Getting Your Hands on Lots of Data, Fast

Alex Squires

UNIVERSITY OF BIRMINGHAM

Where is all this data and how can I get my hands on it?

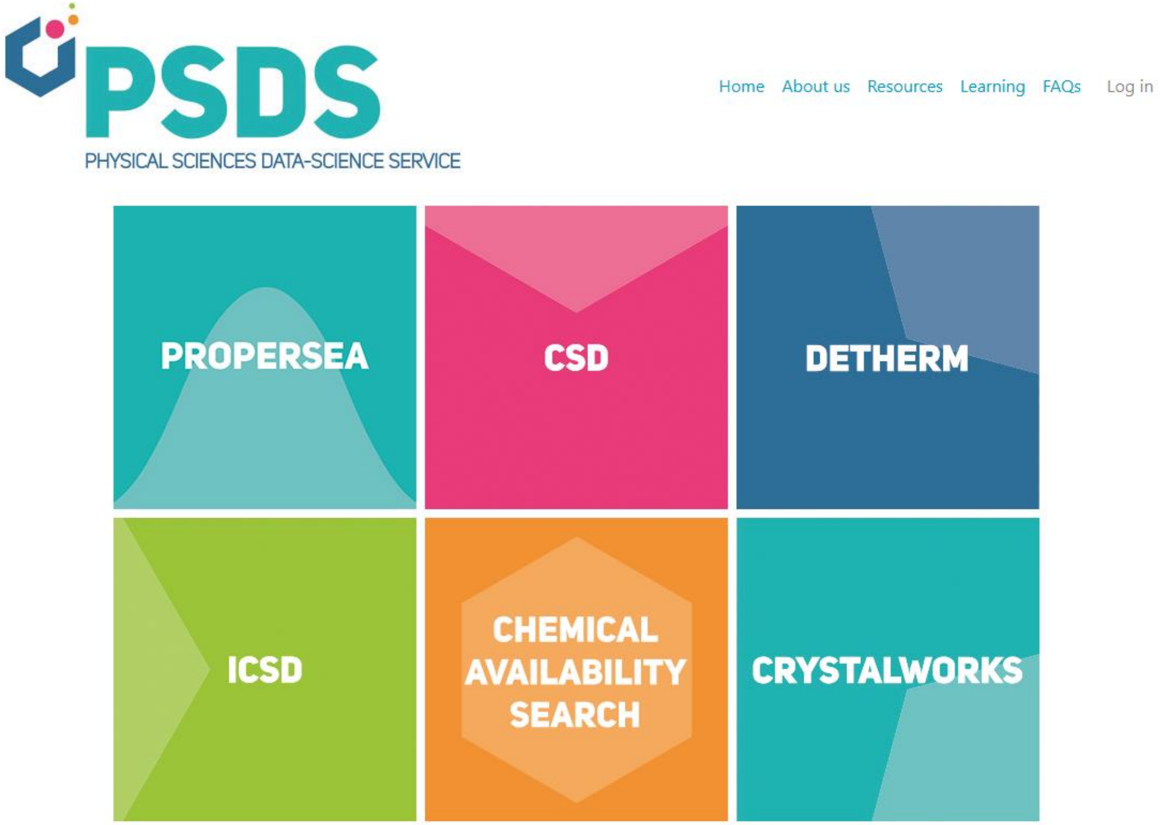
- Experimental Chemical Databases (CSD, ICSD etc.)
- Journals that focus on the publication of data sets (Nature’s Scientific Data)
- Data repositories (Nomad, Zenodo etc.)
- Computational Databases (Materials Project, AFLOW, OQMD etc)

Experimental Databases

In the UK, we can access a lot of these via the PSDS, including the ICSD and the CSD, which give structural information.

These are quite comprehensive and well-curated but often interacting with them programmatically requires proprietary access.

Demo on request!



