National Initiatives for Open Science in Europe

Argos for machine actionable and FAIR Data Management Plans

Elli Papadopoulou
Athena Research Center

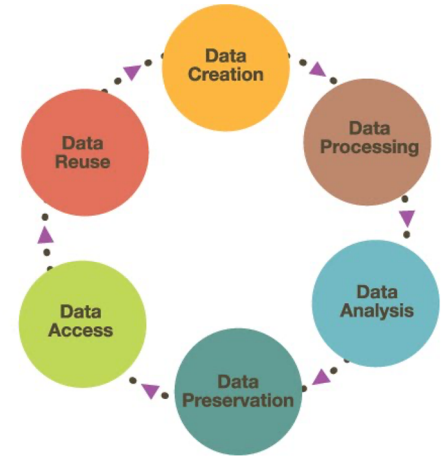
elli.p@athenarc.gr

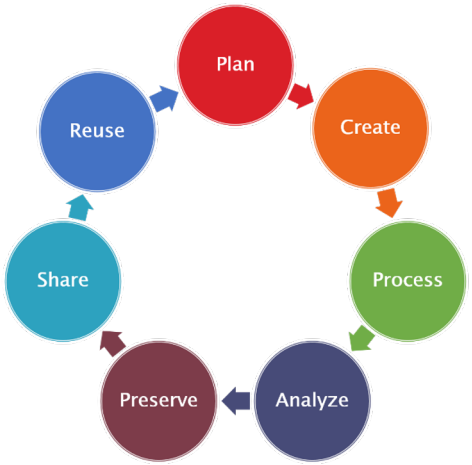
Overview

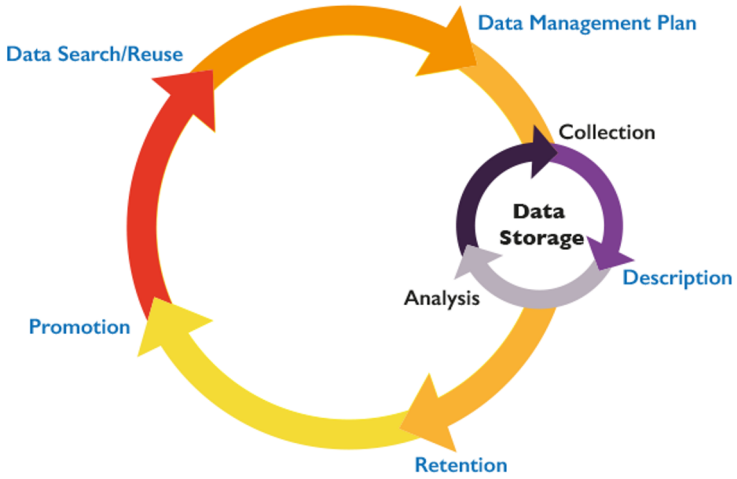
- Research Data Management and Data Management Plans
- DMPs in Horizon Europe
- The DMP Landscape
- ARGOS FAIR DMPs

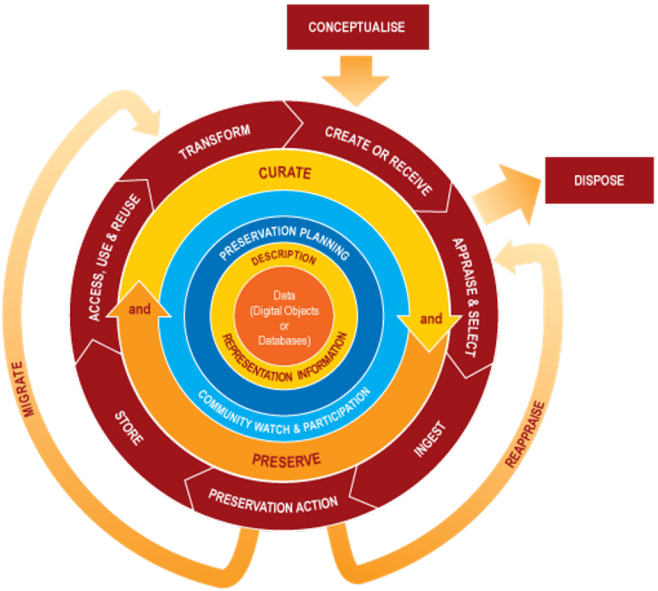
Research Data Management
and
Data Management Plans

