Deliverable D1.2

PROJECT DATA MANAGEMENT PLAN (DMP)

Version 1.1

Document Information

Document version	1.1
Contract Number	823844
Project Website	https://cheese-coe.eu/
Contractual Deadline	30/04/2019 (M6)
Dissemination Level	PU
Deliverable nature	R
Author	Arnau Folch and Nadia Tonello (BSC)
Contributors	
Reviewers	Michael Badar (TUM)

The ChEESE project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No 823844

Change Log

Version	Description of Change
V1.0	Initial draft for internal review
V1.1	Final version

Index

Executive Summary
This deliverable presents the data management plan (DMP) of the ChEESE project, which describes the data management life-cycle for all codes and datasets to be collected, processed and/or generated during the lifetime of the project. The document describes the contents and organization of the DMP, considering that the actual data (and metadata) description will be furnished as a "live" Annex that will be periodically updated during the development of the project. Specifically, this deliverable describes:

- How the different types of datasets that will be generated, collected, and processed during the project will be managed during and after it. This affects mainly the ChEESE Pilot Demonstrators (PDs) and, to a lesser extent, other research activities.
- Which methodologies and which standards (if any) will be applied to manage each of the ChEESE datasets.
- How the datasets will be stored and handled during the lifetime of the project and after its conclusion, as well as how the datasets will be made (openly) accessible.

1. Introduction

The ChEESE project is part of the H2020 Open Research Data Pilot (ORD pilot), aimed at improving and maximizing access and re-use of research data, as well as at taking into account the need to balance openness and protection of scientific information, commercialization and Intellectual Property Rights (IPR), privacy concerns, security, and related data management and preservation questions. This DMP describes how data (in its broadest sense) will be managed during the project, but it is not a final closed document. The DMP described in this report (D2.1) will be updated over the course of the project to account for generation/acquisition of new datasets, implementation of consortium policies, and/or other external factors.

This document follows the Horizon 2020 "FAIR" DMP template and follows the FAIR data guiding principles; i.e. that data must be Findable, Accessible, Interoperable, and Re-usable.

2. Data Summary and structure of the ChEESE DMP Annex

As stated in the GA, ChEESE will prepare 10 Solid Earth community flagship European codes for the upcoming pre-Exascale and Exascale supercomputers and will further develop 12 Pilot Demonstrators (PDs) for scientific problems requiring of Exascale computing on near real-time seismic simulations and full wave-form inversion, ensemble-based volcanic ash dispersal, faster than real-time tsunami simulations, and physics-based hazard assessments. PDs can be considered as small-scale demonstrators for optimizing and testing codes on Exascale hardware prototypes and for addressing the Exascale challenges. It is foreseen that around 8 out of 12 PDs will further increase their Technology Readiness Levels (TRLs) enough to enable services on urgent computing, early warning, and hazard assessment.

The flagship codes (including satellite software components, mini apps and other tools derived from codes) and the PDs will be the main sources of project data, that will include:

- The flagship source codes themselves.
- Datasets generated to evaluate code performance.
- Experimental datasets to be used by PDs.
- Datasets generated from the execution of models.

For this reason, the ChEESE DMP has been organized around a (dynamic) spreadsheet annex document, with one dedicated sheet for each of the PDs (i.e. 12 sheets in total). As an illustrative example, Figure 1 shows the sheet for PD1 (urgent seismic simulations). As observed, the PD1 sheet contains a table that lists all the datasets needed/produced by this PD, as well as the different flagship codes potentially involved. In turn, each of these codes and datasets are linked to other sheets that fully define them (Figures 2 and 3; Sections 2.1 and 2.2), including the related FAIR metrics (Section 3).

	Pilot 1	**Urgent Seismic simulations**							
	Codes	ExaHyPE	Salvus	SPECFEM3D	SeisSol				
	Area	CS							
	Initial TRL	3							
	Target TRL	6							
	Related service	Urgent computing							
	Leader	ETH							
	End users	Civil protection, insurance companies							
	Description	Fast affectation maps for earthquakes are obtained mostly by employing ground motion prediction equations (GMPE) which are empirical relationships that relate maximum shaking with the distance from the event source. Together with site effects from the upper-most geological layers, it gives a quick but rough approximation of the affectation at a wide area. Physical 3D modelling of seismic waves can result in a much more detailed map of shaking for a particular event, although this requires a lot of computing time, urgent access to supercomputing resources, and it is very sensitive to uncertainties in the velocity model of the subsurface and to source parameters.							
	Data Sets								
	Name	*Description*	*Data Category*	*Repository location*	*F*	*A*	*I*	*R*	*References*
DataSet 1	Static (Historical) Seismological data	Historical seismic data	Scientific data	EUDAT EPOS AHEAD*, SHEEC, DISS, GEH, RCMT, CPTI15, EFEHR, DBMI15, ISIDe, INGV, IMO	0,67	0,50	0,50	0,50	No
DataSet 2	Geologic model	Urgent computing's geologic model input	Scientific data	IMO and INGV	0,00	0,00	0,00	0,00	No
DataSet 3	Preexisting faults models	Seismogenic fault systems input data	Scientific data - models	IMO INGV he European Database of Seismogen	1,00	0,00	0,00	0,00	No
DataSet 4	Seismic Events	Data coming from external WebServices containing the location, magnitude and focal mechanisms of an earthquake event.	Scientific Data	ORFEUS* ISC GEOFON IRIS EMSC IMO	1,00	0,50	1,00	1,00	Stations and RawTimeSeries
DataSet 5	Seismological stations	Data coming from external WebServices containing the set of stations of an affected earthquake zone	Scientific Data	ORFEUS* ISC GEOFON IRIS EMSC	0,50	0,33	1,00	1,00	No
DataSet 6	Raw data time series	Data coming from external WebServices containing the raw data time series for a set of given earthquake events	Scientific Data	ORFEUS* ISC GEOFON IRIS EMSC	1,00	0,50	1,00	1,00	No
DataSet 7	Computational Mesh	Computational mesh of the study region(s). These meshes could be different depending on the used flagship code (SeisSol, ExaHype, SPECFEM 3D, Salvus)	Scientific data	LMU ETH TUM	0,00	0,00	0,33	0,67	Geological and Fault datasets
DataSet 8	Synthetic seismic time series	Simulated seismic time series. They are the main output of the flagship codes (SeisSol, ExaHype, SPECFEM 3D, Salvus)	Scientific Data - syntetized data	LMU ETH TUM	0,00	0,00	0,00	0,33	No
DataSet 9	Maps of Potential Damage	Peak velocity and acceleration maps, shaking time maps, spectral acceleration at key locations	Scientific Data - syntetized data	LMU ETH TUM	0,00	0,00	0,33	0,33	No
DataSet 10	Shake movies	Ground shake movie from simulations output	Scientific Data - syntetized data	LMU ETH TUM	0,00	0,00	0,33	0,33	No

Figure 1. PD1 sheet example, taken from the DPM spreadsheet document (Annex). The pilot involves 4 flagship codes and 10 different datasets. Datasets and codes are linked to the corresponding dataset and code sheets respectively. Note that the dataset Table contains also information on the FAIR metric (see Section 3).

