Status of the NOMAD Archive

NOMAD SAC Meeting Barcellona 4-5 Oct. 2016

Fawzi Mohamed FHI Berlin

NOMAD

NOVEL MATERIALS DISCOVERY

NOMAD Archive: Goal

NOMAD Archive: Goal

NOVEL MATERIALS DISCOVERY

NOMAD REPOSITORY

NOMAD Archive

Make data accessible for the simulation codes used in the community

WP2: NOMAD ENCYCOPEDIA

WP3: NOMAD *VISUALIZATION*

WP4: BIG DATA ANALYSIS

External users

NOMAD Archive: What is it

NOVEL MATERIALS DISCOVERY

NOMAD

NOVEL MATERIALS DISCOVERY

NOMAD Archive: What is it

- A growing collection of files
  - Containing calculation data organized with **metadata**
  - With clear **identifiers** to enable workflows and automatic processing
  - Open access

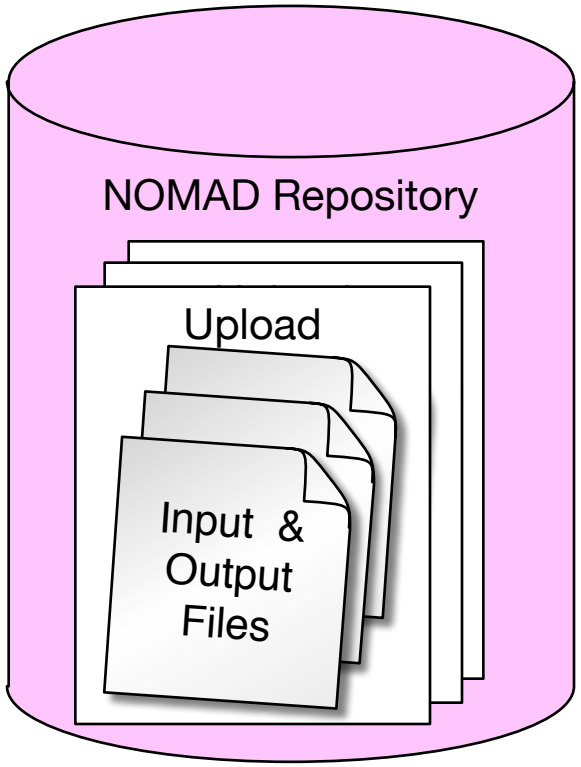
NOMAD

NOVEL MATERIALS DISCOVERY

NOMAD Archive: How is it built

NOMAD Repository

NOVEL MATERIALS DISCOVERY



- http://nomad-repository.eu
- Source of our data
- established to host organize and share materials data
- Keeps data for at least 10 Years
- Open access
- Joint effort by the groups of
  - Matthias Scheffler, FHI Berlin
  - Claudia Draxl, HU Berlin
  - Max Planck Computer & Data Facility (MPCDF), Garching, headed by Stefan Heinzel.

