NOMAD: Data & Pipelines

14.02.2017

Fawzi Mohamed FHI Berlin

NOMAD
NOVEL MATERIALS DISCOVERY

Data & Pipelines

- Raw Data
  - Zip archives (Bag it format)
  - Text or binary data in various formats
- Parsed Data
  - Json or HDF5
  - Metadata representation
- Normalized Data
  - (Json) HDF5 (+ Parquet?)
  - Metadata representation

- Parsing
  - Raw Data → Parsed Data
- Normalization
  - Parsed Data → Normalized Data
- Query
  - Normalized Data → Query result

NOMAD
NOVEL MATERIALS DISCOVERY

Metadata

- ~dictionary, ontology, schema
- Unique name for each entry
- Sections
  - Collect values that belong together (Dictionary)
  - Nested (form a tree)
- Concrete values
  - Multidimensional array of
  - Boolean, string, int/long, real/float
  - Or reference to a section (relational model)

MetaData

NOMAD
NOVEL MATERIALS DISCOVERY

Values: Data
Structure 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

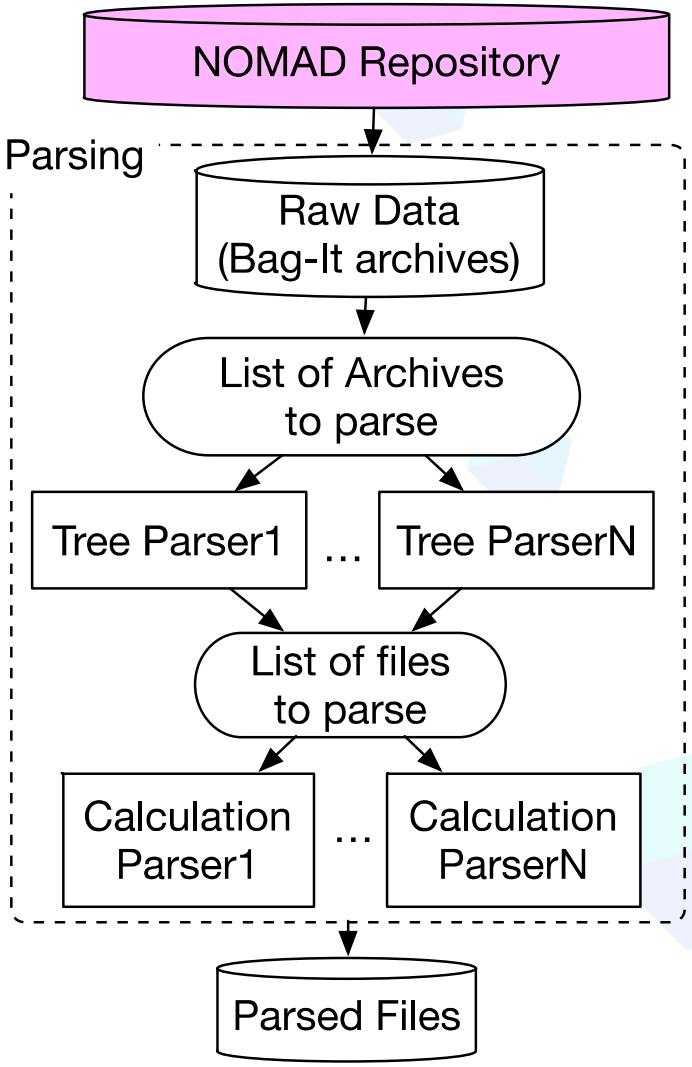
Parse Events

- Open Section
- Close Section
- Add Value

- Partially ordered
- Fully ordered

NOMAD
NOVEL MATERIALS DISCOVERY

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

Parser Developers


Mikkel Strange


Micael Oliviera

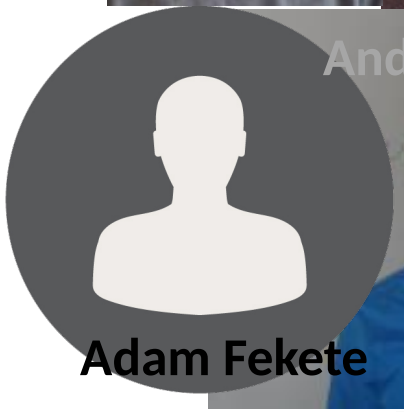

Martina Stella

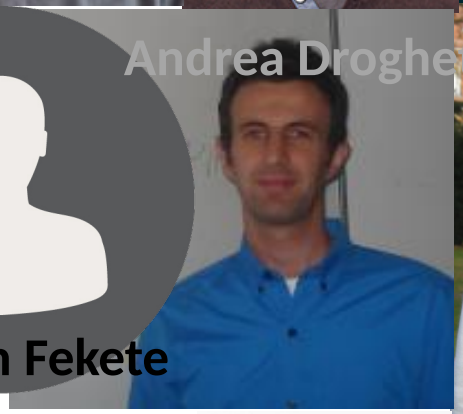

Massimo Riello


Wael Chibiani


Franz Knuth


Aliaksei Mazheika


Adam Fekete

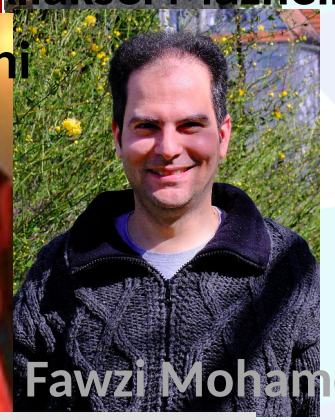

Andrea Droghetti


Honghui Shang


Sebastián Alarcón Villaseca


Daria Tomecka


Lorenzo Pardini


Fawzi Mohamed


