OpenAIRE
Research Data Management
Briefing paper

Understanding Research Data Management

February 2016

H2020-EINFRA-2014-1
Topic: e-Infrastructure for Open Access
Research & Innovation action
Grant Agreement 643410

Purpose of this document

The Open Research Data Pilot in Horizon 2020 aims to make the research data generated by selected projects open. With this briefing paper OpenAIRE aims to orient all stakeholders new to the topic of openness and research data management, including researchers, research officers, research managers, OpenAIRE National Open Access Desks and the EC's National Contact Points. As a general document, it does not differentiate between academic disciplines with their different workflows, standards and data formats. As an introduction, it does not aim at exhausting the topic of research data management. Readers wishing to go beyond this general introduction will find the final section on training materials beneficial.

Table of Content

HORIZON2020 REQUIREMENTS & OPENAIRE SUPPORT 2

DATA, DATA MANAGEMENT AND REUSABILITY 2

WRITING THE DATA MANAGEMENT PLAN 5

ARCHIVING THE DATA AT THE END OF THE PROJECT AND PROVIDING ACCESS 6

ROLES AND RESPONSIBILITIES IN RDM 7

DATA MANAGEMENT TRAINING MATERIALS 9

Rights

Horizon2020 requirements & OpenAIRE support

If a Horizon 2020 project is part of the Open Research Data Pilot, the Principal Investigator must:

- ☐ Develop and keep up-to-date a Data Management Plan.
- ☐ Deposit his/her data in a research data repository.
- ☐ Make sure third parties can freely access, mine, exploit, reproduce and disseminate it.
- ☐ Make clear what tools will be needed to use the raw data to validate research results, or provide the tools themselves.

OpenAIRE offers support for the pilot, for instance by providing information about Research Data Management:

- ☐ Factsheet Open Research Data Pilot
- ☐ Webinar Open Research Data Pilot (June 9, 2015), with slides
- ☐ Web pages
  - What is the Open Research Data Pilot?
  - How to create a Data Management Plan
  - How to select a data repository
- ☐ Frequently Asked Questions

The current document aims to connect and extend these materials. Section 3 (data management planning) and 4 (archiving the research data for reuse) are practical, whereas sections 2 (how to make research data reusable) and 5 (stakeholders in data management) provide more background information. The document concludes with references to organisations where you can find training materials for those who want to know more about research data management.

Data, data management and reusability

Newton stood on the shoulders of giants. Today's researchers and scientists can use terabytes of research data to advance science and meet societal challenges. However, whether one is able to reuse research data in an effective and efficient way will increasingly depend on whether the data has been properly managed and shared with appropriate metadata and documentation.

"Data management starts on Day One": it is part of a researcher's life, and timely planning how data will be captured, collected, used, managed, stored, sustainably archived, and disseminated is of the essence. You can read more about the data management plan in section 3. Here we explain that the notion of managing data also includes some background documentation.

Definitions of "research data" and "(research) data management" (RDM) abound. The following broad descriptions should be usable in all academic domains:

Research data means data in the form of facts, observations, images, computer program results, recordings, measurements or experiences on which an argument, theory, test or hypothesis, or another research output is based. Data may be numerical, descriptive, visual or tactile. It may be raw, cleaned or processed, and may be held in any format or media.

Data management means all the processes and activities required to manage data throughout the research life-cycle for current and future research purposes and users"[1].

Increasingly data management is explicitly seen as part of good research practice (e.g. FTC[2], NWO[3], RCUK[4]), both to enable replicating a study and to enable reusing and sharing data. These purposes can be considered as a minimal and an aspirational level of science, respectively. It is essential that even data that should never be made openly available be managed properly.

In the guidelines for the Open Research Data Pilot the European Commission focuses on research data that is available in digital form.[5]

Clearly, the types of data vary across academic domains, and different RDM activities may be called for. What all domains have in common, however, is that in order to be reusable, the data need to be accompanied by metadata and most likely other documentation as well. Furthermore, in the context of the EC's Open Research Data pilot, the results are by default Open Access.