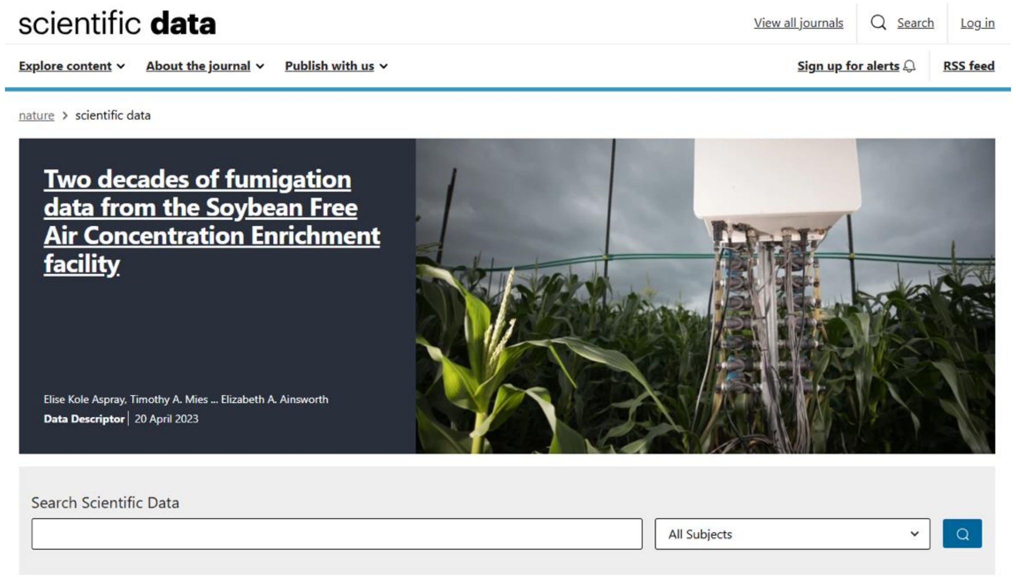
UNIVERSITY OF BIRMINGHAM

Publications that focus on data itself

These are relatively well-curated (as well-curated as most journals!), and the fact they have an associated publication is a useful form of documentation, but they can be very specific

Examples:

- A Quantum-Chemical Bonding Database for Solid-State Materials
- A band-gap database for semiconducting inorganic materials calculated with hybrid functional



Two decades of fumigation data from the Soybean Free Air Concentration Enrichment facility

Elise Kole Aspray, Timothy A. Mies … Elizabeth A. Ainsworth

Data Descriptor | 20 April 2023

UNIVERSITY OF BIRMINGHAM

Data repositories

Zenodo and NOMAD are two great examples of data repositories, these are uploaded by users and so quality will be variable and they will be highly specific, but NOMAD in particular is very exciting because you can upload your *raw* data, and it will generate rich metadata for you. There are great places to check to see if someone else has already done the hard work so you don’t have to!

