# The NOMAD Laboratory

EUSpec Meeting ESCDF further steps
Budapest 18 Oct. 2016

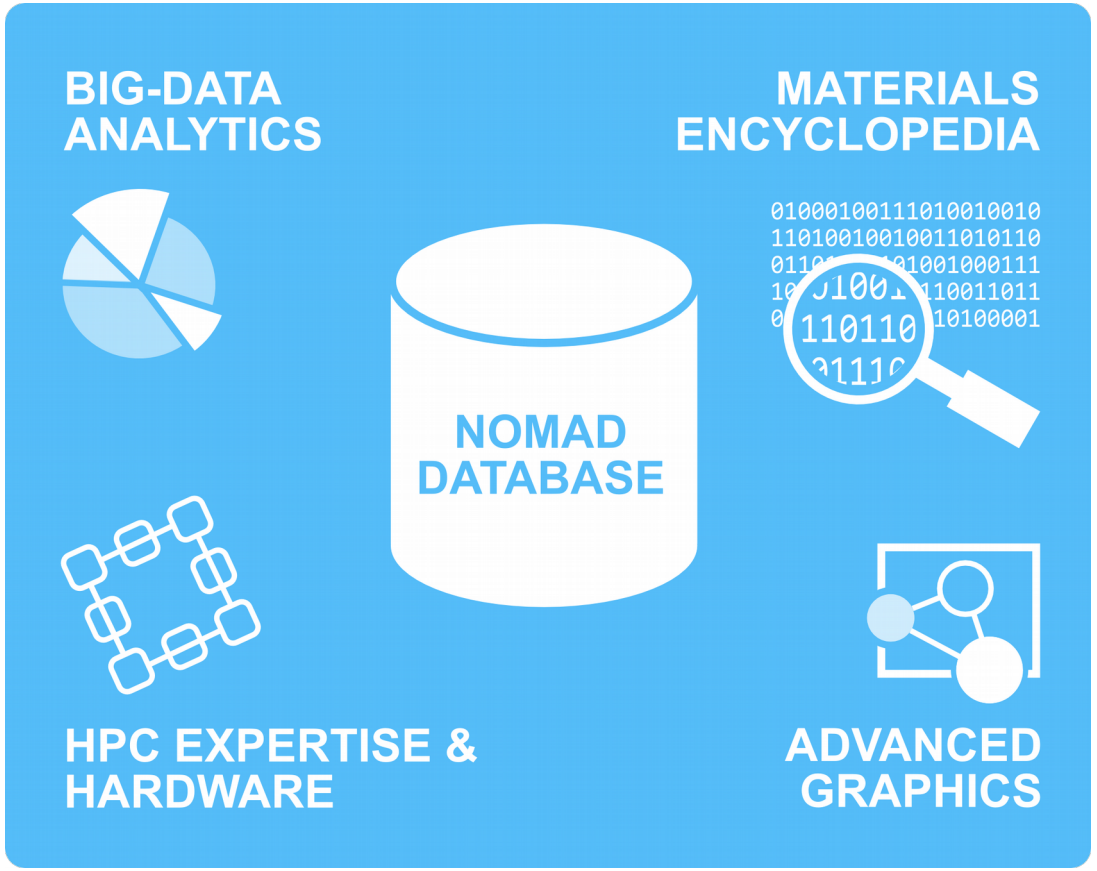
Fawzi Mohamed FHI Berlin

NOMAD

NOVEL MATERIALS DISCOVERY

**The Novel Materials Discovery (NOMAD) Laboratory** develops a Materials Encyclopedia and Big-Data Analytics and Advanced Graphics Tools for materials science and engineering.

Eight complementary computational materials science groups and four high-performance computing centers form the core of this Centre of Excellence.



**Matthias Scheffler,** FHI MPS, Berlin

**Kristian Thygesen** Tech. U., Lyngby

**Ciaran Clissman** pintail Ltd. Dublin

**Arndt Bode** Leibniz Comp. Ctr.

**Jose Maria Cela,**BSC, Barcelona

**Alessandro De Vita** King's Col. London

**Angel Rubio** MPI MPSD, Hamburg

**Claudia Draxl** Humboldt U, Berlin

NOMAD

NOVEL MATERIALS DISCOVERY

**Risto Nieminen** Aalto U. Helsinki

**Kimmo Koski** CSC – IT Center Helsinki

**Francesc Illas** U. of Barcelona

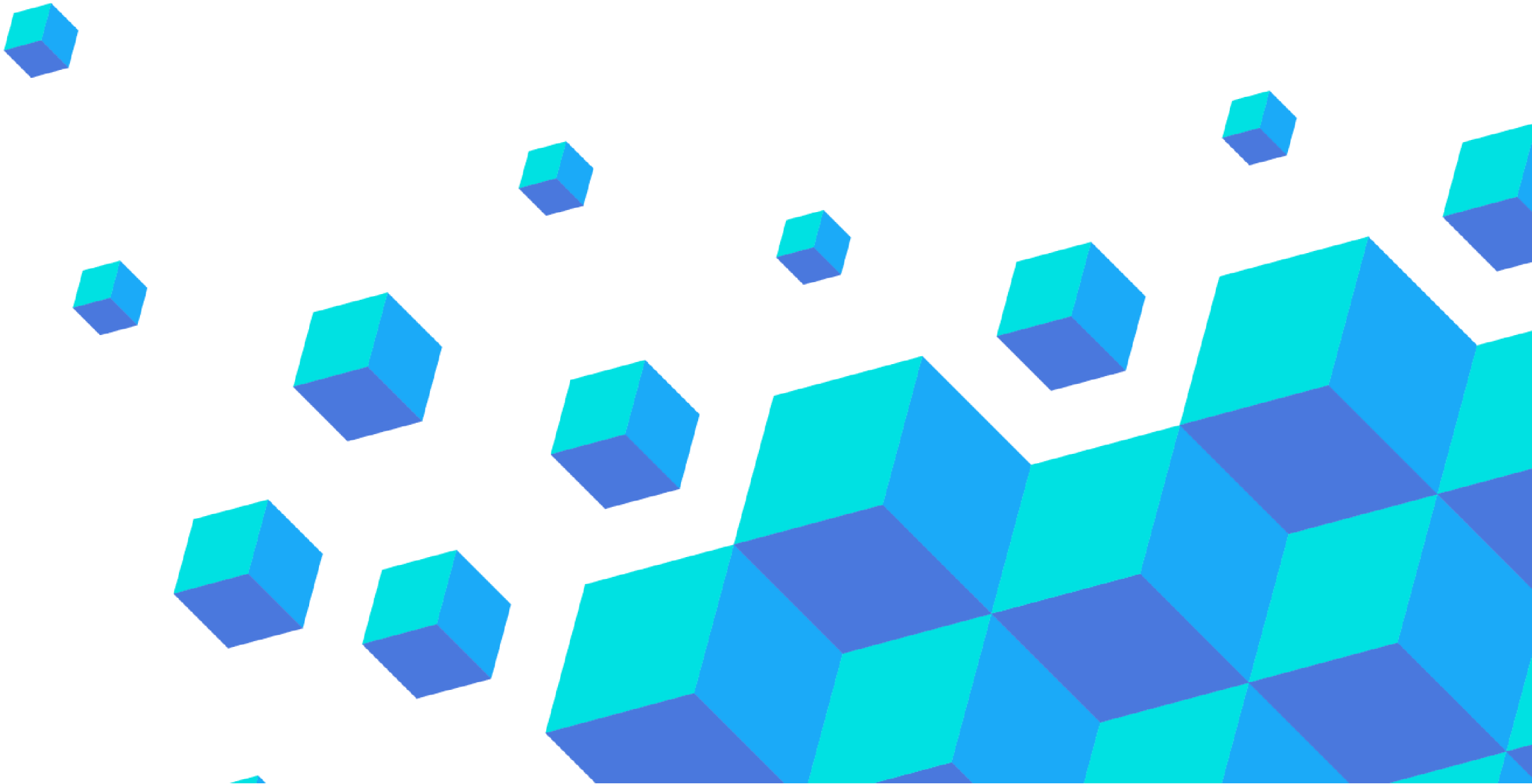
**Stefan Heinzel** MPS Comp. & Data, Garching

