D1.2 Data management Plan

Lead partner:	EGI Foundation
Version:	1
Status:	Final
Dissemination Level:	Public
Keywords:	Data
Document Link:	https://documents.egi.eu/document/3763

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.

COPYRIGHT NOTICE

EGI-ACE receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 101017567.

DELIVERY SLIP

Date	Name	Partner/Activity
From:	Małgorzata Krakowian	EGI Foundation/WP1
Moderated by:	Małgorzata Krakowian	
Reviewed by:	Marta Gutierrez Andrea Cristofori	EGI Foundation EGI Foundation

DOCUMENT LOG

Issue	Date	Comment	Author
v.1	16/07/2021	First version based on the input from WPs	M. Krakowian S. Andreozzi G. La Rocca H. Bui A. Paolini E. Fernandez A. Manzi
1	4/08/2021	Final version addressing review comments	M. Krakowian

TERMINOLOGY

https://confluence.egi.eu/display/EGIG

Contents

Executive summary

This document defines data management plan for research data generated or collected by the EGI-ACE 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.

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.

[1] https://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

2 Data management plans per WP

2.1 WP1 Project Management

WP1 will collect and manage data related to the project management activities that are required by European commission. The detailed plan how to manage outputs of the projects as well as the confidentiality is described in D1.1.

Contact	Hien Bui
Data description	
Types of data	1. Project Documentation (Metrics, Risks, Procedures, Plans, Meetings, Presentations) 2. Deliverables 3. Effort and financial data
Origin of data	All the data was produced and provided by project members.
Scale of data	<1GB
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. Public and Confidential deliverables are shared within the consortium and with the EC. a. Public deliverables are accessible to everyone via EGI-ACE website and Zenodo platform 3. 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 Coordination and cooperation

Contact	Sergio Andreozzi
Data description	

Types of data	T2.1 will produce the following types of data:
	
	• Contact information for liaisons with other initiatives for strategy alignment
	• Data from surveys or interviews
	
	T2.2 will produce the following types of data:
	
	• Service portfolio information
	• Service Management System (policies, processes, procedures, databases)
	• Results from FAIR data assessment of services
	
	T2.3 will produce the following types of data:
	
	• Customer database
	  ○ Contacts of the Early Adopters that may results in new Data Space providers in EOSC.
	  ○ Contacts of the scientific applications selected via open calls that may results in new users of the EOSC Compute Platform.
	• Training materials, documentations and guidelines
	
	T2.4 will produce the following types of data:
	
	• Communication and dissemination data
	  ○ Contact information including First Name, Last Name, Email Address, Organisation, Best way to describe yourself (stakeholder type)
	  ○ Website Statistics including page views, sessions, users, and visit source

Origin of data	T2.1: • Interviews or surveys • Personal contacts T2.2: • Service providers and service portfolio manager of the project • Service Management System process managers and process members • Surveys and interviews with Data Space providers T2.3: • Trainers producing training material • Recording of training events • Data is created by the team members working in the WP2, 3, 4 and 6. T2.4: • Google analytics

	• HootSuite • MailerLite • Registration systems for events like Indico
Scale of data	GB
Standards and metadata	T2.1: • Mostly free text or text filled in pre-defined templates, to be stored in file which format is pdf, docx, or stored in the Confluence space (HTML) T2.2: • Free text structured into tables, stored in project confluence • Text in Confluence • Mostly free text or text filled in pre-defined templates, to be stored in file which format is pdf, docx, or stored in the Confluence space (HTML) T2.3: • Standards and metadata: plain text, .pdf, .docx, .pptx, scripting, docker containers, etc. T2.4: • Mostly free text or text filled in pre-defined templates, to be stored in file which format is pdf, docx, or stored in the Confluence space (HTML
Data sharing	
Target groups	T2.1: • EGI-ACE project consortium, INFRAEOSC-07 projects, EOSC-Future project, EOSC Association T2.2: • Consortium members; Public part with visitors of the EOSC Portal/Marketplace • INFRAEOSC-07 projects, EOSC-Future project, relevant EOSC Association Task Forces T2.3: • Scientific communities, • Consortium members • Early Adopters of the EOSC Compute Platform • Consultants and trainers T2.4: • Most of the data is for internal use

	• Metrics about uptake of communication channels will be shared in deliverables for the EC review
Scientific Impact	T2.1: • No relevant for scientific publications T2.2: • Scientists will read/find service descriptions in the EOSC Portal/Marketplace and will decide about use • SMS: Not relevant for scientists • Results could be considered for publication in a technical related conference/journal T2.3: • The information can be used to develop skills in researchers, but it is not suitable for scientific publications T2.4: • No relevant for scientific publications
Approach to sharing	T2.1: • Data will be aggregated and published via the deliverables with an open license • Contact information will stay private to the project consortium T2.2: • Subset of data to be shared in EOSC Portal/Marketplace • In project confluence; In EGI Confluence • Aggregate data will be published via the deliverables with an open license T2.3: • Shared within the project consortium to support work • Training materials and guidelines will be made available in indico pages and repository[3] • Contact information will become part of the EGI Service Management System T2.4: • Aggregate data will be published via the deliverables with an open license • Contact information will become part of the EGI Service Management System
Archiving and preservation	Once the project is finished, all the materials will be preserved by EGI for at least 5 years.

[3] http://docs.egi.eu

2.3 WP3 Infrastructure Services (VA), WP4 Platform Services (VA), WP6 Federated Access Services (VA)

Contact	Enol Fernandez del Castillo WP3 WP4 Andrea Manzi WP6
Data description	

Types of data	• The EGI Foundation SMS supporting the project, protected and stored in Confluence and other tools:
	  ○ Services, service endpoints, and service contacts
	  ○ Incidents and Service requests created as GGUS tickets (ticketing management system – Help desk)
	  ○ Monitoring and service performance reports
	  ○ Agreements for service delivery in EGI's DocumentDB
	• Deliverables
	• Jira tickets covering different aspects of the project activities
	• Names and email addresses of people with effort in WP3
	• Project meeting information in indico and minutes in Google docs and confluence
	• Documentation and guidelines
	• User information for providing access to the services
	• VA metrics to monitor the usage of services as defined in the project

Origin of data	Created as a result of work done within the project
Scale of data	< 1GB
Standards and metadata	plain text, pdf, MS Word, MS Excel, MS Powerpoint, Google Docs, online access as provided by the Service Management Tools
Data sharing	
Target groups	• Participants within the project and collaborators • members of EGI Federation
Scientific Impact	n/a
Approach to sharing	Shared within the project consortium to support work
Archiving and preservation	Once the project is finished, all the WP information will be preserved by EGI for at least 5 years.

2.4 WP5 Federated data spaces (VA)

WP5 is composed of Data spaces which approach to the data management is diverse, this is why each task is defined separately. The project will look into FAIR Data Maturity Model defined by RDA Working group [4]to better define and assess data management approach of the data spaces. Result will be provided with next version of Data management plan deliverable.