RDM lifecycles

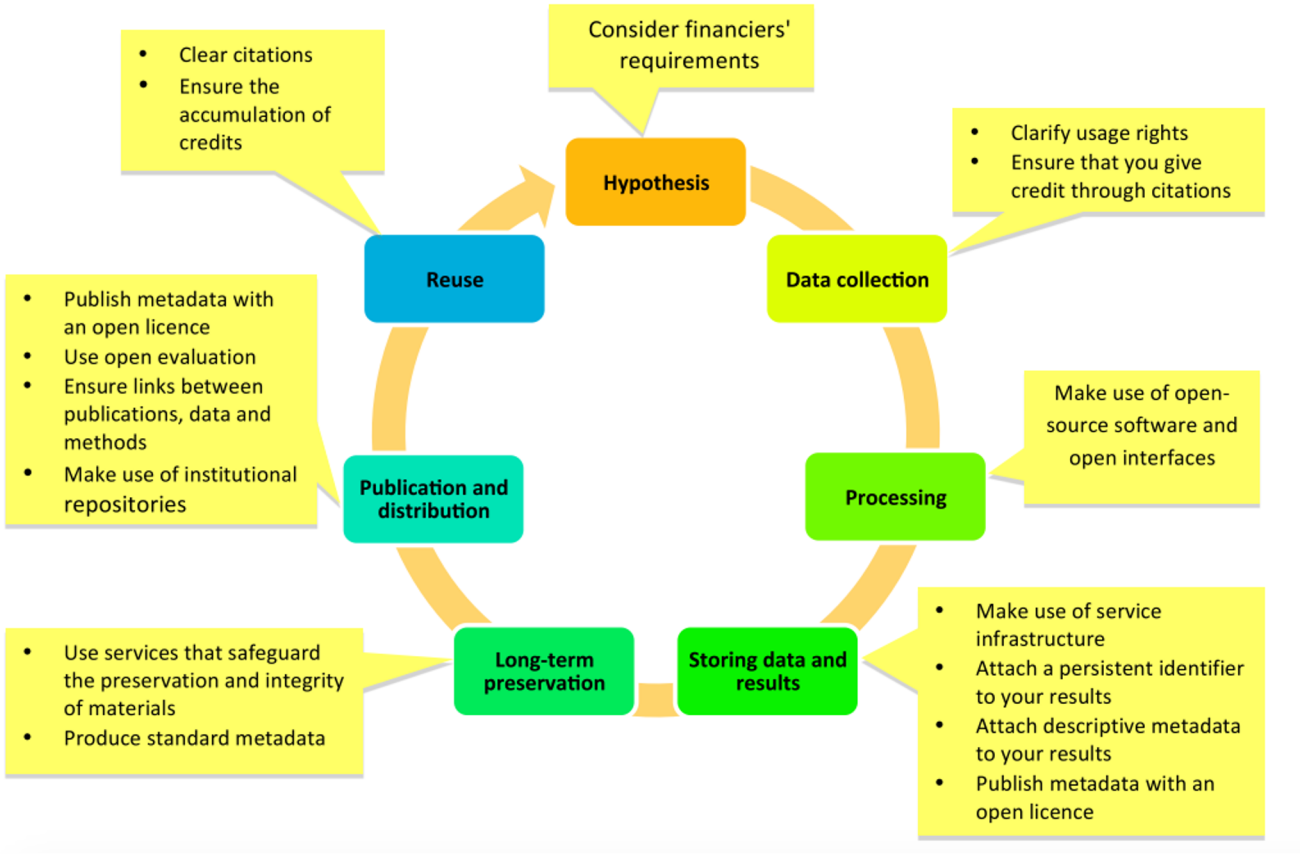








Open Science Practices



(Open Science and Research Initiative, 2014)

Zooming in – The basics

- What is research data?

what has been used or generated (including software) during research process and support/validate its findings

- Why manage research data?
  - Data are understandable, re-usable and reproducible
  - Avoid data loss
  - Get credit
  - Avoid fraudulent/ bad science

Plan – Costing RDM

Plan data management of research activities following research data lifecycle steps

- Costing RDM
  - **Preparing** (DMP)
  - **Data collection**, eg database, formatting, transcription, etc
  - **Data documentation**, eg data description, metadata
  - **Data storage and back-up**
  - **Data access and security**, eg TTP, encryption
  - **Data sharing & reuse**, eg anonymization, copyright, cleaning, digitization
  - **Overall**, eg roles & responsibilities

https://www.openaire.eu/rdm-costs/

Plan – DMPs 1/2

What is a DMP?

Deliverable and “living” document

documents processes undertaken throughout data management lifecycle, including costs

What is not a DMP?

Research assessment method

Example of a DMP

Data Management Plan Information

Data Management Plan for IntelComp. D8.2.