Carl Poelking


Adriel Dominiguez

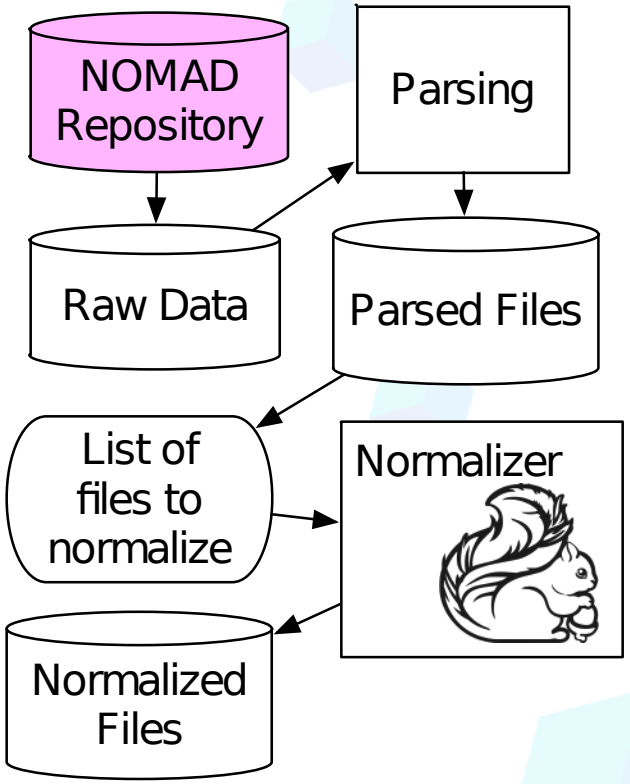

Rosenndo Valero


Ask Hjorth Larsen


Lauri Himanneh


Sami K. Kivisto


Henning Glawe



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

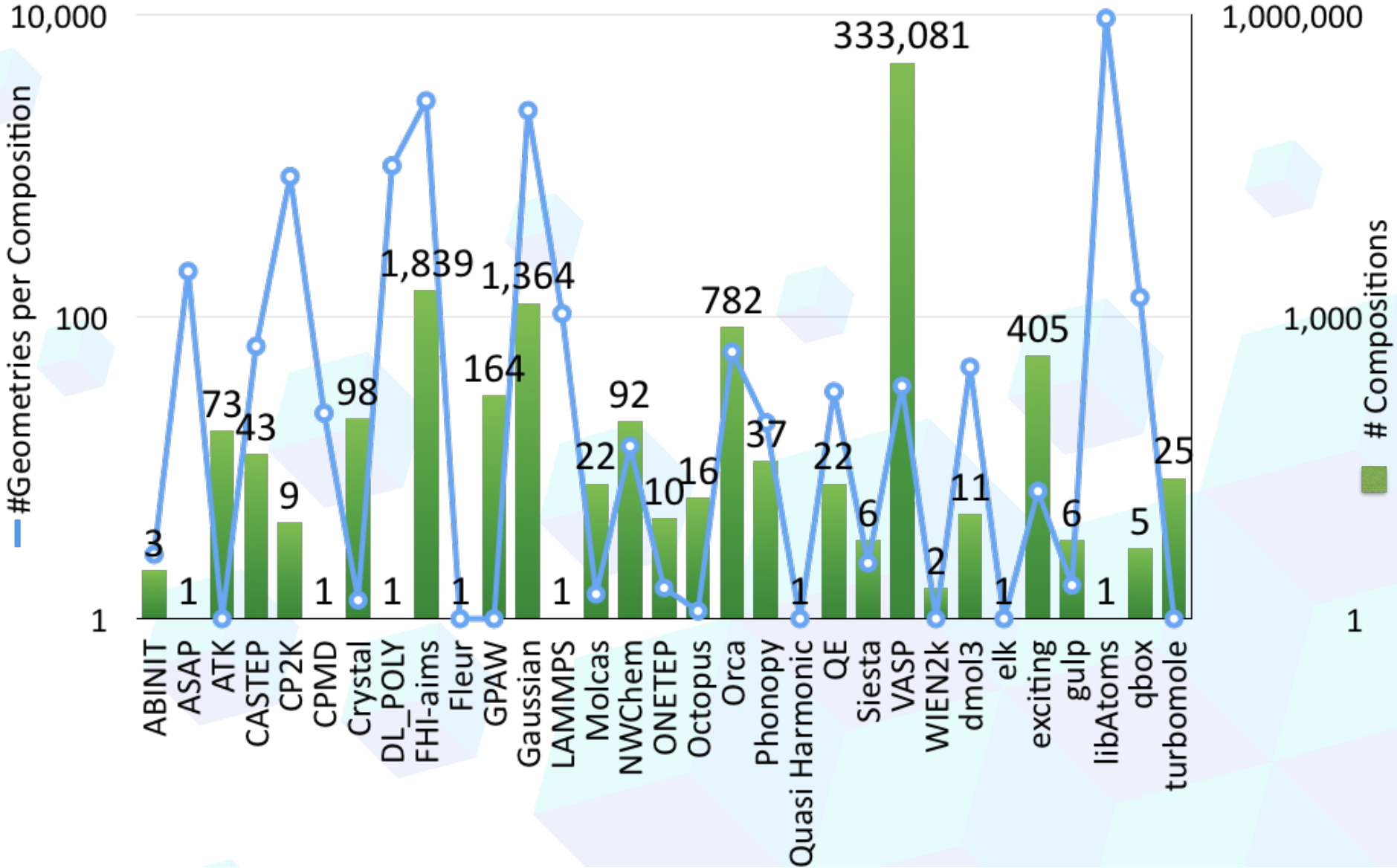
- Normalization is an infrastructure to apply automatically some transformations and store their result along with the parsed data

NOMAD Archive in Numbers

generated parsing **3761** Zip Archives
corresponding to **8 TeraBytes** of compressed data.
Normally the content of the zip archives are repetitive text files
with excellent compression ratio **60-80%**.
This produces **2.7 TeraBytes** of parsed Hdf5 files
with data classified using **279** public metadata
1736 code specific metadata

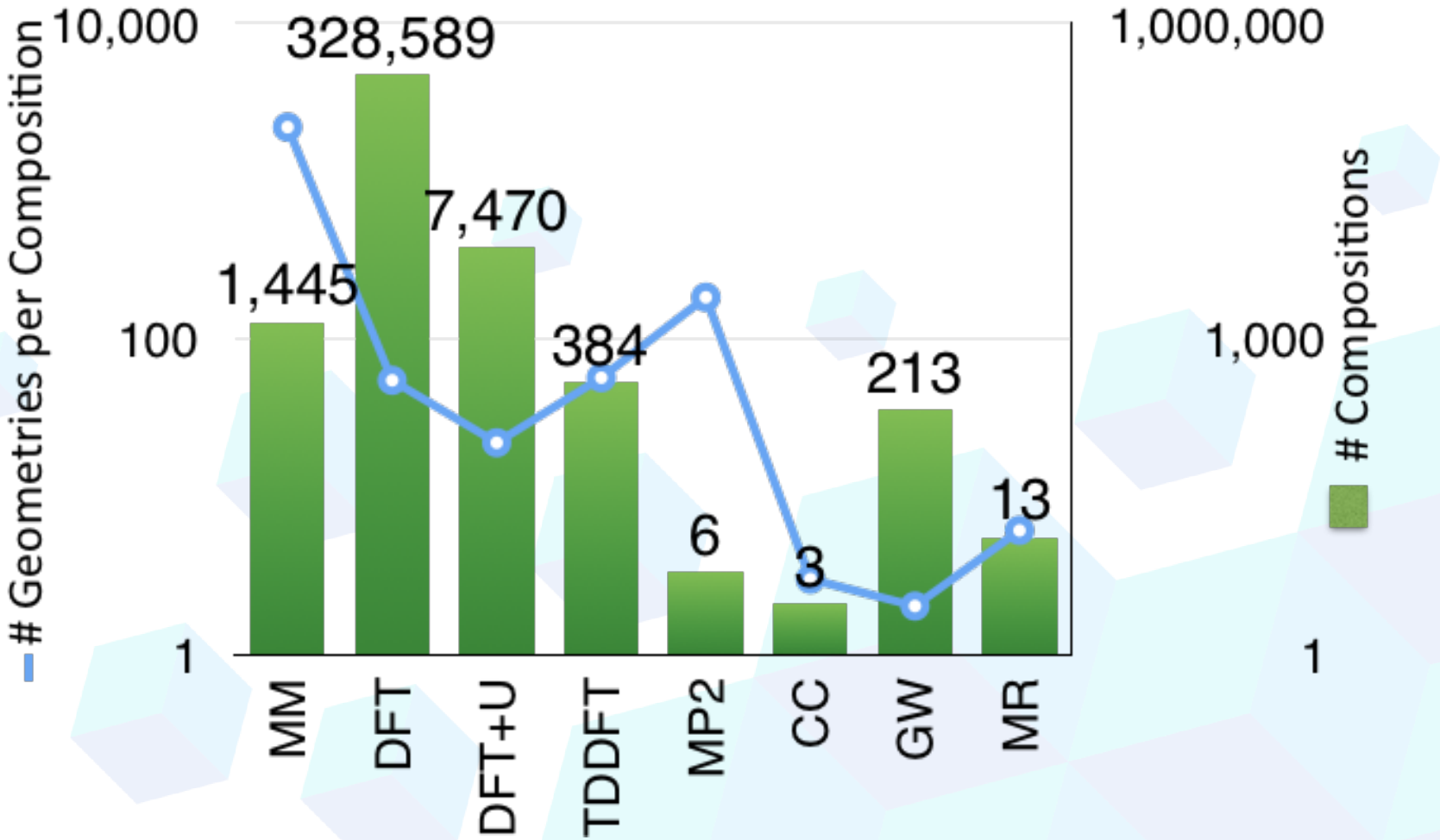
NOMAD

NOVEL MATERIALS DISCOVERY



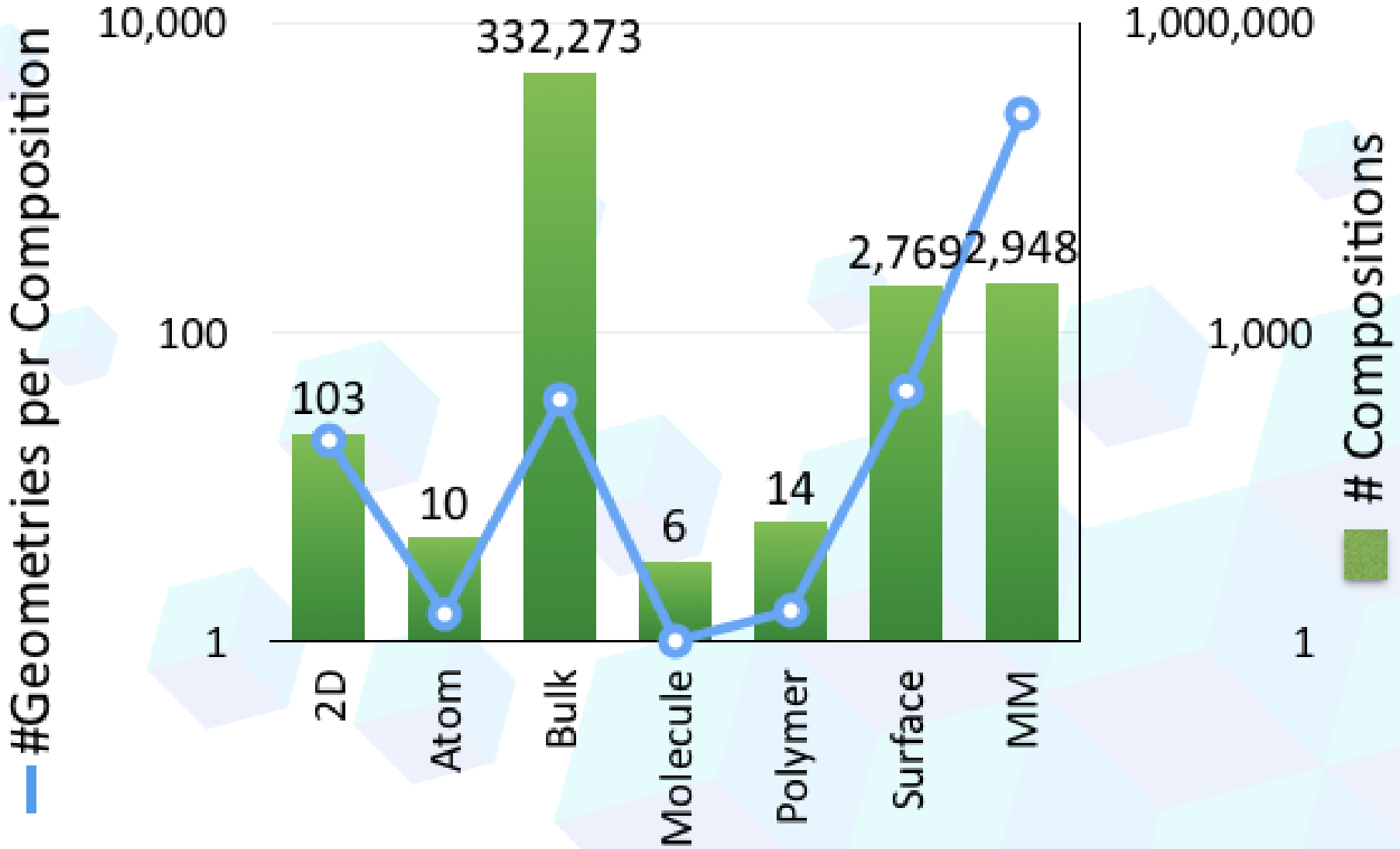
NOMAD

NOVEL MATERIALS DISCOVERY



NOMAD

NOVEL MATERIALS DISCOVERY



NOMAD
NOVEL MATERIALS DISCOVERY

Next Steps

- HDF5 python interface
- Support when analysis is preformed (Bug fixing)
- 10 extra Parsers April 2018 (M30)
- James Kermode (TINKER, NAMD, AMBER, Gromos, Gromacs, CHARMM)
- Aiida
- LM Suite (TB-LMTO-ASA) (KCL?),
- WEST

NOMAD
NOVEL MATERIALS DISCOVERY

NOMAD Metadata

Metadata:

279 public metadata

1736 code specific metadata

ESCDF
HJSON
NIST, EMMC

OPTIMADE REST API

Normalization

NOMAD
NOVEL MATERIALS DISCOVERY

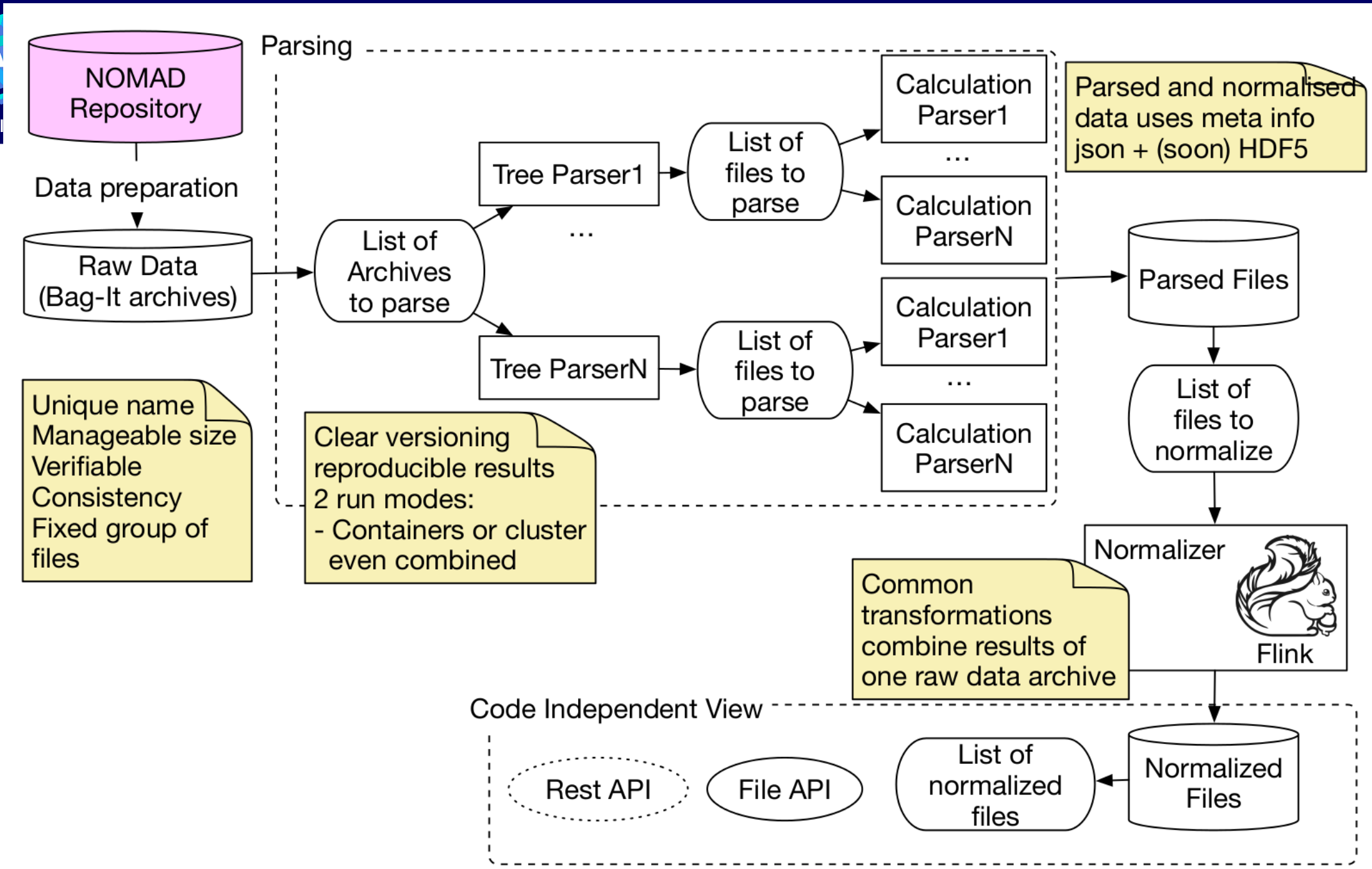
Normalization

- Messy part, might have dependencies

  Example:
- Checksum for geometry
- Connection to repository
- Statistics

- Loop on archives/calculations/single point/goemetries
- Apply transformation
- Store result

- OpenContext/closeContext event


NOMAD Repository
Data preparation
Raw Data (Bag-It archives)
Parsing
List of Archives to parse
Tree Parser1
...
Tree ParserN
List of files to parse
List of files to parse
Calculation Parser1
...
Calculation ParserN
Calculation Parser1
...
Calculation ParserN
Parsed and normalised data uses meta info json + (soon) HDF5
Parsed Files
List of files to normalize
Normalizer
Flink
Unique name
Manageable size
Verifiable
Consistency
Fixed group of files
Clear versioning
reproducible results
2 run modes:
- Containers or cluster
even combined
Common transformations combine results of one raw data archive
Code Independent View
Rest API
File API
List of normalized files
Normalized Files

NOMAD
NOVEL MATERIALS DISCOVERY

Query

- Read only
- Loop on data, select subset
- Return nomad uri + data table