**Daan Frenkel** U. Cambridge

NOMAD

NOVEL MATERIALS DISCOVERY

# NOMAD Archive: Goal



# NOMAD Archive: Goal

NOVEL MATERIALS DISCOVERY

NOMAD REPOSITORY

NOMAD Archive

Make data of the simulation codes used in the community accessible

*WP2*: NOMAD ENCYCLOPEDIA

*WP3*: NOMAD *VISUALIZATION*

*WP4*: BIG DATA ANALYSIS

*External users*

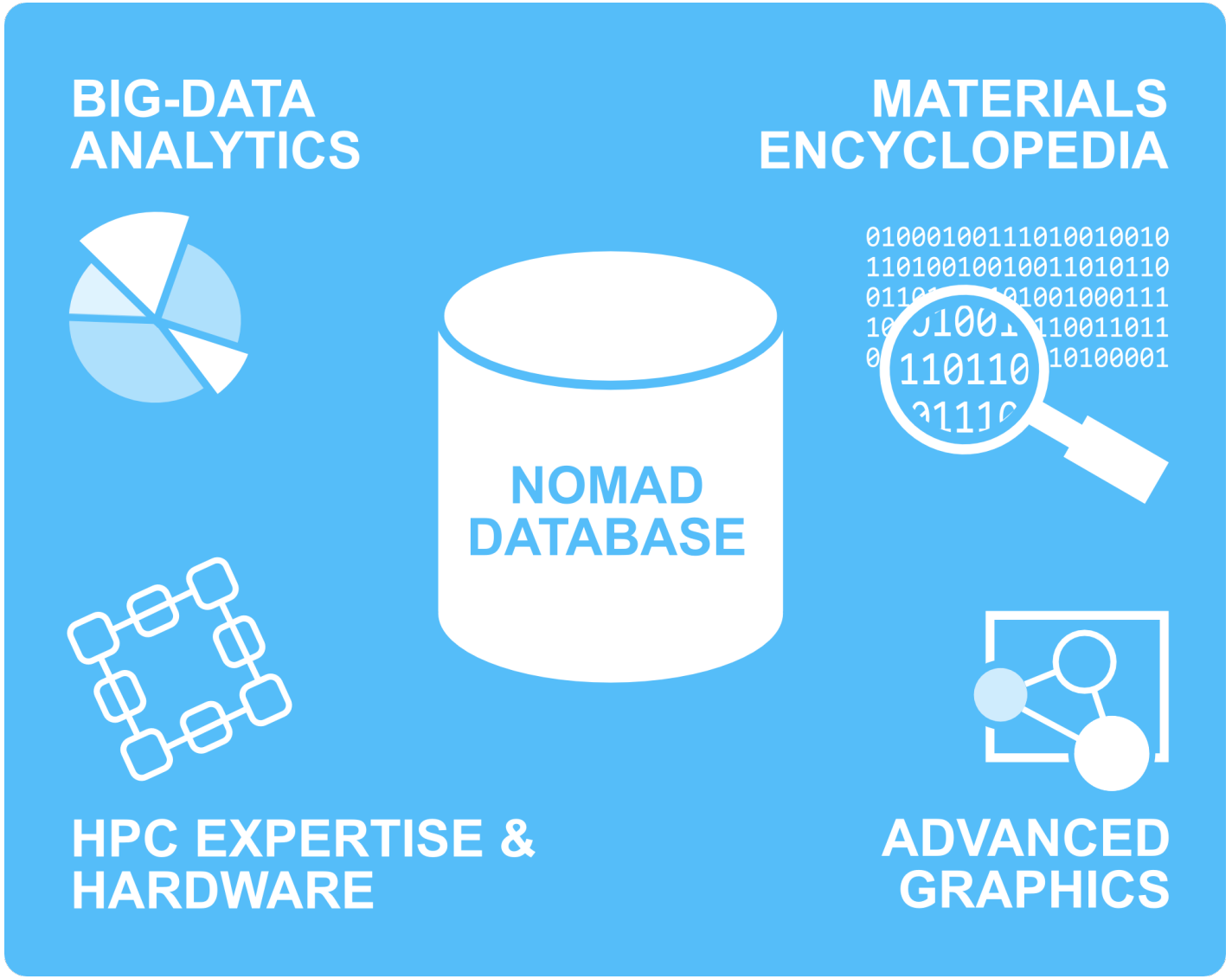
# NOMAD Archive: What is it

NOVEL MATERIALS DISCOVERY

# NOMAD Archive: What is it

NOVEL MATERIALS DISCOVERY

The core of the NOMAD Database



NOMAD

NOVEL MATERIALS DISCOVERY

# NOMAD Archive: What is it

A growing collection of files

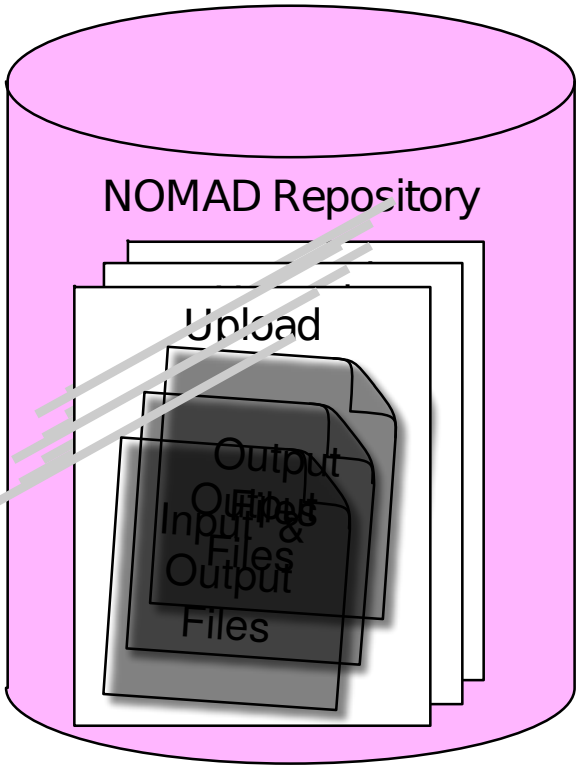
- Containing calculation data organized with **metadata**
- With clear **identifiers** to enable workflows and automatic processing
- Open access

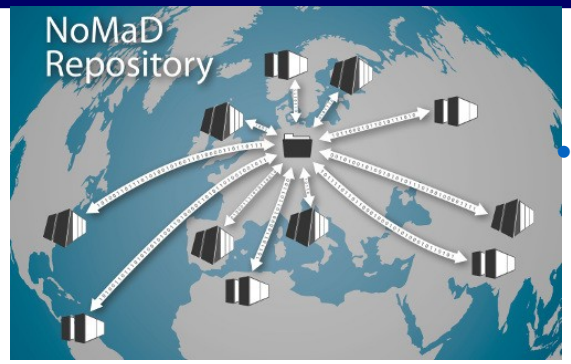
NOMAD
NOVEL MATERIALS DISCOVERY

# NOMAD Archive: How is it built

# NOMAD Repository

NOVEL MATERIALS DISCOVERY




- http://nomad-repository.eu

Claudia Draxl
HUB

Matthias Scheffle
FHI

- Source of our data
- Established to host organize and share materials data
- Keeps data for at least 10 Years
- Open access and restricted data
  - >3M entries, >2.8M open access
- We use only open access data
- 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.

