

STRATA-AI

"STANDARDS, TRAINING & REASONING FOR AI IN SUBSURFACE CHARACTERISATION"

WP Lead Team

N.I. Prasianakis (PSI), D. Eastwell (NWS), R. Szóke (HUN-REN EK), R. Lo Frano (UniPi), J. Pwavodi (BRGM), S. Grigull (SKB), M. De Lucia (GFZ), A. Laikari (VTT), O. Severýn (TUL), J. Cochard (Andra)



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

PROBLEM STATEMENT

- **GDF site characterisation produces exceptionally large, heterogeneous and interdependent datasets**, spanning borehole logs, seismic measurements, core imagery, laboratory analyses, hydraulic tests and descriptive text, with different resolutions, spatial supports, formats and levels of uncertainty.
- **Conventional integration remains resource-intensive and difficult to scale**, relying on discipline-specific workflows, expert interpretation and long-form reporting to derive coherent geological models from disparate data modalities.
- **AI and machine learning create a new technical opportunity to analyse these data jointly**, learning cross-modal relationships that support data fusion, inconsistency detection, missing-data prediction, spatial interpolation and uncertainty-aware geological interpretation.
- **GDF datasets provide a uniquely valuable environment in which to develop and test these methods**, because their quality assurance, controlled spatial referencing and focused sampling of candidate host-rock systems create a stronger basis for traceable, scientifically defensible model development.
- **A coordinated international project is required to convert this opportunity into reusable capability**, combining representative datasets, specialist expertise and shared resources to establish standards, validate methods against independent data, and demonstrate measurable improvements in the efficiency, consistency and evidential strength of site characterisation.

VISION

To establish a scientifically defensible, AI-ready framework that integrates multimodal borehole data, reveals cross-domain subsurface relationships and enables traceable, uncertainty-aware prediction in support of GDF site characterisation and geological modelling.

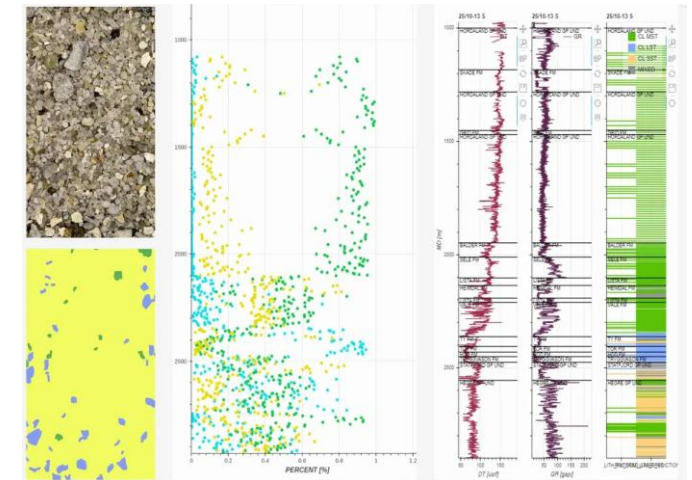
- **Establish a validated AI-ready framework for integrating multimodal subsurface datasets**, enabling quantitative analysis across geological, geophysical, geochemical and observational data domains.
- **Transform disparate site-characterisation data into a unified knowledge space**, where relationships between measurements, imagery, text and interpretations can be systematically identified, interrogated and validated.
- **Exploit machine learning to reveal previously hidden correlations within GDF host-rock datasets**, improving understanding of subsurface properties, heterogeneity and controlling geological processes.
- **Enable predictive modelling of incomplete subsurface datasets**, supporting data consistency assessment, uncertainty reduction and enhanced evidence generation for safety cases and engineering design.
- **Demonstrate how AI can maximise the value of globally unique GDF characterisation datasets**, creating a scalable digital workflow that accelerates geological interpretation and supports evidence-based decision-making for disposal facilities.

