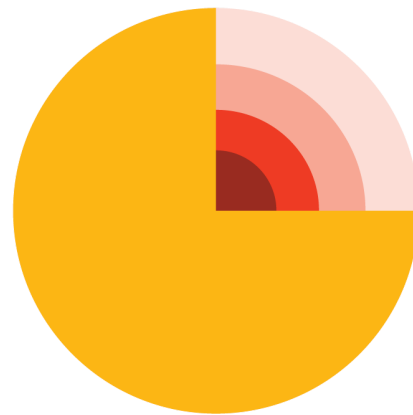




ChEESE

Deliverable D1.2

**First Data Management
Plan (DMP)**



Funded by the European Union. This work has received funding from the European High Performance Computing Joint Undertaking (JU) and Spain, Italy, Iceland, Germany, Norway, France, Finland and Croatia under grant agreement No 101093038.



D1.2 First Data Management Plan (DMP)

Document Information

Contract Number	101093038
Project Website	https://cheese2.eu
Contractual Deadline	31/08/2023 (M7)
Activity	WP1
Dissemination Level	PUBLIC
Nature	R

I. Copyright

Copyright Members of ChEESE collaboration, 2023/2026.

II. Delivery Slip

Role	Name	Partner/Activity
Author	Arnau Folch	GEO3BCN-CSIC
Contributor 1	Tomaso Esposti-Ongaro	INGV
Reviewer 1	Sara Barsotti	IMO
Reviewer 2	Velerio Vinciarelli	EPOS ERIC

III. Version control

Version	Date	Description of change	Author/Partner
V1	05/07/2023	Initial draft	Arnau Folch
V2	19/07/2023	Final version	Arnau Folch, Tomaso Esposti-Ongaro

IV. List of Acronyms

NOTE: the items in bold are further detailed in Annex 1 (Glossary of concepts) at the end of this document.

Acronym	Description
CC	Creative Commons
CERIF	Common European Research project Information Format
CI/CD	Continuous Integration / Continuous Deployment
CoE	Center of Excellence
CSA	Coordination Support Action
DA	Digital Asset
DCAT	Data Catalogue Vocabulary
DevOps	Development Operations
DMP	Data Management Plan
DOI	Digital Object Identifier
EOSC	European Open Science Cloud
EPOS	European Plate Observing System
ERIC	European Research Infrastructure Consortium
FAIR	Findable, Accessible, Interoperable, Reproducible
FAIR-EVA	FAIR data Evaluator Validator and Advisor
FIP	FAIR Implementation Profiles
GA	Grant Agreement
JePL	Jenkins Pipeline Library
JSON	JavaScript Object Notation (file format)
KER	Key Exploitable Results

KPI	Key Performance Indicator
maDMP	Machine-actionable Data Management Plan
PD	Pilot Demonstrator
RDA	Research Data Alliance
SC	Simulation Case
SE	Solid Earth
SQAaaS	Software Quality Assessment as a Service
TCS	Thematic Core Service (EPOS)
TRL	Technology Readiness Level
UML	Unified Modeling Language
WP	Work Package

Index

Executive summary.....	5
1. Introduction.....	5
1.1. Categories of ChEESE-2P digital assets.....	6
1.2. Types of repositories.....	8
1.3. Structure of this document.....	9
2. Management of software assets.....	10
2.1. Software repositories.....	10
2.2. The EOSC-synergy SQAaaS platform.....	12
3. Management of scientific data.....	14
3.1. Data repositories.....	14
3.2 The EOSC-synergy FAIR EVA tool.....	15
3.3. The F-UJI assessment tool.....	17
4. Management of metadata.....	17
4.1. Metadata catalogues and descriptors.....	17
4.1.1. Metadata model for software.....	18
4.1.2. Metadata model for datasets.....	18
4.2. Machine-actionable DMP.....	19
5. Management of other data.....	20
6. Ethical aspects.....	20
7. Timeline for the DMP.....	20
References.....	21
Annex 1. Glossary of concepts.....	22

Executive summary

This deliverable presents the first version of the Data Management Plan (DMP) of the ChEESE-2P project, which describes the management life-cycle for all the digital assets to be collected, processed, and generated during the lifetime of the project (2023-2026). The document describes the overall contents and organisation of the DMP, considering that the software, datasets and metadata descriptors will be continuously generated and updated during the course of the project and compiled in the final version of the DMP (i.e. deliverable D1.4, due by M34). Specifically, this first version of the DMP focusses on:

- How the different categories of assets generated, collected, and processed during the project will be managed during (and after) ChEESE-2P. This affects mainly software and the datasets resulting from the Simulation Cases (SCs) and the service enabling on EuroHPC systems (WP6, from M13 to M48). However, to a lesser extent, other datasets are also expected from other research activities.
- Which methodologies and which standards will be applied to manage the ChEESE-2P digital assets, ensuring that the metadata descriptors are interoperable with the EPOS service catalogue.
- How the assets will be stored and handled during the lifetime of the project and after its conclusion, as well as how they will be made (openly) accessible through different types of trusted open repositories.
- How the Software Quality as a Service (SQaaS) and the FAIR EVA tools resulting from the EOSC-synergy project will be integrated in the FAIRness assessment of the digital assets (other options may also be considered).

1. Introduction

The Grant Agreement (GA) of the ChEESE-2P Center of Excellence (CoE) requires that a Data Management Plan (DMP) is defined during the setup phase of the project (M6) and then updated regularly in order to:

- Deposit the data assets generated during the project in trusted repositories, in compliance with European Open Science Cloud (EOSC) requirements (see the [EOSC declaration](#)).
- Ensure open access (via the repositories) to the deposited data and under the latest available (currently 4.0) versions of either the Creative Commons Attribution International Public License (CC BY) or the Creative Commons Public Domain Dedication (CC 0); see Annex 1 for further details on the CC licence types.
- Ensure that the metadata of the deposited data is also in line with the FAIR principles (in particular machine-actionable) and provides information on:
 - Datasets (description, date of deposit, authors, embargo, etc).

