The NOMAD Archive

- Make atomistic simulations available for analysis
- Naming and provenance crucial to allow queries and workflows
  - Clear definition and naming of the source of the data
  - Versioning and reproducibility
  - Clear definition and naming of data itself (metadata)

Summary:

- Scheme of the metadata structure
- Scheme of the Archive
- Statistics of parsing

Metadata Example 1: Single point evaluation

section_run

program_name FHI-aims

program_version 081912

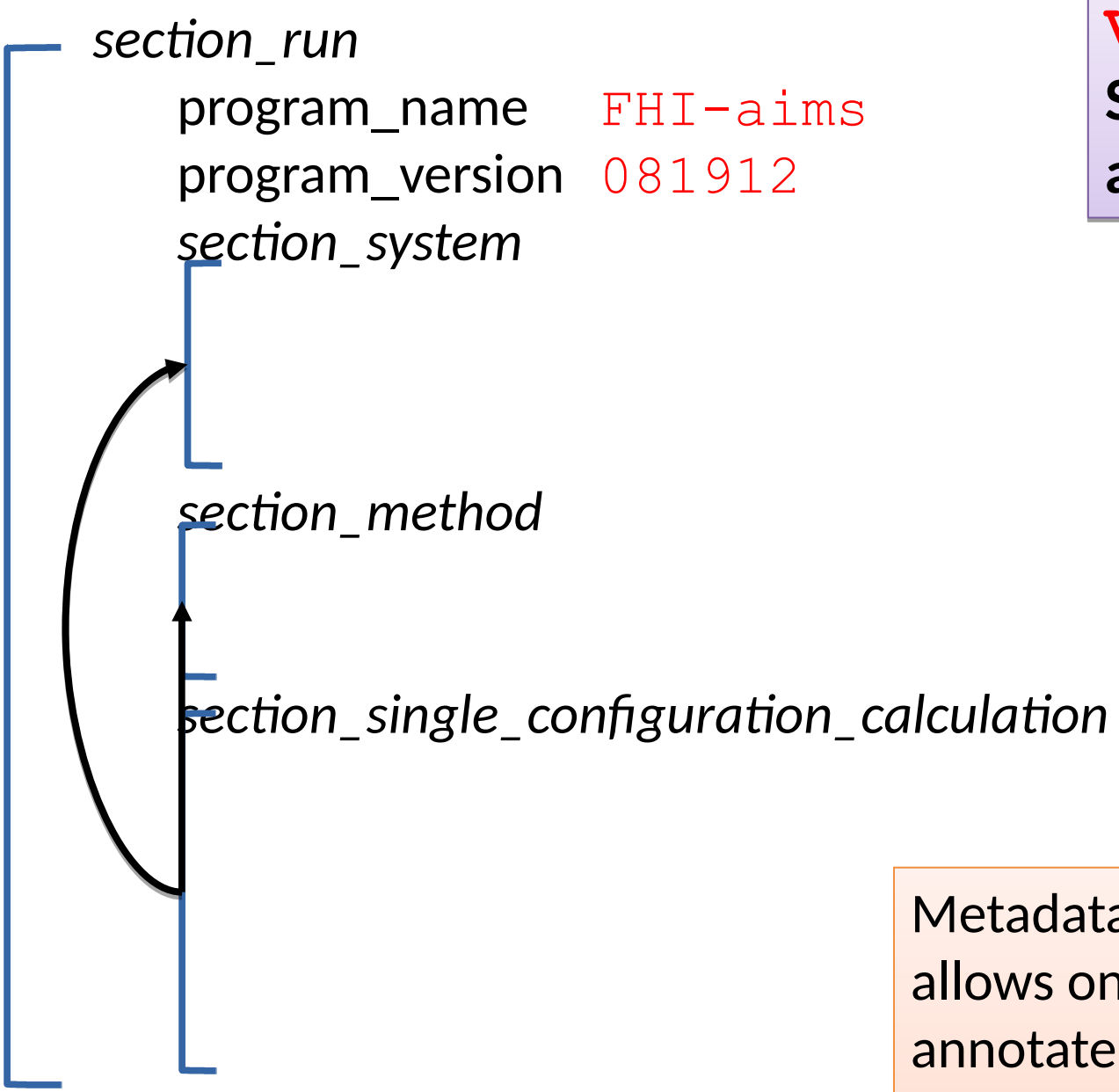
Values: Data

Structures and names: Metadata

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

Metadata Example 1: Single point evaluation

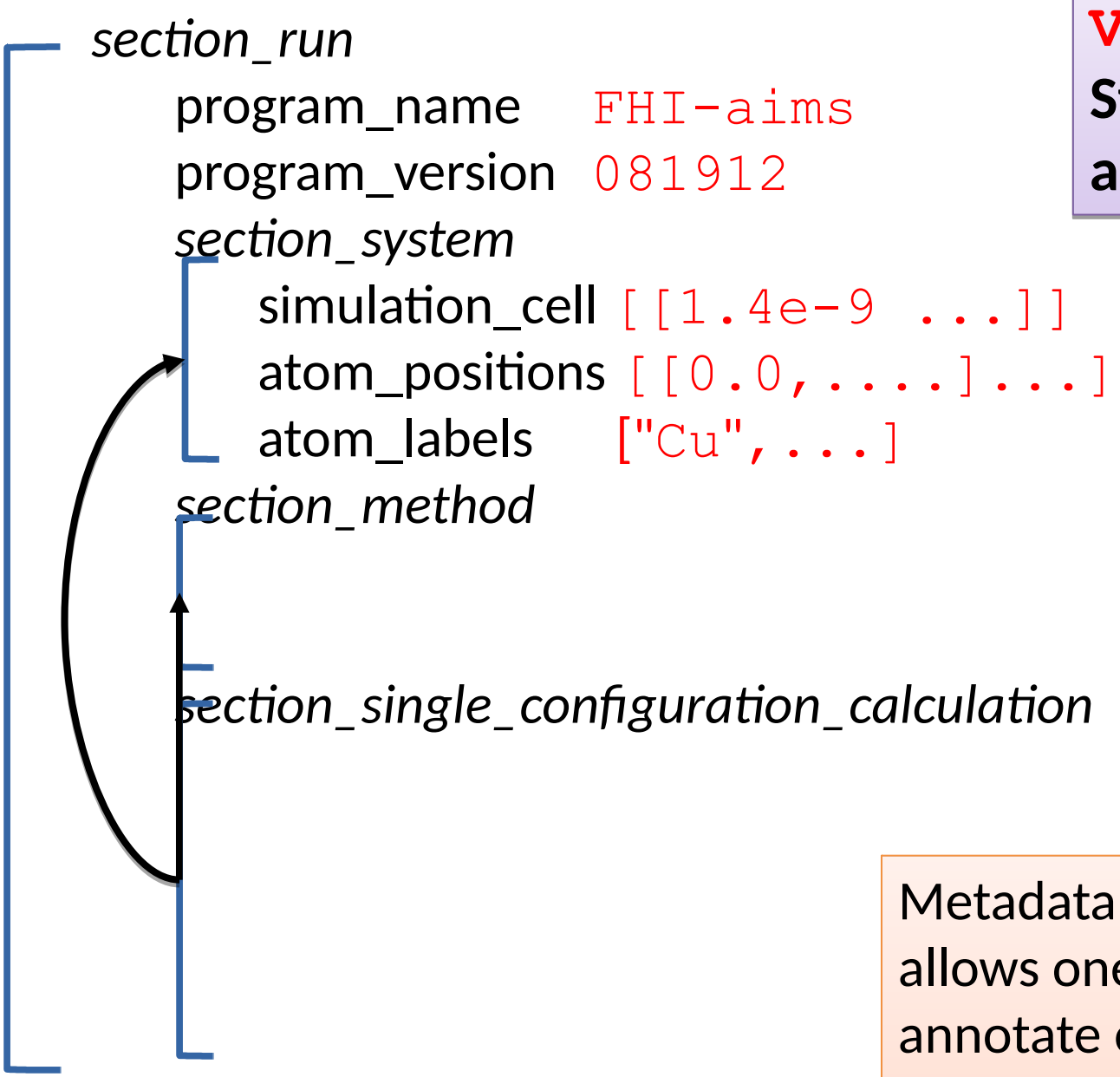
Values: Data
Structures and names: Metadata



