Data Management Plans

(Version 2, December 2016)

(Doc.16/61) (B6SR\GT SR\GDR\PGD_v2Publica_desembre16-EN.docx, 22.12.16)

This document is intended to support researchers in creating their Data Management Plans. It is specifically aimed at projects financed under the EU's Horizon 2020 programme to create a FAIR data management plan (version 3.0, July 2016).

Key to the numbering:

- A number indicates the fields that are required in Horizon 2020.
- A capital letter indicates the elements that should be taken into account when filling in each field.
- A lowercase letter indicates the descriptions of each element and a sample of real examples.

This document was prepared by the CSUC's Working Group to Support Research, which is composed of representatives from the following universities: University of Barcelona, Universitat Autònoma de Barcelona, Universitat Politècnica de Catalunya, Pompeu Fabra University, University of Girona, University of Lleida, Universitat Rovira i Virgili, Open University of Catalonia, University of Vic-Central University of Catalonia, Ramon Llull University and Universitat Jaume I.

The examples cited are examples[1] of Data Management Plans that are available online.

Digital version: http://hdl.handle.net/2072/270395.

---

[1] Actris (Grant 654109), Citilab (Grant 635898), ConnectingGEO (Grant 641538), DIMENSION (Grant 688003), DR-BOB (Grant 696114), EGI-Engage (Grant 654142), FREME (Grant 644771), iCirrus (Grant 644526), MAGIC (Grant 689669) MAMI (Grant 688421), MMT (Grant 645487), RAMCIP (Grant 643433), SatisFactory (Grant 636302), Solidus (Grant 649489), Step (Grant 649493), Tandem (Grant 654206), UMobile (Grant 645124), U-Turn (Grant 635773), WaterInnEU (Grant 641821)

1. Data summary

1.A State the purpose of the data collection/generation

1.A a) Description

A short introduction text explaining the purpose of the data collection/generation.

1.A b) Real example

Ex. 1 Its purpose is to summarize the main characteristics of the tools and clearly identify the owner, the functionalities and the application and target markets, as well as the stage of commercial development. The PSS are the basis for the presentation of the tools in the marketplace and will be the source of information about them when browsing the platform.

Ex. 2 The marketplace platform is the core of the project. It will be a web based platform that will connect the outcomes developed in previous EU funded activities (currently collected by the EIP Water marketplace or the WISE-RTD portal) with the already existing data available at European level (INSPIRE directive, Eurostat Open Data Portal, etc.) and also to the SMEs and companies that are able to deploy products and offer services to the users from these tools and data. The platform will also offer an independent marketplace supported by technical and commercial expertise as a service for users (river basin district managers) that will allow them to access products and services best fitting their priorities, capabilities and procurement processes (based on the previous products).

1.B Explain the relation to the objectives of the project

1.B a) Description

A short introduction text explaining the relation of the data collected/generated to the objectives of the project.

1.B b) Real example

Ex. 1 The inventory dataset is meant as a starting point for the project marketplace and is serving as input for selecting products, solutions and companies that are being integrated in the prototype of the marketplace.

Ex. 2 In the initial steps of the project, a database with an inventory of datasources, companies and solutions to support the development of the project has been created based on the information found in specialized websites.

1.C Specify the types and formats of data generated/collected

1.C a) Description

Description of the content and scope of the data. Research data are generated for various reasons and through various processes, and may be of the following types:

- Observational: data captured in real time (neuroimages, sample data, sensor data, survey data, etc.).
- Experimental: data captured by laboratory equipment (gene sequences, chromatograms, magnetic field data, etc.).
- Simulation: data generated from test models (climate, mathematical, economic, etc.).
- Derived or compiled: data that are reproducible but difficult to reproduce (text and data mining, 3D models, compiled databases, etc.).
- Reference: conglomerated datasets (databases of gene sequences, chemical structures, spatial data portals, etc.).
- Others.

Format of the data (text, numeric, image, etc.) must also be indicated.

1.C b) Real example

Ex. 1 A range of experimental, simulation and theoretical data will be collected on excel spreadsheets for easy accessibility.

1.D Specify if existing data is being re-used (if any)

1.D a) Description

If you reuse a dataset, specify the source from which it was extracted for example from a relevant repository. If purchasing or reusing existing data sources, explain how issues such as copyright and IPR have been addressed.
When creating new data sources, explain why existing data sources cannot be reused.