- Funding recognition (grant project name, acronym and number, etc).
- Licensing terms.
- Persistent identifiers for the assets, related publications, and other research outputs.

To meet the requirements above, ChEESE-2P includes a dedicated task in WP1 (task T1.5; Data management and automatic FAIR validation) devoted to design and progressively implement a DMP in line with these contractual obligations and, in addition, to align with the open science best-practices by means of an EOSC-enabled data/software quality approach able to assess (quantify) the level of FAIRness of the digital assets. To this purpose, task T1.5 contemplates an automatic assessment of data/software FAIRness with quantitative indicators relying either on:

1. Tools developed during the H2020 [EOSC-synergy project](#) (GA No 857647) for both software and data. Dockerized and online (the so-called service platform) versions of the EOSC-synergy validation tools already exist for easier deployment and testing in integrated environments, and will be connected to the ChEESE-2P data/software repositories (and services) to perform the corresponding FAIRness assessments.
2. The F-UJI platform, an automatic FAIR validation (data only) developed by the H2020 [FAIRsFAIR project](#) (GA No 831558) which is also promoted by CASTIEL-2 as a transversal tool for all CoEs.

It is important to note that this DMP embraces ALL the Digital Assets (DAs) in ChEESE-2P, i.e. it includes not only the datasets but also software (codes and workflows), services, and other derived data products. On the other hand, and throughout this DMP, it is necessary to distinguish between:

- The management, curation, and stewardship of the DAs themselves, which can be stored in many distributed repositories and may be accessed via various pathways, including from a dedicated metadata catalogue(s).
- The management of metadata through metadata catalogues describing the digital assets above, which typically are centralised in a unique repository and associated with a user portal access.

1.1. Categories of ChEESE-2P digital assets

As stated in the GA, ChEESE-2P is preparing 11 open-source Solid Earth (SE) community flagship codes in terms of performance, scalability, deployment, and portability across current EuroHPC pre-exascale systems and emerging hardware architectures. During the project, the flagship codes will be used to farm a new generation of 9 Pilot Demonstrators (PDs) and to run 15 derived Simulation Cases (SCs) that address both capability and capacity computational challenges. PDs reaching the highest Technology Readiness Level (TRL) will also serve to enable services on socially-relevant aspects of geohazards, like urgent

computing for disaster response, early warnings, and hazard assessments. The 11 flagship codes (including satellite software components, mini-apps, and other tools derived from codes) and the 9 PDs and related SCs will be the main sources of project digital assets, which can be grouped in the following 4 categories (see Table 1):

1. **Software digital assets** (category 1), including the flagship source codes themselves but also the scripts and codes used to execute the workflows and service-related use-cases and tests implemented in WP4 and WP6 respectively.
2. **Scientific data digital assets** (category 2), including observations, test datasets (e.g. generated during the evaluation of code performances and co-design activities in WP2/WP3, flagship code audits, scalability tests, performance monitoring results, or co-design with the mini-apps), and (larger) datasets generated during the execution of the SCs in WP6.
3. **Metadata digital assets** (category 3) describing the categories above and a machine-actionable version of this DMP (maDMP) that will be included in deliverable D1.4 (final version of the DMP, due by month M34).
4. **Other digital assets** (category 4), including the project deliverables and a miscellaneous of internal reports and documents not directly included in the project deliverables list.

DA category	DA code	DA name	Comments
Software	1.1	Codes	Flagship applications (source codes)
	1.2	Workflows	Resulting from task T4.1
	1.3	Services	Resulting from task T6.4
Scientific datasets	2.1	Experimental	Data coming from observations, measurements, or produced by detectors/sensors or by any other experimental device
	2.2	Benchmark	Datasets (experimental or synthetic) used to validate models and/or their performance
	2.3	Sinthetic	Data generated by a simulation, particularly by the SCs
Metadata	3.1	Catalogues	Any (meta)data describing software and scientific data properties
	3.2	maDMP	Machine-actionable DMP
Other	4.1	Internal reports	Meeting minutes, internal notes to document the evolution of the project, calendar, resources management, mailing lists, etc

	4.2	Documents	Any documentation, either public or private, such as code documentation, technical notes, etc, not directly mentioned in the project deliverables list
	4.3	Deliverables	Project output documents

Table 1. Summary of the different categories of Digital Assets (DAs) included in the ChEESE-2P DMP.

1.2. Types of repositories

ChEESE-2P will place the DAs across different open repositories depending on the needs and requirements. Table 2 summarises the main characteristics of the repositories and Figure 1 sketches the relationship between the DAs and the type of repositories (centralised or distributed) where they will be deposited.