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

Metadata Example 1: Single point evaluation

Values: Data
Structures and names: Metadata



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

Metadata Example 1: Single point evaluation

Values: Data
Structures and names: Metadata

section_run
- program_name FHI-aims
- program_version 081912
- *section_system*
  - simulation_cell [[1.4e-9 ...]]
  - atom_positions [[0.0,....]...]
  - atom_labels ["Cu",...]
- *section_method*
- *section_single_configuration_calculation*

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

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

Metadata Example 1: Single point evaluation

Values: Data
Structures and names: Metadata

section_run
- program_name FHI-aims
- program_version 081912
- *section_system*
  - simulation_cell [[1.4e-9 ...]]
  - atom_positions [[0.0,....]...]
  - atom_labels ["Cu",...]
- *section_method*
  - basis_set fhi_aims_tight
  - XC_method DFT_GGA_PBE
- *section_single_configuration_calculation*

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

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

Metadata Example 1: Single point evaluation

Values: Data
Structures
and names: Metadata

section_run
- program_name FHI-aims
- program_version 081912
- *section_system*
  - simulation_cell [[1.4e-9 ...]]
  - atom_positions [[0.0,....]...]
  - atom_labels ["Cu",...]
- *section_method*
  - basis_set fhi_aims_tight
  - XC_method DFT_GGA_PBE
- *section_single_configuration_calculation*
  - energy_total -1.344e-20

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

Metadata Example 1: Single point evaluation

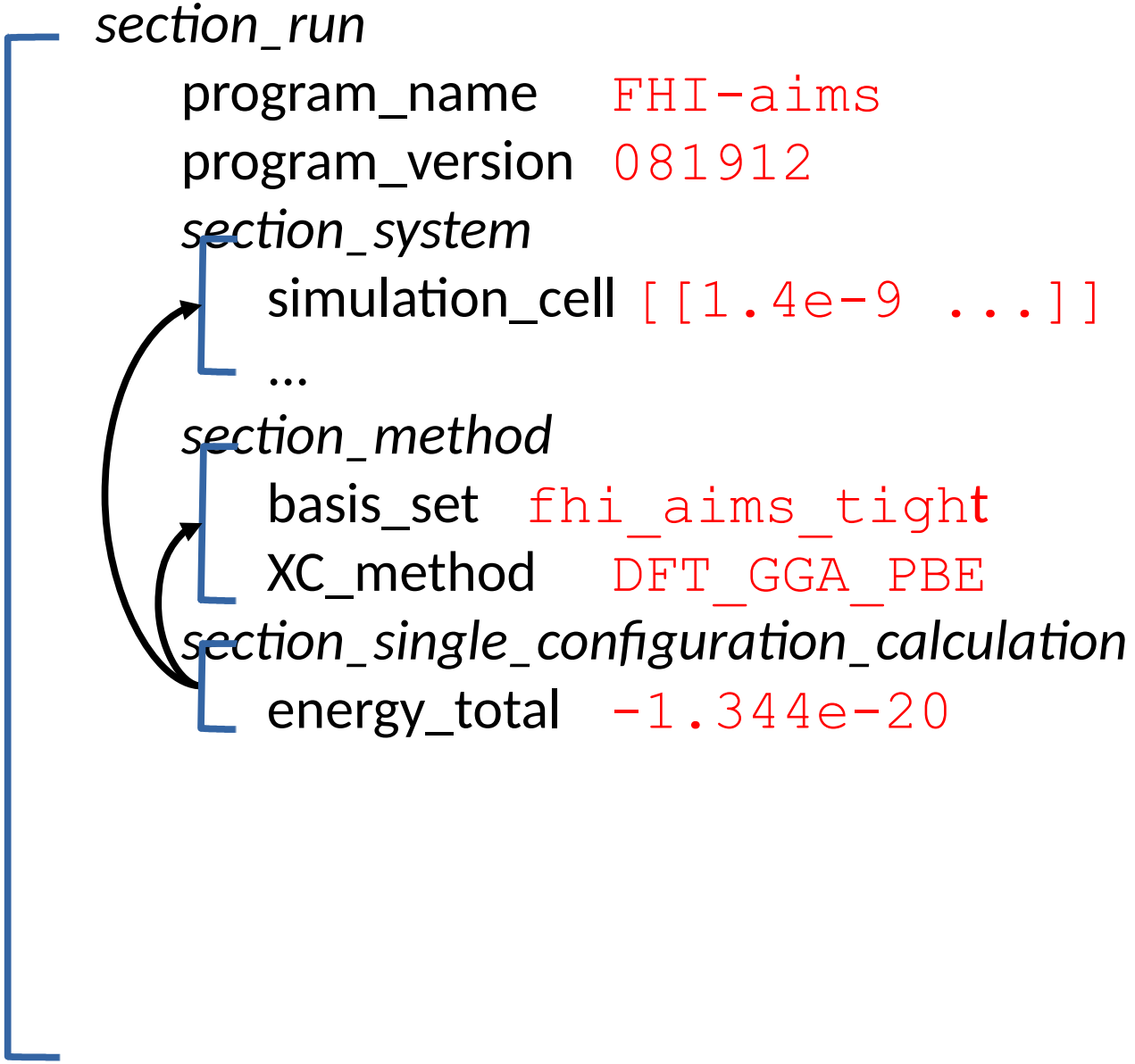
Values: Data
Structures and names: Metadata

section_run
- program_name FHI-aims
- program_version 081912
- *section_system*
  - simulation_cell [[1.4e-9 ...]]
  - atom_positions [[0.0,....]...]
  - atom_labels ["Cu",...]
- *section_method*
  - basis_set fhi_aims_tight
  - XC_method DFT_GGA_PBE
- *section_single_configuration_calculation*
  - *section_scf_iteration*
    - energy_total_scf_iteration -1.326e-20
  - *section_scf_iteration*
    - energy_total_scf_iteration -1.344e-20
- energy_total -1.344e-20

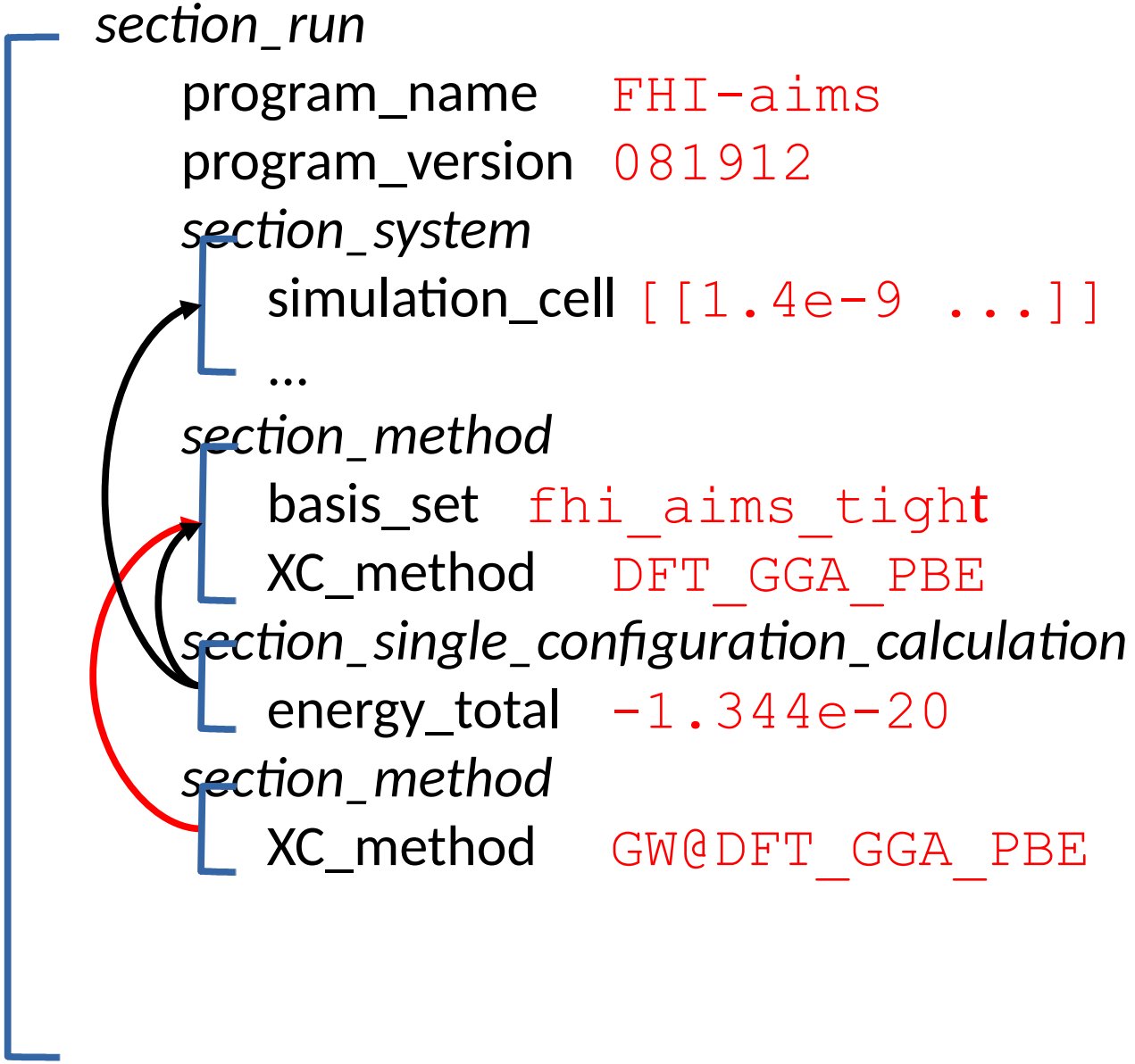
SI Units:
- lengths: m
- energies: J
- …