IntelComp sets out to build an innovative Cloud Platform that will offer Artificial Intelligence based services to public administrators and policy makers across Europe for data- and evidence-driven policy design and implementation in the field of Science, Technology, and Innovation (STI). Large STI datasets are processed on the High Performance Computing (HPC) environment part of the European Open Science Cloud (EOSC) initiative. Public administration, STI stakeholders and civil society produce a great amount of dynamic, multilingual, and heterogeneous data so understanding and analysing this data is crucial for evidence-based policy making. The objective of IntelComp is to deliver a platform that provides tools for assisting the whole spectrum of STI policy, i.e., agenda setting, modelling design, implementation, monitoring and evaluation. IntelComp will focus on domains aligned with the European Agenda and the Horizon Europe Missions: Artificial Intelligence, Climate Change and Health. This deliverable is the first version of the Data Management Plan (DMP) of the Intelcomp project produced in the context of Work Package 8 "Project Management and Coordination". It provides an overview of the activities characterizing data management in the project based on the Horizon 2020 policy and FAIR guidelines. This version of the DMP details the types of datasets collected, generated and used within the project, focusing on re-used datasets, and draws a first picture of how they are expected to be handled by / in the IntelComp STI Data Space. In particular the DMP is structured as follows: - Section 1 presents general information on the project and the consortium, - Section 2 provides the dataset description, including how the consortium plans to comply with the FAIR data principles, and - Section 3 describes the (initial) reused datasets. The DMP is produced using Argos DMP service and will be kept as a living document throughout the project's lifecycle with new versions published in Zenodo IntelComp Community when necessary. You may access the machine actionable version of the Argos DMP in zenodo.

Funder

European Commission||EC

Grant

A Competitive Intelligence Cloud/HPC Platform for AI-based STI Policy Making

Organisations

EVERIS SOLUCIONES TECNOLOGICAS SLU, ELLINIKO IDRYMA EREVNAS KAI KAINOTOMIAS, BARCELONA SUPERCOMPUTING CENTER-CENTRO NACIONAL DE SUPERCOMPUTACION, OPENAIRE AMKE, TECHNOPOLIS CONSULTING GROUP BELGIUM, TILDE SIA, FUNDACION ESPANOLA PARA LA CIENCIAY LA TECNOLOGIA, F.S.P., FECYT, Haut Conseil De L'evaluation De La Recherche Et De L'enseignement Superieur (HCERES), Carlos III University of Madrid, ATHINA-EREVNITIKO KENTRO KAINOTOMIAS STIS TECHNOLOGIES TIS PLIROFORIAS, TON EPIKOINONION KAI TIS GNOSIS, Communication & Information Technologies Experts Anonymos Etaireia Symvouleftikon Kai Anaptyxiakon Ypiresion (CITE), ZENTRUM FUR SOZIALE INNOVATION GMBH, MINISTERIO DE ECONOMIA, INDUSTRIA Y COMPETITIVIDAD

Researchers

Datasets

Title: All data (input and output)

Template: Horizon 2020

This is an overview of how input and output data are intended to be handled and managed in the project.

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: IntelComp aims to provide a data lake of Science Technology and Innovation (STI) digital objects to support data- and evidence-driven policymaking. The tools developed facilitate STI actions, such as agenda setting, modelling design, implementation, monitoring and evaluation. The data lake will collect existing or create new datasets and reuse them according to needs in the different phases of work, from the development of the database to content analysis, visualisations and their exchange with other databases and systems.

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

["text mining", "static", "peer-reviewed data sets", "likely published or curated", "derived or compiled (e.g., text mining, 3D models)", "reference or canonical (e.g., static, peer-reviewed data sets, likely published or curated, such as gene sequence databanks or chemical structures)"]

The project will reuse and further contextualize existing data in order to satisfy the innovative ways of their analysis, such as through AI models. New data will be the outcome of merged, deduplicated and enhanced aggregated or crawled content.

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

["Text files", "Numerical", "Models"]

The project will reuse and further contextualize existing data in order to satisfy the innovative ways of their analysis, such as through AI models. New data will be the outcome of merged, deduplicated and enhanced aggregated or crawled content. Other types include: Metadata - text files: CSV, XML, JSON and Artefacts - PDF, JATS XML, plain text, MS Word, HTML.

1.1.4 What is the origin of the data?

["Primary data", "Secondary data", "Other"]

1.1.5 What is the expected size of the data?

TB (terabyte)

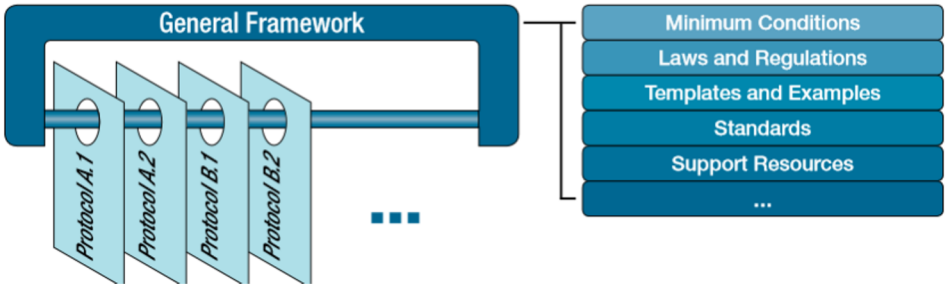
Comment: The size of all data that will be hosted in the Data Lake can not be determined now since the project is at the beginning of its lifetime where certain specifications are put in place.

