D1.6 Data Management Plan

Lead Partner:	EGI Foundation
Version:	1
Status:	Rejected by EC
Dissemination Level:	Public
Document Link:	https://documents.egi.eu/document/3497

Deliverable Abstract
A report that specifies how research data will be collected, processed, monitored and catalogued during the project lifetime. For each dataset, it describes the type of data and their origin, the related metadata standards, the approach to sharing and target groups, and the approach to archival and preservation.

EOSC-HUB RECEIVES FUNDING FROM THE EUROPEAN UNION'S HORIZON 2020 RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO. 777536.

COPYRIGHT NOTICE

DELIVERY SLIP

	Name	***Partner/Activity***	***Date***
From:	Małgorzata Krakowian	EGI Foundation/WP1	
Moderated by:	Malgorzata Krakowian		
Reviewed by	Catalin Condurache Debora Testi	EGI Foundation CINECA/WP7	4/09/2019 5/09/2019
Approved by:	AMB		

DOCUMENT LOG

Issue	***Date***	***Comment***	***Author***
v.0.1	3/09/2019	First version based on input from WP8	M. Krakowian
FINAL	5/09/2019	Final version	M. Krakowian

Contents

Executive summary

This document defines data management plan for research data generated or collected by the Competence Centres (CC) in WP8. The document provides details of each relating to type, origin and scale of data, standards and metadata, data sharing (target groups, impact and approach) and archive and preservation, according to the suggested template (see Annex 1 of the guideline document provided by the EC). All CCs have provided an update of data management plan.

1 Introduction

Research data is defined as information, in particular, facts or numbers, collected to be examined and considered and as a basis for reasoning, discussion, or calculation. In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings, and images[1]. The focus of the Open Research Data Pilot in Horizon 2020 is on research data that is available in digital form[2].

The Open Research Data Pilot applies to two types of data:

1) the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible;
2) other data (e.g. curated data not directly attributable to a publication, or raw data), including associated metadata.

The obligations arising from the Grant Agreement of the projects are (see article 29.3): Regarding the digital research data generated in the action ('data'), the beneficiaries must:

1) deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following: the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible; other data, including associated metadata, as specified and within the deadlines laid down in the 'data management plan';
2) provide information — via the repository — about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and — where possible — provide the tools and instruments themselves).

As an exception, the beneficiaries do not have to ensure open access to specific parts of their research data if the achievement of the action's main objective, as described in Annex 1, would be jeopardised by making those specific parts of the research data openly accessible. In this case, the data management plan must contain the reasons for not giving access.

This document describes the data management plan[3] for the research data after 1.5 year of EOSC-hub project. For each dataset, it describes the type of data and their origin, the related metadata standards, the approach to sharing and target groups, and the approach to archival and preservation.

This document is an updated version of D1.5 Data Management Plan (June 2018).

[1] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

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

[3] Data management plan: document detailing what data the project will generate, whether and how it will be exploited or made accessible for verification and re-use, and how it will be curated and preserved.

2 Datasets

Following section describes data management plans for each WP8 Competence Centre (CC). It provides details of each relating to type, origin and scale of data, standards and metadata, data sharing (target groups, impact and approach) and archive and preservation. All CCs have provided update on data management plan defined initially in D1.5.

2.1 Disaster Mitigation Competence Centre Plus (DMCC+)

Task	T8.8
Contact	Eric Yen, Simon Lin
Data description	
Types of data	1. Observation data for target hazard events: global model data; gridded data in GRIB format; satellite data; radar data; etc. 2. Geographical data of impact areas of target hazard events: topographical data, land use, bathymetry, etc. 3. Simulation results: 3D gridded data, images, video, visualization data
Origin of data	Observation data comes from agencies such as NCEP, NASA, ECMWF and local weather bureau or related government agencies.
Scale of data	Depend on temporal and spatial resolution. Weather simulation by WRF (per simulation): observation and static data scale per simulation is O(100MB); Simulation result is O(TB) per simulation. Tsunami and Storm Surge (per simulation): input data scale is O(10MB); simulation result is O(GB).
Standards and metadata	Data format: GRIB/GRIB2; NetCDF Metadata: no domain specific metadata standard is applied. Darwin Core metadata scheme is used. GRIB/GRIB2[4] NetCDF[5]
Data sharing	
Target groups	Communities of research, education, hazard risk estimation and analysis, disaster management, etc.
Scientific Impact	Simulation event is reproducible. Patterns and correlations in high resolution 3D gridded data could be explored. More accurate event simulation is achieved and could be used for case studies of the same type of disaster.
Approach to sharing	Web services, searchable data catalogue, APIs, etc. All data services and sharing are provided by the simulation portal (or science gateway) according to the workflow. The goal is to make the data

