

Decentralized materials research data management, curation & dissemination for accelerated discovery

Dr Matthew Evans

<https://ml-evs.science>



FMRS24 Boston, USA, 3rd December 2024

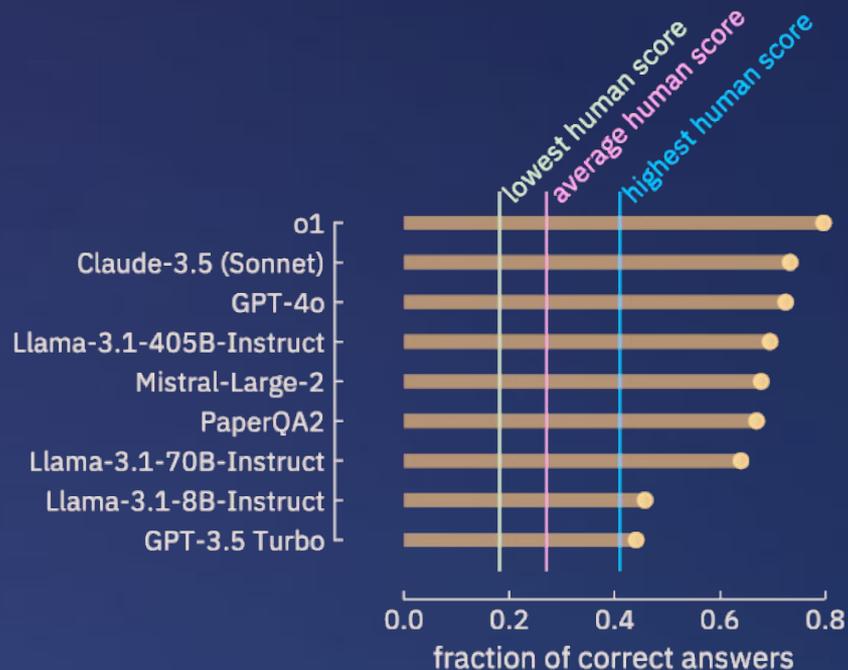
BI01: Democratizing AI in Materials Science—A Pathway to Broaden the Impact of Materials Research

Democratizing AI

- Access to compute (GPU-poor/rich)
- Access to expertise
- Access to models
- Getting your data seen
- Bringing your own tools to the AI

Are large language models superhuman chemists?

Adrian Mirza ^{1,2,*}, Nawaf Alampara ^{1,*}, Sreekanth Kunchapu ^{1,*},
Martino Ríos-García ^{1,3,*}, Benedict Emoekabu ¹, Aswanth Krishnan ⁴,
Tanya Gupta ^{5,6}, Mara Schilling-Wilhelmi ¹, Macjonathan Okereke ¹,
Anagha Aneesh ¹, Mehrdad Asgari ⁷, Juliane Eberhardt ⁸,
Amir Mohammad Elahi ⁹, Hani M. Elbeheiry ¹⁰, María Victoria Gil ³,
Christina Glaubitz ¹, Maximilian Greiner ¹, Caroline T. Holick ^{1,14},
Tim Hoffmann ^{1,14}, Abdelrahman Ibrahim ¹, Lea C. Klepsch ^{1,14},
Yannik Köster ¹, Fabian Alexander Kreth ^{11,12}, Jakob Meyer ¹, Santiago Miret ¹³,
Jan Matthias Peschel ¹, Michael Ringleb ^{1,14}, Nicole Roesner ^{1,14},
Johanna Schreiber ^{1,14}, Ulrich S. Schubert ^{1,2,10,14}, Leanne M. Stafast ^{1,14},
Dinga Wonanke ¹⁵, Michael Pieler ^{16,17}, Philippe Schwaller ^{5,6}, and
Kevin Maik Jablonka ^{1,2,11,14,18}



Broadening impact via decentralized platforms

OPTIMADE

Federated access to
crystal structure
data with universal
search for materials
discovery & design

data^{lab}

An open source data
management platform
for experimental
materials chemistry
and beyond

datatractor + MADICES

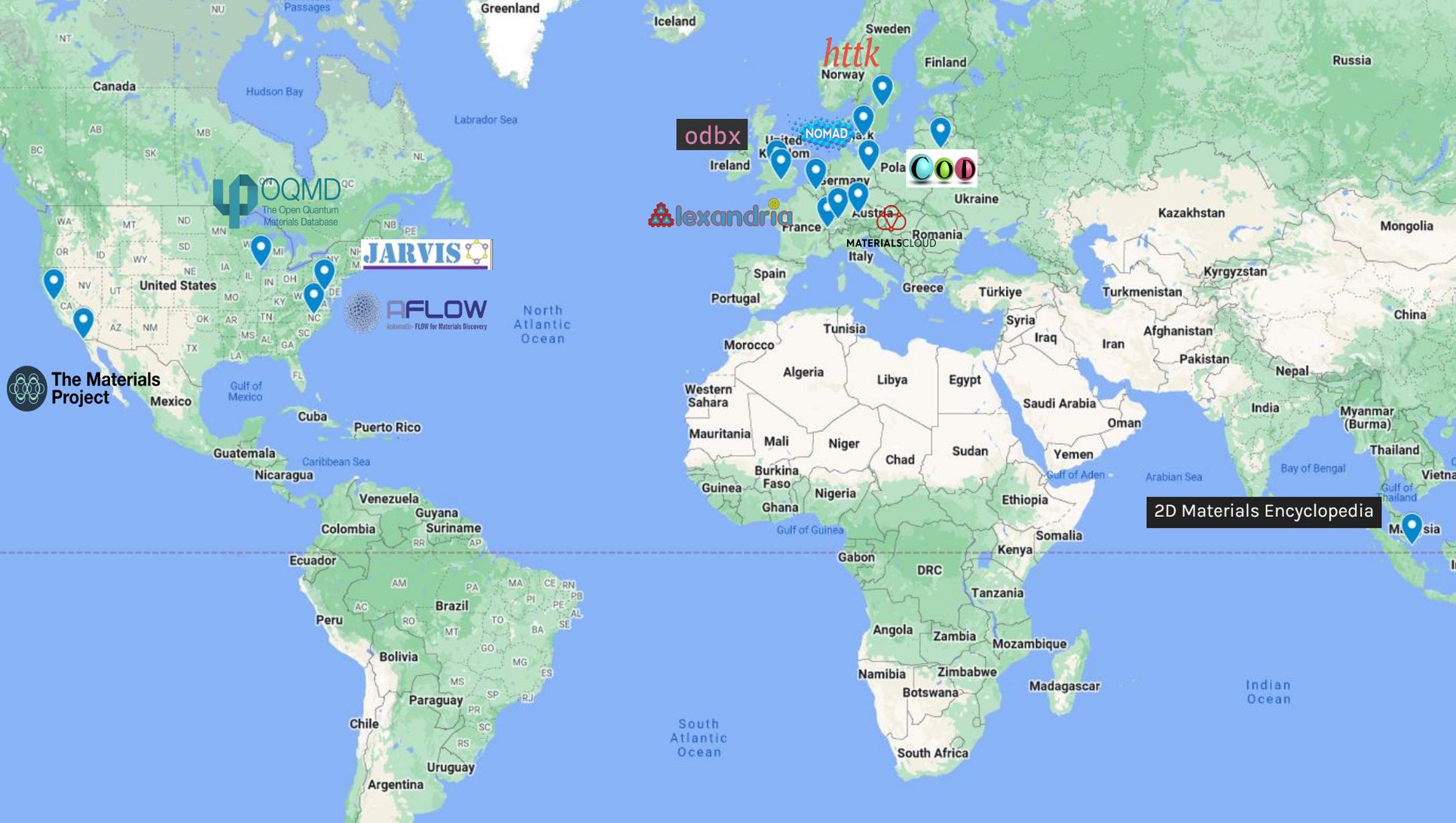
What is OPTIMADE?



“The Open Databases Integration for Materials Design (OPTIMADE) consortium aims to make materials databases interoperational by developing a common REST API.”

- ❖ 7 annual workshops (est. 2016)
- ❖ 70+ authors/attendees
- ❖ 35,671 words
- ❖ 25 registered providers, 60M+ crystal structures