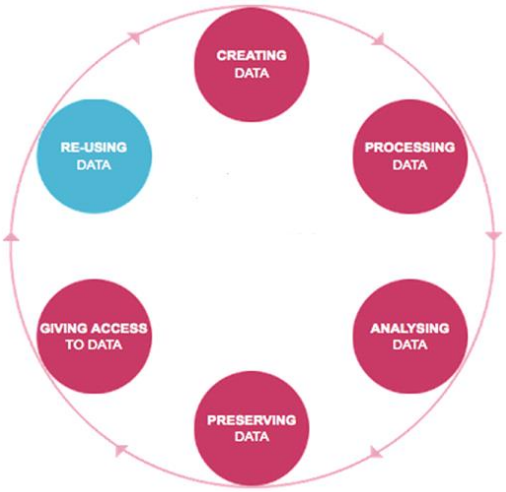

Figure 1. Research Data Life Cycle – UK Data Archive

[1] Queensland University of Technology. (2013). Management of Research data.
[2] The Portuguese Foundation for Science and Technology.
[3] The Netherlands Organisation for Scientific Research.
[4] Research Councils UK data policies.
[5] See footnote 5 in Horizon 2020: Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020.

Metadata is information about the research data; it should enable others to find data. Starting with the dataset's title or name metadata is "standardised structured information explaining the purpose, origin, time references, geographic location, creator, access conditions and terms of use of a data collection"[6]. Several disciplines have such standards; see the disciplinary metadata standards directory provided by the Research Data Alliance[7]. If a domain doesn't have standardised metadata yet, using the Dublin Core standard[8] or the DataCite Schema[9] is recommended. These are generic formats, widely used by online search portals to "harvest" metadata from repositories and archives. Repository managers in research institutions and data archives can provide more information about this topic[10]. Researchers should not aim to define metadata on their own (unless it's part of their research topic, of course), as this goes against the grain of exchanging data and making the data interoperable with other data sources.

In order to understand and use the research data, **documentation** such as code books, lab journals or informed consent forms may be needed, or **software** to access and analyse the research data. In other cases syntax queries used in the statistical analysis or the configuration particulars of a measuring tool are essential for reuse. Usually it is the principal investigator who is responsible for creating and archiving this whole package consisting of data, metadata, documentation and software (more about responsibilities below). Basically, everything needed to replicate a study should be part of the package, including intermediate or processed versions of the data, when these versions reflect design and/or analytical steps in the research process. In practice there can be reasons for omitting particular (intermediate versions of) data, for instance, when regenerating them is cheaper than archiving. Such selection decisions[11] should be documented, along with the methods for recreating the data.

The Open Research Data Pilot applies to both the data and any metadata, documentation, software and tools needed to understand and reuse the data. At a base level, the Open Research Data Pilot applies to all data needed to validate results in scientific publications, but researchers may also choose to share other curated and/or raw data that they feel has value to others and can be shared.

[6] RDNL course *Essentials 4 Data Support.*

[7] *Research Data Alliance (RDA) Metadata Standards Directory Working Group. Please note the earlier version maintained by the DCC.*

[8] *Dublin Core Metadata Initiative.*

[9] *DataCite Schema.*

[10] *OpenAIRE has also a factsheet for repository managers who consider to join OpenAIRE's European repository network.*

[11] *See, e.g., DCC How to select what data to keep.*

Researchers should specify which data will be shared openly in the Data Management Plan. What does "open" mean here? Again, there are several definitions, but the pilot Guidelines on Data Management state "**Open access to research data** refers to the right to access and re-use digital research data. Openly accessible research data can typically be accessed, mined, exploited, reproduced and disseminated free of charge for the user."