WP LEADERSHIP TEAM

Task	Title	Task Lead	Task Co-Lead
1	WP Management/ Coordination	Nikolaos Prasianakis PSI, RE, Switzerland	David Eastwell, NWS, WMO, UK
2	Knowledge Management	Réka Szőke, HUN-REN EK, RE, Hungary	Rosa Lo Frano, UniPi, RE, Italy
3	Develop a comprehensive set of best practice, processes, standards and structure for borehole data including:	Joshua Pwavodi, BRGM, RE, France	Susanne Grigull, SKB, WMO, Sweden
4	Identification and assessment of AI/ML techniques for multimodal subsurface data analysis and interpretation	Marco De Lucia, GFZ, RE, Germany	Arto Laikari, VTT, TSO, Finland
5	Workflow validation, demonstration and proof-of-concept	Otto Severyn TUL, RE, Czech Republic	Jean Cochard, ANDRA, WMO, France

DATA SOURCES

- **STRATA-AI Project will focus on high-quality subsurface datasets sourced from European WMO partners**
- **Data being assembled in collaboration with ANDRA, SKB and Mont Terri**
- **STRATA-AI will focus on borehole derived datasets that have a spatial expression predominantly:**
 - 1) 1D Log data, lithological data (High Priority)
 - 2) Depth correlated point measurements (mechanics, hydraulics also from cores)
 - 3) Core photographs/ scans / analysis
 - 4) Mineralogy measurements
 - 5) Drilling reports (Drillers/ Interpretive)
 - 6) Borehole trajectory
 - 7) Seismic (VSP)



From:

[Interpreted cuttings samples from all released wells on the NCS | Data packages](#)

TASK-2

The main goal of Task 2 is to **capture, consolidate & transfer knowledge** relevant for the SRA topic, to **contribute to knowledge transfer** to the **EURAD-2 community & beyond**, through the EURAD-2 KM programme.

2.1: Knowledge capture Development of documentation materials (SotA, gap analysis)

Will work **in collaboration with the WP-KM**.

Will **capture, structure & consolidate** the knowledge generated **within STRATA-AI & relevant knowledge from previous EURAD, national & international activities** into forms that are reusable by the wider EURAD community.

Includes the

- preparation of the **State-of-the-Art reports**,
- a **structured gap analysis** on the role, opportunities, limitations and implementation needs of AI/ML methods for multimodal subsurface datasets.

Contributors: [HUN-REN EK] [A21] [LEI] [PSI] [RISE] [TUS] [UniPi] [VTT] [SSM] [NRG PALLAS]

2.2: Knowledge transfer & Training materials

Will **translate the knowledge captured** in Subtask 2.1 & **the technical outputs of tasks 3–5 into practical transfer formats** for the EURAD-2 community & beyond.

It will **prepare & deliver** targeted **training modules, webinars, face-to-face & online training sessions, workshop material, guidance notes & reusable teaching resources** on AI-ready subsurface data structures, multimodal data integration workflows, explainability, uncertainty, and responsible use and limitations of AI/ML methods in geological and borehole data evaluation for site characterisation.

Contributors: [HUN REN EK] [FZJ] [A21] [UniPi] [GFZ] [PSI] [BRGM] [TUL] [ANDRA] [RISE] [COVRA] [VTT]

Lead: HUN-REN EK (R. Szőke) **Co-lead:** UniPi (Rosa Lo Frano) **Effort:** x PM **Duration:** M25 – M60

Deliverables: SOTA report (M30) Updated SOTA report (M56)

