

ChESEE

Data Management Plan (DMP): status at M21

Second General Assembly
Reykjavík, 17-19 September 2024

Arnau Folch
Geociencias Barcelona (GEO3BCN-CSIC)

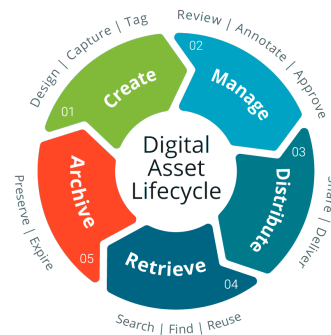


Project funded by EuroHPC under the grant agreement No 101093038.



ChEESE Data Management Plan (DMP)

- **DMP**: a methodology to cover and to document the whole life-cycle of the Digital Assets (DA) to be collected, processed, and generated during the lifetime of a project (and beyond)
- The **life-cycle of a DA** includes several phases:
 1. Create (generation, capture)
 2. Manage (review, annotation with metadata)
 3. Distribute (share and deliver through repositories)
 4. Retrieve and reuse (related to FAIRness of the DAs)
 5. Archive (longer term preservation)
- **Machine-actionable Data Management Plan** (ma-DMP):
 - DMP based on machine-actionable file formats (JSON, XML, etc)
 - Semantics based on controlled vocabularies and existing standards (**metadata model**)
 - Machine-actionable files allow machines performing automated queries (and services) to retrieve information from DMPs (e.g. identify investigators, DOIs, funding recognition, etc)
- **ChEESE task T1.5**: Data management plan and automatic FAIR validation
 - First version by M6 (**D1.2**: "First version of the DMP")
 - Final version by M34 (**D1.4**: "Final version of the DMP")