<https://www.optimade.org/providers-dashboard>



OQMD
The Open Quantum
Materials Database

JARVIS
Automated FLOW for Materials Discovery

AFLOW
Automated FLOW for Materials Discovery

odbx

httk
Norway

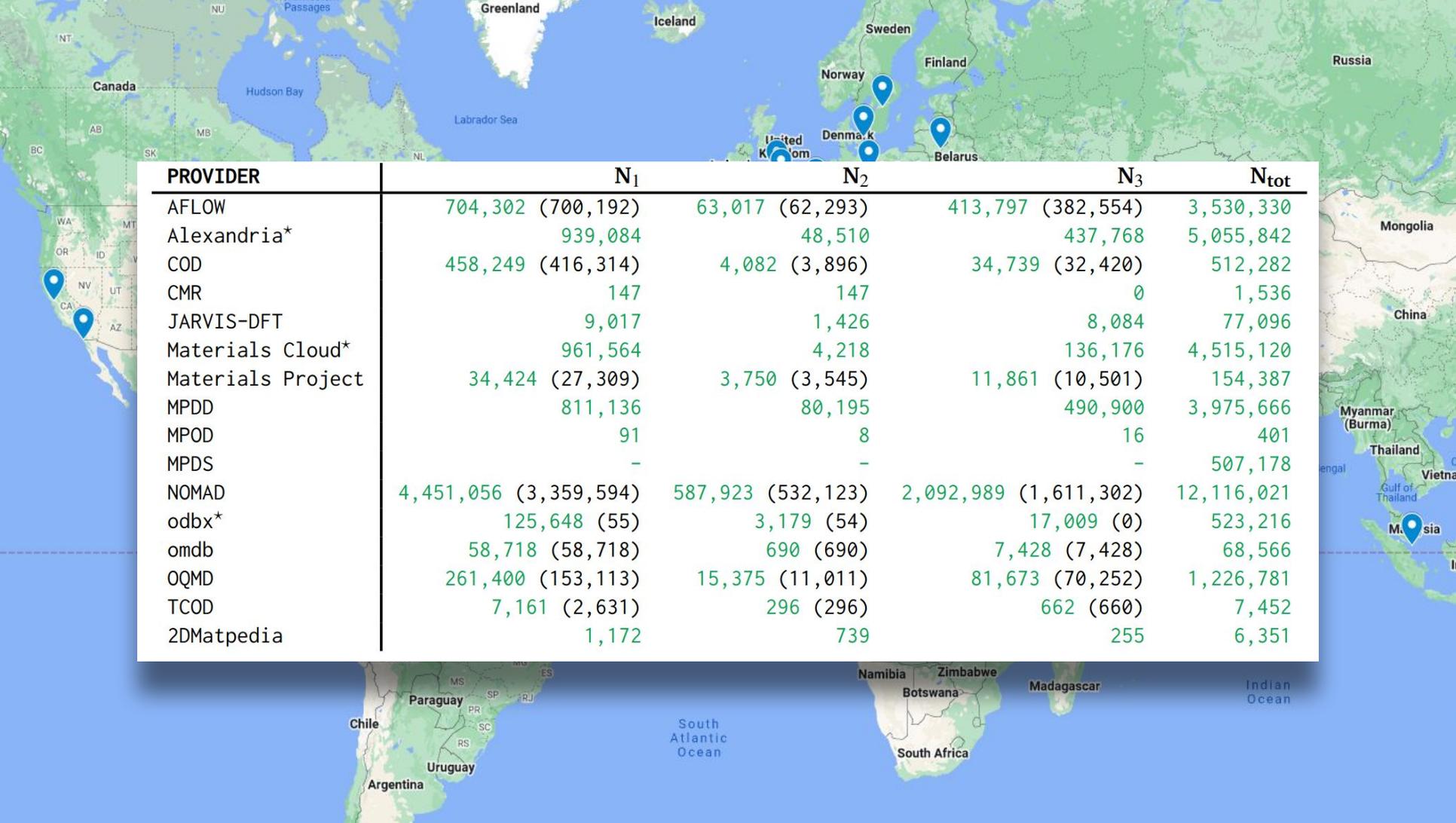
NOMAD
UK

alexandria

MATERIALS CLOUD

2D Materials Encyclopedia

The Materials Project



PROVIDER	N_1	N_2	N_3	N_{tot}
AFLOW	704,302 (700,192)	63,017 (62,293)	413,797 (382,554)	3,530,330
Alexandria*	939,084	48,510	437,768	5,055,842
COD	458,249 (416,314)	4,082 (3,896)	34,739 (32,420)	512,282
CMR	147	147	0	1,536
JARVIS-DFT	9,017	1,426	8,084	77,096
Materials Cloud*	961,564	4,218	136,176	4,515,120
Materials Project	34,424 (27,309)	3,750 (3,545)	11,861 (10,501)	154,387
MPDD	811,136	80,195	490,900	3,975,666
MPOD	91	8	16	401
MPDS	-	-	-	507,178
NOMAD	4,451,056 (3,359,594)	587,923 (532,123)	2,092,989 (1,611,302)	12,116,021
odbx*	125,648 (55)	3,179 (54)	17,009 (0)	523,216
omdb	58,718 (58,718)	690 (690)	7,428 (7,428)	68,566
OQMD	261,400 (153,113)	15,375 (11,011)	81,673 (70,252)	1,226,781
TCOD	7,161 (2,631)	296 (296)	662 (660)	7,452
2DMatpedia	1,172	739	255	6,351

Developments and applications of the OPTIMADE API for materials discovery, design, and data exchange†

Cite this: *Digital Discovery*, 2024, 3, 1509

DOI: 10.1039/d4dd00039k

Article | Open Access

OPTIMADE

Casper W. Andersen

Abhijith Gopalan

Markus Scheidgen

Curtarolo, Daniel

Hautier, Vincent

Kumbhar, Mohan

Francesco Ricci

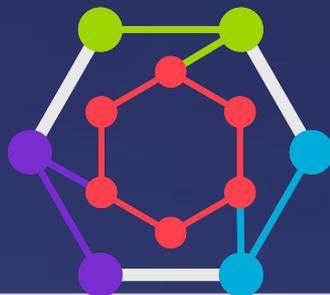
Chris Wolverton

Scientific Data

Matthew L. Evans,^{id a b} Johan Bergsma,^{id ‡ c} Andrius Merkys,^{id d} Casper W. Andersen,^{id e} Oskar B. Andersson,^{id f} Daniel Beltrán,^g Evgeny Blokhin,^{id h i} Tara M. Boland,^{id j} Rubén Castañeda Balderas,^{id k} Kamal Choudhary,^{id l} Alberto Díaz Díaz,^{id k} Rodrigo Domínguez García,^{id k} Hagen Eckert,^{id m n} Kristjan Eimre,^{id o} María Elena Fuentes Montero,^{id p} Adam M. Krajewski,^{id q} Jens Jørgen Mortensen,^{id j} José Manuel Nápoles Duarte,^{id p} Jacob Pietryga,^{id r} Ji Qi,^s Felipe de Jesús Trejo Carrillo,^{id k} Antanas Vaitkus,^{id d} Jusong Yu,^{o ah} Adam Zettel,^{id m n} Pedro Baptista de Castro,^{id t} Johan Carlsson,^{id u} Tiago F. T. Cerqueira,^{id v} Simon Divilov,^{id m n} Hamidreza Hajiyani,^{§ u} Felix Hanke,^{id w} Kevin Jose,^x Corey Oses,^{id y} Janosh Riebesell,^{id x z} Jonathan Schmidt,^{id aa} Donald Winston,^{id ab} Christen Xie,^s Xiaoyu Yang,^{id ac ad ae} Sara Bonella,^{id c} Silvana Botti,^{id af} Stefano Curtarolo,^{id m n} Claudia Draxl,^{id ag} Luis Edmundo Fuentes Cobas,^{id k} Adam Hospital,^{id g} Zi-Kui Liu,^{id q} Miguel A. L. Marques,^{id af} Nicola Marzari,^{id o ah} Andrew J. Morris,^{id ai} Shyue Ping Ong,^{id s} Modesto Orozco,^{id g} Kristin A. Persson,^{id z aj} Kristian S. Thygesen,^{id j} Chris Wolverton,^{id r} Markus Scheidgen,^{id ag} Cormac Toher,^{id n ak} Gareth J. Conduit,^{id x an} Giovanni Pizzi,^{id o ah} Saulius Gražulis,^{id d al} Gian-Marco Rignanese^{id * a b am} and Rickard Armiento^{id * f}

Fekete,
se ✉,
stefano
ni, Geoffroy
nehal
cell,
faroquiers,

Indian
Ocean



RESOURCE
REPRESENTATION

DATABASE AND
DOMAIN-AGNOSTIC
FILTERING GRAMMAR

PROVIDER
FEDERATION

MACHINE-ACTIONABLE
PROPERTY
DEFINITIONS

RESOURCE LAYOUT
(URL formats)



RESOURCE
REPRESENTATION

FILTERING
GRAMMAR

PROPERTY
DEFINITIONS

RESOURCE
LAYOUT

PROVIDER
FEDERATION

```
{
  "data": [
    {
      "id": "mp-7000",
      "type": "structures",
      "attributes": {
        "immutable_id": "645d2ba4bcd30f748b475981",
        "last_modified": "2023-03-11T14:56:30Z",
        "elements": ["O", "Si"],
        "nelements": 2,
        "elements_ratios": [0.3333333333333333, 0.6666666666666666],
        "chemical_formula_descriptive": "O6Si3",
        "chemical_formula_reduced": "O2Si",
        "chemical_formula_hill": "O6Si3",
        "chemical_formula_anonymous": "A2B",
        "dimension_types": [1, 1, 1],
        "nperiodic_dimensions": 3,
        "lattice_vectors": [
          [4.914966, -1e-8, 0],
          [-2.45748252, 4.2564861, 0],
          [0, 0, 5.43130114]
        ],
        "nsites": 9,
        "species_at_sites": ["Si", "Si", "Si", "O", "O", "O", "O", "O", "O"],
      }
    }
  ]
}
```



RESOURCE
REPRESENTATION

FILTERING
GRAMMAR

PROPERTY
DEFINITIONS

RESOURCE
LAYOUT

PROVIDER
FEDERATION

```
/v1/structures?filter=elements HAS ANY "C",  
"Si", "Ge", "Sn", "Pb"
```

```
/v1/structures?filter=elements HAS ANY "C",  
"Si", "Ge", "Sn", "Pb" AND nelements=2
```

```
/v1/structures?filter=elements HAS ANY "C",  
"Si", "Ge", "Sn" AND NOT elements HAS "Pb" AND  
elements LENGTH 3
```

RESOURCE
REPRESENTATION

FILTERING
GRAMMAR

PROPERTY
DEFINITIONS

RESOURCE
LAYOUT

PROVIDER
FEDERATION



```
{  
  "$id": "https://schemas.optimade.org/defs/v1.2/properties/optimade_  
  ↪ /structures/nsites",  
  "$schema": "https://schemas.optimade.org/meta/v1.2/optimade_  
  ↪ /property_definition.json",  
  "title": "number of sites",  
  "x-optimade-type": "integer",  
  "x-optimade-definition": {  
    "label": "nsites_optimade_structures",  
    "kind": "property",  
    "version": "1.2.0",  
    "format": "1.2",  
    "name": "nsites"  
  },  
  "type": [  
    "integer",  
    "null"  
  ],  
  "description": "An integer specifying the length of the  
  ↪ `cartesian_site_positions`  
  ↪ property.\n\n**Requirements/Conventions**:\n\n- MUST be equal to  
  ↪ the lengths of the list properties elements and elements_ratios,  
  ↪ if they are provided.\n\n**Query examples**:\n\n- Match only  
  ↪ structures with exactly 4 sites: `nsites=4`\n\n- Match structures  
  ↪ that have between 2 and 7 sites: `nsites>=2 AND nsites<=7`",  
  "examples": [  
    42  
  ],  
  "x-optimade-unit": "dimensionless"  
}
```



RESOURCE
REPRESENTATION

FILTERING
GRAMMAR

PROPERTY
DEFINITIONS

RESOURCE
LAYOUT

PROVIDER
FEDERATION

base URL

endpoint

<https://optimade.materialsproject.org/v1/structures>

?filter=[chemical_formula_reduced="O2Si"](#)

OPTIMADE filter

RESOURCE
REPRESENTATION

FILTERING
GRAMMAR

PROPERTY
DEFINITIONS

RESOURCE
LAYOUT

PROVIDER
FEDERATION



```
▼ 10:
  type: "links"
  id: "mcloudarchive"
  ▼ attributes:
    name: "Materials Cloud Archive"
    ▼ description: "Databases contributed by the community to the Materials Cloud Archive"
    base_url: "https://www.materialscloud.org/optimade/archive"
    homepage: "https://archive.materialscloud.org"
    link_type: "external"
  ▼ 11:
    type: "links"
    id: "mp"
    ▼ attributes:
      name: "The Materials Project"
      ▼ description: "An open database of computed materials properties to accelerate materials discovery and design"
      base_url: "https://providers.optimade.org/index-metadbs/mp"
      homepage: "https://www.materialsproject.org"
      link_type: "external"
```

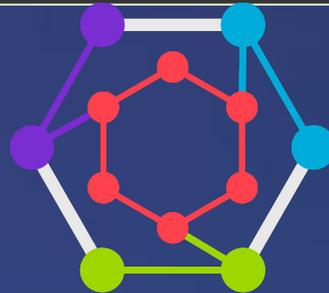
DATA-DRIVEN
WORKFLOWS

INTEROPERABILITY OF
REPRESENTATIONS

AUTOMATED
VALIDATION

DATA TO COMPLIANT
DATABASE PIPELINES

MULTI-PROVIDER
CLIENTS & QUERIES





optimade-python-tools

An open source Python package for consuming and implementing OPTIMADE APIs.



Evans, M. L., Andersen, C. W. A. et al.,
Journal of Open Source Software, 6(65), 3458, 2021
[10.21105/joss.03458](https://doi.org/10.21105/joss.03458)

- ❖ Spin-up OPTIMADE API with “no code”
 - Built with pydantic and FastAPI
 - Annotated data models with data validation
 - Auto-generated OpenAPI 3.0 and JSONSchema
 - EBNF grammar implementation with filter transformers for MongoDB & Elasticsearch
 - Mappers between existing formats (ASE, pymatgen, CIF) and OPTIMADE, supporting aliases etc.
- ❖ Client for asynchronously querying multiple databases
- ❖ Used by Materials Project, NOMAD, *odbx*, 2DMatPedia and Materials Cloud
- ❖ Provides tools for validating remote implementations

~ ▶  optimade_optimade-python-tools_exercise ▶ 05:54:25  24s

⌘

How can we screen the
burgeoning number of
hypothetical materials for
real applications?

