Workflows in NOMAD

23.5.2019

Fawzi Mohamed
fawzi@kitabi.eu

NOMAD CoE

- EU project Nov 2015 - Nov 2018 -> Fairdi
- 8 research centers
- 4 super computing centers

UNIVERSITY OF CAMBRIDGE

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

DTU Technical University of Denmark

mpsd

KING'S College LONDON

CSC
CSC-TIETEEN TIETOTEKNIIKAN KESKUS

Aalto University

RECHENZENTRUM GARCHING

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Universitat de Barcelona

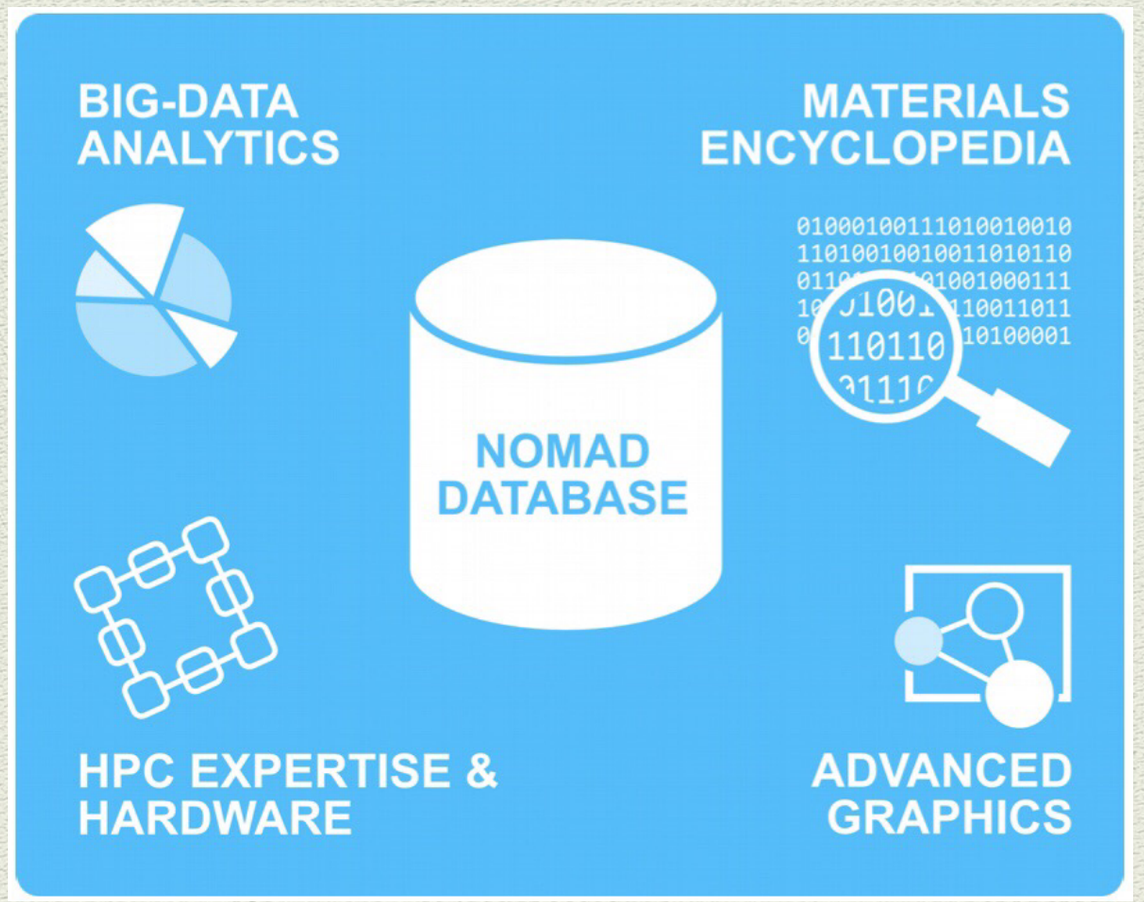
Leibniz-Rechenzentrum
der Bayerischen Akademie der Wissenschaften