T5.1 OpenRiskNet/NanoCommons Virtual Environment

Contact	Barry Hardy
Reference Data description	
Types of data	N/A
Origin of data	N/A
Scale of data	N/A
Scientific Data description	
Types of data	Scientific data produced by projects working in the area of chemical or nanomaterials safety assessment. This includes chemical, biological and toxicological data generated by new approach methods including in silico and in vitro approaches.
Origin of data	Partners on EU projects.
Scale of data	Smaller complex datasets.
FAIR Data	
a.) Making data findable, including provisions for metadata	*Are the data produced and/or used in the Data Space discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?* Yes *What naming conventions do you follow (if any)?* We use defined data templates with metadata agreed on from a combination of project discussions, workshops, available ontologies and collaborative work to develop terms for missing metadata. We currently support partners involved in data generation on EU projects to prepare their data according to such templates.

[4] https://www.rd-alliance.org/groups/fair-data-maturity-model-wg

	Will search keywords be provided that optimize possibilities for re-use? Yes, that is possible. *What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how* We use defined data templates with metadata agreed on from a combination of project discussions, workshops, available ontologies and collaborative work to develop terms for missing metadata.
b.) Making data openly accessible	*Specify which data will be made openly available by the Data Space* On data associated with projects, partners can choose when to make data open and FAIR and the data will then be available in the Data Space. *Specify how the data will be made available to end-users* Through web interfaces Application Programming Interfaces (APIs) *Specify what methods or software tools are needed to access the data* Web browsers, common programming languages accessing APIs e.g., python *Specify where the data and associated metadata, documentation and code are deposited* Currently in project-based databases that we and partners host. *Specify how access will be provided in case there are any restrictions* Access will be provided when data owners provide open access.
c.) Making data interoperable	*Are the data produced by the Data Space interoperable?* Yes, we work towards that goal, although there are nearly always ongoing issues to resolve between resources and their providers to achieve that goal operationally. *What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?*

	Approaches taken on OpenTox, OpenRiskNet, EU-ToxRisk NanoCommons and aligned with community ontology initiatives.
d.) Increase data re-use	*Specify whether the data produced and/or used in the Data Space is use-able by third parties, in particular after the end of the project.* Yes, data should be usable by Third Parties. *Specify when the data will be made available for re-use (if applicable)* When data owners provide access. *Specify whether a sharing mechanism is in place (if applicable)* Yes, we follow a process supporting both restricted and open access, and status updates by data owners. *Specify how the data will be licenced to permit the widest reuse possible (if applicable)* Creative Commons licenses *Are data quality assurance processes described?* Yes, in project reports. *Specify whether there are policies to reinforce the sharing of data in the Data Space (if applicable)* Yes. *Specify whether a long-term preservation and archiving mechanism is in place (if applicable)* Not currently, but the hope is that EOSC could support that goal.
Data Security	
What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data) (if applicable)?	Local Partner ICT policies.
Ethical Aspects	
Are there any ethical or legal	Yes, and these issues are usually addressed in project data management plans and associated ethical reporting.

issues that can have an impact on data sharing?	

T5.1 UseGalaxy.eu

Contact	Björn Grüning
Reference Data description	
Types of data	Everything that Galaxy can process. This goes from txt files to compressed community specialized ones like netcdf or tiff. Audio and Video data are also possible.
Origin of data	Galaxy admin community.
Scale of data	TB
Scientific Data description	
Types of data	Everything that Galaxy can process. This goes from txt files to compressed community specialized ones like netcdf or tiff. Audio and Video data are also possible.
Origin of data	From the Galaxy user.
Scale of data	PB
FAIR Data	
a.) Making data findable, including provisions for metadata	*Are the data produced and/or used in the Data Space discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?* Only if the user wants to. *What naming conventions do you follow (if any)?* None or community dependent. *Will search keywords be provided that optimize possibilities for re-use?* Inside of Galaxy yes. *What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.* Entire provenance information of the data. EDAM ontology. Abstract CWL.

b.) Making data openly accessible	*Specify which data will be made openly available by the Data Space.* None, if the user does not wish to. *Specify how the data will be made available to end-users.* Shared data will be discoverable via Galaxy. *Specify what methods or software tools are needed to access the data.* Galaxy. *Specify where the data and associated metadata, documentation and code are deposited.* Not provided *Specify how access will be provided in case there are any restrictions.* Not provided
c.) Making data interoperable	*Are the data produced by the Data Space interoperable?* Inside of Galaxy yes. Outside of Galaxy it needs to be exported in standard formats like Compute Objects. *What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?* EDAM.
d.) Increase data re-use	*Specify whether the data produced and/or used in the Data Space is use-able by third parties, in particular after the end of the project.* Yes. *Specify when the data will be made available for re-use (if applicable).* When the Galaxy user decides it. *Specify whether a sharing mechanism is in place (if applicable).* Yes, inside of Galaxy every user can decide to share everything with groups, roles or everyone. *Specify how the data will be licenced to permit the widest reuse possible (if applicable).* Decision made by the Galaxy user.

	Are data quality assurance processes described? No. *Specify whether there are policies to reinforce the sharing of data in the Data Space (if applicable).* No. *Specify whether a long-term preservation and archiving mechanism is in place (if applicable).* In planning phase.
Data Security	
What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data) (if applicable)?	Georedundant backups of data are in place.
Ethical Aspects	
Are there any ethical or legal issues that can have an impact on data sharing?	No entirely clear. As the user decides everything on its own.

T5.1 Virtual Imaging Platform

Contact	Sorina Camarasu Pop
Reference Data description	
Types of data	The Virtual Imaging Platform is a computing platform. Data management is mainly provided to handle inputs and outputs for the computations. We encourage the use of external data management platforms (with which VIP is interoperable) for advanced data management features and long term storage. As such, VIP does not provide reference data beyond a few datasets provided as input examples for the scientific applications available in VIP. Very heterogeneous, depending on each of the available applications. It can be text files (mainly for configuration purposes), but also all kind of images (mainly Nifti, but also Dicom, raw/mhd, etc) and Matlab files.

Origin of data	End users upload it from their personal computers or external data platforms.
Scale of data	1-100 MB most of the times, but sometimes also up to several GB.
Scientific Data description	
Types of data	The Virtual Imaging Platform produces heterogeneous types of data, according to the specificities of each of the multiple applications available. It can be text files, but also all king of images (mainly Nifti, but also raw/mhd, etc), ROOT files and probability maps.
Origin of data	End users upload it from their personal computers or external data platforms.
Scale of data	1-100 MB most of the times, but sometimes also up to several GB.
FAIR Data	
a.) Making data findable, including provisions for metadata	*Are the data produced and/or used in the Data Space discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?* We encourage the use of external data management platforms (with which VIP is interoperable) for advanced data management features, such as the use of metadata and DOIs. *What naming conventions do you follow (if any)?* The Logical File System implemented in VIP, based on the Dirac File Catalog, allows users and groups of users to have their own 'home' folders as in a unix-based system. Beyond this structure, users are free to use their own naming conventions. *Will search keywords be provided that optimize possibilities for re-use?* N.A. *What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how* N.A.
b.) Making data openly accessible	*Specify which data will be made openly available by the Data Space*