ML-accelerated high-throughput workflows for optical materials discovery & design: *re2fractive*

with Victor Trinquet, Cameron Hargreaves, Pierre-Paul De Breuck & Gian-Marco Rignanese (MODL-IMCN UCLouvain)

MODL-IMCN, UCLouvain



[modl-uclouvain/re2fractive](https://github.com/modl-uclouvain/re2fractive)

Finding materials with the largest refractive index for a given band gap: *re2fractive*

Leveraging open source tools -- broaden impact by investing in maintenance



60M+
Hypothetical
inorganic crystal
structures*



Property
prediction from
small datasets
with kitchen sink
featurisation and
ensemble
uncertainties

ATOMATE2

Robust
automated
DFT+beyond
workflows

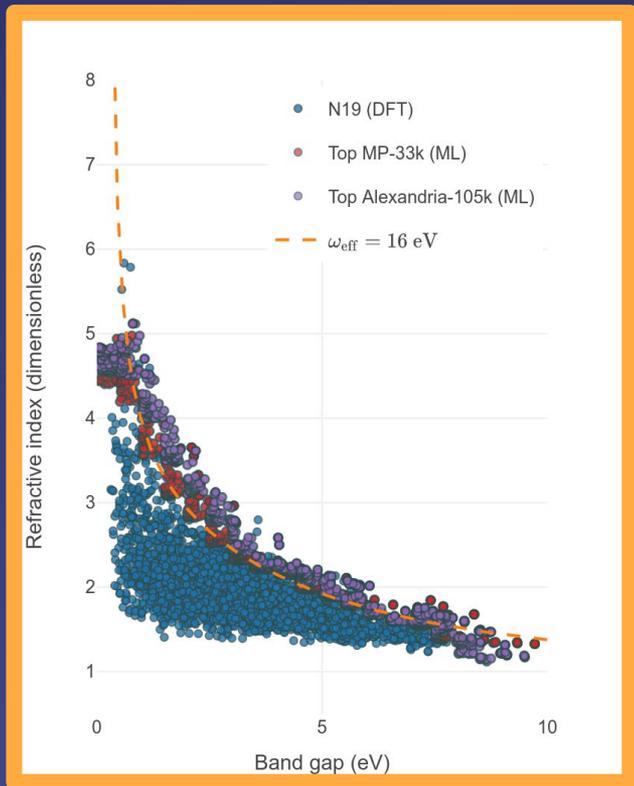


jobflow-remote

Automated
submission and
job handling

Publish repeatable workflows

Finding materials with the largest refractive index for a given band gap: *re2fractive*



Active learning loop

- I. Starting from ~4000 materials (Naccarato et al, Phys. Rev. Mat. 2019), train MODnet model for refractive index
- II. Screen databases with appropriate annotations in OPTIMADE (4.2m structures)
- III. Filter OPTIMADE entries for stable materials with non-zero band gap (~150k structures), featurize and predict
- IV. Select a mix of a) most promising and b) highest uncertainty predictions, compute them with DFT
- V. Retrain model with the new data and repeat until Pareto front is sufficiently explored

"Optical materials discovery and design via federated databases and machine learning", Trinquet, **Evans et al**, Faraday Discussions (2024)
[10.1039/D4FD00092G](https://doi.org/10.1039/D4FD00092G)

Finding materials with the largest refractive index for a given band gap: *re2fractive*

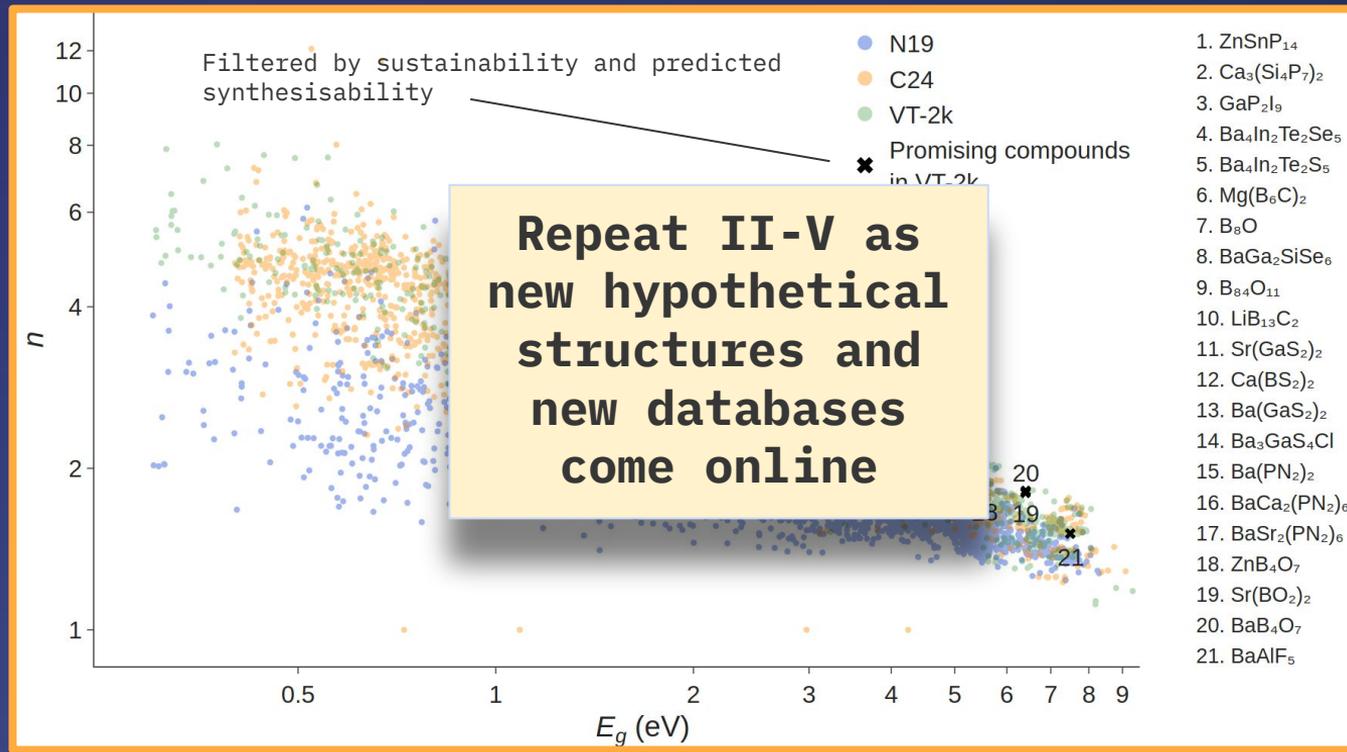
Repeat II-V as new hypothetical structures and new databases come online

Design with this in mind!

- I. Starting from ~4000 materials (Naccarato et al, Phys. Rev. Mat. 2019), train MODnet model for refractive index (#1 matbench leaderboard for dielectric task)
- II. Screen databases with appropriate annotations in OPTIMADE (4.2m structures)
- III. Filter OPTIMADE entries for stable materials with non-zero band gap (~150k structures), featurize and predict
- IV. Select a mix of a) most promising and b) highest uncertainty predictions, compute them with DFT
- V. Retrain model with the new data and repeat until Pareto front is sufficiently explored

"Optical materials discovery and design via federated databases and machine learning", Trinquet, **Evans et al**, Faraday Discussions (2024)
[10.1039/D4FD00092G](https://doi.org/10.1039/D4FD00092G)

Finding materials with the largest refractive index for a given band gap: *re2fractive*



"Optical materials discovery and design via federated databases and machine learning", Trinquet, **Evans et al**, Faraday Discussions (2024)

[10.1039/D4FD00092G](https://doi.org/10.1039/D4FD00092G)

Why stop at refractive index?

- Decentralised AL workflows scanning OPTIMADE databases daily for potential new compounds
- Created datasets themselves are then hosted via OPTIMADE, tracking provenance of initial structures
- Self-reinforcing loops for multi-fidelity property prediction can help prioritise new compounds for synthesis
- AL especially necessary when properties are rare + expensive to compute
- We have initialised a new campaign for non-linear optical response

Data Descriptor | [Open access](#) | Published: 11 July 2024

Second-harmonic generation tensors from high-throughput density-functional perturbation theory

[Victor Trinquet](#) , [Francesco Naccarato](#), [Guillaume Brunin](#), [Guido Petretto](#), [Ludger Wirtz](#), [Geoffroy Hautier](#) & [Gian-Marco Rignanese](#) 

[Scientific Data](#) **11**, Article number: 757 (2024) | [Cite this article](#)

"Optical materials discovery and design via federated databases and machine learning", Trinquet, **Evans et al**, Faraday Discussions (2024)
[10.1039/D4FD00092G](https://doi.org/10.1039/D4FD00092G)

optimade-maker

with Kristjan Eimre, & Giovanni Pizzi

From static data to an API

- Enhancing discoverability, explorability and expressiveness of structural data on the MCloud
- User annotates their MCloud entry with a simple config file
- Additional services watch for such entries and run a processing pipeline to convert to our OPTIMADE JSONL format
- Container launches to expose the structural data and properties as public OPTIMADE APIs

```
database_description: >-
  This database contains some example CIFs.

data_paths:
- structures.zip
- data.tar.gz

entries:
- entry_type: structures
  entry_paths:
  - structures/cifs/*.cif
  property_paths:
  - data/data.csv
  - data/data2.csv
  property_definitions:
  - name: energy
    title: Total energy per atom
    description: The total energy per atom as computed by DFT
    unit: eV/atom
    type: float
  - name: property_b
    title: Property B
    description: Alias for some more complicated property_b
    maps_to: https://schemas.optimade.org/v1.1/dft/dft_hull_distance
  - name: description
    title: Entry description
    description: Provides a human-readable description for this particular entry_type
    type: str
# - entry_type: references
# entry_paths:
# - refs.bib
```



optimade-maker

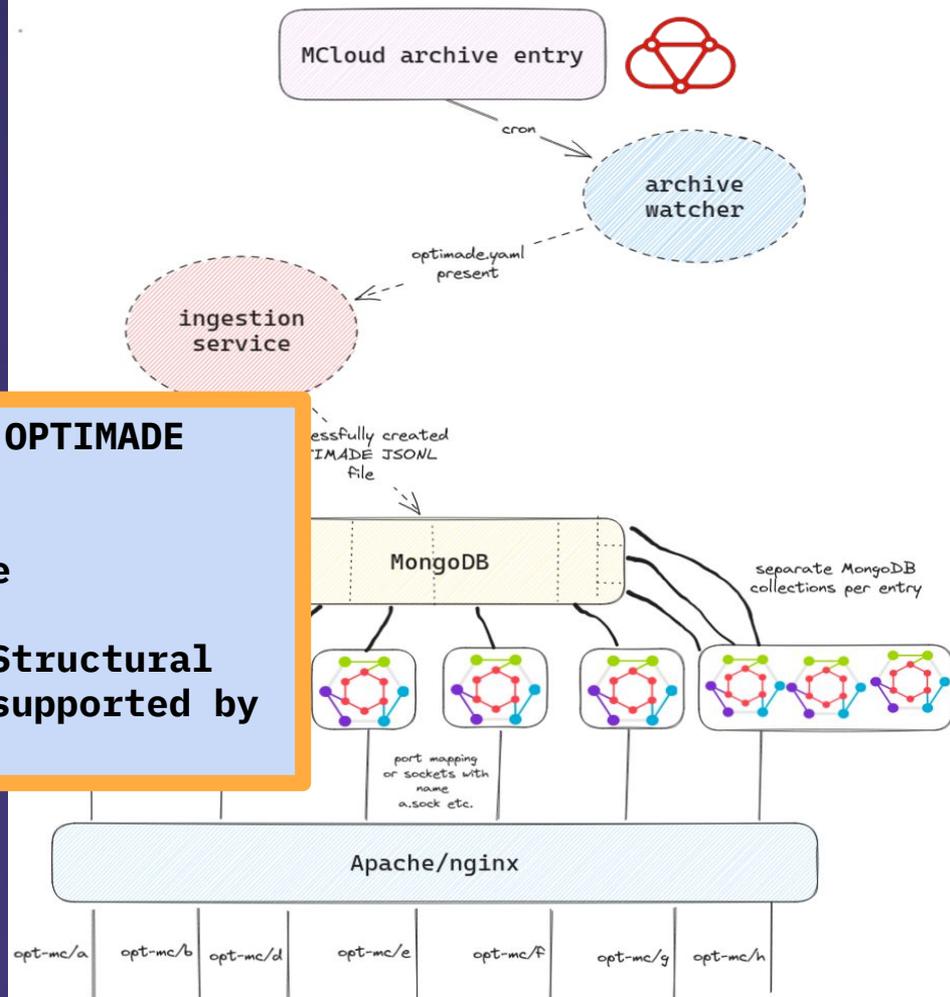
with Kristjan Eimre, & Giovanni Pizzi

From static data to an API

- Enhancing discoverability, explorability and expressiveness of structural data in the MCloud
- User annotates their MCloud with a simple config file
- Additional services watch for changes and run a processing pipeline to convert our OPTIMADE JSONL format
- Container launches to expose the structural data and properties as public OPTIMADE APIs

