Consorci de Serveis Universitaris de Catalunya

Plans de Gestió de Dades

Versió 1, Juliol 2016

(Doc. 16/31) (B6SR\GT Suport Recerca\PlansdeGestiodeDades_versio1Publica_juliol16.pdf, 13.07.16)

Aquest document vol servir de recolzament als investigadors a l'hora de crear els seus Plans de Gestió de Dades; concretament, per als projectes finançats en el marc de l'Horitzó 2020 de la Unió Europea.

A continuació es mostra:

- Amb una lletra majúscula, els diferents camps que es requereixen a Horitzó 2020
- Amb un número, els elements que caldria tenir en compte a l'hora d'omplir cada camp
- Amb una lletra minúscula, les descripcions de cada element i una mostra d'exemples reals

Aquest document ha estat elaborat pel Grup de Treball de Suport a la Recerca del CSUC que està format per representants de les universitats següents: Universitat de Barcelona, Universitat Autònoma de Barcelona, Universitat Politècnica de Catalunya, Universitat Pompeu Fabra, Universitat de Girona, Universitat de Lleida, Universitat Rovira i Virgili, Universitat Oberta de Catalunya, Universitat de Vic-Universitat Central de Catalunya, Universitat Ramon Llull i Universitat Jaume I.

Els exemples citats són una mostra[1] de diferents Plans de Gestió de Dades disponibles a la xarxa.

Versió digital: http://hdl.handle.net/2072/266522.

[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. Dades del projecte

A.1 Identificador del projecte

A.1a) Descripció

Grant number

Acrònim

A.2 Coordinador del projecte

A.2a) Descripció

Nom i cognoms del coordinador del projecte

Institució

A.3 Dades de contacte del coordinador

A.3a) Descripció

Correu electrònic

A.4 Autor/s del Data Management Plan

A.4a) Descripció

Nom i cognoms

Institució

A.5 Dades de contacte dels autor/s del Data Management Plan

A.5a) Descripció

Correu electrònic

B. Referència del conjunt de dades (dataset)

B.1 Referència del conjunt de dades

B.1a) Descripció

Nom i Identificador del conjunt de dades (dataset) que es produeix.

B.1b) Exemple real

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. Descripció del conjunt de dades (dataset)

C.1 Descripció del dataset

C.1a) Descripció

Text curt introductori que explica el contingut del dataset

C.1b) Exemple real

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 in the shop-floor. [...]

C.2 Col·lectiu a qui pot ser d'interès el conjunt de dades (dataset)

C.2a) Descripció

Indicar per a quin col·lectiu poden ser d'interès les dades.

C.2b) Exemple real

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

C.3 Origen de les dades produïdes

C.3a) Descripció

En el cas que les dades es generin dins el projecte, indicar l'origen de les dades.

C.3b) Exemple real

Ex. 1 Dataset produced by simulation tools and/or by real life trials will be used as means to quantify the performance advantages the project architecture offers compared to current practices

Ex. 2

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

(2) Dataset2: This data originates 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 Origen de les dades recol·lectades

C.4a) Descripció

En el cas que les dades siguin recol·lectades indicar la font d'on s'han extret. Per exemple: repositoris temàtics, grups de recerca externs, etc.

C.4b) Exemple real

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 which formats, as well as the terms of reuse. To date, some parts of this important metadata can be found in various repositories, datahub.io being the most accepted one, although various domain-specific repositories exist

C.5 Tipologia de les dades de recerca (dades observacionals, experimentals, computacionals...)

C.5a) Descripció

Descripció del contingut i abast de les dades. Indicar el format de les dades (text, numèric, imatge, etc.). Les dades de recerca es generen per diferents motius o processos i poden ser dades de tipus:

- Observacionals: dades capturades en temps real (neuroimatges, dades de mostres, dades de sensors, dades d'enquesta, etc.)
- Experimentals: dades capturades en equips de laboratori (seqüències de gens, cromatogrames, dades de camps magnètics, etc.).
- Simulació: dades generades a partir de models de prova (climatològiques, matemàtiques o models econòmics, etc.)
- Derivades o compilades: dades reproduïble però de difícil reproducció (text i mineria de dades, models 3D, bases de dades compilada , etc.)
- De referència: conglomerat o conjunt de dades (bancs de dades de seqüències de gens, estructures químiques, portals de dades espacials, etc.)
- Altres

C.5b) Exemple real

Ex. 1 There will be a range of experimental, simulation and theoretical data 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. Estàndards i metadades

D.1 Quines normes o metodologies s'utilitzen?

D.1a) Descripció

Referència a estàndards de metadades de la disciplina. Si no existeixen, descripció de quines metadades es crearan i com.

D.1b) Exemple real

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 Com s'anomenen i s'estructuren les carpetes i els arxius?

D.2a) Descripció

Descriure com s'organitzaran les dades: estructura I nom dels fitxers

D.2b) Exemple real

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 Com s'identifiquen fàcilment les diferents versions d'un conjunt de dades (dataset)?

D.3a) Descripció

Descriure com es gestionarà el control de versions

D.3b) Exemple real

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

D.4 Com es capturen / creen les metadades?

