D1.6 Data Management Plan

Lead Partner:	EGI Foundation
Version:	2
Status:	Under EC review
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
V1	5/09/2019	Final version	M. Krakowian
V2	13/02/2020	Final version after addressing Project review feedback	M. Krakowian

Contents

Executive summary

This document defines data management plan for research data generated or collected by the EOSC-hub project. 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).

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 Data management plans per WP

2.1 WP1

Contact	Malgorzata Krakowian
Data description	
Types of data	1. Project Documentation (Metrics, Risks, Procedures, Plans, Meetings, Presentations) 2. Deliverables 3. Service management system 4. Effort and financial data
Origin of data	All the data was produced and provided by project members.
Scale of data	<100MB
Standards and metadata	plain text, pdf, docx, pptx
Data sharing	
Target groups	The target group is all project members and project office.
Scientific Impact	Not applicable
Approach to sharing	1. Shared within the consortium to support work 2. All deliverables are shared within the consortium and also with EC. Public deliverables are accessible to everyone via EOSC-hub website 3. Shared within the consortium to support work 4. Shared with Project office and management boards to support work, as well as with the EC.
Archiving and preservation	Once the project is finished, all the WP1 information will be preserved by EGI for at least 5 years.

2.2 WP2

Contact	Tiina Maria Kupila-Rantala
Data description	
Types of data	1. Related to D2.2, 12 recorded interviews (2 recordings failed) 2. Related to D2.2, notes of 14 interviews 3. Service Portfolio Management data a. Documentation on onboarding b. Data on services and providers collected from providers

Origin of data	1, 2 Data has been collected from representatives of 14 Horizon 2020 projects contributing to EOSC: Archiver PCP, DEEP-HybridDataCloud, eInfraCentral, EOSC Nordic, eXtreme-DataCloud, EOSC-Pillar, EOSC-synergy, ExPAaNDS, NI4OS-Europe, FAIRsFAIR, FREYA, GÉANT, OCRE, RDA Europe 4.0 3: submitted material from ~180 services from over 100 providers
Scale of data	1. Recordings are of the size of 20-30 MB each. 2. Notes of the interviews are of the size of 25 kB each 3. Around 9mb for the DT data
Standards and metadata	1, 2 N/A 3: Emerging metadata standard scribing services in EOSC, being defined by EOISC-hub, osc N
Data sharing	
Target groups	1, 2 European Commission and EOSC-hub Consortium members in format of the EOSC-hub Strategy Plan D2.2 3. The full data set os for the EEOSC operators (Currently EOSC-hub, later iNFRAEOSSC-03 or the EOSC aisbl)
Scientific Impact	None
Approach to sharing	1, 2 As agreed with the interviewees, the data has been used only for the purpose of preparing the D2.2. The data sets will not be shared or used beyond the context of gathering without further permissions of the interviewees. 3. A subset which is marked as public to be exposed through the catalogue and via EOSC portal, the rest private and held to enable EOSC operations.
Archiving and preservation	1, 2 WP2 Lead CSC maintain the collected data sets within CSC servers 3. Held by EGI as part of T2.2 leadership, to be passed on to INFRAEOSC-03, Enhance and an EOSC aisbl.

2.3 WP3

Contact	Sara Garavelli
Data description	
Types of data	1. Contact information including First Name, Last Name, Email Address, Organisation, Best way to describe yourself (stakeholder type) 2. EOSC-hub Website Statistics including page views, sessions,

	users, and visit source
Origin of data	1. F2f meetings, EOSC-hub Website Webforms (event registration, mailing list subscription, survey submissions, Early Adopter Programme applications etc.) 2. Google analytics
Scale of data	• <10mb • Email contacts database is estimated to reach 600-800 contacts by project end
Standards and metadata	Not applicable
Data sharing	
Target groups	• Stakeholder information: For engagement by WP3 members and the project management • European Commission: for reporting purposes • EOSC-hub Consortium members: for reporting and operational purposes • OCRE, GÉANT, OpenAIRE: for proessing the Early Adopter Programme applications
Scientific Impact	Scientific impact is not applicable to the type of data reported
Approach to sharing	• Access to the data sets, due to their sensitive nature and in respect of GDPR, is provided only to the appropriate and earlier agreed target groups. Web data via Google spreadsheet • Stakeholder data for engagement: Stored in a Confluence DB (https://confluence.egi.eu/display/EOSC/Stakeholder+DB) restricted to EOSC-hub participants
Archiving and preservation	WP3 Lead Trust-IT maintains data within its servers. Additionally, tools used during the operation of WP3 also store contact data i.e. Mailchimp, Drupal Once the project is finished, the WP3 confluence information will be preserved by EGI for at least 5 years.