Used to create OPTIMADE APIs for:

- Alexandria
- re2fractive
- GNome
- Cambridge Structural Database (supported by PSDI)



Well-defined
abstraction: crystal
structure

Flat(ish) data
structure

**Extensible property
definitions**

Well-defined search

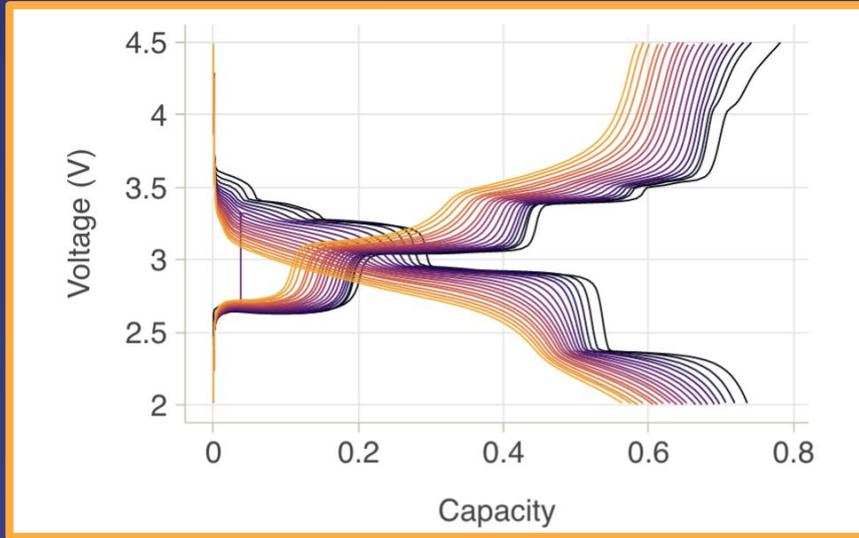
Tools for federation

What about
experimental
data?

Data management tech tree



Managing experimental lab data



What information would need to be recorded to make this information useful after the student leaves?

What information would need to be recorded to make this useful to a machine learning algorithm?

What type of cell is this? What are the electrodes? What batch are they from? Active mass? Where was it cycled? What instrument? Cycling parameters? Surface area? When was it made? Who made it? When? What batch of electrolyte? Was it made in a glovebox? Which one? Have other measurements been made on this cell? Has this result been repeated with other cells? Is there any characterization on the electrodes? On the active material? On the electrolyte? What was the temperature of the room? Were there any spikes? What does the dV/dQ look like? Who "owns" this data?

Experimental data is only useful in its context

Na₂O₂ from
ECO glovebox

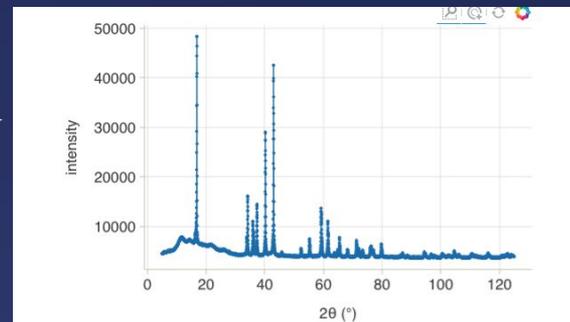
NiO (from
collaborator)

Jmas-1-4a

Jmas-1-4b

Jmas-1-4c

Jmas-1-4d



Super P

PVDF

NMP

Jmas-1-4a_e1

NaPF₆

PC

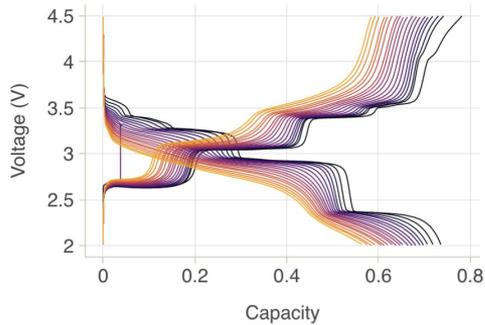
electrolyte

Na metal

Jmas1-4a-e1-c1

Jmas1-4a-e1-c2

Jmas1-4a-e1-c3



Initial motivations: *data*lab

 [datalab-org/datalab](https://github.com/datalab-org/datalab)

- Automating tedious data handling processes
- Reproducible, robust science
- Enhanced dissemination
- Putting the FUN in data management
- Decentralized data unification



Prof Joshua
Bocarsly
(Cambridge 
UHouston)



Josh Bocarsly



Logout

About | [Samples](#) | Collections | Inventory | Graph View

Add an item

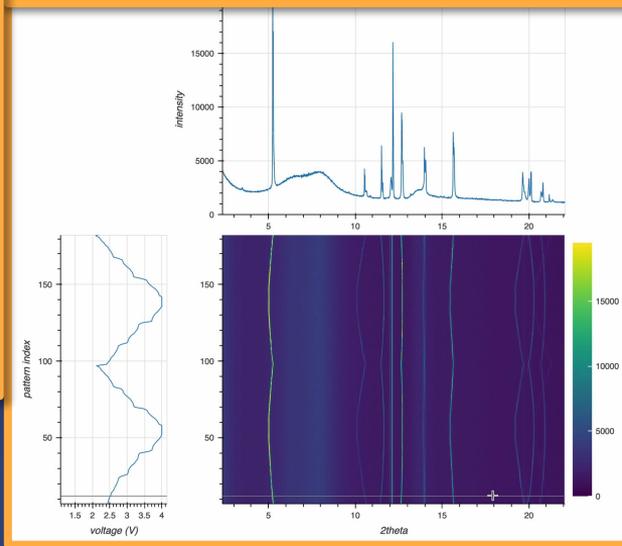
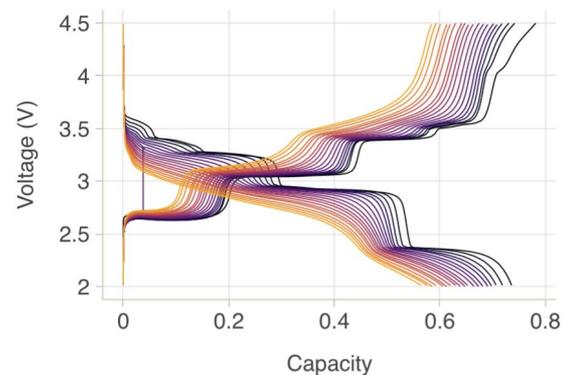
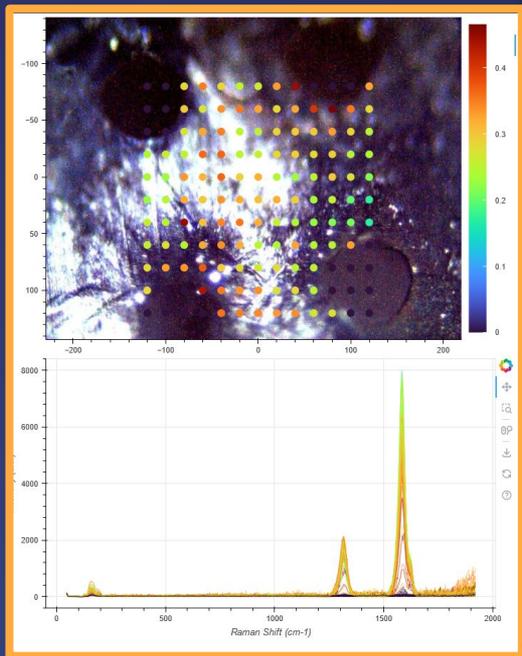
Add batch of samples

ID	Type	Sample name	Formula	Date	Collections	Creators	# of blocks
cell_13b	sample			2023-09-12			1 ×
test-02	cell	test cell		2023-09-08			2 ×
test-01	sample	test		2023-09-08			1 ×
jdb1003	sample	a sample of sodium solution		2023-09-07			1 ×
Ni3Al_1	sample	solid state sample of Ni-Al		2023-09-05			1 ×
jdb_test1	sample	NaCoO2 made from solid state synthesis		2023-09-05			1 ×
mp2028_T15	cell	Titration SEI 15 cell		2023-09-05			1 ×
mp2028_T14	cell	Titration SEI T14 cell		2023-09-05			1 ×
mp2028_T13	cell	Titration SEI T13 cell		2023-09-05			1 ×
Umi-33_PTFE_SP_38-14-48_SSF-OC1	cell	Umi-33_PTFE_SP_38-14-48_SSF-OC1		2023-09-03			0 ×
Umi-34_PTFE_SP_38-14-48_SSF-cell1	cell	Umi-34_PTFE_SP_38-14-48_SSF-cell1		2023-09-03			0 ×
Umi-33_PTFE_SP_38-14-48_SSF-cell1	cell	Umi-33_PTFE_SP_38-14-48_SSF-cell1		2023-09-03			0 ×

Modular "Data block" approach

Currently, we support:

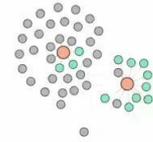
- Electrochemical cycling
- EIS
- Materials synthesis
- Cell preparation
- Images (common + raw)
- Videos
- Powder X-ray diffraction
- Raman (1D + 2D maps)
- NMR
- Mass spectrometry/TGA
- Your method?



Purposefully simple to add new "blocks":

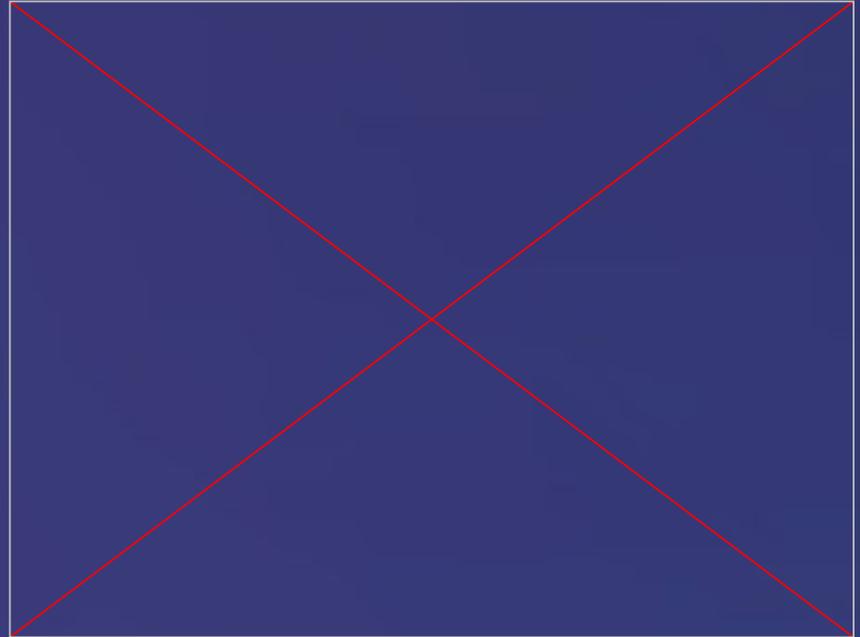
- File loader
- Interactive plot
- Baseline corrections, background subtraction, schemas, models, validation, search, comparative analysis...