1.D b) Real example

Ex. 1 The database will incorporate data from existing sources including the Roman Amphorae digital resource (http://dx.doi.org/10.5284/1028192).

1.E Specify the origin of the data

1.E a) Description

If the data are generated within the project, state the source of the data.
If the data are collected, state the source from which they were extracted.

1.E b) Real example

Ex. 1 (data generated within the project) Dataset produced by simulation tools and/or by real life trials will be used as a means to quantify the performance advantages that the project architecture offers compared with current practices.

Ex. 2 (data collected) The data sources and solutions come from European research projects funded by the EU Commission and so have been collected from official portals, primarily those from the EC and from the project websites themselves.

1.F State the expected size of the data (if known)

1.F a) Description

State the approximate volume of the datasets. Consider the implications of data volumes in terms of storage, backup, cost and access. Estimate the volume of data in MB/GB/TB and how this will grow to make sure any additional storage and technical support required can be provided.

1.F b) Real example

Ex. 1 It is not a big file (250 KB), as the information recorded in it is in the text format and the water related EU funded project and the companies are limited.

1.G Outline the data utility: to whom will it be useful

1.G a) Description

State the group/s who may be interested in the data.

1.G b) Real example

Ex. 1 The dataset will be valuable for benchmarking algorithms for object recognition, robotics navigation and grasping.

2. FAIR data

2.1 Making data findable, including provisions for metadata

2.1.A Outline the discoverability of data (metadata provision)

2.1.A a) Description

Reference to metadata standards of the discipline. If there are none, description of the metadata that will be created and how.

2.1.A b) Real example

Ex. 1 The metadata standard used to describe the dataset will be the Dublin Core Schema, as it is a flexible and common used standard and is also the one adopted by the European OpenAIRE repository.

2.1.B Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

2.1.B a) Description

Explain how the data and metadata are assigned to a globally unique and eternally persistent identifier (DOI, Handle…)

2.1.B b) Real example

Ex. 1 The institutional repository provides a unique URL to access the document with the format https://repository/record/1234.

Ex. 2 The repository assigns Handle/DOIs for persistent identification and citability of the dataset.

2.1.C Outline naming conventions used

2.1.C a) Description

Describe how the data will be organized: the structure and name of the files.

2.1.C b) Real example

Ex. 1 The Project dataset identification follows the naming: Data_<WPno>_<serial number of dataset>_<dataset title>. Example: Data_WP2_1_User generated content.

Ex. 2 Files will be structured in terms of project and lead partner and publication id and figure and filenames.

2.1.D Outline the approach towards search keywords

2.1.D a) Description

State how content search keywords will be created to optimize retrieval and reuse.

2.1.D b) Real example

Ex. 1 ORD has to be findable easily, rapidly and identically. Therefore, exact and standard measures have to be used to identify the data sets. This can include the definition and use of naming conventions, search keywords, version numbers, metadata standards and standard data identifiers.

2.1.E Outline the approach for clear versioning

2.1.E a) Description

Describe how version control will be organized.

2.1.E b) Real example

Ex. 1 Version control mechanisms should be established and documented before any data are collected or generated.

2.1.F Specify standards for metadata creation (if any). If there are no standards in your discipline describe what type of metadata will be created and how

2.1.F a) Description

State the metadata standards that will be used. We recommend using metadata standards that are specific to the discipline. Consult metadata standards.

If metadata standards are not used, state what metadata will be generated (manually or automatically) and how.

2.1.F b) Real example

Ex. 1 Metadata are created manually by depositors in the deposit form at the repository.

Ex. 2

(1) The data are expected to be provided in ANSI SQL, XML or text (ASCII) format. For this dataset, data citation and metadata practices derived from the community will be considered.

(2) There are no standards for these logs. A possible solution is project servers such as AAA servers. In this case, the logs would include the attributes defined by "project".

Ex. 3 Each file associated with data will be accompanied with unique specified metadata in order to allow ease of access and re-usability. Below, the form to be followed is presented.

Ex. 4 Standards such as the Dublin Core and ISO/IEC 11179 Metadata Registry (MDR), which addresses issues in the metadata and data modelling space, will be taken into account.