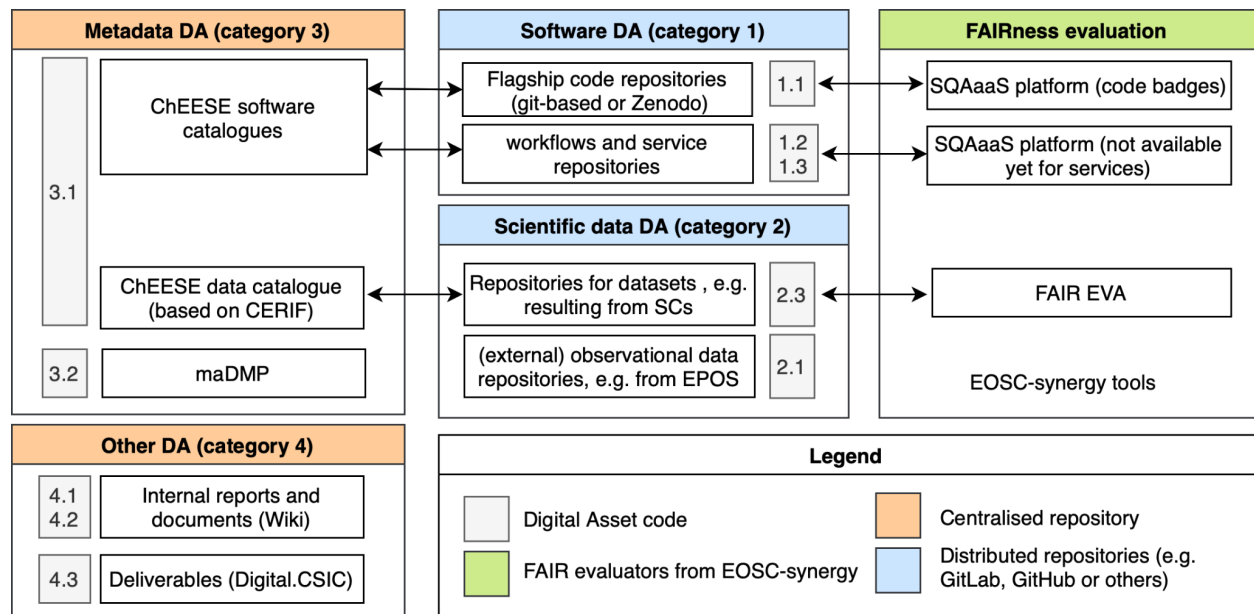


Figure 1. Digital asset categories and their relationship with metadata catalogues and the EOSC tools that will be used to evaluate the assets quality and FAIRness level (green box).

Property	git-based (GitLab/GitHub)	Zenodo	Digital.CSIC	Other
Open	yes	yes	yes	case-based
DevOps (versioning, CI/CD)	yes	no	no	no
DOI	no	yes	yes	possible
large storage	no	no	no	yes
Proprietary (under full control)	possible (requires local installation)	no	yes	yes
long-term storage	unknown	unknown	yes	yes
Main purpose	software	software datasets	documents	(large) datasets

Table 2. List of repositories to be used to host the DAs.

1.3. Structure of this document

The rest of this document presents the initial version of the ChEESE-2P DMP and it is structured as follows:

- **Section 2** contains the DMP for software digital assets (i.e. DA category 1 in Figure 1), including the distributed software repositories and a summary of the EOSC-synergy Software Quality Assessment as a Service (SQaaS) utility that will be used to evaluate the quality and level of FAIRness of the software repositories.
- **Section 3** contains the DMP for the scientific datasets (i.e. DA category 2 in Figure 1), including the distributed data repositories and an overview of the EOSC-synergy FAIR-EVA and the F-UJI utilities.
- **Section 4** contains the DMP for metadata (i.e. DA category 3 in Figure 1), including the metadata models for software/datasets and related catalogues, which will be based on extending the Common European Research project Information Format (CERIF) metadata schema, the same used in the European Plate Observing System (EPOS) research infrastructure. Section 4 also presents the concept of machine-actionable DMP.
- **Section 5** contains the DMP for other types of digital objects (i.e. DA category 4 in Figure 1).
- Finally, **Section 6** sketches the timeline towards the final version of this DMP (i.e. towards deliverable D1.4 due by month M34).