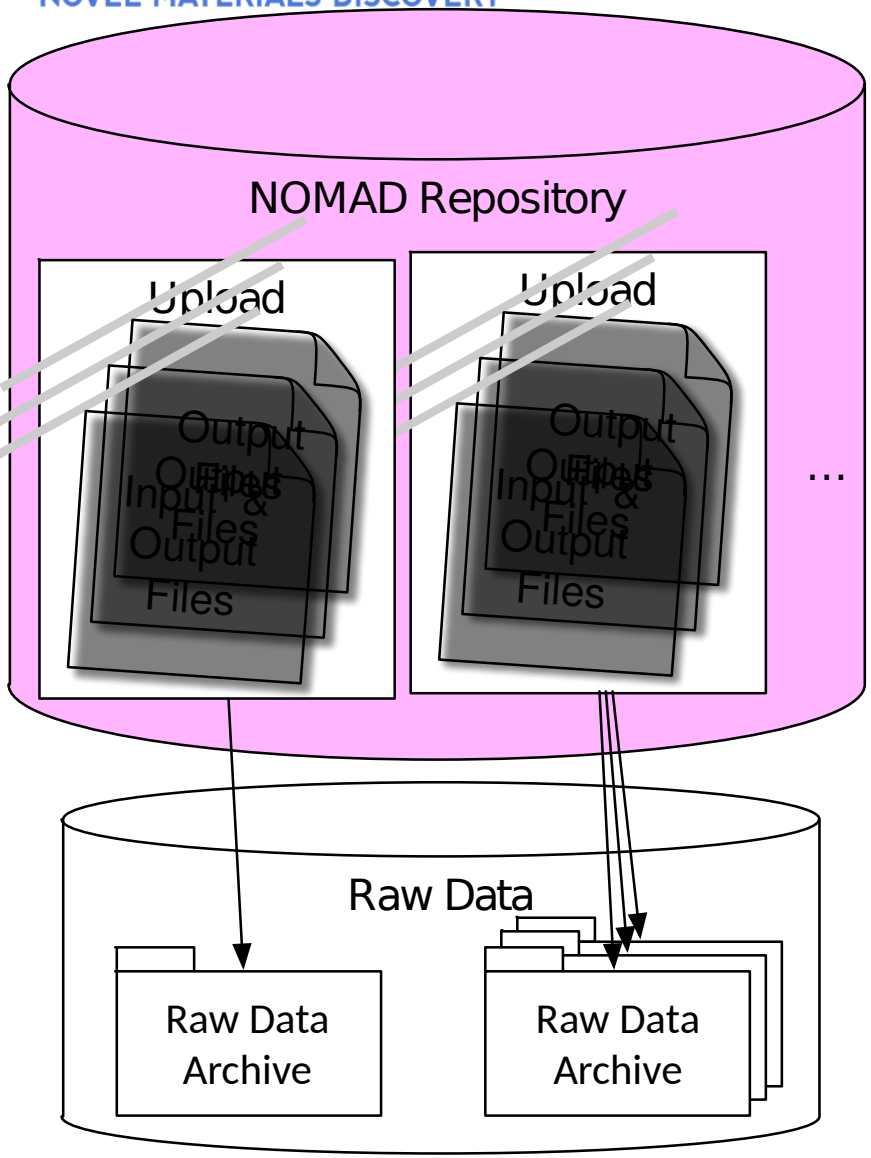
Jungho Shin
HUB

Lorenzo Pardini
HUB

Thomas Zastrow
MPCDF

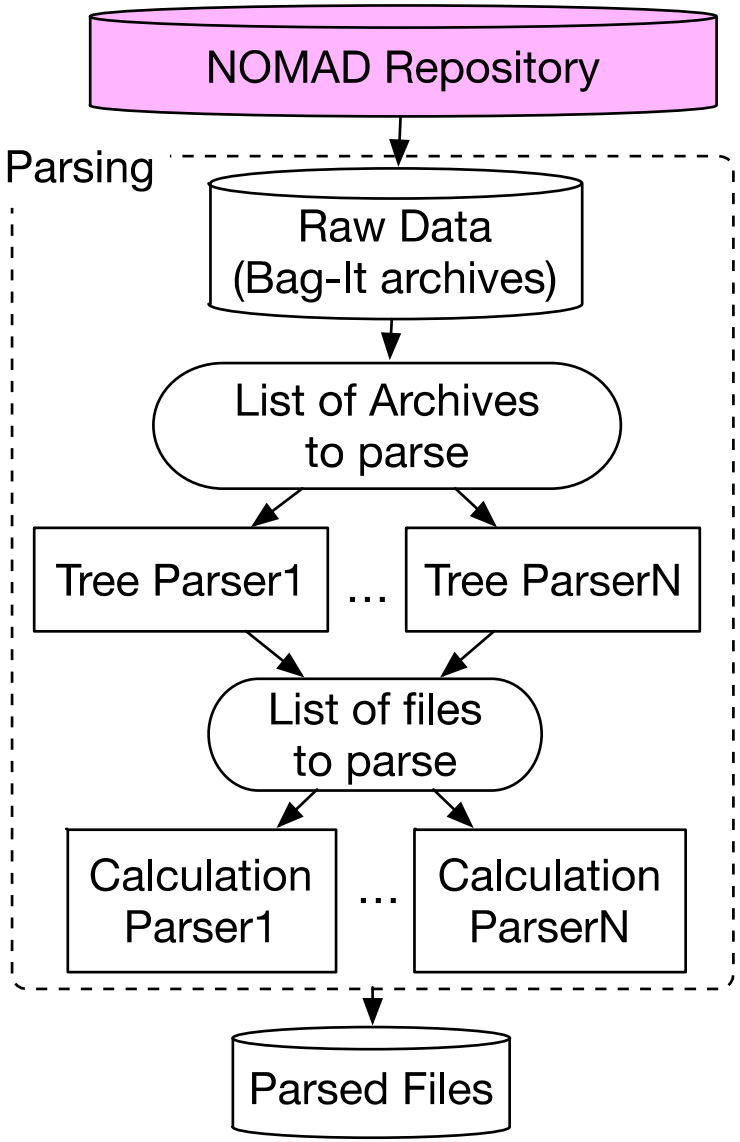
# Raw Data

NOVEL MATERIALS DISCOVERY




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

# Parsing

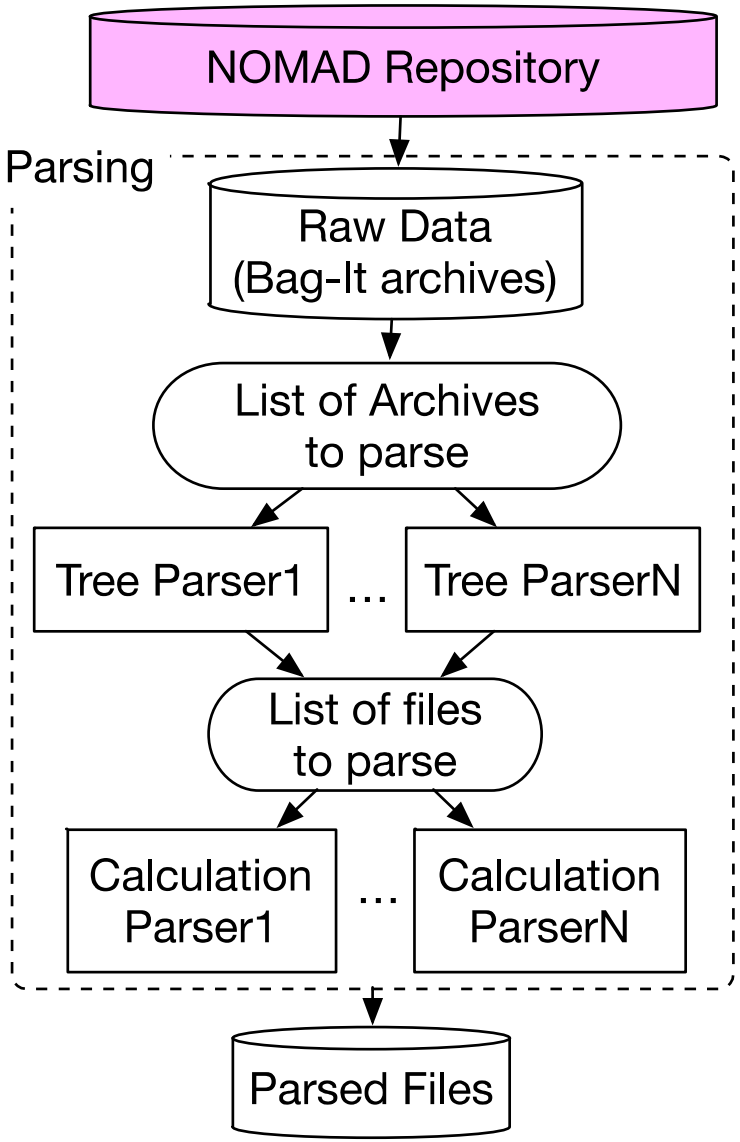



- Parsers **interpret** all calculation data
- **Organize** it according to the metadata structure
- Data not extracted is invisible
