DIAMOND: materials database infrastructure

I. Setoain[1], V. Bergeaud[1], C. Herrera Contreras[2], T. Deutsch[2], A. Amrani[1], J.P. Poli[3]

[1]Université Paris-Saclay, CEA, LIST, F-91190, Palaiseau, France
[2]Université Grenoble Alpes, CEA, IRIG-MEM, F-38000, Grenoble, France
[3]CEA, LIST, 91191 Gif-sur-Yvette cedex, France

PROGRAMME DE RECHERCHE MATÉRIAUX ÉMERGENTS

Objectives

- Establish a **national materials database** infrastructure to integrate experimental and simulation data, relying on the **TGCC-Cloud** facility.
- Initial data will include numerical simulations and characterization data (linked to 2FAST project).
- Expansion to include MOF datasets and machine learning training data (DIAMOND WP3).

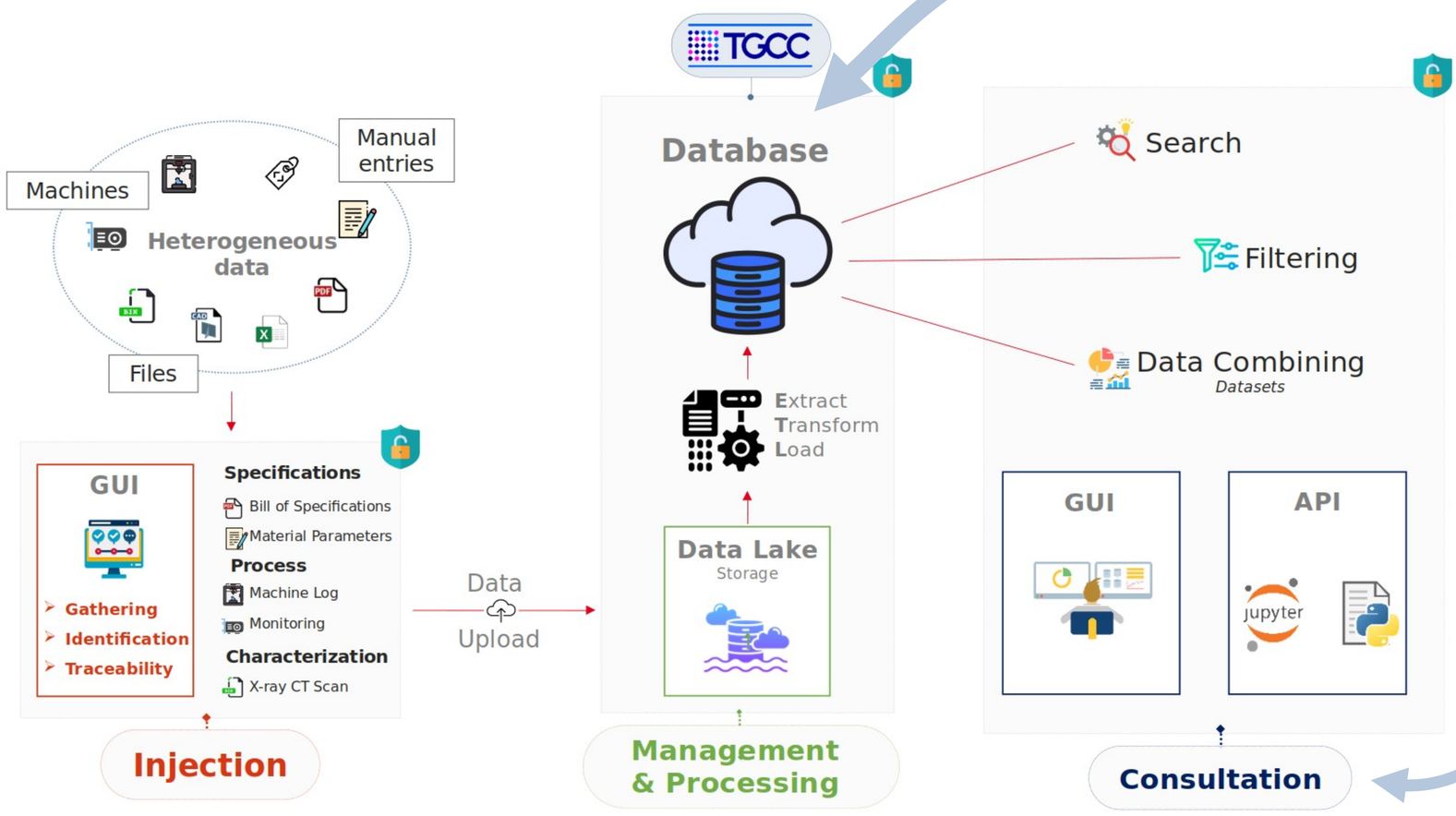
Strategy

DIAMOND database should address:

Data very different sizes, variety and ontologies	Flexible approach for heterogeneous data	Different levels of storage: local and distant

1. Deploy **storage system** + **API** for datasets consultation and extraction.
2. First **database** for standardized (ontologized) **numerical simulations** (BigDFT code).
3. **AI tools** for non-experts to analyse experimental and simulation data

Database architecture



Ontologies

To **define concepts**, their **relationships** and their **properties**, allowing different systems and datasets to communicate and interact seamlessly.

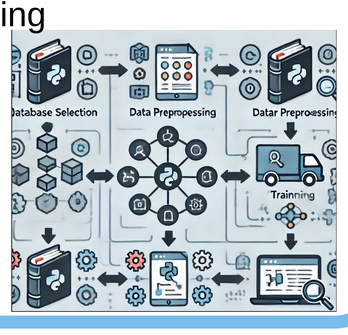
Two main ontologies are used for our numerical simulation database :

EMMO : Foundational ontology for the formal description of concepts in physics, chemistry and materials science

OSMO : Ontology for Simulation, Modelling and Optimization

IA data analysis

Jupyter notebook solution for:
1. Database selection
2. Data preprocessing (clean, etc)
3. Feature engineering
4. Training
5. Validation



Conclusions

This project establishes a **robust**, **flexible** materials **database infrastructure** hosted on the TGCC, uniting experimental and simulation data for seamless integration.
It provides an architecture for :

- Data injection
- Gestion and treatment
- Data Consultation

Key features:

- **Authenticated Access:** Ensures secure, multi-level data storage, both local and remote.
- **Structured API:** Allows data search, filtering, and retrieval, supporting efficient dataset consultation and extraction.
- **Ontologies** at database creation: Standardizes data to enhance interoperability and maintain consistency across datasets.
- **Jupyter Notebook** Integration: Provides flexible, user-friendly tools for AI-driven data analysis.

Acknowledgements

Hosting infrastructure: TGCC-Cloud (CEA Bruyères-le-Chatel)

This work was supported by a grant from the French government managed by the National Research Agency under the France 2030 program with reference **ANR-22-PEXD-0015**.