NOMAD
NOVEL MATERIALS DISCOVERY

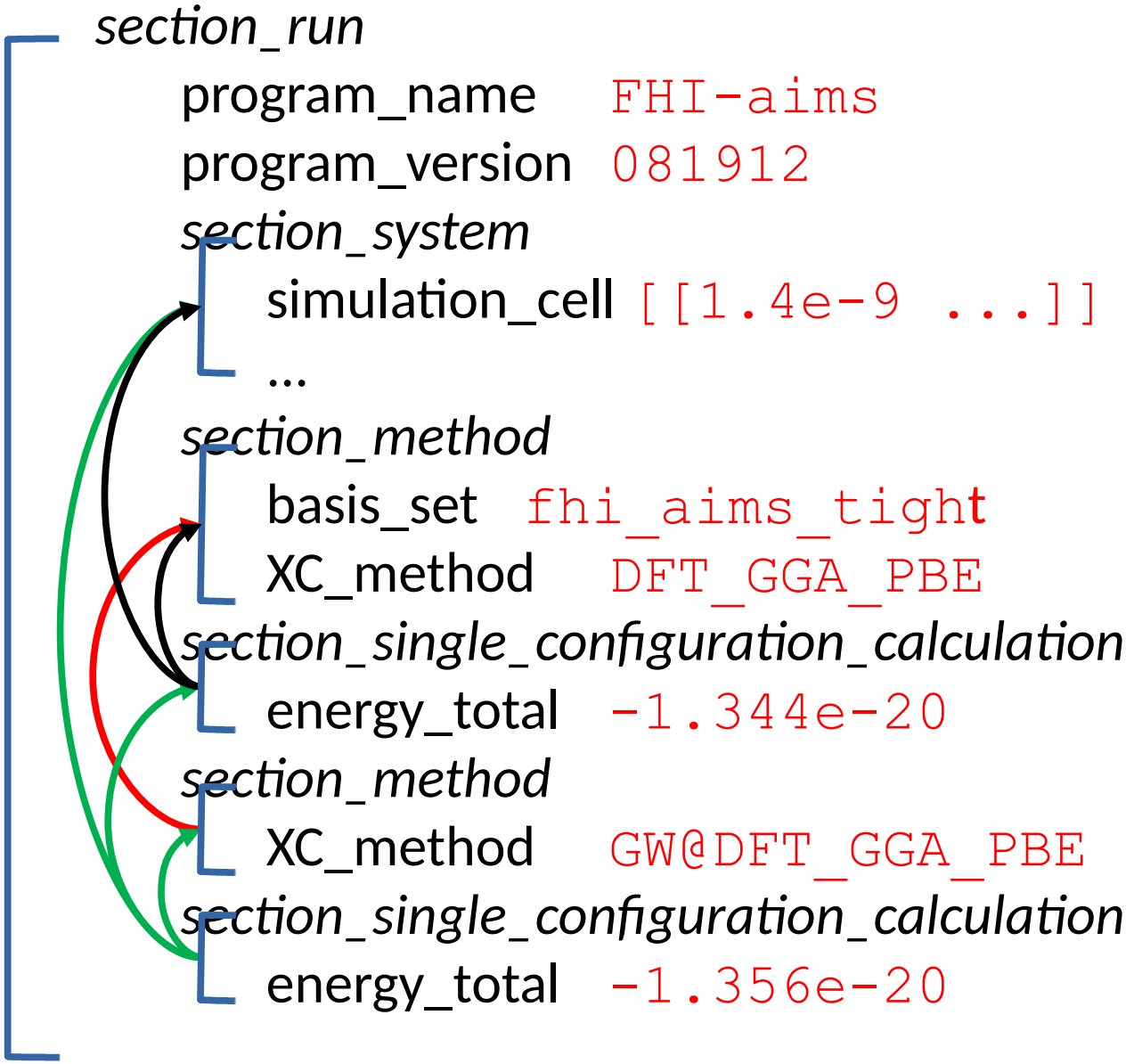
Metadata Example 2: Perturbative calculation