2.1. Dataset sheet

For each dataset in a PD, a dataset sheet exists (Figure 2) with information on the following data items:

Item	Comments/explanation
Name	Descriptive name to identify the dataset
Description	Short description of the contents
Data Category	Data category code (see Table 2 for the corresponding codes)
License	Chosen among the most appropriated and most open ones
Repository location	Institutional or public repository name
Author	Data author(s) name(s)
Naming Conventions	File names structure and conventions
Versioning	How and where the version of the dataset can be found
Format	Standards, definitions, ontologies, etc.
Size	Total estimated size, or single file size and number of expected files
Storage	Physical support selected, dependent on availability needs
Archive path	Folders structure
Associated metadata	Selected metadata standards and to metadata set
Provenance	Structured dataset origin information
Backups needs	Periodicity, subsets backup needs analysis, etc.
Access permissions	Lifecycle dependency: selected groups, or public
Legal/ethical restrictions	Privacy and security issues
Reproducibility	If yes: connection to code and environment
Data transfer needs	Replicas and periodic transfers to/from other repositories
Long term preservation	Needs at 3-5-7-10 years (if any)
Metadata management	Way to access metadata when data are not available
Resources need	Analysis of resources needs at each step of data lifecycle
References to other datasets	If applicable, explain which and why

Table 1. List of items in a dataset sheet and their definition.

Data category	Code	Name	Comments
Scientific data	1.1	Models	
	1.2	Experimental	Data coming from observation, measurements or produced by detectors/sensors or by any other experimental device and or activity
	1.3	Synthetic	Data generated by a simulation and/or are not obtained by direct measurement
	1.4	Test	Datasets (experimental or synthetic) used to validate models
Software	2.1	Libraries	
	2.2	Applications	
	2.3	Services	
	2.4	APIs	
Administrative docs	3.1	Documents	Any documentation, either public or private, such as code documentation,

			technical notes, etc., not directly mentioned in the project deliverables list.
	3.2	Internal reports	Meeting minutes, internal notes to document the evolution of the project, such as calendar, resources management, mailing lists, etc.
	3.3	Deliverables	Project output documents
Other	4.1	Metadata	Any data describing data properties. If they contain scientific information, they can also be classified as scientific data

Table 2. Summary of the different data categories

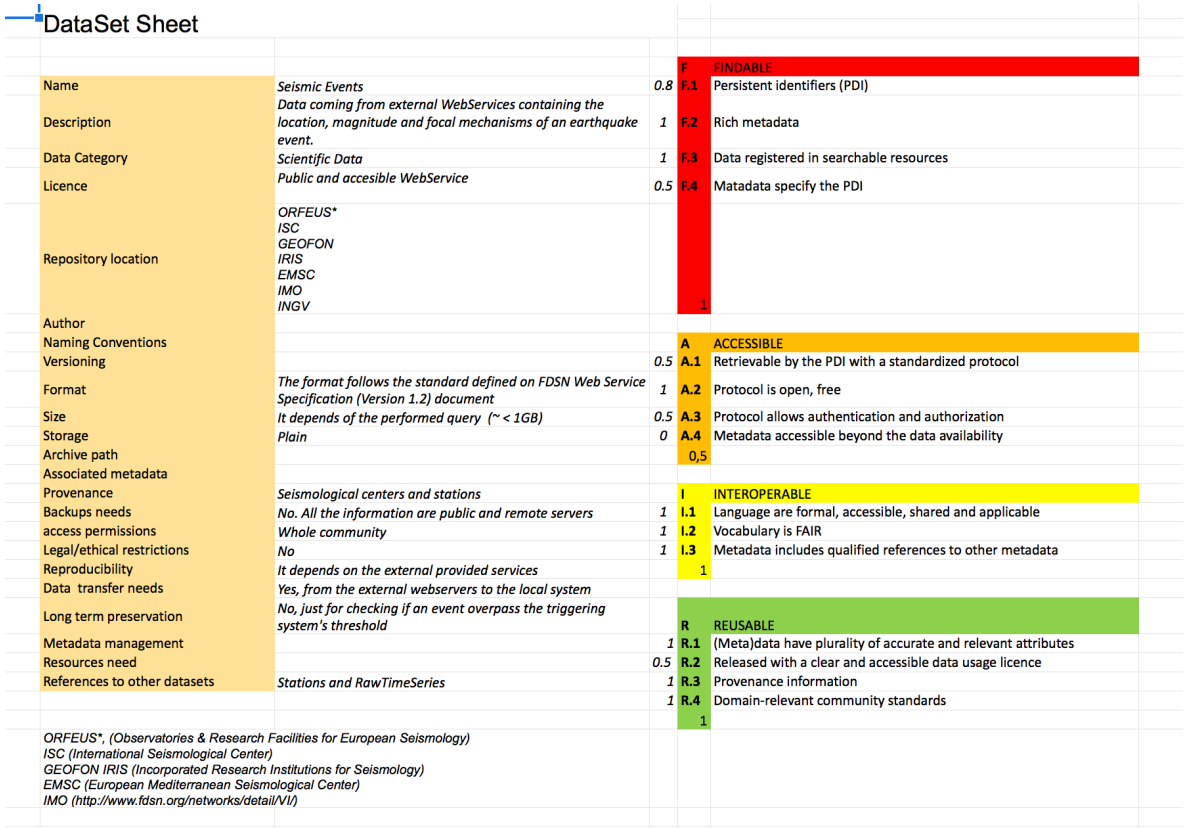

Figure 2. Dataset 4 sheet (example for seismic events data) for PD1 as in the DMP annex document.

2.2. Software sheet

Similar to the dataset sheets, one software sheet per code (or related software component) will contain information on the following aspects:

Item	Comments
Reference name of the program or workflow	Name of the code
Description	Brief description
Programming language	specify
Rules and best coding practices	conventions for filenames, link to an external manual, if exists (ex: PEP8, etc.)
Access permissions and license	lifecycle dependency: only specific groups of collaborators, all partners, whole community, etc.
Code size if relevant (to be updated)	
Repository type	GitHub, GitLab, Bitbucket, SourceForge
Repository structure	Branches, tags, etc.
Provenance information	Containers, virtual environments
Backup and Archiving needs	If any
Legal/ethical restrictions	If any
Versioning control and rules/workflows managing	Specify the repository
Code transfer needs and security	If any
Long term preservation needs	Only if applies to a given official release version
Documentation and inline comments rules	specify
Metadata management	(available even when the software is not)
Resources need	at each step of the lifecycle

Table 3. List of items in a software sheet.

