Data Management Planning

EC Horizon 2020 Pilot on Open Research Data applicants

Version 3.5 June 2021

University of Bristol

Research Data Service

Image: Global European Union, S. Solberg, Wikimedia, CC BY 3.0

University of BRISTOL

SUMMARY

The European Commission recognise that research data is as important as the publications it supports. As part of the stipulation that open access to scientific publications is mandatory for all scientific publications resulting from Horizon 2020 funded projects, projects must also aim to deposit the research data needed to validate the results presented in the deposited scientific publications, known as "underlying data".[1] In order to effectively supply this data, projects need to consider at an early stage how they are going to manage and share the data they create or generate.

Expectations of particular note:

- As of 2017, the Open Research Data (ORD) Pilot covers all thematic areas of Horizon 2020.
- Projects can opt out of the Pilot if they believe there are good reasons for keeping their data closed. Participation in the Pilot is not part of the evaluation of proposals and projects will not be penalised for opting out.
- All projects which are successfully funded under the Open Research Data Pilot are expected to produce an initial DMP deliverable within the first six months of the project.
- List your file types, formats, and the reuse potential for other researchers.
- Where possible use existing metadata standards which will allow for potential integration with other datasets.
- Explain how your data will be shared, and the level of access to be provided (and why).
- Use a repository service to deposit your data and where possible make it and the associated metadata accessible to third parties, free of charge.
- Arrange backup and storage procedures which are most suited to the partners and nature of your project.
- Complete more detailed DMPs at regular intervals throughout your project when changes occur or as a minimum in the run up to mid-term and final reviews.

INTRODUCTION

This guide has been designed to assist any researcher who is planning a Horizon 2020 application and whose proposal includes research data. It provides advice on the Commission's requirements, the usefulness of submitting a Data Management Plan (DMP) as part of your project, and the points a DMP should cover.

Pilot on Open Research Data

In December 2013, the European Commission announced their commitment to open data through the Pilot on Open Research Data,[2] as part of the Horizon 2020 Research and Innovation Programme.[3]

[1] See 'Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020' http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[2] EC Press Release, 'Commission launches pilot to open up publicly funded research data', 16 December 2013 http://europa.eu/rapid/press-release_IP-13-1257_en.htm

[3] http://ec.europa.eu/programmes/horizon2020/en

The Pilot's aim is to "improve and maximise access to and re-use of research data generated by projects for the benefit of society and the economy". The Pilot applies to two types of data; that which is needed to validate results in scientific publications, and any other deemed valuable by the project. Participating projects are expected to make both of these types available for use by other researchers, industries and citizens. During the 2014-2016 work programme the Pilot only applied to certain key areas of Horizon 2020, which accounted for approximately 20% of the overall programme. As of 2017, the ORD Pilot has been extended to all the thematic areas of Horizon 2020.

Projects can choose to opt out of the Pilot to protect intellectual property and personal data, because of security concerns, or if the research will be compromised by making data open. The EC is however keen to emphasise that participating in the Pilot does not mean opening up all your research data, but instead following the principle of "as open as possible, as closed as necessary", and generally encouraging good data management as part of best research practice.

Requirements and expectations

The EC's 'Guidelines on FAIR data management in Horizon 2020' document[4], referred to throughout this guidance, provides information on how applicants and beneficiaries of projects under the Pilot should address the issue of data management. The document should be read in conjunction with this guide.

At the research proposal stage all projects participating in the ORD Pilot are required to provide a short outline of their data management policy under the 'Impact' criterion (as participation is not part of the evaluation criterion a fully developed Data Management Plan is not required at this stage). The outline should cover the following issues:

- What standards will be applied?
- How will data be exploited and/or shared/made accessible for verification and reuse? If data cannot be made available, why?
- How will data be curated and preserved?

It should also reflect the current state of any consortium agreements around data management and be consistent with IPR requirements. Costs associated with open access to research data can be claimed as eligible costs of a Horizon 2020 grant, so the application should include resource and budgetary planning for data management where necessary.