- Writing a parser cannot be automatized and requires a person with *scientific knowledge*

- **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 Status

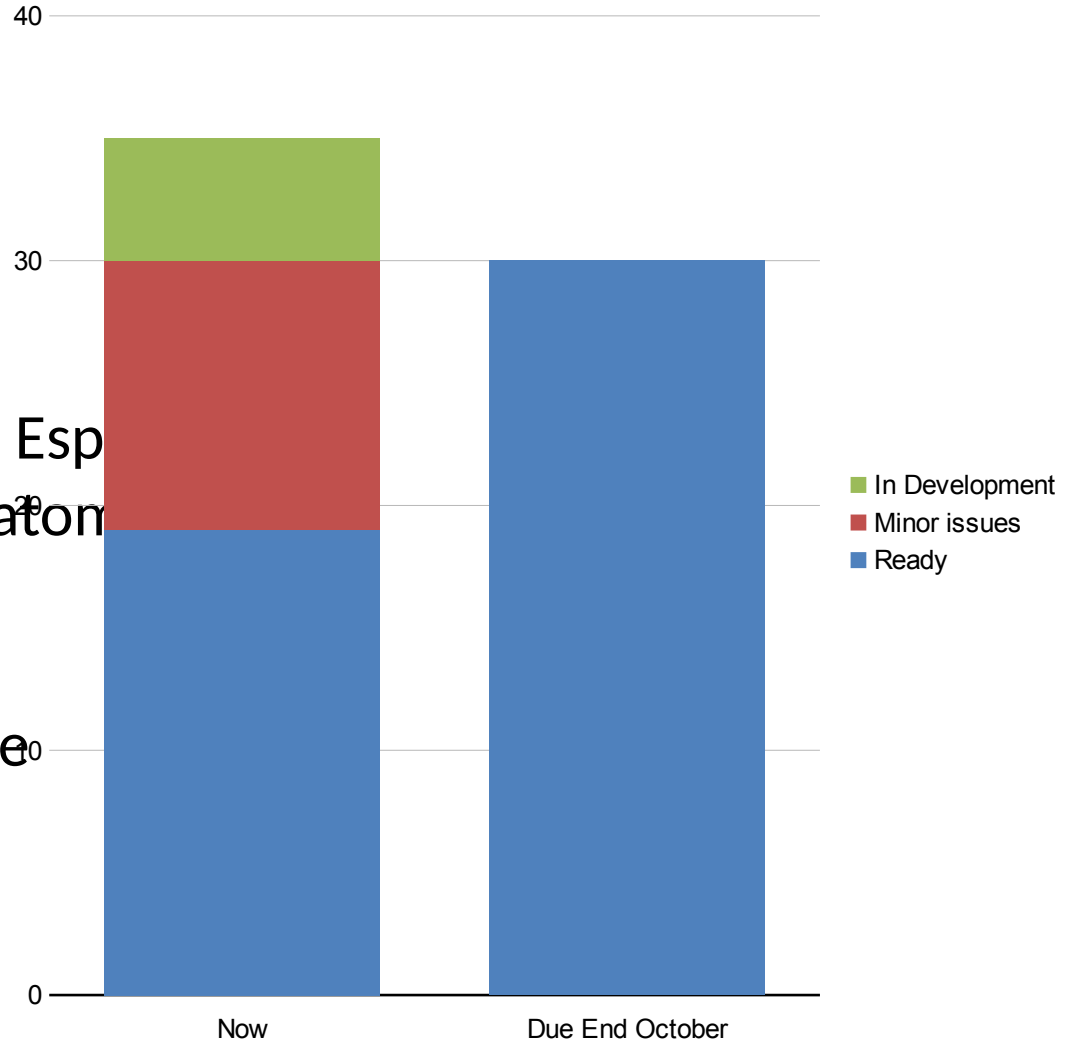
NOVEL MATERIALS DISCOVERY




## Ready:

- abinit
- CASTEP
- cp2k
- CPMD
- FHI-aims
- FLEUR
- Gaussian
- GPAW
- NWChem
- octopus
- onetep
- qBox
- Quantum Esp
- QUIP /libatom
- SIESTA
- Smeagol
- turbomole
- VASP
- WIEN2k