MAX-PLANCK-GESELLSCHAFT

NOMAD CoE Goals

- more than 200,000 materials “known” to exist basic properties known for very few
- highly likely that new materials with superior properties exist
- big-data analytics tools to identify trends and anomalies in material data
- First Step
  - Make data available
  - Make data addressable
  - Make data organizable

http://nomad-coe.eu



Archive: Make Data available

NOMAD REPOSITORY

NOMAD Archive

Make data of the simulation codes used in the community accessible

WP2: NOMAD ENCYCLOPEDIA

WP3: NOMAD *VISUALIZATION*

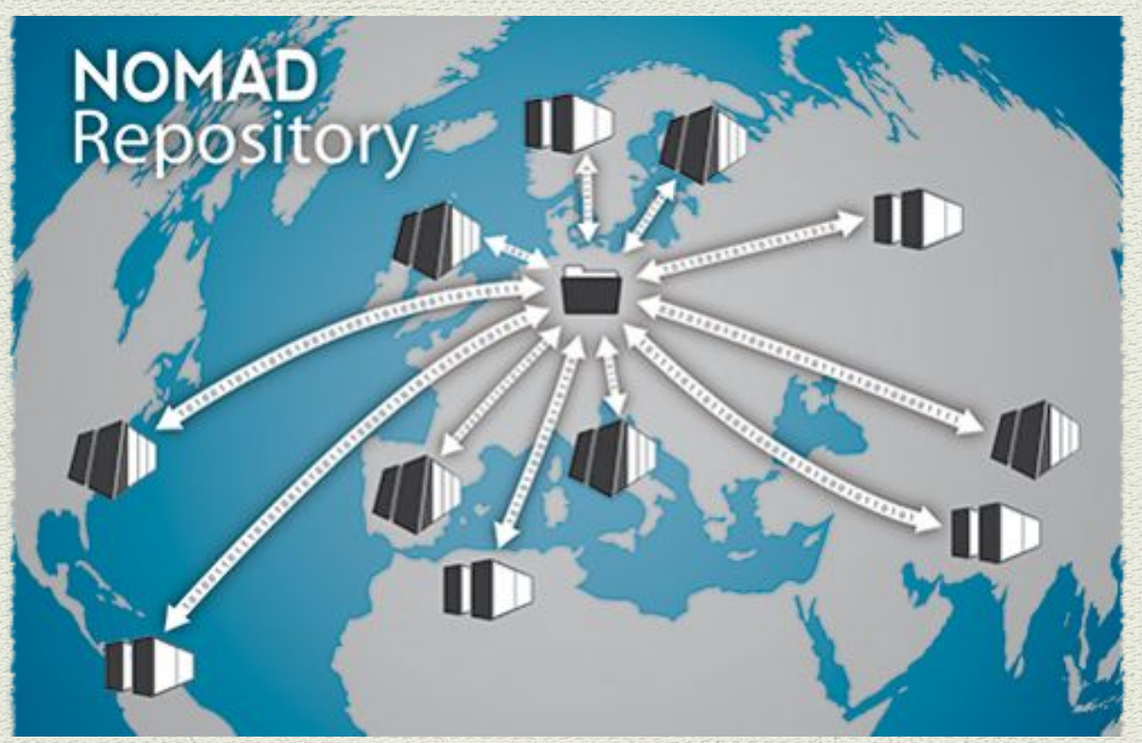
WP4: BIG DATA ANALYSIS

External users

NOMAD Repository

https://nomad-repository.eu

- repository of computational data
- Storage for at least 10 years
- support 40 computational codes
- FHI, HUB, MPCDF



Big Data

- Big data is not only amounts of data...
- ...but also robust tools and workflows.
- not to just producing big amounts of data but to do so
  - in a reliable way
  - so that problems can be found, fixed,
  - ... and get the correct data integrated

Naming and Provenance

- very important to achieve these goals
- required
  - about the source of the data
  - to describe the data itself

Philosophy

- Scientific openness
- Open Access to everything
- Allow others to develop alternatives
- Track origin (cite original contribution)
- API should be used because of its convenience, not because it is the only option
- Low level, bulk sharing of data, via REST API (static files): `http://data.nomad-coe.eu`