	Data belongs to end users (not to VIP). They may decide to make it publicly available either though their group folder in VIP or through external data platforms. *Specify how the data will be made available to end-users* Within VIP, data is made available through the File Transfer tool available in the portal. *Specify what methods or software tools are needed to access the data* Web browser. *Specify where the data and associated metadata, documentation and code are deposited* N/A *Specify how access will be provided in case there are any restrictions* N/A
c.) Making data interoperable	*Are the data produced by the Data Space interoperable?* Not provided *What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?* Not provided
d.) Increase data re-use	*Specify whether the data produced and/or used in the Data Space is use-able by third parties, in particular after the end of the project.* Not provided *Specify when the data will be made available for re-use (if applicable)* Not provided *Specify whether a sharing mechanism is in place (if applicable)* Not provided *Sharing can be done with group members through the group folder or through external platforms (note that data belongs to the user)* Not provided

	Specify how the data will be licenced to permit the widest reuse possible (if applicable) Not provided *Are data quality assurance processes described?* Not provided *Specify whether there are policies to reinforce the sharing of data in the Data Space (if applicable)* Not provided *Specify whether a long-term preservation and archiving mechanism is in place (if applicable)* Not provided
Data Security	
What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data) (if applicable)?	Not provided
Ethical Aspects	
Are there any ethical or legal issues that can have an impact on data sharing?	Yes, some of the data may be medical data (even if psedo-anonymized).

T5.1 WeNMR: A worldwide e-Infrastructure for NMR spectroscopy and Structural biology

Contact	Alexandre Bonvin
Reference Data description	
Types of data	N.A. Only some datasets are provided with tutorials as part of pre-calculated output of WeNMR services
Origin of data	N/A

Scale of data	N/A
Scientific Data description	
Types of data	The various WeNMR services generate mostly structural data in the form of 3D coordinates of biomolecules. 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
FAIR Data	
a.) Making data findable, including provisions for metadata	*Are the data produced and/or used in the Data Space discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?* The generated data belong to the users and are not stored on WeNMR portals *What naming conventions do you follow (if any)?* Coordinate files are typically PDB or mmCIF formatted files (see http://www.wwpdb.org). *Will search keywords be provided that optimize possibilities for re-use?* N.A. *What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how* N.A.
b.) Making data openly accessible	*Specify which data will be made openly available by the Data Space* N.A. *Specify how the data will be made available to end-users* End-users can download the results of their calculations directly from the WeNMR portals, either as 3D coordinate files (e.g. PDB or mmCIF format), or as data archives. Those are only stored for a limited period of time on the portals. *Specify what methods or software tools are needed to access the data*

	Web browser *Specify where the data and associated metadata, documentation and code are deposited* N.A. - this is the responsibility of the end user. Code is maintained by the WeNMR providers using Github. *Specify how access will be provided in case there are any restrictions* N.A.
c.) Making data interoperable	*Are the data produced by the Data Space interoperable?* Yes. The structural data follow well-defined standards from the Protein Data Bank *What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?* See above.
d.) Increase data re-use	*Specify whether the data produced and/or used in the Data Space is use-able by third parties, in particular after the end of the project.* N.A. *Specify when the data will be made available for re-use (if applicable)* N.A. *Specify whether a sharing mechanism is in place (if applicable)* 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 *Specify how the data will be licensed to permit the widest reuse possible (if applicable)* N.A. *Are data quality assurance processes described?* N.A. *Specify whether there are any policies within the Data Space to reinforce the sharing of data (if applicable)*

	End user responsibility - in the field of structural biology, deposition in public databases is often a requirement from scientific journals *Specify whether a long-term preservation and archiving mechanism is in place (if applicable)* N.A. (not done by WeNMR)
Data Security	
What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data) (if applicable)?	N/A
Ethical Aspects	
Are there any ethical or legal issues that can have an impact on data sharing?	N/A

T5.2 OpenCoastS

Contact	Anabela Oliviera
Reference Data description	
Types of data	Inputs: grid files, boundary conditions, parameter files (all in SCHISM formats, in ASCII, binary and netcdf)
Origin of data	Inputs: provided by the users, confidential
Scale of data	Dependent on the user's files resolution
Scientific Data description	
Types of data	Outputs: model results for the several 2D or 3D variables (water levels, velocity, salinity, temperature, all netcdf files)
Origin of data	Outputs: Generated within the service, confidential
Scale of data	(size) of average simulation per day: 2D simulations - 1Gb; 3D simulations - 10 Gb

FAIR Data	
a.) Making data findable, including provisions for metadata	Are the data produced and/or used in the Data Space discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)? No, the data (both inputs and outputs) are confidential What naming conventions do you follow (if any)? The file names are always the same for all runs and are established by the model developers (SCHISM) Will search keywords be provided that optimize possibilities for re-use? No, due to the confidential nature of the data. What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how Not applicable
b.) Making data openly accessible	Specify which data will be made openly available by the Data Space None Specify how the data will be made available to end-users Does not apply to the data. The software is available in https://gitlab.com/opencoasts/eosc-hub Specify what methods or software tools are needed to access the data None Specify where the data and associated metadata, documentation and code are deposited Code and its documentation: https://gitlab.com/opencoasts/eosc-hub Documentation on the service: https://opencoasts.ncg.ingrid.pt/ Specify how access will be provided in case there are any restrictions Regarding the software, it is opensource, so be accessed and re-used. The deployments data is confidential.

c.) Making data interoperable	Are the data produced by the Data Space interoperable? Yes since it follows the standard file formats (netcdf,...). What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable? Usage of model formats are already standard
d.) Increase data re-use	Specify whether the data produced and/or used in the Data Space is use-able by third parties, in particular after the end of the project. 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. Specify when the data will be made available for re-use (if applicable) Input and outputs are property of the users. Ncwms layers data sharing is possible if users are willing to share. Specify whether a sharing mechanism is in place (if applicable) Yes, the code is open source. This will allow everyone to download it at will. Specify how the data will be licenced to permit the widest reuse possible (if applicable) The data is confidential. The code has an Apache Licence v2 Are data quality assurance processes described? Data quality is the responsibility of the user. Code quality is guaranteed under a detailed set of tests, for each release. Specify whether there are policies to reinforce the sharing of data in the Data Space (if applicable) Sharing the software code; Sharing the grids openly in a GitHub repository: https://github.com/LNEC-GTI/OPENCoastS-Grids Specify whether a long-term preservation and archiving mechanism is in place (if applicable) For a short period of time, through EUDAT services. Preservation is done by the users.
Data Security	

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data) (if applicable)?	N/A
Ethical Aspects	
Are there any ethical or legal issues that can have an impact on data sharing?	None for ethical issues. Legal issues may be at stake for some user input files, but the OPENCoastS team may not be aware of that. The use cases data is confidential to address this another user concerns

T5.3 PROMINENCE

Contact	Shaun de Witt
Reference Data description	
Types of data	none
Origin of data	N/A
Scale of data	N/A
Scientific Data description	
Types of data	Modelling data.
Origin of data	Data will be the output of running JOREK simulations and a trained neural network to optimise parametric analysis.
Scale of data	~1 Mo files, ~50TB
FAIR Data	
a.) Making data findable, including provisions for metadata	*Are the data produced and/or used in the Data Space discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?* TBC - In principle we could provide the training data set, which will be smaller set of objects, as a FAIR object (or objects), or we could create a research object with links to workflow and codes and make this available as a RO. *What naming conventions do you follow (if any)?*