35 Parsers in active development



# Parser Developers



Mikkel Strange


Micael Oliviera


Martina Stella


Massimo Riello


Wael Chibiani


Franz Knuth


Aliaksei Mazheika


Adam Fekete

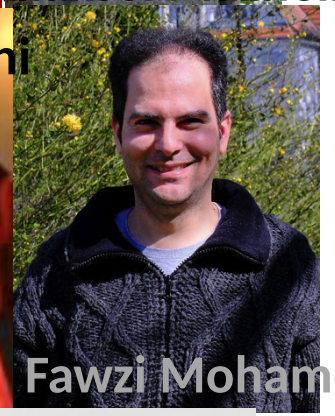

Andrea Droghetti


Honghui Shang

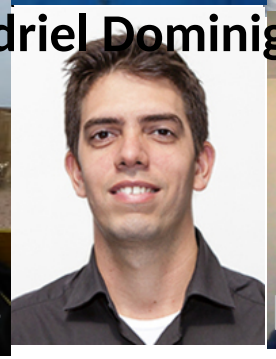

Sebastián Alarcón Villaseca


Daria Tomecka

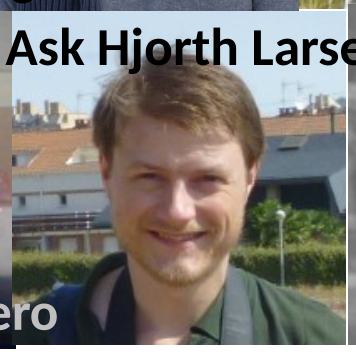

Lorenzo Pardini


Fawzi Mohamed


Carl Poelking


Adriel Dominiguez


Rosenndo Valero


Ask Hjorth Larsen


Lauri Himanneh


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

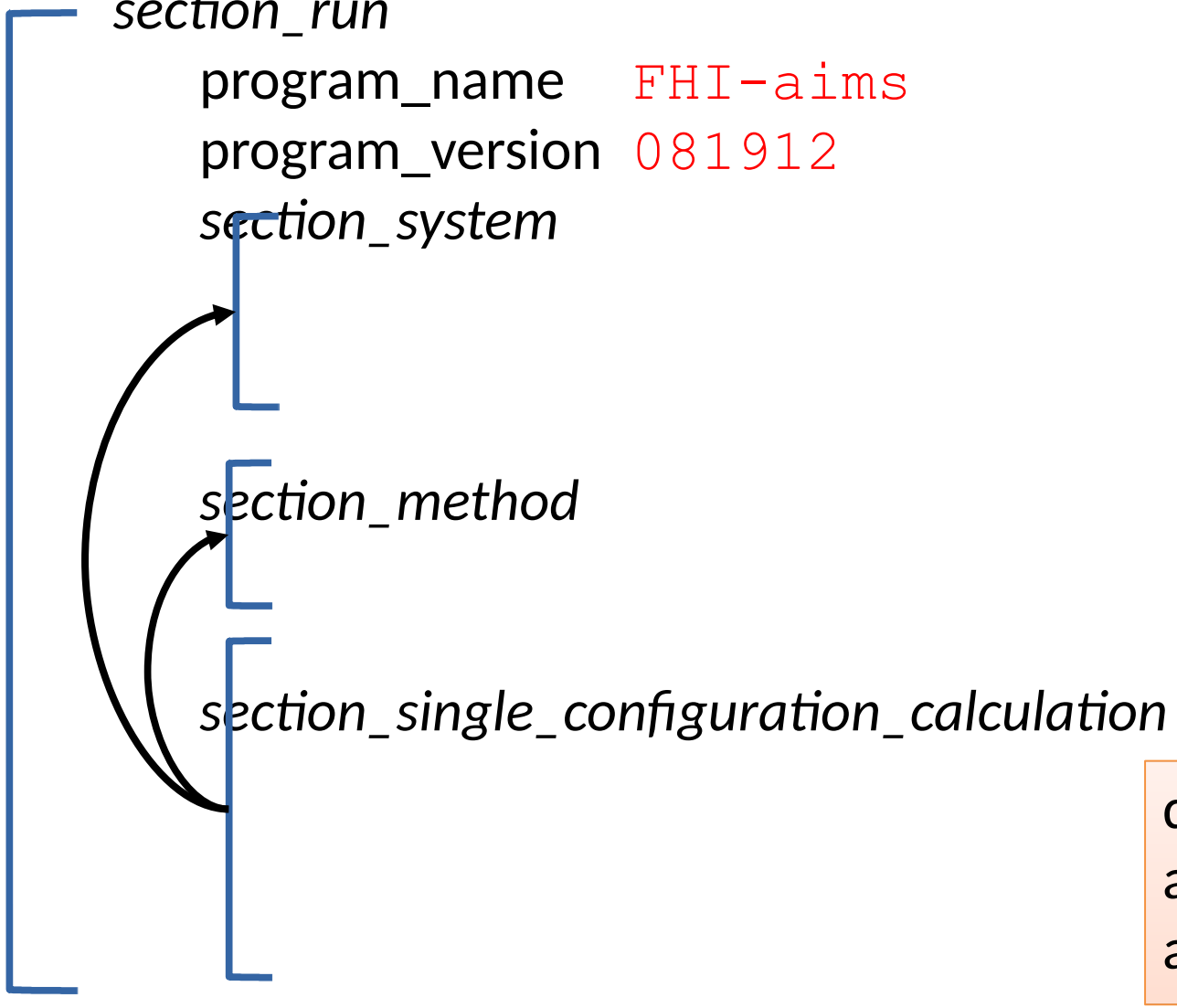
**Values: Data**
**Structures and names: Metadata**

describes the data
allows one to express queries
annotates operations performed on the data