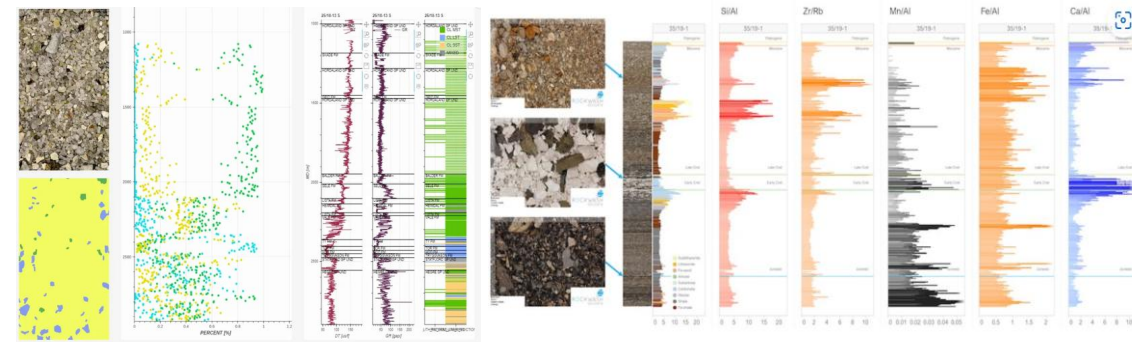
"Immediate" Actions

#	Decision/action	Responsible	Proposed timing
1	Agree SOTA chapter outline / gap analysis focus (based on Template-3)	WPL, TLs	October
2	Nominate chapter leaders, DOERS & contributors	TLs <i>(nomination, volunteering)</i>	October
3	Confirm PM budget & distribution principles	T2 with EURAD-2 KM	October
5	Check ProjectPlace access for all WP contributing partners	WPL / T2	October
6a	Communicate T2 budget & AI-use principles to all contributors	WPL/T2	early November
6b	Appoint reviewers	TLs/WPL	end of November
7	Initiate Subtask 2.2 planning , assign leads/contributors	T2 with TLs	end of November
8	Send orientation communication to all T2 contributing partners	T2	<i>agenda point for the kick-off at PSI</i>

TASK -3

WP-3 · TECHNICAL · 1/3 · SUBTASK 3.1

REQUIREMENTS, SCREENING & AGGREGATION



1

Central
repository

2

Ideal dataset

3

Raw data from
owners

4

Consult WMOs

5

Screen

6

Stage

7

Gap analysis

DATA PRIORITY

1 Logs + lithology · High

2 Mechanics & hydraulics

3 Core imagery & analysis

4 Mineralogy

5 Drilling reports

6 Trajectory

7 VSP

8 Seismic

ALSO IN SCOPE

Core descriptions

Fractures · FMI

Permeability

Monitoring

Chemistry

Thermal

Depth & formation

Reports

REQUIREMENTS DEFINED

Metadata

Quality

Confidentiality

AI/ML readiness

TASK-3

WP-3 · TECHNICAL · 2/3 · SUBTASK 3.2

ARCHETYPE DATABASE: DESIGN, POPULATE, SERVE

STANDARDS REVIEW · NWS, BRGM, RISE, VTT

Open-source data structures & models; non-proprietary standards from hydrocarbon, geothermal, CCUS, hydrogen

ECOFRAME ALIGNMENT → EURAD-2

Data lifecycle · interoperability · traceability · long-term usability

PROVIDE

- ANDRA · SKB · GFZ · SÚRAO · NAGRA (+ others)
- Multiple datasets per organisation
- Most suitable inputs → archetype



CURATE

- Transform · harmonise · standardise
- Agreed schemas, formats, vocabularies
- Complete modality inventory

LEI, TUL, BRGM, SCK CEN, NRG Pallas, RISE, Paris Mines, PSI



STORE

- Multimodal schema
- Consistent relationships
- Spatial · temporal · hierarchical
- zarr-fuse storage (TUL)

TUL, VTT, GFZ, CIEMAT, Paris Mines, NRG Pallas, GRS



SERVE

- JSON · XML · HDF5 · Zarr · Xarray
- Structured + unstructured
- AI/ML training, analysis, visualisation
- Blind-test subset → Task 5

DMP Org clouds (WMO · TSO · RE)



EURAD-2 / EU Dataspaces



Zenodo (open)



Secure pipelines: HUN REN EK, VTT

TASK-3

WP-3 · TECHNICAL · 3/3 · SUBTASK 3.3

VALIDATE, THEN DEVELOP THE REFERENCE TOOL

EARLY-STAGE TECHNICAL VALIDATION

Layer	What is tested	Who
Ingest · store · retrieve	Consistent access; no loss of structure, meaning or metadata	Validation team
Analytical	Data convolution, integration, cross-domain queries: geological · geophysical · geotechnical	Validation team
Statistical	Validity, consistency, plausibility: completeness, distributional integrity, logical relationships	Validation team
Traceability & QC	Traceability, confidentiality-aware validation, quality control	Team

