

DATA MANAGEMENT FRAMEWORK FOR EURAD

Yevheniia Kudriashova, SSTC NRS, Ukraine
George-Dan Miron, PSI, Switzerland

EURAD 2 Annual Event , Bologna
Session 2 - Data Management 10 September 2025



Co-funded by the European Union under Grant Agreement n° 101166718

INTRODUCTION

A unified framework for EURAD's data challenges

The Archival & Knowledge challenge



The Goal: to ensure the **long-term value, reproducibility, and shareability** of the final scientific datasets.

This is about the **legacy and impact** of our research.

The Operational & Integration challenge



The Goal: to manage and integrate **live, real-time data streams** from diverse partner technologies with full traceability.

This is about the **efficiency and scalability** of the operations.

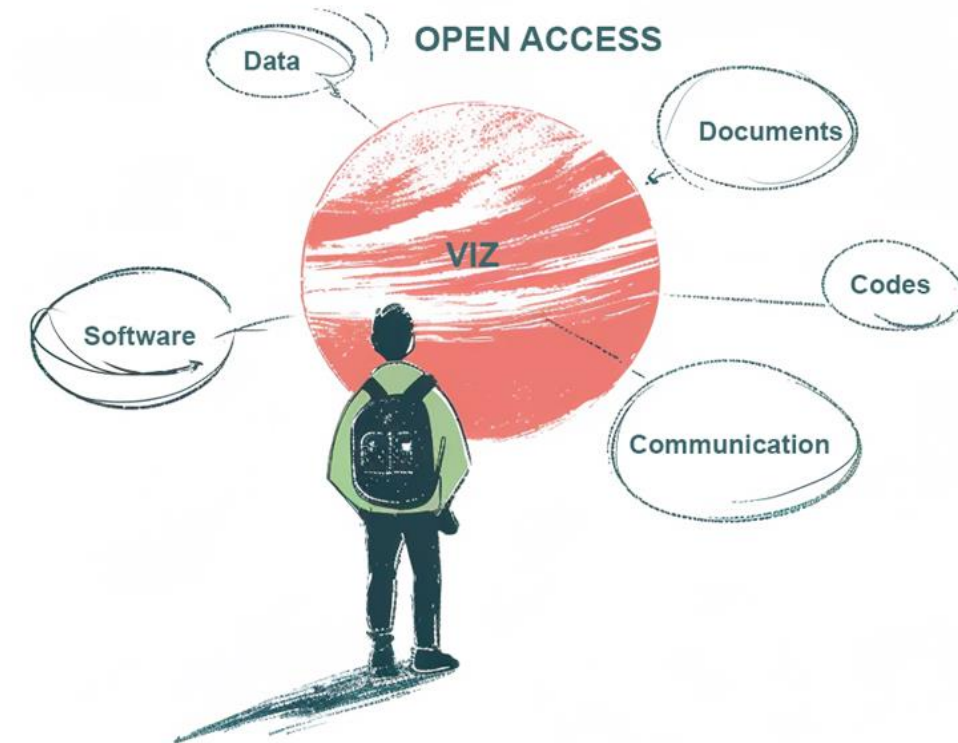
WHAT IS HAPPENING IN EURAD

- **EURAD-2 Data Management Plan (DMP)**

- Based on the "Horizon 2020 DMP" template provided by the European Commission.
- Framework for effective data management across the EURAD-2.
- Provides (general) guidelines within EURAD partnership for data collection, storage, sharing, and preservation while ensuring compliance with FAIR (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable) principles – data handling during and after the project




- **EURAD-2 Knowledge Management Platform (KMP)**

- Ideas on potential KMP, to make knowledge easier to find, share, and use.
- A hub for all relevant knowledge – data – tools – learning A hub with links to already developed resources and tools (use what we have in a smart way)






DMP FAIR DATA: FINDABLE, ACCESSIBLE

Example repositories

Rank	Repository	Strengths	Weaknesses
 1	Zenodo	Free, open, trusted (CERN), code/data/papers, DOIs	Limited collaborative tools, no institutional tier
 2	OSF (Open Science Framework)	Full project workflows, integrations, preprints, open	Interface less polished, not ideal for large datasets
 3	Figshare	Widely used, multimedia friendly, DOIs, institutional tier	Free tier has limited storage; commercial ownership