3. FAIR data

The FAIR Guiding Principles (Wilkinson et al.; 2016; DOI: 10.1038/sdata.2016.18) describe distinct considerations for contemporary data publishing environments with respect to supporting both manual and automated deposition, exploration, sharing, and reuse. A metric to quantify the degree of "FAIRness" of each dataset in ChEESE has been defined. It results on a normalized value (between 0 and 1) for each of the 4 FAIR components. In turn, this (0,1) value results from assigning a flag value 0/1 to each of the FAIR subcomponents defined by Wilkinson et al. (2016) and listed in Table 4.

F	FINDABLE		
F.1	Persistent Identifiers (PDI)	(meta)data are assigned a globally unique and persistent identifier	0/1
F.2	Rich metadata	data are described with rich metadata (defined by subcomponent R.1 below)	0/1
F.3	Data registered in searchable resources	(meta)data are registered or indexed in a searchable resource	0/1
F.4	Metadata specifies the PDI	metadata clearly and explicitly include the identifier of the data it describes	0/1
A	ACCESSIBLE		
A.1	Retrievable by the PDI with a standardized protocol	(meta)data are retrievable by their identifier using a standardized communications protocol.	0/1

A.1.1	Open, free protocol	the protocol is open, free, and universally implementable	0/1
A.1.2	Authentication and Authorization	the protocol allows for an authentication and authorization procedure, where necessary	0/1
A.2	Metadata availability	Metadata are accessible beyond the data availability	0/1
I	INTEROPERABLE		
I.1	formal, accessible, shared and applicable language	(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation	0/1
I.2	FAIR vocabulary	(meta)data use vocabularies that follow FAIR principles	0/1
I.3	Metadata references	Metadata includes qualified references to other metadata	0/1
R	REUSABLE		
R.1	Relevant metadata	(Meta)data have plurality of accurate and relevant attributes	
R.1.1	Usage license	(meta)data are released with a clear and accessible data usage license	0/1
R.1.2	Provenance	(meta)data are associated with detailed provenance	0/1
R.1.3	Community standards	(meta)data meet domain-relevant community standards	0/1

Table 4. Definition of the different FAIR components and flag value (0/1) used to quantify the degree of fairness of each dataset.

3.1. Making ChEESE data Findable

ChEESE datasets suited for publication will be easily citable and easily findable with the assignation of Persistent Identifiers.

- The codes will be stored in repositories which permit versioning and tags for the identification of official releases and the connection with their outputs.
- Whenever possible, a rich metadata model and the register in disciplinary repositories will be used to allow other scientists to find the datasets produced by the project.
- Given the variety of the data of the project, the specific solutions and data models adopted for each dataset and software will be found in the corresponding sheet of the DMP.

3.2. Making ChEESE data openly Accessible

The open-data will be made accessible as follows:

- The source-codes of the flagship codes and related software components licensed as open-source will be included in a web repository for codes and toolkits at the end of the project (D4.13; due by M36). This will be useful for archiving project results. Additionally, and in collaboration with the European Plate Observing System (EPOS), ChEESE will promote and facilitate the integration of HPC services to widen the access to codes to the Solid Earth user's Community.

- Datasets access will depend on the different case and it will be described in the corresponding dataset sheet. Restriction of access will be guaranteed in cases ethical issues arise. Metadata will be made available as soon and as long as possible, independently on the accessibility of data.

3.3 Making ChEESE data Interoperable

The choice of metadata standards and of the way to access the data is still under discussion between the consortium members and with EPOS. Whenever possible, data coming from other resources will also be described in the DMP. Metadata standards will be chosen to guarantee the maximum interoperability inside EPOS.

3.4 Increase ChEESE data Re-use

The ChEESE open-datasets will be licensed under some Creative Commons data licensing (see Table 5) to let the widest reuse possible of it, since this license allows both commercial and non-commercial use of the data without any restriction. If necessary, an embargo on the data may exist to guarantee publication of results for a maximum of 1 year after the conclusion of the project. In any case, this will be specified in the corresponding dataset sheet.

		Allowed			
Creative Commons	**Description**	**Modification of the content**	**Commercial Use**	**Free cultural works**	**Open definition**
CC0	Free content, no restrictions	yes	yes	yes	yes
BY	Attribution	yes	yes	yes	yes
BY-SA	Attribution + Share Alike	yes	yes	yes	yes
BY-NC	Non Commercial	yes	no	no	no
BY-ND	No Derivatives	no	yes	no	no
BY-NC-SA		yes	no	no	no
BY-NC-ND		no	no	no	no

Table 5. Data licensing options. See this link for details.

4. Allocation of Resources

There is no additional cost for making the ChEESE datasets identified in Section 2 FAIR:

- The source code of the open-source software components and tools will be included in the repository (D4.13).
- The code performance evaluation datasets will be maintained at BSC facilities and included in publications.
- The rest of the open-data will be stored at the project site for at least three years after the end of the project. The infrastructure and personnel funds granted from the European Community will cover the storage, hardware and staff time to manage the servers on which the data will be stored.

5. Data security

Each dataset will be evaluated separately and exceptional security measures will be identified and applied. Regular backups for preventing loss of information will be used.

6. Ethical aspects

Early warning and hazard assessment can potentially have ethical implications: the diffusion of hazard results or a warning message can be risky for public order, and have social and economic impacts if not correctly interpreted and evaluated by competent authorities. To mitigate the impact, the project will distribute the ultimate simulations results (application to real cases) only under specific conditions and to the appropriate stakeholders, while the scientific results and products, like the models, etc., will be openly accessible. The limitations and conditions of distribution of each dataset will be indicated in the corresponding dataset sheet.

7. Engagement with EUDAT

ChEESE will provide solutions for data management and movement. In particular, it will foster the use of EUDAT services to store and publish research data (B2SHARE), distribute and store large volumes of data based on data policies (B2SAFE) and to transfer data between data resources and external computational facilities (B2STAGE), exploiting data citation (B2HANDLE), that for EUDAT hosted data is managed through Persistent Identifiers (PIDs), and metadata enrichment (B2NOTE).

Annex I

The management of the data information is going to be carried out to permit the researchers to keep it updated and spot issues, common requirements and solutions. The structure and the final format of the DMP may change in the future to fit the needs of the project.