NOMAD
NOVEL MATERIALS DISCOVERY

# Metadata

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

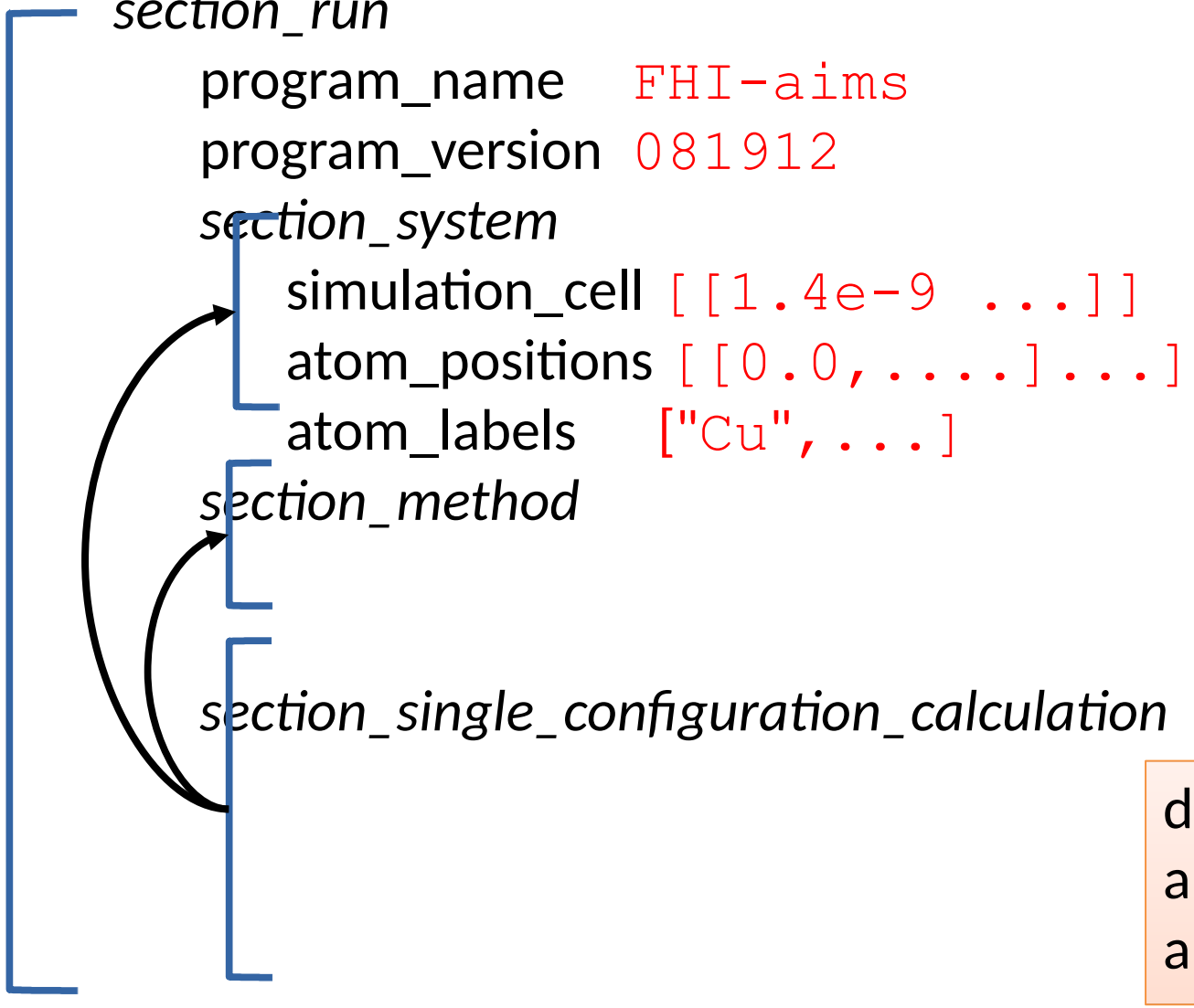



describes the data
allows one to express queries
annotates operations performed on the data

NOMAD
NOVEL MATERIALS DISCOVERY

# Metadata

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

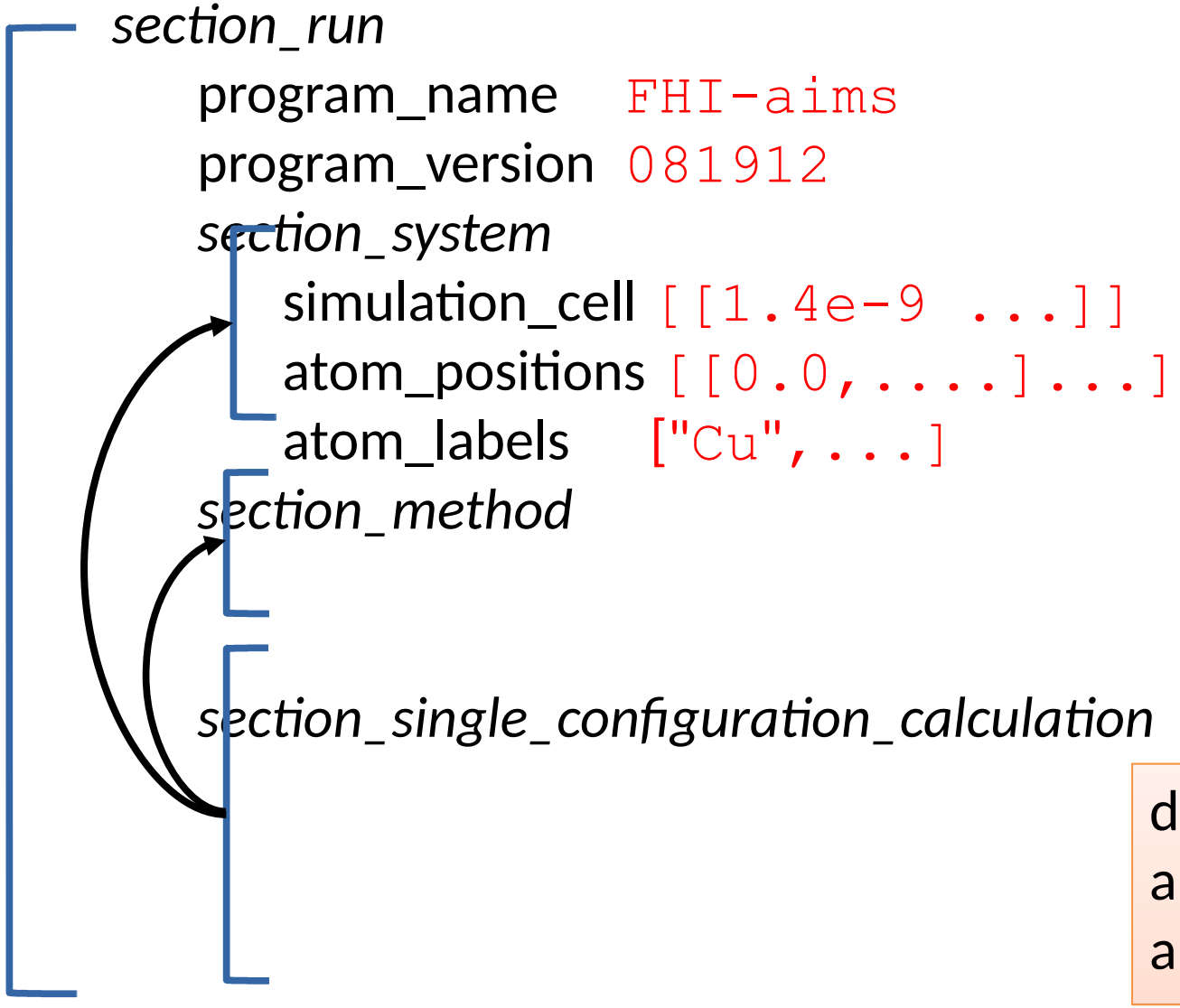



describes the data
allows one to express queries
annotates operations performed on the data

NOMAD
NOVEL MATERIALS DISCOVERY

# Metadata

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




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

describes the data
allows one to express queries
annotates operations performed on the data

