Consorci de Serveis Universitaris de Catalunya

Data Management Plans

(Version 1, July 2016)

(Doc. 16/36) (B6SR\GT Suport Recerca\PlansdeGestiodeDades_versio1Publica_juliol16-EN.pdf, 28.07.16)

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.

Key to the numbering:

- A capital letter indicates the fields that are required in Horizon 2020.
- A number 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/2072/266523.

[1] Actris (Grant 654109), Citilab (Grant 635898), ConnectingGEO (Grant 641538), EGI-Engage (Grant 654142), FREME (Grant 644771), iCirrus (Grant 644526), MMT (Grant 645487), RAMCIP (Grant 643433), SatisFactory (Grant 636302), Step (Grant 649493), Tandem (Grant 654206), UMobile (Grant 645124), U-Turn (Grant 635773)

A. Project data

A.1 Project identifier

A.1a) Description

Grant number

Acronym

A.2 Project coordinator

A.2a) Description

Name and surname of project coordinator

Institution

A.3 Contact details of the coordinator

A.3a) Description

E-mail

A.4 Author/s of the Data Management Plan

A.4a) Description

Name and surname of the author/s

Institution

A.5 Contact details of the author/s of the Data Management Plan

A.5a) Description

E-mail

B. Data set reference and name

B.1 Data set reference and name

B.1a) Description

Name and identifier of the dataset produced.

B.1b) Real example

Ex. 1 Object recognition data set - DS1.CRT.01

Ex. 2 The project dataset identification follows the naming: Data_<WPno>_<serial number of dataset>_<dataset title>. Example: Data_WP2_1_User generated content.

Ex. 3 Project following the alliance principles described in DMP model will employ a standard identification mechanism for each data set. A PID (Persistant Identifier) or the ISLRN (International Standard Language Resource Number) will be used to identify the dataset.

C. Description of the dataset

C.1 Description of the dataset

C.1a) Description

A short introductory text explaining the content of the dataset.

C.1b) Real example

Ex. 1 This dataset contains the posts and the contact details of all subscribers to the project collaborative platform.

Ex. 2 MMT is a translation memory obtained by finding and aligning at the sentence level parallel documents from the web.

Ex. 3 Dataset for incident detection, along with high-level activities and business processes monitoring (e.g. activities occurring at the shop-floor, etc.), obtained with thermal and depth cameras mounted at specific locations on the shop-floor. [...]

C.2 The group who may be interested in the dataset

C.2a) Description

State the group who may be interested in the data.

C.2b) Real example

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

C.3 Origin of the data produced

C.3a) Description

If the data are generated within the project, state the source of the data.

C.3b) Real example

Ex. 1 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

(1) Dataset1: Data collected through literature review, and questionnaire survey among project pilot partners (countries).

(2) Dataset2: These data originate though qualitative and standardized open-ended interviews with 28 Young European Citizens and 12 Public Authorities/Policy Makers. Datasets generated for WP2 comply with the University Research Data Policy (https://intranet...)

C.4 Origin of the data collected

C.4a) Description

If the data are collected, state the source from which they were extracted. For example: Thematic repositories, external research groups, etc.

C.4b) Real example

Ex. 1 Linked Open Data offers the unique chance of accessing vast amounts of machine-readable, semantically annotated data. However, access is still limited by additional knowledge required for data discovery. Data consumers have to know where datasets of interest are located, what kind of data they contain, where to access them, in what formats, and the terms of reuse. To date, some parts of these important metadata can be found in various repositories: datahub.io is the most accepted one, but various domain-specific repositories exist.

C.5 Types of research data (observational, experimental, computational data, etc.)

C.5a) Description

Description of the content and scope of the data. Indicate the format of the data (text, numeric, image, etc.). 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.

C.5b) Real example

Ex. 1 A range of experimental, simulation and theoretical data will be collected on excel spreadsheets for easy accessibility. In cases where the experiments or simulations generate large volumes of data, only data selected as useful for analysis will be archived. No existing data will be used: all data will be generated during the project.

D. Standards and metadata

D.1 What rules or methodologies are used?

D.1a) Description

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

D.1b) Real example

Ex. 1 Generic metadata standard: Dublin core http://dublincore.org (used by the X repository).

Ex. 2 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.

D.2 How are the folders and files named and structured?

D.2a) Description

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

D.2b) 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.

D.3 How are the different versions of a dataset easily identified?

D.3a) Description

Describe how version control will be organized.

D.3b) Real example

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

D.4 How are the metadata captured/created?

D.4a) Description

State how the metadata are generated (manually or automatically).