Claudia Draxl
HUB

Matthias Scheffler
FHI

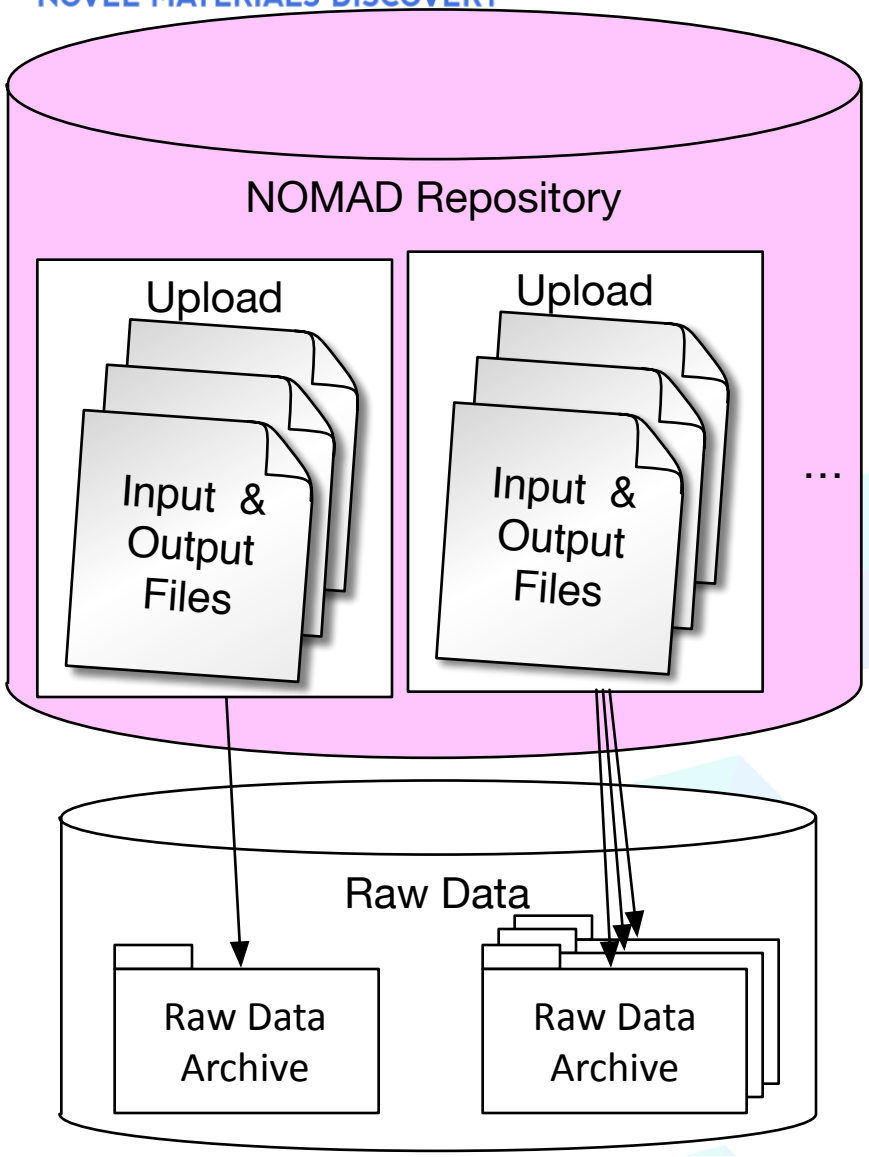
Jungho Shin
HUB

Lorenzo Pardini
HUB

Thomas Zastrow
MPCDF

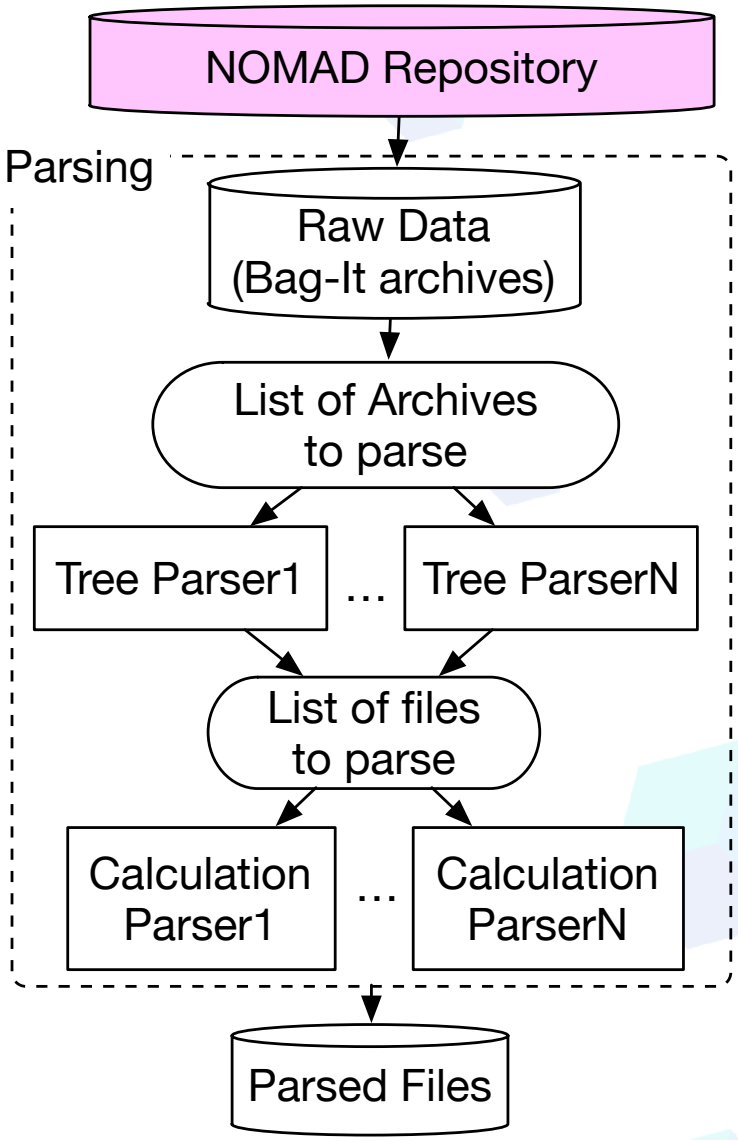
Raw Data

NOVEL MATERIALS DISCOVERY



- *BagIt* format
  - Zip archive
  - Checksums (verifiable integrity)
  - Leading format for digital archiving and exchange
- Limited size (~15GB) by splitting large uploads
- **Identifier** based on the contents (reproducible)
- Good unit for processing
- Open access: http://data.nomad-coe.eu/raw-data

Parsing



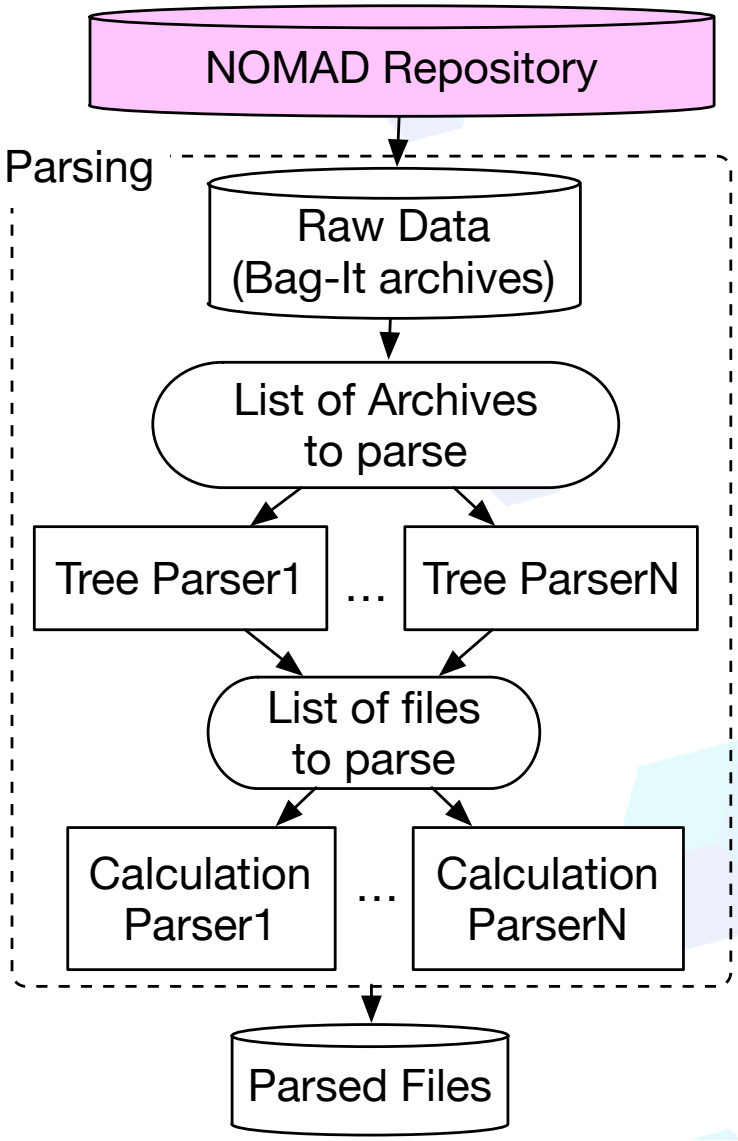
- **Parallel** execution
  - *Tree Parser* identifies the files
  - *Calculation Parser* performs the parsing and generates the parsed files
- Parsing is **pure**: the same version on the same data should give the same result