Metadata Example 2: Perturbative calculation

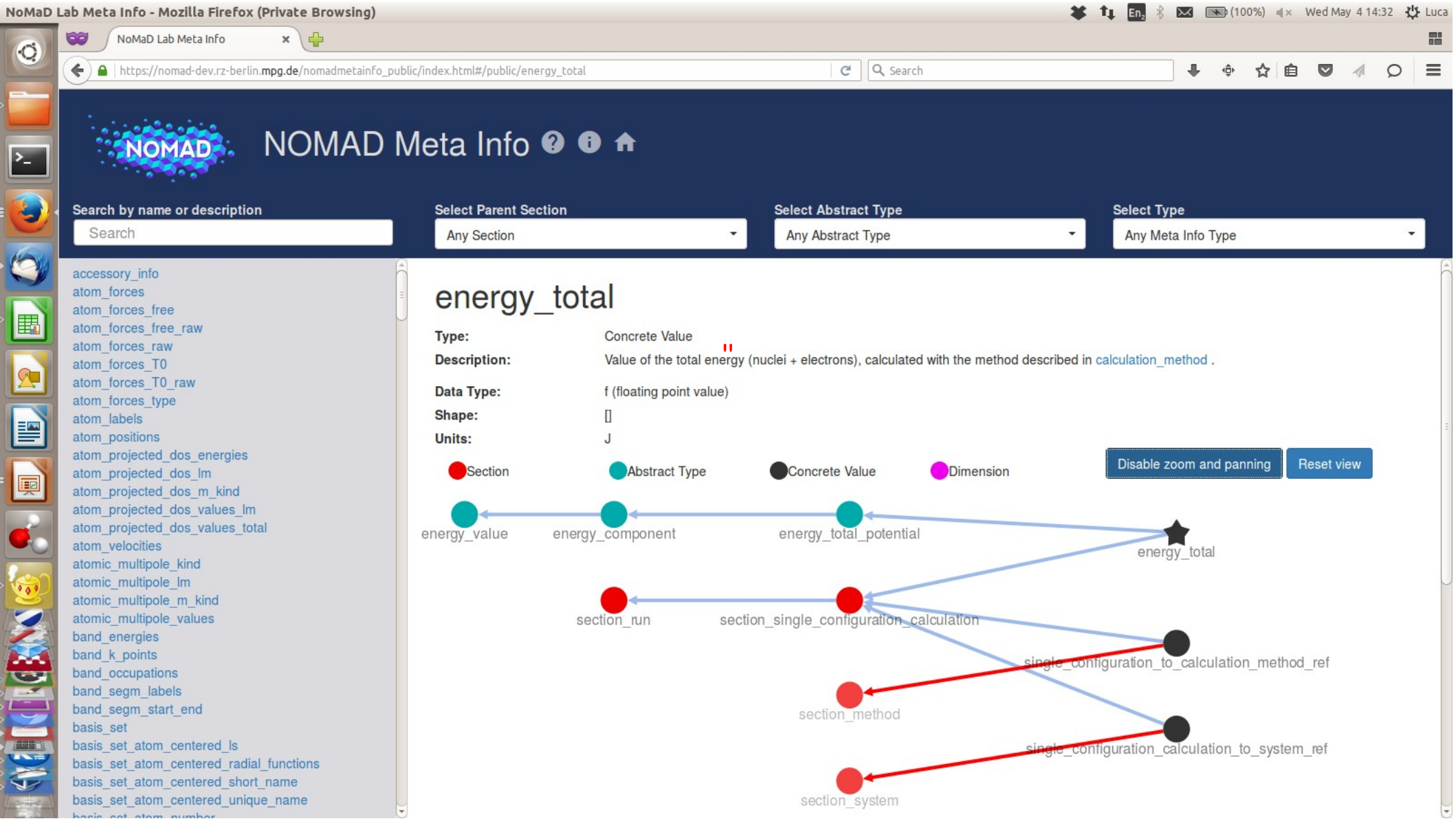


Metadata Example 2: Perturbative calculation

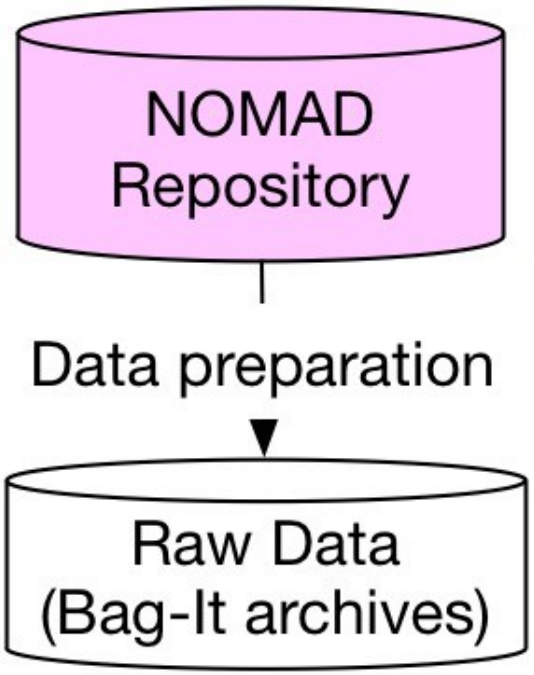


Current code-independent metadata definition:
https://nomad-dev.rz-berlin.mpg.de/nomadmetainfo_public/index.html

Current code-independent metadata definition:
https://nomad-dev.rz-berlin.mpg.de/nomadmetainfo_public/index.html



Structure of the Archive I: Data Preparation



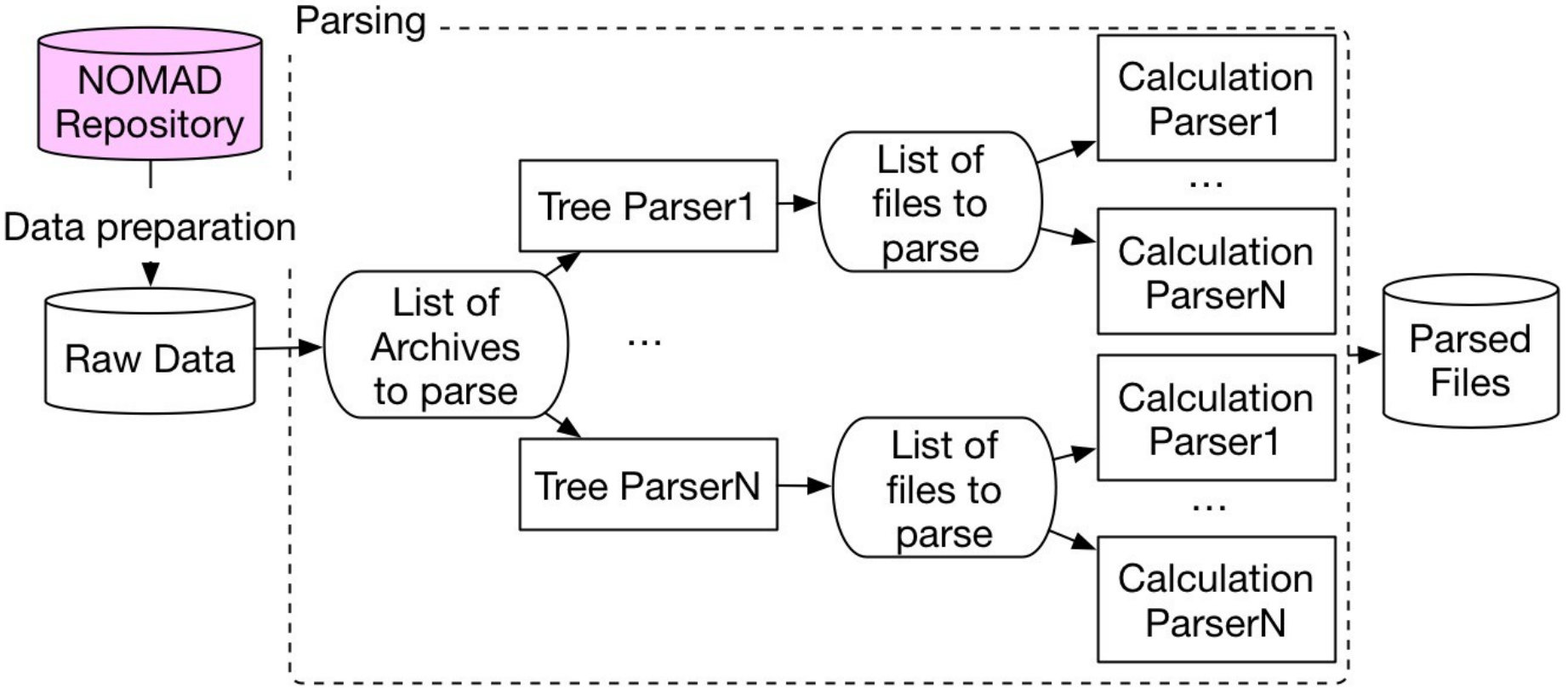
NOMAD Repository:

- >20M files, >3M calculations
- >930 uploads (>250 open access)
- Upload size: 0K-6TB
- Difficult to verify, transfer to other places, know a manageable subset that can be parsed independently

Raw Data archives (on May 13, 2016: 4.8 TB)

