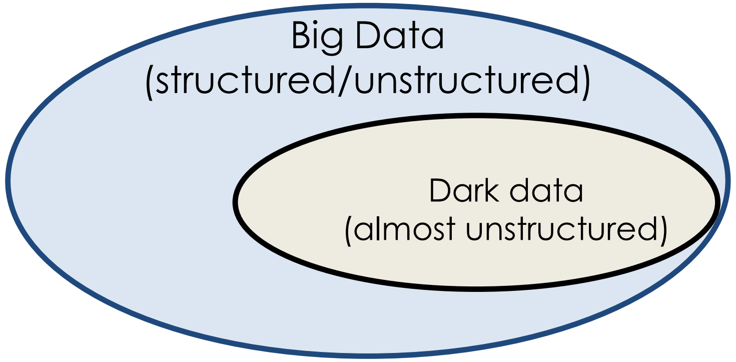
Importance of Data in Engineering Sciences - Automated Metadata Extraction

Nadiia Huskova, Thomas Bönisch

contributed by Björn Schembera
High Performance Computing Center Stuttgart (HLRS)

In engineering sciences and high-performance computing, research data management poses a number of challenges:

- need to analyze big data
- often no full understanding of generated data
- lack of relevant metadata describing the process
- lack of resources/operation costs



Dark data…

…. is automatically collected during routine activities, but is not used in any way to obtain information or make decisions.

Characteristics:

- ✓ not tagged with metadata
- ✓ no longer technically accessible
- ✓ not understandable, available or usable

Examples:

- ✓ **Log files (servers, systems, architecture, etc.)**
- ✓ Previous employee data
- ✓ Geolocation data

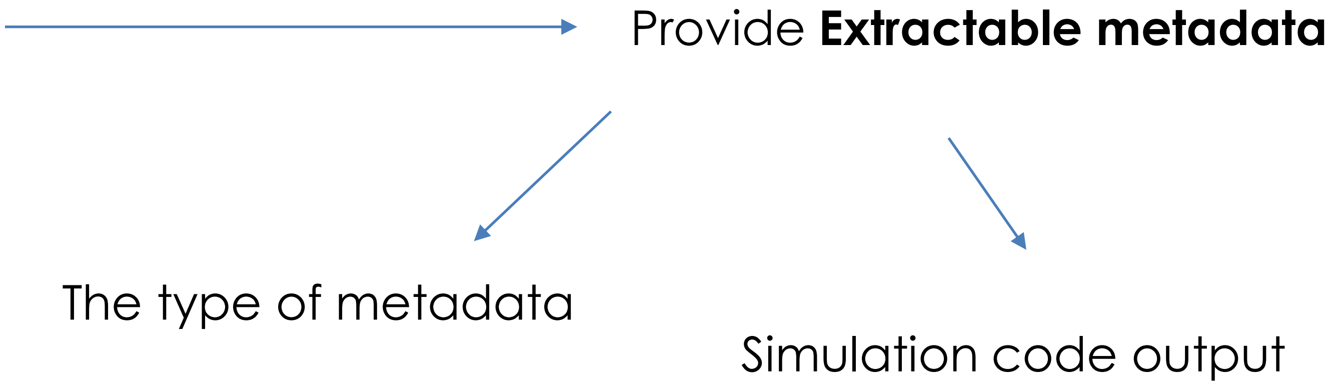
Case study: Thermodynamics

- Thermodynamics deals with computer-aided modelling, among other things of molecules and their movements.

- High-performance computers and Cluster systems used to calculate the trajectories of the molecules, generated with the simulation code GROMACS*.

Goal: receive a description of the data

- as comprehensive as possible
- enable the mechanical recording of the metadata for the individual simulation runs