Theoretical databases

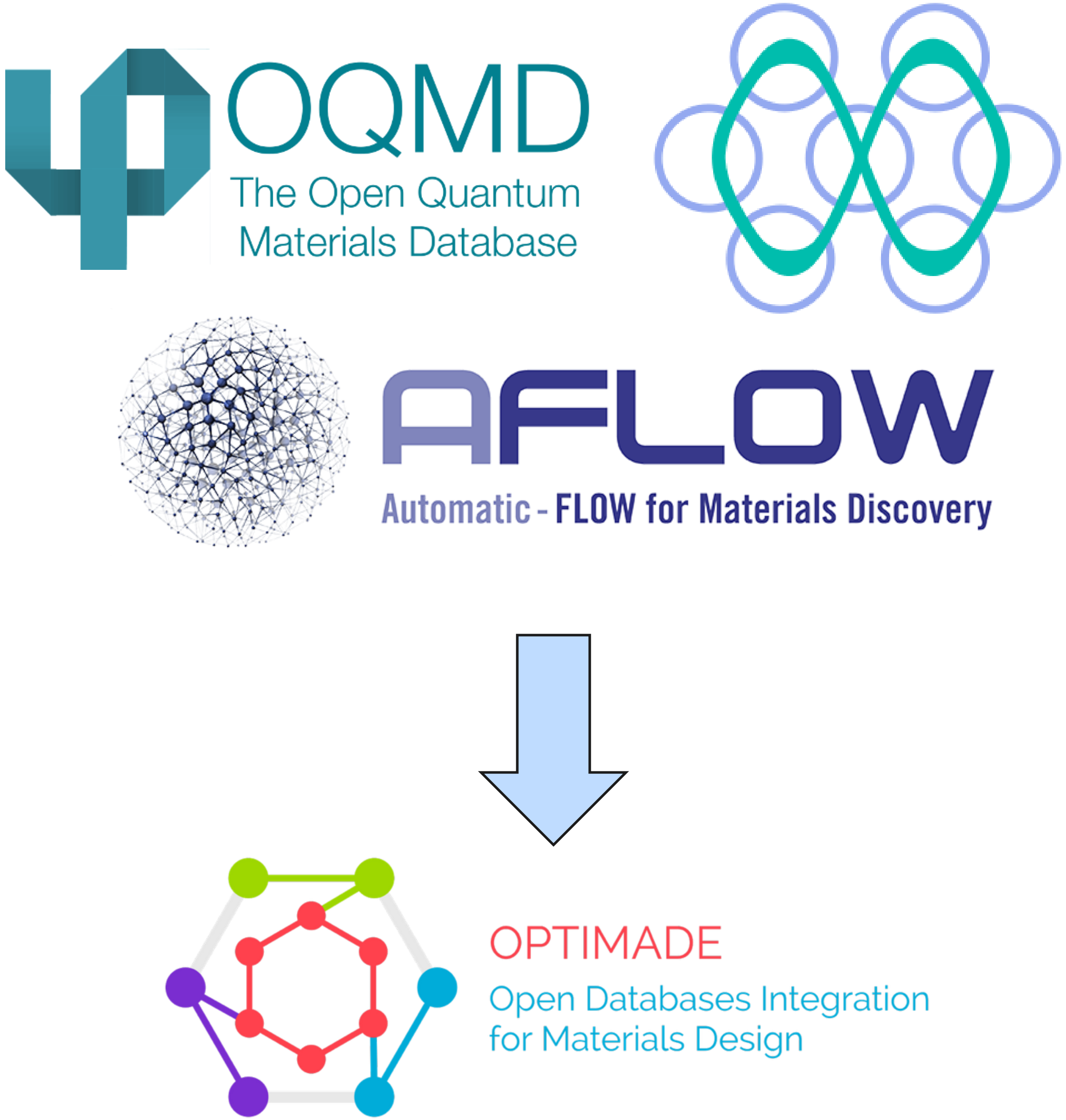
These are quite comprehensive, but not necessarily accurate. Usual DFT caveats—absolute energies may shift by tenths of an eV, and band gaps are typically underestimated, especially with high-throughput DFT! These are best for looking at trends, so for “big data” approaches, they can still be useful/exciting!

OQMD
The Open Quantum
Materials Database

AFLOW
Automatic - FLOW for Materials Discovery

Side-note: OPTIMADE

Optimade is an initiative to standardise and group all of the theoretical databases under one universal standard via consistent database structure, it contains properties for around 60,000,000 structures!


OQMD
The Open Quantum Materials Database
AFLOW
Automatic - FLOW for Materials Discovery
OPTIMADE
Open Databases Integration for Materials Design

Okay great, but I know how to click some buttons and download a structure, how do I get my hands on a massive dataset?

Rest APIs

A REST API (Representational State Transfer Application Programming Interface) is a standardized way for programs to talk to remote data servers over the internet. It allows users to access, query, and retrieve structured data — like material structures or computed properties.

Why Use REST APIs?

- **Programmatic access:** No more manual downloads
- **Scalable queries:** Pull data for thousands of materials as easily as pulling tens
- **Reproducibility**: Store query scripts with your codebase

Let’s take an example

Finding new solar absorbers on the Materials Project