Growth of the Bocarsly group data



*data*lab is open-source software (MIT)

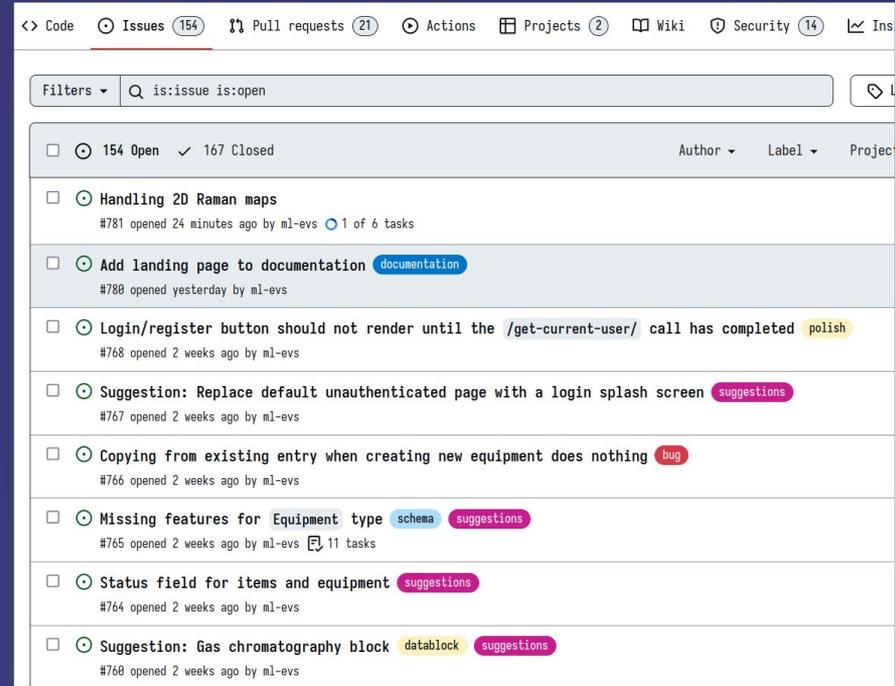
- Collaborative development is performed on GitHub using modern best practices for open source software development.
 - ◆ ~30,000 lines of code
 - ◆ 439 merged PRs
- Automated CI with testing of server and GUI
- Simplified deployment via Docker
- Automated cloud deployments with Ansible & Terraform
- Decentralizing maintenance costs via the community



github.com/data-lab-org/data-lab

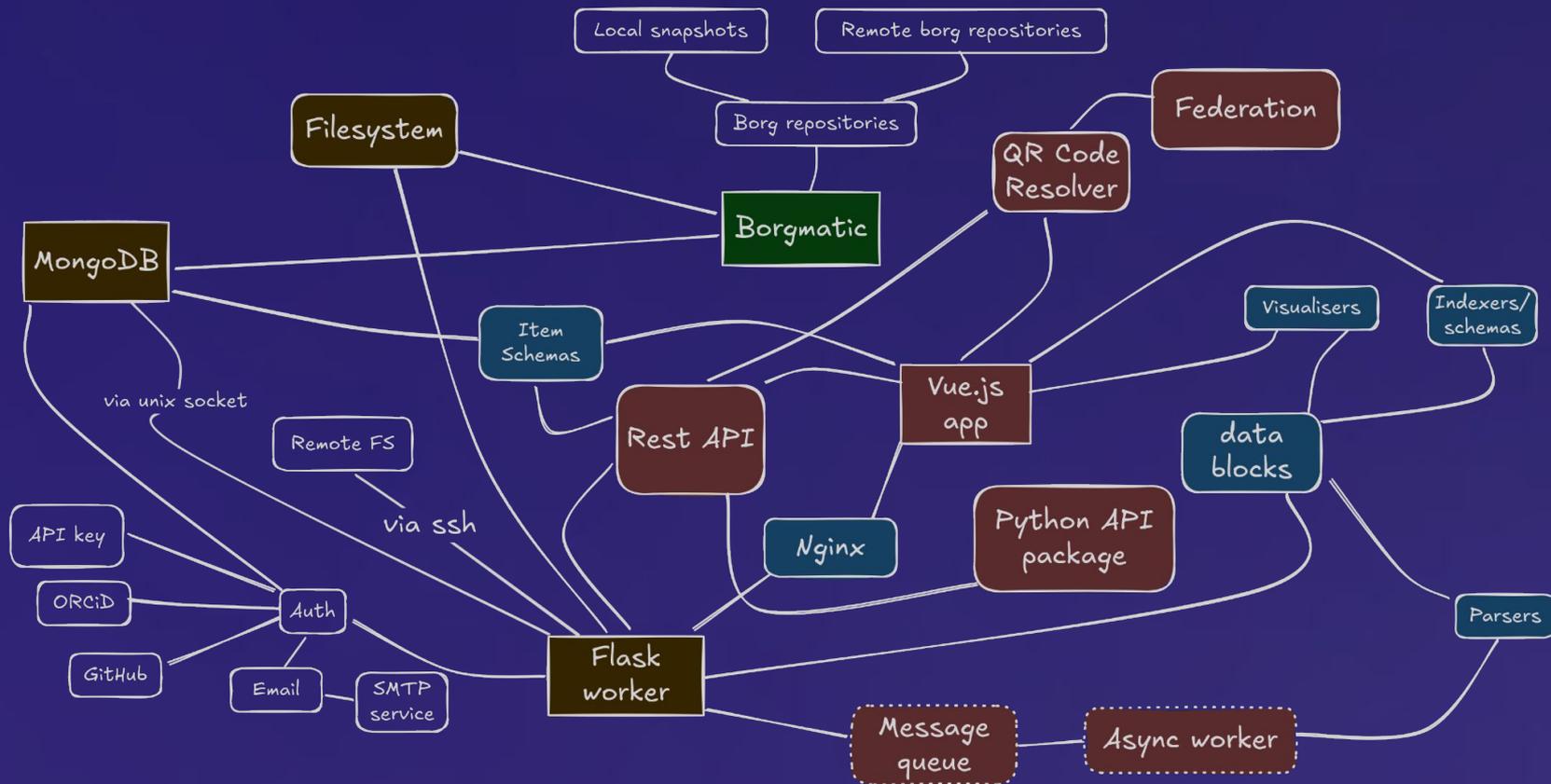
*data*lab is open-source software (MIT)

- Collaborative development is performed on GitHub using modern best practices for open source software development.
 - ◆ ~30,000 lines of code
 - ◆ 439 merged PRs
- Automated CI with testing of server and GUI
- Simplified containerized deployment via Docker (instructions online)
- Automated cloud deployments with Ansible & Terraform