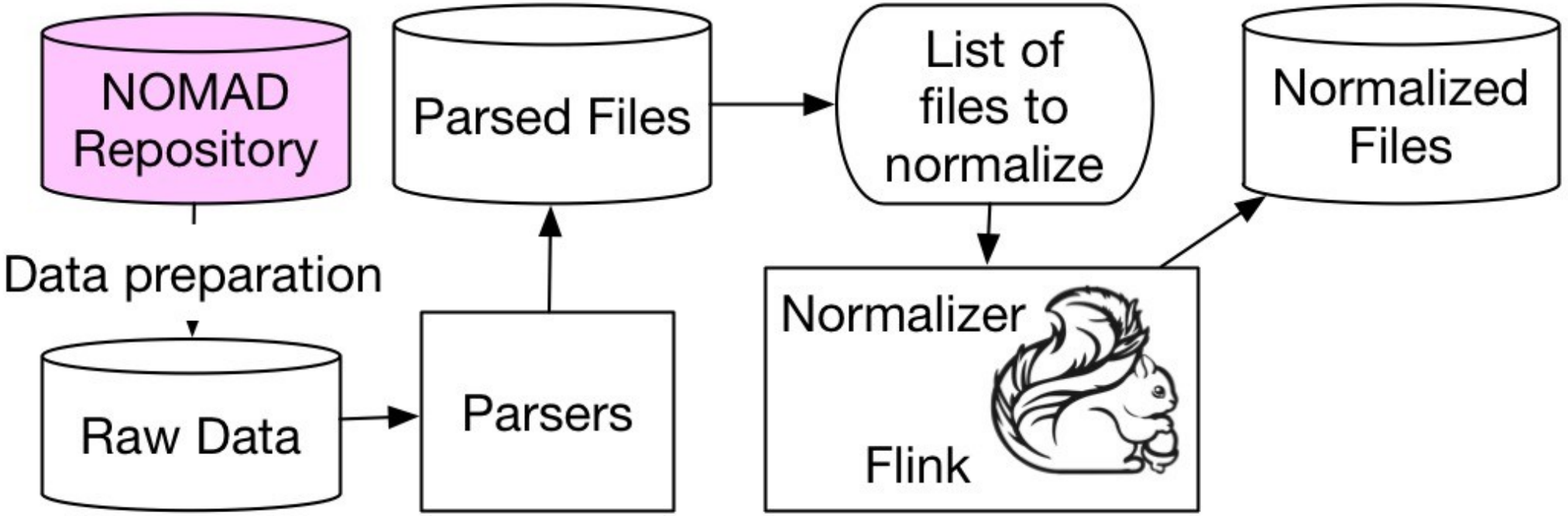
- Manageable size: compressed archives with 10K-29GB uncompressed data each
- 13.5.2016: >860 archives >70% of the open access data packed
- Unique name (only from archive content)
- Names uniformly distributed (can be used for cheap work distributing)
- Every archive can be verified independently (Bag-It format)
- Well defined group of file to parse together

Structure of the Archive II: Parsing



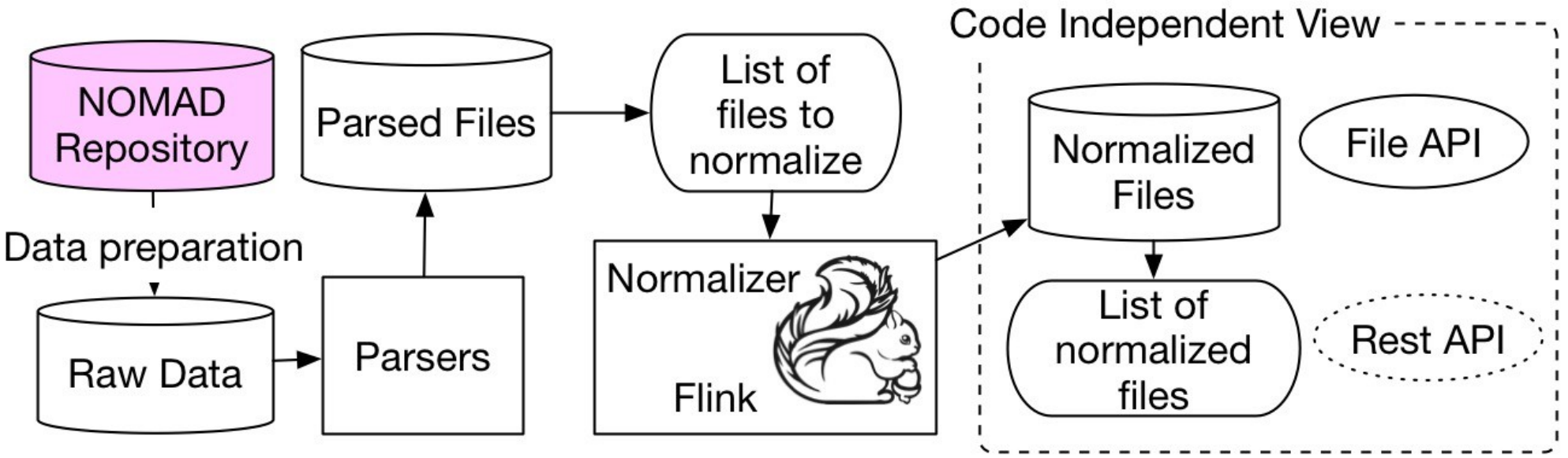
- Two run modes: containers or cluster, they can be combined (flexible parallelization)
- Raw data archive + parser version → reproducible results
- Well structured repositories, automatic testing → support of many parser developers

Structure of the Archive III : Normalization



- Parsed files and Normalized files are json + (soon) HDF5 files that use the meta info
- Normalization performs common transformations and combined results of the same raw data archive.
- Examples of normalization (currently implemented): type of exchange-correlation treatment, material classification (bulk, molecule/cluster, surface)

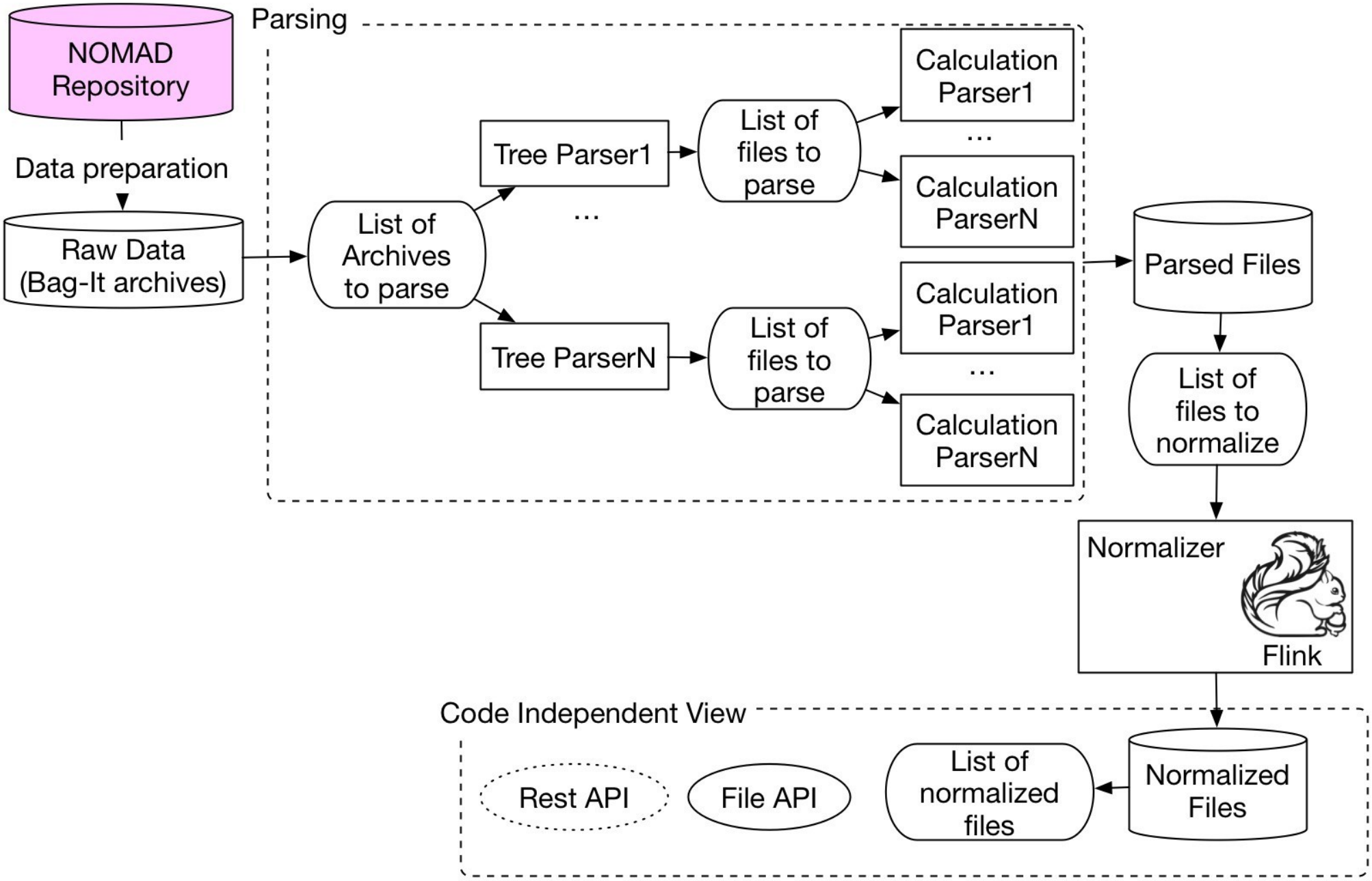
Structure of the Archive IV : External View