Writing the Data Management Plan

A Data Management Plan or DMP is a valuable instrument for defining up front what data and associated metadata and tools will be used, delivered and possible shared in a project. "DMPs will be useful whenever researchers are creating or reusing data, especially where the research involves multiple partners, countries, etc."[12] In the pilot, the first version of the DMP must be delivered within the first 6 months of the project. Support staff at the institute of the PI or of the researcher who is responsible for RDM will usually be involved at this stage, for instance when it comes to facilities for safely storing data or future access to the data (more about roles and support in section 5). By requiring the DMP early in the project, funders stimulate that the necessary arrangements for data management will be in place at an early stage. This is more efficient than for instance starting to think about relevant metadata when a project is near completion. Researchers can use the Horizon 2020 DMP template in the DCC's DMPonline tool (information about setting up a DMP can be found here). In addition to the OpenAIRE information listed above, Research Data Netherlands (RNDL) provides a video clip about "The what, why and how of data management planning" for data supporters and researchers; among other things it explains step by step how to complete a DMP (though not the Horizon 2020 template). Completed sample DMPs – not only for Horizon 2020 – are available from RDNL and DCC. A video clip on answers to possible concerns researchers may have about sharing their data.

The DMP is a living document.

It evolves and gains more precision and substance during the lifespan of the project. It should be updated at least during mid-term and final reviews to fine-tune it to the data generated and the uses identified by the project consortium. Apart from these formal requirements it can be helpful to periodically review the data management plan throughout the project to check how well the stated goals for data management are being met.

[12] *OpenAIRE Webinar Open Research Data Pilot in H2020 (Martin Donnelly, FOSTER/DCC). June 9, 2015.*

Archiving the data at the end of the project and providing access

The H2020 Open Research Data Pilot aims to make the research data **accessible** with as few restrictions as possible, while at the same time protecting sensitive data from inappropriate access. At the end of the project pilot participants must deposit the package of data, metadata, documentation and tools in a research data repository. A data repository is a digital archive collecting, preserving and displaying datasets, related documentation and metadata. Repositories and archives typically use terms like "preservation" and "curation" rather than "archiving" or "storage": long-term accessibility implies expertise and services to convert data to new formats and to add value to the data, for instance by new functionality to query the data.

When data preservation standards or norms exist in the participant's discipline, these should be followed. In deciding where to archive data there may be a number of choices about who will look after them. The choice may be straightforward if there is an established data management facility in the project's domain or in the participant's institution. In order of preference:

1. Use an external data archive or repository already established for your research domain to preserve the data according to recognised standards in your discipline. More information for **selecting a data repository**.
2. If available, use an institutional research data repository, or your research group's established data management facilities.
3. Use a cost-free data repository such as Zenodo.

4. Search for other research data repositories in http://re3data.org/

Preferably, the chosen service would be certified as a Trustworthy Digital Repository, for instance against the guidelines of the Data Seal of Approval.

It is advisable that researchers contact the long-term data repository of their choice when writing the first version of a DMP. Repositories may offer guidelines for sustainable data formats and metadata standards, as well as support for dealing with sensitive data and licensing. Making research data openly accessible is best done using explicit licences or waivers[13]. A lot of data repositories also accept publications, and allow linking between publications and their underlying data. This increases visibility and potential reuse of either.

Roles and responsibilities in RDM

Open Science encourages heterogeneous stakeholder groups to work together for a shared societal goal. It's worth bearing in mind that RDM and data management planning similarly involve multiple stakeholder types:

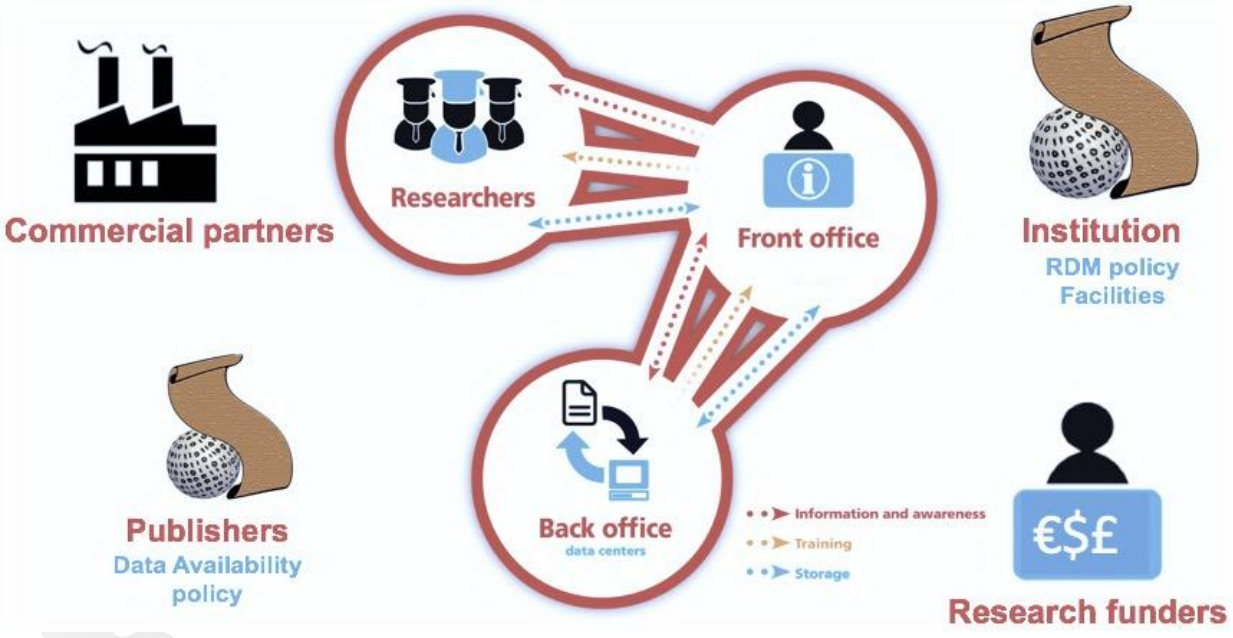

Figure 2: RDM Stakeholder Overview

[13] *For information about licenses see e.g. http://creativecommons.org/licenses/ or http://opendefinition.org/guide/data/.*

- ☐ The principal investigator – ultimately responsible for the data and for data management
- ☐ Researchers, research assistants and/or data managers – involved in day-to-day data management
- ☐ The institution's management – draft and enforce data policies; raise data awareness
- ☐ The institution's research office consisting of library, IT and legal services – provide external data, tools, secure storage and access; expertise on rights management and ethics, data citation, metadata, access and licenses, funder requirements; raise data awareness
- ☐ Research funders – encourage good data practices; invest in data infrastructure; raise data awareness
- ☐ Project partners in academic and other research institutions as well as commercial partners
- ☐ Academic publishers – impose requirements on the availability of data underlying submitted and/or published papers; provide identifiers to cite papers and link to related data
- ☐ Research data repositories – preserve data long term; provide persistent identifiers and data discovery service
- ☐ National Open Access Desks – provide expertise on open access to data and publications, in the context of Horizon 2020 and often beyond

However different, these stakeholders all play a role in planning, carrying out, assessing and benefiting from good data management. Therefore it is essential that researchers share their DMP with all involved in their project.

Data management training materials

Increasingly, RDM is the topic of training and courses for researchers – PhD students in particular – and for those who support them, such as librarians, or collaborate with them, such as data scientists. The organisations below provide good starting points for those who want to know more about RDM.

- **DCC** maintains a summary of RDM training materials from various organisations. Users are encouraged to share their tailored materials with the DCC so they can disseminate them to a wider audience.
- **RDNL** offers a blended-learning course *Essentials 4 data supporters*. Unless otherwise stated, all materials are licensed as CC BY-SA 4.0. All material under this licence can be freely used, as long as Research Data Netherlands is credited as the author.
- **DataOne** offers *Data Management Education Modules* in Powerpoint format (2012). Materials are licensed as CC0 and users may enhance and reuse them for their own purposes.