	None *Will search keywords be provided that optimize possibilities for re-use?* Yes, but only for the trained neural network *What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how* TBD
b.) Making data openly accessible	*Specify which data will be made openly available by the Data Space* Training data or research object (as described above). Trained NN as a downloadable entity *Specify how the data will be made available to end-users* Download *Specify what methods or software tools are needed to access the data* TBC - if we use zenodo then no tools will be necessary *Specify where the data and associated metadata, documentation and code are deposited* JOREK: Under source control. Source is open but requires authorisation. PROMINENCE: https://github.com/prominence-eosc *Specify how access will be provided in case there are any restrictions* User will need to request access to JOREK through the web page.
c.) Making data interoperable	*Are the data produced by the Data Space interoperable?* TBC - what format does JOREK output in? *What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?* None
d.) Increase data re-use	*Specify whether the data produced and/or used in the Data Space is use-able by third parties, in particular after the end of the project.*

	Training Data/Research object - yes, but not with ease *Specify when the data will be made available for re-use (if applicable)* See above *Specify whether a sharing mechanism is in place (if applicable)* No *Specify how the data will be licenced to permit the widest reuse possible (if applicable)* CC-BY-SA 4.0 *Are data quality assurance processes described?* No *Specify whether there are policies to reinforce the sharing of data in the Data Space (if applicable)* No *Specify whether a long-term preservation and archiving mechanism is in place (if applicable)* Intend to use Zenodo
Data Security	
What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data) (if applicable)?	N/A
Ethical Aspects	
Are there any ethical or legal issues that can have an impact on data sharing?	N/A

T5.3 LOFAR Science Products

Contact	Agnes Mika

Reference Data description	
Types of data	Observation data in raw formats (visibilities and time-series)
Origin of data	Observation data is generated by the LOFAR instrument, and the processing cluster managed by the ILT Observatory.
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 50 petabyte in the LTA for several millions of data-products.
Scientific Data description	
Types of data	Processed data (calibrated data, radio astronomical images, pulsar profiles, etc.)
Origin of data	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. The LOFAR Data Space in EGI-ACE will generate calibrated and other scientific data products from the reference data on EOSC compute facilities co-located with the data archive, both as a centrally managed activity to add value to the data archive, and as a user-requested service for specific science projects.
Scale of data	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.
FAIR Data	
a.) Making data findable, including provisions for metadata	*Are the data produced and/or used in the Data Space discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?* 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, the FITS format and the HDF5 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. For LOFAR, a catalogue is maintained of all data in the archive that contains rich metadata to allow data discovery. Unique identifiers have from the start ben generated and are ensured by ASTRON to be unique and persistent. In the

	DICE project, Handle-type PID's will be generated for the science level data products and collections. *What naming conventions do you follow (if any)?* Naming conventions for LOFAR data products are defined in the Interface Control Documents. *Will search keywords be provided that optimize possibilities for re-use?* Keywords indicating instrument, science type, and type of data will be provided for data publicly released and made available through a Virtual Observatory (IVOA compliant) service and/or a general purpose data repository. What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how The metadata in the catalogue for the LTA is filled using XML documents that comply to a custom schema. An effort is underway to publish metadata for publicly released data products in accordance with IVOA standards, notably the obscure model. When data is (also) registered in a general-purpose science repository, the supported schema's will be built on, extended with a schema for (radio-) astronomical metadata, to be mappable to the IVOA schema in use.
b.) Making data openly accessible	*Specify which data will be made openly available by the Data Space* All data made available by the Data Space in accordance with the LOFAR Data Policy which states that all data owned by all ILT data will (eventually) be made publicly available. A (default 1 year) proprietary period may apply for data requested in a science project proposal and specifically generated for it. *Specify how the data will be made available to end-users* The data is ingested into the LOFAR archive which can be queried openly although access limitations (in particular exclusive access for an initial period) for retrieving data may apply. Publicly released data is registered in a public Virtual Observatory compliant service and/or in a general-purpose Science data repository. *Specify what methods or software tools are needed to access the data* Data that is available for immediate access can be retrieved using standard web protocols and optionally through special high-performance tools (data in the LOFAR LTA can currently also be retrieved using the SRM and GridFTP

	protocols). Data may be stored on a nearline medium (tape), in which case a web-based staging service is to be used before data is available for retrieval. A special purpose API exists for interacting with the staging service. *Specify where the data and associated metadata, documentation and code are deposited* Data in the LOFAR LTA is deposited in a dCche storage environment at one of the (currently three: SURF, FZJ, PSNC) research data centers that support the LTA. Metadata is deposited in the LTA catalogue which is built on a high performance database hosted by the University of Groningen. Metadata can further be deposited in an ASTRON provided Virtual Observatory service and in a SURF hosted general purpose science data repository. *Specify how access will be provided in case there are any restrictions* For all portals where data can be discovered, in case direct/open access is not allowed, either because an exclusive access policy still applies, or the data needs to be staged from tape first, a link is provided to access the staging service which requires a user to log in for identity and authorization checking. Users that do not yet have an identity registered with the LOFAR organization will be required to create one.
c.) Making data interoperable	*Are the data produced by the Data Space interoperable?* Data and metadata are provided in standard formats and include references to related data and processes. *What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?* For data, we will provide it in one of the standard data formats (primarily FITS, CASA/Measurement Set, and HDF5, and provide ICD's that define the specific structure and provided metadata attributes. Metadata will be provided following the IVOA standards and any provided by the general purpose science data repository, e.g. Dublin Core, DataCite, etc.
d.) Increase data re-use	*Specify whether the data produced and/or used in the Data Space is use-able by third parties, in particular after the end of the project.* Yes. Data will (eventually) be public and available for use by anybody *Specify when the data will be made available for re-use (if applicable)*

	In accordance with the LOFAR DATA Policy: Depending if a proprietary period applies or not: In principle one year after the final data ingest for a science project has occurred, or immediately upon publication. *Specify whether a sharing mechanism is in place (if applicable)* Users will be able to share data through Persistent Identifiers and/or unique persistent data paths. Curated data collections are registered in public repositories (IVOA and/or general purpose) which are harvested by indexing services and allow users to query and access the data through standard tools. *Specify how the data will be licenced to permit the widest reuse possible (if applicable)* It is under consideration to associate a license with all published/released data. This will likely be a type that allows free re-use under condition of including credits (attribution). *Are data quality assurance processes described?* In general, the data quality assurance of released data collections is described in accompanying documentation for which a reference is than provided. *Specify whether there are policies to reinforce the sharing of data in the Data Space (if applicable)* 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 can be queried publicly and the Data Space aims to improve public and open sharing of data by building on EOSC-based FAIR data services. Specify whether a long-term preservation and archiving mechanism is in place (if applicable) The LOFAR Long Term Archive (LTA) provides a centrally managed data archive, ensuring long term preservation of observation data. The ASTRON Science Data Center supports the LOFAR community in accessing and processing the data.
Data Security	
What provisions are in place for data security (including data recovery as well as secure storage and transfer of	Given the size of the data, in general a single replica is maintained in the archive. Since the full provenance is described, it is possible to regenerate data, including re-observing a field. For unique data that cannot be regenerated, it is considered to store multiple replicas. The Data Space does not store sensitive data.