D.4a) Descripció

Indicar com es generen les metadades (manual, automàticament..)

D.4b) Exemple real

Ex. 1 Metadata creation is carried out manually by depositors in the deposit form at the repository

D.5 Quins estàndards de metadades utilitzarà i per què?

D.5a) Descripció

Indicar els estàndards de metadades que s'utilitzarà. Es recomana utilitzar estàndards de metadades específics de la disciplina. Es pot consultar: http://www.dcc.ac.uk/resources/metadata-standards. Si no s'utilitzen estàndards de metadades, indicar com i quines metadades es crearan.

D.5b) Exemple real

Ex. 1

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

(2) There are no standards for these logs. A possible solution is project servers 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 their 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. Dades compartides

E.1 Com i quan estaran disponibles les dades per als altres (dins el grup, resta de grups, públic)?

E.1a) Descripció

Cal descriure com es compartiran les dades, és a dir qui tindrà accés al conjunt de dades (dataset). Es pot crear un procediment per tal que temporalment les dades es facin accessibles a la resta de membres del grup, dels socis del projecte, i al públic en general. Cal indicar si les dades es posaran en accés obert i en quin termini raonable. Una possibilitat és oferir-les conjuntament amb les publicacions. Si cal establir períodes d'embargament, és aquí on cal especificar-los.

També cal incloure els requeriments tècnics per accedir i reutilitzar les dades, si són necessaris. Per exemple, si cal un programari específic.

E.1b) Exemple real

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 would be released in waves as they become available or 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 associated paper and will be made accessible on request and under conditions agreed on a case-by-case basis, and after agreement of the project consortium.

E.2 Com es permetrà la reutilització?

E.2a) Descripció

En el cas que les dades es posin a l'abast d'altres investigadors i el públic en general cal especificar quin grau de reutilització es permetrà. Aquest grau de reutilització vindrà marcat per l'establiment de llicències d'ús. La CE proposa l'ús de les llicències CC BY o CC0 de Creative Commons, però n'hi ha d'altres.

- Creative Commons (6 llicències estàndard + 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) Exemple real

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 company

E.3 Les dades requereixen alguna restricció? Si parcialment o totalment les dades no es poden posar en obert indiqueu-ne el motiu

E.3a) Descripció

En principi les dades s'han de posar a disposició d'altres investigadors i el públic en general amb el menor nombre possible de restriccions. Tanmateix poden existir diversos motius que impedeixin compartir-les: motius ètics, protecció de dades de caràcter personal, implicació de drets de propietat intel·lectual i/o industrial, interessos comercials, etc. Cal especificar les raons per les quals no es compartirà un conjunt de dades (dataset).

E.3b) Exemple real

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 external to 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. Concerning the results of the project, these 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 on it. Furthermore, if specific portions of it (e.g. metadata, statistics, etc.) are decided to become of 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 that most of the data generated can 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 En quin repositori seran dipositades les dades?

E.4a) Descripció

Cal indicar el repositori on s'emmagatzemaran les dades i indicar els tipus de repositori (institucional o temàtic). Es pot consultar quins repositoris hi ha per a una determinada disciplina (llistat de repositoris temàtics a Figshare: https://figshare.com/articles/Scientific_Data_recommended_repositories_June_2015/1434640).

És recomanable utilitzar un enllaç permanent (DOI, handle) a les dades en el repositori per tal que es puguin citar correctament, per exemple en una publicació.

E.4b) Exemple real

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 either due to participant confidentiality concerns or 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:

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

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 into a 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. Arxiu i preservació

F.1 Quin és el pla de conservació a llarg termini per al conjunt de dades (dataset)? Per exemple: dipòsit en un repositori de dades

F.1a) Descripció

Especifiqueu el pla de preservació de les dades de recerca indicant la institució que assumirà l'arxiu i preservació del projecte, si es farà durant o al final d'aquest i el període de temps que es preservaran les dades.

F.1b) Exemple real

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 Es necessitaran recursos addicionals (programari, maquinari, emmagatzematge, etc.)?

F.2a) Descripció

Indiqueu si són necessaris recursos addicionals per preparar les dades per al dipòsit i en cas afirmatiu, quins.

F.2b) Exemple real

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 University

Ex. 3

(1) Probably two dedicated hard disk drives will be allocated for the dataset; one dedicated to the public part and one to the private.

(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 Es requereixen coneixements especialitzats addicionals?

F.3a) Descripció

Detalleu els coneixements específics que requereix el personal per dur a terme les tasques d'arxiu i preservació.

F.3b) Exemple real

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 Quin és l'espai d'emmagatzematge necessari?

F.4a) Descripció

Indiqueu el volum aproximat dels diferents datasets especificant, si s'escau, la tipologia de les dades.

F.4b) Exemple real

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 Es preveuen costos addicionals per garantir l'arxiu i la preservació?

F.5a) Descripció

Indiqueu si preveieu costos associats per l'arxiu i la preservació i com teniu previst cobrir-los

F.5b) Exemple real

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

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

Ex. 3

(1) Probably a dedicated hard disk drive will 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 project.
(3) The cost will be covered at the local hosting institute as a part of the standard network system maintenance.