Abstract

This report summarises the work of WP1 on the Data Management considerations and initial plan for the DEEP-HybridDataCloud project. This document describes the types of data that will be generated or collected during the project, the standards that will be used and the ways in which the data may be exploited and shared including the data security and ethical aspects.
Copyright Notice
Copyright © Members of the DEEP-HybridDataCloud Collaboration, 2017-2020.

Delivery Slip

<table>
<thead>
<tr>
<th>From</th>
<th>Name</th>
<th>Partner/Activity</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lara Lloret Iglesias</td>
<td>CSIC/WP1</td>
<td>28/04/2018</td>
</tr>
<tr>
<td>Reviewed by</td>
<td>Fernando Aguilar</td>
<td>CSIC</td>
<td>26/04/2018</td>
</tr>
<tr>
<td></td>
<td>Alessandro Costantini</td>
<td>INFN</td>
<td></td>
</tr>
<tr>
<td>Approved by</td>
<td>Steering Committee</td>
<td></td>
<td>30/04/2018</td>
</tr>
</tbody>
</table>

Document Log

<table>
<thead>
<tr>
<th>Issue</th>
<th>Date</th>
<th>Comment</th>
<th>Author/Partner</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1.0</td>
<td>8/02/2018</td>
<td>First template version</td>
<td>Álvaro López / CSIC</td>
</tr>
<tr>
<td>V2.0</td>
<td>15/03/2018</td>
<td>TOC</td>
<td>Lara Lloret / CSIC</td>
</tr>
<tr>
<td>V3.0</td>
<td>18/03/2018</td>
<td>First version</td>
<td>Lara Lloret / CSIC</td>
</tr>
<tr>
<td>V4.0</td>
<td>25/04/2018</td>
<td>Update on review</td>
<td>Lara Lloret / CSIC</td>
</tr>
<tr>
<td>V5.0</td>
<td>26/04/2018</td>
<td>Update on review</td>
<td>Lara Lloret / CSIC</td>
</tr>
<tr>
<td>V6.0</td>
<td>27/04/2018</td>
<td>Use Cases Update</td>
<td>Werner Dubitzky / HMGU</td>
</tr>
<tr>
<td>V7.0</td>
<td>28/04/2018</td>
<td>Improved Executive Summary</td>
<td>Álvaro López / CSIC</td>
</tr>
<tr>
<td>V8.0</td>
<td>28/04/2018</td>
<td>Final Version</td>
<td>Lara Lloret / CSIC</td>
</tr>
</tbody>
</table>
# Table of Contents

Executive summary............................................................................................................................4  
1. Introduction....................................................................................................................................5  
2. Use Cases Data description............................................................................................................5  
  2.1. Deep Learning applications for biological image data from the back of the human eye (HMGU)........................................................................................................................................6  
  2.2. Plant classification with Deep Learning techniques (CSIC)......................................................6  
  2.3. Deep Learning application for monitoring through satellite imagery (CSIC).........................6  
  2.4. Massive Online Data Streams (CESNET, CSIC, IISAS)............................................................7  
3. FAIR data........................................................................................................................................7  
  3.1. Making data findable, including provisions for metadata:..........................................................7  
  3.2. Making data openly accessible:..................................................................................................8  
  3.3. Making data interoperable:.........................................................................................................8  
  3.4. Increase data re-use (through clarifying licenses):......................................................................8  
4. Allocation of resources.....................................................................................................................8  
5. Data security.....................................................................................................................................9  
6. Ethical aspects...................................................................................................................................9
Executive summary

This document contains the Data Management Plan (DMP) for the DEEP-HybridDataCloud project (DEEP from now on). This plan is included within the context of the WP1/NA1 and describes the types of data that will be generated or collected during the project, the standards that will be used, and the ways in which the data may be exploited and shared including the data security and ethical aspects. An online Data Management Planning tool (DMPonline: https://dmponline.dcc.ac.uk) has been used to elaborate this deliverable. This tool provides a template with the guidelines to assure a complete description of the project DMP following the FAIR principles, i.e. making data Findable, Accessible, Interoperable, and Reusable.

The DMP will be updated during the project lifetime to cope with the data management requirements and related issues that may arise, in particular with the support of new Use Cases.
1. Introduction

DEEP project is aimed at developing and promoting a new generation of e-infrastructures tackling the latest generation technologies, supporting Deep Learning and other intensive computing techniques to exploit very large data sources. This means that the main aim of the project is not generating new data but providing for new communities and users the corresponding services to lower the current adoption barriers for exploiting their data. The DMP will be then based on the data collected/generated by the different Use Cases. To this purpose, DMPonline has been used as Data Management planning tool in order to create, review, and share this document meeting standard requirements.