sensitive data) (if applicable)?	
Ethical Aspects	
Are there any ethical or legal issues that can have an impact on data sharing?	No

T5.4 Disaster mitigation and agriculture

Contact	Eric Yen
Reference Data description	
Types of data	Observation data for target hazard events: global model data; gridded data in GRIB format; satellite data; radar data; etc. Geographical data of impact areas of target hazard events: topographical data, land use, bathymetry, etc. Outcomes of a case study: images, videos, publications, documents, simulation results, etc.
Origin of data	Observation data comes from agencies or reliable open sources such as NCEP, NASA, EMCWF and local weather bureau or related government agencies.
Scale of data	Observation and static data scale are O(10-100MB) for each case study. Output of a simulation results from a case study ranges from 10GB to 1TB.
Scientific Data description	
Types of data	Simulation results: 3D gridded data, images, video, visualization data. Case study: sets of data including observation data and geographic data required for simulation, results of simulations, publications, and related documents.
Origin of data	Simulation results are generated by users for each case study. Observation data and environmental data are the input. Whole analysis of a case study usually is published as an academic paper.
Scale of data	Depend on temporal and spatial resolution:

	Weather simulation by WRF (per simulation): Simulation result is O(TB) per simulation. Tsunami and Storm Surge (per simulation): Simulation result is O(GB).
FAIR Data	
a.) Making data findable, including provisions for metadata	*Are the data produced and/or used in the Data Space discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?* 1. Metadata: Darwin Core metadata scheme is used as the generic & exchange metadata scheme for all types of data in this data space. Domain specific metadata standard will be applied if it is proposed by the user community. 2. Data format: GRIB/GRIB2; NetCDF, HDF5, and open standard format for images and videos 3. Metadata management: Depositar (Research Data Management system developed by Academia Sinica) will be provided as one of the metadata management solution. Will investigate the interoperability with EGI/EGI-ACE solutions such as OpenRDM. 4. Data identifier will be embedded for each case study and metadata. *What naming conventions do you follow (if any)?* Based on recommendations of WMO and EUMETNET, etc. *Will search keywords be provided that optimize possibilities for re-use?* Yes. Will be provided by both web site search and metadata management system. *What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how* Two levels of metadata will be provided: case study and individual data set. Structure of case study data will be based on standards such as RDF or OWL. Individual data set includes geo-spatial data and attribute data. The former one covers maps (raster and vector). The latter contains documents, publications, images (e.g., satellite, figures, etc.), videos, weather data, radar data, etc.
b.) Making data openly accessible	*Specify which data will be made openly available by the Data Space*

	In general, all data in a case study will be opened. As most observation data is owned by government agencies, open access to those observation data is up to those agencies. When part of data sets of a case study is not accessible, all other data associated with the case study will be released first. *Specify how the data will be made available to end-users* By web services through the DMCC-Ag website in case study-oriented model. Those data will be searchable by EGI-ACE services as well. *Specify what methods or software tools are needed to access the data* Accessible by https by default *Specify where the data and associated metadata, documentation and code are deposited* ASGC will host the data services and maintain the data access and backup in the long run. *Specify how access will be provided in case there are any restrictions* Restriction reasons will be explained online. If request for approval of access to the data owner is acceptable, the contact point will be disclosed at the same time.
c.) Making data interoperable	*Are the data produced by the Data Space interoperable?* Yes. All data formats used are in open standards. *What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?* Domain recommended vocabularies, ontologies or thesauri with explicit references will be adapted if they are in place. For example, communities such as meteorology, climate, remote sensing, earth science, oceanology, environmental changes, etc.
d.) Increase data re-use	*Specify whether the data produced and/or used in the Data Space is use-able by third parties, in particular after the end of the project.* 1. Simulation events of open accessible case studies are reproducible. 2. Patterns and correlations in high resolution 3D gridded data could be explored. 3. Users could implement their own simulations based on the simulation portals referring to the case study materials.

Specify when the data will be made available for re-use (if applicable)

1. All data services and share materials 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.
2. When a case study is finished based on approval of all involved partners, all materials and the simulation workflow will be released according to FAIR principles in three months.
3. For example, data of tsunami case studies will be available through the iCOMCOT simulation portal at https://icomcot.twgrid.org. Data of meteorological hazard case studies will be provisioned throughout the WRF portal (which is now under reconstruction)

Specify whether a sharing mechanism is in place (if applicable)

According to AAI, user could access to shared materials by HTTP or SFTP (non-web environment).

Specify how the data will be licenced to permit the widest reuse possible (if applicable)

CC-BY 4.0 is the default scheme of license.

Are data quality assurance processes described?

For simulations of case studies, quality assurance of data is ensured through demonstrated functionality and performance over the common platforms, tools and services. Data quality control is verified through compliance with recognized community standards.

Specify whether there are policies to reinforce the sharing of data in the Data Space (if applicable)

Restriction of data access will be checked routinely on an annual basis or when the requests are more than 3 times in the same year. Protocol and services to access shared materials will be evolved along with technology advancement.

Specify whether a long-term preservation and archiving mechanism is in place (if applicable)

ASGC is in 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 center 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.
Data Security	
What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data) (if applicable)?	In addition to data policy, integration of AAI, replications on different storages, ASGC security operation center, long-term backup as well as default https-enabled encrypted data transfer will be implemented.
Ethical Aspects	
Are there any ethical or legal issues that can have an impact on data sharing?	No ethical or legal issues of shared data and case studies.

T5.4 EMSO ERIC data services

Contact	Ivan Rodero
Reference Data description	
Types of data	Harmonized data from EMSO ERIC Regional Facilities and test sites.
Origin of data	EMSO ERIC is a distributed RI based on an integrated system of in situ Eulerian observatories. Its mission is the long-term observation of the deep seafloor and the water column by means of fixed-point multi-sensor platforms deployed in environmentally relevant sites of the European seas.
Scale of data	Multi-node parameters spanning over decades of observations.
Scientific Data description	
Types of data	EMSO's main scientific objective is to understand global environmental processes and stimulate the development of new technologies and knowledge.
Origin of data	It comprises Regional Facilities and test sites distributed from the North Atlantic, Nordic Seas through the Mediterranean to the anoxic Black Sea.

