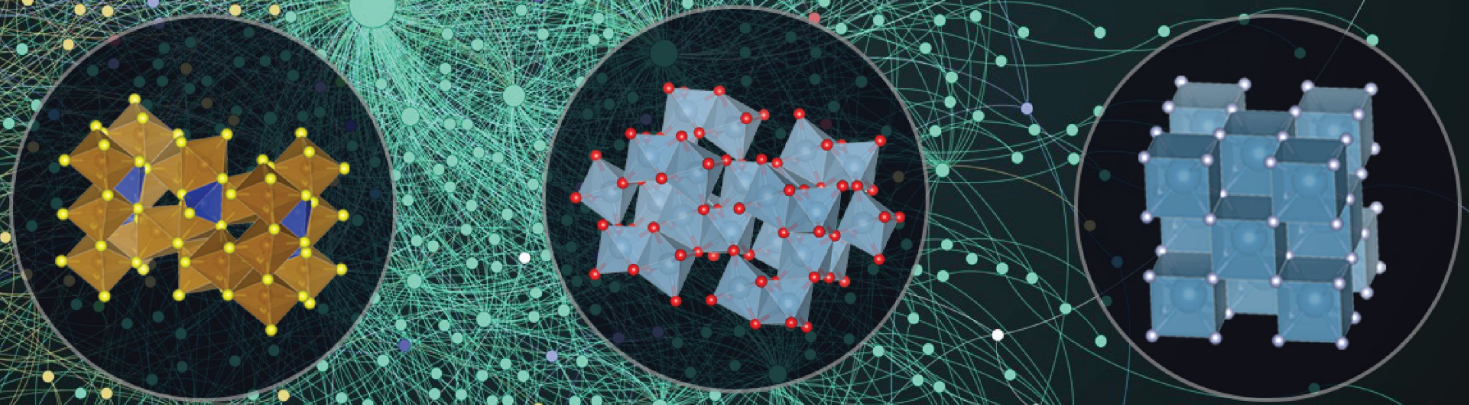
Materials Data: Repositories



Is materials data centralized…. Or dispersed?

vs

Is materials data centralized… or dispersed?

Mostly dispersed

- DFT hubs
- Raw calc archives
- HTE experiments
- Battery cycling
- Crystallography
- Reference / curated
- Community datasets
- Industrial/private

…but increasingly federated via standards

Interoperability layer

- OPTIMADE (federated structures API)
- CIF + domain schemas (structures)
- GEMD / ontologies (experiments)
- Provenance + FAIR metadata

Outcomes for users

- Cross-repo search & filtering
- Reproducible pipelines (provenance)
- Easier ML dataset assembly
- Better citation & re-use

A quick rubric for comparing repositories

Ask these six questions:

- Is it open, mixed, or subscription? (license + redistribution)
- How do you access it? (REST, Python client, bulk download, OPTIMADE)
- How curated is it? (expert review, automated validation, community upload)
- What scale & coverage? (structures vs calculations vs measurements)
- What data types? (CIF/structures, properties, raw files, metadata/provenance)
- How FAIR is it? (persistent IDs/DOIs, provenance, standardized metadata)

Standards & federation (why “one database” is hard)

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

Computational repositories (DFT / potentials / provenance)

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

Counts are snapshots; repositories also expose many more “documents” than structures (e.g., tasks, runs, properties).

Experimental + HTE repositories

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

Experimental data remains more heterogeneous than DFT: schemas + metadata quality are often the limiting factor.