Validation team: BRGM, TUL + GFZ, PSI, TUS, LEI, NRG Pallas, RISE, SCK CEN

DATABASE DEVELOPMENT

Central traceable reference for data & final models

- Public / restricted data layers
- Non-published data: owner + IP-owner permission
- Provenance handling
- Machine-readable interfaces
- Long-term usability

TUL + VTT, BRGM, GRS, CIEMAT, NRG Pallas

 **Test case:** initial archetype dataset → **Evidence:** fit for downstream analysis, modelling & decision support

TASK-4 AI/ML TECHNIQUES FOR MULTIMODAL SUBSURFACE DATA ANALYSIS AND INTERPRETATION

Objective: select, implement and evaluate AI/ML methods extracting **traceable** geological, geochemical, geotechnical and structural insights from the Task 3 reference borehole datasets for site characterisation.

4.1 Methods & use cases

- Shortlist from SOTA (2.1): multimodal transformers, CNNs, geostatistics, graph/generative models
 - Use cases driven by geoscience and TSO/WMO needs: processing, fusion, uncertainty
 - Evaluation metrics for applicability, quality, trustworthiness
- **Documented shortlist (M4, M40)**

4.2 Workflows & training

- Harmonisation pipelines; multi-borehole spatial integration
- Traceable training with ablation and sensitivity tests
- Missing-data prediction, inconsistency detection, cross-modal correlation
- Physics-informed ML, LLM agentic workflows; Git / Zenodo sharing

4.3 Validation, XAI & UQ

- Explainability (SHAP, saliency) and uncertainty quantification
 - Extraction of useful information: correlations, anomalies
 - Explicit limits and domain of validity of each model
- **Evidence pack; open-source model release**

Lead: GFZ (M. De Lucia) **Co-lead:** VTT (A. Laikari) **Effort:** 140 PM **Duration:** M28 – M60

Inputs: Task 2 (SOTA), Task 3 (reference datasets) **Outputs to:** Task 5 (PoC), WP GEOSCALE, HERMES

TASK-5 WORKFLOW VALIDATION, DEMONSTRATION AND PROOF-OF-CONCEPT

Aim: Test the selected AI/ML workflows end-to-end on representative and blind datasets, and show their added value for site characterisation and modelling.

Subtasks:

- **5.1 Independent validation:** Apply trained models to data not used in training, such as withheld boreholes, removed data and synthetic sets. Predict missing data and modalities, and assess robustness, generalisability and limitations.
- **5.2 Sensitivity & explainability:** Identify which data features and assumptions drive the outputs. Quantify uncertainty and how it propagates. Communicate results clearly to domain experts.
- **5.3 Proof-of-concept demonstration:** Evaluate ML outputs as model-ready inputs for flow and transport modelling, covering rock types, mineralogy, permeability, diffusivity, pore water and temperature. Benchmark them against traditional methods.