Scale of data	Multi-node parameters spanning over decades of observations.
FAIR Data	
a.) Making data findable, including provisions for metadata	*Are the data produced and/or used in the Data Space discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?* Yes. *What naming conventions do you follow (if any)?* OceanSites specifications and CF standards. *Will search keywords be provided that optimize possibilities for re-use?* Use of OceanSites and vocabularies. *What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how* Harmonized metadata across EMSO ERIC facilities following a common approach.
b.) Making data openly accessible	*Specify which data will be made openly available by the Data Space* Harmonized EMSO ERIC data in different formats such as NetCDF and CSV. *Specify how the data will be made available to end-users* Via web portal, ERDDAP, APIs and tools associated with the data management back-end. *Specify what methods or software tools are needed to access the data* Direct access via the EMSO ERIC data portal and services associated with it. REST queries for API-based access. *Specify where the data and associated metadata, documentation and code are deposited* They are initially deposited at the system established at the facility level (e.g., National Data Center, repository, etc.) and later on the EMSO ERIC data portal and configuration management system (Gitlab). *Specify how access will be provided in case there are any restrictions*

	There are no restrictions foreseen this point.
c.) Making data interoperable	*Are the data produced by the Data Space interoperable?* Yes, they follow standard formats and specifications, which are interoperable. *What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?* OceanSites specifications and CF standards.
d.) Increase data re-use	*Specify whether the data produced and/or used in the Data Space is use-able by third parties, in particular after the end of the project.* Published results will be re-usable by third parties. *Specify when the data will be made available for re-use (if applicable)* N/A *Specify whether a sharing mechanism is in place (if applicable)* Data can be shared via direct link or PIDs. *Specify how the data will be licenced to permit the widest reuse possible (if applicable)* EMSO ERIC facilities use open licenses. EMSO ERIC is currently working on a harmonized policy. *Are data quality assurance processes described?* Yes, when available at the facility level. *Specify whether there are policies to reinforce the sharing of data in the Data Space (if applicable)* N/A Specify whether a long-term preservation and archiving mechanism is in place (if applicable) N/A
Data Security	
What provisions are in place for data security (including data	Redundant deployment for fail-over, business continuity and disaster recovery. Use of cloud-based firewall and server-level security in place following to best practices. Use of authentication and authorization.

recovery as well as secure storage and transfer of sensitive data) (if applicable)?	
Ethical Aspects	
Are there any ethical or legal issues that can have an impact on data sharing?	N/A

T5.4 ENES Data Space

Contact	Fabrizio Antonio
Reference Data description	
Types of data	N/A
Origin of data	N/A
Scale of data	N/A
Scientific Data description	
Types of data	Input data: gridded “community-relevant” data in NetCDF format including the most downloaded datasets/variables (CMIP6, CORDEX) Generated output: gridded data or other kinds of output such as diagrams, scripts, tables, maps and graphs User scripts: Jupyter notebooks, Python scripts, configuration files, workflow documents
Origin of data	Community-provided data sources: e.g., gridded multi-variable climate data from the ESGF/CMIP data archive. User-provided data sources: aside from community sources, users may provide data out of input data collections or DataHub and their curation policies apply.
Scale of data	Input data may be in the order of hundreds of TB. A first selection of data collections will be downloaded from the ESGF/CMIP data sources and made available to users (around 100TB); new datasets could be added incrementally to address users’ needs.

FAIR Data	
a.) Making data findable, including provisions for metadata	*Are the data produced and/or used in the Data Space discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?* Data and metadata conform to CMIP community agreements, thus enabling search and discovery, through a dedicated science portal deployed on the EGI infrastructure or the ESGF portals. In addition, the persistent identification of ESGF-published datasets is accomplished through persistent identifiers (PIDs) or DOI's, which allow for persistent references to data including versioning and replica information. *What naming conventions do you follow (if any)?* NetCDF CF conventions, CMIP community agreements: standardized variables, file-level metadata, domain metadata, CMIP6 Data Reference Syntax (DRS) and controlled vocabularies (CVs). *Will search keywords be provided that optimize possibilities for re-use?* Data search will be allowed on the dedicated science portal through search keywords based on CF compliant metadata. *What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how* N.A.
b.) Making data openly accessible	*Specify which data will be made openly available by the Data Space* “community-relevant” data including the most downloaded datasets/variables both at global and regional scales (CMIP6, CORDEX). *Specify how the data will be made available to end-users* The most relevant datasets/variables from CMIP6 and CORDEX will be downloaded from the ESGF/CMIP data sources by using the synda command line tool and made available to users enabling data analysis and visualization activities. *Specify what methods or software tools are needed to access the data* Users can access an Open (data) Science environment through EGI Check-in and exploit a core set of software

	modules for running data manipulation, analysis, and visualization. *Specify where the data and associated metadata, documentation and code are deposited* Documentation and training material concerning the ENES Data Space will be stored on a dedicated Github repository. *Specify how access will be provided in case there are any restrictions* N.A.
c.) Making data interoperable	*Are the data produced by the Data Space interoperable?* The data space will mainly rely on data collections from ESGF/CMIP in NetCDF format, which is the standard de facto in the field of Earth System Modelling and climate sciences. This guarantees that the files are interoperable between a wide set of tools. Moreover, metadata-level conventions (CF-conventions) also ensure that the file content can be easily recognized by climate-based tools. *What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?* NetCDF CF conventions, CMIP community agreements, CMIP6 Data Reference Syntax (DRS) and controlled vocabularies (CVs).
d.) Increase data re-use	*Specify whether the data produced and/or used in the Data Space is use-able by third parties, in particular after the end of the project.* Data in the ENES Data Space is publicly available from the ESGF data archive. Specific terms of use apply to each collection of data. Data stored within the data space is made available to any user upon accessing the service. Output products are not directly archived, but they may be shared per user's discretion. *Specify when the data will be made available for re-use (if applicable)* Data is available for reuse according to the specific terms of use. *Specify whether a sharing mechanism is in place (if applicable)* The ENES Data Space will integrate Onedata enabling data sharing both inside and outside the climate community. *Specify how the data will be licenced to permit the widest reuse possible (if applicable)*

	Specific terms of use apply to each collection of data; for example CMIP6 data are distributed under some type of creative common license (https://pcmdi.llnl.gov/CMIP6/TermsOfUse/TermsOfUse6-1.html); CORDEX are also freely accessible but could be restricted to non-commercial use (http://is-enes-data.github.io/cordex_terms_of_use.pdf). *Are data quality assurance processes described?* N.A. *Specify whether there are policies to reinforce the sharing of data in the Data Space (if applicable)* N.A. *Specify whether a long-term preservation and archiving mechanism is in place (if applicable)* The ENES Data Space 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.
Data Security	
What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data) (if applicable)?	N.A.
Ethical Aspects	
Are there any ethical or legal issues that can have an impact on data sharing?	N.A.

T5.4 GBIF Cloud data space

Contact	Rui Figueira
Reference Data description	
Types of data	Biodiversity data containing species occurrence records.