2.2 Making data openly accessible

2.2.A Specify which data will be made openly available. If some data is kept closed provide rationale for doing so

2.2.A a) Description

Description of whether and how data will be shared, including access procedures, embargo periods (if any), and definition of whether access will be widely open or restricted to specific groups. If some cannot be made openly available you must justify why.

2.2.A b) Real example

Ex. 1 When no embargo period applies and a data package related to a case study has been marked as public, it will be made openly available. Only data gathered by partners outside of the project work plan and protected by IPR, or inside the work plan but containing confidential information (e.g. related to personal interviews) will be kept closed for privacy reasons.

Ex. 2 To protect the Personal data of participants, this dataset will not be open access in any moment, and it will be restricted to the partners of the project.

Ex. 3 As described in Section 1, due to privacy concerns project will not provide general access to raw data sets to external users. These raw data are stored inside project with "copyright project consortium, all rights reserved".

2.2.B Specify how the data will be made available

2.2.B a) Description

Describe how the data will be shared, i.e. who will have access to the dataset. You can create a procedure to temporarily make the data accessible to other group members, project partners, and the general public. You should state whether the data will be open access and in what reasonable period. One possibility is to offer them together with the publications. If embargo periods are required, this is where you need to specify them.

2.2.B b) Real example

Ex. 1 Timeliness of Data Sharing. The data sharing should occur in a timely fashion. This means that the data resulted from the research conducted in the project should become available close to the project results themselves. Furthermore, it is reasonable to expect that the data will be released in waves as they become available or as main findings from waves of the data are published.

Ex. 2 Embargo: None.

Ex. 3 Potential users will find out about the data through publications and the website. Data will be made available on publication of the associated paper and will be made accessible on request, under conditions agreed on a case-by-case basis, and after agreement of the project consortium.

2.2.C Specify what methods or software tools are needed to access the data. Is documentation about the software needed to access the data included. It's possible to include the relevant software (e.g. In open source code)

2.2.C a) Description

You must also include any technical requirements for access to and reuse of data. For example, whether you need special software.

2.2.C b) Real example

Ex. 1 There is not included documentation about the software needed to access the data, but it consists in PDF files, which is a common used format and widely distributed software.

Ex. 2 All the data needed to create and maintain the marketplace is being made openly accessible through the GitHub repository, along with the corresponding technical documentation.

2.2.D Specify where the data and associated metadata, documentation and code are deposited

2.2.D a) Description

State the repository in which the data and associated metadata, documents and code will be stored. It can be the same repository or different repositories for the different types of content, for instance code could be deposit in a specific repository for code. There is available a document on recommendations to select research data repositories from CSUC (in Catalan).

It is important to use a repository that provides permanent links (DOI, handle) to data in order to facilitate findability and citation.

2.2.D b) Real example

Ex. 1 The created dataset will be shared using a data management portal that is going to be created and maintained by the project. The public version of the data will be shared within the portal as well. Of course, the data management portal will be equipped with authentication mechanisms, so as to handle the identity of the persons/organizations that download them, as well as the purpose and the use of the downloaded dataset.

Ex. 2 Data will be shared via a repository held and managed by the lead participant, the University (http://www.example.edu).

Ex. 3 The consortium agreed to deposit the data and publications generated by the project in Zenodo, unless for a specific project there is a subject specific repository that is considered more relevant.

Ex. 4 Deposit the research data in an online research data repository. In deciding where to store project data, the following choice will be performed, in order of priority:

- An institutional research data repository, if available

- An external data archive or repository already established in the project research domain (to preserve the data according to recognised standards)

- The European sponsored repository: Zenodo

- Other data repositories (searchable here: re3data), if the previous ones are ineligible

2.2.E Specify how access will be provided in case there are any restrictions

2.2.E a) Description

In case public access to data is restricted for any justified reason please specify if data would be accessible to an individual partner, to all partners or under request. Specify procedures of how to request access to restricted data and under which conditions it would be granted. Moreover specify if restrictions will be lifted after a period of time.

2.2.E b) Real example

Ex. 1 Data availability is ore categorised at this stage in one of three ways:

- Open Data that is shared for re-use or that underpins a scientific publication.

- Consortium Confidential data that is accessible to all partners and the Energy Expert Group, but retained within the consortium and subject to the project NonDisclosure Agreement (NDA).

- Private Data that is maintained by an individual partner for their own purposes.