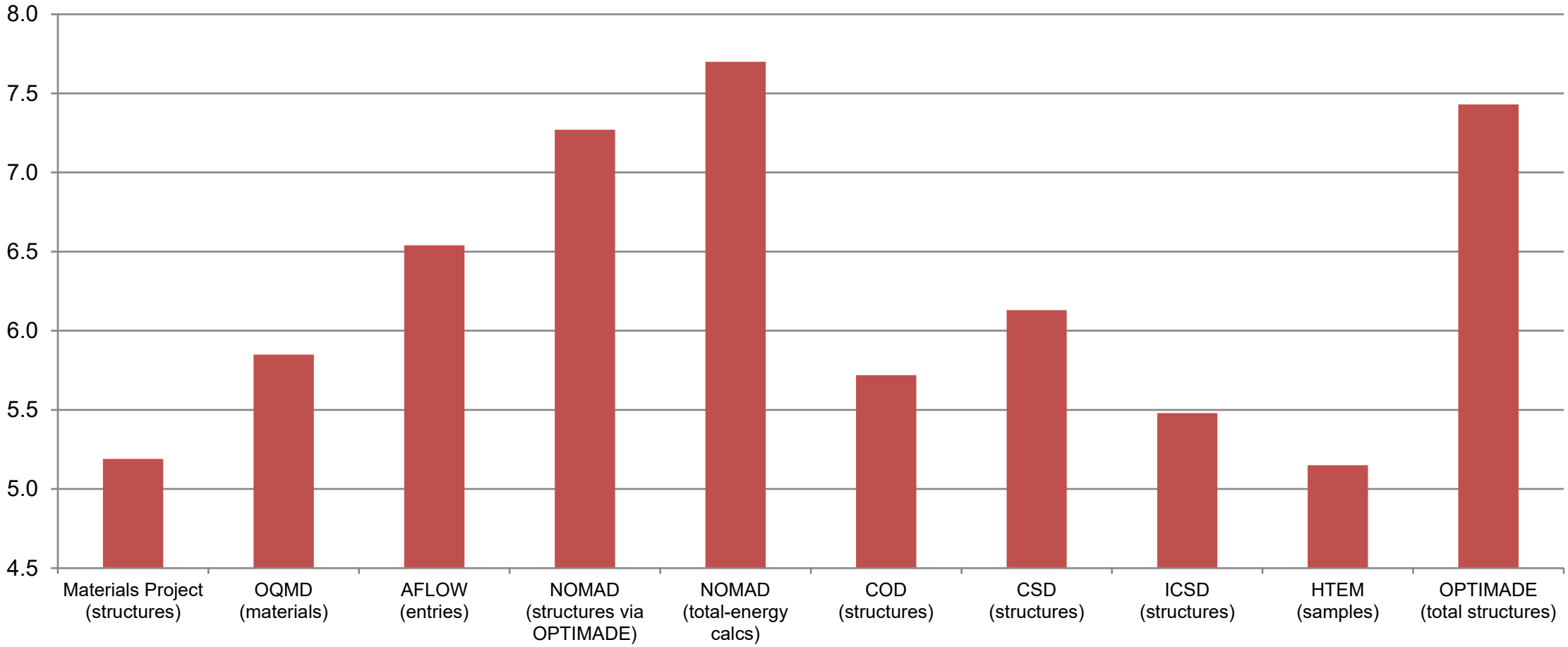
Curated reference data (often subscription)

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)

These resources are often the “ground truth” for validated experimental structures/properties, but have redistribution constraints.

Scale snapshot (order-of-magnitude, 2025–2026)

Approximate counts vary by definition (structures vs calculations vs measurements). Here we show commonly quoted figures.



Numbers compiled from repository/provider dashboards and documentation; see speaker notes for sources.

Practical access patterns

Recommended workflow:

- Start with an open API + client library (mp-api, optimade-python-tools, MDF Forge).
- Use OPTIMADE when you need cross-provider structure discovery; use native APIs for richer properties.
- Track IDs + provenance early (material IDs, DOIs, calculation hashes, instrument metadata).
- Respect rate limits and pagination; prefer vectorized/batched queries over loops.
- Be explicit about licenses and redistribution (especially subscription/reference data).
- Expect biases: positives-only reporting, inconsistent measurement conditions, missing failed experiments.

More places to explore…

Computational (open)

- Materials Project — materialsproject.org
- AFLOW — aflow.org
- OQMD — oqmd.org
- NOMAD — nomad-lab.eu
- JARVIS — jarvis.nist.gov
- Materials Cloud — materialscloud.org

Experimental + publishing

- HTEM-DB — htem.nrel.gov
- Materials Data Facility — materialsdatafacility.org
- Materials Commons / PRISMS — materialscommons.org / prisms-center.org
- NanoMine — materialsdata.nist.gov (MaterialsMine / NanoMine)
- Battery Archive + BEEP — batteryarchive.org

Reference (subscription)

- CSD (CCDC) — ccdc.cam.ac.uk
- ICSD (FIZ Karlsruhe) — icsd.fiz-karlsruhe.de
- MPDS / Pauling File — mpds.io
- SpringerMaterials — materials.springer.com
- ICDD PDF — icdd.com

Standards, federation, benchmarks

- OPTIMADE — optimade.org (providers dashboard)
- CIF — IUCr resources
- GEMD — Citrine gemd-docs
- Matbench — matbench.materialsproject.org
- Community lists — “awesome materials informatics” repos

This slide intentionally replaces long raw URL lists with an organized navigation map.

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