Origin of data	Data is uploaded by publishers in Darwin Core Archive format (compressed zip file with csv tables with data, a xml file csv table structure, and a xml with metadata). Other data includes shapefiles with geographic information for context.
Scale of data	Estimated values: • text: 25 GB (50 M records) • images: 75GB (450 K images) • geographic information: 20 GB
Scientific Data description	
Types of data	Data is reinterpreted during ingestion in order to enable facets and search services. Raw data is kept in parallel with the reinterpreted data, and is made available to final users.
Origin of data	Reinterpreted data of biodiversity occurrences, include: • scientific names matched to the taxonomic backbone taxonomy • data quality issues flagged, up to 90 properties tested during data ingestion. Flags are: failed, warning, passed, missing; unchecked • indexed data to create facets for user search. Filters are possible for 44 properties. More filters can be created from custom lists of properties provided by users.
Scale of data	Estimated values: • text: 25 GB (50 M records) • images: 75GB (450 K images) • geographic information: 20 GB
FAIR Data	
a.) Making data findable, including provisions for metadata	*Are the data produced and/or used in the Data Space discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?* Data records use Darwin Core standard format. Each record receives a unique identifier (at dataset level, many times a GUID). Additionally, each record receives a unique GBIFID at the time of publication through GBIF.org Dataset metadata uses EML metadata standard format. Each dataset receives a DOI What naming conventions do you follow (if any)? The data standard Darwin Core uses controlled vocabularies for several terms of the standard.

	Will search keywords be provided that optimize possibilities for re-use? Taxon names, geographic names, type of dataset (occurrence, sampling event, checklist) *What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how* Each dataset is described by its metadata, using the EML metadata standard
b.) Making data openly accessible	*Specify which data will be made openly available by the Data Space* Occurrence species data *Specify how the data will be made available to end-users* Accessible through the data platform, file download, webservices (API) *Specify what methods or software tools are needed to access the data* Data can be accessed directly through the data platform, but it is possible to use the API service to access data using R or other tool *Specify where the data and associated metadata, documentation and code are deposited* GBIF.org for global access. datos.gbif.es and dados.gbif.pt for national data portals. *Specify how access will be provided in case there are any restrictions* NA
c.) Making data interoperable	*Are the data produced by the Data Space interoperable?* Yes *What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?* Darwin Core standard (https://dwc.tdwg.org/terms/) and Darwin Core extensions (https://tools.gbif.org/dwca-validator/extensions.do). DwC extensions are part of the DwC-A file format implementation, and their terms may or may not be DwC terms. their approval and registration are not part of the TDWG procedures

	EML metadata standard (https://eml.ecoinformatics.org/), with GBIF profile (https://github.com/gbif/ipt/wiki/GMPHowToGuide)
d.) Increase data re-use	*Specify whether the data produced and/or used in the Data Space is use-able by third parties, in particular after the end of the project.* All data can be reused by third parties, including after the end of the project. *Specify when the data will be made available for re-use (if applicable)* All data is made available for re-use by open and free means to all through the internet. *Specify whether a sharing mechanism is in place (if applicable)* Any institution holding biodiversity data can become a data publisher, by agreeing with the Data Publisher agreement (https://www.gbif.org/terms/data-publisher). The application form to become a data publisher is available at https://www.gbif.org/become-a-publisher. Applications by data publishers are endorsed by its national GBIF node. Following open discussion on the Community Forum, explore preserving periodic copies of GBIF-mediated data on open public and research cloud infrastructures to both ensure persistence and promote wider and easier use of GBIF; develop recommendations and tools to support best-practice citation of GBIF-mediated data accessed through external cloud environments. *Specify how the data will be licenced to permit the widest reuse possible (if applicable)* Datasets published through GBIF must adopt one of the following licenses: CC0, CC-BY, CC-BY-NC. It is recommended the adoption of one of the first two. The license adopted is a decision of the data publisher. With each data download it is included a recommended citation which includes a DOI specific to this download. GBIF internally manages links of DOIs of downloads to DOIs of datasets, so that each original dataset's contributions to downloads is acknowledged. In the page of each original dataset, each publisher can see metrics of data use, including the number of citations in scientific publications for his dataset (see example https://www.gbif.org/dataset/231c5bcf-1b56-4905-a398-6d0e18f6de1a/activity). *Are data quality assurance processes described?* Best practices on data quality are available to data publishers as a manual on data quality principles manual (https://doi.org/10.15468/doc.jrgg-a190) and a data cleaning

	manual (http://www.gbif.org/document/80528). Training workshops on data quality is regularly provided to data publishers (e.g https://www.gbif.es/talleres/ii-taller-gbifes-online-calidad-bases-datos-biodiversidad/ , https://www.gbif.pt/node/525). During data ingestion, more than 90 parameters are verified for quality issues. *Specify whether there are policies to reinforce the sharing of data in the Data Space (if applicable)* No policies, as the current ones enable full sharing of data. *Specify whether a long-term preservation and archiving mechanism is in place (if applicable)* N/A
Data Security	
What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data) (if applicable)?	Occurrence data about protected or endangered species is considered sensitive, especially about the specific location. Other sensitive data concerns personal data of observers or collectors. GBIF provides a guide on how to generalize and document sensitive data: https://doi.org/10.15468/doc-5jp4-5g10. These good practices are recommended to data publishers.
Ethical Aspects	
Are there any ethical or legal issues that can have an impact on data sharing?	As mentioned in Data Security, a guide on how to publish sensitive data is recommended to data publishers: https://doi.org/10.15468/doc-5jp4-5g10

T5.4 SeaDataNet WebOcean Data Analysis

Contact	Dick M.A. Schaap
Reference Data description	
Types of data	As part of the use case a data space with a volume of established and major marine data collections will be provided which users can use as input for further analyses. These public data collections can be considered as a 'treasure box'. The data collections concern: • SeaDataNet: Harmonised marine data collections for Temperature and Salinity

	• EMODnet Chemistry: Harmonised marine data collections for eutrophication, contaminants, and marine litter • Euro-ARGO: Collections of data from Argo floats
Origin of data	Original observation data sets have been collected in-situ by hundreds of organisations and then gathered and stored by data centres, which are nodes in the SeaDataNet pan-European infrastructure for marine data management. Then subsets have been made for further processing and harmonization by groups of experts in SeaDataNet respectively EMODnet Chemistry to generate these harmonized aggregated data collections as products for selected parameters. In the case of Euro-ARGO all Argo data sets have been gathered and validated and processed by experts and made available as validated data collections by the ARGO GDAC, of which one is hosted at IFREMER (France) as part of the Euro-Argo ERIC
Scale of data	The data collections cover several decades and are global for SeaDataNet and Argo, and European seas for EMODnet Chemistry
Scientific Data description	
Types of data	The public provided data collections, see above, will serve as input for the use case, whereby users can make use of the analytical functionalities of the WebODV application to make subsets of the data for selected date, place, parameter, organization, etc and to make graphics of analytical processes. So the WebODV output will be CSV or NetCDF files with subsets of data AND graphs.
Origin of data	The output data sets and graphs are products from the subsetting or analytical processing of the input data. In addition, users might include their own marine data sets as extra input.
Scale of data	Global, many decades, multiple parameters. See reference data above
FAIR Data	
a.) Making data findable, including provisions for metadata	*Are the data produced and/or used in the Data Space discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?* The reference data collections have DOIs and landing pages. Within the data collections all data are accompanied with standard metadata descriptions, following SeaDataNet CDI standards (SeaDataNet and EMODnet Chemistry) and