2.4 WP4

Contact	Matthew Viljoen
Data description	
Types of data	• The EOSC SMS created within the project, protected and stored in Confluence • Processes, procedures, policies and people associated with maintaining these. • Services and service contacts

	• Jira tickets covering different aspects of the SMS and project • Project meeting minutes in Indico
Origin of data	Created as a result of work done within the project
Scale of data	<100MB
Standards and metadata	Not applicable
Data sharing	
Target groups	Participants within the project and collaborators
Scientific Impact	None directly. Indirectly through optimal service delivery via a functioning SMS
Approach to sharing	Shared within the project consortium and to follow-on projects (infraeosc03)
Archiving and preservation	Once the project is finished, the WP4 confluence/JIRA information will be preserved by EGI for at least 5 years.

2.5 WP5

Contact	Pavel Weber
Data description	
Types of data	WP5 collects storage accounting data, service monitoring metrics, customer information like identity, affilication, VO/group membership as well as order, feature requests and incidents information. WP5 also collects and analyses data in the scope of surveys e.g. AAI survey in 2019 to identify the usa cases and customer requirements to the central federation services.
Origin of data	• Surveys, • EOSC Portal/Marketplace service orders • Helpdesk • AAI IdP/SP Proxies • F2F Meetings • Data repositories • Generated tutorials, webinars, service documentation for customers
Scale of data	<200GB
Standards and metadata	Not applicable
Data sharing	
Target groups	Service Providers, service owners, operators, process managers (e.g. ISRM for Helpdesk and SOCRM for Marketplace),

	stakeholders, EC
Scientific Impact	• Indirect impact on the thematic and common services which are integrated with federation services like AAI, enabling access, accounting and monitoring of the services as well as data sharing between different storage platforms. • Generation of webinars, tutorials and service documentation facilitates integration of scientific platforms and research infrastructures in EOSC, and as a result simplifies scientific data sharing and usage of the EOSC infrastructure for research.
Approach to sharing	Data manged in WP5 is managed and shared according to GDPR and available to authorized agents (digital and personal access) via • Confluence wiki • JIRA • Google docs • Service API • Federation Transport Layer (Messaging Service)
Archiving and preservation	Data stored on EOSC collaborative servers, EC Portal, archived and backuped by service owners according the availability and continuity plan. Once the project is finished, the WP5 confluence information will be preserved by EGI for at least 5 years.

2.6 WP6

Contact	John Alan Kennedy
Data description	
Types of data	WP6 collects various statistics/metrics to monitor the uptake and usage of services provided by EOSC hub. No personal user data is collected in this context.
Origin of data	• Through access to the services themselves (w.r.t access to services etc) - service administrators gather statistics • Via the portal
Scale of data	Few MBs
Standards and metadata	Not applicable
Data sharing	
Target groups	European Commission to provide evidence of service uptake, future infrastructure projects to aid with planning and focus/scope decisions.
Scientific Impact	Not applicable

Approach to sharing	The data will be shared in deliverables and milestone documents.
Archiving and preservation	Data published in deliverables will be stored in the EC portal etc. The EOSC collaborative services (wiki pages, web sites). Once the project is finished, the WP6 confluence information will be preserved by EGI for at least 5 years.

2.7 WP7

2.7.1 OPENCoastS

Task	T7.5
Contact	Anabela Oliveira deputy: Alberto Azevedo, technological development deputy: João Rogeiro
Data description	
Types of data	Inputs: grid files, boundary conditions, parameter files (all in SCHISM formats, in ASCII, binary and netcdf) Outputs: model results for the several 2D or 3D variables (water levels, velocity, salinity, temperature, all netcdf files)
Origin of data	Inputs: provided by the users, confidential Outputs: Generated within the service, confidential
Scale of data	(size) of average simulation per day: 2D simulations - 1Gb; 3D simulations - 10 Gb
Standards and metadata	ASCII; Unstructured Netcdf outputs, WMS: ncwms layers
Data sharing	
Target groups	Researchers, end-users and the general public interested in coastal management
Scientific Impact	The free availability of the service allows everyone to build their own forecast systems. The service can contribute to improve coastal management, harbor operations, coastal recreation, etc.
Approach to sharing	Input and outputs are property of the users. Ncwms layers data sharing is possible if users are willing to share.
Archiving and preservation	For a short period of time, through EUDAT services. Preservation is done by the users.

