Common Data

Leiden, 25. October 2016

Matthias Scheffler & Fawzi Mohamed FHI Berlin

NOMAD
NOVEL MATERIALS DISCOVERY

Common Data

Why

- Sharing
- Tools

Issues

- No perfect solutions
- Trade-offs are required

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 obvius 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

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

- Aflow, OQMD, CCCBDB did share their raw input and outputs through NOMAD Repository
- Open sharing is one of the reasons for this meeting

NOMAD
NOVEL MATERIALS DISCOVERY

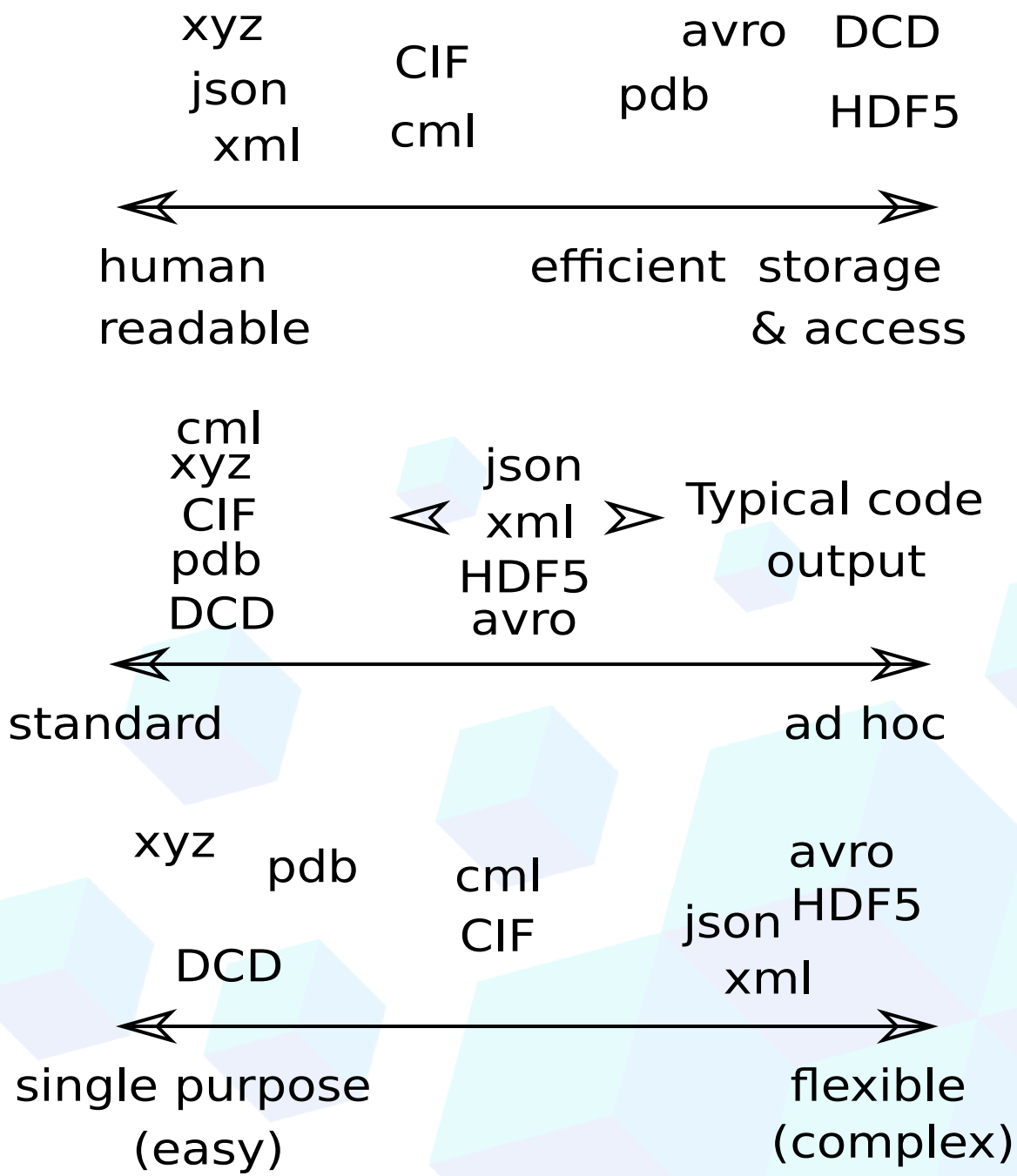
Common File Formats

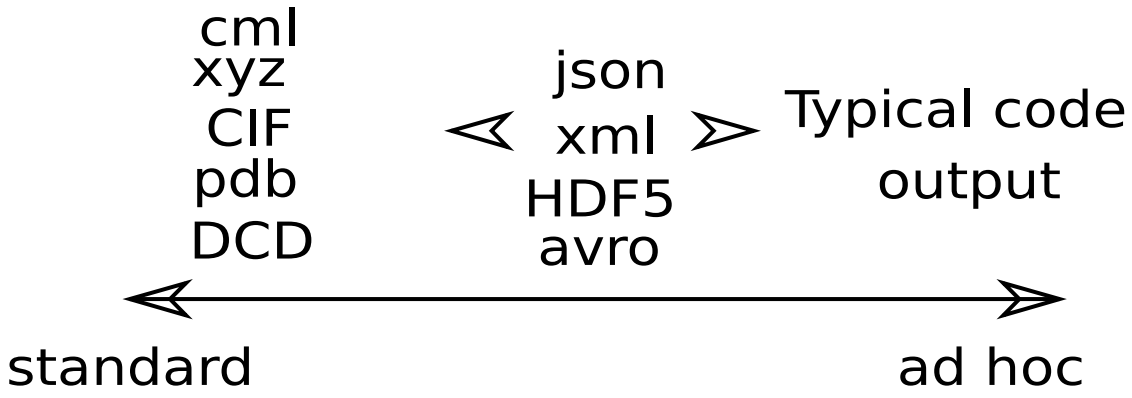
Pros

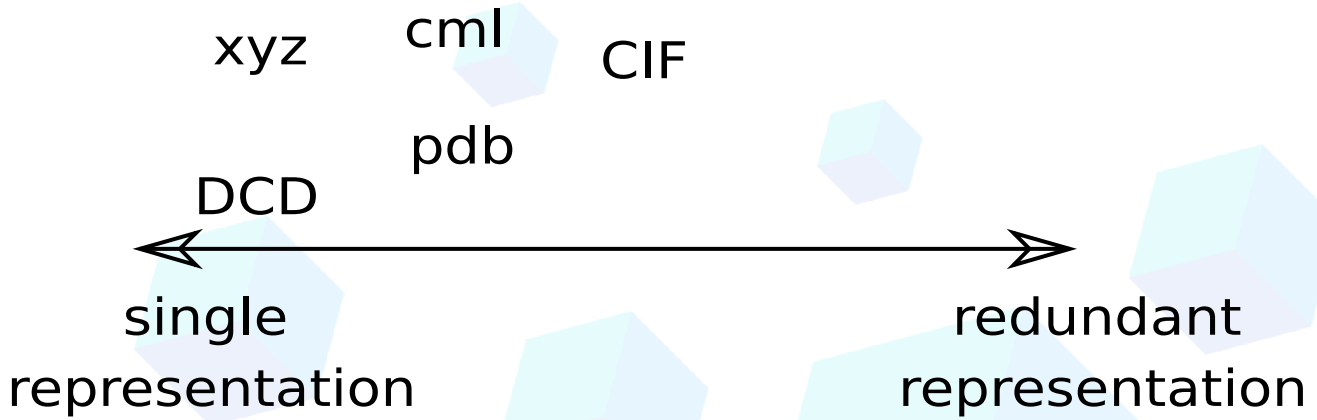
+ Clear definition
+ Simple exchange
+ Implicit relationships

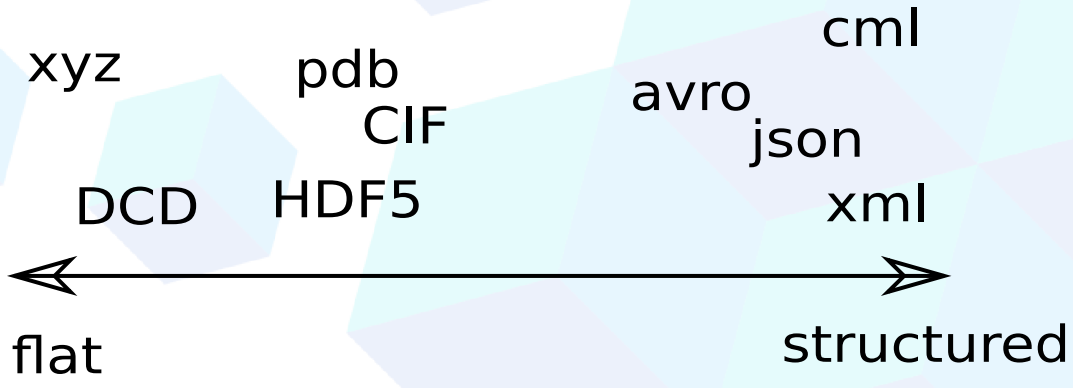
Contras

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


xyz
json
xml
CIF
cml
avro
pdb
DCD
HDF5
human readable
efficient storage & access
cml
xyz
CIF
pdb
DCD
json
xml
HDF5
avro
Typical code output
standard
ad hoc
xyz
pdb
DCD
cml
CIF
avro
HDF5
json
xml
single purpose (easy)
flexible (complex)


cml
xyz
CIF
pdb
DCD
json
xml
HDF5
avro
Typical code output
standard
ad hoc


xyz
cml
CIF
pdb
DCD
single representation
redundant representation


xyz
pdb
CIF
DCD
HDF5
cml
avro
json
xml
flat
structured

NOMAD
NOVEL MATERIALS DISCOVERY

Relational Databases

- One true view
- Very difficult to merge
- Sharing of them is good to be more open (OQMD):
  - alleviates the problem that APIs normally constrain usage, and make bulk download difficult
- Unfortunately is not way to build a robust ecosystem, or help derived work based on them (due to the first two points)

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 and JSON storage
  - Very different trade offs

NOMAD
NOVEL MATERIALS DISCOVERY

Miscellaneous

- CIF, tables, relational model, some similarities, structure modeled indirectly

- Aflow: Website, API, clear identifiers, raw data, (directly and through NOMAD)
- OQMD: Website, DB dumps, raw data (NOMAD)
- CCCBDB: Website, raw data through (NOMAD)
- Materials project: Website, API

NOMAD
NOVEL MATERIALS DISCOVERY

Common Data, discussion

- Should standarize the metadata?
- How?
- Do we need new file formats?
- For which purpose?