Although DEEP project is not a data oriented project, some Use Cases will operate with their own data and generate new one. The decisions concerning such new data, in particular the different aspects of data management (publication, type of license, etc.) are in charge of each Community. Even so, the DEEP project will encourage the adoption, among the Research Communities, of Open Repositories and Open Licenses whenever possible.

The project includes five different Use Cases that will manage data at different levels. The Use Cases are detailed in the deliverable D2.1, and are the following:

- Deep Learning applications for biological image classification (HMGU)
- Plant classification with Deep Learning techniques (CSIC)
- Deep Learning application for monitoring through satellite imagery (CSIC)
- Massive Online Data Streams (CESNET, CSIC, IISAS)
- Post-Processing of large amounts of data (CSIC)

The rest of the document is structured as follows

- Section 2 contains a description of the selected Use Cases and the description of the data they manage.
- Section 3 covers the different parts of the FAIR principles and how they are going to be filled within the DEEP project.
- Section 4 describes the allocation of resources used to assure the proper data management.
- Section 5 and 6 will describe data security and ethical aspects included in the DMP.

2. Use Cases Data description

The data managed within the context of the project is related to the following project objective:

**Objective 1.** Focus the interest of intensive computing techniques on the analysis of very large datasets, considering demanding cases from different research communities, in the context of the future generation of e-infrastructure.
In respect to the above Objective, the types, formats, sizes and origins of these large datasets will depend on each Use Case.

2.1. Deep Learning applications for biological image data from the back of the human eye (HMGU)

This use case employs Deep Learning to classify the stage and reconstruct progression of retinopathy based on labelled image data taken from the back of the human eye.

The dataset used for this use case consists of ca. 35,000 labelled and 48,000 unlabelled colour fundus photographies (JPEG images) of the human retina. Together, the entire data set is 80 GB in size. The each label describes a disease stage or score from 0 (Healthy) to 1 (Severe).

The retinopathy used in this use case comes from the Kaggle repository (https://www.kaggle.com/c/diabetic-retinopathy-detection). For training the application those data should partially be copied to a local server allowing fast access from the GPU.

2.2. Plant classification with Deep Learning techniques (CSIC)

This Use Case aims at using near state-of-the-art convolutional network architectures techniques for plant classification from a citizen science perspective oriented to large-scale biodiversity monitoring.

The collected data will be images in standard graphic formats (JPG or PNG). Along with the images the metadata (author, data, original URL, species, ID, etc...) will also be saved in JSON format. The size of the training dataset will be of the order of 10 TB.

The biodiversity data used for the training are collected from portals like iNaturalist (https://www.inaturalist.org/) or Natusfera (http://natusfera.gbif.es/), that enable users to easily upload their observation photos. Additionally one could merge data from other (minor) sources like PlantNet (https://identify.plantnet-project.org/) to extend the species database. For training the application those data should partially be copied to a local server allowing fast access from the GPU.

2.3. Deep Learning application for monitoring through satellite imagery (CSIC)

The goal of this Use Case is to use the latest techniques in Machine Learning/Deep Learning (ML/DL from now on) to perform pattern recognition in satellite images combined with other in-situ measurements.

For this Use Case the collected data are divided in two types:

- **Satellite data**: Data can usually be exported as images (TIFF). During the preprocessing these images are transformed into CSV or NetCDF4-HDF5 formats.
- **Environmental measurements**: Used to validate the predictions made using satellite observations. They usually come in CSV format.
The data capacity will be in the order of TBs. Each satellite image file corresponding to an area of 10000 km² weights about 1 GB.

The satellite data is retrieved from ESA (https://scihub.copernicus.eu/dhus) or NASA (https://earthexplorer.usgs.gov). For training the application those data should partially be copied to a local server allowing fast access from the GPU.

### 2.4. Massive Online Data Streams (CESNET, CSIC, IISAS)

The collected datasets are monitoring log files that will be used to perform misuse detection in real-time with external intelligence modules built based on the heuristic analysis techniques using ML/DL techniques. The files formats are provided by monitoring tools listed in the Use Case description in D2.1. The incoming raw data (stream) is required to be cleaned, reduced and transformed online in on-the-fly manner. The datasets size will be of the order of 1-10TB.

The collected data (logs, intrusion detection system alerts, etc.) can be divided in:

- **Logs**: logs data produced by services running on servers.
- **Metrics**: periodical measurement of system loads/usages (CPU, IO, disk, mem, processes ..) of servers.
- **Network**: network monitoring.