2. Management of software assets

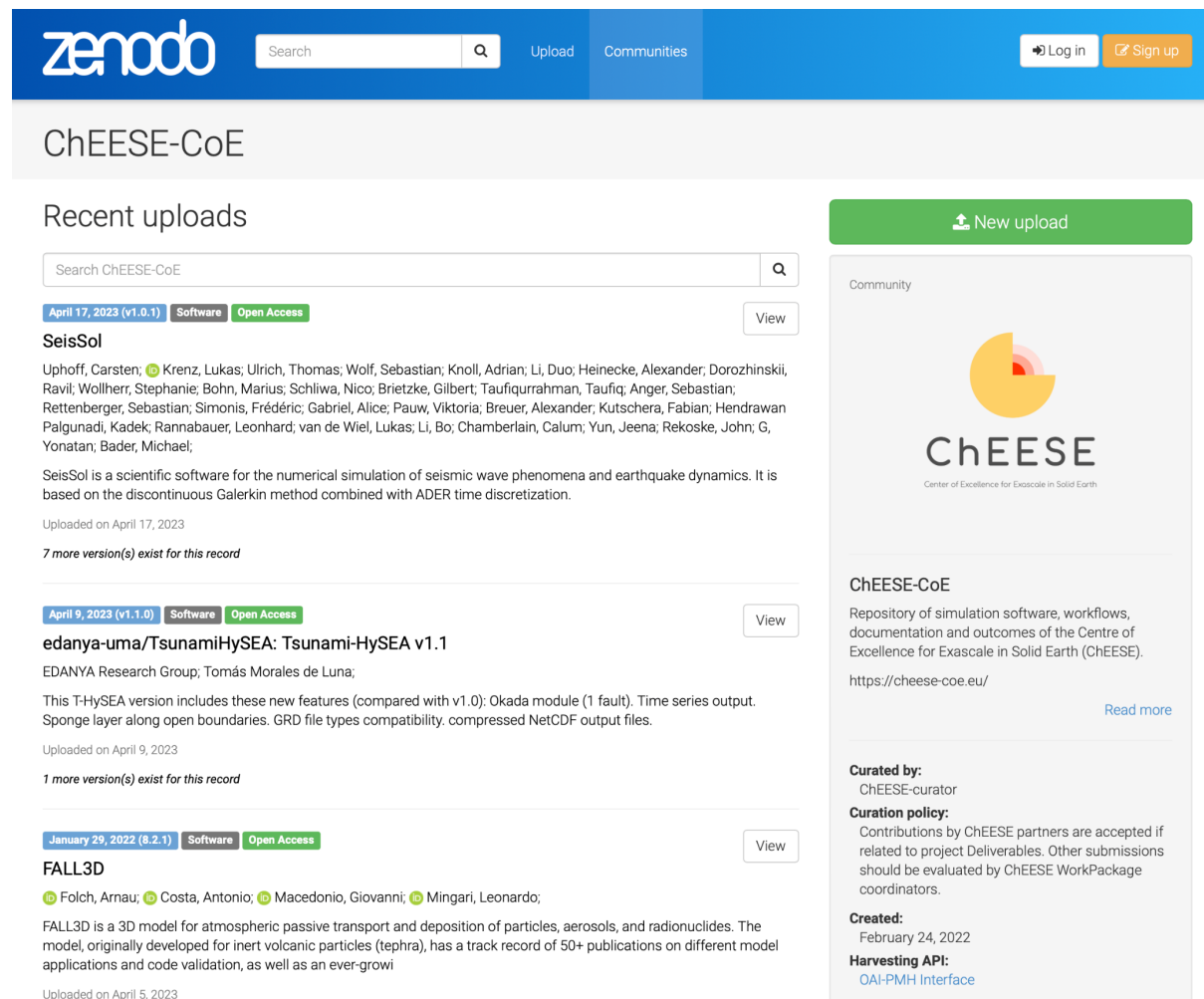
2.1. Software repositories

ChEESE-1P (2018-2022) carried out activities to improve the accessibility to computing applications including the creation of a Zenodo collection public repository (Figure 2) to store flagship code releases, workflows, and other satellite simulation toolkits. This single entry point centralised repository was a first step towards a potential integration of ChEESE products as a digital Research Infrastructure, into the EPOS ERIC, and will be leveraged by the Geo-INQUIRE project (www.geo-inquire.eu) as a basis to offer virtual and/or transnational access to ChEESE-related software and workflows. However, in the context of ChEESE-2P, it is important to consider the following aspects:

- ChEESE-2P has 5 new flagship codes that were not present in the first project phase (Tandem, OpenPDAC, LaMEM, pTatin3D, and Elmer/ICE).
- All the ChEESE-2P flagship codes also have “native” and independent git-based public repositories (on GitHub/GitLab/bitbucket, see Table 3 for details) which, in most cases, are also accessible through dedicated web portals and are consolidated references for the SE scientific community.
- A major advantage of the Zenodo repository is that it allows registering two persistent identifiers (two DOIs), one representing a specific version of the code and another representing all of the versions of it. The DOI associated with specific code versions (i.e. with a given code release or commit tag) can be used in citations to ensure reproducibility whereas, in contrast, the generic DOI is typically used for version-independent code citations. However, note that, given the attribution of a persistent identifier, the Zenodo repository contains only “frozen” versions (releases) of a code (production versions). This strategy is unpractical during code development stages and/or for the CI/CD DevOps that ChEESE will implement in the EuroHPC systems in collaboration with the CASTIEL-2 Coordination Support Action (CSA), for which the “native” git-based code repositories are necessary (e.g. to implement CI/CD pipelines through GitLab/GitHub runners).
- Finally, it may also happen that some components of the software repositories (beyond the codes) are not open-source or might be not yet publicly available during the project development phase (embargo).

For all these reasons, ChEESE-2P will maintain this duplicity of software repositories, relying on the native git-based repositories for the development and for the CI/CD DevOps actions and on the Zenodo collection (with associated DOIs) for the current (and future) production code releases. Whichever the case, all software repositories (native and Zenodo) will be linked to a centralised software catalogue built upon the same software metadata model (see Section 4.1.1). Metadata for the software components provides a way to achieve its full identification, thus making software citation viable and is key towards

preservation, discovery, reuse, and attribution of the software components. The metadata catalogue will be machine-actionable (JSON file) and updated when needed, e.g. in case of a new code version release.



The screenshot shows the Zenodo interface for the ChEESE-CoE community. The header includes the Zenodo logo, a search bar, and links for Upload, Communities, Log in, and Sign up. The main content area is titled 'ChEESE-CoE' and features a 'Recent uploads' section. This section lists three software releases: 'SeisSol' (April 17, 2023), 'edanya-uma/TsunamiHySEA: Tsunami-HySEA v1.1' (April 9, 2023), and 'FALL3D' (January 29, 2022). Each entry includes a search bar, version information, a 'View' button, and a brief description of the software. On the right side, there is a 'New upload' button and a community profile for 'ChEESE-CoE', which includes a description of the repository, a link to the website, and information about curation and creation.

Figure 2. Screenshot of the ChEESE-1P Zenodo collection repository (see <https://zenodo.org/communities/cheese-coe>).