- Parsers **interpret** all calculation data
- **Organize** it according to the metadata structure
- Data not extracted is invisible
- Writing a parser cannot be automatized and requires a person with *scientific knowledge*

NOMAD

NOVEL MATERIALS DISCOVERY

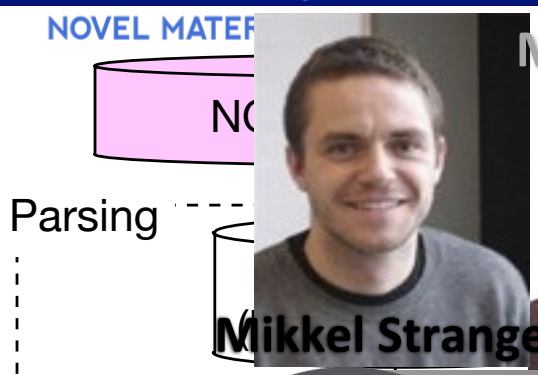
Parsers Ready



- FHI-aims
- VASP
- Turbomole
- qBox
- ORCA
- exciting
- WIEN2k
- ELK
- Gaussian
- SIESTA
- LAMMPS
- CASTEP
- onetep

- DL_POLY
- QUIP/libatoms/GAP
- CP2K
- Crystal
- CPMD
- NWChem
- Octopus
- Quantum Espresso
- Smeagol
- abinit
- GPAW
- Mopac

Parser Developers



Mikkel Strange


Micael Oliviera


Martina Stella

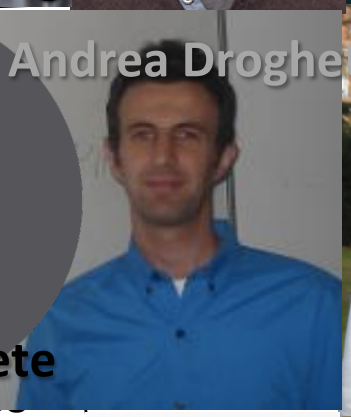

Massimo Riello


Wael Chibiani


Franz Knuth


Aliaksei Mazheika


Adam Fekete

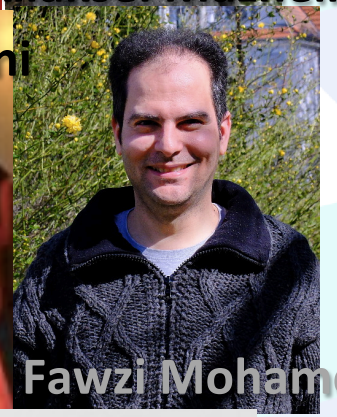

Andrea Droghetti


Honghui Shang


Sebastián Alarcón Villaseca


Daria Tomecka


Lorenzo Pardini


Fawzi Mohamed


Carl Poelking


Adriel Dominiguez


Rosendo Valero


Ask Hjorth Larsen


Lauri Himanen


Sami K. Kivisto


Henning Glawe

NOMAD

NOVEL MATERIALS DISCOVERY

Metadata

- **Metadata** is the conceptual model of our data
- Format independent
- Describes both the data and its structure

NOMAD
NOVEL MATERIALS DISCOVERY

Metadata

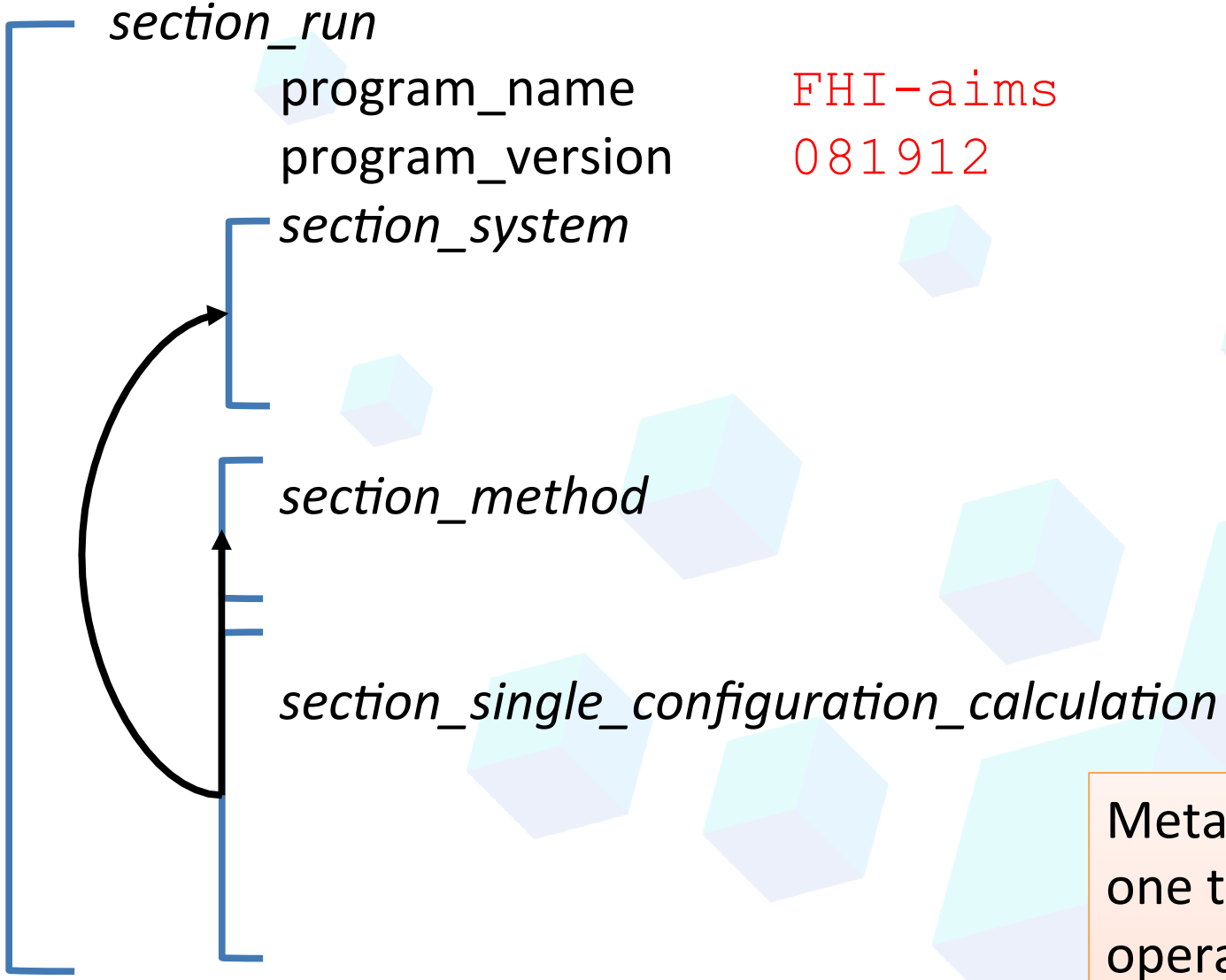
section_run
- program_name `FHI-aims`
- program_version `081912`