The project members can access the DMP shared spreadsheets document https://docs.google.com/spreadsheets/d/17jK4vz5JiJQkMrDcvKAHGQWlVrmZYVfatRXQimVl_dg/edit?usp=sharing reported in the following pages, and update it whenever necessary.

	Pilot 1	**Urgent Seismic simulations**							
	Codes	ExaHyPE	Salvus	SPECFEM3D	SeisSol				
	Area	CS							
	Initial TRL	3							
	Target TRL	6							
	Related service	Urgent computing							
	Leader	ETH							
	End users	Civil protection, insurance companies							
	Description	Fast affectation maps for earthquakes are obtained mostly by employing ground motion prediction equations (GMPE) which are empirical relationships that relate maximum shaking with the distance from the event source. Together with site effects from the upper-most geological layers, it gives a quick but rough approximation of the affectation at a wide area. Physical 3D modelling of seismic waves can result in a much more detailed map of shaking for a particular event, although this requires a lot of computing time, urgent access to supercomputing resources, and it is very sensitive to uncertainties in the velocity model of the subsurface and to source parameters.							
	Data Sets								

	Name	*Description*	*Data Category*	*Repository location*	*F*	*A*	*I*	*R*	*References*
DataSet 1	Static (Historical) Seismological data	Historical seismic data	Scientific data	*See Data repositories B13:B24	1,00	0,50	0,50	0,50	No
DataSet 2	Geologic model	Urgent computing's geologic model input	Scientific data	IMO and INGV	0,00	0,00	0,00	0,00	No
DataSet 3	Preexisting faults models	Seismogenic fault systems input data	Scientific data - models	IMO INGV *See Data repositories B25	1,00	0,00	0,00	0,00	No
DataSet 4	Seismic Events	Data coming from external WebServices containing the location, magnitude and focal mechanisms of an earthquake event.	Scientific Data	* See Data repositories B8:B14	1,00	0,50	1,00	1,00	Stations and RawTimeSeries
DataSet 5	Seismological stations	Data coming from external WebServices containing the set of stations of an affected earthquake zone	Scientific Data	* See Data repositories B8:B14	0,50	0,33	1,00	1,00	No
DataSet 6	Raw data time series	Data coming from external WebServices containing the raw data time series for a set of given earthquake events	Scientific Data	* See Data repositories B8:B14	1,00	0,50	1,00	1,00	No
DataSet 7	Computational Mesh	Computational mesh of the study region(s). These meshes could be different depending on the used flagship code (SeisSol, ExaHype, SPECFEM 3D, Salvus)	Scientific data	LMU ETH TUM	0,00	0,00	0,33	0,67	Geological and Fault datasets

DataSet 8	Synthetic seismic time series	Simulated seismic time series. They are the main output of the flagship codes (SeisSol, ExaHype, SPECFEM 3D, Salvus)	Scientific Data - syntetized data	LMU ETH TUM	0,00	0,00	0,00	0,33	No
DataSet 9	Maps of Potential Damage	Peak velocity and acceleration maps, shaking time maps, spectral acceleration at key locations	Scientific Data - syntetized data	LMU ETH TUM	0,00	0,00	0,33	0,33	No
DataSet 10	Shake movies	Ground shake movie from simulations output	Scientific Data - syntetized data	LMU ETH TUM	0,00	0,00	0,33	0,33	No

Pilot 5	**Physics-based probabilistic seismic hazard assessment (PSHA)**			
Codes	ExaHyPE	SeisSol	A WP-ODC	
Area	CS			
Initial TRL	4			
Target TRL	7			
Related service	Hazard			
Leader	LMU			
End users	Civil protection, insurance companies			
Description	PSHA is widely established for deciding safety criteria for making official national hazard maps, developing building code requirements, safety of critical infrastructure (e. g. nuclear power plants) and determining earthquake insurance rates by governments and industry. However, PSHA currently rests on empirical, time-independent assumptions known to be too simplistic and conflict with earthquake physics. Validation Scenario: Northern Iceland. This PD will develop general concepts for enabling physics-based seismic hazard assessment with state-of-the-art multi-physics earthquake simulation software (SeisSol, AWP-ODC) and conduct 3D physics-based seismic simulations to improve PSHA for a validation scenario provided by IMO (Northern Iceland). We will make the pilot readily applicable to supplement established methods by stakeholders, for example for the e-science environment of EPOS, the USGS Science Applications for Risk Reduction (SAFRR), for Nuclear Power Plants, insurance companies, etc.			

Data Sets

	Name	*Description*	*Data Category*	*Repository location*	*F*	*A*	*I*	*R*	*References*
B	Geologic model	Urgent computing's geolc	Scientific data	IMO and INGV	0	0	0	0	No
C	Preexisting faults model	Seismogenic fault system	Scientific data - mc	IMO INGV *See Data repositories E	1	0	0	0	No
D									
E									
F									
G									
H									
I									

DataSet Sheet

			F	FINDABLE	Comments
Name	*Static (Historical) Seismological data*	*1*	**F.1**	Persistent identifiers (PDI)	
Description	*Historical seismic data*	*0.5*	**F.2**	Rich metadata	
Data Category	*Scientific data*	*1*	**F.3**	Data registered in searchable resources	
Licence	*Public*	*0.5*	**F.4**	Matadata specify the PDI	
Repository location	**See Data repositories B13:B24*		1		
Author					
Naming Conventions			**A**	ACCESSIBLE	
Versioning		*1*	**A.1**	Retrievable by the PDI with a standardized protocol	
Format		*1*	**A.2**	Protocol is open, free	
Size	*~ 1 GB*	*0*	**A.3**	Protocol allows authentication and authorization	
Storage	*Local*	*0*	**A.4**	Metadata accessible beyond the data availability	
Archive path			0,5		
Associated metadata					
Provenance			**I**	INTEROPERABLE	
Backups needs	*Yes*	*1*	**I.1**	Language are formal, accessible, shared and applicable	
access permissions	*Public*	*0*	**I.2**	Vocabulary is FAIR	
Legal/ethical restrictions	*No*	*0.5*	**I.3**	Metadata includes qualified references to other metadata	
Reproducibility	*Data origin*		0,5		
Data transfer needs	*Replicas*				
Long term preservation	*Yes*		**R**	REUSABLE	
Metadata management	*Internal DB*	*0*	**R.1**	(Meta)data have plurality of accurate and relevant attributes	
Resources need		*0*	**R.2**	Released with a clear and accessible data usage licence	
References to other datasets	*No*	*1*	**R.3**	Provenance information	
		1	**R.4**	Domain-relevant community standards	
			0,5		
Purpose - Processing					