D.4b) Real example

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

D.5 What metadata standards will be used and why?

D.5a) Description

State the metadata standards that will be used. We recommend using metadata standards that are specific to the discipline. Consult metadata standards: http://www.dcc.ac.uk/resources/metadata-standards. If metadata standards are not used, state what metadata will be created and how.

D.5b) Real example

Ex. 1

(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. 2 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. 3 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.

E. Shared data

E.1 How and when will the data be available to others (within the group, other groups, the public)?

E.1a) 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.

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

E.1b) 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.

E.2 How will reuse be allowed?

E.2a) 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:

- Creative Commons (6 standard licenses + CC0)
  - http://creativecommons.org/licenses
- Open Data Commons
  - http://opendatacommons.org/
  - Public Domain Dedication and License (PDDL)
  - Attribution License (ODC-By)
  - Database License (ODC-ODbL)

E.2b) Real example

Ex. 1 Data sharing:

License: CC-BY-SA 3.0

ODRL license description: http://purl.org/NET/rdflicense/cc-by-sa3.0de

Openness: DBpedia is an open dataset, licensed under CC-BY-SA 3.0.

Ex. 2 Data sharing:

License: The data set is copyright-protected.

Openness: The data set is not openly available, since it is intellectual property of the company.

E.3 Do the data require any restrictions? If part or all of the data cannot be open access, please state the reason

E.3a) 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.

E.3b) 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.

E.4 In what data repository will the data be deposited?

E.4a) Description

State the repository in which the data will be stored and whether it is institutional or thematic. You can check the repositories available for a particular discipline (the list of thematic repositories in Figshare: https://figshare.com/articles/Scientific_Data_recommended_repositories_June_2015/1434640).

We recommend using a permanent link (DOI, handle) to the data in the repository so that they can be correctly cited, for example, in a publication.

E.4b) Real example

Ex. 1 Methods for Data Sharing. Raw data or resulted data that are governed by any IPRs or confidentiality issues will be added to a data enclave. Data enclaves are considered controlled, secure environments for datasets that cannot be distributed to the general public due to participant confidentiality concerns, third-party licensing, or use agreements that prohibit redistribution.

An additional raw-data collection issue is the provision of data required during the pilots of the project, such as basic data required for a use-case. This kind of data will be inserted to the project platform either manually by the user, or in batches using the defined system interfaces. Either way, the confidentiality and integrity of these data will be guaranteed by the security encryption scheme that

will be defined in the respective deliverable regarding the non-functional requirements of the platform.

On the other hand, data that are eligible for public distribution may be disseminated through the following channels:

- Scientific papers
- Lectureships in case of universities
- Interest groups created by the partners of the project
- Dissemination though the dissemination and exploitation channels of the project to attract more interested parties

Appropriate repositories such as OpenAIRE will be used for storing the results of the project and providing access to the scientific community.

Ex. 2 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. 3 Data will be shared via a repository held and managed by the lead participant, the University (http://www.example.edu).

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: http://zenodo.org/.
- Other data repositories (searchable here: http://www.re3data.org), if the previous ones are ineligible.

F. Archiving and preservation

F.1 What is the long-term conservation plan for the dataset? For example: deposit in a data repository

F.1a) Description

Specify the level of preservation of the research data by indicating the institution that will archive and preserve the project, whether this will be done during the project or at the end, and the time during which the data will be preserved.

F.1b) Real example

Ex. 1 Data will be stored at the coordinator's repository (www.example.edu), KAR, and will be kept for 5 years after the end of the project. Where requested, data will be kept for 2 more years.

F.2 Will additional resources (software, hardware, storage, etc.) be needed?

F.2a) Description

State whether additional resources are required to prepare the data for depositing and, if so, what resources.

F.2b) 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

F.3 Will additional expertise be required?

F.3a) Description

List the expertise required by staff to archive and preserve the data.

F.3b) Real example

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

Ex. 2 Where dedicated resources are needed, these should be outlined and justified, including any relevant technical expertise, support and training that is likely to be required and how it will be acquired.

F.4 What storage space is necessary?

F.4a) Description

State the approximate volume of the datasets and, if applicable, the type of data.

F.4b) Real example

Ex. 1 The dataset is expected to be several gigabytes.

Ex. 2 The volume of data is estimated to be about 10 Gb for all pilots.

Ex. 3 Videos and pictures – 8 GB; informed consents – 18 pages; questionnaires – 789 pages

F.5 Are additional costs foreseen to ensure archiving and preservation?

F.5a) Description

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

F.5b) 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.