Open source code repositories

Rank	Repository	Strengths	Weaknesses
 1	GitHub + Zenodo	Collaborative development + DOI archiving; gold standard	Requires setup for DOI via Zenodo; GitHub alone doesn't mint DOIs
 2	GitLab (Self-hosted or GitLab.com)	Powerful CI/CD; open-core; flexible project visibility	UI steeper than GitHub; public trust lower
 3	Codeberg (Forgejo-based)	Fully open-source; EU-hosted; privacy-first	Smaller community; less integrations

DMP FAIR DATA: INTEROPERABLE

- **Established, normative standards as the foundation.**
 - Reuse common conventions for units, date/time formats, and naming (referencing standards from **ISO, IEEE, etc.**).
 - This prevents reinventing the wheel and ensures baseline compatibility.
- **Layer domain-specific vocabularies and schemas for the unique context.**
 - **Metadata** (data about data) is crucial for machines and humans to understand the dataset's context, structure, and format.
 - Metadata should follow a **schema (blueprint)** of how the metadata and data are structured
 - This blueprint should use a **dictionary – ontology of terms and concepts**
 - These should be developed like "code" (common repository, peer review, and regular releases)
- **Leverage existing work.**
 - *Example Schema:* The **RepMet** (Radioactive Waste Repository Metadata Management) Initiative.
 - *Example Ontology:* The **PLEIADES** (nuclear decommissioning ontology).
- **Crucially, these must be defined by the experts producing the data**, with support from IT specialists.
- **This development and maintenance work must be considered in the funding when WPs are being planned.**

DMP FAIR DATA: REUSABLE

Provide the Context (The "What & Why")



- Rich metadata:**

the who, what, when, where, and why of the data.

- "Readme" file:**

a simple, human-readable guide to the dataset.

- Documentation:**

explain the structure, units, and any special values.
(*Without context, data is just a collection of numbers*)

Ensure Transparency (The "How")



- Documented workflow:**

show how raw data was processed and analysed.

- Open and commented code:**

publish the scripts used, so the process can be verified and trusted.

- Full provenance:**

clearly state the data's origin and history, avoid closed source, proprietary codes, use versioning control

Assign Permissions & Responsibility (The "Rules & Who")



- Clear data license:**

explicitly state how others can reuse the data

- Defined ownership:**

a specific team or person who is responsible for the final data quality / data curation.

- Dedicated resources:** planned time and effort in the WP.

THE BADS: WHY GOOD DATA GOES BAD

Problem #1: The Silo Effect

Data is created but remains trapped and inconsistent.

Symptoms We've Seen:

- ✓ Data lives in personal emails, not shared PP platform.
- ✓ Key context and decisions are buried in private meeting notes or local files, not the shared workspace where others can find them.
- ✓ No common format or "referential" exists between experimental and modelling teams.
- ✓ Official repositories are abandoned.

Root Cause:

Workflows default to individual convenience, not project-wide best practice. Individual habits wins over collective best practice. No real data management.

This leads to inconsistent, untraceable data that is impossible to integrate.

Problem #2: The Lost & Found

Knowledge is generated but never becomes public or findable.

Symptoms We've Seen:

- ✓ "Public" documents (milestones, reports) are not actually made public.
- ✓ A final report is published, but the underlying raw data is lost, preventing re-analysis and locking in a single, subjective interpretation.
- ✓ Valuable inputs, like the survey raw data, are collected but never shared back.
- ✓ At the end of a project, the final data is not published externally.

Root Cause:

The "last mile" of sharing is often an afterthought. There is no formal process or assigned owner responsible for the data publication step.

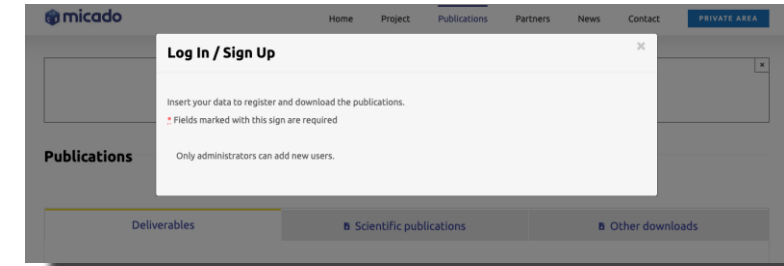
This leads to a massive loss of value and a failure to build a collective knowledge base.

