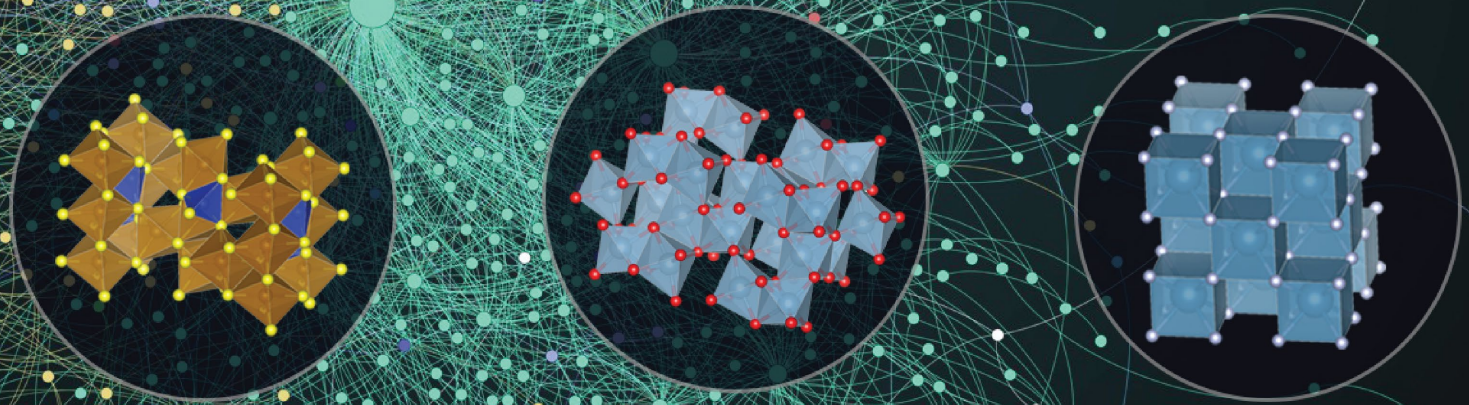
Materials Data: Repositories



Is materials data centralized…. Or dispersed?

vs

Is materials data centralized…. Or dispersed?

Mostly dispersed…

- DFT Hubs
- Raw calc archives
- HTE experiments
- Battery & Catalysis
- Structure Databases
- Reference / Curated
- Community Datasets
- Industrial / Private

But increasingly federated with standards

Interoperability layer:
APIs (Rest, OPTIMADE)
Schemas & ontologies (cif, GEMD)
Provenance and FAIR metadata

Outcomes for users
Cross repo searching
Reproducibility pipelines
Easier ML data assembly
Better citation and reuse

Important questions to ask about repos

Ask these six questions

1. Is it open, mixed, or subscription?
2. How do you access it? (REST, OPTIMADE, python client, bulk download?)
3. How curated is it? (expert review, automated validation, community upload)
4. What scale and coverage is there? (structures, calculations, measurements other?)
5. What data types (CIF, properties, raw files, metadata)
6. How FAIR is it (persistent DOIs, provenance, standardized metadata)

Important questions to ask about repos

CIF (structures)

- Widely used interchange format for crystallography
- Enables cross-database reuse of structures
- Often paired with quality/metadata fields

OPTIMADE (federated API)

- Standard REST interface for structures/properties
- Provider discovery via /links and /info
- >26M structures served across registered providers (snapshot)

GEMD / experimental schemas

- Captures process–structure–property context
- Helps encode synthesis + measurement metadata
- Useful for LIMS → ML pipelines

Provenance + FAIR

- Provenance (e.g., AiiDA/NOMAD) enables reproducibility
- Persistent IDs/DOIs for citation
- Metadata normalization for interoperability