https://zenodo.org/record/5109933#.Yp1AhahBw2w

Plan – DMPs 2/2

- Depends on the funder/institution requirements
- Differences in research communities
  - Formats, standards, documentation etc

-> Minimum requirements: Science Europe – **DDPs** (Domain Data Protocols)



DMPs in Horizon Europe

Open Science mainstreamed

Publications

- Immediate Open Access
- Intellectual Property Rights
  - Copyright
  - TDM

Research Data

- Immediate access to underlying data
- FAIR data
- DMP “living documents”
- Preservation

Practices

- Incentives for Open Science practices, with appropriate metrics
  - Metadata under CC0
  - Data: CC-BY or CCO license

OpenAIRE

European Commission

ATHENA Research & Innovation Information Technologies

Horizon Europe Template 1/2

1. Data Summary

Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.

What types and formats of data will the project generate or re-use?

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

What is the expected size of the data that you intend to generate or re-use?

What is the origin/provenance of the data, either generated or re-used?

To whom might your data be useful ('data utility'), outside your project?

4. Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)?

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

Who will be responsible for data management in your project?

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

5. Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

6. Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

7. Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

Horizon Europe Template 2/2

2. FAIR data

2.1. Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Will metadata be offered in such a way that it can be harvested and indexed?

2.2. Making data accessible

Repository:

Will the data be deposited in a trusted repository?

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Data:

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Will the data be accessible through a free and standardized access protocol?

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

How will the identity of the person accessing the data be ascertained?

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

Metadata:

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

2.3. Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

Will your data include qualified references1 to other data (e.g. other data from your project, or datasets from previous research)?

2.4. Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Will the data produced in the project be useable by third parties, in particular after the end of the project?

Will the provenance of the data be thoroughly documented using the appropriate standards?

Describe all relevant data quality assurance processes.

Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

3. Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

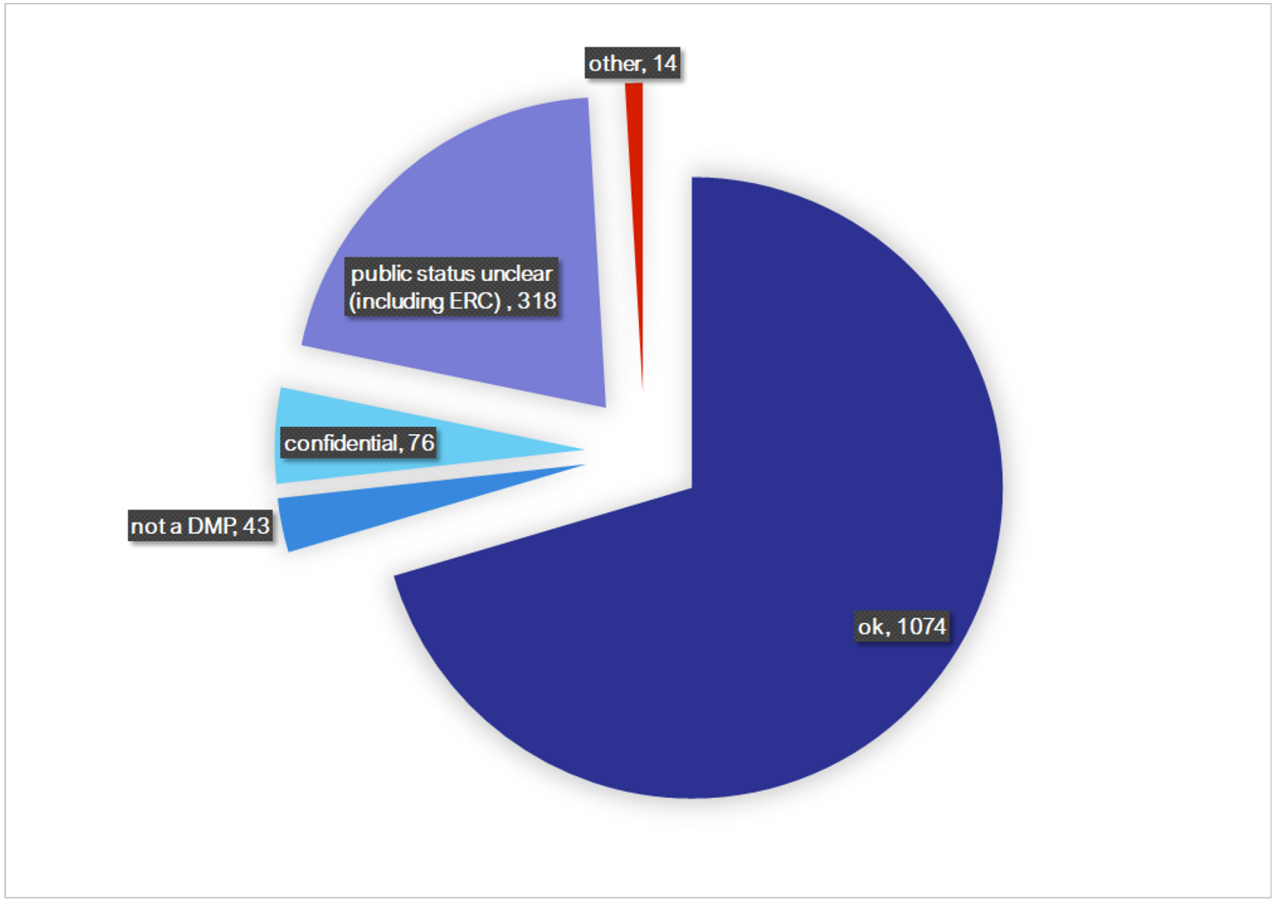
Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