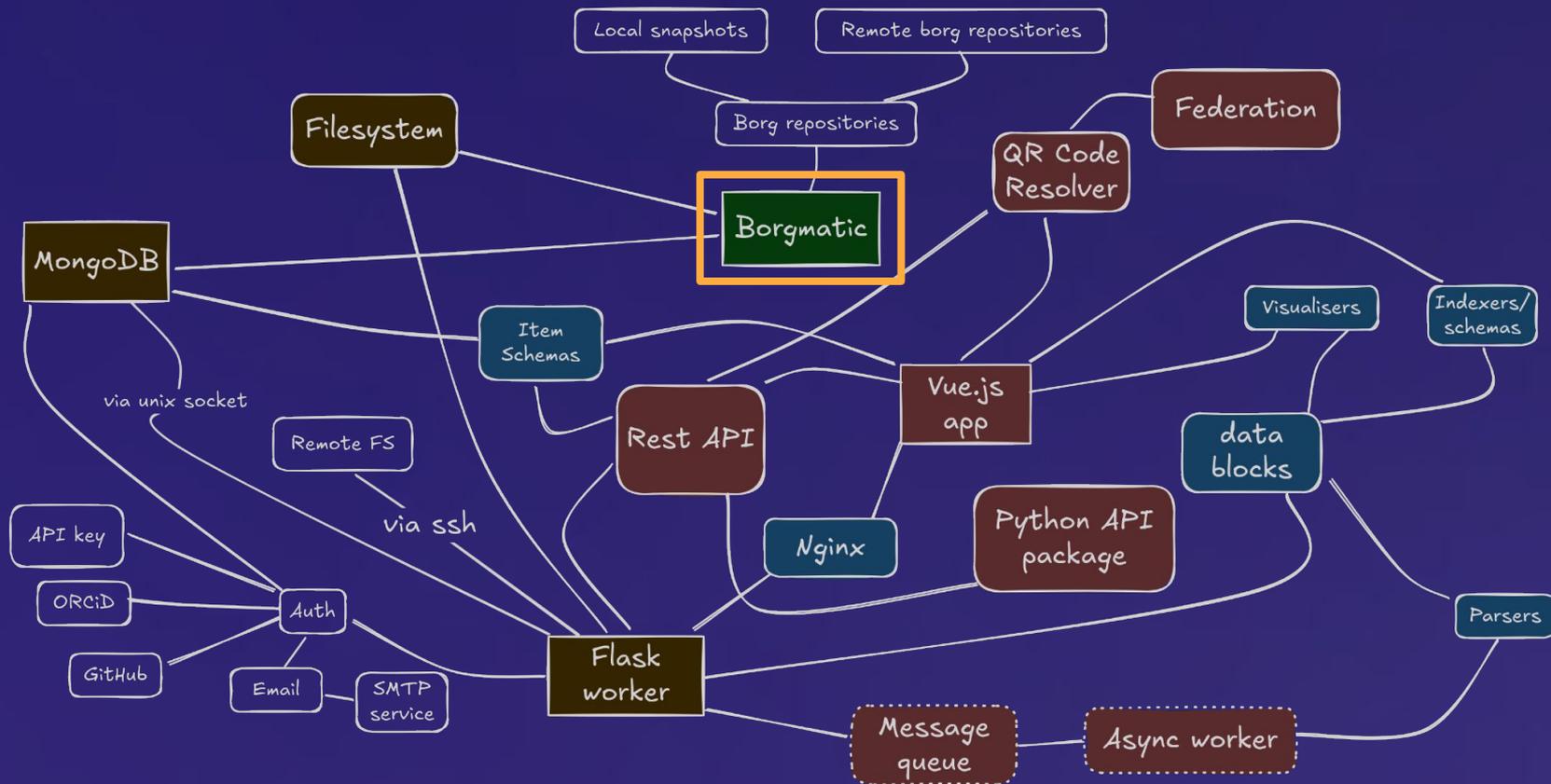


github.com/data-lab-org/data-lab

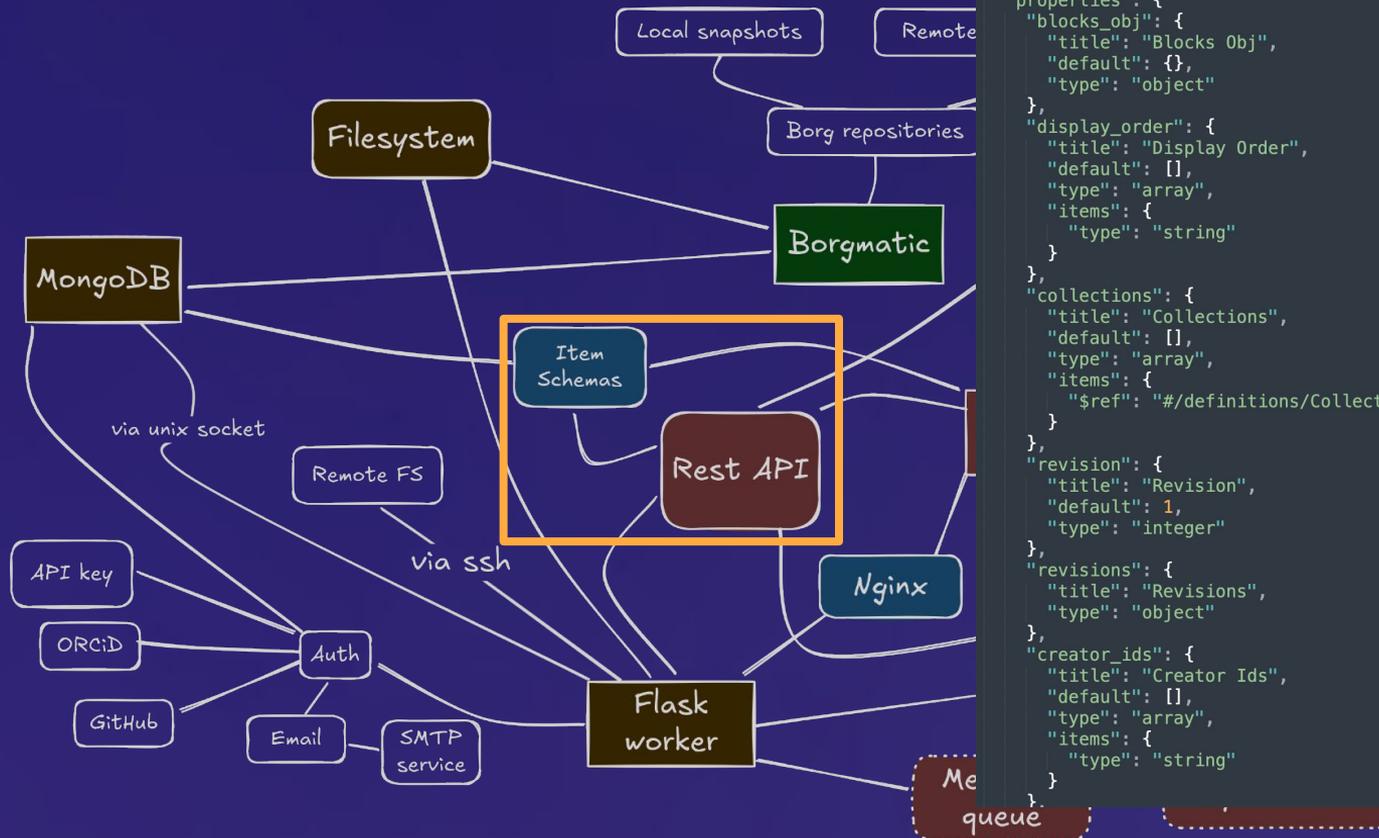
dataLab Architecture



dataLab Architecture

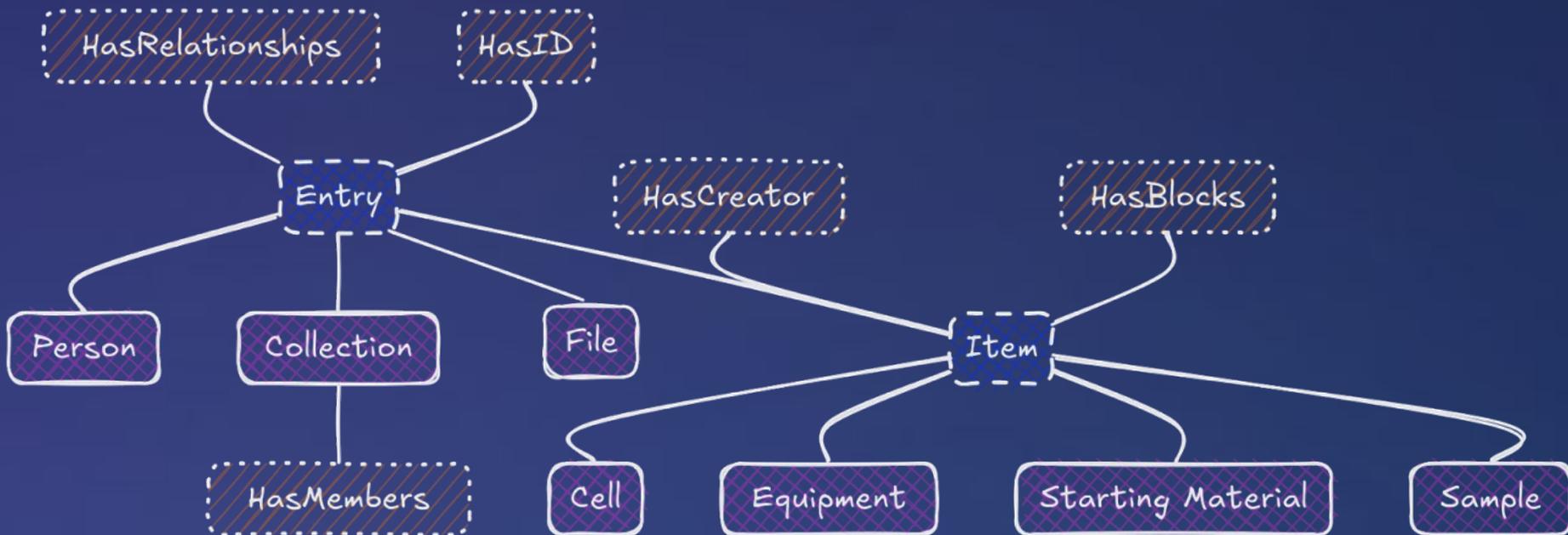


dataLab Architecture

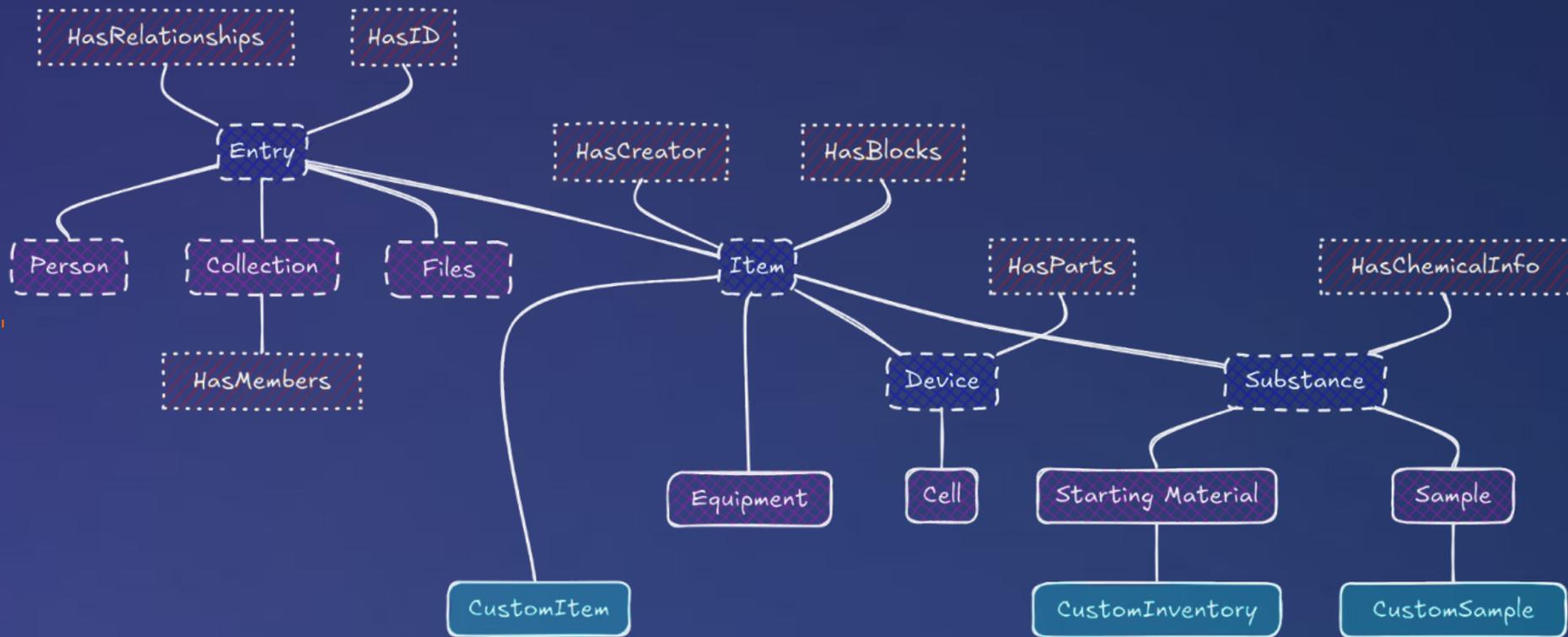


```
{
  "title": "Sample",
  "description": "A model for representing an experimental sample.",
  "type": "object",
  "properties": {
    "blocks_obj": {
      "title": "Blocks Obj",
      "default": {},
      "type": "object"
    },
    "display_order": {
      "title": "Display Order",
      "default": [],
      "type": "array",
      "items": {
        "type": "string"
      }
    }
  },
  "collections": {
    "title": "Collections",
    "default": [],
    "type": "array",
    "items": {
      "$ref": "#/definitions/Collection"
    }
  },
  "revision": {
    "title": "Revision",
    "default": 1,
    "type": "integer"
  },
  "revisions": {
    "title": "Revisions",
    "type": "object"
  },
  "creator_ids": {
    "title": "Creator Ids",
    "default": [],
    "type": "array",
    "items": {
      "type": "string"
    }
  }
}
```

dataLab data model: traits-based

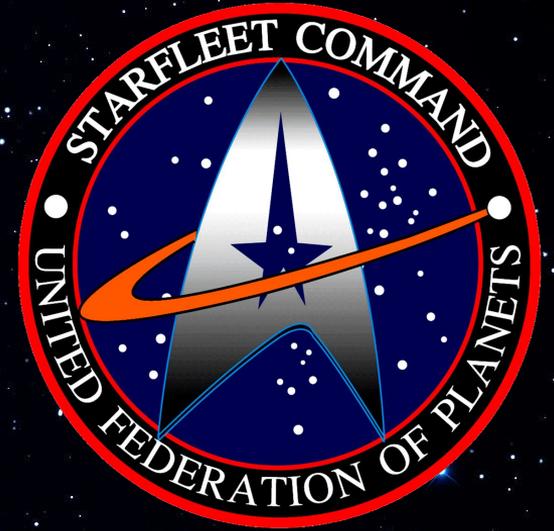


dataLab data model: custom schemas



*data*lab Federation

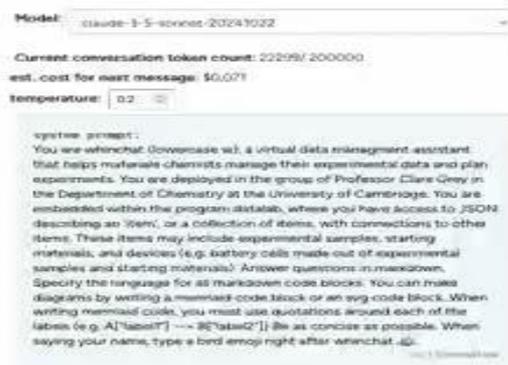
- Share and discover data between collaborators
- Share analysis workflows & schemas
- Share infrastructure: QR codes and persistent labelling
 - Decentralize & preserve safety information
- Private machine learning
 - opt-in with automatic data credit to train new models
- Robust offsite encrypted backups



- Grey group (Cambridge)
- Bocarsly group (UHouston)
- Cliffe group (Nottingham)
- Titirici group (Imperial)
- ConCat lab (TUBerlin)
- 20+ person startup (USA)
- Menkin group (Cambridge)
- Materials Research Lab (UCSB)

Whinchat

- Chat interface to *dataLab* (LLM hackathon 2023)
- Stuff structured data from a collection of samples into context window of an LLM, then interrogate



