



**EOSC-hub**

## D1.5 Data Management Plan

<b>Lead Partner:</b>	EGI Foundation
<b>Version:</b>	1
<b>Status:</b>	Rejected by EC
<b>Dissemination Level:</b>	Public
<b>Document Link:</b>	<a href="https://documents.egi.eu/document/3379">https://documents.egi.eu/document/3379</a>

### Deliverable Abstract

A report that will specify how research publications and data will be collected, processed, monitored, catalogued, and disseminated 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.



**COPYRIGHT NOTICE**

This work by Parties of the EOSC-hub Consortium is licensed under a Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>). The EOSC-hub project is co-funded by the European Union Horizon 2020 programme under grant number 777536.

**DELIVERY SLIP**

	<i><b>Name</b></i>	<i><b>Partner/Activity</b></i>	<i><b>Date</b></i>
<b>From:</b>	Małgorzata Krakowian	EGI Foundation	
<b>Moderated by:</b>	Małgorzata Krakowian		
<b>Reviewed by</b>	Yin Chen	EGI Foundation	7/08/2018
	Mark van de Sanden	SurfSara	8/08/2018
<b>Approved by:</b>	AMB		

**DOCUMENT LOG**

<i><b>Issue</b></i>	<i><b>Date</b></i>	<i><b>Comment</b></i>	<i><b>Author</b></i>
v.0.1	4/07/2018	First version	M. Krakowian
FINAL	31/10/2018	Final version	M. Krakowian

---

## Contents

1	Introduction.....	5
2	Datasets.....	7
2.1	Disaster Mitigation Competence Centre Plus (DMCC+) .....	7
2.2	EISCAT_3D.....	8
2.3	ELIXIR.....	9
2.4	EPOS-ORFEUS .....	9
2.5	Fusion.....	11
2.6	ICOS.....	12
2.7	Marine.....	12
2.8	Radio Astronomy Competence Center (RACC).....	13

## **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 initial 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<sup>1</sup>. The focus of the open research data pilot in Horizon 2020 is on research data that is available in digital form<sup>2</sup>.

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 initial data management plan<sup>3</sup> for the research data that will be generated within EOSC-hub. 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.

The EOSC-hub project activities will also generate data arising from user surveys or from managing the hub that are used to define requirements to create or improve services. Because this data drives research and innovation in the services and solutions that EOSC-hub provides and therefore can also underpin scientific publications, those data are also considered in the scope of this data management plan.

---

<sup>1</sup>[http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access\\_en.htm](http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm)

<sup>2</sup> 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](http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf)

<sup>3</sup> 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.

As recommended in Guidelines on Data Management in Horizon 2020, this document will be further developed in D1.6 Data Management Plan (June 2019) with more detailed information related to the discoverability, accessibility and exploitation of the data.

## 2 Datasets

Within the EOSC-hub project, research data will be mainly generated or collected by the WP8 Competence Centres (CC). The following sections provide 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 initial data management plan.

### 2.1 Disaster Mitigation Competence Centre Plus (DMCC+)

<b>Task</b>	T8.8
<b>Contact</b>	Eric Yen <a href="mailto:Eric.Yen@twgrid.org">Eric.Yen@twgrid.org</a> , Simon Lin <a href="mailto:Simon.Lin@twgrid.org">Simon.Lin@twgrid.org</a>
<b>Data description</b>	
<b>Types of data</b>	<ol style="list-style-type: none"> <li>1. Observation data for target hazard events: global model data; gridded data in GRIB format; satellite data; radar data; etc.</li> <li>2. Geographical data of impact areas of target hazard events: topographical data, land use, bathymetry, etc.</li> <li>3. Simulation results: 3D gridded data, images, video, visualization data</li> </ol>
<b>Origin of data</b>	Observation data comes from agencies such as NCEP, NASA, ECMWF and local weather bureau or related government agencies.
<b>Scale of data</b>	<p>Depend on temporal and spatial resolution.</p> <p>Weather simulation by WRF (per simulation): observation and static data scale per simulation is O(100MB); Simulation result is O(TB) per simulation.</p> <p>Tsunami and Storm Surge (per simulation): input data scale is O(10MB); simulation result is O(GB).</p>
<b>Standards and metadata</b>	<p>Data format: GRIB/GRIB2; NetCDF</p> <p>Metadata: no domain specific metadata standard is applied. Darwin Core metadata scheme is used.</p>
<b>Data sharing</b>	
<b>Target groups</b>	communities of research, education, hazard risk estimation and analysis, disaster management, etc.
<b>Scientific Impact</b>	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.
<b>Approach to sharing</b>	Web services, searchable data catalogue, APIs, etc.
<b>Archiving and preservation</b>	Centralized archive is provided by ASGC for the moment. Will be implemented by distributed storage services over the regional infrastructure. The solution is under evaluation.