Projects successfully funded will then be required to produce a first version of a DMP as a deliverable by month six of the project at the latest. A template for this is provided.[5]

More detailed versions of the DMP should be submitted as additional deliverables at later stages of

[4] 'Guidelines on FAIR Data Management in Horizon 2020', v3.0. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[5] 'Guidelines on FAIR Data Management in Horizon 2020', v3.0, p.6. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

the project, as a minimum by the mid-term and final reviews, but also when any significant changes occur such as the generation of new data sets or changes in consortium agreements.

As well as the University's internal Research Data Service,[6] additional support in research data management for Horizon 2020 projects will be provided by projects funded under the e-infrastructures call. In particular, the OpenAIRE project has been established to support researchers in meeting EC requirements for Open Access and the Open Data Pilot.[7] This includes a network of local representatives who can offer advice. Their 'Open Research Data Pilot Factsheet'[8] and 'Research Data Management Briefing'[9] documents should be read in conjunction with this guide.

Planning for data management and sharing in large consortium projects can be complicated; make sure you consult with the RED Contracts team[10] and, if appropriate, seek ethical approval.[11] If you are unsure about the implications of handling and sharing data across various institutions and countries, the Research Data Service[12] can also offer advice.

First Data Management Plan

The first DMP deliverable requires information based around a number of key issues, listed in Annex 1 of the guidelines document. The EC specify that you should develop a single DMP for your project to cover its overall approach, but where there are specific issues for individual datasets these should be clearly spelled out. The template provided is a set of questions (see below) that should be answered at a level appropriate to the project. Detailed answers to all the questions are not expected at this stage; the DMP is a living document and it is anticipated that further granularity will be added as the project progresses. Ensure your DMP therefore has a clear version number and where possible a timetable for updates.

It is recommended that projects participating in the ORD Pilot use the DMPonline tool[13] to produce their DMP. This provides the template and guidance for the first DMP, as well as for the later plans. OpenAIRE provide further guidance on using the tool.[14]

1. *Data summary*

Provide an explanation of the different types of data which will be produced, including file formats where possible. It may be difficult for you to predict accurately the nature and extent of the datasets your project will generate at the proposal or early project stage, which is why the EC requests more detailed DMPs as your project develops, as not all data or potential uses are clear from the start. You won't therefore be expected to list everything which is

[6] Research Data Service data.bris.ac.uk/research/
[7] What is the Open Research Data Pilot? https://www.openaire.eu/what-is-the-open-research-data-pilot
[8] OpenAire Open Research Data Pilot factsheet https://www.openaire.eu/or-data-pilot-factsheet
[9] OpenAire Open Research Data Management Briefing Paper https://www.openaire.eu/briefpaper-rdm-infonoads
[10] http://www.bristol.ac.uk/red/contracts/
[11] http://www.bristol.ac.uk/red/research-governance/ethics/
[12] Research Data Service data.bris.ac.uk/research
[13] DMPonline https://dmponline.dcc.ac.uk/
[14] 'What is a Data Management Plan (DMP) and how do I create one?', https://www.openaire.eu/what-isa-data-management-plan-and-how-do-i-create-one

subsequently produced but try to highlight which datasets the consortium agrees are the most significant and likely to have long-term value. You are also expected to state the origins of any collected data, for whom the data might be useful, and what is the purpose of the data collection/generation and its relation to the objectives of the project. Highlighting whether similar data already exists and the possibilities for integration and reuse will also be helpful.

2. FAIR data

H2020 use the FAIR Data Principles[15] as a general concept, in that data is expected to be findable, accessible, interoperable and re-usable.

2.1 Making data findable, including provisions for metadata

Metadata is 'data about data' and is the information that enables data users to find and/or use a dataset. In your DMP you should outline plans for documenting your research data, to meet both your own needs and those of later users.

In attempting to organise and document your data it may help to imagine a secondary data user trying to make sense of your data in your absence, after the end of your project. If presented with only the data itself, a secondary user may be faced with the difficult task of 'unpicking' it. How will they make sense of your file and folder naming conventions? Has any special software been used to create your data? What extra information would they need to make maximum use of your data? Detail in this section the naming conventions and versioning you plan to use to make data understandable to others.

Some subjects have their own metadata standards, which should be used. The Research Data Service can help you identify these in your DMP, or the Research Data Alliance provides a Metadata Standards Directory that can be searched for discipline-specific standards and associated tools.[16] If these are not available, what new metadata will need to be created and how?