Values: **Data**
Structures and names: **Metadata**

Metadata describes the data, and allows one to express queries, and annotate operations performed on the data

NOMAD
NOVEL MATERIALS DISCOVERY

Metadata



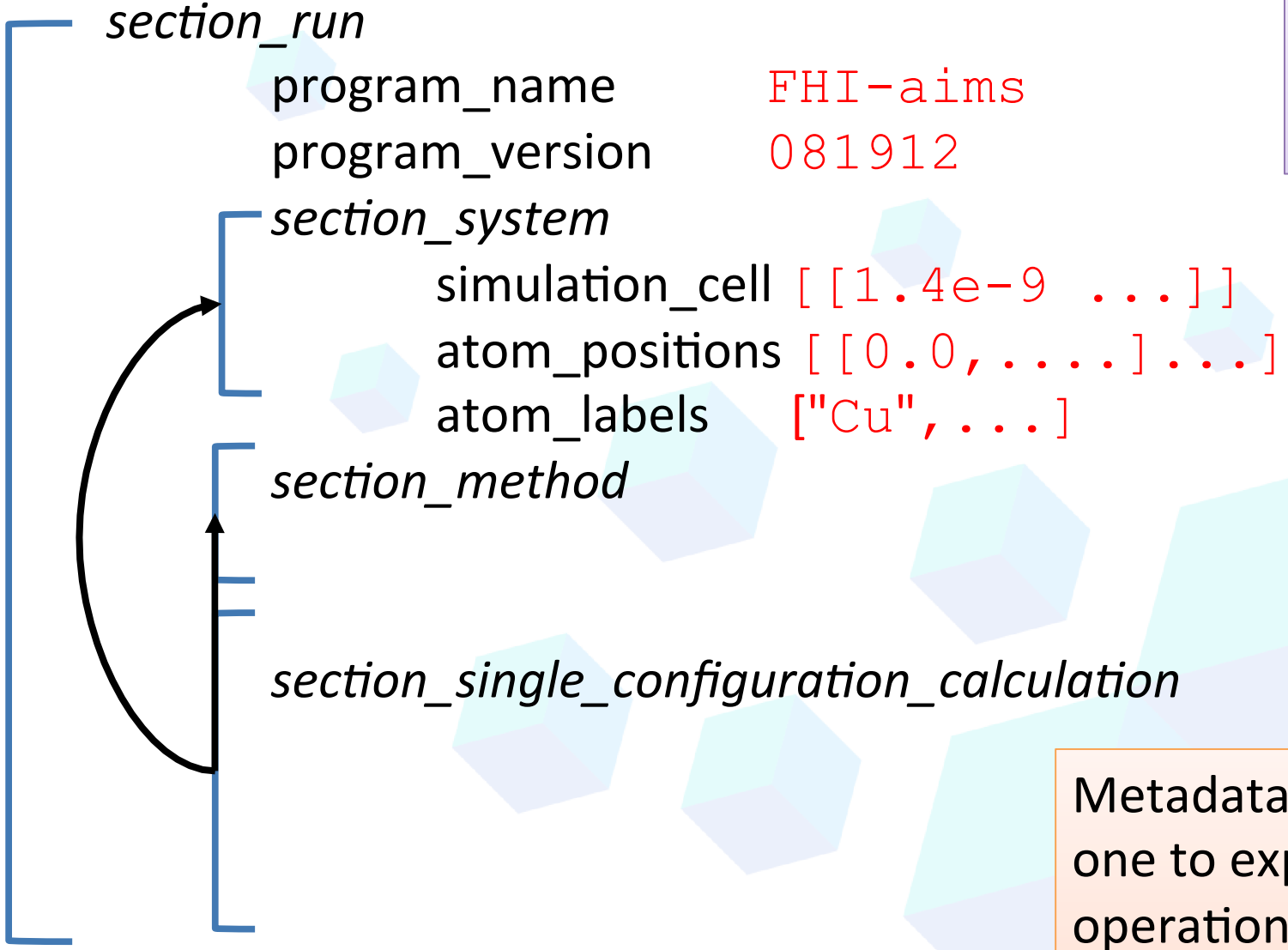
Values: Data
Structures and names: Metadata

Metadata describes the data, and allows one to express queries, and annotate operations performed on the data