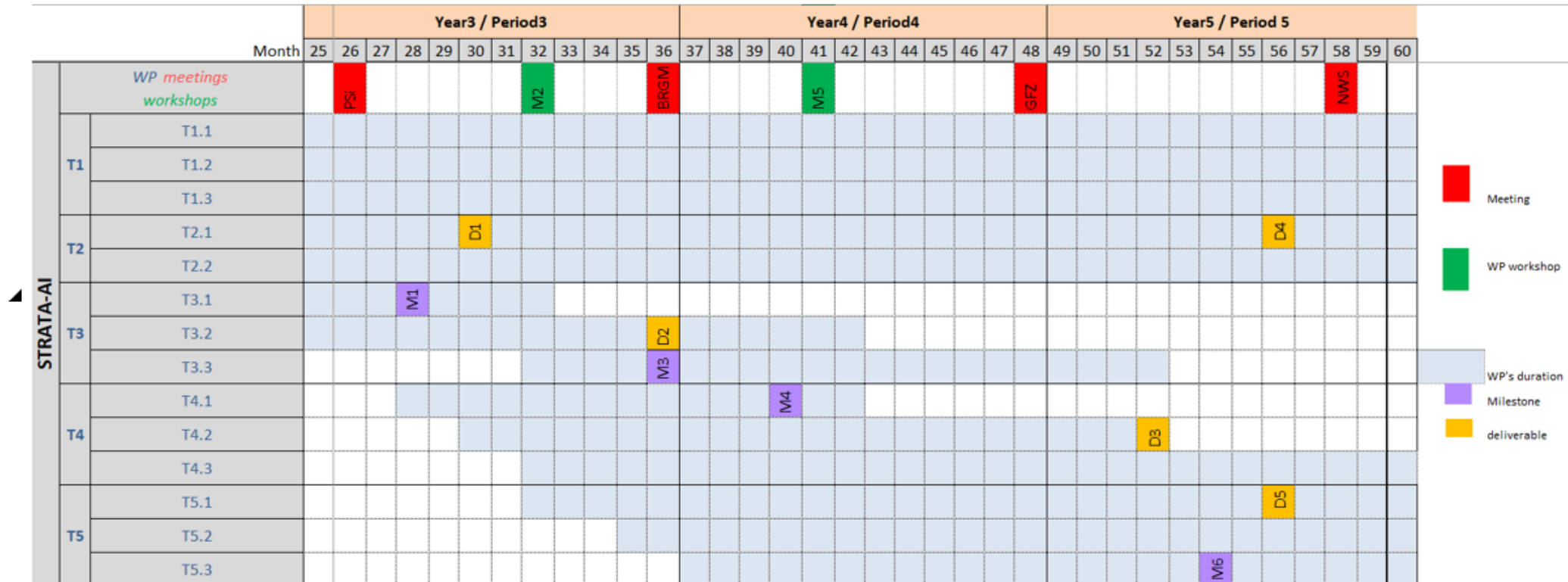
Collaboration: Work with WP-GEOSCALE and WP-HERMES. Implementers and regulators (ANDRA, SKB, SÚRAO, NAGRA, etc.) give end-user feedback.

Outputs:

Validated and benchmarked models, plus reusable, documented workflows and code for the EURAD community.

TIMELINE

Gantt chart of WP STRATA-AI





SCHEDULED EVENTS

- Work has started according to schedule.
- Minimal viable borehole dataset is currently prepared (ANDRA, SKB, ...). Details will be discussed during the online Task 3 Kick-off meeting on the 22nd October. On the same day the Task-2 will also convene online.
- ProjectPlace will be used for data-sharing in the first phase.
- STRATA-AI In-Person Kick-off Meeting will be held at PSI on 19-20 November 2026

DELIVERABLES

No.	Deliverable	Summary	Lead	Type	Dissemination	Delivery
1	Initial State of the Art: AI/ML for Multimodal Borehole Data Integration	Review of data structures and AI/ML methods for integrating multimodal borehole data.	HUN-REN EK	Report	Public	Month 30
2	Reference Borehole Datasets	Reference repository, data framework and workflows for structuring, standardising and integrating borehole datasets for future R&D.	BRGM	Dataset	Public; sensitive where required	Month 36
3	AI/ML Models, Pipelines and Proof-of-Concept Workflow Assessment	Pre-trained models, validated pipelines and reusable code supporting geological modelling and site characterisation, including testing recommendations and limitations.	GFZ	Report / Other	Public	Month 52
4	Final State of the Art	Final synthesis of STRATA-AI progress, including best practice for AI-ready data, metadata, traceability, uncertainty reporting and workflow transferability.	HUN-REN EK	Report	Public	Month 56
5	Outcomes and Impacts Report for Member States and End Users	Synthesis of outcomes, added value, limitations, domains of validity and recommendations for future site-characterisation and modelling applications.	PSI	Report	Public	Month 56