Ex. 2 According to the article 28.1 "Obligation to exploit the results" of the GA, this dataset will be open access from 30th May 2022, in order to respect the right of each country/partner involved to exploit its own results. However, the dataset will be open access to all Consortium members after June 2018, at the end of the project. However, data published before the period mentioned above will be automatically open access, in order to made data accessible for verification and re-use

Ex. 3 Project can license these data for use by other researchers on a case-by-case basis, after these researchers have come to an agreement with the project-consortium to access the raw data and not expose any PII in any derived results.

2.3 Making data interoperable

2.3.A Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability

2.3.A a) Description

Explain what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

2.3.A b) Real example

Ex. 1 The documents are based on XML according to a DTD. The vocabulary is represented in SKOS. The RDF data is based on an OWL ontology.

Ex. 2 Where possible, standard codes has been followed, for example in the case of country identification, registered following the ISO 3166-1-alpha-2 codes.

2.3.B Specify whether you will be using standards vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability. If not, will you provide mapping to more commonly used ontologies

2.3.B a) Description

Specify whether you will be using standards vocabularies for all types of data. If this is not your case, indicate their correspondence with the most common classifications of your specialty.

2.3.B b) Real example

Ex. 1 Other types of data have been registered following internal codifications, clearly specified within the file

2.4 Increase data re-use (through clarifying licences)

2.4.A Specify how the data will be licenced to permit the widest reuse possible

2.4.A a) Description

If the data are made available to other researchers and the general public, you need to specify what degree of reuse is allowed. This level of reuse will be marked by the establishment of licenses. The EC proposes the use of Creative Commons CC BY or CC0 licences, but there are others.

2.4.A b) Real example

Ex. 1 The deliverables associated to the dataset are licensed through an All rights reserved license as they are working papers not intended to be re-used. Nevertheless the database should be shared as a possible reusable dataset. For this reason, when deposited to the repository, an Attribution-NonCommercial license (by-nc) will be requested. The data is currently available for re-use from the project website and will also be findable and reusable through the final depositing repository (the institutional one or Zenodo) and from OpenAire, the latest by the end of the project.

2.4.B Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed

2.4.B a) Description

Describe when will you make the data available for re-use. You can create a procedure to temporarily make the data accessible to other group members, project partners, and then to general public. You should state whether the data will be open access and in what reasonable period. One possibility is to offer them together with the publications. If embargo periods are required, this is where you need to specify them.

2.4.B b) Real example

Ex. 1 The data will remain re-usable after the end of the project by anyone interested in it, with no access or time restrictions.

2.4.C Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project. If the re-use of some data is restricted, explain why

2.4.C a) Description

In principle the data should be made available to other researchers and the general public with the fewest possible restrictions. However, there may be several reasons for not sharing them: ethical reasons, protection of personal data, involvement of intellectual and/or industrial property rights, commercial interests, etc. You must specify the reasons why a dataset will not be shared.

2.4.C b) Real example

Ex. 1 IPRs and Privacy Issues. Data access and sharing activities will be rigorously implemented in compliance with the privacy and data collection rules and regulations, as they are applied nationally and in the EU, as well as with the H2020 rules. Raw data collected through the interviews from externals the consortium sources may be available to the whole consortium or specific partners upon authorization of the owners. This kind of data will not be available to the public. The results of the project will become publicly available based on the IPRs, as described in the Consortium Agreement.

Ex. 2 The full dataset will be confidential and only the members of the consortium will have access to it. Furthermore, if it is decided to make specific portions of it (e.g. metadata, statistics, etc.) widely open access, a data management portal will be created that should provide a description of the dataset and link to a download section. Of course, these data will be anonymized so as not to have any potential correlation and identification of the ethical issues with their publication and dissemination.

Ex. 3 Each archived data set will have its own permanent repository ID and will be easily accessible. We expect most of the data generated to be made available without restrictions and only data sets subject to IPR and confidentiality issues will be restricted. Where this is going to be the case, agreements will be made based on the individual data sets. Requests for the use of the data by externals will be approved by the project consortium.

2.4.D Describe data quality assurance processes

2.4.D a) Description

Describe what are your data quality assurance processes. How/when internal data quality assessments will be implemented?

The data quality can be ensured by different measures. These include validation of the sample, replication and comparison with results of similar studies and control of systematic distortion.

2.4.D b) Real example

Ex. 1 The quality of the dataset is guaranteed by the platform functioning.