THE BADS: WHY GOOD DATA GOES BAD

May we help you?

You tried to visit a webpage or a website that no longer exists.

OTHER DECOMMISSIONING
RELATED PROJECTS



See website: <https://www.chance.eu>

Safe interim storage and final disposal of radioactive waste (RW) requires effective characterization and quality control of the waste. The CHANCE project therefore aims to address the specific and complex issue of the characterization of conditioned radioactive waste (CRW) by means of non-destructive analytical (NDA) techniques and methodologies. Characterization issues within CHANCE encompass both



This site can't be reached

Check if there is a typo in insiderh2020.eu.

DNS_PROBE_FINISHED_NXDOMAIN

Reload

eurad 2

STEP BY STEP IMPROVEMENTS

How it was (EURAD1)

Planview ProjectPlace		Workspaces	EURAD - DONUT
Overview Conversations Plan Boards Documents Whiteboards			
Filter Add			
Name ↓			
2nd WP meeting Jan 2021			
ACED DONUT Workshop			
Annual Meeting May 2020			
benchmarks questionnaire			
Benchmark-ML-Geochem			
Benchmark: Multiphase Reactive Transport			
Benchmarks THM bentonite			
D4.5			
Description of work			
DELIVERABLE			
END USERS			
EURADWASTE			
FEBRUARY 2022 MEETING			
File before Sub			
Final Annual Event			
Final deliverable D4.7			

How it's going (EURAD2)

Planview ProjectPlace		Workspaces	EURAD-2 - KM
Overview Conversations Plan Boards Documents Whiteboards			
Filter Add			
Name ↓			
01_BACKGROUND_DOCUMENTS			
02_TASKS_SUBTASKS			
TASK_01			
TASK_02			
TASK_03			
TASK_04			
TASK_05			
TASK_06			
03_MEETINGS			
04_DELIVERABLES_MILESTONES			
05_DISSEMINATION			
06_END_USERS_STAKEHOLDERS			
Public documents			

Planview ProjectPlace		Workspaces	EURAD-2 - ICARUS
Overview Conversations Plan Boards Documents Whiteboards Meetings			
Filter Add			
Name ↓			
01_BACKGROUND_DOCUMENTS			
02_TASKS_SUBTASKS			
TASK_01			
TASK_02			
TASK_03			
TASK_04			
TASK_05			
03_MEETINGS			
EURAD-2 Annual events			
TASK_01			
TASK_02			
TASK_03			
TASK_04			
TASK_05			
WPO5_EVENTS			
04_DELIVERABLES_MILESTONES			
D05_1_ Initial SotA			
Milestone5_1_SF_Isotopes_WAC			
05_DISSEMINATION			