2.7.2 ECAS

Task	T7.3
Contact	Tobias Weigel, deputy: Sandro Fiore

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	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).
Data sharing	
Target groups	Research and academic community, public and private sector (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.

2.7.3 WeNMR

Task	T7.6
Contact	Alexandre Bouvin, deputy: Antonio Rosato
Data description	
Types of data	User-Specific input data for the various portals in form of text files representing various kind of experimental restraints (information) and coordinates (PDB or mmCIF format)
Origin of data	End user or in some case the PDB database (http://www.wwpdb.org)

Scale of data	1-100 MB of input data, up to several GB of output data
Standards and metadata	Coordinate files are typically PDB or mmCIF formatted files (see http://www.wwpdb.org)
Data sharing	
Target groups	Data belong to the user, no sharing mechanism in place. Results of computations might be deposited in standard public databases for structural biology like http://www.wwpdb.org, https://pdb-dev.wwpdb.org or https://data.sbgrid.org
Scientific Impact	N.A. (not done by WeNMR)
Approach to sharing	End user responsibility - deposition in public databases (often a requirement from scientific journals)
Archiving and preservation	N.A. (not done by WeNMR)

2.8 WP8

2.8.1 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]

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

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 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.

2.8.2 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.

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

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

	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 (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.

2.8.3 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:	Facilitate collaborations by facilitating access to large reference

Scientific Impact	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).

2.8.4 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.

2.8.5 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

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

	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.

2.8.6 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

2.8.7 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 metadata	Argo data files comply with NetCDF CF1.6 convention (Climate and 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.

2.8.8 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

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

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

	orders of magnitude smaller than the observation data.
Standards and metadata	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 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/

2.9 WP9

Contact	Sy Holsinger
Data description	
Types of data	1. Pilot Contact information including First Name, Last Name, Email Address, Organisation, skype account, Twitter account, Linkedin account. 2. Satisfaction information regarding usage of EOSC-DIH services 3. Pilot description (name of the company, products and solutions) with the services / infrastructure used, TRL 4. Potential customer information (name of the company, name of the contact and email) 5. Collaboration agreements, including Name of the company, description and objectives and task to carry out.
Origin of data	1. Pilot contributors, MoU/Collaboration Agreements 2. Surveys at the end of the pilot 3. Via email within WP9, or via Trello 4. F2F meetings / conferences / Trello 5. Via email within WP9
Scale of data	• <10mb • Email contacts database is estimated to reach 50-80 by the end of the project
Standards and metadata	Data is following spreadsheets template agreed within WP9
Data sharing	
Target groups	• European Commission: for reporting purposes • EOSC-hub Consortium members: for reporting and operational purposes • DIH communities
Scientific Impact	Scientific impact is not applicable to the type of data reported.
Approach to sharing	1. Shared within the WP9 members through confluence webpages and private (WP9) spreadsheets stored in Google Drive. 2. Stored in WP9 Google Drive. 3. Website, confluence and Drive, Trello 4. Confluence, Trello 5. Website, confluence and Drive
Archiving and preservation	1. Once the project is finished, the webpage will be hosted by EGI and the confluence information will be preserved by EGI for at least 5 years. 2. Once the project is finished, the Google Drive will be preserved by EGI for at least 5 years. 3. Once the project is finished, the webpage will be hosted by EGI

	and the Drive and Confluence information will be preserved by EGI for at least 5 years. Any relevant information remaining in Trello will be moved to Confluence. 4. Once the project is finished, Confluence information will be preserved by EGI for at least 5 years. Any relevant information remaining in Trello will be moved to Confluence. 5. Once the project is finished, the Webpage will be hosted by EGI and the Drive and Confluence information will be preserved by EGI for at least 5 years.