NOMAD
NOVEL MATERIALS DISCOVERY

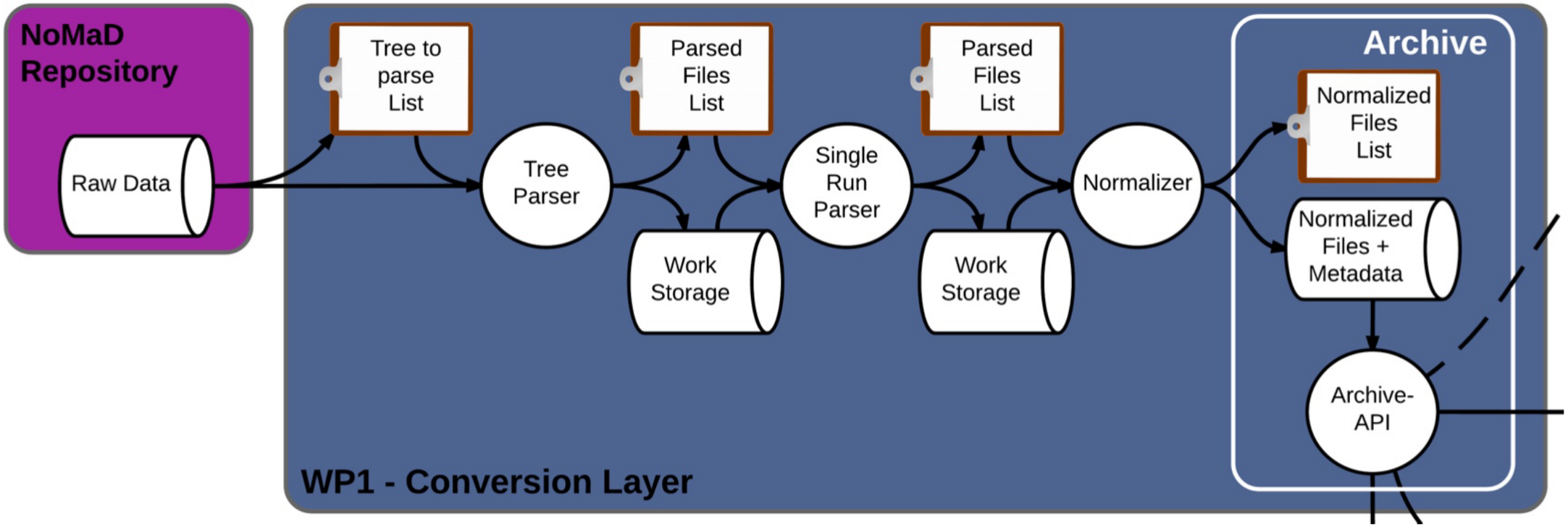
NOMAD: Data

- Atomistic simulations (mostly ab initio)
  - AFLOW, OQMD, CCCBDB and Materials Project largest contributors
- >30 Different Program (file formats)
- 9,274 Zip Archives for parsing: 16.5 TB of data (compressed)
- Data extracted with parsing: 5.6 TB of HDF5 files (compressed)
- Data classified using 168 public metadata of the NOMAD Metadata and 2,360 code-specific metadata
- Number of parsed quantities 871,497,996

44,179,006	**39,402,326**	**220,918**	**4,517,016**
Total-Energy Calculations	Bulk Crystals	Surfaces	Molecules/Clusters
40,481,615	**281,135**	**1,936,325**	**91**
Different Geometries	Chemical Compositions	Band Structures	Phonon Calculations

RabbitMQ, Kubernetes and Docker

- We use RabbitMQ as queuing system and docker container for the workers
- `http://kubernetes.io` is used to coordinate and scale the worker



NOMAD
NOVEL MATERIALS DISCOVERY

Source Data

- public data from the NOMAD Repository
- archives using the BagIt format
- Name depends only on the content of the archive (archiveGid -> R+28-char checksum)
- host independent uris: nmd://<archiveGid>/path
- clear naming of program that transforms the data (git describe)

Expressions for derived data

- parsed data comes from archive X (raw data naming and nomad uri), and was produced parser Y version Z1

- we can say: “replace (or compare) that data with the one using parser Y version Z2”

Describing the data: Meta Info

- our conceptual model of the data
- language and format independent
- defines how the data we extract is organized
- important for both human and machine

Metadata - MetaInfo

- data about the data
- depends on what is data
- rfc5013: The Dublin Core Metadata Element Set
- avoid misconceptions: MetaInfo

NOMAD Meta Info

- Data: values of physical quantities or settings of a computational method
  - Numbers, strings, booleans
  - References to other groups of data (relational model)
  - Multidimensional arrays of these types (useful to support scientific data)
- Metadata:
  - Describe and organize the data
  - Unique name for all concrete quantities, and the groups used to organize them (using alphanumeric characters and ”_”)
  - Hierarchical organization (tree)
  - Single public name-space
  - Semantic versioning