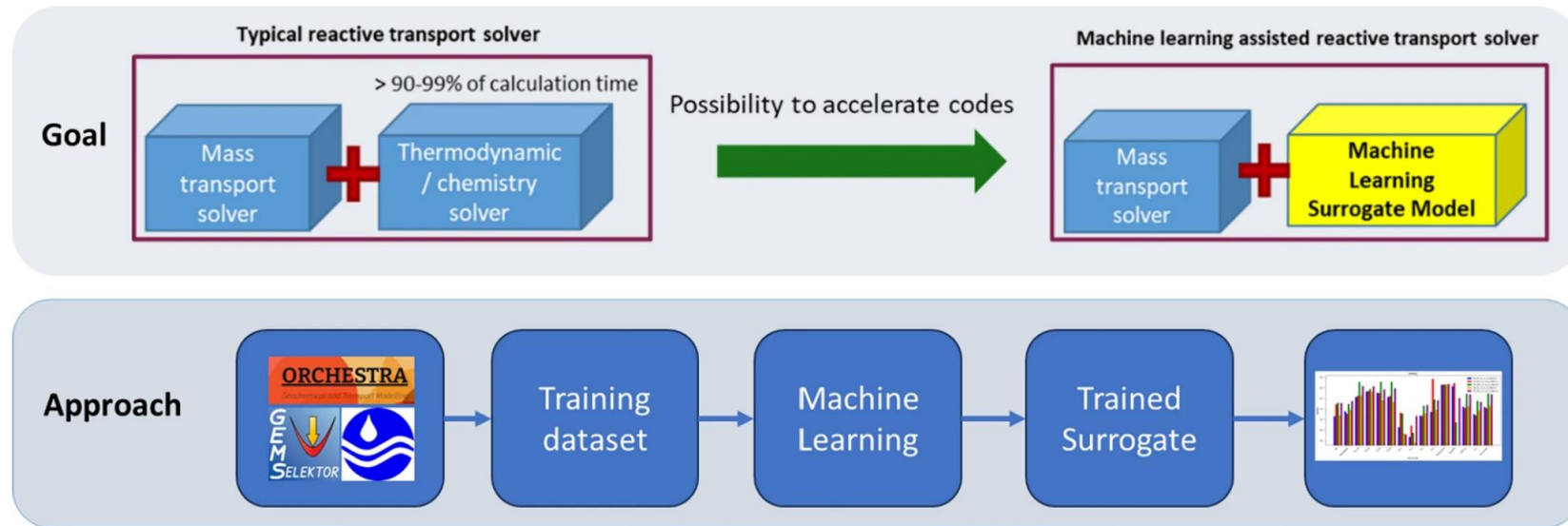
Planview ProjectPlace		Workspaces	EURAD-2 - DITOCO2030
Overview Conversations Plan Boards Documents Whiteboards Meetings			
Filter Add			
Name ↓			
01_BACKGROUND_DOCUMENTS			
02_TASKS_SUBTASKS			
TASK_01_WP_BOARD			
TASK_02_KM			
TASK_03_CURRENT_PRACTICES_DT			
TASK_04_GAP_ANALYSIS			
TASK_05_RECOMMENDATIONS			
TASK_06_CS_INTERACTIONS			
03_MEETINGS			
04_DELIVERABLES_MILESTONES			
05_DISSEMINATION			
06_END_USERS_STAKEHOLDERS			
Interim Progress Report			
Public documents			

THE GOODS (EXAMPLE FROM EURAD 1)

- **DONUT WP** (Development and Improvement Of Numerical methods and Tools for modelling coupled processes)
 - Prasianakis et al. (2025) Geochemistry and machine learning: methods and benchmarking. Environmental Earth Sciences (2025) 84:121 <https://doi.org/10.1007/s12665-024-12066-3> (**Open Access**)
 - A positive example of knowledge and data management
 - No other way to reach the project goals without having a good DMP
 - During the project meetings decisions were made on DMP – storage, structure of data, dictionary
- **Goals: framework for production of high-quality consistent training datasets from different geochemical codes that can be used in different ML techniques → test the methods with appropriate metrics and provide guidance and future perspectives**

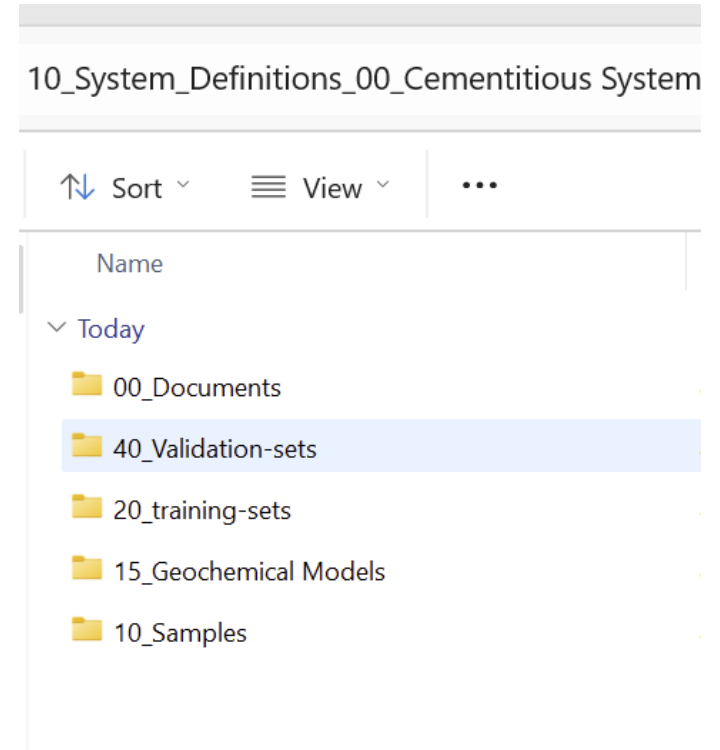
TASKS AND CHALLENGES

- Compare and use the output of **3 geochemical codes** Phreeqc, Orchestra and GEMS → produce training datasets
- Train surrogate models using **several methods/codes**: Neural Networks: Matlab), Tensor Flow, PyTorch; Gaussian Processes: scikit-learn 2.0 / PhreeqcRM, GP and Random Forests; Decision trees (DecTree); PCE;
- Make the workflow **reproducible for future extensions** (other or more complex systems)