NOMAD
NOVEL MATERIALS DISCOVERY

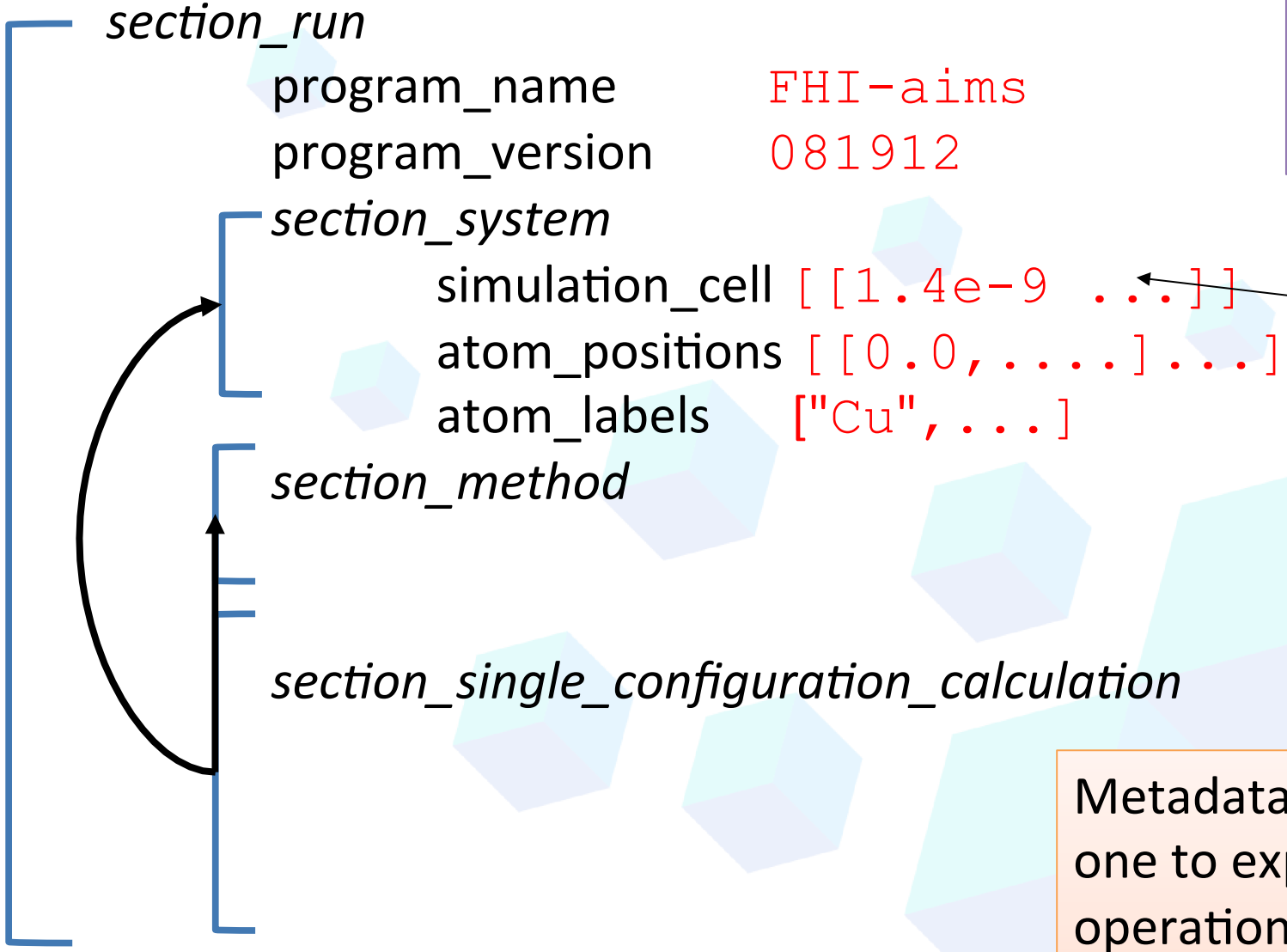
Metadata

Values: Data
Structures and names: Metadata



Metadata describes the data, and allows one to express queries, and annotate operations performed on the data

Metadata



Values: Data
Structures and names: Metadata

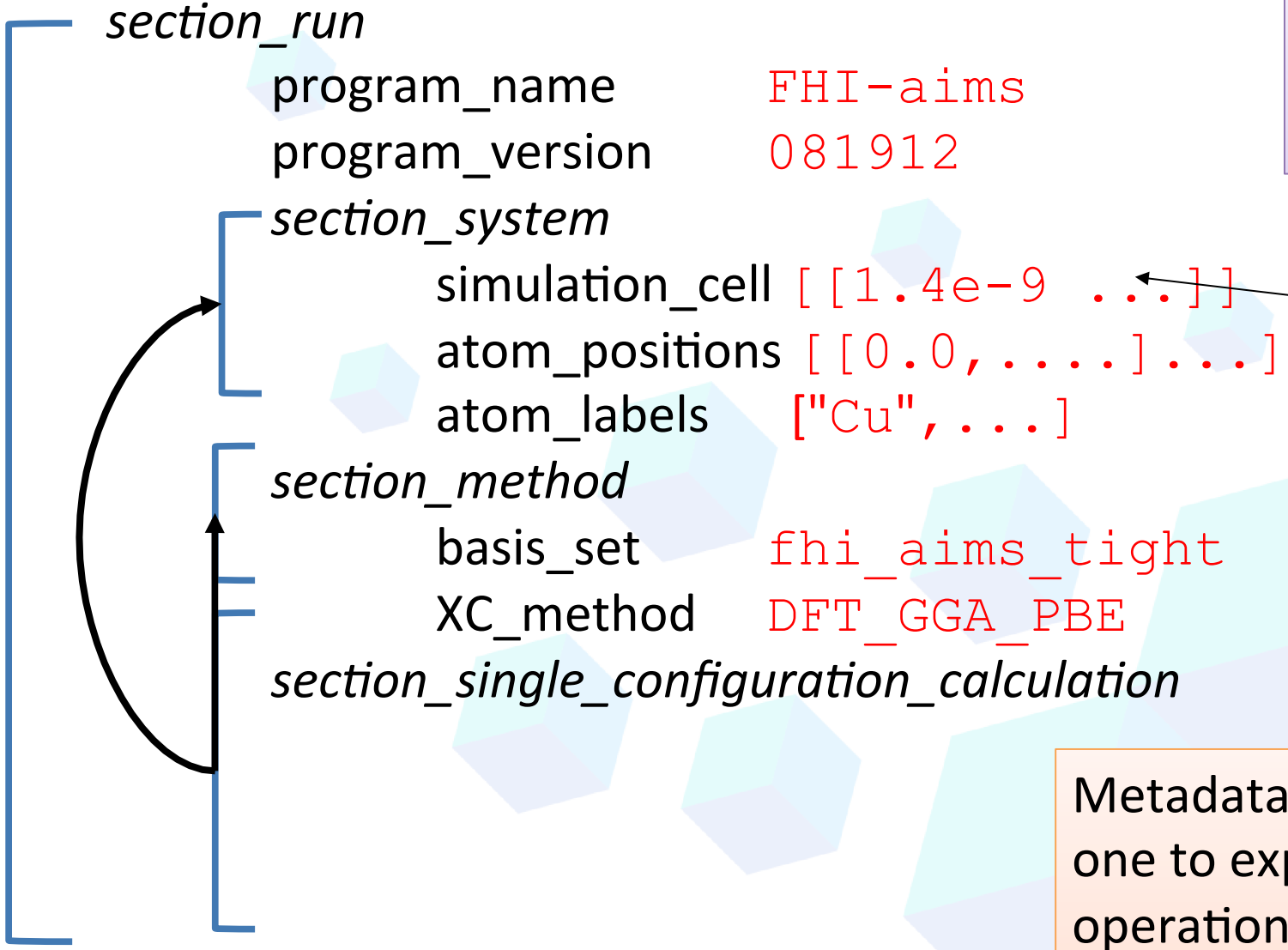
SI Units:
- lengths: m
- energies: J
- …