The DMP Landscape

DMPs distribution

Public DMPs (more than 1500) available through CORDIS

-> unclear how they could be re-distributed



https://open-research-europe.ec.europa.eu/articles/1-42

DMPs content

Major challenges (qualitative interviews only)

The following major challenges were raised by the interviewees in the qualitative interviews:

- reading and analyzing partner input and turning it into one understandable document, in particular at the beginning of the project, when there was little experience
- where to put the focus and how much details to give – internal procedures or output; also whether to tackle any data or data underlying publications (the latter strongly preferred)
- understanding the technicalities
- how to create the DMP from scratch with zero experience
- Understanding the requirements and convincing partners to submit thorough information (done through peer pressure). This is easier in newer projects since DMPs are more accepted
- Covering all partners, some of them in non-EU countries where different national policies apply (e.g. on protecting vulnerable groups)

https://phaidra.univie.ac.at/detail/o:1165751

DMPs exploitation

The standard does not necessarily highlight and/or solve the problem of DMPs recording a pool of information, *collectively about all project's datasets*, thus posing obstacles in the evaluation and exploitation of individual datasets.

DMPs can be harvested in many ways, depending on how they have been published. They can be found *classified with diverse labels*, such as articles, reports etc. This means that repository providers need to specify the resource types of DMPs[158] in their systems, and promote them widely so that researchers are aware of and use them.

https://op.europa.eu/en/publication-detail/-/publication/56cc104f-0ebb-11ec-b771-01aa75ed71a1/language-en

-> Individual dataset information difficult to identify from others

-> Resource_type

Data reusability

-> Re-used data difficult to be identified

Table 42. Identifying datasets produced by Horizon 2020 projects

ASSUMPTIONS
A1. All datasets **reported in SyGMa** are produced by the project.
A2. All datasets that **reference the project in their metadata** (as harvested from OpenAIRE) are produced by the project.
A3. Datasets linked to projects via OpenAIRE's **inference system** (text-mined) are not necessarily produced by the project.

The first two assumptions are straightforward, in the sense that in both cases there is no incentive to report a dataset-project link unless the former was a project output. With respect to assumption A3, because OpenAIRE aims to ***link*** projects to research outputs, the inference system is currently *agnostic* towards the semantic relationship between a project and the linked dataset.

Thus, **for those Horizon 2020 datasets not found in EC-Shared or in the harvested OpenAIRE data,** it is not possible to verify that they were created by the projects. For the purposes of this study, we have therefore discarded **1,579 ORG** datasets that are linked to Horizon 2020 only via text mining.

ARGOS FAIR DMPs

Argos

an **open source**, **configurable** and **extensible** tool for **planning Research Data Management** (RDM) activities according to Open Access & FAIR data policies.

- **Templating system**

Dynamic vs static parts

- Access points (APIs);