## 2.2 EISCAT\_3D

<b>Task</b>	T8.4
<b>Contact</b>	Ingemar Häggström <a href="mailto:ingemar.haggstrom@eiscat.se">ingemar.haggstrom@eiscat.se</a> , Carl-Fredrik Enell <a href="mailto:carl-fredrik.enell@eiscat.se">carl-fredrik.enell@eiscat.se</a>
<b>Data description</b>	
<b>Types of data</b>	Different levels of EISCAT radar data of the geospace environment
<b>Origin of data</b>	EISCAT radar sites
<b>Scale of data</b>	Several stages of buffers, from ms (~0.5 PB) to some months (20 PB). Final archive sizes 2-4 PB/year. Spatial scales for cm to many km.
<b>Standards and metadata</b>	Scientific metadata following geospace community standards with EISCAT enhancements, like Madrigal Engineering metadata following instrumental standards, like RFRadar Cross discipline metadata following environmental and astronomical standards, like ENVRI General metadata following data infrastructure standards, like B2Find
<b>Data sharing</b>	
<b>Target groups</b>	Scientists from associate countries, affiliated institutes, all world, long tail
<b>Scientific Impact</b>	Large collaborations
<b>Approach to sharing</b>	Data and processing services. Lower levels of data have embargo times of up to 4 years, open to the association. High levels of data with scientific parameters open after quality control, quick-look data open from time zero.
<b>Archiving and preservation</b>	General archives to be preserved over the EISCAT lifetime, 30 years, after which the associates have committed to



	<p>continued storage.</p> <p>The data used in the CC development are mainly residing in the EISCAT archives, B2Share and registered metadata into B2Find.</p>
--	---

## 2.3 ELIXIR

<b>Task</b>	T8.1
<b>Contact</b>	Steven Newhouse <a href="mailto:steven.newhouse@ebi.ac.uk">steven.newhouse@ebi.ac.uk</a> , Susheel Varma <a href="mailto:susheel.varma@ebi.ac.uk">susheel.varma@ebi.ac.uk</a>
<b>Data description: Types of data</b>	<p>Public reference datasets: Genes, Protein, metabolite expression, protein sequences, molecular structures, chemical biology, reactions, interactions and pathways, systems biology.</p> <p>Some container and workflow execution data will be generated but will be discarded after integration testing between ELIXIR and EOSC-Hub</p>
<b>Data description: Origin of data</b>	ELIXIR Data Platform & EMBL-EBI
<b>Data description: Scale of data</b>	Scale: Spatial - Molecules to Systems Biology; Temporal - $\mu$ s to years; Storage: KB to PB
<b>Standards and metadata</b>	Technical Metadata ( data object id, uri path, size, version, checksum, metadata links)
<b>Data sharing: Target groups</b>	Life scientists and cross-domain researchers
<b>Data sharing: Scientific Impact</b>	Facilitate collaborations by facilitating access to large reference datasets
<b>Data sharing: Approach to sharing</b>	Public Datasets are freely accessible by any. Data caches in an institutional site may be restricted and will be managed by the institution.
<b>Archiving and preservation</b>	Eternal 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

<b>Task</b>	T8.5
<b>Contact</b>	Sara Ramezani <a href="mailto:sara.ramezani@surfsara.nl">sara.ramezani@surfsara.nl</a> , Luca Trani <a href="mailto:Luca.Trani@knmi.nl">Luca.Trani@knmi.nl</a> , Javier Quinteros (GFZ) <a href="mailto:javier@gfz-potsdam.de">javier@gfz-potsdam.de</a>

<b>Data description</b>	
<b>Types of data</b>	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
<b>Origin of data</b>	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
<b>Scale of data</b>	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.
<b>Standards and metadata</b>	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
<b>Data sharing</b>	
<b>Target groups</b>	Seismology researchers and solid Earth-scientists. Currently EIDA servers ~1900 unique users p/y
<b>Scientific Impact</b>	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.
<b>Approach to sharing</b>	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.
<b>Archiving and preservation</b>	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

<b>Task</b>	T8.2
<b>Contact</b>	de Witt, Shaun <a href="mailto:shaun.de-witt@ukaea.uk">shaun.de-witt@ukaea.uk</a>
<b>Data description</b>	
<b>Types of data</b>	<p>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.</p> <p>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.</p>
<b>Origin of data</b>	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.
<b>Scale of data</b>	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.
<b>Standards and metadata</b>	Only local metadata and formats currently exist; there is currently no standard metadata or format for fusion data. During the course of the project, it is anticipated that a metadata standard will emerge with ontology mapping to existing local standards.
<b>Data sharing</b>	
<b>Target groups</b>	Public, researchers from other domains
<b>Scientific Impact</b>	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.
<b>Approach to sharing</b>	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.
<b>Archiving and preservation</b>	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.