IS IT FAIR? FINDABLE, ACCESSIBLE

- **Internal Storage:** During the project, the files were stored on an institutional server (SwitchDrive CH), >1500 files and 10Gb data
- **Shared open access storage:** At the end of the project the results and workflow was described in the publication and all data, scripts and codes are provided as a standalone package on Zenodo (<https://doi.org/10.5281/zenodo.14904784>)
- **Consistent folder and file naming system;** documents contain workflow, data description
- **Narrative description:** of the work and data are provided in the reports on the EURAD –DONUT website and in the published manuscript
- **Suggestion:** update the EURAD project website with links to the publication(s), data repository



IS IT FAIR? INTEROPERABLE

- **A defined format for input/output** - of the geochemical models, training datasets, metric assement and visualization
- **Schema for file naming and data format** described in tables inside a word file. Contains dictionary for the used variables, data ranges, and units.
- **Suggestion:** use of standard schemas, put schemas in data folders

CEMENTITIOUS SYSTEM: 6 LEVELS OF COMPLEXITY

System	Acronym	Oxides	Hydrates	CaO	SiO ₂	CO ₂	Al ₂ O ₃	SO ₃	K ₂ O	H ₂ O
Primitive	P	CaO SiO ₂	Portlandite AmorfSi C-S-H ¹	0.1-1.8	0.2-0.7					0.05-0.15
Primitive+C	Pc	CaO SiO ₂ CO ₂	P + calcite	0.1-1.8	0.2-0.7	0.001-0.96				0.05-0.15
Minimal	M	CaO SiO ₂ Al ₂ O ₃	P + gibbsite katoite ² chabazite straetlingite	0.9-1.4	0.3-0.6		0.03-0.07			0.05-0.15

Systems

Aqueous activity model (in GEMS calculations): Davies 0.3, $\gamma = 1.0$ for neutral species

Primitive cement CaO-SiO₂-H₂O

Components: Ca Si H O

Geochemical model:

Pure minerals: Portlandite, Amor-SI

Solid-Solution:

CSH (ideal): CSHQ_JenD, CSHQ_JenH, CSHQ_TobD, CSHQ_TobH

Input

Variable heading	Description	Unit	Min	Max
SiO ₂	Amount of SiO ₂	mole	0.2	0.7
CaO	Amount of CaO ₂	mole	0.1	1.8
H ₂ O	Mass of water	kg	0.05	0.15
T	Temperature	°C	25	25

Output

Variable heading	Description	Unit
T	Temperature	C
CaO	Amount of CaO	mole
SiO ₂	Amount of SiO ₂	mole
H ₂ O	Mass of water	Kg
pH	pH	
MassWater	Mass of water after reaction	kg
Ca_aq	Amount of Ca in solution	mole
Si_aq	Amount of Si in solution	mole
O_aq	Amount of O in solution	mole



Generated sample input to geochemical codes

```
10_PC_02_LHS_500_54854_01_s1 × +
File Edit View

# {'SiO2': [0.2, 0.7], 'CaO': [0.1, 1.8], 'H2O': [0.05, 0.15]}
# Random seed 54854
# Sampling method LHS
SiO2  CaO  H2O
3.936958e-01 1.570795e+00 1.307169e-01
4.770984e-01 6.715356e-01 1.318965e-01
2.693271e-01 3.750896e-01 1.049152e-01
4.740024e-01 6.920741e-01 5.157852e-02
2.255677e-01 9.621644e-01 9.101669e-02
6.520770e-01 1.193130e+00 5.003403e-02
5.150537e-01 1.425687e+00 1.150510e-01
6.676499e-01 5.451852e-01 1.400306e-01
3.905742e-01 1.746112e+00 8.150850e-02
2.978511e-01 2.852676e-01 6.245972e-02
2.782536e-01 6.346480e-01 7.743922e-02
```

Generated training dataset output from geochemical codes

```
10_PC_02_LHS_500_54854_01_s1 × +
File Edit View

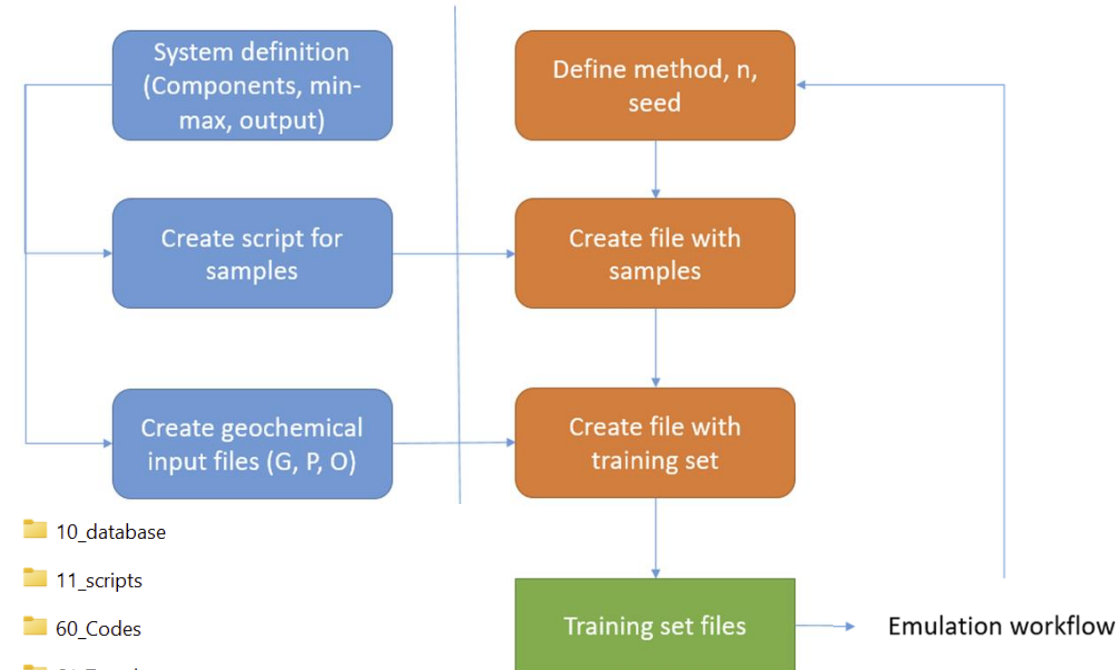
T, CaO, SiO2, H2O, pH, MassWater, Ca_aq, Si_aq, O_aq, H_aq, Ca_s, Si_s, O_s, H_s, Portlandite, AmorfSi, mCSHQ, Ca_ss, Si_ss, H2O_ss, V_s, Gel_water
2.500000e+01,1.570795e+00,3.936958e-01,1.307170e-01,1.247261e+01,9.354582e-02,1.900788e-03,2.899352e-06,3.806572e-03,3.795970e-03,1.568894e+00,3.936929e-01,4.41769
2.500000e+01,6.715356e-01,4.770984e-01,1.318966e-01,1.221850e+01,1.156570e-03,8.028254e-06,2.323804e-03,2.298356e-03,6.703790e-01,4.770904e-01,2.84851
2.500000e+01,3.750896e-01,2.693271e-01,1.049153e-01,1.219657e+01,9.255699e-02,9.219865e-04,7.197989e-06,1.853786e-03,1.830808e-03,3.741676e-01,2.693199e-01,1.59788
2.500000e+01,6.920741e-01,4.740024e-01,5.157858e-02,1.228912e+01,2.902366e-02,3.660626e-04,1.718686e-06,7.359507e-04,7.289015e-04,6.917080e-01,4.740007e-01,2.89133
2.500000e+01,9.621644e-01,2.255677e-01,9.101679e-02,1.247261e+01,6.859942e-02,1.393899e-03,2.126388e-06,2.791995e-03,2.783686e-03,9.607705e-01,2.255656e-01,2.65486
2.500000e+01,1.193130e+00,6.520770e-01,5.003409e-02,1.247261e+01,1.384319e-02,2.812841e-04,4.290542e-07,5.650115e-04,5.617385e-04,1.192849e+00,6.520766e-01,4.50561
2.500000e+01,1.425687e+00,5.150537e-01,1.150511e-01,1.247261e+01,7.775892e-02,1.580010e-03,2.410088e-06,3.164509e-03,3.155358e-03,1.424107e+00,5.150513e-01,4.52266
2.500000e+01,5.451852e-01,6.676499e-01,1.400308e-01,1.124618e+01,1.189537e-01,1.542629e-04,6.872505e-05,4.333888e-04,2.793516e-04,5.450309e-01,6.675812e-01,3.05006
2.500000e+01,1.746112e+00,3.905742e-01,8.150859e-02,1.247261e+01,4.124926e-02,8.381575e-04,1.278467e-06,1.679636e-03,1.673842e-03,1.745274e+00,3.905729e-01,4.76031
2.500000e+01,2.852676e-01,2.978511e-01,6.245979e-02,1.153502e+01,5.194449e-02,1.113697e-04,1.747901e-05,2.519278e-04,2.072001e-04,2.851562e-01,2.978336e-01,1.46446
2.500000e+01,6.346480e-01,2.782536e-01,7.743931e-02,1.247261e+01,5.973481e-02,1.213771e-03,1.851423e-06,2.431456e-03,2.423962e-03,6.334342e-01,2.782517e-01,2.17147
2.500000e+01,1.108103e+00,2.676318e-01,1.275921e-01,1.247261e+01,1.015977e-01,2.064419e-03,3.148906e-06,4.134092e-03,4.122749e-03,1.106039e+00,2.676287e-01,3.08214
2.500000e+01,1.129375e+00,5.715560e-01,9.379731e-02,1.247261e+01,6.056980e-02,1.230738e-03,1.877285e-06,2.465415e-03,2.457845e-03,1.128144e+00,5.715541e-01,4.11442
```

IS IT FAIR? REUSABLE

- Description of the **workflow, defined metrics**
- **Scripts** to generate sampling datasets (fixed seed value)
- **Geochemical codes** with chemical system definition and thermodynamic database.
- Additional **information** on geochemical models, constraints found in readme files.
- Each partner was responsible for the data curation, following the agreed upon formats.
- Reused in **EURAD-2 HERMES**

Workflow

Creation of the training sets



10_database
11_scripts
60_Codes
61_Template
00_Generate-Training.py
10_PC_File-Sample.dat
10_PC_File-Sample.in
10_PC_File-Sample.out
10_PC_Single-Sample.in
10_PC_Single-Sample.out
Readme.docx
run.bat



eurad²

KEY MESSAGES

Embrace Radical Transparency and Sharing: data is not a private byproduct; it is a core, shared deliverable. This requires a default-to-open mindset, ensuring our work is visible, verifiable, and contributes to the collective knowledge base.

Future-Proof the Work: technology evolves, and the data practices must too. We must actively keep up with advancements in data formats, schemas, databases, and open-source tools to avoid creating obsolete data silos.

FAIR principles are the goal, but they don't happen automatically. Making them a reality requires a practical and sustained commitment.

Allocate resources: Data curation must be a budgeted line item in every WP, with dedicated time assigned. It is a primary task, not an afterthought to report writing.

Empower the experts (Bottom-Up): the scientists generating the data must define the specific formats and metadata they need, using the EURAD-2 DMP as their guide. This ensures the system is practical and gets active feedback.

Organize collaboration, information flow: the DMP should be up to date based on discussions between WP, between data producers, users, it experts, different initiatives (NEA, other fields, ...)

OUR PATH FORWARD: FROM PRINCIPLES TO POLICY

1. Mandate work package DMPs

- Every future WP proposal submitted to EURAD **should** include its own specific Data Management Plan.
- This plan will start from the general EURAD-2 DMP template but should be tailored to the project's specific data and objectives (where, how, real collective data management during project lifetime).
- Fulfillment of this requirement will be a factor in the assessment of the proposal.

2. Enforce DMP accountability

- The final report for every Work Package must contain a dedicated section reporting on the fulfilment of its DMP.
- This section will provide a narrative of the data management activities and, most importantly, direct links to the public repositories where the final datasets are stored.

3. Build a central knowledge hub and toolkit

- To develop and maintain a central catalogue of approved tools, methods, and best practices for the DMP and KMP (from general to useful for waste disposal).
- This will include templates for data schemas, links to recommended software, and clear "how-to" guides. This reduces duplicated effort and provides a clear, supported starting point for all projects.