Import / Export; RDA standard

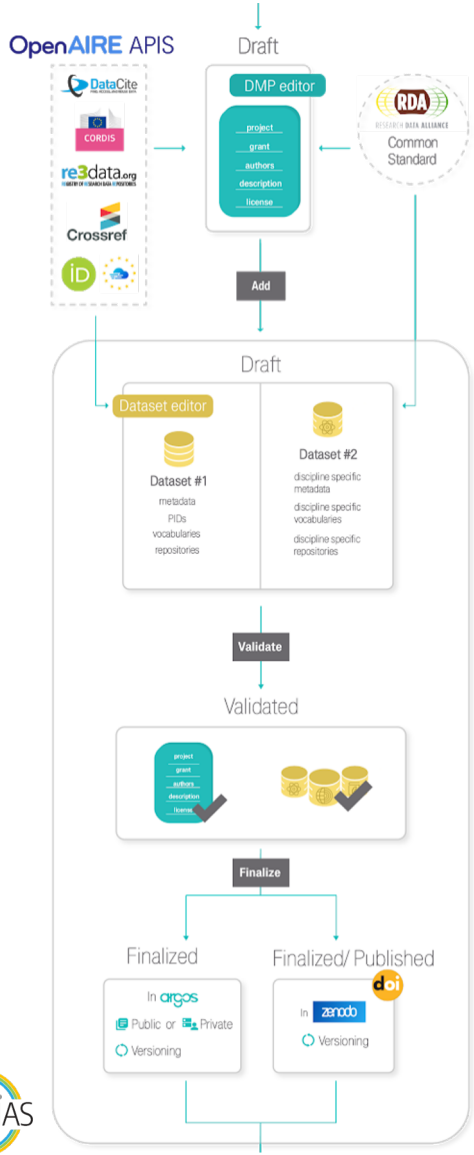
https://code-repo.d4science.org/MaDgiK-CITE/argos

– not just a tool!

- **Full DMP Lifecycle**
  - generate & publish DMPs according to Open and FAIR principles
- **Machine actionable DMP (ma-DMP) outputs**
- **Data Domain Protocols**
  - create many dataset profiles in a single DMP
    - -> e.g. new vs re-used vs sensitive vs discipline specific
- **Contextualized and exploitable DMP data**
  - connect with and enhance reference services and data sources (OpenAIRE, EOSC, etc).
- **Standardization of global practices and collaborations**
  - –> e.g. RDA DMP Common Standard; DMPs exposed in repositories with appropriate resource_type

Write & publish

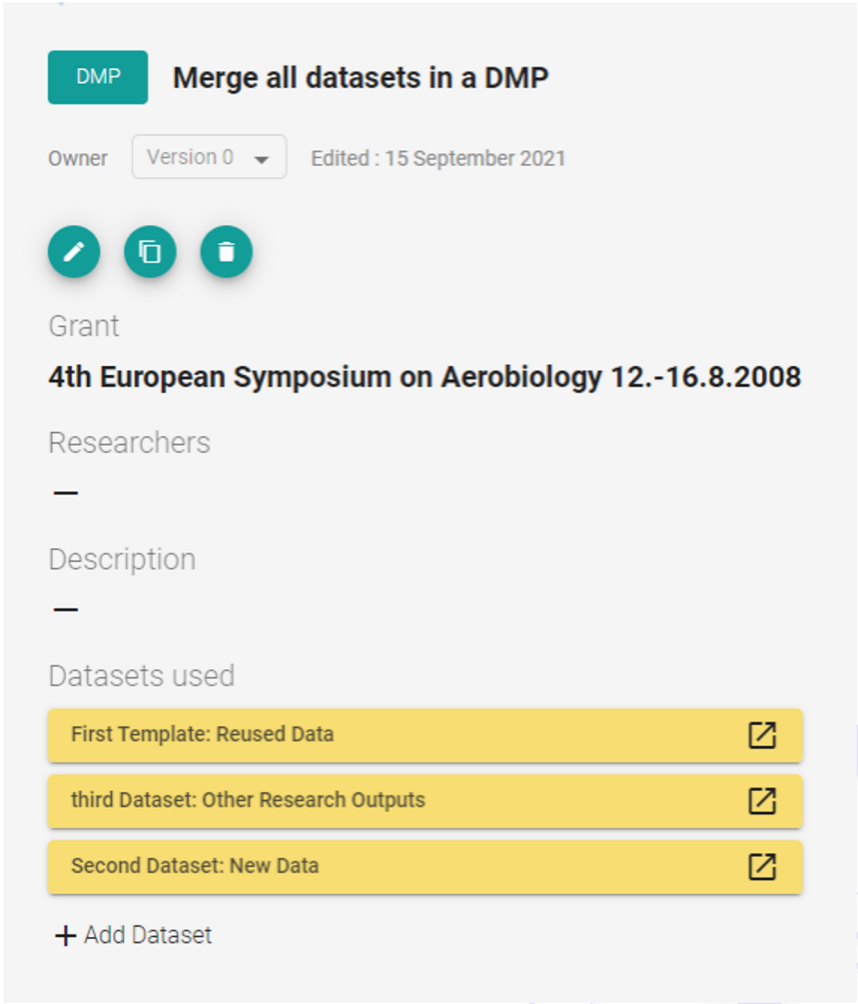
- Full DMP publication
- Two editors to create DMPs
  - Full DMP editor
  - Quick wizard to add new datasets to DMPs
- First validation
- Finalization
- ma-DMP outputs



- Rich documents
- Discoverable through OpenAIRE
- Versioned (provenance)
- Accessible: PIDs (ORCIDs & DOIs)
- Reusable: Licenses
- Preserved in Zenodo

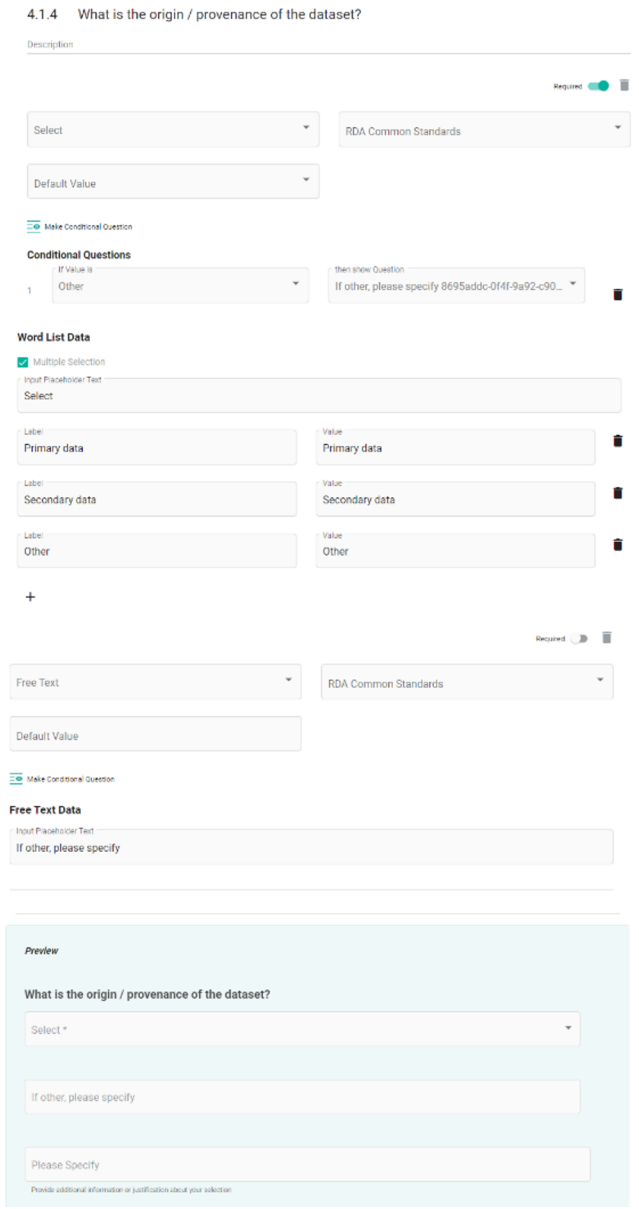
Create dataset profiles

- Different data questions
- Configurable APIs
- Tailored instructions



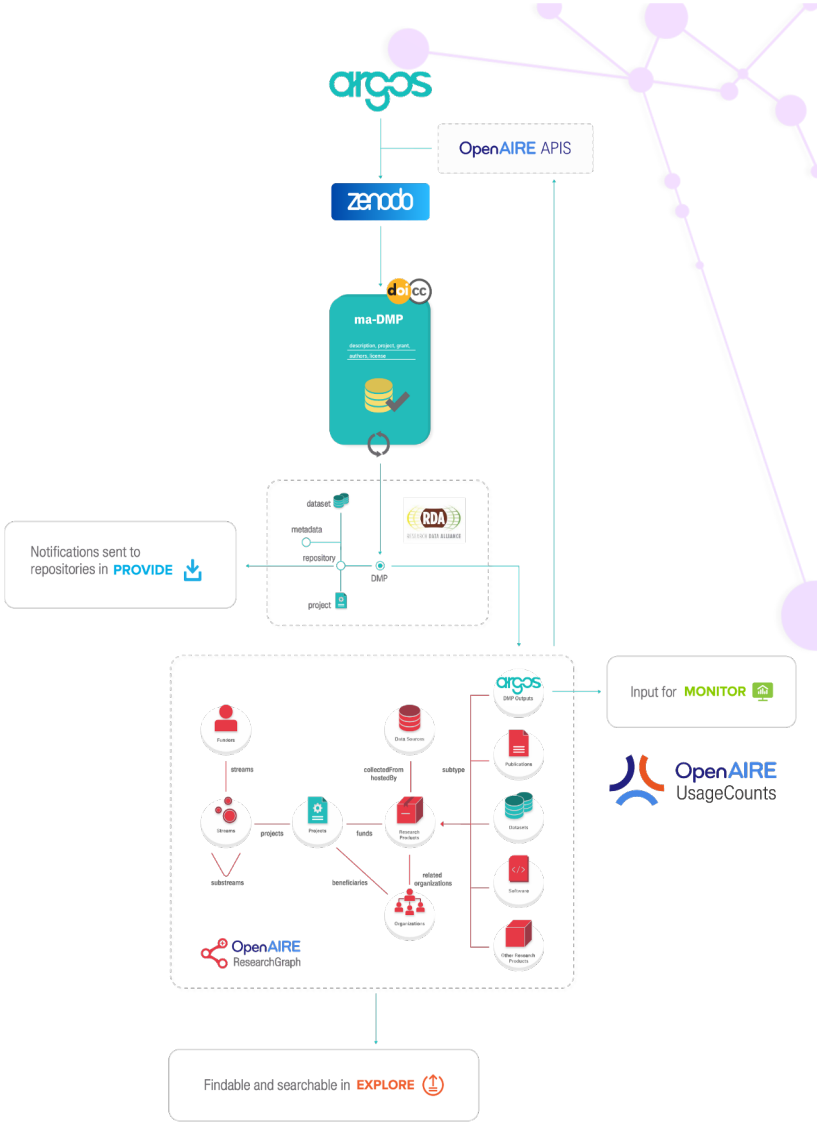
Design machine-actionable templates

- Many inputs to create a question
- Long list of input types
  - Boolean, Multiple choice, Free text, custom APIs…
- Collection of static APIs
- Conditional questions
- Multiplicity
  - x the question can be answered with different input