Metadata describes the data, and allows one to express queries, and annotate operations performed on the data

NOMAD
NOVEL MATERIALS DISCOVERY

Metadata

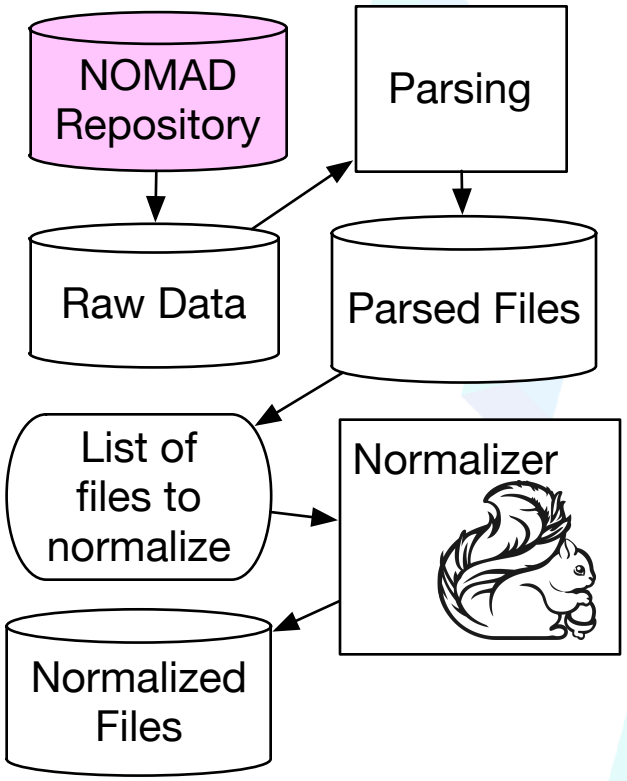
Values: Data
Structures and names: Metadata



SI Units:
- lengths: m
- energies: J
- …

Metadata describes the data, and allows one to express queries, and annotate operations performed on the data

Normalization



Standardization

- Parsers **standardizes** the data, **avoiding** to **loose information**

Normalization

- The varius users (WP2, WP3, WP4,...) may define **derived quantities** (normalized representations,...) that can be generally useful for analysis or visualization
- Normalization is an infrastructure to apply automatically some transformations and store their result along with the parsed data