Code num	Code name	Lead	Platform	Code Repository
1	SeisSol	LMU	github	https://github.com/SeisSol/SeisSol
2	SPECFEM3D	CNRS	github	https://github.com/SPECFEM
3	ExaHyPE	TUM	gitlab	https://gitlab.lrz.de/exahype
4	Tandem	LMU	github	https://github.com/TEAR-ERC/tandem
5	xSHELLS	CNRS	bitbucket	https://bitbucket.org/nschaeff/xshells
6	HySEA	UMA	github	https://github.com/edanya-uma/TsunamiHySEA
7	FALL3D	CSIC	gitlab	https://gitlab.com/fall3d-suite/fall3d/
8	OpenPDAC	INGV	github	https://github.com/demichie/OpenPDAC-10
9	LaMEM	UM	gitlab	https://github.com/UniMainzGeo/LaMEM
10	pTatin3D	SU	bitbucket	https://bitbucket.org/ptatin/ptatin3d
11	ELMER/ICE	CSC	github	https://github.com/ElmerCSC/elmerfem

Table 3. Current native git-based repositories for the ChEESE-2P flagship codes. Some specific releases for codes that were in ChEESE-1P are also mirrored in the ChEESE Zenodo collection.

2.2. The EOSC-synergy SQAaaS platform

The EOSC-Synergy project (2019-2022; funded under the H2020 topic INFRA-EOSC-05-2018-2019) defined a quality-driven framework aimed at fostering the adoption of EOSC software and services, identified widely adopted best practices for software and services and, finally, compiled them into quality criteria that are both open and technology agnostic. The quality assessment tools developed by EOSC-Synergy (Orbiz et al., 2022a, 2022b) can be leveraged through the so-called Software and Services Quality Assurance as a Service platform (SQAaaS), making quality verification and validation easily accessible to developers through a web user interface and API. In addition, the underlying Jenkins Pipeline Library (JePL) component of the SQAaaS can be used separately to create and execute CI/CD pipelines enabling adherence to the quality baselines without the need to use the complete platform. The SQAaaS platform harnesses the potential of the open badges specification to issue digital badges rewarding quality achievements and compliance to the quality baselines. EOSC-Synergy has three types of digital badges covering software, services and FAIR data (Figure 3) and each badge type has three levels (bronze, silver and gold) certifying different levels of compliance (Figure 4).

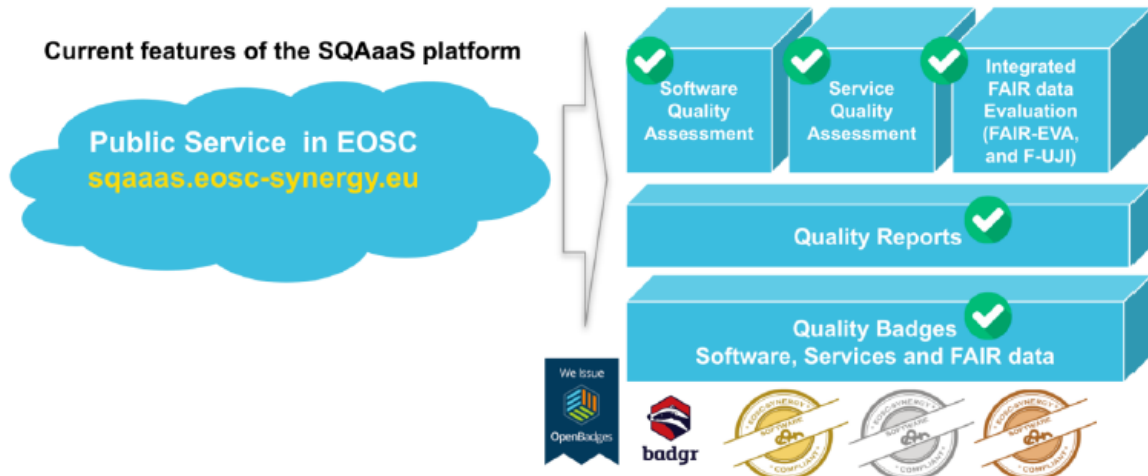


Figure 3. Concept of the EOSC-synergy SQAaaS platform. The integrated capabilities of SQAaaS for quality assessment are not only for software, but also for services and to support FAIR data evaluation and rewarding by issuing digital badges. There is no similar tool in the research or in the commercial sector with such integrated capabilities.

	Bronze	Silver	Gold
Accessibility (QC.Acc)	✓	✓	✓
Code Management (QC.Man)			✓
Code Metadata (QC.Met)		✓	✓
Code Style (QC.Sty)			✓
Code Workflow (QC.Wor)			✓
Delivery (QC.Del)			✓
Documentation (QC.Doc)	✓	✓	✓
Licensing (QC.Lic)	✓	✓	✓
Security Static Analysis (QC.Sec)			✓
Unit Testing (QC.Uni)			✓
Versioning (QC.Ver)		✓	✓

<https://indigo-dc.github.io/sqa-baseline/>



Figure 4. List of criteria to award the three types of badges (gold, silver and bronze) in the EOSC-synergy SQAaaS.

As a self-imposed seal of quality, ChESEE-2P requires that all code repositories must have a silver/gold badge before starting its CI/CD on the EuroHPC systems. This has been pioneered by the FALL3D code repository, which has been already awarded with a gold badge (Figure 5) and will serve as a model for the rest of the codes (see DMP implementation timeline in Section 7).

FALL3D-suite/FALL3D

Congratulations! The following badge has been awarded



Code Accessibility	100
Documentation	80
Licensing	100
Code metadata	100
Security	100
Code Style	100
Versioning	100

Bronze badge ✓ ✓ Licensing ✓ Code Accessibility ✓ Documentation	Silver badge ✓ ✓ Code metadata ✓ Versioning	Gold badge ✓ ✓ Security ✓ Code Style
Code Accessibility ✓ ✓ QC.Acc01 Is the source code managed with a Version Control System? ✓ Source code uses Git for version control More Info	Code metadata ✓ ✓ QC.Met01 Is the software citable? ✓ Software metadata successfully validated More Info	Licensing ✓ ✓ QC.Lic01 Is code licensed to be freely used, modified and distributed by others? ✓ An Open Source license found in the code repository: GPL-3.0 More Info ✓ QC.Lic01.1 Is LICENSE file placed in the root path of the code repository? ✓ LICENSE file is visible at the root path of the code repository: LICENSE More Info ✓ QC.Lic02 Does the code use a license approved by Open Source Initiative? ✓ License GPL-3.0 is approved by the Open Source Initiative More Info ✓ QC.Lic02.1 Does the code use an Open Source Initiative's popular-recognized license? ✓ License GPL-3.0 is listed under the Open Source Initiative popular category More Info

Figure 5. Example of the FALL3D repository metrics obtained following the EOSC-synergy SQAaaS platform analysis, which yielded a gold badge award.