DataSet Sheet

			F	FINDABLE
Name	*Geologic model*	*0*	**F.1**	Persistent identifiers (PDI)
Description	*Urgent computing's geologic model input*	*0*	**F.2**	Rich metadata
Data Category	*Scientific data*	*0*	**F.3**	Data registered in searchable resources
Licence	*Public*	*0*	**F.4**	Metadata specify the PDI
Repository location	*IMO and INGV*		0	
Author				
Naming Conventions			**A**	ACCESSIBLE
Versioning		*0*	**A.1**	Retrievable by the PDI with a standardized protocol
Format		*0*	**A.2**	Protocol is open, free
Size	*< 1GB*	*0*	**A.3**	Protocol allows authentication and authorization
Storage	*Local*	*0*	**A.4**	Metadata accessible beyond the data availability
Archive path			0	
Associated metadata				
Provenance	*Yes*		**I**	INTEROPERABLE
Backups needs	*No*	*0*	**I.1**	Language are formal, accessible, shared and applicable
access permissions	*Public*	*0*	**I.2**	Vocabulary is FAIR
Legal/ethical restrictions	*No*	*0*	**I.3**	Metadata includes qualified references to other metadata
Reproducibility	*Data origin*		0	
Data transfer needs	*replicas*			
Long term preservation	*No*		**R**	REUSABLE
Metadata management	*Internal DB*	*0*	**R.1**	(Meta)data have plurality of accurate and relevant attributes
Resources need		*0*	**R.2**	Released with a clear and accessible data usage licence
References to other datasets	*No*	*0*	**R.3**	Provenance information
		0	**R.4**	Domain-relevant community standards
			0	

DataSet Sheet

			F	FINDABLE
Name	*Preexisting faults models*	*1*	**F.1**	Persistent identifiers (PDI)
Description	*Seismogenic fault systems input data*	*1*	**F.2**	Rich metadata
Data Category	*Scientific data - models*	*1*	**F.3**	Data registered in searchable resources
Licence		*1*	**F.4**	Matadata specify the PDI
Repository location	*IMO* *INGV* **See Data repositories B25*			1
Author				
Naming Conventions			**A**	ACCESSIBLE
Versioning		*0.8*	**A.1**	Retrievable by the PDI with a standardized protocol
Format		*0.8*	**A.2**	Protocol is open, free
Size	*< 1GB*	*0*	**A.3**	Protocol allows authentication and authorization
Storage	*Local*	*0*	**A.4**	Metadata accessible beyond the data availability
Archive path			0	
Associated metadata				
Provenance	*Yes*		**I**	INTEROPERABLE
Backups needs	*No*	*0*	**I.1**	Language are formal, accessible, shared and applicable
access permissions	*Restricted*	*0*	**I.2**	Vocabulary is FAIR
Legal/ethical restrictions	*No*	*0*	**I.3**	Metadata includes qualified references to other metadata
Reproducibility	*Data download*		0	
Data transfer needs	*Periodic update*			
Long term preservation	*No*		**R**	REUSABLE
Metadata management	*Internal DB*	*0*	**R.1**	(Meta)data have plurality of accurate and relevant attributes
Resources need		*0*	**R.2**	Released with a clear and accessible data usage licence
References to other datasets	*No*	*0*	**R.3**	Provenance information
		0	**R.4**	Domain-relevant community standards
			0	

DataSet Sheet

			F	FINDABLE
Name	*Seismic Events*	*0.8*	**F.1**	Persistent identifiers (PDI)
Description	*Data coming from external WebServices containing the location, magnitude and focal mechanisms of an earthquake event.*	*1*	**F.2**	Rich metadata
Data Category	*Scientific Data*	*1*	**F.3**	Data registered in searchable resources
Licence	*Public and accesible WebService*	*0.5*	**F.4**	Matadata specify the PDI
Repository location	** See Data repositories B8:B14*		1	
Author				
Naming Conventions			**A**	ACCESSIBLE
Versioning		*0.5*	**A.1**	Retrievable by the PDI with a standardized protocol
Format	*The format follows the standard defined on FDSN Web Service Specification (Version 1.2) document*	*1*	**A.2**	Protocol is open, free
Size	*It depends of the performed query (~ < 1GB)*	*0.5*	**A.3**	Protocol allows authentication and authorization
Storage	*Plain*	*0*	**A.4**	Metadata accessible beyond the data availability
Archive path			0,5	
Associated metadata				
Provenance	*Seismological centers and stations*		**I**	INTEROPERABLE
Backups needs	*No. All the information are public and remote servers*	*1*	**I.1**	Language are formal, accessible, shared and applicable
access permissions	*Whole community*	*1*	**I.2**	Vocabulary is FAIR
Legal/ethical restrictions	*No*	*1*	**I.3**	Metadata includes qualified references to other metadata
Reproducibility	*It depends on the external provided services*		1	
Data transfer needs	*Yes, from the external webservers to the local system*			
Long term preservation	*No, just for checking if an event overpass the triggering system's threshold*		**R**	REUSABLE
Metadata management		*1*	**R.1**	(Meta)data have plurality of accurate and relevant attributes
Resources need		*0.5*	**R.2**	Released with a clear and accessible data usage licence
References to other datasets	*Stations and RawTimeSeries*	*1*	**R.3**	Provenance information
		1	**R.4**	Domain-relevant community standards
			1	

DataSet Sheet

			F	FINDABLE
Name	*Seismological stations*	*0*	**F.1**	Persistent identifiers (PDI)
Description	*Data coming from external WebServices containing the set of stations of an affected earthquake zone*	*1*	**F.2**	Rich metadata
Data Category	*Scientific Data*	*1*	**F.3**	Data registered in searchable resources
Licence	*Public and accesible WebService*	*0*	**F.4**	Matadata specify the PDI
Repository location	** See Data repositories B8:B14*		0,5	
Author				
Naming Conventions			**A**	ACCESSIBLE
Versioning		*0*	**A.1**	Retrievable by the PDI with a standardized protocol
Format	*The format follows the standard defined on FDSN Web Service Specification (Version 1.2) document*	*1*	**A.2**	Protocol is open, free
Size	*It depends of the performed query (~ < 1 GB)*	*0.5*	**A.3**	Protocol allows authentication and authorization
Storage	*Plain*	*0*	**A.4**	Metadata accessible beyond the data availability
Archive path			0,333	
Associated metadata				
Provenance	*Seismological centers and stations*		**I**	INTEROPERABLE
Backups needs	*No. All the information are public and remote servers*	*1*	**I.1**	Language are formal, accessible, shared and applicable
access permissions	*Whole community*	*1*	**I.2**	Vocabulary is FAIR
Legal/ethical restrictions	*No*	*1*	**I.3**	Metadata includes qualified references to other metadata
Reproducibility	*It depends on the external provided services*		1	
Data transfer needs	*Yes, from the external webservers to the local system*			
Long term preservation	*No, just for checking if an event overpass the triggering system's threshold*		**R**	REUSABLE
Metadata management		*1*	**R.1**	(Meta)data have plurality of accurate and relevant attributes
Resources need		*0.5*	**R.2**	Released with a clear and accessible data usage licence
References to other datasets	*No*	*1*	**R.3**	Provenance information
		1	**R.4**	Domain-relevant community standards
			1	

DataSet Sheet

			F	FINDABLE
Name	*Raw data time series*	*0.5*	**F.1**	Persistent identifiers (PDI)
Description	*Data coming from external WebServices containing the raw data time series for a set of given earthquake events*	*1*	**F.2**	Rich metadata
Data Category	*Scientific Data*	*1*	**F.3**	Data registered in searchable resources
Licence	*Public and accesible WebService*	*0.5*	**F.4**	Matadata specify the PDI
Repository location	** See Data repositories B8:B14*		1	
Author				
Naming Conventions			**A**	ACCESSIBLE
Versioning		*0.5*	**A.1**	Retrievable by the PDI with a standardized protocol
Format	*The format follows the standard defined on FDSN Web Service Specification (Version 1.2) document*	*1*	**A.2**	Protocol is open, free
Size	*It depends of the performed query (~ < 1GB)*	*0.5*	**A.3**	Protocol allows authentication and authorization
Storage	*Plain*	*0*	**A.4**	Metadata accessible beyond the data availability
Archive path			0,5	
Associated metadata				
Provenance	*Seismological centers and stations*		**I**	INTEROPERABLE
Backups needs	*No. All the information are public and remote servers*	*1*	**I.1**	Language are formal, accessible, shared and applicable
access permissions	*Whole community*	*1*	**I.2**	Vocabulary is FAIR
Legal/ethical restrictions	*No*	*1*	**I.3**	Metadata includes qualified references to other metadata
Reproducibility	*It depends on the external provided services*		1	
Data transfer needs	*Yes, from the external webservers to the local system*			
Long term preservation	*No, just for checking if an event overpass the triggering system's threshold*		**R**	REUSABLE
Metadata management		*1*	**R.1**	(Meta)data have plurality of accurate and relevant attributes
Resources need		*0.5*	**R.2**	Released with a clear and accessible data usage licence
References to other datasets	*No*	*1*	**R.3**	Provenance information
		1	**R.4**	Domain-relevant community standards
			1	

DataSet Sheet

			F	FINDABLE
Name	*Computational Mesh*	*0*	**F.1**	Persistent identifiers (PDI)
Description	*Computational mesh of the study region(s). These meshes could be different depending on the used flagship code (SeisSol, ExaHype, SPECFEM 3D, Salvus)*	*0*	**F.2**	Rich metadata
Data Category	*Scientific Data - syntetized data*	*0*	**F.3**	Data registered in searchable resources
Licence		*0*	**F.4**	Matadata specify the PDI
Repository location	*LMU* *ETH* *TUM*		0	
Author	*LMU* *ETH* *TUM*			
Naming Conventions			**A**	ACCESSIBLE
Versioning		*0*	**A.1**	Retrievable by the PDI with a standardized protocol
Format	*Each code will create the mesh following the requierments of the Computational Seismology code (SeisSol, ExaHype, SPECFEM 3D, Salvus)*	*0.3*	**A.2**	Protocol is open, free
Size	*It depends on the Computational Seismology code (< 50 GB)*	*0*	**A.3**	Protocol allows authentication and authorization
Storage		*0.8*	**A.4**	Metadata accessible beyond the data availability
Archive path			0	
Associated metadata				
Provenance	*LMU* *ETH* *TUM*		**I**	INTEROPERABLE
Backups needs		*1*	**I.1**	Language are formal, accessible, shared and applicable
access permissions	*ChEESE Partners*	*0*	**I.2**	Vocabulary is FAIR
Legal/ethical restrictions	*No*	*0*	**I.3**	Metadata includes qualified references to other metadata
Reproducibility	*Yes*		0,333	
Data transfer needs	*Yes, from their location to a HPC center*			
Long term preservation	*Yes, could be used to future simulations*		**R**	REUSABLE
Metadata management		*1*	**R.1**	(Meta)data have plurality of accurate and relevant attributes
Resources need		*0*	**R.2**	Released with a clear and accessible data usage licence
References to other datasets	*Geological and Fault datasets*	*0.8*	**R.3**	Provenance information
		1	**R.4**	Domain-relevant community standards
			0,666	

DataSet Sheet

			F	FINDABLE
Name	*Synthetic seismic time series*	*0*	**F.1**	Persistent identifiers (PDI)
Description	*Simulated seismic time series. They are the main output of the flagship codes (SeisSol, ExaHype, SPECFEM 3D, Salvus)*	*0*	**F.2**	Rich metadata
Data Category	*Scientific Data - syntetized data*	*0*	**F.3**	Data registered in searchable resources
Licence		*0*	**F.4**	Matadata specify the PDI
Repository location	*LMU* *ETH* *TUM*		0	
Author	*LMU* *ETH* *TUM*			
Naming Conventions			**A**	ACCESSIBLE
Versioning		*0*	**A.1**	Retrievable by the PDI with a standardized protocol
Format	*All flagship codes should generate this data on the same format.*	*0*	**A.2**	Protocol is open, free
Size	*It depends on the problem solving (~ < 20 GB)*	*0*	**A.3**	Protocol allows authentication and authorization
Storage		*0*	**A.4**	Metadata accessible beyond the data availability
Archive path			0	
Associated metadata				
Provenance	*LMU* *ETH* *TUM*		**I**	INTEROPERABLE
Backups needs	*Yes*	*0.5*	**I.1**	Language are formal, accessible, shared and applicable
access permissions	*Whole community*	*0*	**I.2**	Vocabulary is FAIR
Legal/ethical restrictions	*No*	*0*	**I.3**	Metadata includes qualified references to other metadata
Reproducibility	*Yes*		0	
Data transfer needs	*Yes, from HPC center to the Static data center (e.g. EUDAT)*			
Long term preservation	*Yes, according to the ChEESE project directives*		**R**	REUSABLE
Metadata management		*0*	**R.1**	(Meta)data have plurality of accurate and relevant attributes
Resources need		*0*	**R.2**	Released with a clear and accessible data usage licence
References to other datasets	*No*	*0.8*	**R.3**	Provenance information
		1	**R.4**	Domain-relevant community standards
			0,333	