NOMAD
NOVEL MATERIALS DISCOVERY

# Metadata

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

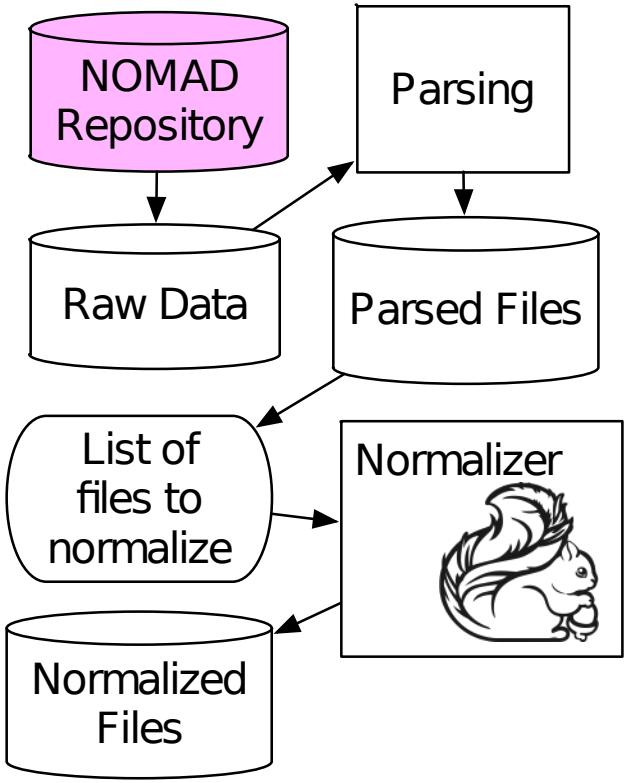
*section_run*
- program_name FHI-aims
- program_version 081912
- *section_system*
  - simulation_cell [[1.4e-9 ...]]
  - atom_positions [[0.0,....]...]
  - atom_labels ["Cu",...]
- *section_method*
  - basis_set fhi_aims_tight
  - XC_method DFT_GGA_PBE
- *section_single_configuration_calculation*

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

describes the data
allows one to express queries
annotates operations performed on the data

# Normalization

NOVEL MATERIALS DISCOVERY




## Standardization

- Parsers standardize the data,
- avoid the loss of information

## Normalization

- The work packages define **derived quantities** (normalized representations,…)
- These 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