NOMAD
NOVEL MATERIALS DISCOVERY

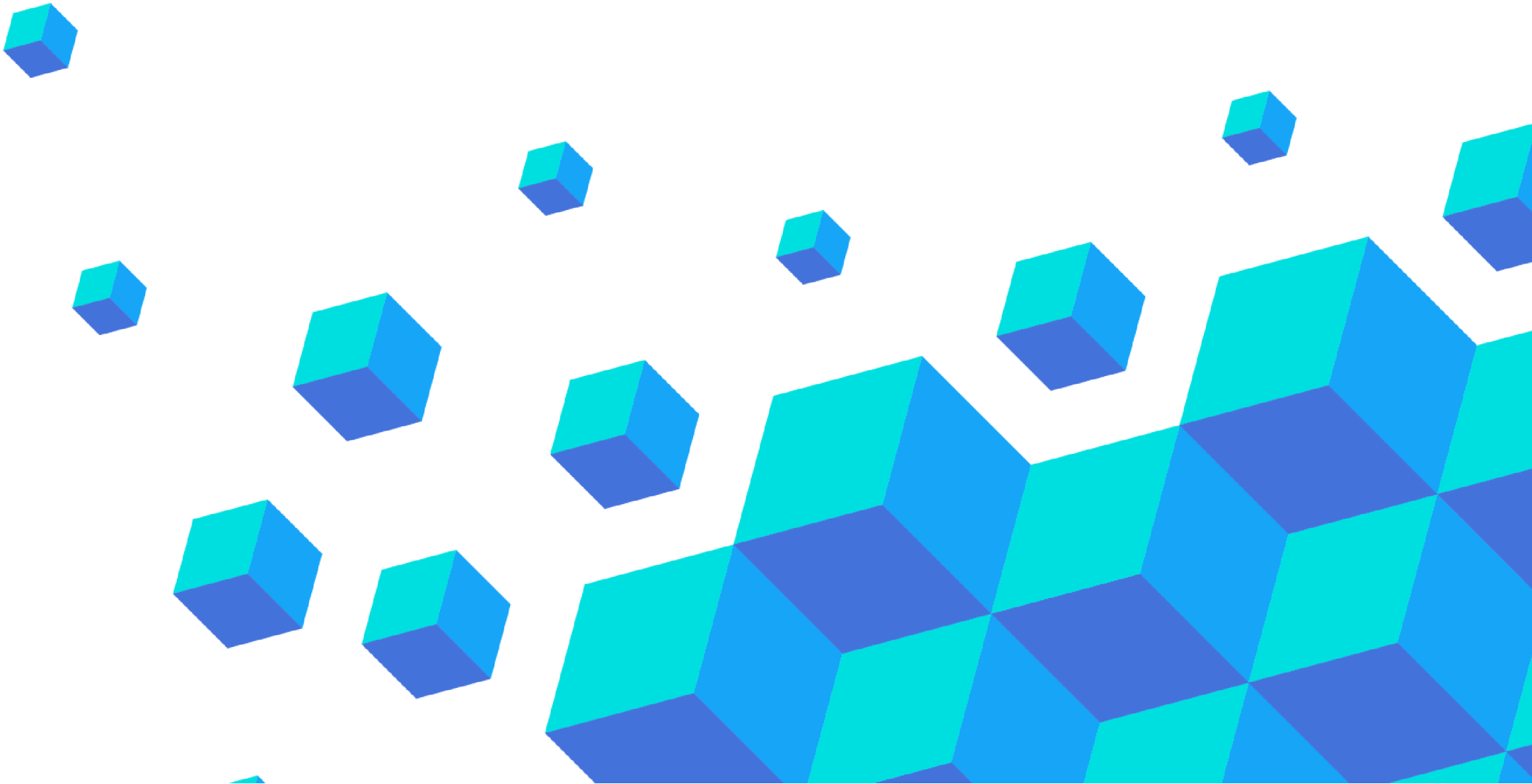
Normalized Files

- Parsed and Normalized files are stored using
  - JSON: human readable, nice to use in the web
  - HDF5: efficient binary representation, indexed access
- RawData identifier used to group calculations
- HDF5 files available on
  - http://data.nomad-coe.eu/normalized