DataSet Sheet

				F	FINDABLE
Name	*Maps of Potential Damage*	*0*	**F.1**	Persistent identifiers (PDI)	
Description	*Peak velocity and acceleration maps, shaking time maps, spectral acceleration at key locations*	*0*	**F.2**	Rich metadata	
Data Category	*Scientific Data - syntetized data*	*0*	**F.3**	Data registered in searchable resources	
Licence		*0*	**F.4**	Matadata specify the PDI	
Repository location	*LMU* *ETH* *TUM*		0		
Author	*LMU* *ETH* *TUM*				
Naming Conventions			**A**	ACCESSIBLE	
Versioning		*0*	**A.1**	Retrievable by the PDI with a standardized protocol	
Format	*All flagship codes should generate this data on the same format.*	*0.5*	**A.2**	Protocol is open, free	
Size	*It depends on the problem solving (~ < 1 GB)*	*0*	**A.3**	Protocol allows authentication and authorization	
Storage		*0*	**A.4**	Metadata accessible beyond the data availability	
Archive path			0		
Associated metadata					
Provenance	*LMU* *ETH* *TUM*		**I**	INTEROPERABLE	
Backups needs	*Yes*	*1*	**I.1**	Language are formal, accessible, shared and applicable	
access permissions	*Whole community*	*0*	**I.2**	Vocabulary is FAIR	
Legal/ethical restrictions	*Yes*	*0*	**I.3**	Metadata includes qualified references to other metadata	
Reproducibility	*Yes*		0,333		
Data transfer needs	*Yes, from HPC center to the Static data center (e.g. EUDAT)*				
Long term preservation	*Yes, according to the ChEESE project directives*		**R**	REUSABLE	
Metadata management		*0*	**R.1**	(Meta)data have plurality of accurate and relevant attributes	
Resources need		*0*	**R.2**	Released with a clear and accessible data usage licence	
References to other datasets	*No*	*0.5*	**R.3**	Provenance information	
		1	**R.4**	Domain-relevant community standards	
			0,333		

DataSet Sheet

			F	FINDABLE
Name	*Shake movies*	*0*	**F.1**	Persistent identifiers (PDI)
Description	*Ground shake movie from simulations output*	*0*	**F.2**	Rich metadata
Data Category	*Scientific Data - syntetized data*	*0*	**F.3**	Data registered in searchable resources
Licence		*0*	**F.4**	Matadata specify the PDI
Repository location	*LMU* *ETH* *TUM*		0	
Author	*BSC* *LMU* *ETH* *TUM*			
Naming Conventions			**A**	ACCESSIBLE
Versioning		*0*	**A.1**	Retrievable by the PDI with a standardized protocol
Format	*MP4 or similar*	*0.5*	**A.2**	Protocol is open, free
Size	*~ GB*	*0*	**A.3**	Protocol allows authentication and authorization
Storage		*0*	**A.4**	Metadata accessible beyond the data availability
Archive path			0	
Associated metadata				
Provenance	*BSC* *LMU* *ETH* *TUM*		**I**	INTEROPERABLE
Backups needs	*Yes*	*1*	**I.1**	Language are formal, accessible, shared and applicable
access permissions	*Whole community*	*0*	**I.2**	Vocabulary is FAIR
Legal/ethical restrictions	*Yes*	*0*	**I.3**	Metadata includes qualified references to other metadata
Reproducibility	*Yes*		0,333	
Data transfer needs	*Yes, from HPC center to the Static data center (e.g. EUDAT)*			
Long term preservation	*Yes, according to the ChEESE project directives*		**R**	REUSABLE
Metadata management		*0*	**R.1**	(Meta)data have plurality of accurate and relevant attributes
Resources need		*0*	**R.2**	Released with a clear and accessible data usage licence
References to other datasets	*Yes, DataSet I*	*0.5*	**R.3**	Provenance information
		1	**R.4**	Domain-relevant community standards
			0,333	

Software Sheet

Reference name of the program or workflow	**Alert trigger system**
Description	Workflow developed to control the alert trigger system. That consists on: a earthquake (EQ) events monitoring process and a decision protocol for triggering the EQ alert.
Programming language	Python and Unix bash scripts
Rules and best coding practices	PEP8 style for Python
Access permissions and license	None
Code size if relevant (to be updated)	
Repository type	GIT on a private repository server (e.g. GiTLab)
Repository structure	GIT repository structure (master, branches and tags for stable versions)
Provenance information	Developed at BSC
Backup and Archiving needs	
Legal/ethical restrictions	N/A
Versioning control and rules/workflows managing	Major.Minor.Patches (e.g. v1.4.10)
Code transfer needs and security	N/A
Long term preservation needs	Long term (unlimited)
Documentation and inline comments rules	Doxygen is used for both inline and outside documentation.
Metadata management	
Resources need	A server with internet access, running the software in background.

Formats definition

Pilot	Scientific Files	Source code	Metadata	Documents
PD1			FDSN (Federation of Digital Seismograph Networks)	
PD1			EIDA (European Integrated Data Archive)	
PD1			SRF (Standard Rupture Format)	

Data Categories Definition

Find your dataset category from the listed below.
If you need a new sub-category, you can add more at the end of the corresponding category list.

1	Scientific Data		2	Software		3	Administrative docs		4	Other	
	Name	*Definition*		*Name*	*Definition*		*Name*	*Definition*		*Name*	*Definition*
1,01	**models**		2,01	**libraries**		3,01	**Documents**	Any documentation, either public or private, such as code documentation, technical notes, etc., not directly mentioned in the project deliverables list.	4,01	**metadata**	Any data describing data properties. If they contain scientific information, they can also be classified as scientific data
1,02	**experimental data**	Data coming from measurements or produced by detectors or by any other experimental activity	2,02	**applications**		3,02	**Internal reports**	Meeting minutes, internal notes to document the evolution of the project, such as calendar, resources management, mailing lists, …			
1,03	**syntetized data**	Data produced from simulation	2,03	**services**		3,03	**Deliverables**	Project outputs documents			
1,04	**test data**	Data used to test software	2,04	**APIs - source code**							

Data Repositories Information

Data repositories are listed below.
If you need, you can add more at the end of the list.