[4] https://www.nco.ncep.noaa.gov/pmb/docs/grib2/grib2_doc/

[5] https://www.unidata.ucar.edu/software/netcdf/docs/index.html

	open according to FAIR principles. For example, data of tsunami case studies are available through the iCOMCOT simulation portal[6]. Data of meteorological hazard case studies will be provisioned throughout the WRF portal (which is now under reconstruction). A new simulation portal for storm surge will be available in early 2020.
Archiving and preservation	ASGC is coordinating the archive and preservation of all the data in those portals. For the moment, all data have multiple copies (according to data policy defined by the scientific group) in various file systems under Ceph on disks. The next step is to make the replications distributed in multiple sites. At least one copy will be stored in the local centre of the data owner (or case study owner). Archive data will be verified annually to ensure the data integrity based on checksum at this moment.

2.2 EISCAT_3D

Task	T8.4
Contact	Ingemar Häggström, Carl-Fredrik Enell
Data description	
Types of data	Input data: gridded multi-variable data Generated output: gridded data or diverse output such as diagrams, scripts, tables User scripts: Python scripts, configuration files
Origin of data	Community-provided data sources: For example, gridded multi-variable climate data from the ESGF/CMIP data pool. Other community data sources will be integrated later. But in any case, these typically have existing curation mechanisms. User-provided data sources: Aside from community sources, the user may provide data out of B2DROP, B2SHARE, DataHub and their curation policies apply.
Scale of data	Input data may be in the order of multiple MB/GB, but is not provided by ECAS directly, but served via services connected to ECAS (ESGF/CMIP and other community data sources).
Standards and metadata	Data and metadata conform to CMIP community agreements (standardized variables, file-level metadata, and domain metadata).
Data sharing	
Target groups	Research and academic community, public and private sector

[6] https://icomcot.twgrid.org

	(restrictions may apply for some data concerning commercial use), wider public, training, citizen scientists
Scientific Impact	Facilitate large-scale, multi-model climate data analysis; enable users to provide derived data products and information diagrams; support training for the next generation of data users.
Approach to sharing	Output data may be shared via EOSC data services (B2DROP, B2SHARE, DataHub) per user's discretion.
Archiving and preservation	ECAS does not archive outputs directly but relies on the connected data sharing services to do so. Input data is subject to standing curation policies of ESGF/CMIP.

2.3 ELIXIR

Task	T8.1
Contact	Steven Newhouse, Susheel Varma
Data description: Types of data	Public reference datasets: Genes, Protein, metabolite expression, protein sequences, molecular structures, chemical biology, reactions, interactions and pathways, systems biology. Some container and workflow execution data will be generated but will be discarded after integration testing between ELIXIR and EOSC-Hub.
Data description: Origin of data	ELIXIR Data Platform & EMBL-EBI
Data description: Scale of data	Scale: Spatial - Molecules to Systems Biology; Temporal - µs to years; Storage: KB to PB
Standards and metadata	Technical Metadata (data object id, uri path, size, version, checksum, metadata links)
Data sharing: Target groups	Life scientists and cross-domain researchers
Data sharing: Scientific Impact	Facilitate collaborations by facilitating access to large reference datasets.
Data sharing: Approach to sharing	Public Datasets are freely accessible by anyone. Data caches in an institutional site may be restricted and will be managed by the institution.
Archiving and preservation	External data archives will last longer than the duration of the project and will be overseen by the Data Management Plan of the repositories holding the data (EMBl-EBI & ELIXIR Nodes).

