# Data Lakehouse: Building the Foundation for AI/ML Ready Scientific Data

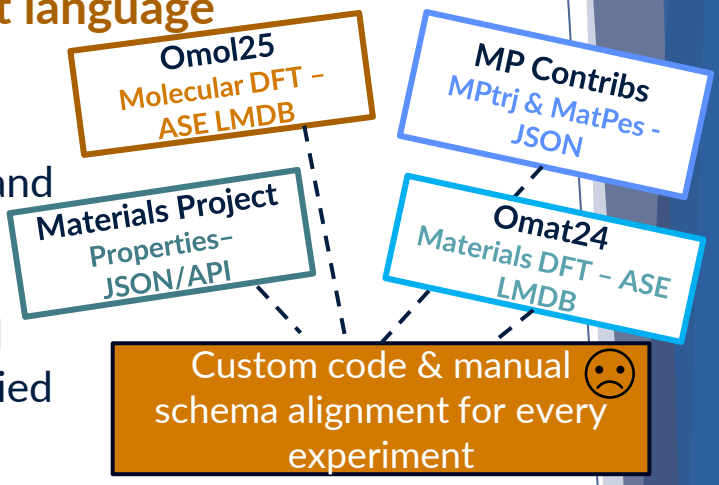
https://www.psdi.ac.uk/

A. Belozerov, V. Bunakov, S. Meacham, A. Pawula Hewage, T. L. Underwood

*Science and Technology Facilities Council (STFC), Rutherford Appleton Laboratory, Harwell Campus, Didcot, OX11 0QX*

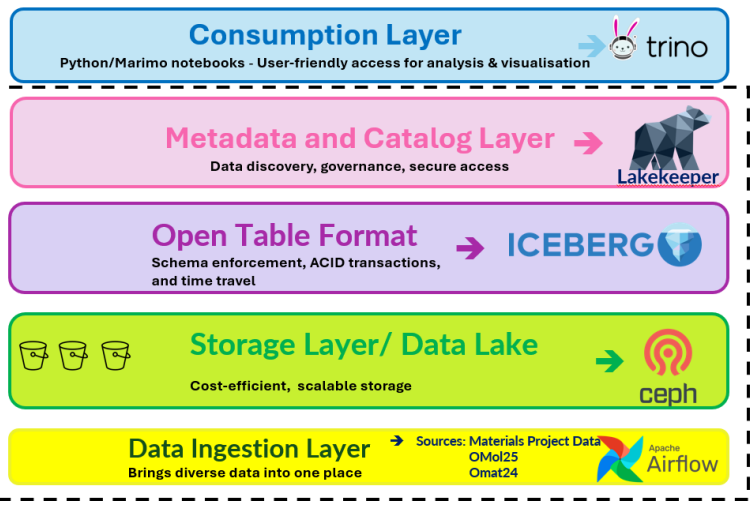
## Every dataset speaks a different language

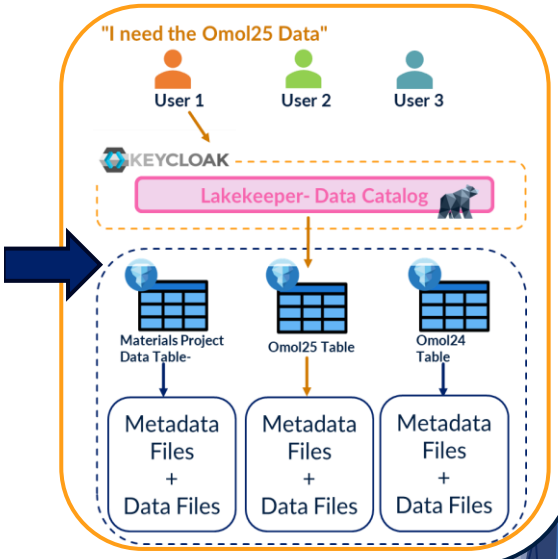
Researchers increasingly combine diverse datasets from multiple repositories to train AI/ML models and perform cross-dataset analyses. This is demanding: datasets must be downloaded, parsed, cleaned, linked and aggregated, with each one queried in its own syntax.




**PSDI Data Lakehouse was built to eliminate this burden - making scientific data interoperable and AI/ML ready**

## PSDI Data Lakehouse Architecture




# "Query first, download never"

**PSDI data Lakehouse offers programmatic access (a machine interface) to datasets that can be queried in the Cloud without the need to download them**

### Why does it matter for researchers?

- ✓ No custom parsers or per-format libraries
- ✓ One SQL & Python interface for all datasets
- ✓ Fully open source

**Getting started – Scan here**
https://github.com/PSDI-UK/lakehouse-tutorials

**More resources - Scan to explore**

PSDI acknowledges the funding support by the EPSRC grants EP/X032701/1, EP/X032663/1 and EP/W032252/1