Data should also be findable. What search keywords will you provide to optimise reuse, and will you use identifiers such as DOIs so your data can be easily located?

2.2 Making data openly accessible

Give a description of which data will be shared, and the ways in which this will be done. The EC require clarity on the level of access that will be provided; will the data be widely and publicly open or will there be restrictions on who can access the data? If the latter is the case, you will need to outline what access procedures will be put in place. Information on any software and other tools necessary for enabling re-use of datasets should also be provided.

The EC require where at all possible projects to deposit their datasets in a repository where third parties are able to access and reuse the data and associated metadata free of charge. However, the Commission recognise that there will be some instances where complete open access will not be possible, and the DMP is the place to explain those reasons. Some

[15] The FAIR Data Principles https://www.force11.org/group/fairgroup/fairprinciples

[16] RDA Metadata Standards Directory http://rd-alliance.github.io/metadata-directory/standards

repositories offer restricted access levels that allow data to be made available to bona fide researchers through an application process.

You will need to identify the repository where your data will be stored, the type of repository it is (e.g. institutional, discipline-specific), and the licenses and access procedures in place. The OpenAIRE project have developed the general purpose Zenodo repository[17], which can be used for the depositing and publishing of H2020 datasets. The repository can also provide DOIs for code developed in GitHub.

The University of Bristol also has its own Research Data Repository[18] which researchers from any discipline may wish to use. This repository can provide ongoing access to research data for extended periods of time and issue unique DOIs for deposited datasets. Both open and restricted access levels are available. Contact the Research Data Service as early as possible if you believe you'll need to make use of Bristol's data repository.

2.3 Making data interoperable

Use this section to explain how your data will be exchangeable and reusable between other researchers, institutions and countries. One way to ensure this would be to use open file formats and software where possible. A significant barrier to sharing any research digitally is the widespread use of highly specialised file formats. In order to use any digital file, a number of digital technologies must be available, which are known as technological 'dependencies'. These may be fairly common technologies such as a desktop PC, the Windows 10 operating system and Adobe Reader DC 15 software. Or the technology required might be rare and hard to acquire, or even unique (for example any software package made by a single vendor).

You should address this problem by minimising the number of technological dependencies involved in using your digital output/technology as much as possible.

Where dependencies are inevitable you should favour 'open' technologies rather than proprietary ones. Proprietary technologies are owned by a vendor or group of vendors. Commercial pressures may lead to the withdrawal of a particular piece of commercial hardware or software, in favour of a new and possibly incompatible replacement. In contrast, 'open' technologies are supported by a community of users and do not have the same commercial vulnerabilities.

When selecting a file format, your own research needs must come first. If you find you need to use an unusual or non-standard format (one that isn't widely used) you should consider converting it into a more widely re-usable format, once you have finished using the data exclusively for your own purposes. If you're unsure which file formats are 'open' and/or widely used, the Research Data Service can help.

This section should also state if you are using any metadata vocabularies or standards that will make your data interoperable. These could be ones that

[17] Zenodo https://www.zenodo.org/

[18] data.bris Research Data Repository https://data.bris.ac.uk/data/

exist specifically for your discipline or more general standard vocabularies that allow interdisciplinary interoperability.

2.4 Increase data re-use (through clarifying licences)

Explain what licences will be attached to the data that will enable reuse. Where possible the EC expect data to be made openly available, but more restrictive licences will be permissible if you explain your reasons in the DMP. If you are depositing your data in a repository there will be a licence attached to the data when it is made available; in some cases you will be asked to select the most suitable licence from a number of options. If you are unsure what licence should be applied to your data, the EUDAT B2SHARE tool[19] includes a built-in licence wizard that facilitates the selection of an adequate licence for research data.

If for any reason there will need to be an embargo period on your data, for example to allow time to publish results or seek patents, you should state this is the case (and why) at this stage. The general expectation is that data will be made available as soon as possible. You should also clarify how long data will be available for reuse once it is shared.

Describe any quality assurance processes you have in place. Quality should be considered whenever data is created or altered, for instance at the time of data collection, data entry or digitisation. This is particularly important when working in a large consortium, to ensure consistency. You should provide information about the procedures you will carry out to ensure that data quality is maintained, such as allocating time to validate data or entering values into prepared databases or transcription templates.