Repository

Short Name	*Extended name*	*Location*	*URL*	*Permissions*	*...*
ORFEUS	Observatories & Research Facilities for European Seismology		http://www.orfeus-eu.org/data/eida/webservices/		
ISC	International Seismological Center		http://www.isc.ac.uk/fdsnws/event/1/		
GEOFON			http://geofon.gfz-potsdam.de/fdsnws/dataselect/1/version		
IRIS	Incorporated Research Institutions for Seismology		http://service.iris.edu/fdsnws/		
EMSC	European Mediterranean Seismological Center		https://www.emsc-csem.org/Earthquake/seismicity/real_time.php		
IMO	Icelandic Meteorological Office		http://www.fdsn.org/networks/detail/VI/		
INGV	Istituto Nazionale di Geofisica e Vulcanologia		http://www.fdsn.org/networks/detail/TV/		
EUDAT			https://eudat.eu/catalogue		
EPOS	European Plate Observing System		ttps://www.epos-ip.org/tcs/seismology		
AHEAD	European Archive of Historical EArthquake Data		https://www.emidius.eu/AHEAD/services/		
SHEEC	The SHARE European Earthquake Catalogue		https://www.emidius.eu/SHEEC/		
DISS	Database of Individual Seismogenic Sources		http://diss.rm.ingv.it/diss/		
GEH	Global Historical Earthquake Archive		https://www.emidius.eu/GEH/		
RCMT	European-Mediterranean Regional Centroid-Moment Tensors		http://rcmt2.bo.ingv.it/		
CPTI15	Parametric Catalogue of Italian Earthquakes		https://emidius.mi.ingv.it/CPTI15-DBMI15/index_en.htm		
EFEHR	Earthquake Catalogue Europe		http://www.efehr.org/en/hazard-data-access/Intro/		
ISIDe	Italian Seismological Instrumental and Parametric Data-Base		http://cnt.rm.ingv.it/en/iside		
SHARE	The European Database of Seismogenic Faults		http://diss.rm.ingv.it/share-edsf/		

DataSet Sheet Template - do not modify!

Valid values: 1(totally compliant), 0.5(partially/ongoing), 0(not compliant)

		Value			Comments
			F	FINDABLE	
Name	Descriptive name to identify the dataset		F.1	Persistent identifiers (PDI)	
Description	Short description of the contents		F.2	Rich metadata	
Data Category	Data category code (see Table Data Category for the corresponding codes)		F.3	Data registered in searchable resources	
Licence	Chosen among the most appropriated and most open ones		F.4	Metadata specify the PDI	
Repository location	Institutional or public repository name and URL, if available				
Author	Data author(s) name(s)				
Naming Conventions	File names structure and conventions	Value	A	ACCESSIBLE	
Versioning	How and where the version of the dataset can be found		A.1	Retrievable by the PDI with a standardized protocol	
Format	Standard formats and contents standards, definitions, ontologies, … Link to description of format document. General or specific format - libraries or parsing code		A.2	Protocol is open, free	
Size	Total or single file size * n. of files		A.3	Protocol allows authentication and authorization	
Storage	Physical support		A.4	Metadata accessible beyond the data availability	
Archive path	Folders structure				
Associated metadata	reference to metadata standards				
Provenance	Structured dataset origin information	Value	I	INTEROPERABLE	
Backups needs	Periodicity, subsets backup needs analysis, etc.		I.1	Language are formal, accessible, shared and applicable	
access permissions	Lifecycle dependency: only specific groups of collaborators, all partners, whole community, …		I.2	Vocabulary is FAIR	
Legal/ethical restrictions	Privacy and security issues		I.3	Metadata includes qualified references to other metadata	
Reproducibility	If yes: connection to code and environment				
Data transfer needs	Replicas and periodic transfers to/from other repositories				
Long term preservation	Needs at 3-5-7-10 years (if any)	Value	R	REUSABLE	
Metadata management	Way to access metadata when data are not available		R.1	(Meta)data have plurality of accurate and relevant attributes	
Resources need	Analysis of resources needs at each step of data lifecycle		R.2	Released with a clear and accessible data usage licence	
References to other datasets	If applicable, explain which and why		R.3	Provenance information	
			R.4	Domain-relevant community standards	

Software Sheet - do not modify

Reference name of the program or workflow	Name of the code
Description	Brief description of the functionality and applicability of the software
Author	specify
Programming language	specify the programming language(s)
Rules and best coding practices	conventions for filenames, link to an external manual, if exists (ex: PEP8, etc.)
Access permissions and license	lifecycle dependency: only specific groups of collaborators, all partners, whole community, etc.
Code size	if relevant (to be updated)
Repository type	GitHub, GitLab, Bitbucket, SourceForge
Repository structure	Branches, tags, etc.
Provenance information	Containers, virtual environments
Backup and Archiving needs	If any
Legal/ethical restrictions	If any
Versioning control and rules/workflows managing	Specify the repository
Code transfer needs and security	If any
Long term preservation needs	Only if applies to a given official release version
Documentation and inline comments rules	specify
Metadata management	(available even when the software is not)
Resources need	Requirements for sw at each step of the lifecycle (access to repository, computational needs, accessibilities, permissions, ...)

FAIR DATA definitions - do not modify

GoFAIR - principles definition link

To evaluate the FAIRness of a dataset, a punctuation should be given to each of the following points (elaborated by www.force11.org)
Valid values: 1 (totally compliant), 0.5 (partially/ongoing), 0 (not compliant)
To be noted that not all the datasets must reach the maximum FAIRness, but in principle the output should, unless there is a valid motivation.

F	FINDABLE	A	ACCESSIBLE	I	INTEROPERABLE	R	REUSABLE
F.1	(Meta)data are assigned a globally unique and persistent identifier	A.1	(Meta)data are retrievable by their identifier using a standardised communications protocol	I.1	(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	R.1	Meta(data) are richly described with a plurality of accurate and relevant attributes
F.2	Data are described with rich metadata (defined by R1)	A.2	The protocol is open, free, and universally implementable	I.2	(Meta)data use vocabularies that follow FAIR principles	R.2	(Meta)data are released with a clear and accessible data usage license
F.3	(Meta)data are registered or indexed in a searchable resource	A.3	The protocol allows for an authentication and authorisation procedure, where necessary	I.3	(Meta)data include qualified references to other (meta)data	R.3	(Meta)data are associated with detailed provenance
F.4	Metadata clearly and explicitly include the identifier of the data they describe	A.4	Metadata are accessible, even when the data are no longer available			R.4	(Meta)data meet domain-relevant community standards

Data licensing

		Allowed			
Creative Commons	**Description**	**Modification of the content**	**Commercial Use**	**Free cultural works**	**Open definition**
CC0	Free content, no restrictions	yes	yes	yes	yes
BY	Attribution	yes	yes	yes	yes
BY-SA	Attribution+ShareAlike	yes	yes	yes	yes
BY-NC	NonCommercial	yes	no	no	no
BY-ND	NoDerivatives	no	yes	no	no
BY-NC-SA		yes	no	no	no
BY-NC-ND		no	no	no	no

Definitions of common license types

data.world: Common license types for datasets

For details, click here

Public Domain	The work has been dedicated to the public domain by waiving all rights to the work worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law.
Attribution	You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
Share-alike	If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.
Non-commercial	You may not use the material for commercial purposes.
Database Only	License applies to the database only and not its contents or data.
No Derivatives	No Derivative Works. You may not alter, transform, or build upon this work.