- RDA compatibility

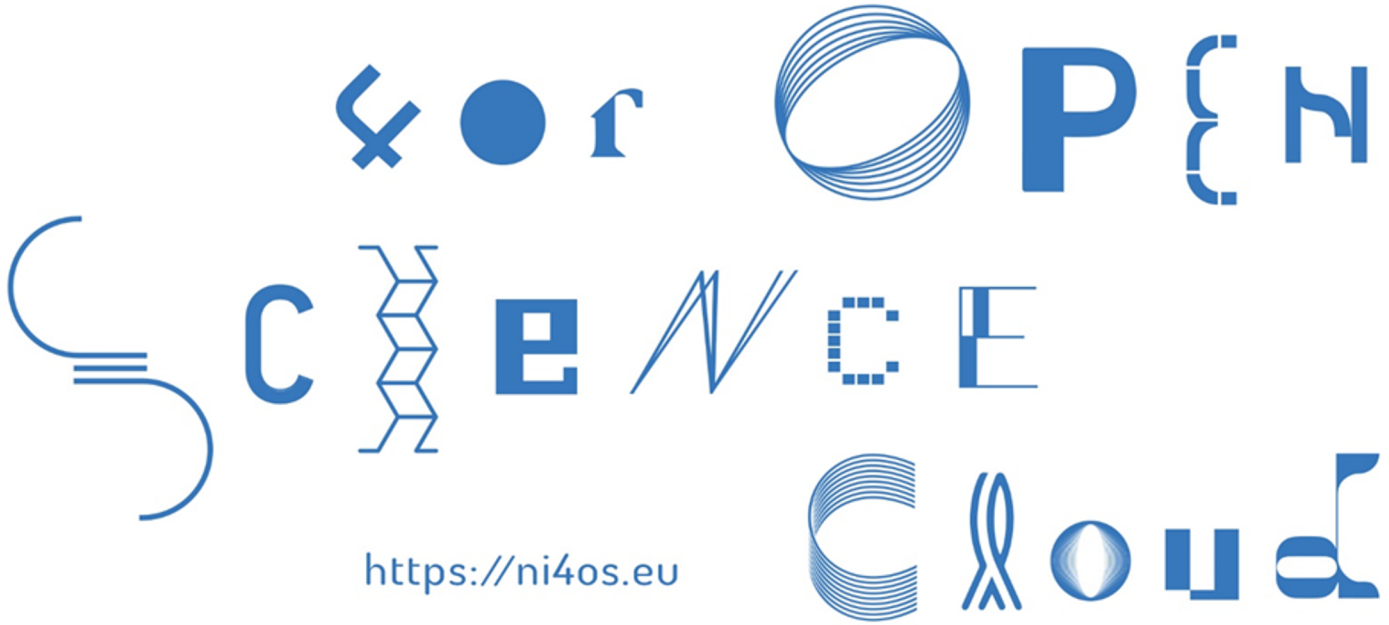


Connect workflows & benefit from OpenAIRE pool of data

- Create subsets of Argos in OpenAIRE that meet your needs.
- Publish in Zenodo.
  - Or integrate with own repositories
- Notify repository managers for new datasets.
  - Integrations to enable pre-filling of DMPs for re-used data
- Exploit DMP entities in the Research Graph.
  - Create links between outputs and entities.
- Combine with validated OpenAIRE data and provide statistics.
  - Define indicators; Add to dashboards.
- Add DMPs under the project's page.



Thanks!



@NI4OS_eu @NI4OS