Toward the rest of the NOMAD Laboratory (i.e., the Encyclopedia and the Data Analytics) the archive is visible:

- as a set of files updated atomically (the normalized and parsed files)
- with a REST API (currently accessing mostly the metadata), but to be extended

Structure of the Archive : Overview



Statistics of parsing

(on May 13, 2016)

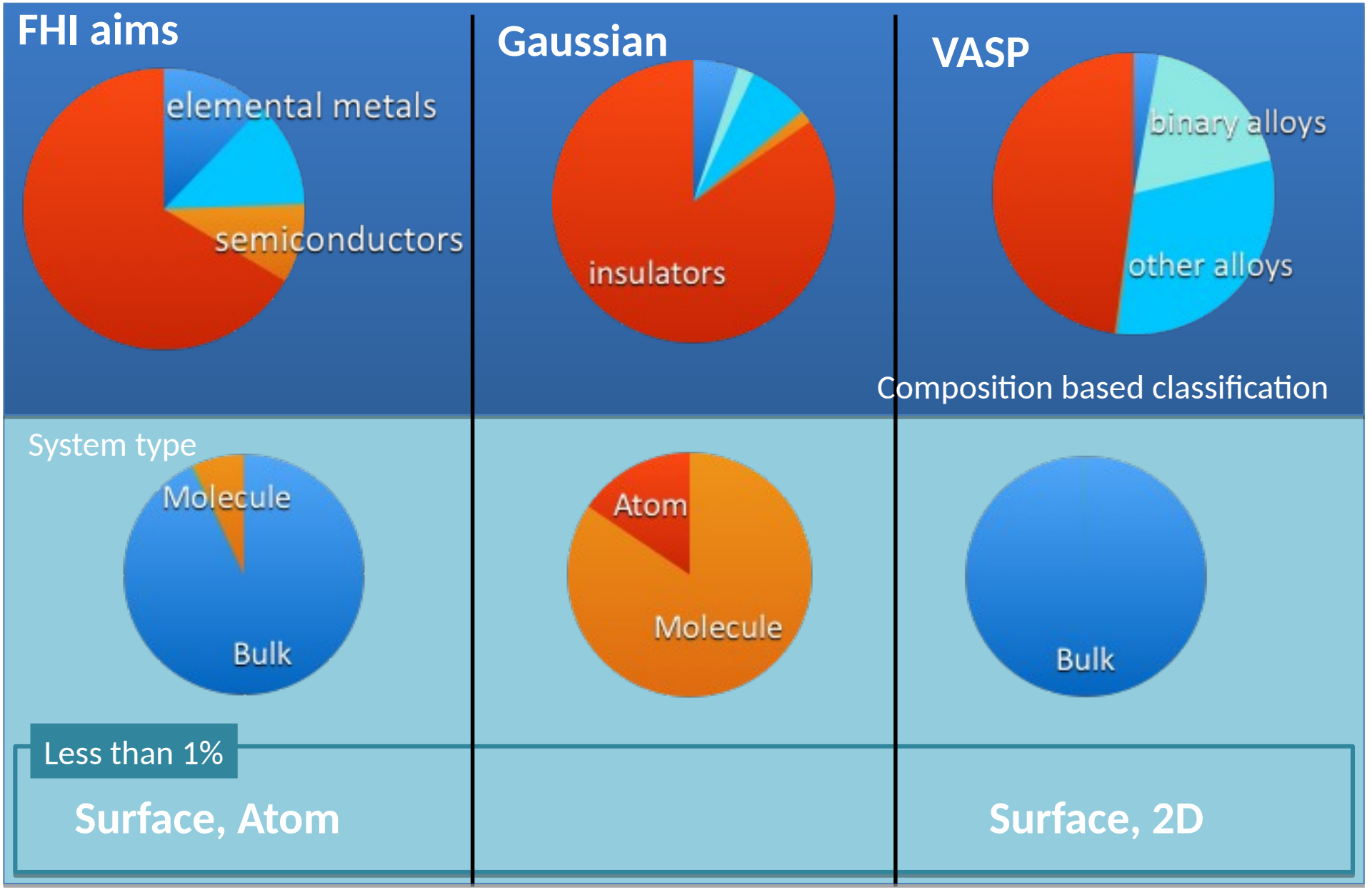
- 17 parsers in active development
- >1M successfully parsed output files
- 660 raw data achives >55% of the total open access data
- >10M single point energy evaluations
- >100k different material compositions
- >60k band structures
- >1.3M eigenvalues

Details of parsers with most data

(on May 13, 2016)

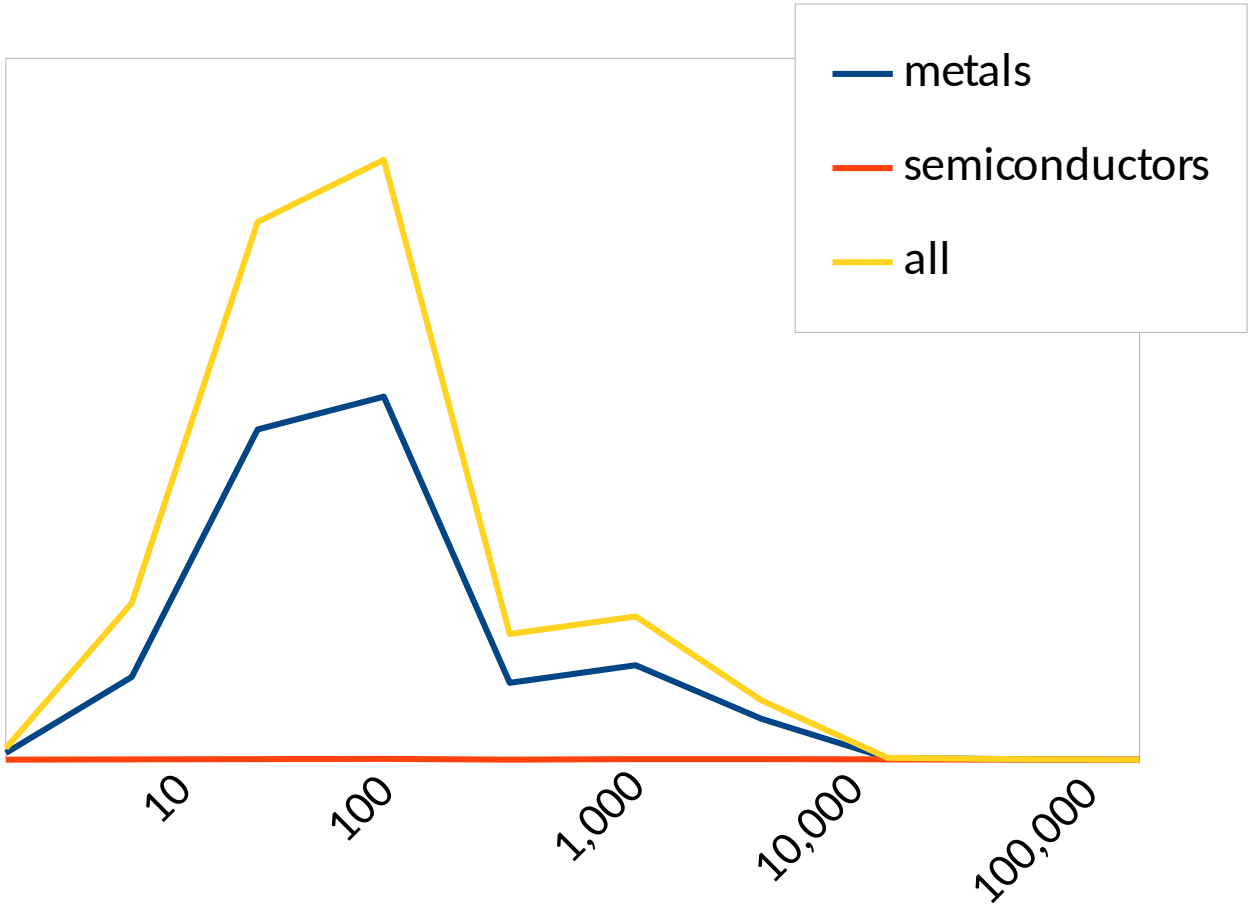
Details of parsers with most data

(on May 13, 2016)