Whinchat

- Chat interface to *datalab* (LLM hackathon 2023)
- Stuff structured data from a collection of samples into context window of an LLM, then interrogate



Jablonka *et al*, *Digital Discovery*, 2023, 2

Hello! I'm whinchat 🐦, your virtual data management assistant. It looks like you've provided a comprehensive JSON dataset for a series of sodium cobalt oxide (NaCoO_2) coin cells and related samples, all made to test their electrochemical performance.

raw | formatted

In these experiments, is there a relationship between the battery performance and the synthesis or processing conditions?

raw | formatted

whinchat 🐦 here. Let me analyze the relationships between synthesis and performance:

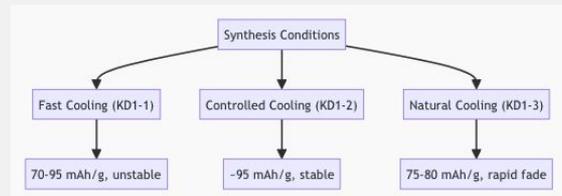
The samples (KD1-1, KD1-2, KD1-3) were all synthesized similarly at 850°C under O_2 , but with some key differences:

- KD1-1: 20 hours dwell, fast cooling (4 hours)
- KD1-2: 24 hours dwell, controlled cooling ($5^\circ\text{C}/\text{min}$)
- KD1-3: 24 hours dwell, natural cooling (15 hours)

Looking at cell performance:

- KD1-1 cells: ~70-95 mAh/g, significant degradation
- KD1-2 cells: Best performer with cell KD1-2_e1_c3 showing ~95 mAh/g with "reasonably stable cycling"
- KD1-3 cells: 75-80 mAh/g, all showing rapid capacity fade

The trend suggests that the controlled cooling rate (KD1-2) produced the most stable material, while both fast cooling (KD1-1) and very slow cooling (KD1-3) resulted in materials with poorer cycling stability. This could be related to the formation of optimal crystal structure or sodium content during the cooling process.



raw | formatted

YeLLowhaMMer

datalab-org/yellowhammer (with Josh Bocarsly & Yue Wu)

- Agentic interface to *datalab* -- still chat driven, but now uses our Python API directly (LLM hackathon 2024)
- Can automate complicated (not complex) data analysis tasks, e.g.
 - Search/RAG
 - Writing tools that we don't have time for
 - Multimodal/historic data input from spreadsheets, physical lab notebooks or even video
- Saves us writing UI workflows for complicated features!

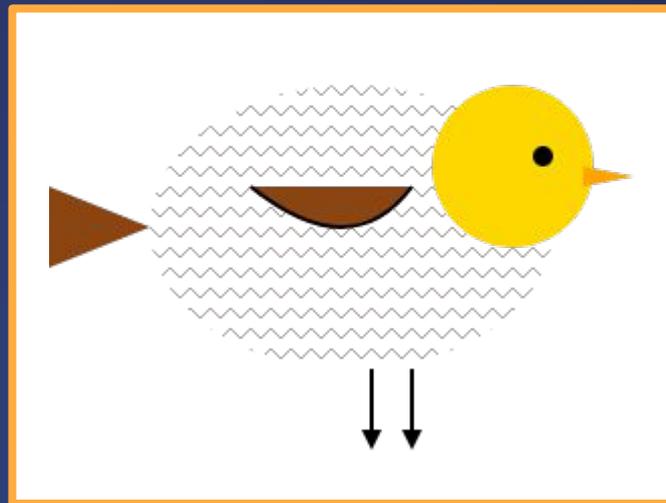


Zimmerman *et al* [arXiv:2411.15221](https://arxiv.org/abs/2411.15221) (2024)

YeLLowhaMMer

datalab-org/yellowhammer (with Josh Bocarsly & Yue Wu)

- Agentic interface to *dataLab* -- still chat driven, but now uses our Python API directly (LLM hackathon 2024)
- Can automate complicated (not complex) data analysis tasks, e.g.
 - Search/RAG
 - Writing tools that we don't have time for
 - Multimodal/historic data input from spreadsheets, physical lab notebooks or even video
- Saves us writing UI workflows for complicated features!

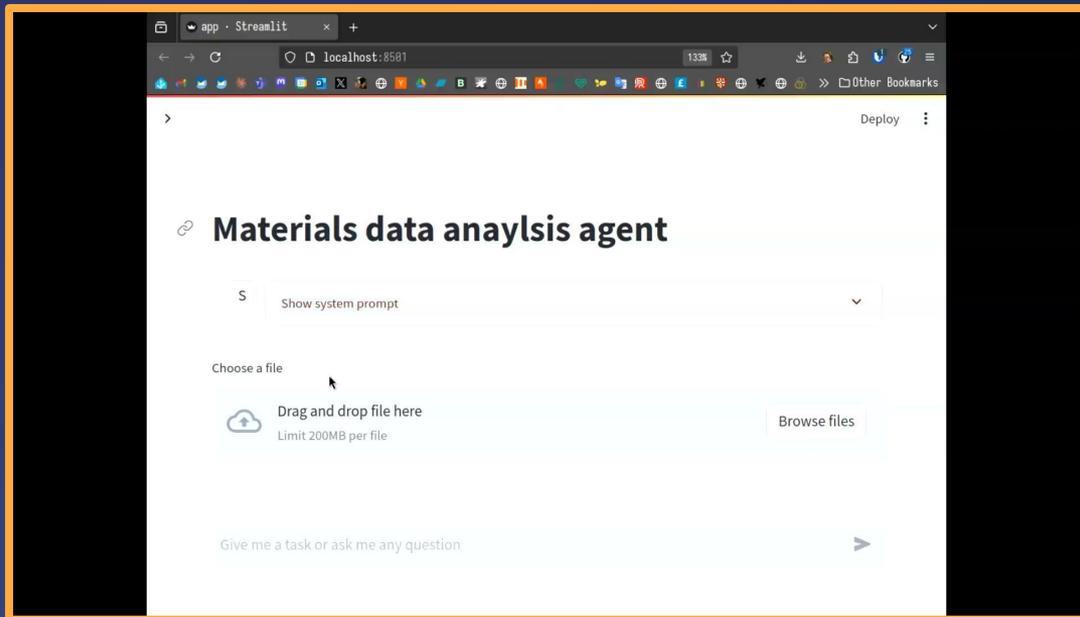


Zimmerman *et al* [arXiv:2411.15221](https://arxiv.org/abs/2411.15221) (2024)

YeLLowhaMMer

datalab-org/yellowhammer (with Josh Bocarsly & Yue Wu)

- Agentic interface to *datalab* -- still chat driven, but now uses our Python API directly (LLM hackathon 2024)
- Can automate complicated (not complex) data analysis tasks, e.g.
 - Search/RAG
 - Writing tools that we don't have time for
 - Multimodal/historic data input from spreadsheets, physical lab notebooks or even video
- Jupyter "magic" `%%llm` interface reduces complexity



Zimmerman *et al* [arXiv:2411.15221](https://arxiv.org/abs/2411.15221) (2024)

YeLLowhaMMer

datalab-org/yellowhammer (with Josh Bocarsly & Yue Wu)

- Jupyter “magic” `%%llm` interface reduces complexity
- Can make sure to track the provenance of AI-generated data for human inspection
- UI components for guiding the AI through common workflows
- Designing our tools for AI:
 - `/llms.txt` and agent-readable documentation that is always up to date
 - API schemas and introspective endpoints useful for tool-calling
 - Expose in decentralized way?

Zimmerman *et al* [arXiv:2411.15221](https://arxiv.org/abs/2411.15221) (2024)

*data*lab Roadmap + sustainability

- Plugin ecosystem
- Deployment-level schema customisation
- Sidecar Jupyter “Hub”
- Secure some sustainable funding (<https://datalab.industries> & <https://matgenix.com>)
- Proper/automatic exports to archive servers and integrations with other ELNs
- Dashboards of live experiments
- ~~● Ingesting historic data~~
- Integration with lab “knowledge bases”
 - Wiki, Scheduling, Slack
- ~~● Advanced chat block:~~
 - ~~○ Custom plots and API calls~~
- **Deep search via OPTIMADE**
- Integrating ML models fine-tuned on “private” data for e.g., DoE
 - Mroz *et al*, Web-BO Faraday Discussions (2024)

Decentralisation means we
don't have to all use the
same **tools** or **schemas**

but agreeing on **protocols**
can be very powerful

Datatractor (MaRDA extractors working group)

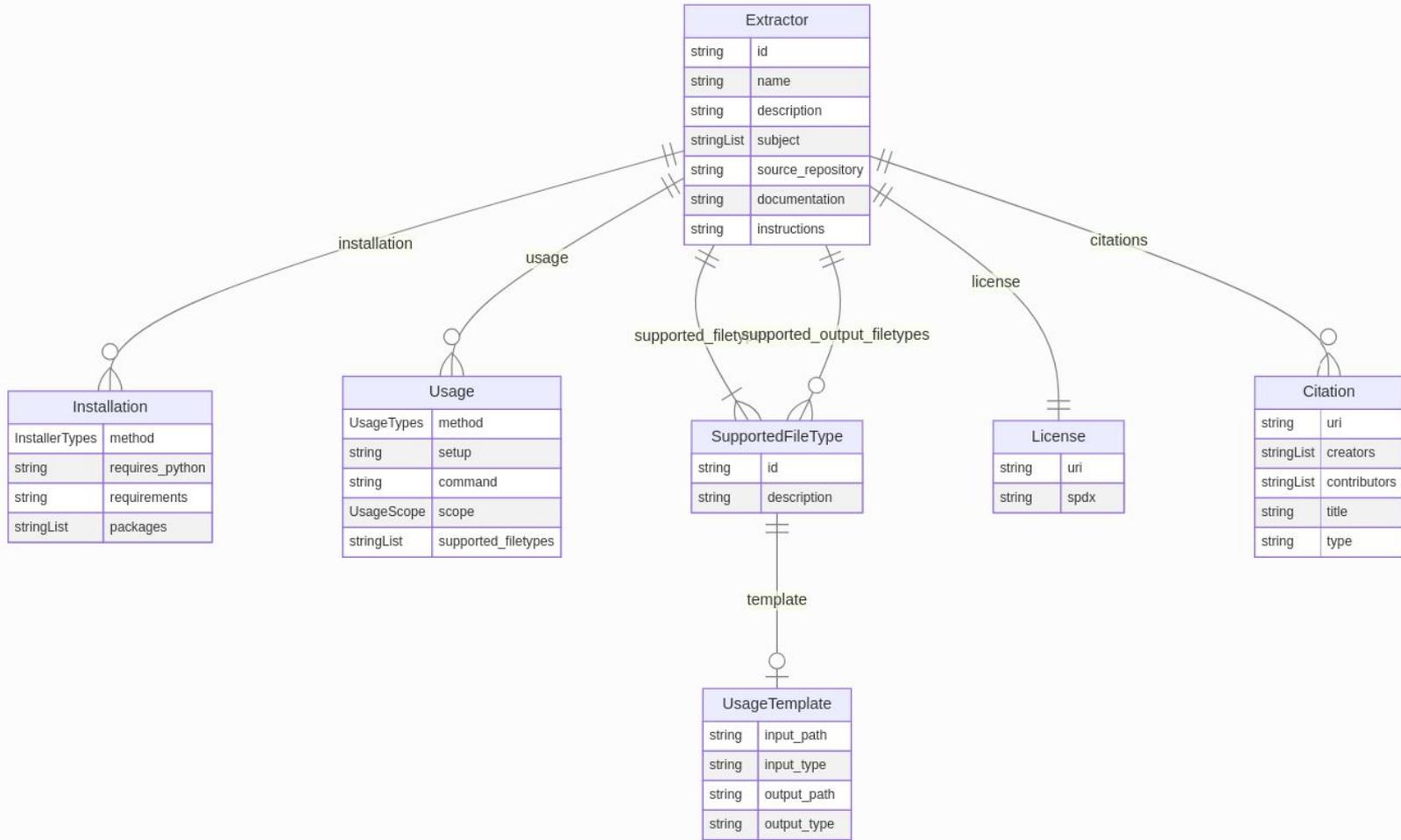
w/*Peter Kraus (TU Berlin)* & *David Elbert (JHU)*

Supporting all possible instrument file format types in a single platform/framework is not possible

How can we communalize or decentralize this work?