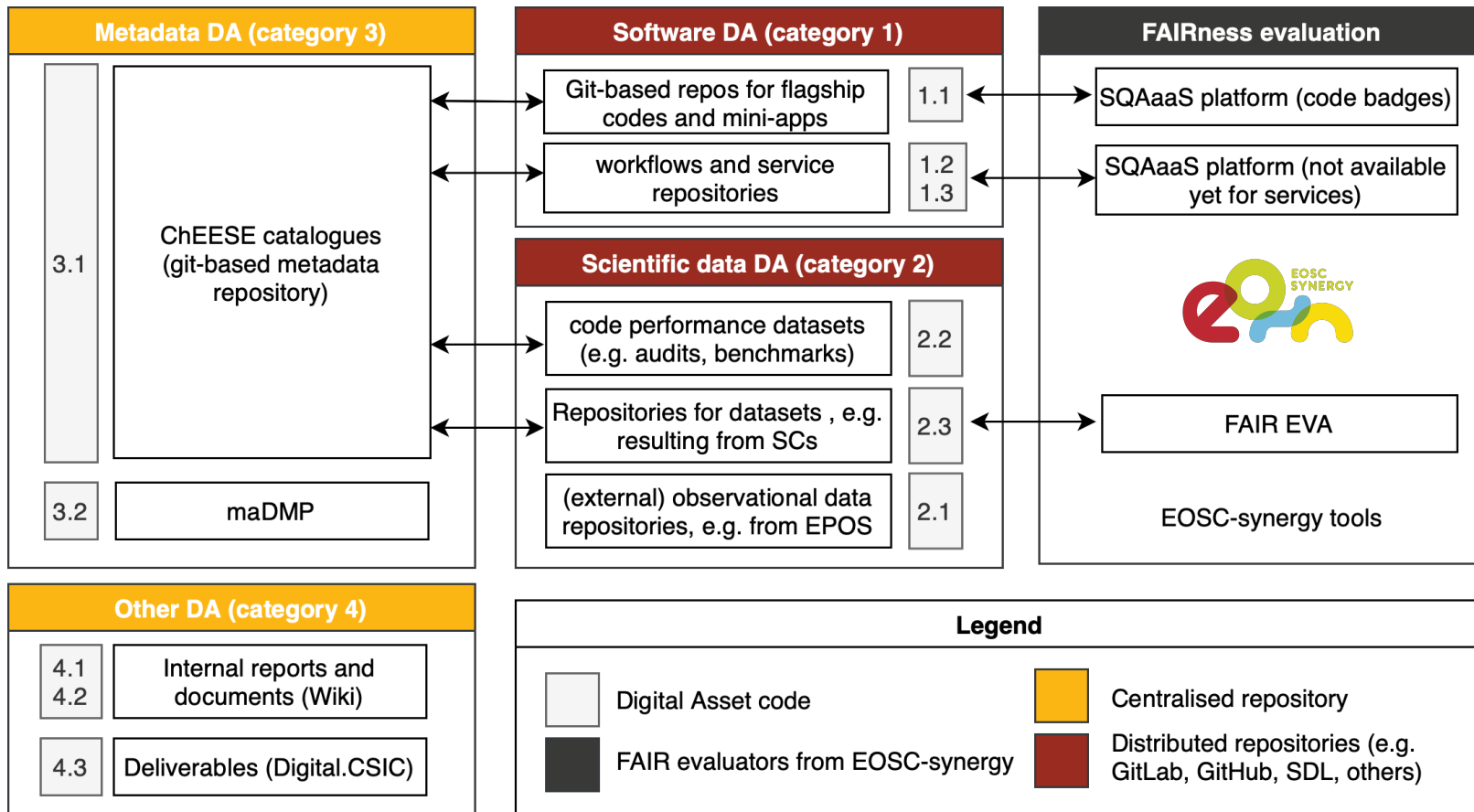


ChEESE DMP: 4 categories of Digital Assets

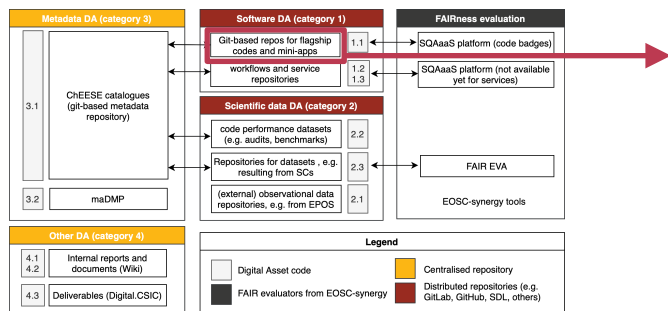
DA category		DA ID	DA name	Comments
1	Software	1.1	Codes	Flagship applications and mini-apps (source codes)
		1.2	Workflows	Resulting from T4.1
		1.3	Services	Resulting from T6.4
2	Scientific datasets	2.1	Observations	Data coming from observations, measurements, or produced by detectors/sensors or by any other experimental device
		2.2	Benchmark	Datasets used to validate models and/or their performance; mostly resulting from W2 activities (e.g. audits, performance portability, etc)
		2.3	Synthetic	Data generated by simulations, particularly from the 15 SCs
3	Metadata	3.1	Catalogues	Any (meta)data describing software and scientific data properties
		3.2	maDMP	Machine-actionable DMP
4	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



Types of repositories



Code (and mini-app) repositories: status at M21

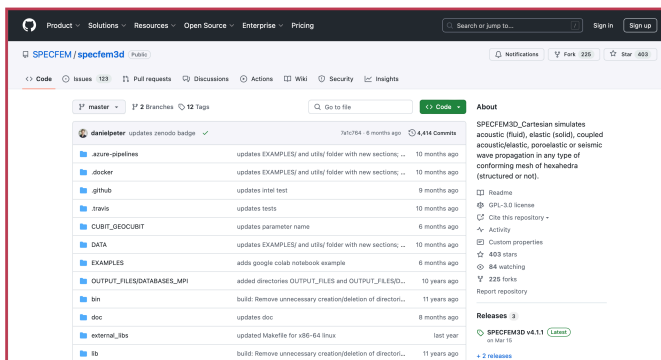


The 11 codes already in native git-based open repositories

- Code versioning
- Licence

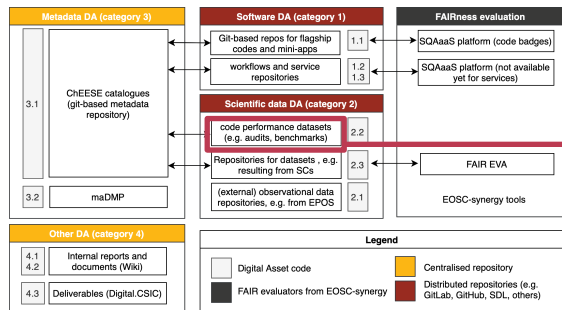


Status of the 7 mini-apps?



Code	Repo	Link
SeisSol	GitHub	https://github.com/SeisSol/SeisSol
SPECfem3D	GitHub	https://github.com/SPECfem
ExaHyPE	GitLab	https://gitlab.lrz.de/hpcsoftware/Peano
Tandem	GitHub	https://github.com/TEAR-ERC/tandem
xSHELLS	GitLab	https://gricad-gitlab.univ-grenoble-alpes.fr/schaeffn/xshells
HySEA	GitHub	https://github.com/edanya-uma/TsunamiHySEA
FALL3D	GitLab	https://gitlab.com/fall3d-suite/fall3d/
OpenPDAC	GitHub	https://github.com/demichie/OpenPDAC-11
LAMEM	GitHub	https://github.com/UniMainzGeo/LaMEM
pTatin3D	BitBucket	https://bitbucket.org/ptatin/ptatin3d/src/master/
Elmer/Ice	GitHub	https://github.com/ElmerCSC/elmerfem

Code performance repositories: status at M21



How to proceed? The example of FALL3D...

Tests

Subgroups and projects Shared projects

- Search (3 character minimum)
- Leonardo
 - Audit
 - Constant
 - Scalability
 - LUMI
 - audit
 - constant
 - scalability
 - MNS
 - Audit
 - scalability

Full test and report added

Heribert Pascual Saldanya authored 10 months ago

Name

test_audit_L00_16GPU
test_audit_L00_1GPU
test_audit_L00_2GPU
test_audit_L00_4GPU
test_audit_L00_8GPU
informe_L00_audit.pdf
output.json.out
test_audit_L00.json



Full test and report added

Heribert Pascual Saldanya authored 10 months ago

Name

..

FALL3D.gpu.cmd

FALL3D_16GPU.err

FALL3D_16GPU.out

test_audit.Fall3d.log

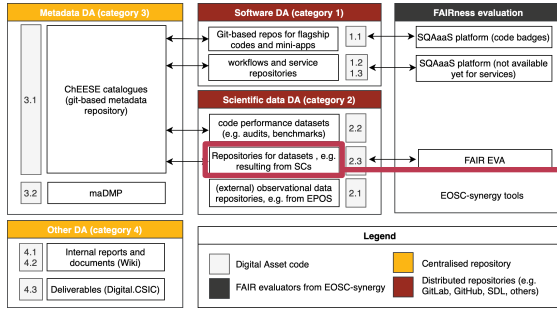
test_audit.grn

test_audit.inp

test_audit.src

```
{
  "test": {
    "type": "test_audit",
    "testName": "Arnaud Folch",
    "note": "Copy and modify this template in each new test_audit_ID folder",
    "comment": "Baseline version without Extrae instrumentation. At M09 of ChESEE-2P project"
  },
  "code": {
    "origin": "decube83/fall-3-d-gpu-cicd.git",
    "version": "8.2.1-gpu",
    "branch": "/cicd/hpascual",
    "commitID": "30349dce61c8ba3b3f20ea79041c543a1471f76",
    "commitAuthor": "decube83 <hpascual@geo3bcn.csic.es>",
    "commitDate": "Mon Nov 06 09:15:29 2023 +0100"
  },
  "system": {
    "host": "Leonardo@CINECA",
    "partition": "booster",
    "nodeCPU": "32 cores Intel Ice Lake ",
    "nodeGPU": "4 NVIDIA Ampere 64GB",
    "nodeRAM": "512 GB"
  },
  "modules": {
    "fortran": "nvhpc/23.1",
    "cuda": "cuda/11.8",
    "mpi": "openmpi/4.1.4--nvhpc--23.1-cuda-11.8",
    "hdf5": "hdf5/1.12.2--openmpi--4.1.4--nvhpc--23.1",
    "netcdf-c": "netcdf-c/4.9.0--openmpi--4.1.4--nvhpc--23.1",
    "netcdf-fortran": "netcdf-fortran/4.6.0--openmpi--4.1.4--nvhpc--23.1",
    "netcdf-parallel": "parallel-netcdf/1.12.3--openmpi--4.1.4--nvhpc--23.1",
    "zlib": "zlib/1.2.13--gcc--11.3.0"
  },
  "compilation": {
    "FCFLAGS": "FCFLAGS=-g -gpuspinned -cuda -fast -MinLine -acc",
    "FFLAGS": "FFLAGS=-g",
    "LDLAGS": "LDLAGS= -gpuspinned -cuda -acc",
    "enableGPU": "--with-gpuspinned",
    "enableACC": "--with-accyes",
    "precision": "--with-n4=no",
    "binary": "/Leonardo_work/EUHPD_002_008/TESTS/test_audit/shared/binaries/Fall3d.gpu.231106_r8.init.time.x"
  }
}
```

Repositories for datasets (mostly from SCs)



How to proceed from M24? **A good option** is the SDL@CINECA...

Simulation Data Lake Challenges

- Definition of the **Data Model**
- Interoperability through the **adoption of a common metadata schema**
- **Proximity of the data to the HPC clusters** for
 - re-use
 - postprocessing
 - comparison with new studies
 - training of AI algorithms

- Synergies with other significant Solid Earth European projects and other initiatives:

- **EPOS** research infrastructure that facilitates the use of data, data products, and facilities from the solid Earth science community in Europe. (<https://www.epos-eu.org/>)
- **DT-GEO** Digital Twin for GEOPhysical extremes (<https://dtgeo.eu/>)
- **ChEESE** Centre of Excellence (CoE) for Exascale in Solid Earth (<https://cheese-coe.eu/>)
- **Destination Earth** (<https://destination-earth.eu/>)
- **M@TE** Model Atlas of the Earth (<https://mate.science/>)
- **AUScope** Australian Geophysical Observing System (<https://www.auscope.org.au/>)

Next slides...



The Simulation Data Lake for geosciences (SDL@CINECA)

FAIRness principles

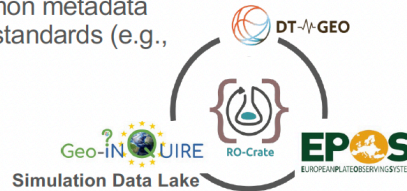
Data needs to be **findable** at different levels

- Whole datasets
- Simulation runs
- Single files

Enable to query the SDL in useful ways.
Metadata needs to include specific information.

Adoption of standards used by the community

Synergies with **DT-GEO** and **EPOS**.
Adoption of a common metadata schema and other standards (e.g., **RO-Crate**).



Data **accessible** at different levels:

- Whole datasets
- Simulation runs
- Single files

API or/and GUI to access data

Interoperable



Reusable

Datasets should include **all the data** and the information needed to **re-run simulations**

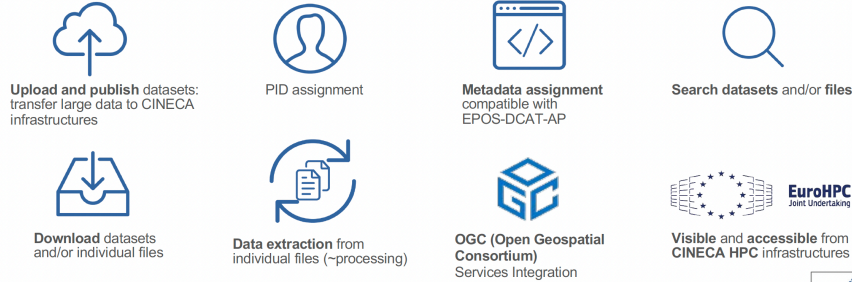
Datasets need to be well characterised



The Simulation Data Lake for geosciences (SDL@CINECA)

Simulation Data Lake functionalities

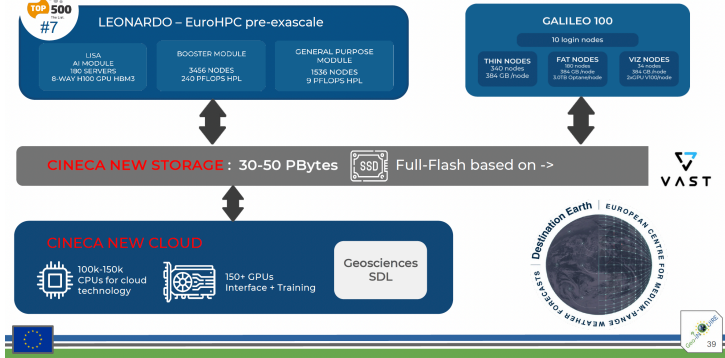
- Store/access simulation datasets, promoting data discoverability, reuse, and experiment repeatability
- Functionalities:



20

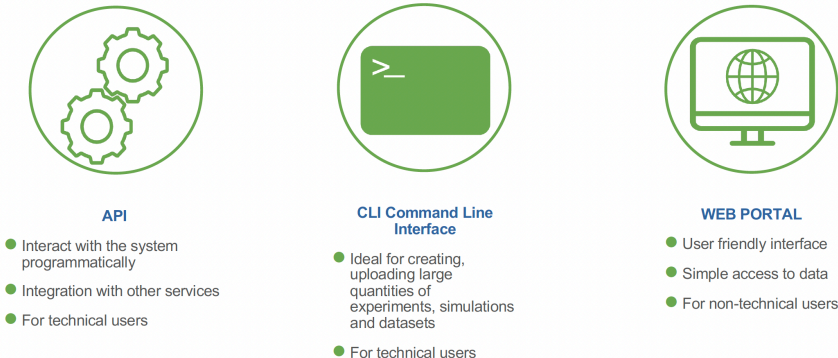
Slides from Gabriella Scipione (CINECA)

The Data close to the HPC resources

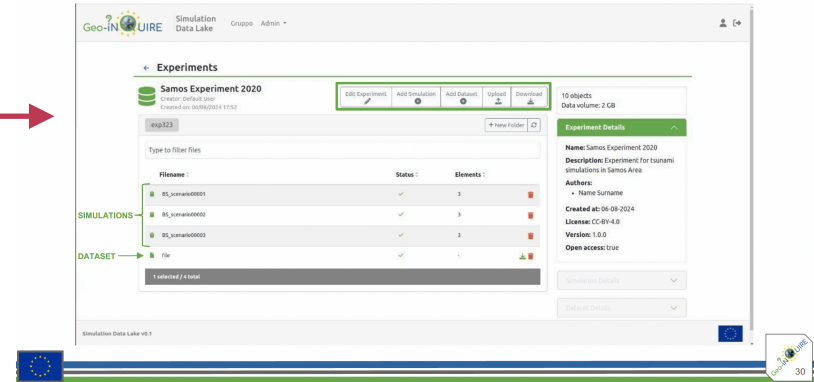


39

Simulation Data Lake Access

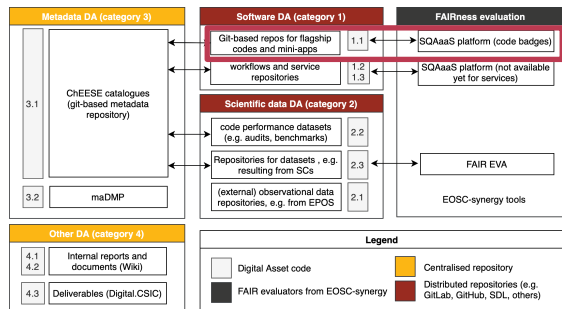


Web Portal: Experiment view



30

FAIRness of code repositories: status at M21



✓ The 11 codes have an EOSC-synergy badge from the SQAaaS (with room to upgrade to “gold” before M48)

⚠ Should we get badges for the 7 mini-app repositories?



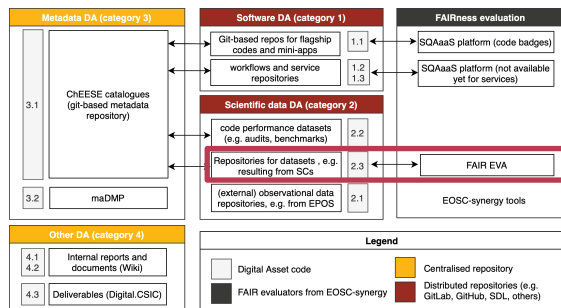
	Bronze	Silver	Gold
Accessibility (QC.Acc)	✓	✓	✓
Code Management (QC.Man)			✓
Code Metadata (QC.Met)		✓	✓
Code Style (QC.Sty)			✓
Code Workflow (QC.Wor)			✓
Delivery (QC.Del)			✓
Documentation (QC.Doc)	✓	✓	✓
Licensing (QC.Lic)	✓	✓	✓
Security Static Analysis (QC.Sec)			✓
Unit Testing (QC.Uni)			✓
Versioning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

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



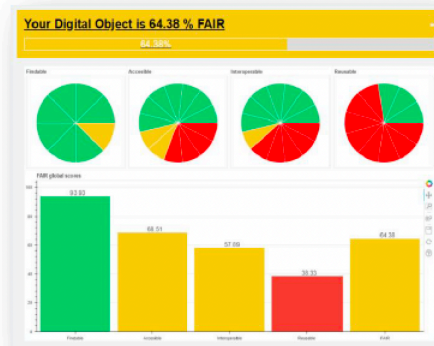
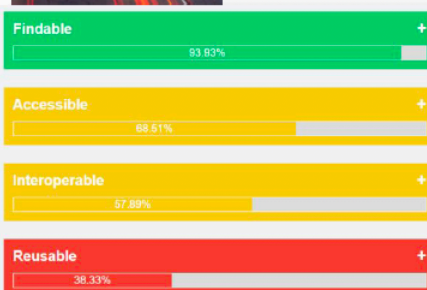
Name	SQAaaS	DOI	Version	License
SeisSol	sqaas software silver	DOI 10.5281/zenodo.11405512	release v1.1.4	License BSD 3-Clause
SPECFEM	sqaas software bronze	DOI 10.5281/zenodo.10823181	release v4.1.1	License GPL v3
Tandem	sqaas software silver		release v1.1.0	License BSD 3-Clause
xSHELLS	sqaas software silver	Reference		License GPL v3
HySEA	sqaas software silver	DOI 10.5281/zenodo.10956965	release v1.1.1	License GPL v2
FALL3D	sqaas software gold	DOI 10.5281/zenodo.6343786	release v8.3	License GPL v3
OpenPDAC	sqaas software silver	DOI 10.5281/zenodo.10721203	release v1.0.2	License GPL v3
LaMEM	sqaas software bronze	DOI 10.5281/zenodo.10718860	release v2.1.4	License MIT
pTatin3D	sqaas software silver	Reference		License GPL v2
ELMER_ICE	sqaas software silver	DOI 10.5281/zenodo.7892181	release release-9.0	License GPL v2
ExaHyPE	sqaas software bronze	DOI 10.5281/zenodo.6387677		License BSD 3-Clause

FAIRness of datasets (from PDs and SCs)



How to proceed from M24? **A good option** is the FAIR Evaluator-Validator-Advisor (FAIR-EVA) developed by EOSC-Synergy; an assessment tool that implements the RDA FAIR indicators

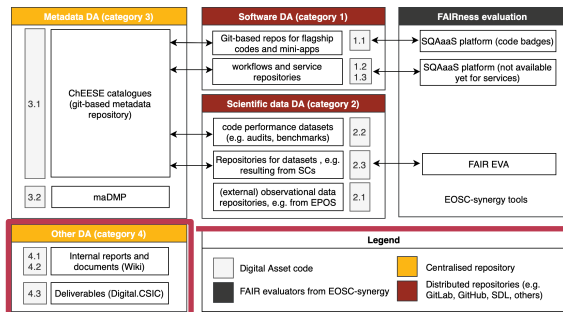
Category		Condition
Findable	F1	(Meta) data are assigned globally unique and persistent identifiers
	F2	Data are described with rich metadata
	F3	Metadata clearly and explicitly include the identifier of the data they describe
	F4	(Meta)data are registered or indexed in a searchable resource
Accessible	A1	(Meta)data are retrievable by their identifier using a standardised 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
	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



DT-GEO is extending the FAIR-EVA service to include EPOS-DCAT-AP metadata scheme



Repositories for other types of DAs: status at M21



Wiki in place and operative



DIGITAL-CSIC for permanent storage of deliverables (with DOI)

DIGITAL-CSIC / Recursos Naturales / Geociencias Barcelona (Geo3Bcn) / (Geo3Bcn) Informes y documentos de trabajo

Por favor, use este identificador para citar o enlazar a este ítem: <http://hdl.handle.net/10261/365841>

COMPARTIR / EXPORTAR:

OpenAIRE Visualizar otros formatos: MARC | Dublin Core | RDF | ORE | MODS | METS | DIDL | DATACITE

☐ Refman ☐ EndNote ☒ Bibtex ☐ RefWorks ☐ Excel ☐ CSV ☐ PDF ☐ DataCite ☐ Send via email



INVITAR A REVISIÓN POR PARES OPEN ACCESS REVIEW Module

Título:	ChEESE-2P Deliverable: D2.1 First code audit report
Autor:	Gracia, Jose; Lanucara, Piero; Bader, Michael; ChEESE-2P
Financiadores:	European Commission
Fecha de publicación:	2024
Serie:	D2.1
Descripción:	Copyright Members of ChEESE collaboration, 2023/2026.
URI:	http://hdl.handle.net/10261/365841
Aparece en las colecciones:	(Geo3Bcn) Informes y documentos de trabajo

Page view(s)
14
checked on 11-sep-2024

Download(s)
11
checked on 11-sep-2024

Google Scholar™
Check

1 resultat (0,06 s)



[\[PDF\] ChEESE-2P Deliverable: D2. 1 First code audit report](#)

J Gracia, P Lanucara, M Bader

2024 - digital.csic.es

Executive Summary

This document describes the result of the first performance assessment campaign of the project ChEESE-2P. The campaign serves three purposes:(1) identify performance bottlenecks to guide optimisation work,(2) set a performance baseline to monitor code evolution throughout the project duration, and (3) produce and document key performance indicators (KPI) related to the progress on flagship codes.

digital.csic.es

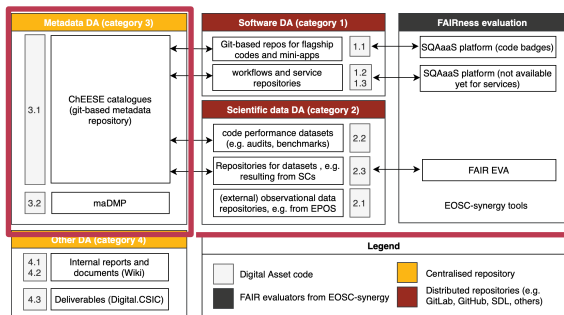
Desa Cita Artículos relacionados

Repositories for other types of DAs: status at M21

Mostrando resultados 1 a 8 de 8						
Derechos	Preview	Fecha Public.	Título	Autor(es)	Tipo	
		2024	ChEESE-2P Deliverable: D1.1 Project Management and Quality Guidelines	García Miraz, María José; Folch, Arnau    Beltrán Torres, Barbara; Montellier, Vadim; ChEESE-2P	informe técnico	
		2024	ChEESE-2P Deliverable: D1.2 First Data Management Plan (DMP)	Folch, Arnau   Esposti-Ongaro, Tomaso; Barsotti, Sara; Vinciarelli, Valerio; ChEESE-2P	informe técnico	
		2024	ChEESE-2P Deliverable: D2.1 First code audit report	Gracia, Jose; Lanucara, Piero; Bader, Michael; ChEESE-2P	informe técnico	
		2024	ChEESE-2P Deliverable: D3.1 Selected set of mini-apps and its requirements	Masini, A.; Raffin, Erwan; Hamitou, Okba; García, M.; Mantovani, F.; Amati, Giorgio; Pascual, H.; ChEESE-2P	informe técnico	
		2024	ChEESE-2P Deliverable: D7.1 Project Website	Vedia, Varvara; Gregorio, Rose; Zamora, Natalia; Filosa, Silvia; ChEESE-2P	informe técnico	
		2024	ChEESE-2P Deliverable: D7.2 Communication and Dissemination Plan	Vedia, Varvara; Gregorio, Rose; Pourhiet, Laetitia Le; Denamiel, Clea; ChEESE-2P	informe técnico	
		2024	ChEESE-2P Deliverable: D7.3 Collaboration plan with complementary projects	García Miraz, María José; Folch, Arnau    Beltrán Torres, Barbara; Lanucara, Piero; ChEESE-2P	informe técnico	
		2024	ChEESE-2P Deliverable: D7.4 Training Plan	Baghetti, María Paz; Vedia, Varvara; Folch, Arnau   Lanucara, Piero; García Miraz, María José; Kaus, Boris; ChEESE-2P	informe técnico	



Repositories for metadata (catalogues): status at M21



A skeleton (draft) exists at <https://gitlab.com/cheese5311126>



Need to discuss several aspects (including metadata model)



ChESEE

©2023 Center of Excellence for Exascale in Solid Earth - Second Phase (ChESEE-2P). Project funded by the European High Performance Computing Joint Undertaking (EuroHPC JU) and Spain, Italy, Iceland, Germany, Norway, France, Finland and Croatia under [Grant Agreement No 101093038](#).

Subgroups and projects Shared projects Inactive

	Search (3 character minimum)	Q	Name	⌵	⌴
➤	Code audits and tests		1	1	1
	ChESEE-2P: database of POP code audits and tests				
➤	Flagship codes		11	1	1
	ChESEE-2P: catalogue of flagship codes				
➤	Mini-apps		7	1	1
	ChESEE-2P: catalogue of mini-apps				
➤	Pilot Demonstrators (PDs)		0	10	1
	ChESEE-2P: catalogue of Pilot Demonstrators (Workflows)				
➤	Simulation Cases (SCs)		0	16	1
	ChESEE-2P: database of Simulation Cases				
	Documentation		1		1 month ago
	ChESEE-2P: repository documentation and templates				
	gitlab-profile		0		1 month ago

Next talk!

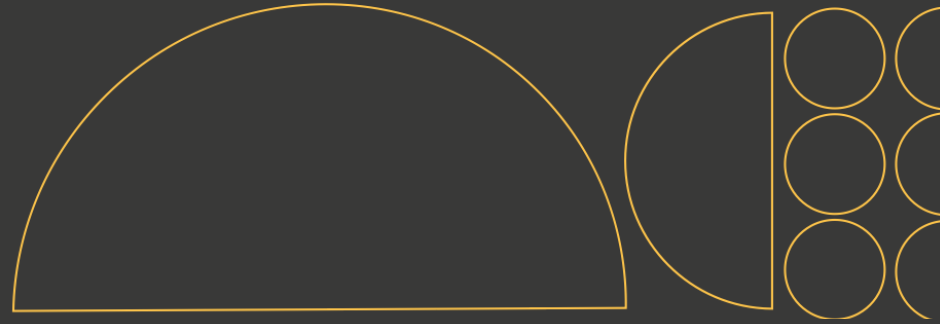


DMP: implementation calendar

ID		Year 1 (2023)	Year 2 (2024)	Year 3 (2025)	Year 4 (2026)
1.1	git-based code repositories and SQAaaS (badges)	✓			
1.2	WF and service repositories			✗	✗
2.2	Code performance datasets		⚠		✗
2.3	Repositories for datasets from PDs, SCs (SDL@CIN)			✗	✗
3.1	Repository for metadata (catalogues)		⚠	✗	✗
3.2	ma-DMP				✗
4.1	Internal reports and documents (Wiki)	✓	✓	✗	✗
4.3	Other DAs (e.g. deliverables)	✓	✓	✗	✗
	FAIR-EVA for datasets (depends on data model)			⚠	
	Final DMP (D1.4)			✗	



Thank you!



@cheese-coe



<http://cheese2.eu>



@cheese-coe



@cheese_coe@techhub.social