- Raw data identifier used to group calculations

- HDF5 files and raw data available on
  **http://data.nomad-coe.eu/**

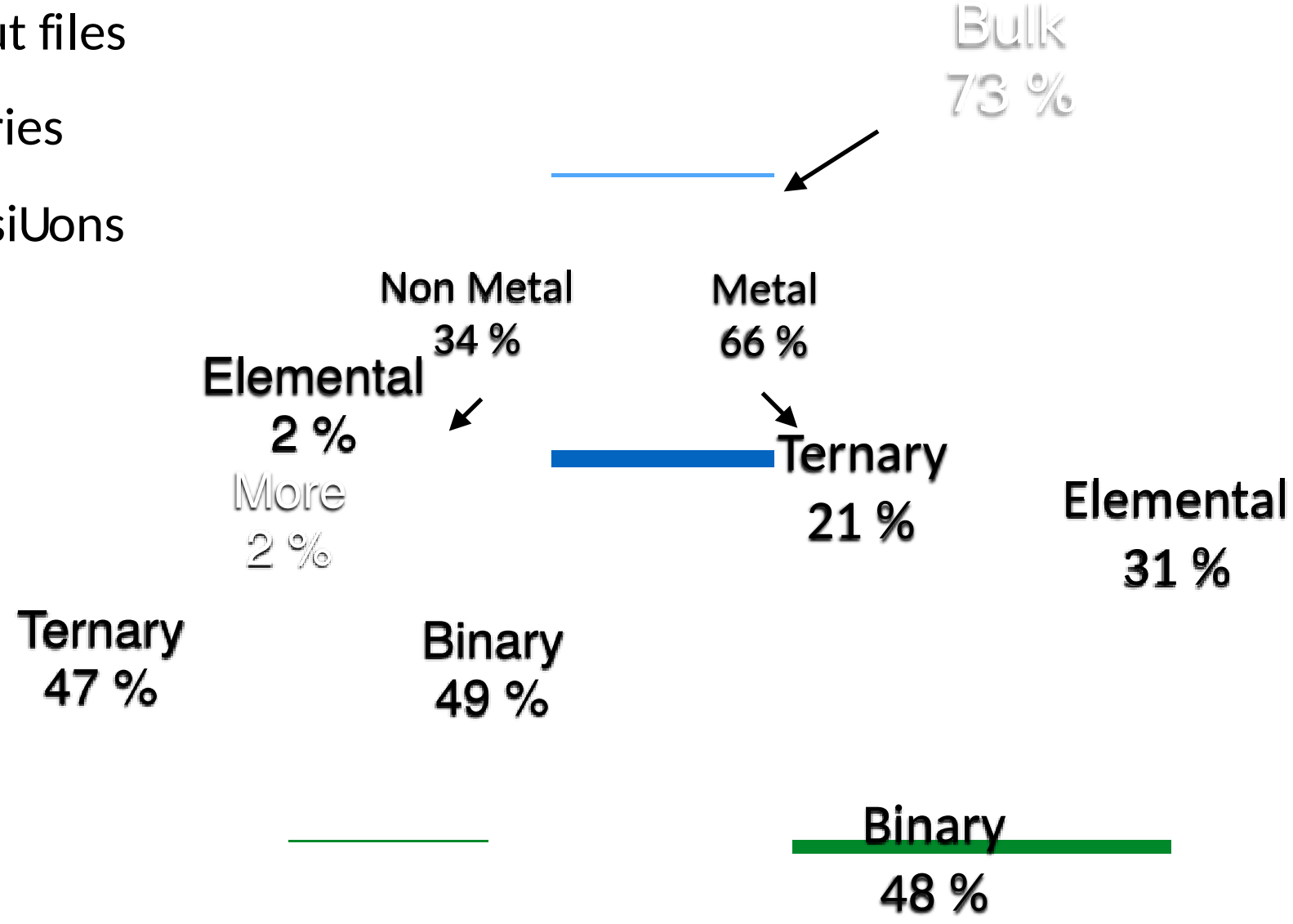
NOMAD
NOVEL MATERIALS DISCOVERY

# NOMAD Archive: What is in it

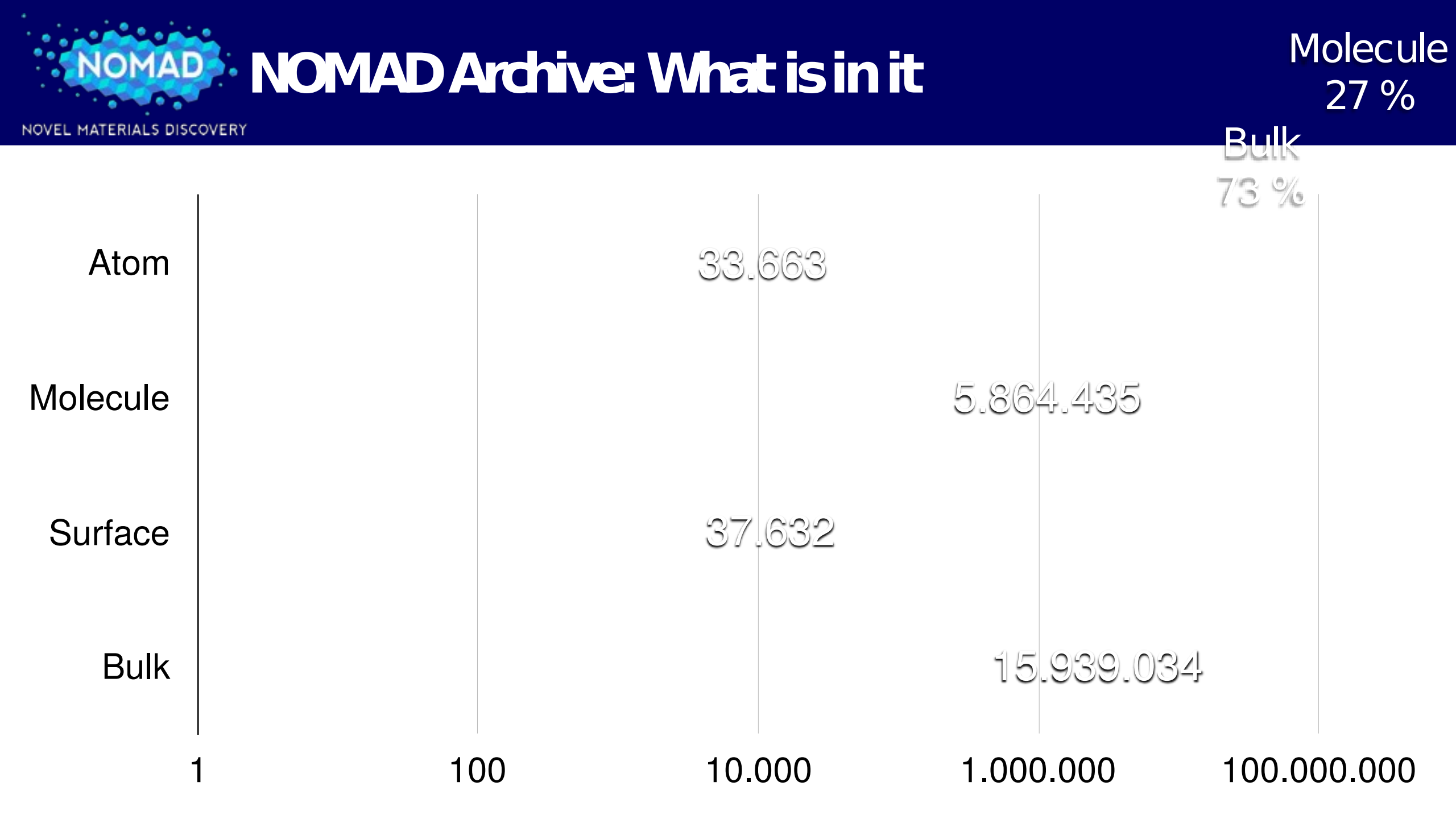
# NOMAD Archive: What is in it

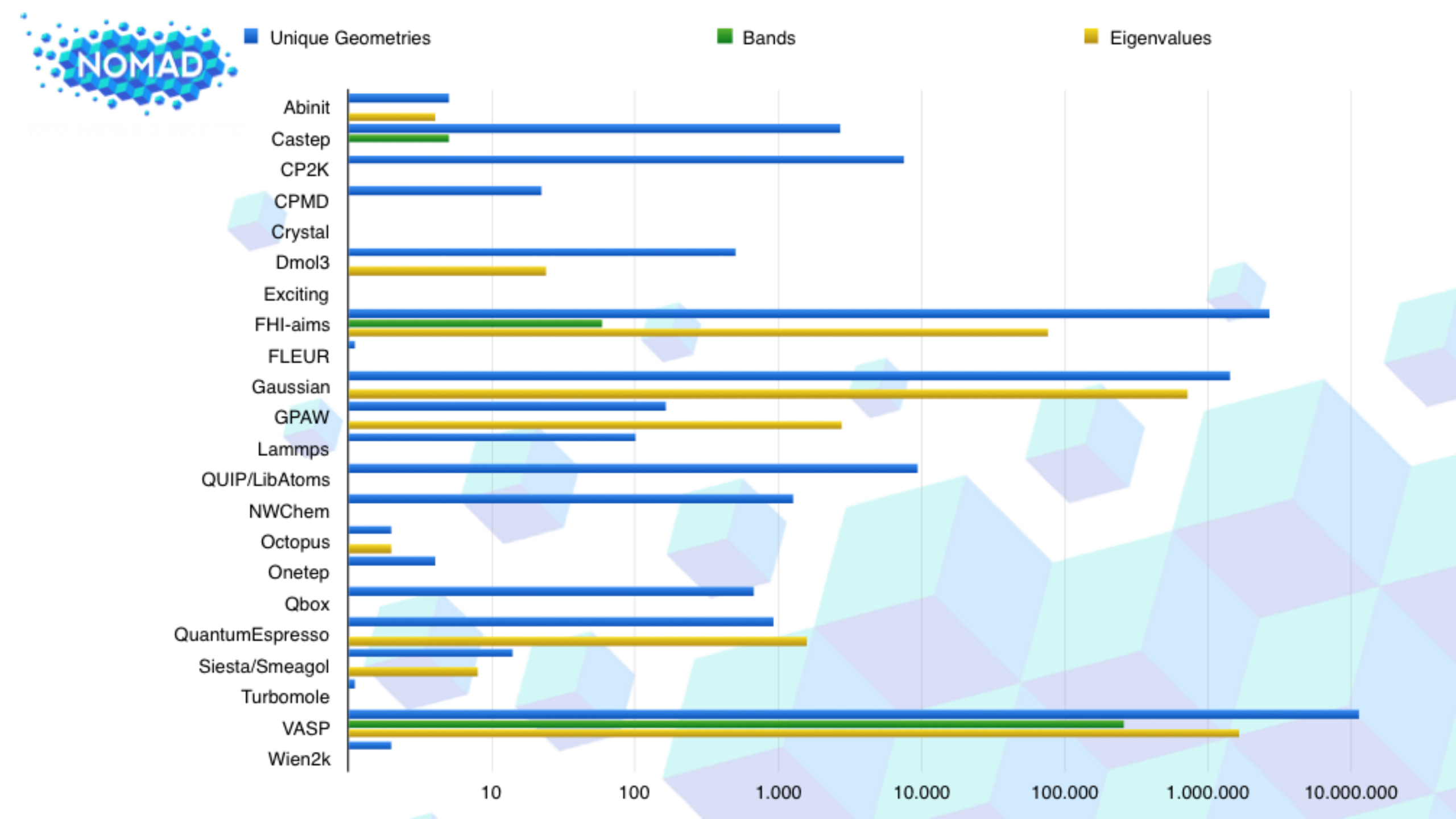
- >2.6M successfully parsed output files
- energy of >15M unique geometries
- >288k different material composiUons

  (>41k reduced)
- >260k  band structures
- >2.4M eigenvalues



# NOMAD Archive: What is in it




Unique Geometries
Bands
Eigenvalues
Abinit
Castep
CP2K
CPMD
Crystal
Dmol3
Exciting
FHI-aims
FLEUR
Gaussian
GPAW
Lammps
QUIP/LibAtoms
NWChem
Octopus
Onetep
Qbox
QuantumEspresso
Siesta/Smeagol
Turbomole
VASP
Wien2k
10
100
1.000
10.000
100.000
1.000.000
10.000.000

NOMAD

NOVEL MATERIALS DISCOVERY

# NOMAD Archive: Conclusions

The core development of the NOMAD Archive is done

- 19 codes are ready and 16 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 fix errors and support new code versions and codes
- Roughly weekly reparse of data with new parsers
- Extend normalization with new transformations used by the encyclopedia (WP2), visualization (WP3) and data analysis (WP4)
- Improve parser speed
- Extend and complete the metadata (force-field codes, excited states,…)

NOMAD

NOVEL MATERIALS DISCOVERY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 676580.