- Common schema for describing file types and extractor code
- Machine-actionable definitions for extractor installation & usage
- Online registry of such file types and extractors
- API package that can automatically install and extract





Datatractor (MaRDA extractors working group)

w/Peter Kraus (TU Berlin) & David Elbert (JHU)

About

`{}` JSON API

 GitHub

File types

Extractors

[agilent-ch](#) (Agilent ChemStation export file)

A binary data file created by Agilent ChemStation OpenLab software, for liquid chromatography files contain the limits of the retention time as well as the signal data.

[More details](#)

[agilent-dx](#) (Agilent OpenLab export file)

A raw data export file for Agilent OpenLab software, actually a zipped archive containing other data file formats.

```
JSON Raw Data Headers
Save Copy Collapse All Expand All Filter JSON
▼ data:
  id: "yadg"
  name: "yet another datagram"
  ▶ description: "yadg is a set of tools to... raw instrument data."
  ▶ subject: [...]
  ▶ citations: [...]
  ▶ license: {}
  ▶ supported_filetypes: [...]
  supported_output_filetypes: null
  source_repository: "https://github.com/dgbowll/yadg"
  documentation: "https://dgbowll.github.io/yadg"
  ▼ usage:
    ▼ 0:
      method: "cli"
      setup: null
      ▼ command: "yadg extract {{ input_type }} {{ input_path }} {{ output_path }}"
      scope: "meta+data"
      supported_filetypes: null
    ▶ 1: {}
    ▶ 2: {}
  ▼ installation:
    ▼ 0:
      method: "pip"
      requires_python: "≥3.10"
      requirements: null
```

yet another datagram)

3-only
github.com/dgbowll/yadg
/dgbowll.github.io/yadg

set of tools to extract and parse raw instrument

ns:

the package into a Python 3.9+ environment with `pip install yadg`. After activating the environment, the `extract` command will be available at the command-line.

file types:

[agilent-mpr](#)

Note: Several fields in file headers are not translated into meaningful parameters.

[agilent-mpi](#)

[agilent-ch](#)

[agilent-dx](#)

Note: Only .ch files contained in the .dx archive are parsed.

[agilent-spe](#)

[agilent-lytical-xrdml](#)

:

maintained by: P. Kraus, N. Vetsch, C. Battaglia, *yadg*: yet another datagram



[datatractor](#)

Evans et al, arXiv:2410.18839 (2024)

Datatractor (MaRDA extractors working group)

w/Peter Kraus (TUBerlin) & David Elbert (JHU)

```
from beam import extract

# extract(<input_type>, <input_path>)
data = extract("./example.mpr", "biologic-mpr")
```

Follows installation and usage instructions then installs in container, piping resulting data back

```
data>>> data
<xarray.Dataset> Size: 196kB
Dimensions: (uts: 721)
Coordinates:
  * uts      (uts) float64 6kB 1.557e+09 ... 1.557e+09
Data variables: (12/31)
  Ns        (uts) int64 6kB 0 0 0 0 0 0 ... 0 0 0 0 0 0
  time      (uts) float64 6kB 1.089e+05 ... 1.521e+05
  time_std_err (uts) float64 6kB 5e-05 5e-05 ... 5e-05 5e-05
  control_V  (uts) float64 6kB 0.3 0.3 0.3 ... 0.3 0.3 0.3
  control_V_std_err (uts) float64 6kB 2e-05 2e-05 ... 2e-05 2e-05
  Ewe       (uts) float64 6kB 0.1464 0.2987 ... 0.2986
  ...
  mode      (uts) int64 6kB 2 2 2 2 2 2 ... 2 2 2 2 2 2
  ox or red (uts) int64 6kB 1 1 1 1 1 1 ... 1 1 1 1 1 1
  error     (uts) int64 6kB 0 0 0 0 0 0 ... 0 0 0 0 0 0
  control changes (uts) int64 6kB 1 1 1 1 1 1 ... 1 1 1 1 1 1
  Ns changes (uts) int64 6kB 0 0 0 0 0 0 ... 0 0 0 0 0 0
  counter inc. (uts) int64 6kB 0 0 0 0 0 0 ... 0 0 0 0 0 0
Attributes:
  provenance: yadg extract
  date: 2024-02-21 16:36:20
  datagram_version: 5.0.2
  yadg_extract_filename: /tmp/tmpo276u26f
  yadg_extract_filetype: filetype='marda:biologic-mpr' timezone='Europe/Lo...
```

Datatractor (MaRDA extractors working group)

w/Peter Kraus (TU Berlin) & David Elbert (JHU)

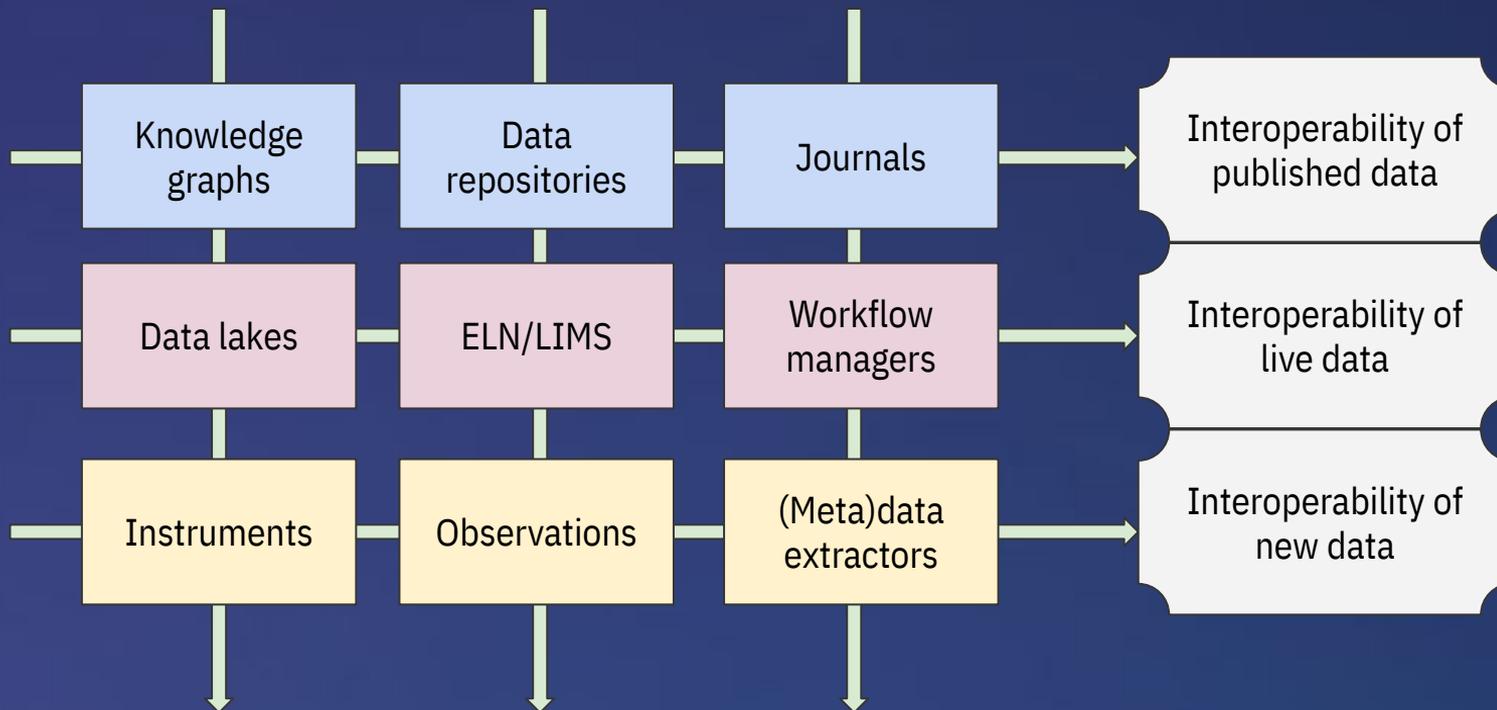
Remaining challenges

- More robust file type detection
- Automatically running registered extractors against example files
- Providing a place in the schema for announcing output schemas
- Dockerized execution environment (with security features, e.g., no access to network or other files)
- Streamlining contributions to registry and deciding policies for versioning (without implementing our own package manager!)

Potential use cases

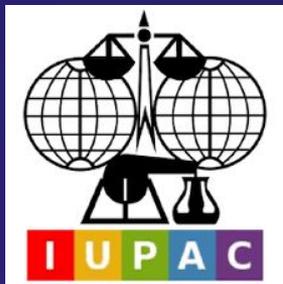
- Create dynamic *dataLab* plugins form datatractor to render data in an ELN
 - And publish *dataLab* plugins as datatractor extractors!
- Extending domain-specific archive APIs
 - Currently allows an `optimade.yaml` config file to create an OPTIMADE API on top of archived data, but could use `datatractor.yaml` in the future
 - Enhances sustainability of platforms like *dataLab*

Interoperability hierarchy



Machine-Actionable Data Interoperability for Chemical Sciences (MADICES)

- CECAM workshop series: 2022 and 2024 -- hopefully 2025
- Practicable standards for interoperability across:



Concluding thoughts

- Federation enables qualitatively distinct applications
- Treat data management as an investment, rather than a cost
- LLMs != data management
- Can your data be discovered accidentally?
- Can you easily repeat and share systematic screening when new data arises?

Call to arms

OPTIMADE:

- Interesting material dataset?
 - Materials data providers
 - Potential users
 - Developers
 - Interested in adopting a generic federated API format?
- + Monthly meetings to discuss extensions, yearly CECAM workshops
- + Check out [Materials-Consortia](#) on GitHub

*data*lab, *data*tractor & MADICES

- Dealing with tricky interconnected data?
 - Interested groups/companies
 - Individual researchers doing this work already, plugin or schema development?
 - Looking for a postdoc? Talk to Josh!
- + Monthly *data*lab meetings
- + Communal slack
- + Demo server demo.data-lab.org.io
- + [data-lab-org](#) and [data-tractor](#) on GitHub
- + Talk to Josh and I this week!
- + MADICES mailing list and pre-meetings

Acknowledgements



Prof Josh Bocarsly
(Houston)



Prof Clare Grey
(Cambridge)



Prof Gian-Marco Rignanesi
(UCLouvain)



Dr David Waroquiers
(Matgenix)

- + OPTIMADE consortium
- + Materials Research Data Alliance (MaRDA)
- + Physical Sciences Data Infrastructure (PSDI)
- + *re2fractive*: Victor Trinquet (UCLouvain)
- + datatractor: Dr Peter Kraus (TU Berlin)
- + Ben Charmes & Dr Yue Wu
- + The Grey Group (Cambridge)

Useful links

OPTIMADE

<https://optimade.org>

Digital Discovery, 2024, **3**, 1509-1533
[10.1039/D4DD00039K](https://doi.org/10.1039/D4DD00039K)

*data*lab ecosystem

<https://datalab-org.io>

<https://datalab.industries>

<https://demo.datalab-org.io>

Datatractor / MaRDA Extractors WG

<https://yard.datatractor.org>

MADICES

<https://madices.github.io>



OPTIMADE



PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE