Metadata schemas: the experience of NOMAD MetaInfo

Luca M. Ghiringhelli

Workshop on NFDI tools / services / synergies between physics-related consortia and others

Erewhon, 5 April 2022

What are metadata?

NOVEL MATERIALS DISCOVERY

Working group at the
NOMAD-FAIRDI workshop: “Shared metadata and data formats for Big-Data Driven Materials Science.” Berlin, July 2019.

Participants:
(d*ata scientists)* Javad Chamanara, Patrick Lambrix, Tatyana Sheveleva ,
(*materials scientists*) Carsten Baldauf, Stefano Cozzini, Christoph Koch, Astrid Schneidewind, Christof Wöll.

Data object (information resource): a row in the data table.

UID	Structure	Method	Total energy
31415	Graphite.xyz	DFT, PBE +TS	-2718281.828 eV

Columns are attributes of the data objects.

What are metadata?

NOVEL MATERIALS DISCOVERY

UID	Structure	Method	Total energy	New structure
31415	Graphite.xyz	DFT, PBE +TS	-2718281.828 eV	Graphite_2.xyz

Columns are attributes of the data objects.
These attributes are **data** or **metadata** depending on **context**.
Administrative: location, access privileges, who, when, where.
Provenance: how, workflow.

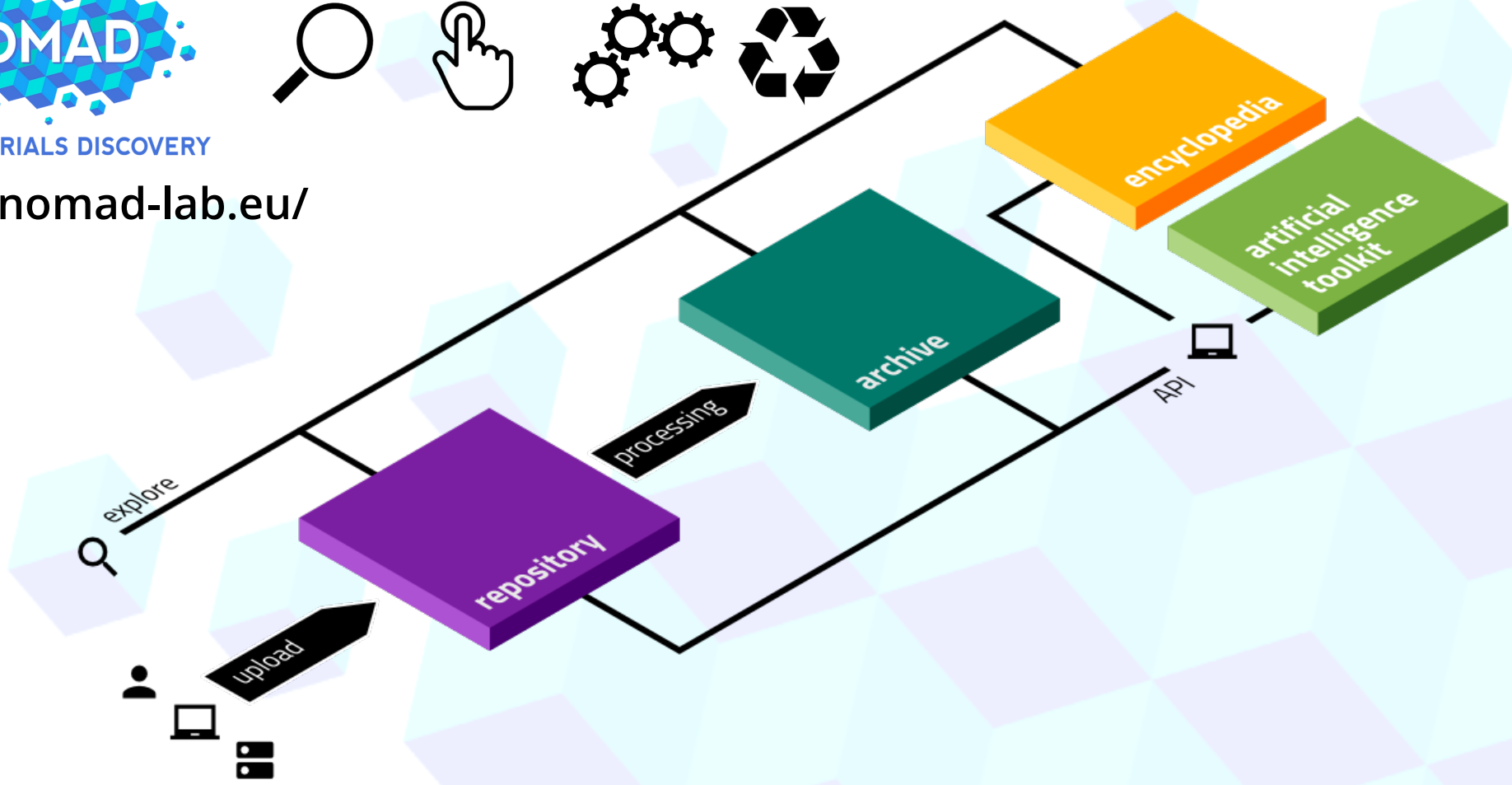
Metadata: The attributes that are necessary to locate, fully characterize, and – ultimately – **reproduce** other attributes that are identified as data. The metadata include a clear and unambiguous description of the data, and their full provenance.