3. Management of scientific data

3.1. Data repositories

The Simulation Cases (SCs) resulting from executing the PDs in production runs (see Table 4) will be the main sources of project datasets, together with smaller datasets resulting from WP2/WP3 code development activities. Datasets will be stored in distributed repositories managed by the data owners and, in some cases, integrated also in EPOS as Thematic Core Services (TCS). Similar to software, datasets will be linked to a centralised ChESEE data catalogue based on the CERIF standard metadata model (see Section 4.1.2), thereby providing a mechanism for end-users (or programs) to discover, understand and obtain access to the datasets based on their characteristics described by attributes.

SC num	SC name	Type of outcome
SC1.1	Anthropogenic noise and seismicity, wave propagation at higher frequencies in Rade d'Hyères	open datasets
SC1.2	Next generation of earthquake simulators for the Hellenic arc and Iceland	open datasets (earthquake catalogues)
SC1.3	Physics-based PSHA and urgent UQ for Iceland	UC service enabling
SC2.1	Seismic tomography of the Adriatic-Ionian region	open datasets
SC3.1	The Global Tsunami Hazard Map	TCS-TSU service in EPOS
SC4.1	Multi-source local high-resolution scenario with emphasis on the 2022 Tonga case	UC service enabling

SC5.1	Towards the European tephra hazard map	TCS-VO service in EPOS
SC5.2	Volcanic dispersal at local scale: a reanalysis of the 2021 La Palma case	UC service enabling
SC6.1	Urgent high-resolution, 3D multiphase flow simulation of phreatic eruptions at Vulcano	UC service enabling
SC6.2	Long-term probabilistic hazard maps for phreatic eruptions at Vulcano island	open datasets
SC7.1	MHD turbulence closure models in planetary cores using AI	open datasets
SC8.1	Magmatic systems and induced landslides: the Etna case	open datasets
SC8.2	3D evolution of tectonic deformation as driver of seismic hazard in the Hellenic arc	open datasets
SC9.1	Eastern Skaftá cauldron GLOF simulation, Iceland	open datasets
SC9.2	Calving glacier hazard in Greenland	open datasets

Table 4. List of the Simulation Cases (SC) resulting from executing PDs in production runs and expected results. Nomenclature: SC1.2 means second SC derived from PD1, SC6.1 means first SC derived from PD6, and so on.

3.2 The EOSC-synergy FAIR EVA tool

The EOSC-Synergy project developed a framework for FAIR data evaluation following the indicators from both the RDA FAIR Maturity Working Group and the FAIRsFAIR project. The FAIR Evaluator, Validator and Advisor (FAIR-EVA; DOI: 10.20350/digitalCSIC/14559) developed in EOSC-Synergy is an assessment tool that implements the RDA FAIR indicators (Table 5) and checks the FAIRness level of digital objects from different repositories or data portals. To this purpose, FAIR-EVA requires the object identifier (preferably persistent and unique identifier) and the repository to check. The tool is modular and extensible allowing the creation of plugins to access and assess different types of repositories and metadata and it is arguably the most comprehensive tool for automated assessment of FAIR compliance. FAIR EVA can be deployed as a stand-alone application or in a docker container, but it is also implemented as a web service with an API that manages the evaluation and a web interface to facilitate accessing and user-friendliness. FAIR EVA will be used to assess the FAIRness level of all ChEESE-2P datasets, thus providing an objective KPI for FAIRness (Figure 6).

Category			Condition
Findable	F1		(Meta) data are assigned globally unique and persistent identifiers

	F2		Data are described with rich metadata
	F3		Metadata clearly and explicitly include the identifier of the data they describe
	F4		(Meta)data are registered or indexed in a searchable resource
Accessible	A1		(Meta)data are retrievable by their identifier using a standardised communication protocol
		A1.1	The protocol is open, free and universally implementable
		A1.2	The protocol allows for an authentication and authorisation where necessary
	A2		Metadata should be accessible even when the data is no longer available
Interoperable	I1		(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
	I2		(Meta)data use vocabularies that follow the FAIR principles
	I3		(Meta)data include qualified references to other (meta)data
Reusable	R1		(Meta)data are richly described with a plurality of accurate and relevant attributes
		R1.1	(Meta)data are released with a clear and accessible data usage licence
		R1.2	(Meta)data are associated with detailed provenance
		R1.3	(Meta)data meet domain-relevant community standards

Table 5. Categories of FAIRness indicators in the FAIR EVA tool (in agreement with the RDA indicators). By default, categories are equally weighted when assessing quantitatively the level of FAIRness (from 0 to 100%). However, weights for the different categories can also be user-defined if needed.



Figure 6. Example of FAIR EVA tool evaluation (FAIRness level of a particular DA).

3.3. The F-UJI assessment tool

F-UJI is a tool developed under the umbrella of the [FAIRsFAIR](#) project (GA No 831558) to assess FAIRness of research data objects at the dataset level based on the FAIRsFAIR Data Object Assessment Metrics (Devaraju and Huber, 2020). A web service exists to check how the output of F-UJI test results look like, but the code (service) can also be installed and run locally from the F-UJI GitHub repository.