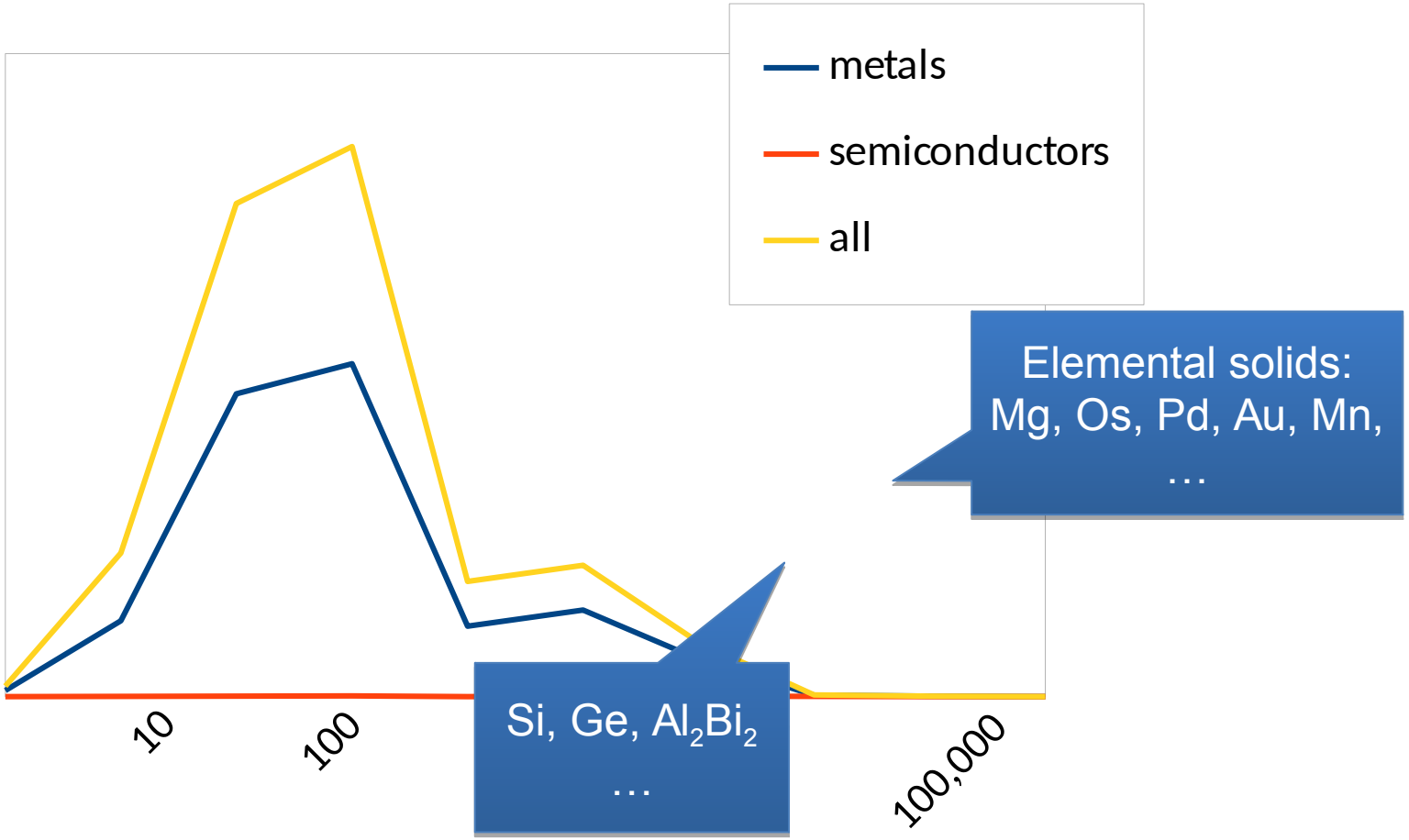
NOMAD
NOVEL MATERIALS DISCOVERY

VASP: How much was every different composition explored



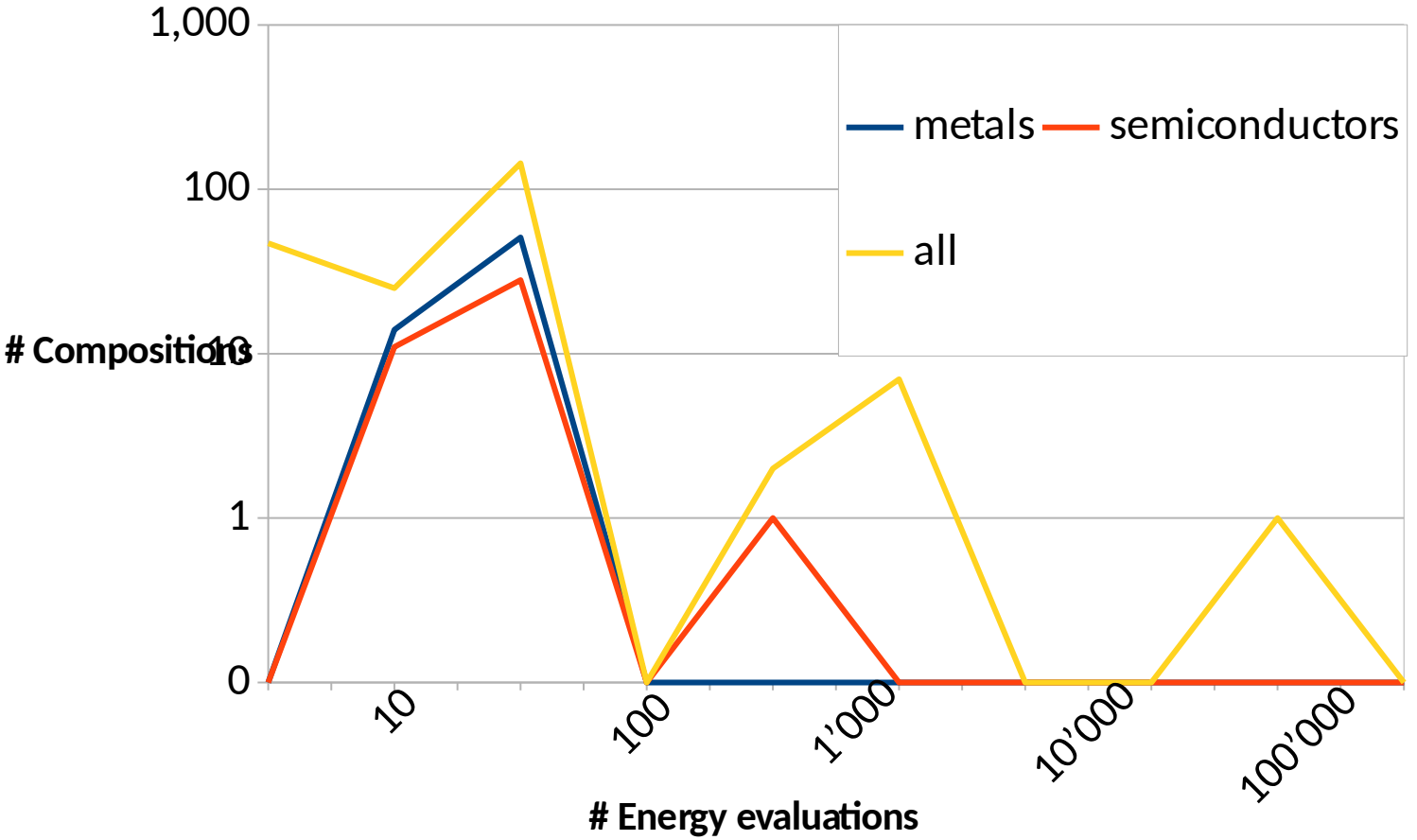
NOMAD
NOVEL MATERIALS DISCOVERY

VASP: How much was every different composition explored



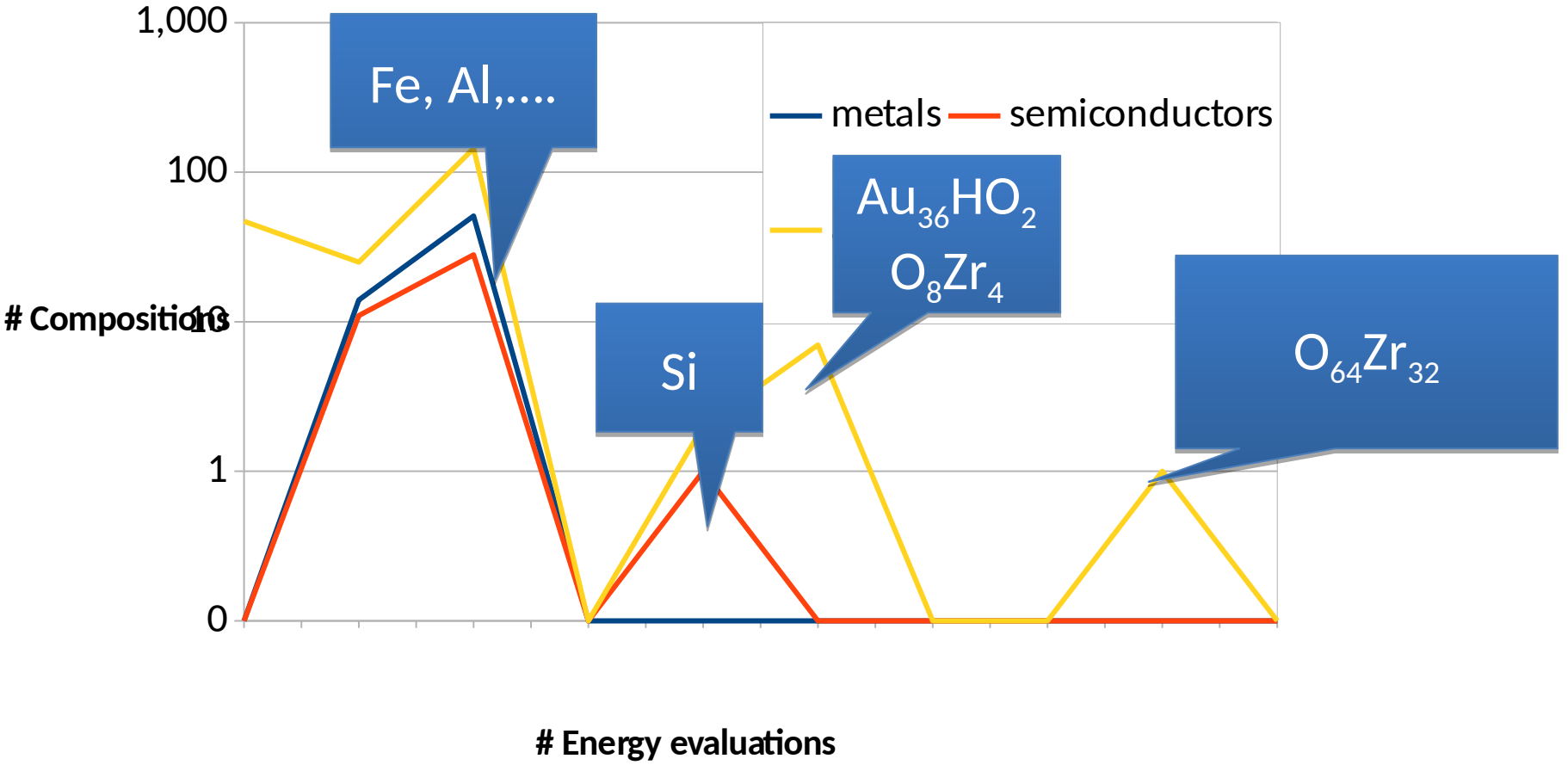
NOMAD
NOVEL MATERIALS DISCOVERY

