NOMAD Metadata for all

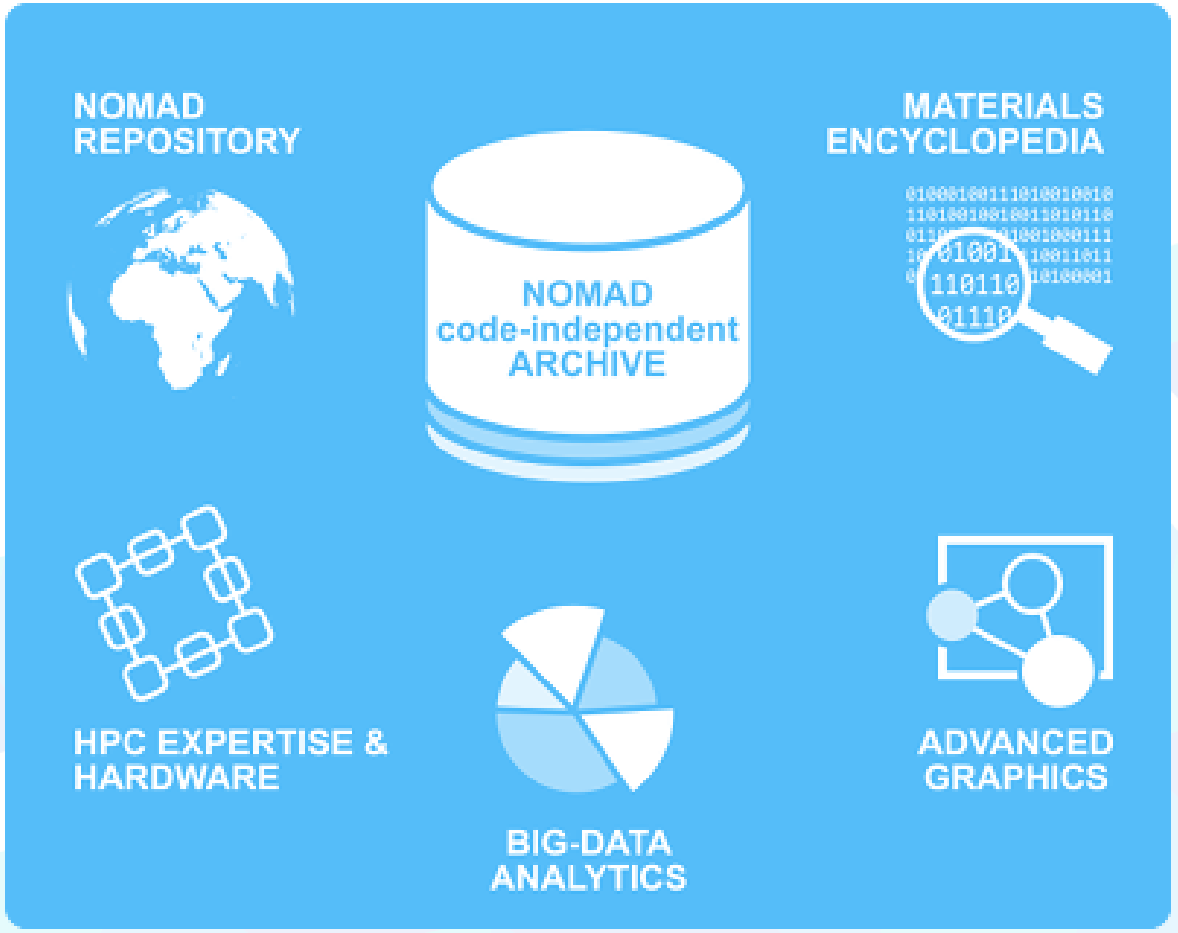
Cambridge, 8 Nov 2017

Fawzi Mohamed FHI Berlin

NOMAD
NOVEL MATERIALS DISCOVERY

NOMAD Center of excellence goals

- 200,000 materials “known” to exist
- basic properties 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

NOMAD
NOVEL MATERIALS DISCOVERY

NOMAD philosophy

- Scientific openess
- 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

NOMAD
NOVEL MATERIALS DISCOVERY

NOMAD philosophy

- Scientific openess
- 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 Metadata: our conceptual model

- defines how the data that we extract is organized, and what it is
- important both for human and for the machine

- Can induce an ontology, but is not an ontology

NOMAD
NOVEL MATERIALS DISCOVERY

Common File Formats & APIs

- Files are a well known and proven technology
- Backup & batch processing of files is commonly done
- File transfer and synch is well undestood

- API must return data
- Web APIs often use JSON, but how to structure the JSON is still non obvious for complex data
- Large transfers via API often difficult

- Common File Formats have still a place with Big Data, Databases and APIs

NOMAD
NOVEL MATERIALS DISCOVERY

File formats Trade-offs

Format	xyz	dcd	cml	cif	xml	json	hdf5	parquet
Human Readable	+	−	+	0	+	+	−	−
No Dependencies Read	+	−	0	−	0	+	−	−
Space Efficient	−	+	−	−	−	−	+	+
Extensible	−	−	+	+	+	+	+	+
Simple	+	+	−	−	−	0	−	−
Future Readability	+	0	0	+	+	+	+	−

NOMAD
NOVEL MATERIALS DISCOVERY

Common File Formats

Pros

+ Clear definition
+ Simple exchange
+ Implicit relationships

Contras

- Trade off cannot be changed
- Different applications have different needs
- In memory/DB storage

NOMAD
NOVEL MATERIALS DISCOVERY

Overcoming the limits of File storage

- Abstract away the actual storage method

NOMAD
NOVEL MATERIALS DISCOVERY

XML Schemas

+ Describe how a document can be structured
+ Well formalized

- Tend to overstructure
- XML oriented

Examples:
- Initial CML
- NIST proposal for Materials Genome
- Some code outputs

NOMAD
NOVEL MATERIALS DISCOVERY

Metadata Dictionaries & Informal Definitions

+ Avoids over structuring
+ Reusable in different schematas

- Requires extra information to be used in practice
- Inaccuracies can limit its usefulness

Examples:
- CML dictionaries
- EMMC metadata

NOMAD
NOVEL MATERIALS DISCOVERY

Metadata & Structure

- NOMAD Metadata
  - http://nomad-coe.eu/
  - Open git repository
- Purposefully limited expressivity
  - data: float, int, booleans, strings, section references and multidimensional arrays of them
  - Grouping (sections), with hierarchical structure
  - Classification (abstract types)
- HDF5, parquet and JSON storage
  - Very different trade offs

NOMAD
NOVEL MATERIALS DISCOVERY

NOMAD Metadata

- 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

Sample NOMAD Data

NOMAD
NOVEL MATERIALS DISCOVERY

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*
  - ...

Values: Data
Structures and names: Meta data

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

Nested sections (hierarchical tree structure)
References to other sections

Defining the NOMAD Metadata

- https://gitlab.mpcdf.mpg.de/nomad-lab/nomad-meta-info
- Description: https://www.nomad-coe.eu/the-project/nomad-archive/archive-meta-info-guide
- 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"]
}
```

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

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).

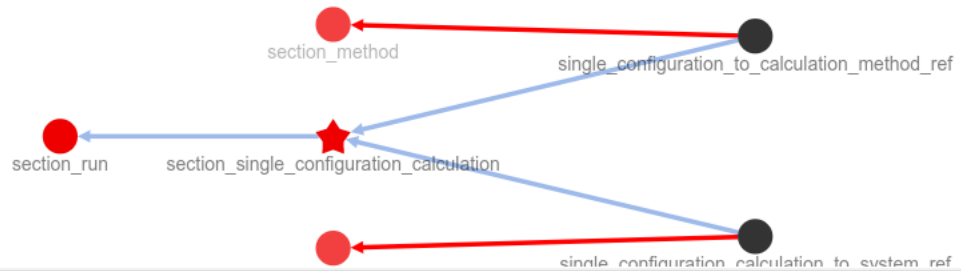
The link between the current *section_single_configuration_calculation* and the related section_system and section_method sections is established by the values stored in single_configuration_calculation_to_system_ref and single_configuration_to_calculation_method_ref , respectively.

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



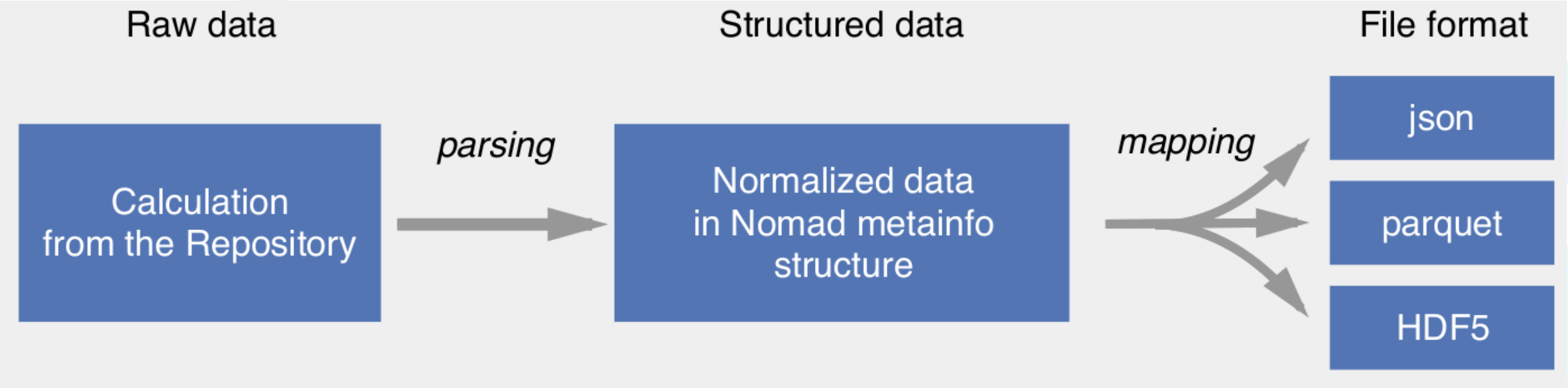
NOMAD
NOVEL MATERIALS DISCOVERY

Storing and Exchanging data

- With the NOMAD Metadata we have a conceptual model, flexible enough to be useful and support several file formats:
- JSON: Human readable & web ready
- HDF5: indexed and optimized for multidimensional arrays
- Parquet: a modern format for Big-Data storage (immutable and efficient columnar storage for hierarchical data)

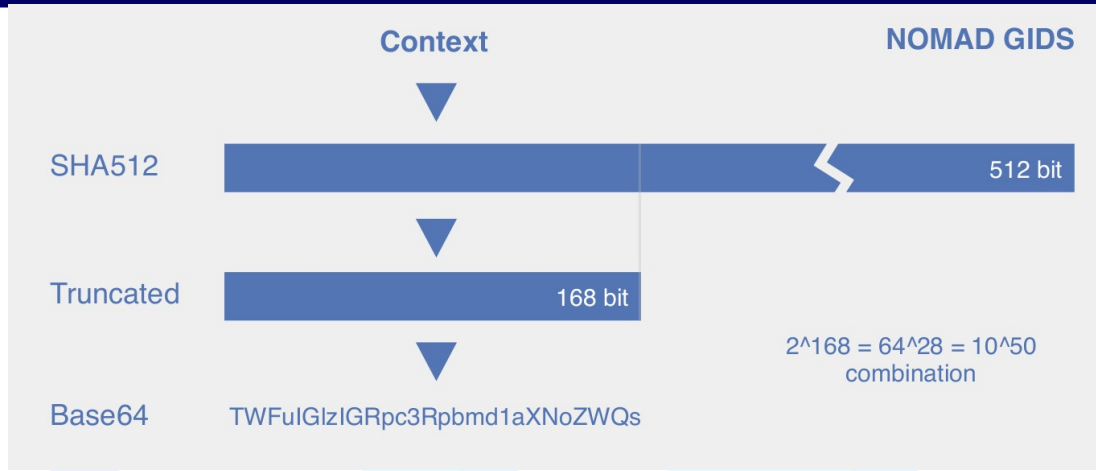
NOMAD
NOVEL MATERIALS DISCOVERY

Storing and Exchanging data



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

- 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

NOMAD
NOVEL MATERIALS DISCOVERY

NOMAD Meta Data: what we have

- A single framework to describe computational material-science data
- Open and versioned machine-readable definition on gitlab hosted at MPCDF https://gitlab.mpcdf.mpg.de/nomad-lab/nomad-meta-info
- Browsable definitions for users https://metainfo.nomad-coe.eu/ui/index.html
- enough structure and definitions to allow meaningful use while being file-format agnostic
- A good calculation-oriented view of *ab initio* calculation data
- Unique identifiers (NOMAD URI) for every piece of data
- API to access data, both in bulk, or any specific subset identified by a NOMAD URI (e.g., via a web browser)

NOMAD Meta Data: what we need to do

- Further extend the calculation-oriented view (a lot is there, but the field is large)
- Improve usefulness of using the NOMAD meta data
  - with better services in NOMAD
  - improving the collaboration with others (ecosystem effect)
- Integrate better with EMMC ontology (if needed)

NOMAD
NOVEL MATERIALS DISCOVERY

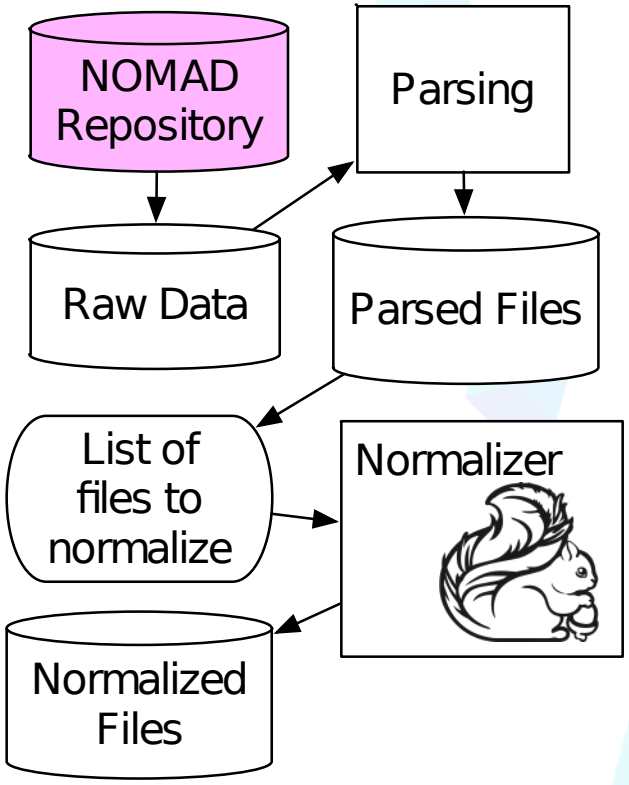
Aknowledgments

- Matthias Scheffler, Luca Ghiringhelli and the whole NOMAD Team

- You for your attention

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

Data Processing



Parsing

- Parsers map input and output files to our metadata representaion
- Should 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

Metadata For Computational Material Science

- Purposes of metadata
  - Describe and organize data
  - Keep related data pieces connected (e.g., calculation method and system geometry)
- Two main points of view can be taken:
- Starting from calculations (*a posteriori*)
  - Flexible and efficient when parsing row inputs/outputs from materials-science codes.
  - Immediately tried to build a code-independent representation
  - Describe also code-specific settings to understand the kind of calculation performed
- Starting from materials (*a priori*)
  - Material → property association is what one is after at the end
  - Easy within a consistent curated dataset
  - For comparing datasets, one needs exact information about calculation method
  - For the latter purpose, the calculation view can be very useful