4. Management of metadata

4.1. Metadata catalogues and descriptors

As shown in Figure 1, the project will create software and data catalogues linked to the respective distributed repositories. These catalogues will be stored in a centralised repository and will be written in a data exchange format (preferably JSON or XML) following a standard metadata scheme, so that will be machine-actionable and syntactically interoperable. The underlying metadata model will be based on (and extended when necessary) the Common European Research Information Format (CERIF), a comprehensive information model for the domain of scientific research developed by the CERIF and CRIS Architectures Task Group of euroCRIS. This is the same base metadata model describing existing digital assets in EPOS and, consequently, metadata produced during the project will be compatible and re-used within the EPOS-ERIC.

4.1.1. Metadata model for software

In ChEESE-1p, the metadata elements to be included in the software catalogue (for both native and Zenodo repositories) were identified. They are shown in Table 6.

Element Name	Child Name	Type	Description
Identification	DOI	string	Digital Object Identifier (for Repo)
	Title	string	Code name or acronym
	Version	string	Release number
	PublicationDate	date	Date of last release
	Repository	string	Link to download or execution page
	isInteractive	boolean	Can be run online
	isPublished	boolean	Published on the web portal
DataIN/ DataOUT	Type	text	Description of the datum
	Format	string	Standard format type (if available)
	Extension	string	File extension
Information	Language	string	Main programming language
	Description	text	Model description
	Algorithm	text	Algorithm Description
	Implementation	text	Implementation description
Documentation	WebPage	string	Link to web page (if available)
	UserGuide	string	Link to User Guide (if available)
	ReferenceGuide	string	Link to Reference Guide (if available)
	Related Identifiers	string	Publication(s) of reference
	Gallery	string	Link to Image and Video gallery (if available)
Copyright	Authorship	string	Name(s) of the author(s)
	Author ID Type	integer	1. ORCID, 2. ResearcherID, 3. Scopus
	Author ID	string	Author ID
	Affiliation	string	Institution
	Licence	string	Text or Link to licence page
	Contacts	string	contact email

Table 6. Elements of the software metadata. In red, the “mandatory” elements requested in Zenodo. In orange, the “recommended” ones. All child elements in black can be inserted in one of the red/orange categories.

Such a scheme will be reviewed, e.g. following the guidelines proposed in the framework of the Research Data Alliance (<https://doi.org/10.15497/RDA00068>) for software FAIRness.

4.1.2. Metadata model for datasets

A specific objective of ChEESE-2P is to align and liaise with other projects, mission initiatives (Digital Agenda, Destination Earth), and Research Infrastructures, particularly in the EPOS catalogue. A major challenge in EPOS is

the integration of multi-disciplinary, multi-organisational, distributed resources and community assets into a single overarching Research Infrastructure, the so-called EPOS Integrated Core Services (ICS). ICS aggregate and harmonise descriptions of datasets, data products, software and services from different domain-specific services - the Thematic Core Services (TCS). The TCS adopt heterogeneous formats, vocabularies, protocols and standards to represent and make their resources available. Digital datasets produced by the ChEESE-2P Simulation Cases can be integrated in the EPOS ecosystem in the form of Level-3 data products (integrated data products coming from complex analyses or community shared products). One of the two main ingredients used by EPOS to achieve interoperability is metadata, that describes services and assets provided by the Thematic Communities. The exchange of metadata between ICS and TCS is crucial to achieve integration and interoperability in EPOS. In order to capture, organise and harmonise information from different sources and to enable semantic interoperability, a data model has been developed and adopted, namely EPOS-DCAT-AP. It extends and builds on the W3C standard Data Catalogue Vocabulary (DCAT), designed to facilitate interoperability between data catalogues published on the Web.

EPOS-DCAT-AP is available on GitHub and includes a Unified Modeling Language (UML) diagram, ontology definition, examples and more details. The actual EPOS catalogue uses as its format CERIF (as recommended by the EU to Member States for research information). CERIF is maintained by EuroCRIS and used widely (also outside Europe) and is embedded in products from Elsevier and Thomson-Reuters as well as being used in OpenAIRE. In EPOS, the metadata catalogue is populated from the heterogeneous metadata formats used within the TCS via EPOS-DCAT-AP in an ingestion pipeline that includes dynamic update by specified TCS managers within a secure environment. The EPOS-DCAT-AP specifications already include concepts to describe software, in particular Software Applications and Software Source Code, which are suited to describe HPC software (not run online). These will be adopted to implement the ChEESE-2P catalogue.

4.2. Machine-actionable DMP

The Common Standards Working Group (WP) of the Research Data Alliance (RDA) has proposed a common standard for machine-actionable Data Management Plans or maDMPs (Miksa et al., 2019). As opposed to the traditional DMPs (i.e. free-form text documents describing the data that is used and produced during the course of a project or research activity), maDMPs express the same information in traditional DMPs but in a machine-actionable way that allows for automatic exchange, integration, validation of information, and interoperability between systems producing or consuming maDMPs. This provides a specification for the whole research activity lifecycle, so that the objectives, key performance indicators (KPIs) and key exploitable results (KERs) specified in a maDMP can be utilised in Science Knowledge Graphs (SKG) such

as that of OpenAIRE (<https://graph.openaire.eu>) and integrated with assessment of FAIRness such as the FAIR Implementation Profiles (FIP) (Schultes, 2020) or other assessment mechanisms such as those of the EOSC-synergy project.