Ex. 2 The data quality is ensured by different measures. These include validation of the sample, replication and comparison with results of similar studies and control of systematic distortion.

2.4.E Specify the length of time for which the data will remain re-usable

2.4.E a) Description

How long do you expect your data to remain re-usable?

2.4.E b) Real example

Ex. 1 For re-usability the data will be stored on the webpage or on a repository system when implemented for at least ten years.

3. Allocation of resources

3.A Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

3.A a) Description

State the approximate cost for making your data FAIR and how you plan to cover them.

3.A b) Real example

Ex. 1 There are no costs associated to the described mechanisms to make the database FAIR and long term preserved.

Ex. 2 The costs for depositing the dataset with the project, and subsequent resources required to make the dataset publicly available have been included within specific Work Packages within the project.

3.B Clearly identify responsibilities for data management in your project

3.B a) Description

Explain the responsibilities for data management in your project

3.B b) Real example

Ex. 1 The project coordinator has the ultimate responsibility for the data management in the project and so, for the Marketplace platform management.

3.C Describe costs and potential value of long term preservation

3.C a) Description

State whether archiving and preservation of the data involves additional costs and how you plan to cover them.

3.C b) Real example

Ex. 1 KAR is managed and supported by a team of experts and is free of charge.

Ex. 2 The cost of preserving the database will be assumed by the CNR.

Ex. 3

(1) A dedicated hard disk drive will probably be allocated for the dataset. No costs are currently foreseen regarding its preservation.

(2) The cost will be covered at the local hosting institute in the context of the project.

(3) The cost will be covered at the local hosting institute as a part of the standard network system maintenance.

4. Data security

4.A Address data recovery as well as secure storage and transfer of sensitive data

4.A a) Description

State what provisions are in place for data security. Includes storage & backup, long term preservation and how to keep research data safe and secure.

4.A b) Real example

Ex. 1

(1) An alert system is implemented to ensure warning messages if there are problems during file transfer from the data originators to the data centre

(2) Due to the data volume, most sites also hold a copy of their own processed data, effectively acting as a second distributed database and additional backup.

Ex. 2 In WP 2 it is planned to develop an observatory for urban logistics, and this will be one mechanism for sharing data. The observatory will be connected to the web site hosted by the University.

Ex. 3

(1) Two dedicated hard disk drives will probably be allocated for the dataset: one dedicated to the public part and one to the private part.

(2) The digital signature of the whole dataset, or the storage of the dataset in a git repository could provide support for the correct duplication and preservation.

5. Ethical aspects

5.A To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if no covered by the former

5.A a) Description

Description of potential ethical issues of collecting, storing, processing and archiving data. Description of procedures of ethical approval related to the project It is important to remark here any point that was mentioned in the Article 34 of the grant Agreement "Article 34 — Ethics and research integrity".

If your research activities involve children, patients, vulnerable populations, the use of human embryonic stem cells, privacy and data protection issues, or research on animals and non-human primates, you must comply with ethical principles and relevant national, EU and international legislation, and here you have to state how your are going to deal with it.

5.A b) Real example

Ex. 1 The ethical aspects related to the personal data collected in this dataset are addressed in the Ethics Requirements version 1.0 document of the project.

Ex. 2 Regarding the protection of personal data of the research participants, the Consortium will meet the following conditions:

- To submit to the REA the copies of ethical approvals for the collection of personal data by each of the competent University Data Protection Officers or National Data Protection authorities.

- To justify (if necessary) the collection and/or processing of personal sensitive data.

- To follow and accomplish the national and EU legislation on the procedures that will be implemented for data collection, storage, protection, retention and destruction.

6. Other issues

6.A Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

6.A a) Description

Explain the national/funder/sectorial/departmental procedures for data management that you are using.

6.A b) Real example

Ex. 1 Through the use of the institutional repository, we are also following these procedures for data management:

Ex. 2 National level: Law 14/2011 of June 1st, on Science, Technology and Innovation (Article 37 Dissemination in open access)

Ex. 3 The project documentation is also stored in the project SharePoint, which is accessible to all project partners.

7. Further support in developing your DMP

7.A Further support in developing your DMP

6.A a) Description

Explain further support in developing your DMP

6.A b) Real example

Ex. 1 This DMP has been created with the tool "Pla de Gestió de Dades de Recerca" (www.dmp.csuc.cat).