	Argo standards (see Argo). Therefore, the data sets resulting from the WebODV analyses and extractions will also contain those metadata per individual data set in the collections. However, it is up to the WebODV users to apply for DOIs for their results, including editing landing pages *What naming conventions do you follow (if any)?* SeaDataNet standards and vocabularies and CF standards *Will search keywords be provided that optimize possibilities for re-use?* The use of SeaDataNet vocabularies and CF standards provides keywords for searches. *What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.* The users can make metadata descriptions for their resulting data products, while going for a DOI and data publishing. This is not included sec within the WebODV use case, but can be done using common facilities in EOSC or Marine domain. Depends largely on the scientific excellence of the results whether the user wants to publish as such.
b.) Making data openly accessible	*Specify which data will be made openly available by the Data Space.* Users of WebODV will store the analysis results and subsets in their private data space. Then they can decide to publish, using common facilities in EOSC or Marine domain. Or decide to keep the results private. *Specify how the data will be made available to end-users.* Users can decide to publish their webODV results. This could be e.g. in SEANOE, which is the data publishing service, operated by SeaDataNet. Or they can decide to make use of such data publishing and citing services, provided by EOSC, arranged via EGI-ACE. *Specify what methods or software tools are needed to access the data.* If users have published their WebODV results, then these can be found in SEANOE or EOSC by their regular services. *Specify where the data and associated metadata, documentation and code are deposited. Users will store their results and documentation first in their private data space.* If users have decided to published their WebODV results, then these can be found in SEANOE or EOSC, including documentation as required by these external services.

	Specify how access will be provided in case there are any restrictions. There are no restrictions foreseen of WebODV results.
c.) Making data interoperable	*Are the data produced by the Data Space interoperable?* Yes, as webODV has a number of output formats which are standard and interoperable. *What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?* SeaDataNet and CF standards
d.) Increase data re-use	*Specify whether the data produced and/or used in the Data Space is use-able by third parties, in particular after the end of the project.* Published results will be re-usable by third parties. *Specify when the data will be made available for re-use (if applicable).* Is depending on choices and analyses made by the WebODV users. *Specify whether a sharing mechanism is in place (if applicable).* Once published in official citation services (see above) sharing is made possible. *Specify how the data will be licenced to permit the widest reuse possible (if applicable).* CC BY 4.0 is most appropriate *Are data quality assurance processes described?* If users publishes results, it is recommended to document quality assurance measures to demonstrate scientific excellence. *Specify whether there are policies to reinforce the sharing of data in the Data Space (if applicable).* No, choice of users to publish or not *Specify whether a long-term preservation and archiving mechanism is in place (if applicable).* No is not part of the use case, but can be arranged by users by publishing in SEANOE or EOSC service

Data Security	
What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data) (if applicable)?	Back-up measures and security for unwanted access of private use space need to be arranged with the cloud provider
Ethical Aspects	
Are there any ethical or legal issues that can have an impact on data sharing?	Not applicable

T5.5 Operas Metrics service and Certification service

Contact	Yoann Moranville
Reference Data description	
Types of data	Metrics and Altmetrics data - retrieved from partners (e.g. Stats directly from website of a partner) + other sources (e.g. Crossref eventdata)
Origin of data	Partners of the service (publishers) + external sources for Altmetrics (e.g. Crossref, Hypothes.is, etc.)
Scale of data	Not provided
Scientific Data description	
Types of data	Only formatting (conversion) of the data is done, no other alteration
Origin of data	Not provided
Scale of data	Not provided
FAIR Data	
a.) Making data findable, including provisions for metadata	*Are the data produced and/or used in the Data Space discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g.*

	persistent and unique identifiers such as Digital Object Identifiers)? No, the events in the database have a unique identifier but only internal. However, all data is only accessible in relation to a DOI (of a monograph) *What naming conventions do you follow (if any)?* A data format was created, but no naming convention *Will search keywords be provided that optimize possibilities for re-use?* No *What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how* Not created, we only reuse and convert into our own data model
b.) Making data openly accessible	*Specify which data will be made openly available by the Data Space* All data is openly available, apart from login credentials. Data is "event (or events) of Metrics or Altmetrics". *Specify how the data will be made available to end-users* Via a public read-only access API (and private write access API for partners to upload their data) The public API is also used by a widget created that can be embedded in any websites *Specify what methods or software tools are needed to access the data* Data from the API is in JSON, so any JSON reader or directly manually via a browser, cURL, etc... *Specify where the data and associated metadata, documentation and code are deposited* Data: https://metrics-api.operas-eu.org Documentation: https://metrics.operas-eu.org Code: Different repositories under https://github.com/hirmeos/ *Specify how access will be provided in case there are any restrictions* For read access, all open

	For write access, only partners providing data have login possibilities to upload their own data
c.) Making data interoperable	*Are the data produced by the Data Space interoperable?* No, but it could become - again the data is not produced by us, but aggregated into a single metadata format *What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?* None
d.) Increase data re-use	*Specify whether the data produced and/or used in the Data Space is use-able by third parties, in particular after the end of the project.* Yes, the data is "use-able" by third parties *Specify when the data will be made available for re-use (if applicable)* Already available *Specify whether a sharing mechanism is in place (if applicable)* What is a "sharing mechanism" here? We don't really share, we provide as open access. *Specify how the data will be licenced to permit the widest reuse possible (if applicable)* All open access *Are data quality assurance processes described?* No *Specify whether there are policies to reinforce the sharing of data in the Data Space (if applicable)* No policies for providing data, API is opened to all *Specify whether a long-term preservation and archiving mechanism is in place (if applicable)* No long-term preservation, data is backed up but not more
Data Security	
What provisions are in place for data security	No sensitive data

(including data recovery as well as secure storage and transfer of sensitive data) (if applicable)?	Data recovery from backups (how often do we rotate the backups is to be confirmed)
Ethical Aspects	
Are there any ethical or legal issues that can have an impact on data sharing?	Unlikely, no

2.5 WP7 Service Delivery and Planning

Contact	Alessandro Paolini
Data description	

Types of data	• The EGI Foundation SMS supporting the project, protected and stored in Confluence and other tools:
	  ○ Processes, procedures, policies and people associated with maintaining these.
	  ○ Services, service endpoints, and service contacts
	  ○ Incidents and Service requests created as GGUS tickets
	  ○ Requests for Change as Jira tickets
	  ○ Monitoring and service performance reports
	  ○ OLAs/UAs/SLAs for service delivery
	• Deliverables
	• Jira tickets covering different aspects of the SMS and project
	• Names and email addresses of people with effort in WP7
	• Project meeting minutes in Indico
	• HTC, Storage, and Cloud accounting data
	• Documentation and guidelines
	• Online surveys on specific topics

Origin of data	Created as a result of work done within the project
Scale of data	< 1GB

Standards and metadata	plain text, pdf, MS Word, MS Excel, MS Powerpoint, Google Docs, online access as provided by the Service Management Tools
Data sharing	
Target groups	• Participants within the project and collaborators • members of EGI Federation
Scientific Impact	Not applicable
Approach to sharing	Shared within the project consortium to support work
Archiving and preservation	Once the project is finished, all the WP7 information will be preserved by EGI for at least 5 years.