## 2.6 ICOS

<b>Task</b>	T8.7
<b>Contact</b>	Alex Vermeulen <a href="mailto:alex.vermeulen@icos-ri.eu">alex.vermeulen@icos-ri.eu</a> , Margareta Hellström <a href="mailto:margareta.hellstrom@nateko.lu.se">margareta.hellstrom@nateko.lu.se</a>
<b>Data description</b>	
<b>Types of data</b>	High frequency eddy covariance flux data for CO <sub>2</sub> and other gases
<b>Origin of data</b>	Measurements stations from the ICOS and LTER networks
<b>Scale of data</b>	Up to 78 stations from ICOS and 20 from LTER provide half hourly updates of 20 Hz data
<b>Standards and metadata</b>	Ecosystem stations use BADM metadata, raw data datafiles are binary or TSV formatted ASCII in a well described community standard. Output files are community standard TSV formatted ASCII files, metadata is ISO19115 and Inspire compliant
<b>Data sharing</b>	
<b>Target groups</b>	Carbon science, climate science, remote sensing satellite data validation, vegetation models tuning and validation, crop model improvements, COPERNICUS, FLUXNET
<b>Scientific Impact</b>	At least 4000 downloads per year, hundreds of articles and ten-thousands of citations per year
<b>Approach to sharing</b>	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)
<b>Archiving and preservation</b>	B2SAFE trusted repository

## 2.7 Marine

<b>Task</b>	T8.3
<b>Contact</b>	Thierry Carval <a href="mailto:tcarval@ifremer.fr">tcarval@ifremer.fr</a>
<b>Data description</b>	
<b>Types of data</b>	Argo floats ocean data, SeaDataCloud marine data
<b>Origin of data</b>	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.

<b>Scale of data</b>	The Argo dataset is a collection of 2 million NetCDF files, of about 250GB  A not yet defined subset of SeaDataCloud will be provided.
<b>Standards and metadata</b>	Argo data files comply with NetCDF CF1.6 convention (Climate and Forecast).  Data and metadata formats are published "Argo user's manual" <a href="http://dx.doi.org/10.13155/29825">http://dx.doi.org/10.13155/29825</a>  The parameters are compliant with SeaDataNet P01 (Parameter Usage Vocabulary) and P06 (data storage units) vocabularies.  SeaDataCloud data files comply with SeaDataNet vocabularies. <a href="http://seadatanet.maris2.nl/v_bodc_vocab_v2/welcome.asp">http://seadatanet.maris2.nl/v_bodc_vocab_v2/welcome.asp</a>
<b>Data sharing</b>	
<b>Target groups</b>	scientists, operational services
<b>Scientific Impact</b>	climate studies, seasonal forecasting, meteo-oceano activities
<b>Approach to sharing</b>	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.
<b>Archiving and preservation</b>	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)

<b>Task</b>	T8.6
<b>Contact</b>	Hanno Holties <a href="mailto:holties@astron.nl">holties@astron.nl</a> , Rob van der Meer < <a href="mailto:meer@astron.nl">meer@astron.nl</a> >
<b>Data description</b>	
<b>Types of data</b>	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.)
<b>Origin of data</b>	Observation data is generated by the LOFAR instrument and the processing cluster managed by the ILT Observatory. It is

	<p>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.</p>
<b>Scale of data</b>	<p>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.</p>
<b>Standards and metadata</b>	<p>Radio-astronomical data can be in one of various data-formats with varying levels of definition and standardization. Among the most used formats are the <a href="#">Measurement Set</a>, the <a href="#">FITS</a> format and the <a href="#">HDF5</a> format. For LOFAR, a set of Interface Control Documents, including a description of the metadata contained in the data formats, can be found on <a href="#">the LOFAR WIKI</a>. The metadata in the catalogue for the LTA is filled using XML documents that comply to a custom <a href="#">schema</a>. A limited-scope ontology for radio-astronomical data-products is being developed in the EOSC pilot project.</p>
<b>Data sharing</b>	
<b>Target groups</b>	<p>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.</p>
<b>Scientific Impact</b>	<p>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.</p>
<b>Approach to sharing</b>	<p>The <a href="#">ILT promotes</a> 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 <a href="#">LTA catalogue</a> can be queried publicly and the RACC</p>

---

	aims to improve public and open sharing of data by building on EOSC-based FAIR data services.
<b>Archiving and preservation</b>	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.