?

- Yes
- *Comment: PDF ; LaTeX*

3.1.1.20 Do supported open-source tools exist for accessing the data?

- Yes
- *Comment: Xpdf, LaTeX*

3.1.1.21 Please describe if data require proprietary tools to access the data?

- NO

3.1.1.22 Will you provide metadata describing the quality of the data?

- No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

- No
- *Comment: Published data does not include personal or confidential data.*

3.1.2.2 Will your data be openly accessible?

- all

3.1.2.4 How will the data be made available?

- ["Project website"]
- https://www.episciences.org/
- Couldn't find it? Insert it manually

3.1.2.5 Please provide URL/Name of used data repositories

- https://arxiv.org/
- arXiv
- https://zenodo.org/
- Zenodo
- https://ir.cwi.nl/
- CWI
- https://hal.archives-ouvertes.fr/
- HAL

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

- secure with backup and recovery
- *Comment: Data is hosted or open archive repositories ; mirrored on Episciences site using CCSD hosting facilities and CCIN2P3 datacenter backup facilities*

3.1.2.7 Are there any methods or tools required to access the data?

- No

3.1.2.10 Will you also make auxiliary data that may be of interest to researchers available?

- after publication
- *Comment: Every version of preprints are available on open access repositories*

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

- No

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

- No

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

- immediately

3.1.4.4 What internationally recognised licence will you use for your data?

- Creative Commons Attribution 4.0 International

3.1.4.5 Do you have documented procedures for quality assurance of your data?

- No

3.1.4.8 Will you provide any support for data reuse?

- Yes

3.1.4.9 How long do you intend to support data reuse?

- More than 10 years

4.1 Allocation of resources

4.1.1 How will the cost of making your data findable, accessible, interoperable and reusable be covered?

- ["Use of national infrastructure", "Collaboration with other Projects"]

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

- No

4.1.3 Identify the people or roles that will be responsible for the management of the project data

- Raphaël Tournoy (orcid:0000-0003-1244-0823)

4.1.4 How do you intend to ensure data reuse after your project finishes?

- ["Institutional archive", "National archive"]

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

- Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

- No

6.1.2 What are the methods used for processing sensitive/personal data?

- ["Anonymising data where necessary", "Privacy constraints and applicable ethical norms", "National laws"]

7.1 Other

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

- No