These data will be retrieved as an online data stream from the data centre under study.

**Post-Processing of large amounts of data (CSIC)**

This Use Case consists in running the Lattice QCD analysis software over all the configurations produced during the Monte Carlo history, and which are stored on a HPC filesystem. The output of the analysis is typically rather small. Encapsulating the analysis software (standard tools for statistical analysis) in a container docker-style is a trivial task. Providing the way for such container to be run, seamlessly, and in such a way that it has access to those data stored in the HPC system, would be of enormous help for this community.

The generated configurations are binary files containing the description of the status of the fields at the checkpoints of the simulation. No input data is needed.

These are typically of the order of a few hundred files, with a size that highly depends on the lattice size and the problem under investigation (from a few hundred of Megabytes, up to one Terabyte per configuration in very large lattices).

### 3. FAIR data

#### 3.1. Making data findable, including provisions for metadata:

The publication of data is out of the scope of the project. Each Use Case will manage their data according to their rules. In case some of the Use Cases would like to publish their data, DEEP will
encourage its publication fulfilling all the FAIR criteria in an Open Repository such as Digital CSIC. This repository assigns a persistent identifier (PID) allowing to identify univocally the digital resource.

3.2. Making data openly accessible:

Some of the derived data produced within the Use Cases could be published in the DEEP testbeds environments or externally. As already mentioned this will depend on each community. Just as an example some possible publication scenarios, including the kind of data and its interest for external communities, are given here for the different Use Cases:

- Deep Learning for retinopathy detection, plant classification and satellite imagery monitoring: the trained neural networks could be openly published to be used out-of-the-box or as the base architecture for a further fine tuning for a particular problem with similar input types.
- Massive Online Data Streams: Log data could be anonymized and openly published to serve as training data for other communities.
- Post-Processing of large amounts of data: The generated data are private since the Use Case is within the context of a collaboration.

The software documentation will be published as defined in the Deliverable D3.1 concerning the initial plan and definition of the software life cycle management process and procedures.

3.3. Making data interoperable:

The Communities are responsible of the data within their Use Case. In case Communities decide to publish the data, they will be encouraged to use Open Repositories such as Digital CSIC which is OAI-PMH compliant and supports basic metadata standards as Dublin Cor

3.4. Increase data re-use (through clarifying licenses):

DEEP does not assure that the data stored in the pilot testbeds will remain available after the end of the project. Methods for data quality assurance depends also in the Use Cases. The different communities will be encouraged to store their generated digital resources (data, neural network weights, etc...) in an Open Repository where the FAIR principles are applied and its re-use is promoted. The type of license for the generated data will be decided by each Use Case. The collected data stored in the project testbeds will keep the data origin licenses.

4. Allocation of resources

The costs of making the generated data under the project context FAIR will be covered by the project itself. The Research Communities participating in the project are in charge to ensure the data management features during the entire project lifetime. It is their responsibility to preserve such features after the end of the DEEP project.
5. Data security

Security of data will be defined by each involved community and it will be strictly related to the proper Use Case. For such reason, the Data Management Plan will be updated in case of need to reflect any data security issues that may arise.

6. Ethical aspects

Ethical aspects and related policy will be defined and described, if needed, by the Research Community in charge for the Use Case. Within the project personal data collecting or processing is not foreseen. In case it should be needed the project will adhere to the law as laid down in the European Directive 95/46/EEC as well as the relevant national laws and regulations, including the General Data Protection Regulation (GDPR) (EU) regulation 2016/679.

The DEEP-HybridDataCloud project is not producing any new data, but will post-process already collected data from existing registries. The project will use externally generated scientific data and metadata to test the developed products against the different Use Cases provided. This data will be used respecting the policies requested by the data controllers and will not be stored permanently in production e-infrastructures. The DEEP-HybridDataCloud project will coordinate with a data protection officer linked with Task 1.2.4 (under WP1) and Work package 7.

Within the DEEP-HybridDataCloud project, ethical aspects are covered by Work package 7, whose deliverable WP7.1 - “POPD – Requirement No. 1” will include the following aspects:

- Justification will be given in case of collection and/or processing of personal sensitive data.
- Detailed information will be provided on the procedures that will be implemented for data collection, storage, protection, retention and destruction and confirmation that they comply with national and EU legislation.
- The DEEP-HybridDataCloud project will explicitly confirm that the data used are publicly available.
- In case of data not publicly available, The DEEP-HybridDataCloud project will confirm that relevant authorisations have been obtained, and are kept on file.

Ethical aspects and related policies will be continuously monitored and evaluated for existing and new Use Cases and this DMP will be updated accordingly.