FHI aims: How much was every different composition explored



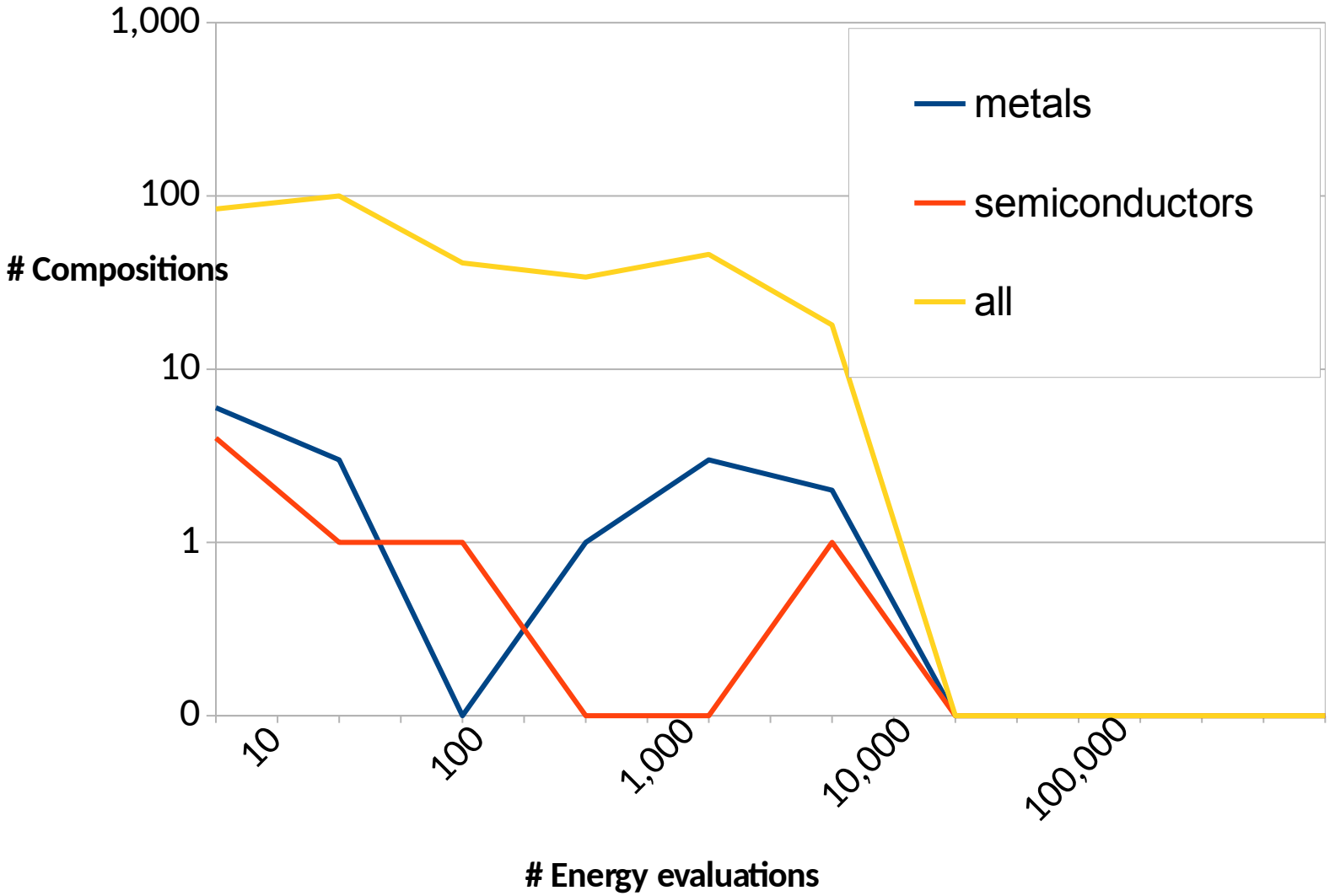
NOMAD
NOVEL MATERIALS DISCOVERY

FHI aims: How much was every different composition explored



NOMAD
NOVEL MATERIALS DISCOVERY

Gaussian: How much was every different composition explored



NOMAD
NOVEL MATERIALS DISCOVERY

Gaussian: How much was every different composition explored