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

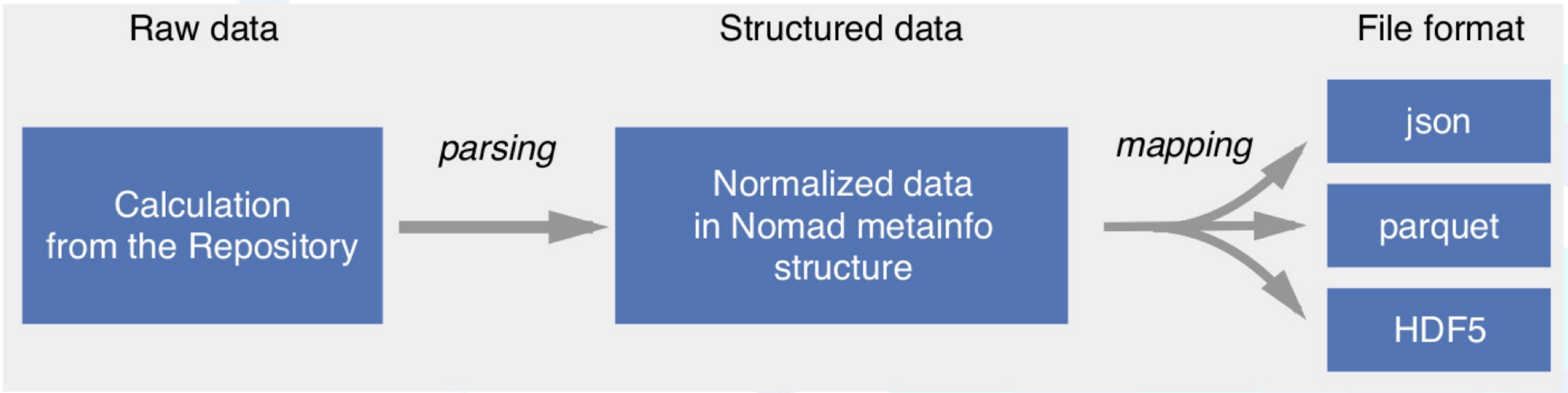
Values: Data
Structure and names: Meta info

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

- group together values with sections
- allow one to many relationships between sections (relational data model)
- group together similar types making them inherit from the same type (all energies inherit from the energy)

NOMAD
NOVEL MATERIALS DISCOVERY

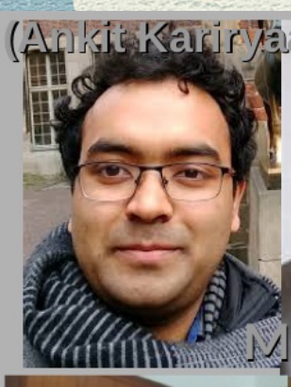
Storing and Exchanging data

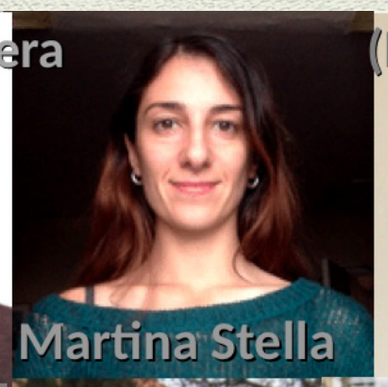


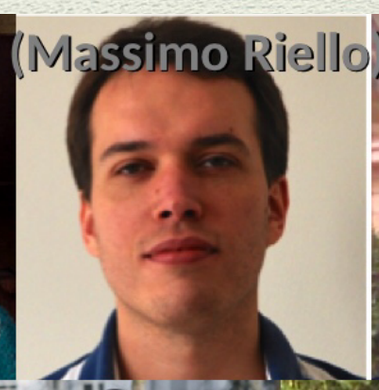
Parsers

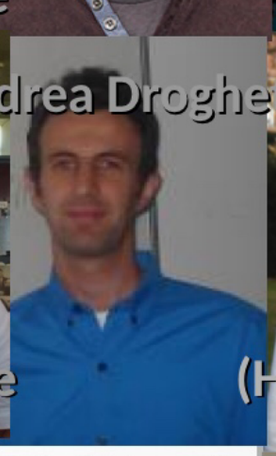
- Map raw data to the model defined by the Meta Info
- Require knowledge of the simulation methods
- Many Codes -> many parsers

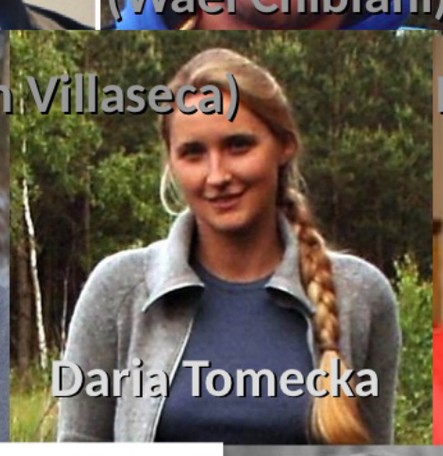
Parser Developers

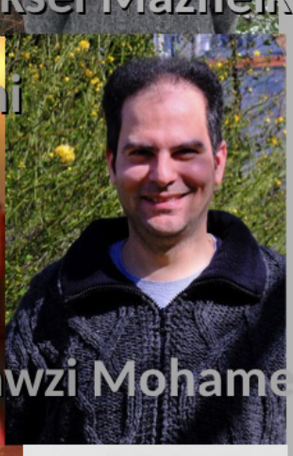


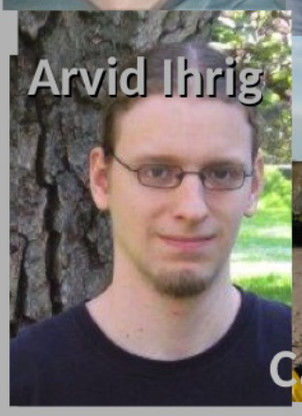


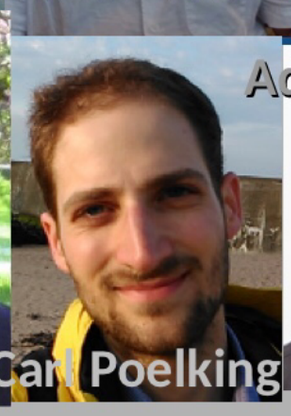


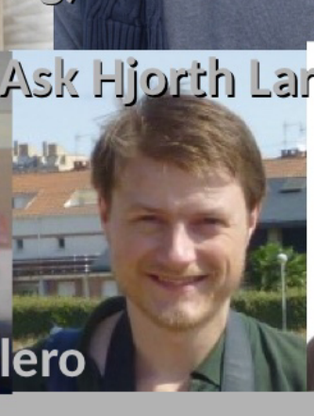




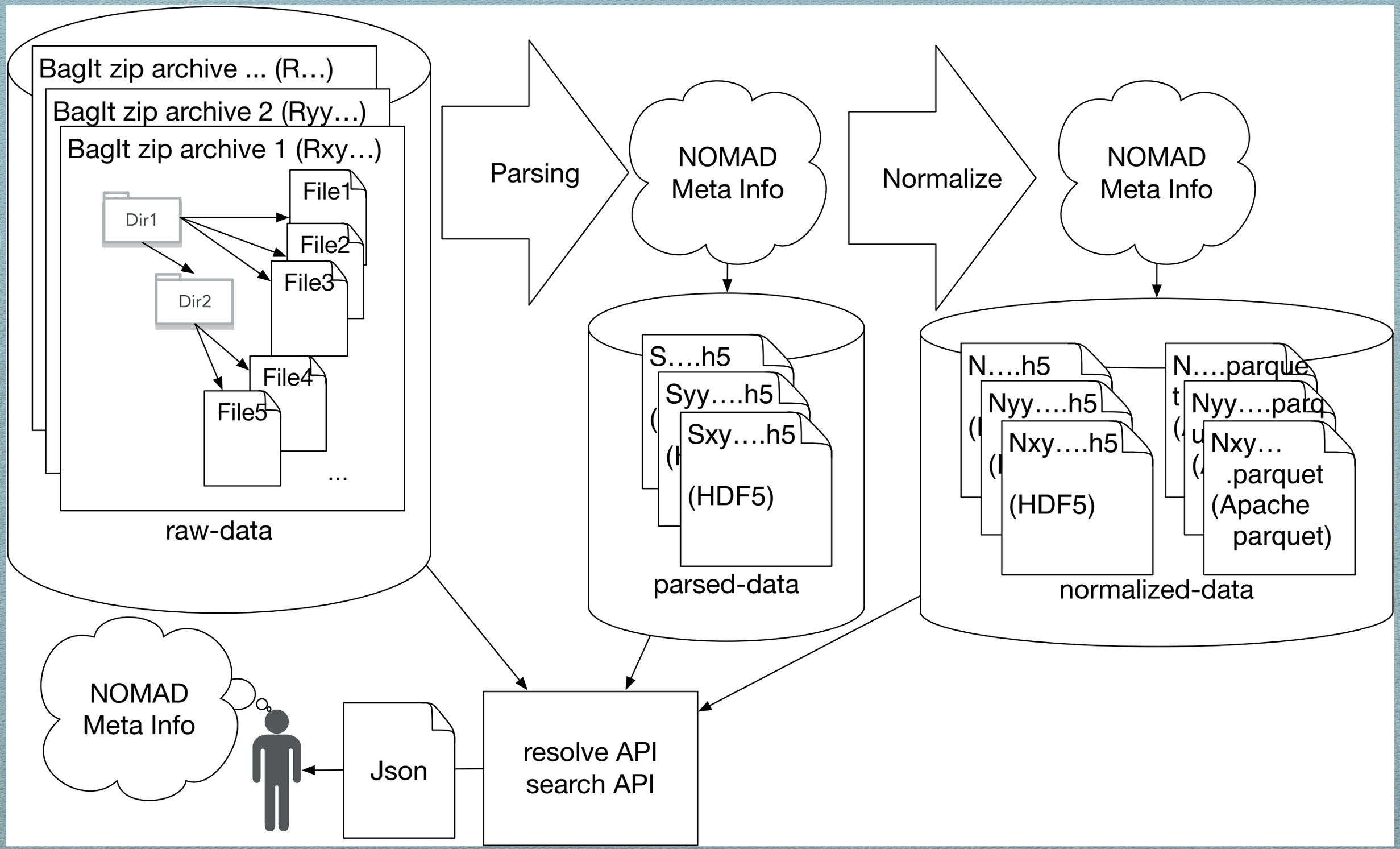








Parsing Workflow



A posteriori

- Difficult to define without being grounded (a priori)
- MetaInfo definitions stems from data in calculation codes

Defining the MetaInfo

- https://gitlab.mpcdf.mpg.de/nomad-lab/nomad-meta-info

- 4 types:
  - Conrete Value
  - Section
  - Dimension
  - Abstract types

```
{
  "meta_name": "program_name",
  "meta_description": "Specifies the name of
  "meta_type": "value_type",
  "meta_data_type": "string",
  "meta_super_section": "section_run",
  "meta_super_types": ["program_info"]
}
```

Open World

- extensible format also from outsiders
- easy to add data
- focus data description first, not validation
- describe data and peer it to its context

Meta Info browser

https://metainfo.nomad-coe.eu/ui/index.html

NOMAD Meta Info

Select Version: common only

Search by name or description: Search

Select Parent Section: Any Section

Select Abstract Type: Any Abstract Type

Select Type: Any Meta Info Type

- accessory_info
- atom_forces
- atom_forces_free
- atom_forces_free_raw
- atom_forces_raw
- atom_forces_T0
- atom_forces_T0_raw
- atom_forces_type
- atom_in_molecule_charge
- atom_in_molecule_name
- atom_in_molecule_to_atom_type_ref
- atom_labels
- atom_positions
- atom_projected_dos_energies
- atom_projected_dos_lm
- atom_projected_dos_m_kind
- atom_projected_dos_values_lm
- atom_projected_dos_values_total
- atom_to_molecule
- atom_type_charge
- atom_type_mass
- atom_type_name
- atom_velocities
- atomic_multipole_kind
- atomic_multipole_lm
- atomic_multipole_m_kind

p7A1nSnZVsRVByfq_gjtrvf7pV94S
type_section

section_single_configuration_calculation

Type: Section

Description: Every *section_single_configuration_calculation* section contains the values computed during a *single configuration calculation*, i.e. a calculation performed on a given configuration of the system (as defined in section_system) and a given computational method (e.g., exchange-correlation method, basis sets, as defined in section_method).

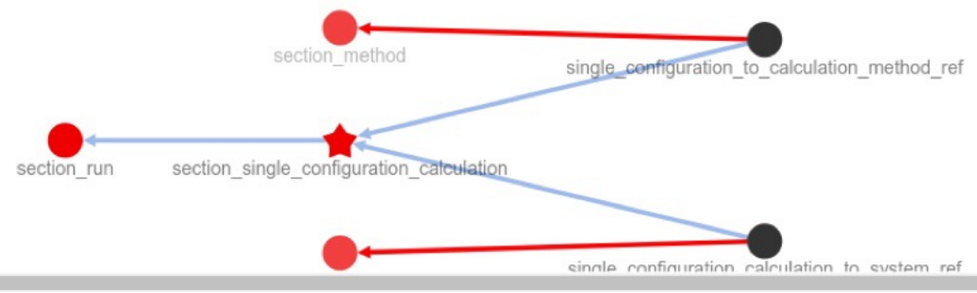
The link between the current *section_single_configuration_calculation* and the related section_system and section_method sections is established by the values stored in single_configuration_calculation_to_system_ref and single_configuration_to_calculation_method_ref , respectively.

The reason why information on the system configuration and computational method is stored separately is that several *single configuration calculations* can be performed on the same system configuration, viz. several system configurations can be evaluated with the same computational method. This storage strategy avoids redundancies.

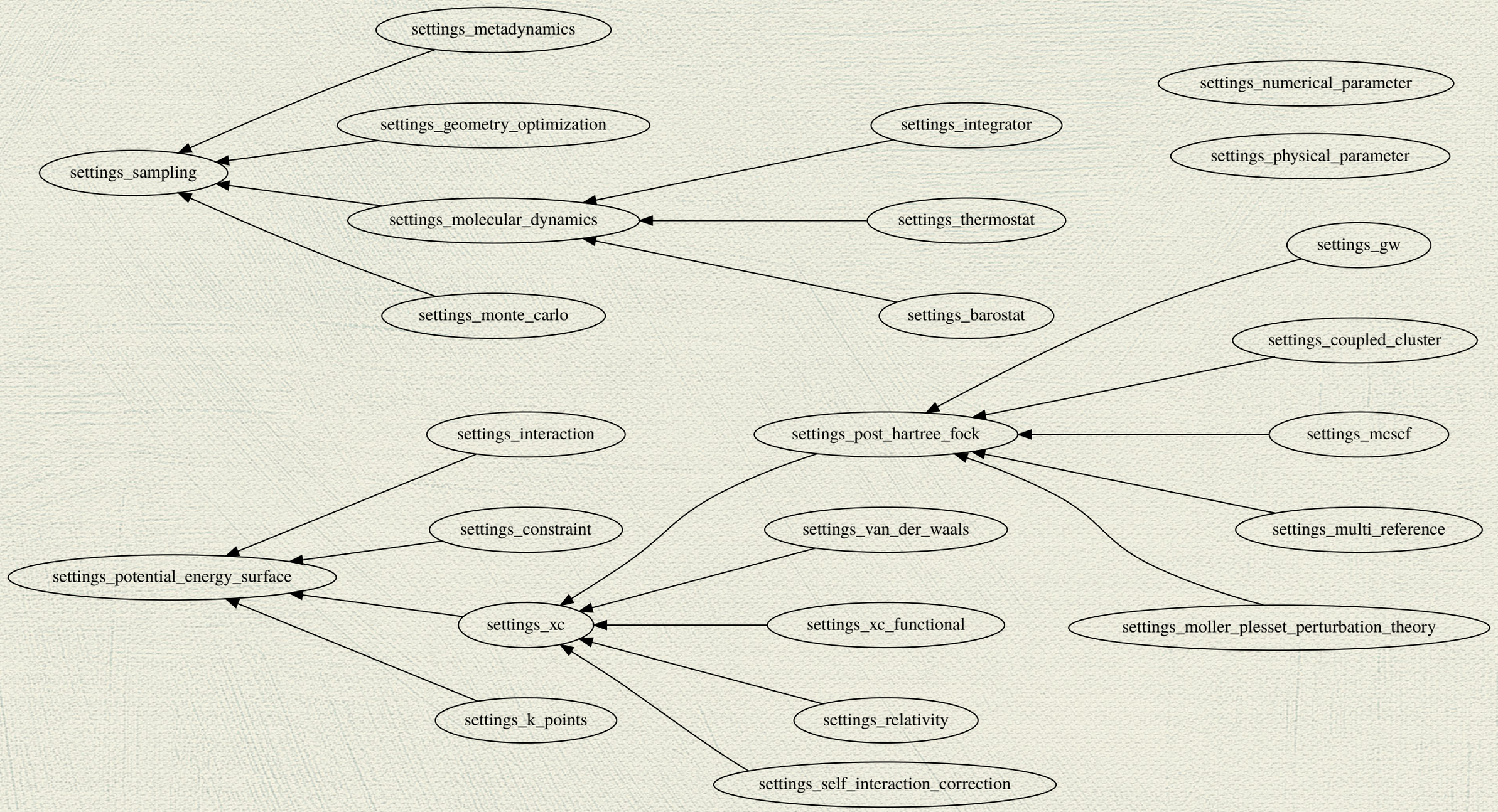
Section Abstract Type Concrete Value Dimension

Enable zoom and panning

Reset view



Abstract Types



section_run
Represents a run of a simulation script or program

section_frame_sequence
represents a sequence of frames (single configuration calculations) exploring the potential energy surface
- time, statistics

section_user_quantity_stats
Statistics on a user defined quantity

0..

frame_sequence_to_sampling_ref
0..
1

frame_sequence_to_frames_ref
0..1
0..

user_quantity_stats_to_user_quantity_definition
1
1

section_sampling
Description of how the potential energy surface is explored
- ensemble type, sampling method,…

section_single_configuration_calculation-
energy, pressure, temperature,...

section_atomic_forces
Forces on the atoms

section_user_quantity
Value of a user defined quantity

section_user_quantity_definition
Name and definition of a user defined quantity

user_quantity_to_user_quantity_definition
0..
1

single_configuration_calculation_system_ref
0..
1

single_configuration_calculation_to_constraints_ref
0..
1

single_configuration_calculation_to_method_ref
0..
1

section_system
Descripion of the system to be simulated
- atom position, simulation cell,...

system_to_topology_ref
0..
1

section_constraints
- thermostat, barostat, fixes

section_constraint

constraints_to_topology
0..
1

section_method
Potential energy surface definition
- theory level, forcefield parameters, numerical accuracy thresholds

section_interation
defines a force field interaction

method_to_topology
0..
1

section_topology
Topology of the system
- atom label, mass, charge, topology

section_atom_type

section_atom_group

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

Values: Data
Structure and names: Meta info

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

Workflows?

- References between calculations and runs:

- section_calculation_to_calculation_refs and calculation_to_calculation_kind...

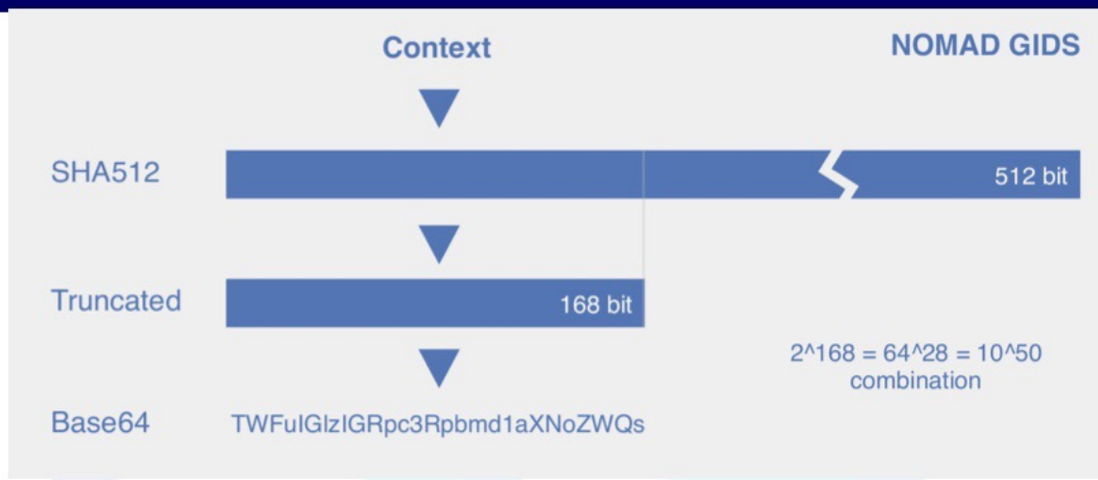
Detection challenges

- checksums to know if the same configuration and method are used
- connect files with the same run -id
- keep enough information
- extract workflow info when available

NOMAD URIs: Data Identifiers

nmd://NWGq3on-jgxo8pw5cr-nEclMNqYW2/
C-Gl21uIOi_OQQQo8NSKu3FRHSH5m/
section_run/0c/section_system/0c/atom_positions/0c

- All pieces of data, organized using the NOMAD metadata, are uniquely identified using their path in the hierarchical structure
- In NOMAD, we use an unique identifier (a checksum using only the uploaded data) to identify calculations and archives
- Combining those two we obtain NOMAD URIs that allow us to uniquely identify any piece of data, e.g., through a web browser
- With them, one can build collections of data, or access single values using our REST API



NOMAD API
NOVEL MATERIALS DISCOVERY

- https://analytics-toolkit.nomad-coe.eu/nomad-query-gui
- Resolve
  - Nomad uri → data values
- Query
  - Metadata based query → nomad uris (data values)

NOMAD
NOVEL MATERIALS DISCOVERY

Usefulness of Metadata

- Like a language
- more used, more data, more tools → more useful
- Being simple increases the usefulness
- Constraints can make it more broadly usable or easier to use in different contexts

Container Manager

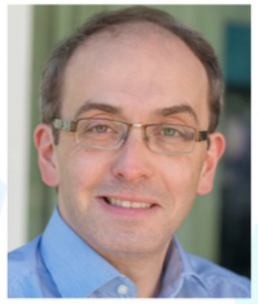
- Make arbitrary web applications running in a separate Docker for each user transparently available
  - Authenticate user
  - Start user specific docker if required
  - Forward all user requests to "his”container
- http://kitabi.eu/nomad/ContainerManager/

Infrastructure

- Low level infrastructure
- Higher level infrastructure
- Documentation
- Repository
- Parser development

(Yazid Hamdi)

Jungho Shin
HUB

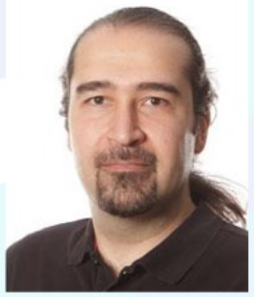


Nicolas Fabas
MPCDF

(Ankit Kariryaa)
FHI

(Alfonso Sastre)
FHI

Fawzi Mohamed
FHI



Berk Onat
Warwick

Markus
Scheidgen
HUB

Arvid Ihrig
FHI

Thomas Zastrow
MPCDF

Aknowledgments

- Matthias Scheffler, Luca Ghiringhelli and the whole NOMAD Team

- You for your attention

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