DFT databases

Repository	Open?	API / Federation	Scale (approx.)	What you get
Materials Project	Open	REST + Python (mp-api) OPTIMADE	154k structures (OPTIMADE)	DFT-derived properties (thermo, electronic, elastic, etc.), IDs/DOIs; rate-limited API
AFLOW	Open	AFLUX/REST (OPTIMADE ecosystem)	3.5M+ entries	High-throughput computed structures + large property coverage; automated validation
OQMD	Open	REST (qmpy) Bulk download	~700k materials (API)	Total energies, formation energies, structures; useful for thermodynamics/phase diagrams
NOMAD	Open	Archive API OPTIMADE	18.7M structures (OPTIMADE) 50M+ calc. (archive)	Raw input/output + normalized metadata (code-independent), rich provenance
JARVIS (NIST)	Open	REST + tools OPTIMADE	80k+ materials "millions of properties"	DFT + FF + ML datasets; optical/phonon/elastic, potentials; curated workflows
Materials Cloud (MC3D)	Open	OPTIMADE AiiDA provenance	111.8k structures (OPTIMADE)	Curated, provenance-rich datasets; relaxed structures derived from experimental CIFs

Experimental and HTE databases

Repository	Open?	API / Access	Scale (approx.)	What you get
HTEM-DB (NREL)	Mixed	Web + API	~140k samples	Thin-film combinatorial experiments: synthesis/process metadata + structural/optoelectronic properties
Materials Data Facility (MDF)	Open	Publish/DOIs Python + REST	Many community datasets	General-purpose publishing & discovery for computational/experimental datasets; large-file transfer support
Materials Commons / PRISMS	Mixed	Web platform (Provenance)	Project-scale	Lab/project data management: raw files + workflow/provenance; collaboration & sharing
NanoMine (MaterialsMine)	Open	Web + schema	182 papers; 1k+ samples	Polymer nanocomposite experimental data with structured metadata for ML
Battery Archive + BEEP	Mixed	Web + Python (BEEP)	Community studies	Cycling curves + metadata; tooling for standardized featurization and comparisons
NIST Materials Data Repo	Open	REST + website	Varies	NIST-hosted datasets + best-practice repository; API available for programmatic access

Structure and curated databases (usually subscription)

Repository	Access model	API	Scale (approx.)	Primary data
CSD (CCDC)	Subscription	CSD Python API + tools	1.36M structures (Jan 2025)	Expert-curated organic / metal–organic crystal structures; rich validation and analysis tooling
ICSD (FIZ)	Subscription	Web tools (subscriber)	≈299k structures	Expert-curated inorganic structures with quality checks; biannual updates
MPDS / Pauling File	Subscription	API key (subscriber)	Handbook-scale	Curated experimental properties + phase diagrams + structures, extracted from literature
SpringerMaterials	Licensed	Web (licensed)	Handbook-scale	Curated property tables (Landolt–Börnstein) + phase diagrams and reference data
ICDD PDF	Subscription	Varies	Large	Powder diffraction reference patterns for phase identification
MatWeb	Mixed	Web	Large	Engineering datasheets; vendor/standard properties (useful, but licensing varies)

We have lots and lots of repos to look through

https://citrineinformatics.github.io/gemd-docs/
https://en.wikipedia.org/wiki/Crystallographic_Information_File
http://crystallography.net/cod/search.html
http://www.crystalimpact.com/pcd/
https://www.icdd.com/
https://www.fiz-karlsruhe.de/en/produkte-und-dienstleistungen/inorganic-crystal-structure-database-icsd
https://matbench.materialsproject.org/
https://github.com/anhender/mse_ML_datasets/tree/v1.0
https://link.springer.com/article/10.1007/s40192-020-00174-4
https://nanohub.org/resources/mastmltutorial
https://citrination.com/search/simple?searchMatchOption=fuzzyMatch
https://www.materialsdatafacility.org/
https://materialsdata.nist.gov/
https://materialsproject.org/
http://www.aflowlib.org/
http://oqmd.org/
https://github.com/sedaoturak/data-resources-for-materials-science
https://github.com/tilde-lab/awesome-materials-informatics
https://github.com/blaiszik/Materials-Databases

Materials Data: Access