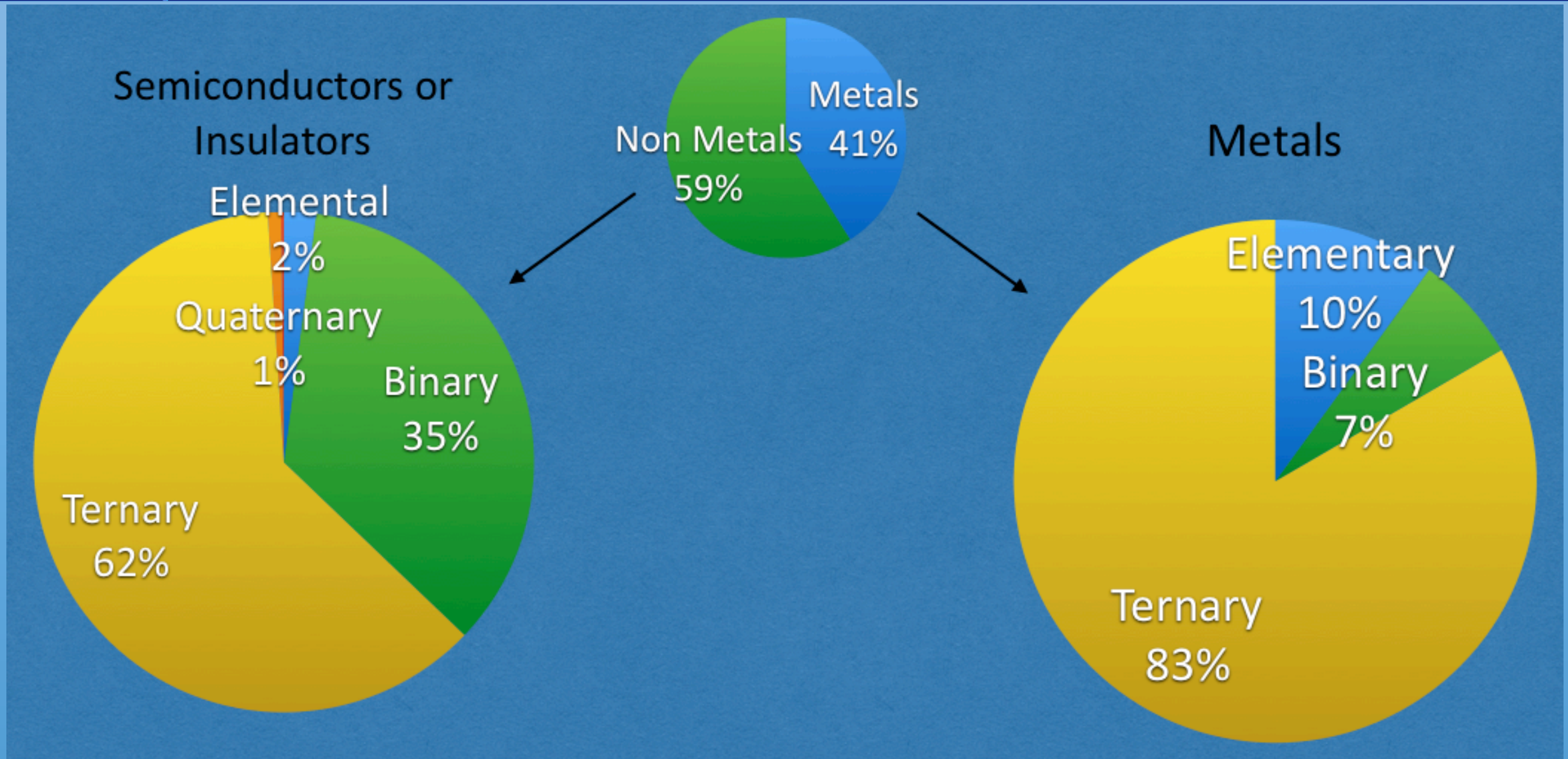
NOMAD

NOVEL MATERIALS DISCOVERY

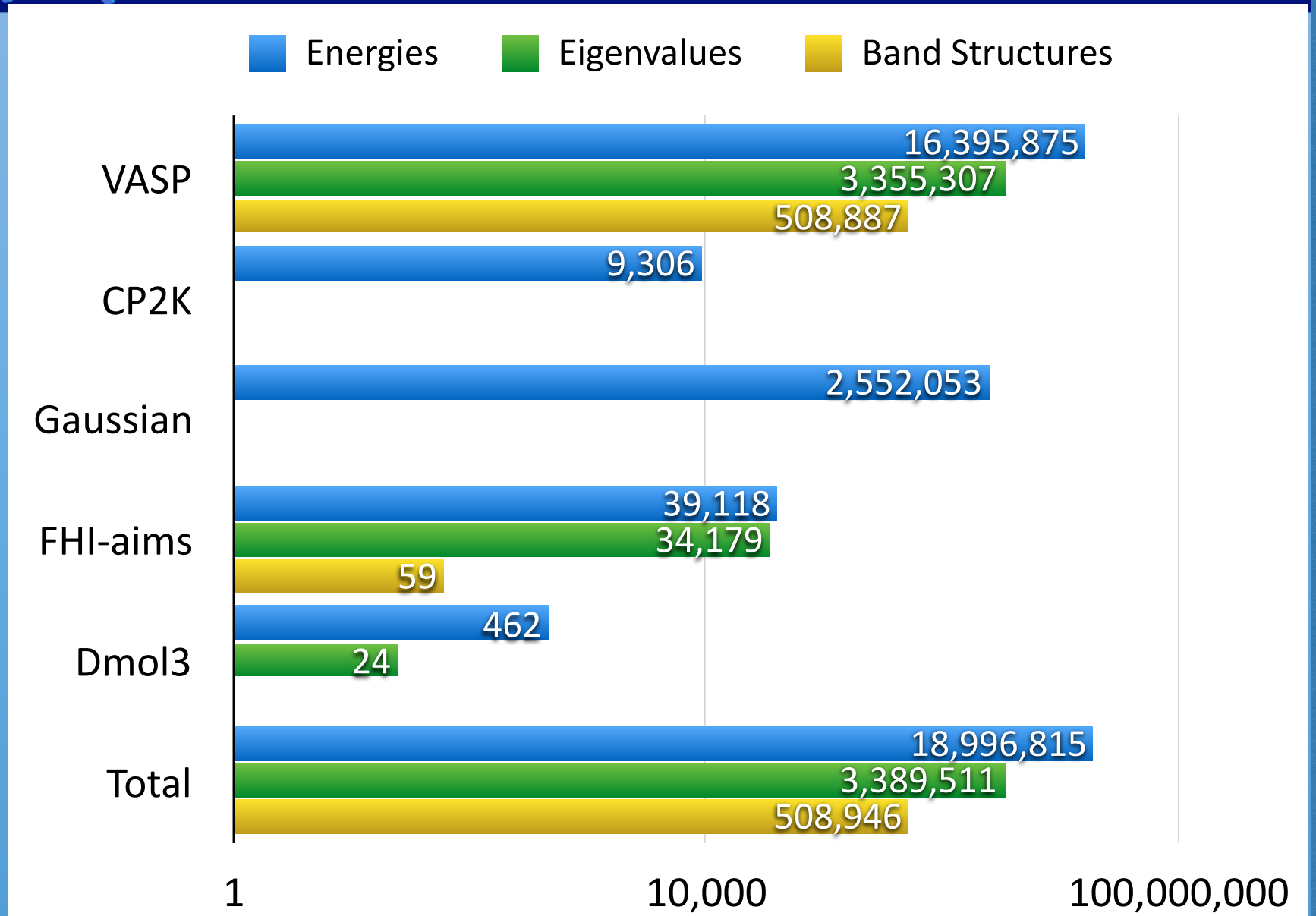
NOMAD Archive: What is in it

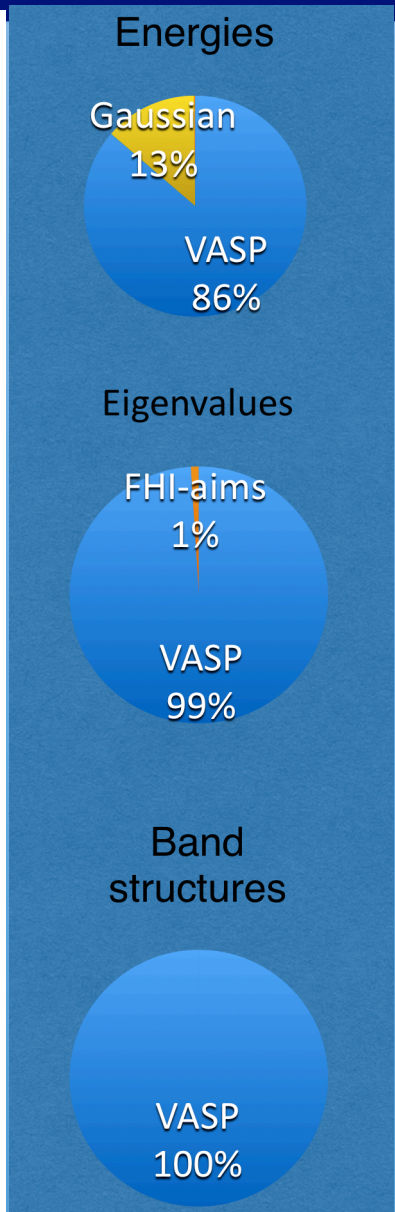


NOMAD Archive: What is in it



NOMAD Archive: What is in it





NOMAD

NOVEL MATERIALS DISCOVERY

NOMAD Archive: Conclusions

- The core development of the NOMAD Archive is done,
  - 25 codes are ready and new ones are in development
  - An extensive metadata description for ab-initio calculations
  - Standard formats to store the data
  - Open access to the archive:
    - http://data.nomad-coe.eu
- Future
  - Parser development to support new code versions and new codes
  - Weekly reparse of data with new parsers
  - Extend Normalization with new transformations used by WP2, WP3 and WP4
  - Improve parser speed
  - Extend metadata to force-field codes