2.4 EPOS-ORFEUS

Task	T8.5
Contact	Sara Ramezani, Luca Trani, Javier Quinteros (GFZ)
Data description	
Types of data	Continuous Seismic Waveforms (SW) will be staged onto the storage facilities of the CC. Seismic Sensors Descriptions (SD) and Quality Indicators (QI) might be exploited to produce User Generated Products (UGeP). Additionally logs and accounting information might be produced and collected for a limited time.
Origin of data	SW, SD and QI will be accessible from the ORFEUS European Integrated Data Archive (EIDA) and from 4 SW replica archives. UGeP, logs and accounting information will be produced by means of the services of the CC.
Scale of data	The primary data originated from EIDA are in the order of: 200+ Seismic Networks, 10K+ Seismic Stations, 400+TB Seismic Waveform Data continuously updated, 300+GB of metadata describing waveforms and quality indicators.
Standards and metadata	Community and international standards for data encodings and metadata. E.g. FDSN MSEED, FDSN StationXML, DOI and DataCite. Additional descriptions might be adopted by UGeP.
Data sharing	
Target groups	Seismology researchers and solid Earth-scientists. Currently EIDA servers ~1900 unique users p/y.
Scientific Impact	Improve access to data and compute facilities and provide robust and easy to use authentication and authorisation mechanisms. Improve visibility and usability of dataset and products beyond the current user community e.g. by targeting EPOS cross-disciplinary users.
Approach to sharing	Most of the data are publicly available through community standard interfaces and tools. Some datasets, e.g. embargoed experimental datasets might require authentication. The CC will provide additional methods to access data e.g. by means of staging services and by exploiting locality to compute resources.
Archiving and preservation	The CC will host SW replicated archives from 4 EIDA primary repositories. Users will be responsible of the preservation of their products generated in the CC. Long term archival might be offered for products of particular interest within EIDA and/or the CC's storage providers. Logs and accounting information might be maintained by the CC for a limited time.

2.5 Fusion

Task	T8.2
Contact	de Witt, Shaun
Data description	
Types of data	Primarily experimental from a number of diagnostics, some synthetic data is possible if real data cannot be released. This data will consist of calibrated science and engineering data with a number of different parameters contained in each file. In addition, some output of model runs is anticipated, but these will be test runs of no scientific value and will be discarded post testing.
Origin of data	Experimental data will have been produced and quality controlled by each tokamak site involved. The data will have come from number diagnostic sensors and have been converted from raw data to physically meaningful parameters, via an initial processing chain. In some cases this contains provenance information.
Scale of data	The data set from the MAST experiment currently consists of ~100TB over 30,000 files. Each object within the file covers several decades in scale both temporally and spatially.
Standards and metadata	Only local metadata and formats currently exist; there is currently no standard metadata or format for fusion data. A separate project, FAIR4Fusion [7]will develop a community metadata standard with ontology mapping to existing local standards.
Data sharing	
Target groups	Public, researchers from other domains
Scientific Impact	For existing researchers, easier access to data (in cases where the data is open). Primary impact is likely to come from cross disciplinary research and/or industrial sector. Use of fusion results in materials science is currently an important research area.
Approach to sharing	Currently licensing varies from site to site. MAST data is currently embargoed for a period of three years and then made accessible through a registration process. During the EOSC-Hub project, a parallel project is addressing making fusion data more accessible and this document will be updated based on the results of this initiative.
Archiving and preservation	Each site is responsible for archival and preservation of its own experimental and modelling data and local procedures exist for this, with local repositories existing at hosting sites.

[7] https://cordis.europa.eu/project/rcn/223667/factsheet/en

2.6 ICOS

Task	T8.7
Contact	Alex Vermeulen, Margareta Hellström
Data description	
Types of data	High frequency eddy covariance flux data for CO2 and other gases.
Origin of data	Measurements stations from the ICOS and LTER networks.
Scale of data	Up to 78 stations from ICOS and 20 from LTER provide half hourly updates of 20 Hz data.
Standards and metadata	Ecosystem stations use BADM metadata, raw data datafiles are binary or TSV formatted ACSII in a well described community standard. Output files are community standard TSV formatted ASCII files, metadata is ISO19115 and Inspire compliant.
Data sharing	
Target groups	Carbon science, climate science, remote sensing satellite data validation, vegetation models tuning and validation, crop model improvements, COPERNICUS, FLUXNET.
Scientific Impact	At least 4000 downloads per year, hundreds of articles and ten-thousands of citations per year.
Approach to sharing	CC4BY of all data levels, from raw to final QC'ed products. Published using Handle PIDs and DOIs. Metadata shared through DATACITE, B2FIND, GEOSS, DATACITE and other portals (of portals).
Archiving and preservation	B2SAFE trusted repository