Currently, there is much research activity in various initiatives around maDMPs so it is fair to say that, at present, they are not yet in a full production status. However, we have the intention to develop further the metadata catalogue schema to accommodate the evolving requirements from the different SE research communities involved in ChEESE-2P and the improved documentation of DAs utilised in the project. To this purpose, the plan is to start with the RDA DMP Common Standard metadata application profile (Miksa et al., 2019), which is a JSON schema general enough to cover a wide range of use cases and which does not constrain to any funder-specific requirements.

5. Management of other data

Regarding other data types (i.e. DA category 4 in Figure 1), management will be as follows:

- Deliverables will be stored at the CSIC (ChEESE leading partner) institutional repository DIGITAL.CSIC, each with an assigned DOI. This will be done only after the different project review periods, once deliverables are approved by reviewers and the Project Officer (i.e. after an eventual revision).
- The rest of DAs are of internal nature, and will be managed by the project Wiki. These assets are not actually contemplated by this DMP.

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 potentially negative impacts, the project may eventually distribute some simulation 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 definition.

7. Timeline for the DMP

The time diagram below illustrates the timings foreseen to implement the different aspects of the final DMP.

	M6	M12	M18	M24	M30	M34	M48 end
SQAaaS for code repositories							
Software catalogue (metadata)							
Production runs (PDs, SCs)							
Data catalogue (metadata)							
FAIR EVA for datasets							
maDMP							
Other DAs (e.g. deliverables)							
Final DMP (D1.4)						x	

References

- Devaraju, A., Huber, R. (2020): F-UJI - An Automated FAIR Data Assessment Tool (v1.0.0). Zenodo. <https://doi.org/10.5281/zenodo.4063720>
- Berman, F. (2019): The Research Data Alliance --The First Five Years, Supplement to: Berman, F., & Crosas, M. (2020). The Research Data Alliance: Benefits and Challenges of Building a Community Organization. Harvard Data Science Review, 2(1), doi:10.1162/99608f92.5e126552.
- Miksa, T., Walk, P., & Neish, P. (2019): RDA DMP Common Standard for Machine-actionable Data Management Plans, doi: 10.15497/rda00039.
- Orviz, P., López García, A., Duma, D.C., Donvito, G., David, M., Gomes, J., Campos, I., Moltó, G., Tykhonov, V. (2022a): A set of common software quality assurance baseline criteria for research projects, doi: 10.20350/digitalCSIC/12543.

- Orviz, P., David, M., Gomes, J., Pina, J., Bernardo, S., Campos, I., Moltó, G., Caballer, M., Tykhonov, V. (2022b): A Set of Common Service Quality Assurance Baseline Criteria for Research Projects, doi: 10.20350/digitalCSIC/12533.
- Schultes, E., Magagna, B., Hettne, K.M., Pergl, R., Suchánek, M., Kuhn, T. (2020): Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence. In: Grossmann, G., Ram, S. (eds) Advances in Conceptual Modeling. ER 2020. Lecture Notes in Computer Science(), vol 12584. Springer, Cham, doi:10.1007/978-3-030-65847-2_13.

Annex 1. Glossary of concepts

- **Common European Research Information Format (CERIF)**

The Common European Research Information Format is a standard developed and maintained by the euroCRIS CERIF Task Group as a means to ensure the maximum possible level of system interoperability.

- **Creative Commons (CC)**

Creative Commons (<https://creativecommons.org/>) is a global nonprofit organisation that enables sharing and reuse of creativity and knowledge through the provision of free legal tools. There are 6 different Creative Commons licences: CC BY, CC BY-SA, CC BY-NC, CC BY-ND, CC BY-NC-SA, CC BY-NC-ND plus the public domain (CC 0). The letter pairs indicate conditions for use: attribution (BY), non-commercial (NC), sharealike (SA) and non-derivative (ND). In particular, CC Attribution (CC BY) is the most open licence, allowing the user to redistribute, to create derivatives, such as a translation, and even use the publication for commercial activities, provided that appropriate credit is given to the author (BY) and that the user indicates whether the publication has been changed.

CREATIVE COMMONS LICENSES		COPY & PUBLISH	ATTRIBUTION REQUIRED	COMMERCIAL USE	MODIFY & ADAPT	CHANGE LICENSE
	PUBLIC DOMAIN	✓	✗	✓	✓	✓
	CC BY	✓	✓	✓	✓	✓
	CC BY-SA	✓	✓	✓	✓	✗
	CC BY-ND	✓	✓	✓	✗	✓
	CC BY-NC	✓	✓	✗	✓	✓
	CC BY-NC-SA	✓	✓	✗	✓	✗
	CC BY-NC-ND	✓	✓	✗	✗	✓

 You can redistribute (copy, publish, display, communicate, etc.)
  You have to attribute the original work
  You can use the work commercially
  You can modify and adapt the original work
  You can choose license type for your adaptations of the work

- **FAIR Implementation Profiles (FIP)**

A FIP is a collection of FAIR implementation choices made by a community of practice for each of the 4 FAIR Principles. Community-specific FAIR Implementation Profiles are themselves captured as FAIR datasets and are made openly available to other communities for reuse.

- **OpenAIRE**

OpenAIRE is a Non-Profit Partnership, to shift scholarly communication towards openness and transparency and facilitate innovative ways to communicate and monitor research.

- **Research Data Alliance (RDA)**

The Research Data Alliance is a global community-driven initiative (Foundation) engaging data producers, users and stewards that promotes open sharing and re-use of data.

- **Zenodo**

Zenodo (www.zenodo.org) is a repository for research data hosted at CERN, allowing users to share their data according to the EU policies and make them citable by their own DOI. Zenodo is open to all disciplines of science, and users can deposit any kinds of data, such as spreadsheets, figures, reports, posters, presentations, and software for free. The submission mask is very user-friendly and offers many options of customising metadata.