The NOMAD Laboratory - https://nomad-lab.eu/

NOVEL MATERIALS DISCOVERY

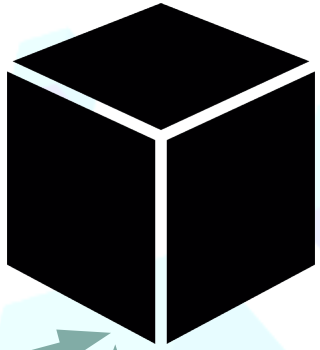
https://nomad-lab.eu/

FAIR since 2014



Computational materials science: basics

NOMAD
NOVEL MATERIALS DISCOVERY



- Code name
- Version
- Libraries

Input structure
- Coordinates
  $Z_i \;\; x_i \, y_i \, z_i$
- Cell vectors
- (Topology)

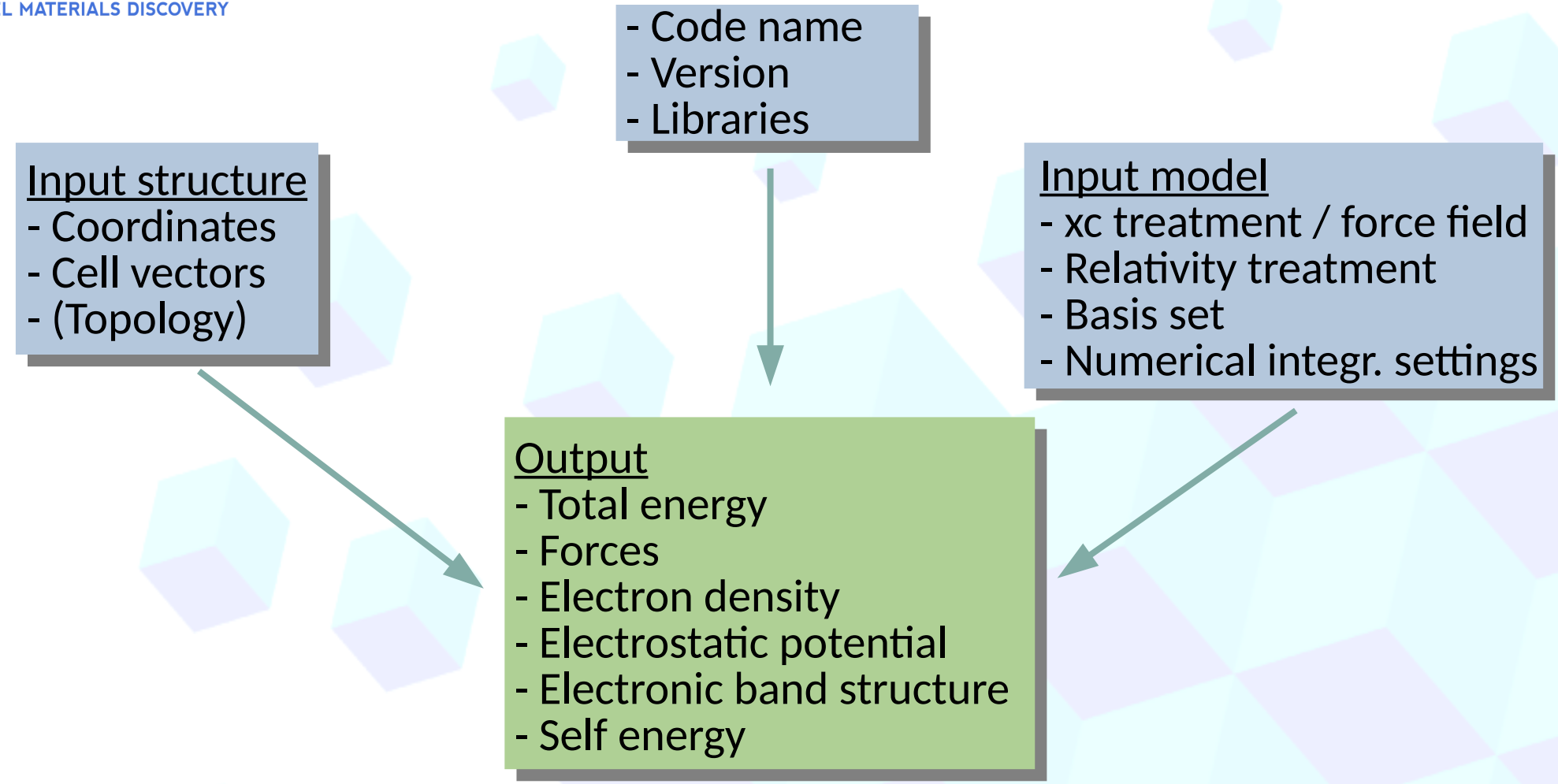
Output
- Total energy
- Forces
- Electron density
- Electrostatic pot.
- El. band structure
- Self energy
- …

Input model
- xc treatment / force field
- Relativity treatment
- Basis set
- Numerical integr. settings

More complex input model
- Model_1 (scf)→ Model_2 (perturbative)
- QM/MM (space partitioning of models)
- …

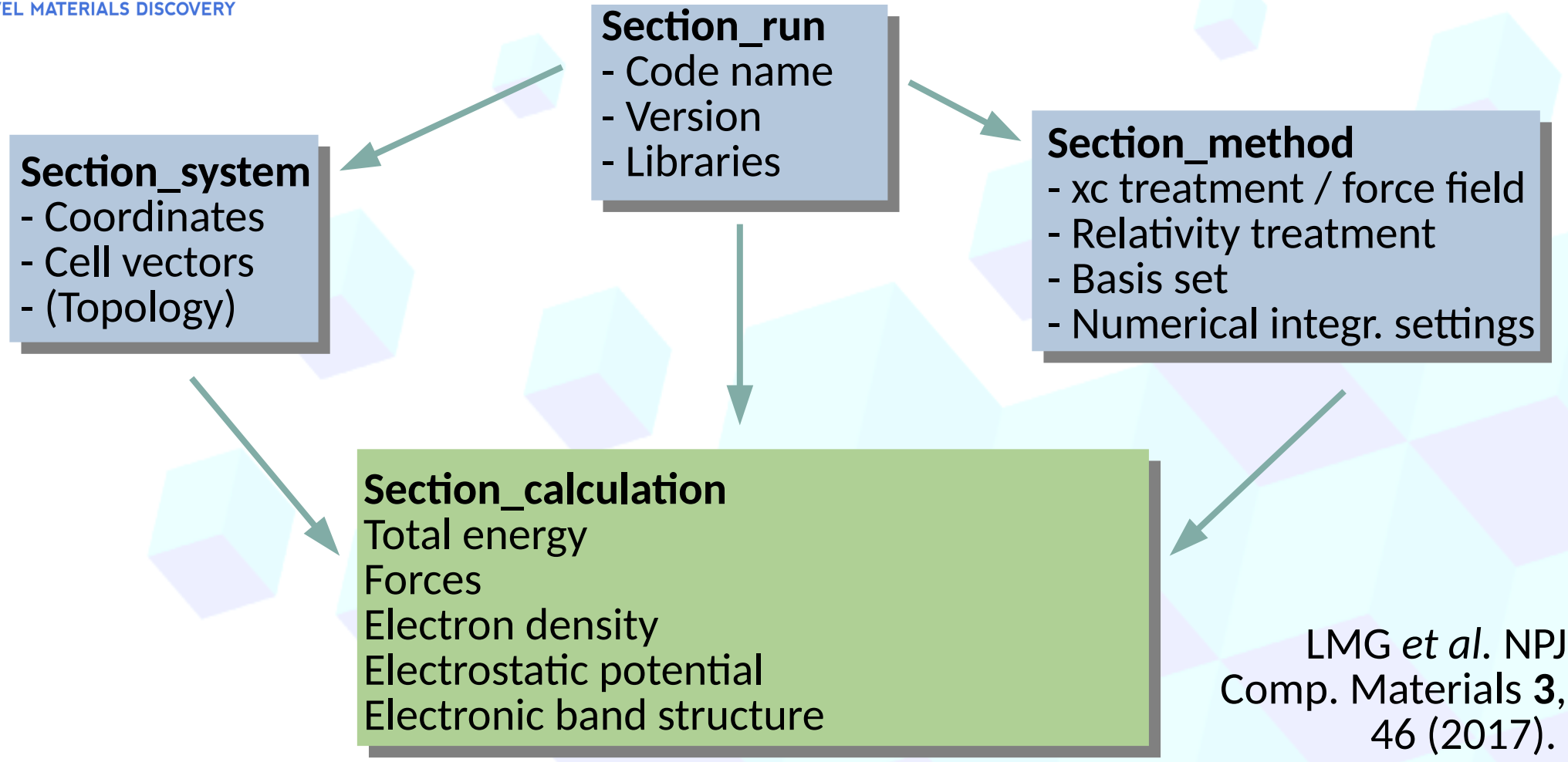
Computational materials science: basics

NOMAD
NOVEL MATERIALS DISCOVERY



NOMAD
NOVEL MATERIALS DISCOVERY

A posteriori, hierarchical scheme



LMG *et al.* NPJ Comp. Materials **3**, 46 (2017).

https://nomad-lab.eu/

NOMAD Laboratory

enables FAIR sharing and use of materials science data

Publish

- Publish your data with our without embargo, get a DOI, and share data with others.
- We support input and output files of most electronic-structure codes.
- Watch our video tutorial on how to upload and publish data.

UPLOAD

Explore

- Search for materials (Encyclopedia) or calculations (Repository).
- All raw and processed data can be downloaded and used under the CC BY 4.0.
- Watch our video tutorials on how to use the Encyclopedia and Repository.

MATERIALS CALCULATIONS

Analyze

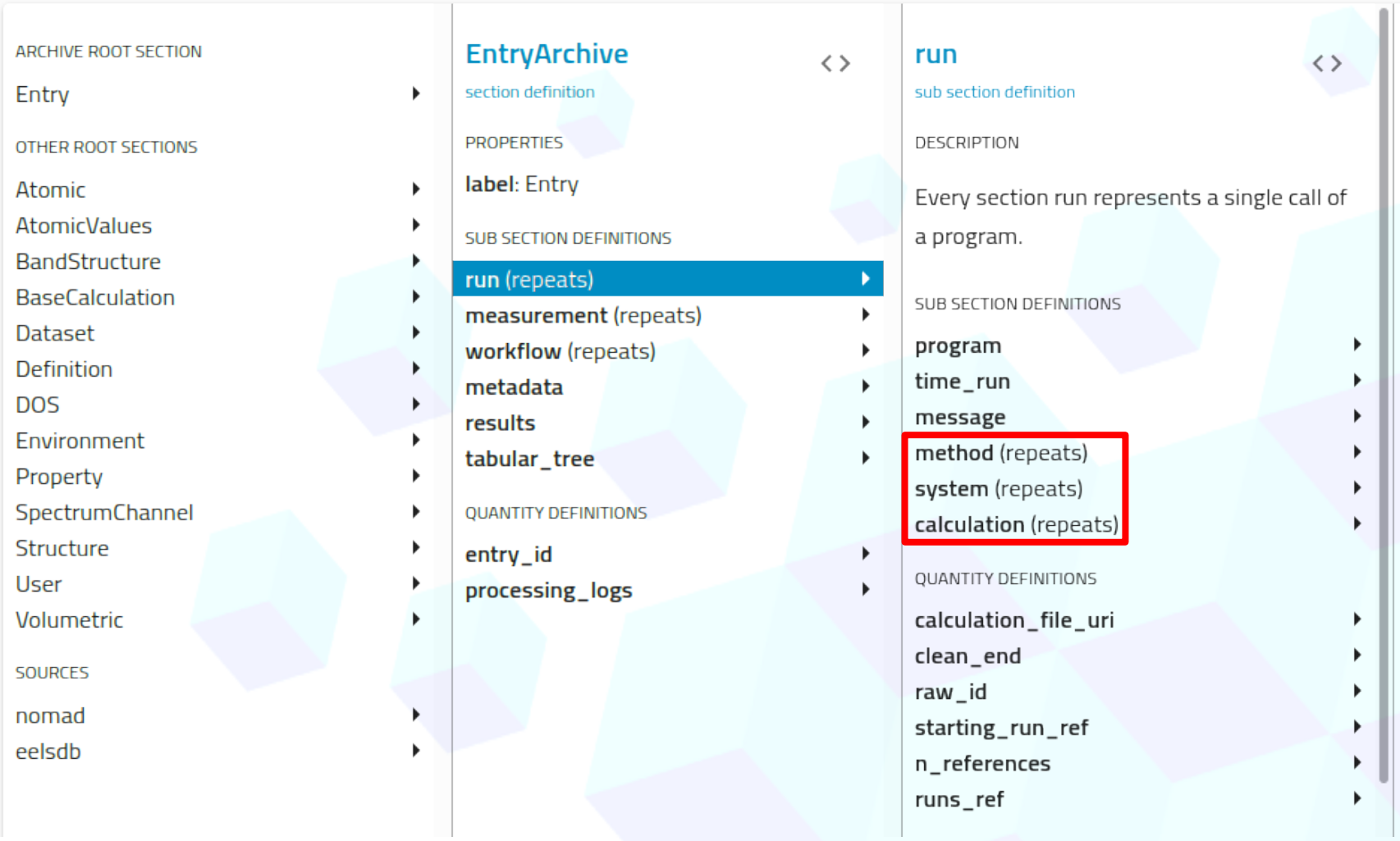
- Analyze data with Jupyter notebooks directly on NOMAD servers (Artificial Intelligence (AI) Toolkit).
- Access all data programmatically via NOMAD API or OPTIMADE API.
- Watch our video tutorial on how to use the NOMAD API.

AI TOOLKIT TUTORIALS

There is a new version of NOMAD (1.0) that we currently provide as a beta version. This installation contains most of NOMAD's data and you can already use it to upload and publish more data. Eventually all data will be migrated to this version. It will become the official NOMAD after a short beta phase. We also provide an empty test version of NOMAD. You can use this to try the upload and publish process without any consequences. We will routinly void the test data.

NOMAD 1.0 BETA NOMAD 1.0 TEST

https://nomad-lab.eu/prod/v1/gui/analyze/metainfo



https://nomad-lab.eu/prod/v1/gui/analyze/metainfo

ARCHIVE ROOT SECTION

Entry

OTHER ROOT SECTIONS

Atomic
AtomicValues
BandStructure
BaseCalculation
Dataset
Definition
DOS
Environment
Property
SpectrumChannel
Structure
User
Volumetric

SOURCES

nomad
eelsdb

EntryArchive

section definition

PROPERTIES

label: Entry

SUB SECTION DEFINITIONS

run (repeats)
measurement (repeats)
workflow (repeats)
metadata
results
tabular_tree

QUANTITY DEFINITIONS

entry_id
processing_logs

run

sub section definition

DESCRIPTION

Every section run represents a single call of a program.

SUB SECTION DEFINITIONS

program
time_run
message
method (repeats)
system (repeats)
calculation (repeats)

QUANTITY DEFINITIONS

calculation_file_uri
clean_end
raw_id
starting_run_ref
n_references
runs_ref

inclue the compound name, atomic positions, lattice vectors, contraints on the atoms, etc.

SUB SECTION DEFINITIONS

atoms
constraint (repeats)
prototype (repeats)
springer_material (repeats)
symmetry (repeats)

QUANTITY DEFINITIONS

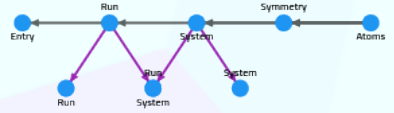
name
type
configuration_raw_gid
is_representative
n_references
sub_system_ref
systems_ref
chemical_composition
chemical_composition_hill
chemical_composition_reduced
chemical_composition_anonymous

positions, lattice vectors, etc.

QUANTITY DEFINITIONS

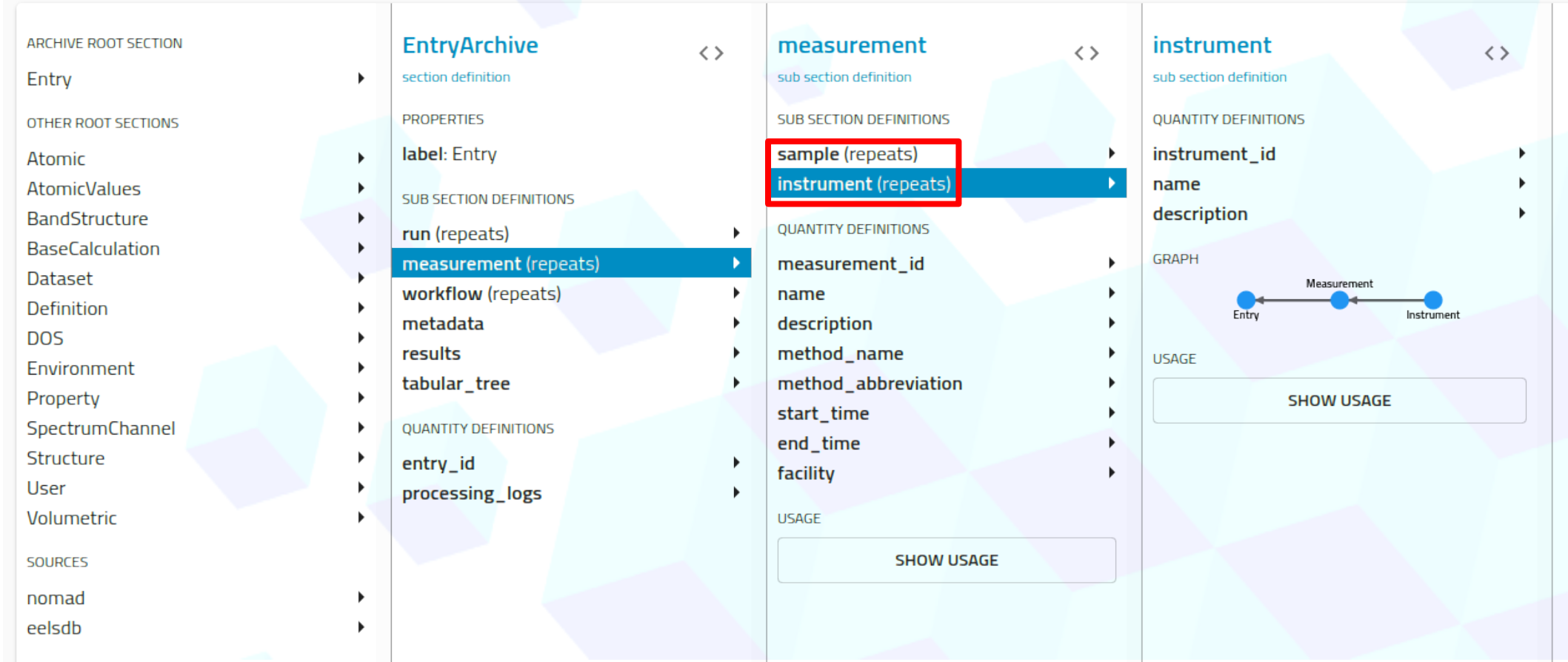
n_atoms
atomic_numbers
equivalent_atoms
wyckoff_letters
concentrations
species
labels
positions
velocities
lattice_vectors
lattice_vectors_reciprocal
local_rotations
periodic
supercell_matrix
symmorphic

GRAPH



https://nomad-lab.eu/prod/v1/gui/analyze/metainfo

Not only calculations!



https://nomad-lab.eu/prod/v1/gui/analyze/metainfo

run

sub section definition

DESCRIPTION

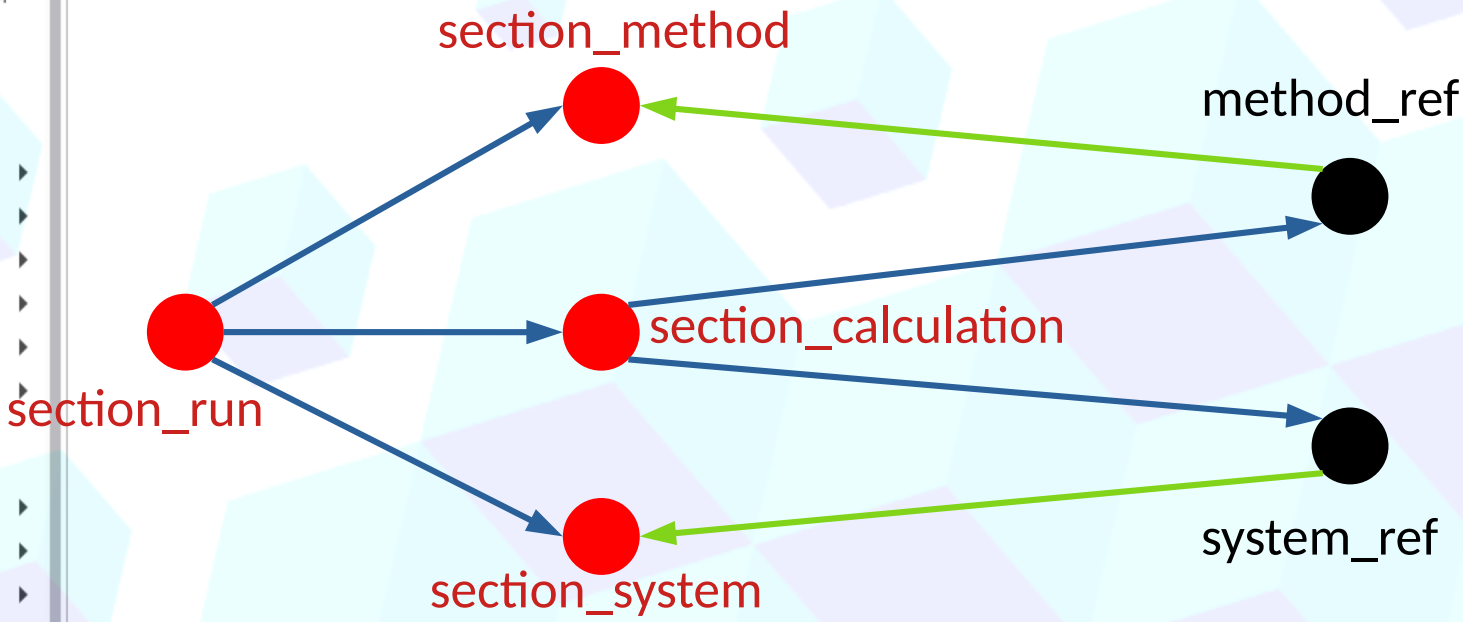
Every section run represents a single call of a program.

SUB SECTION DEFINITIONS

- program
- time_run
- message
- **method** (repeats)
- **system** (repeats)
- **calculation** (repeats)

QUANTITY DEFINITIONS

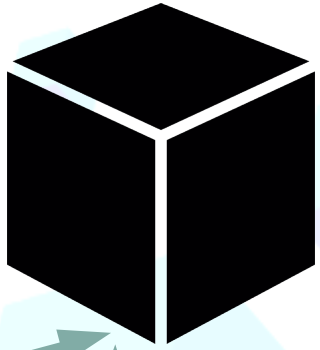
- calculation_file_uri
- clean_end
- raw_id
- starting_run_ref
- n_references
- runs_ref



Computational MatSci: from basics...

NOMAD
NOVEL MATERIALS DISCOVERY

Atomistic-simulation code



- Code name
- Version
- Libraries

Input structure
- Coordinates
  $Z_i \;\; x_i \, y_i \, z_i$
- Cell vectors
- (Topology)

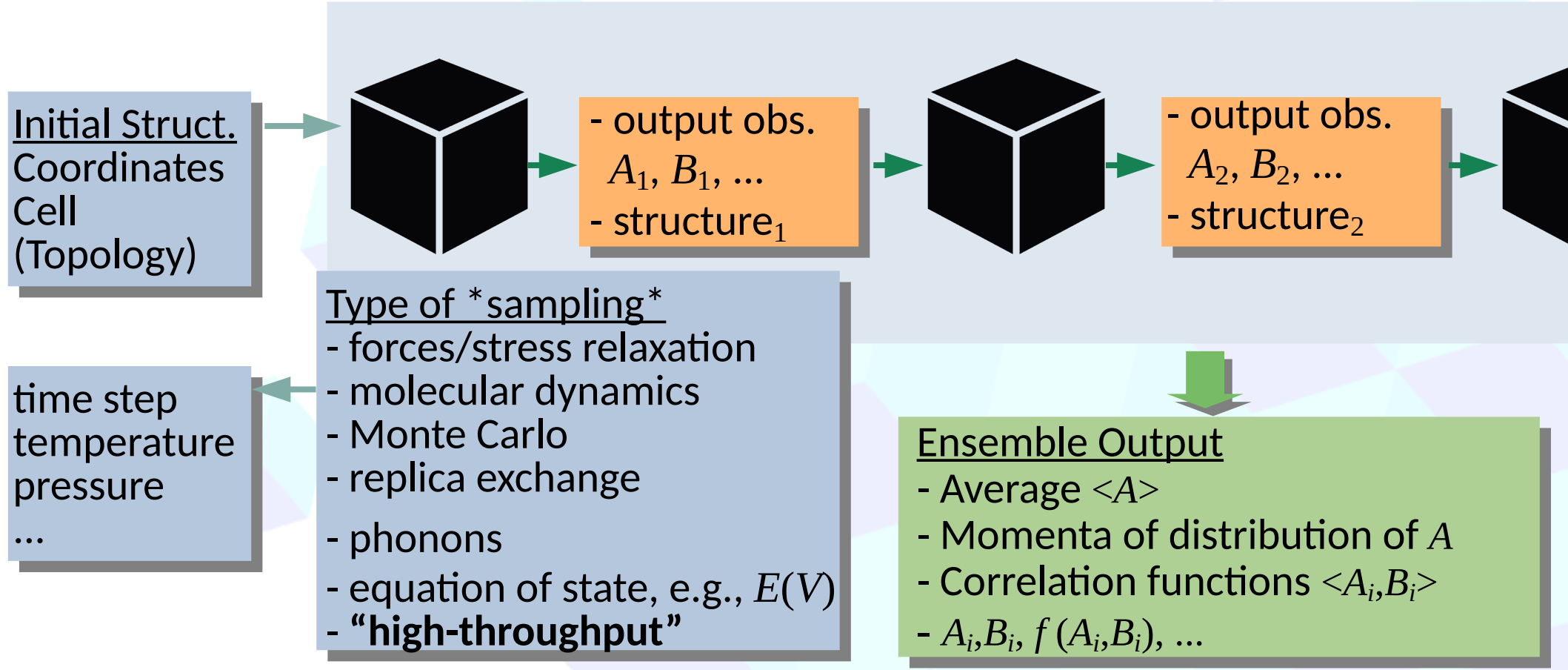
Output
- Total energy
- Forces
- Electron density
- Electrostatic pot.
- El. band structure
- Self energy
- ...