2.10 WP10

Contact	Giacinto Donvito
Data description	
Types of data	• Contact information including First Name, Last Name, Email Address, Organisation, Best way to describe yourself (stakeholder type) • User community use cases • Services information coming from WP5-WP6-WP7
Origin of data	• F2f meetings • EOSC-hub Website Webforms (event registration, mailing list subscription, survey submissions, Confluence Web Pages, Early Adopter Programme applications etc.)
Scale of data	• <10mb • Email contacts database is estimated to reach 60-100 by the end of the project
Standards and metadata	Not applicable
Data sharing	
Target groups	• European Commission: for reporting purposes • EOSC-hub Consortium members: for reporting and operational purposes • OCRE, GÉANT, OpenAIRE, and Communities involved in Early Adopters Programme
Scientific Impact	Scientific impact is not applicable to the type of data reported
Approach to sharing	• Access to the data sets, due to their sensitive nature and in respect of GDPR, is provided only to the appropriate and earlier agreed target groups. • Confluence Web pages • Google documents closed to the WP members

Archiving and preservation	Once the project is finished, the WP10 confluence information will be preserved by EGI for at least 5 years.

2.11 WP11

Contact	Giuseppe La Rocca
Data description	
Types of data	1. information about training events, to display them in training catalog https://eosc-hub.eu/training-events 2. training materials about the available services in the service catalog, and https://eosc-hub.eu/training-material 3. feedback from participants attending the events. No personal data of users is collected.
Origin of data	1. feedback from participants is collected via web forms and paper documents, 2. service providers are providing material for the target groups, 3. trainers organizing the training events
Scale of data	few MB of data
Standards and metadata	Training feedback is following the template prepared by WP11
Data sharing	
Target groups	1. End users and service providers 2. EOSC-hub project - training feedback
Scientific Impact	None
Approach to sharing	Information about upcoming and past events are publicly available in the EOSC Training Registry on the EOSC-hub website
Archiving and preservation	Information about training events and materials are stored in the Trust-IT servers. Feedback from training events are stored in a shared folder of EOSC-hub and summarized in WP11 deliverables. Once the project is finished, the WP11 confluence information will be preserved by EGI for at least 5 years.

2.12 WP12

Contact	Sergio Andreozzi
Data description	
Types of data	Data from market research: transcriptions of interviews,

	spreadsheet containing data from online survey
Origin of data	32 organisations interviewed + 33 organisations participated in an online survey
Scale of data	< 10 MB
Standards and metadata	Natural language associated to questions
Data sharing	
Target groups	The raw data was collected mainly to develop insights in the context of the market research conducted within WP12; the privacy policy stated that only aggregated and anonymised data would be published. Therefore, such raw data is accessible only to partners of the Work Package 12. The aggregated and anonymised data are published in Deliverable D12.1 (public deliverable). The target groups are research manager, procurers, senior managers of service providers and research performing organisations, policy makers, EOSC implementation projects.
Scientific Impact	The derived insights were published in Deliverable D12.1 After the first periodic review, the EC reviewers suggested to consider the opportunity to write a scientific paper or a technical paper. We will consider this in the final part of the project, within T12.3.
Approach to sharing	The results of the analysis is published in D12.1. The raw data is kept private within the consortium to comply with the defined privacy policy.
Archiving and preservation	Deliverable D12.1: https://documents.egi.eu/document/3466 Raw data in EOSC-hub Confluence section private to WP12 partners. Once the project is finished, the WP12 confluence information will be preserved by EGI for at least 5 years.

2.13 WP13

Contact	Malgorzata Krakowian
Data description	
Types of data	WP13 is collecting statistics and metrics related to operation of the services number of users, usage, marketplace and EOSC-hub website views. No personal data of users is collected.
Origin of data	Data are collected from google analytics and from accounting data collected by the provider.

Scale of data	Data is stored and collected in Microsoft .docx and .pdf files of a size 1-2 MB
Standards and metadata	Data is following reporting template agreed with EC in .pdf files.
Data sharing	
Target groups	Data is collected for European Commission to provide evidence that services under Virtual access funding mechanism are used by users.
Scientific Impact	None
Approach to sharing	All data is shared publicly via document db service provide by EGI, EOSC-hub website and also uploaded to EC portal as deliverable.
Archiving and preservation	All the data will be stored as deliverables in EC portal, in EC service document db and EOSC-hub website. Once the project is finished, the WP13 confluence information will be preserved by EGI for at least 5 years.