3. *Allocation of resources*

Outline any anticipated costs for making your data FAIR, and how these will be covered (as already noted, costs related to making research data available are eligible as part of the Horizon2020 grant). For example these might include costs for depositing data in a repository to ensure long-term preservation, or employing a Data Manager to handle large quantities of data. Also use this section to describe who will be responsible for data management within your project. There might be one person who oversees RDM activities on behalf of the whole project, or there could be individuals from each partner institution with responsibilities. Who will decide what data will be kept and for how long?

4. *Data security*

This section requires an outline of the procedures which will ensure the security of your data, including backup and storage. In the case of projects where a number of partners are involved, it is worthwhile deciding at this early stage exactly where everyone's datasets are going to be stored. If data does not need to be shared between partners then robust storage facilities at individual institutions could be used, but make sure that the security and backup procedures of each data holding partner are described within the DMP. Where partners will potentially want to share and compare datasets, a collaborative space may be a more suitable arrangement. Do remember that some

[19] EUDAT B2SHARE https://b2share.eudat.eu/

cloud-based storage options may not be governed by EU legislation (such as the General Data Protection Regulation (GDPR)), and so would not be suitable for storing sensitive data.

It is recommended that, as you create data, you store it in the University's own Research Data Storage Facility (RDSF),[20] managed by the Advanced Computing Research Centre (ACRC).[21] Each research staff member is entitled to 5TB of storage without charge. If your storage quota is used up, or your project requires more storage space, there will be a cost and ACRC should be contacted for guidance before your application is finalised. The back-up procedures, policies and controlled access arrangements used by the RDSF are of a very high standard. Procedures are also in place to allow authenticated, external collaborators to view, add and/or edit data in the RDSF, useful for large consortium projects.

Your DMP should briefly indicate how you'll keep your data safe before it's deposited in a storage facility such as the RDSF. This is particularly important if you're conducting any field research. As a minimum requirement, try to ensure that at all times at least two copies of the data exist and that every copy can easily be accounted for and located if required. If you will be transferring data between partners, how will you ensure this is done securely?

The DMP should clarify how long your data will be stored/preserved for. The RDSF and the public-facing Research Data Repository provide storage for your data for a minimum of 20 years, but other repository services will be different. Provide details of the storage facility or repository's policies and procedures to demonstrate their suitability for holding your data in the long term.

5. *Ethical aspects*

Whilst projects participating in the ORD Pilot are expected to make their data openly available, it is recognised that there are potential issues when data includes sensitive or personal information. Use this section to outline any ethical or legal issues that could have an impact on data sharing, and how you will try to address these. Include reference to any ethics reviews or other deliverables that are relevant. The OpenAIRE project have produced a factsheet on personal data and the ORD Pilot which offers further guidance.[22]

Including the correct information in consent forms is crucial for the potential sharing of data. Obtaining permission to publish data from human research participants is essential even if data is to be anonymised before publication. This is because some risk of re-identification may remain, even after anonymisation, and participants should be made aware that others outside of the research project may be able to view this data. Also, even if a participant has the right to withdraw from a study, it may not be possible to remove their data. The Research Data

[20] Research Data Storage Facility, http://www.bristol.ac.uk/acrc/research-data-storage-facility

[21] Advanced Computing Research Centre, https://www.bristol.ac.uk/acrc/

[22] 'Personal data and the Open Research Data Pilot: how can OpenAIRE help?' https://www.openaire.eu/personal-data-and-the-ordpilot-factsheet

Service has produced a guide to sharing data concerning human participants that includes sample consent form wording.[23]

6. *Other issues*

Use this final section of the DMP to state if you will be using any other national/funder/sectoral/ departmental procedures for data management and what these are.

Detailed data management plans

More elaborate versions of the DMP can be submitted at later stages of the project. If you have thought carefully about how you will manage your data in the early stages of the project, later DMPs allow you to provide a greater degree of detail as more becomes known about the nature and potential use of individual datasets. As is the case with much research, plans and ideas can also change as the project progresses, so these later DMPs are also an opportunity to clarify any changes and how these will in turn affect the future plans for any datasets.

[23] Sharing research data concerning human participants, http://bit.ly/35hldfU