2.7 Marine

Task	T8.3
Contact	Thierry Carval
Data description	
Types of data	Argo floats ocean data, SeaDataCloud marine data
Origin of data	The Argo floats data are collected, quality controlled and distributed by Argo data management team. The SeaDataCloud data are collected; quality controlled and distributed by the European network of national oceanographic data centres.
Scale of data	The Argo dataset is a collection of 2 million NetCDF files, of about 250GB. A not yet defined subset of SeaDataCloud will be provided.
Standards and	Argo data files comply with NetCDF CF1.6 convention (Climate and

metadata	Forecast). Data and metadata formats are published "Argo user's manual" [8] The parameters are compliant with SeaDataNet P01 (Parameter Usage Vocabulary) and P06 (data storage units) vocabularies. SeaDataCloud data files comply with SeaDataNet vocabularies. [9]
Data sharing	
Target groups	Scientists, operational services
Scientific Impact	Climate studies, seasonal forecasting, meteo-oceano activities
Approach to sharing	Argo data are publicly and immediately available in real-time and delayed mode (when available), with no user registration. SeaDataCloud data are distributed under a SeaDataNet licence.
Archiving and preservation	US-NCEI is in charge of the long term preservation of Argo data. SeaDataNet national data centres are in charge of the long term preservation of their national data.

2.8 Radio Astronomy Competence Center (RACC)

Task	T8.6
Contact	Hanno Holties, Rob van der Meer
Data description	
Types of data	The RACC will provide services for the Radio Astronomy community with a particular focus on the International LOFAR Telescope (ILT). Data types include observation data in raw formats (visibilities and time-series) as well as processed data (radio astronomical images, pulsar profiles, etc.).
Origin of data	Observation data is generated by the LOFAR instrument and the processing cluster managed by the ILT Observatory. It is stored in grid-enabled storage facilities that are part of the LOFAR Long Term Archive (LTA) and hosted by the LTA-partners SURFsara, FZJ, and PSNC. The community generates scientific data-products from the observation data on processing clusters that they have access to, either hosted by the LOFAR Observatory, the LTA partners, or anywhere else.
Scale of data	LOFAR observation data volumes are typically large, ranging from hundreds of megabytes to terabytes for a single data-product with a total volume of over 30 petabyte in the LTA for several millions of data-products. Derived data-products can be large as well, covering an even wider range of size per data-product but typically one or more orders of magnitude smaller than the observation data.
Standards and	Radio-astronomical data can be in one of various data-formats with

[8] http://dx.doi.org/10.13155/29825

[9] http://seadatanet.maris2.nl/v_bodc_vocab_v2/welcome.asp

metadata	varying levels of definition and standardization. Among the most used formats are the Measurement Set[10], the FITS[11] format and the HDF5 [12]format. For LOFAR, a set of Interface Control Documents, including a description of the metadata contained in the data formats, can be found on the LOFAR WIKI[13]. The metadata in the catalogue for the LTA is filled using XML documents that comply to a custom schema[14]. A limited-scope ontology for radio-astronomical data-products is being developed in the EOSC pilot project.
Data sharing	
Target groups	The main target group is the (LOFAR) radio-astronomical community. The science level data products that are generated by the LOFAR community will be of interest for a much wider astronomical community that is involved in multi-wavelength research. Additionally, LOFAR data can be of interest to other communities, e.g. for ionospheric and space-weather research.
Scientific Impact	The principal objective of the RACC is to improve the science-generating capabilities of the LOFAR community by leveraging lower threshold access to appropriate large scale computing facilities that can connect at significant bandwidth to the LTA storage. It is expected that the generation and sharing of science ready data-products will increase scientific output significantly as it is known that science output from archives of science-level astronomical data can be a factors higher than that directly from observation data.
Approach to sharing	The ILT promotes open access to archived data, keeping data under embargo for a limited time only to allow the creation of scientific publications from requested observation data. The LTA catalogue [15]can be queried publicly and the RACC aims to improve public and open sharing of data by building on EOSC-based FAIR data services.
Archiving and preservation	The LOFAR LTA provides a centrally managed data archive, ensuring long term preservation. The ILT Observatory supports the LOFAR community in accessing and processing the data.

[10] https://casa.nrao.edu/Memos/229.html

[11] https://fits.gsfc.nasa.gov/fits_standard.html

[12] https://support.hdfgroup.org/HDF5/

[13] https://www.astron.nl/lofarwiki/doku.php?id=public:documents:lofar_documents#dataformats

[14] https://svn.astron.nl/viewvc/LOFAR/trunk/LTA/LTACommon/LTA-SIP.xsd?view=co

[15] https://lta.lofar.eu/