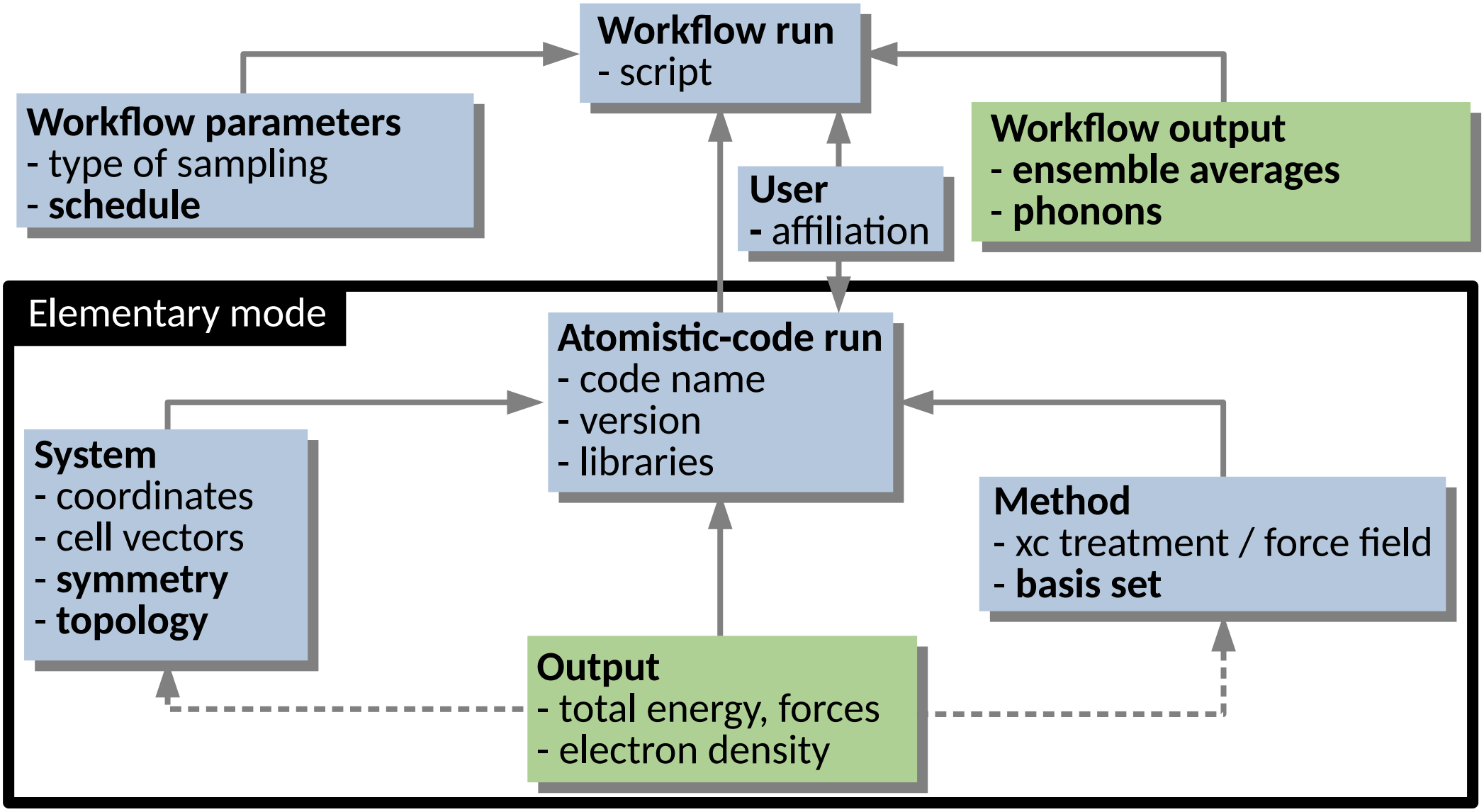
Input model
- xc treatment / force field
- Relativity treatment
- Basis set
- Numerical integr. settings

More complex input model
- Model_1 (scf)→ Model_2 (perturbative)
- QM/MM (space partitioning of models)
- ...

NOMAD
NOVEL MATERIALS DISCOVERY

… to a more complex *workflow* …




Workflow run
- script
Workflow parameters
- type of sampling
- schedule
User
- affiliation
Workflow output
- ensemble averages
- phonons
Elementary mode
Atomistic-code run
- code name
- version
- libraries
System
- coordinates
- cell vectors
- symmetry
- topology
Method
- xc treatment / force field
- basis set
Output
- total energy, forces
- electron density

In summary: FAIRness

NOVEL MATERIALS DISCOVERY

MetaInfo: Metadata for FAIR scientific-data management and stewardship.

Findable: unique names, human-readable descriptions

Accessible: URL, accessible via API

Interoperable: typed, extensible schema → ontologies

Reusable: hierarchical schema → data-analytics

Risks and challenges: Redundant or conflicting metadata

Risks and challenges: Heterogeneous computational methods

Acknowledgments:

MetaInfo: Fawzi Mohamed, Pasquale Pavone, Henning Glawe, Micael Olivera, Benjamin Regler, Bryan Goldsmith, Lauri Himanen, Alvin Noe Ladines, Joseph Rudzinski, Nathan Daelman, Robert Hussein, and more.

NOMAD Repository & Archive: Markus Scheidgen and FAIRmat infrastructure team

NOMAD Lab & FAIRmat: Matthias Scheffler, Claudia Draxl