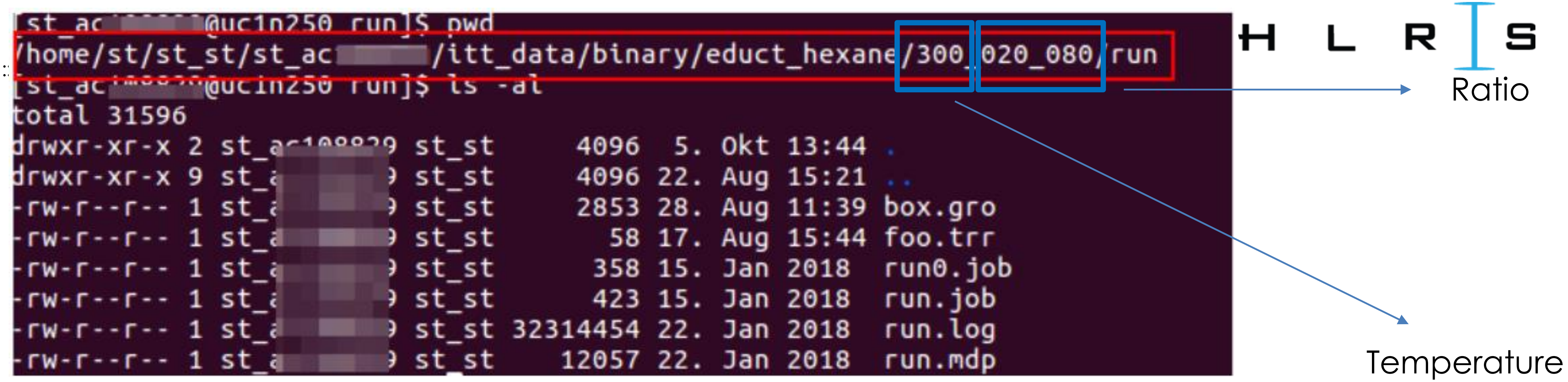


Provide **Extractable metadata**

The type of metadata

Simulation code output

For the GROMACS code, lots of processing metadata and domain-specific information is already available



```
[st_ac[redacted]@uc1n250 run]$ pwd
/home/st/st_st/st_ac[redacted]/itt_data/binary/educt_hexane/300_020_080/run
[st_ac[redacted]@uc1n250 run]$ ls -al
total 31596
drwxr-xr-x 2 st_ac108829 st_st     4096  5. Okt 13:44 .
drwxr-xr-x 9 st_a[redacted]9 st_st     4096 22. Aug 15:21 ..
-rw-r--r-- 1 st_a[redacted]9 st_st     2853 28. Aug 11:39 box.gro
-rw-r--r-- 1 st_a[redacted]9 st_st       58 17. Aug 15:44 foo.trr
-rw-r--r-- 1 st_a[redacted]9 st_st      358 15. Jan 2018  run0.job
-rw-r--r-- 1 st_a[redacted]9 st_st      423 15. Jan 2018  run.job
-rw-r--r-- 1 st_a[redacted]9 st_st 32314454 22. Jan 2018  run.log
-rw-r--r-- 1 st_a[redacted]9 st_st    12057 22. Jan 2018  run.mdp
```

Figure 1. Data organization in directory structures on filesystems. Sample from GROMACS

Example from Dr. Bjorn Schembera

A lot of (semi-structured) metadata is already available

– In job or log files of simulation codes (e.g. nodes, version)

– In non-standardized or standardized file formats (i.e. HDF5 or NetCDF)

Extracting metadata is possible and desirable

Metadata Scheme

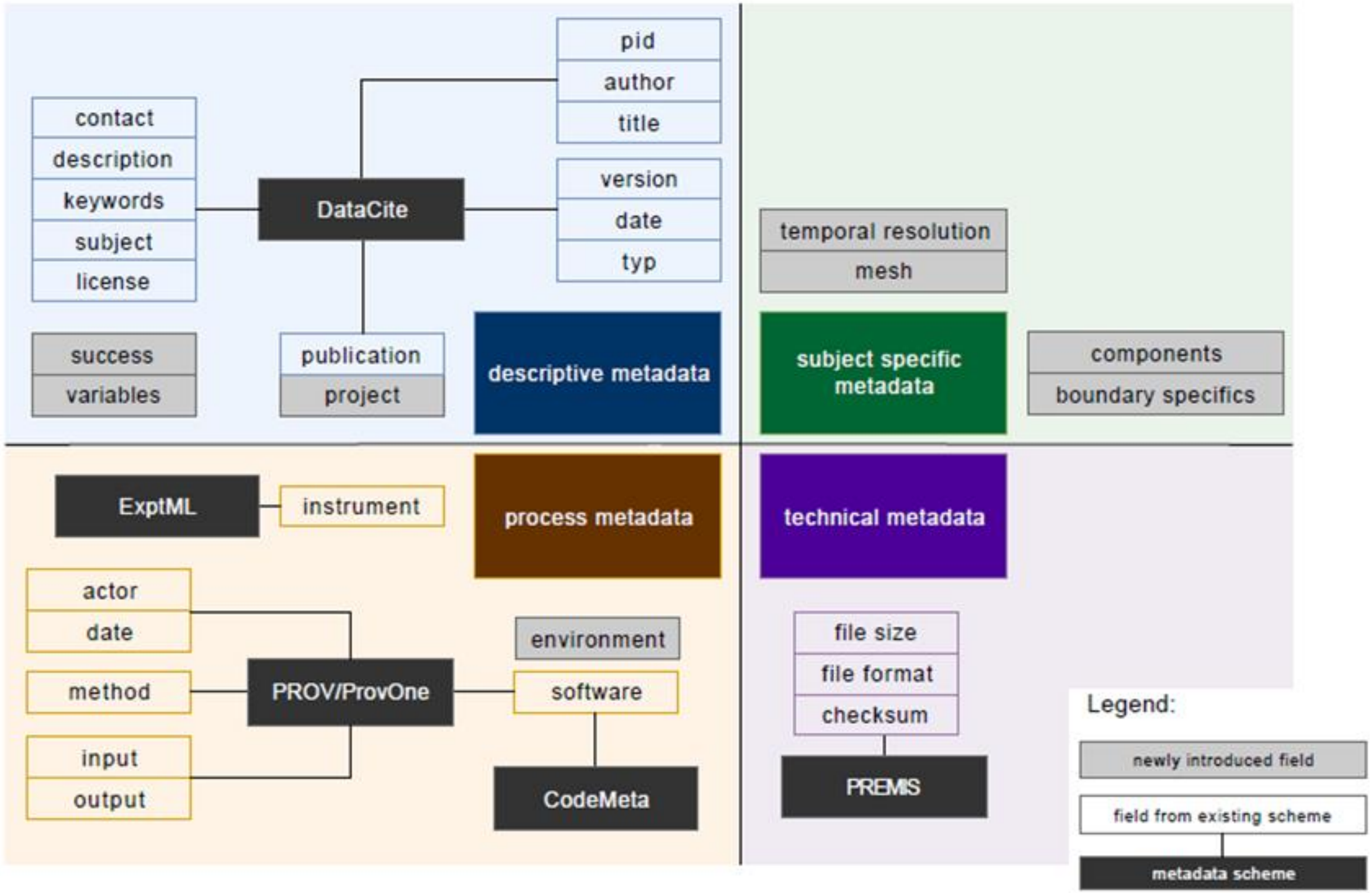
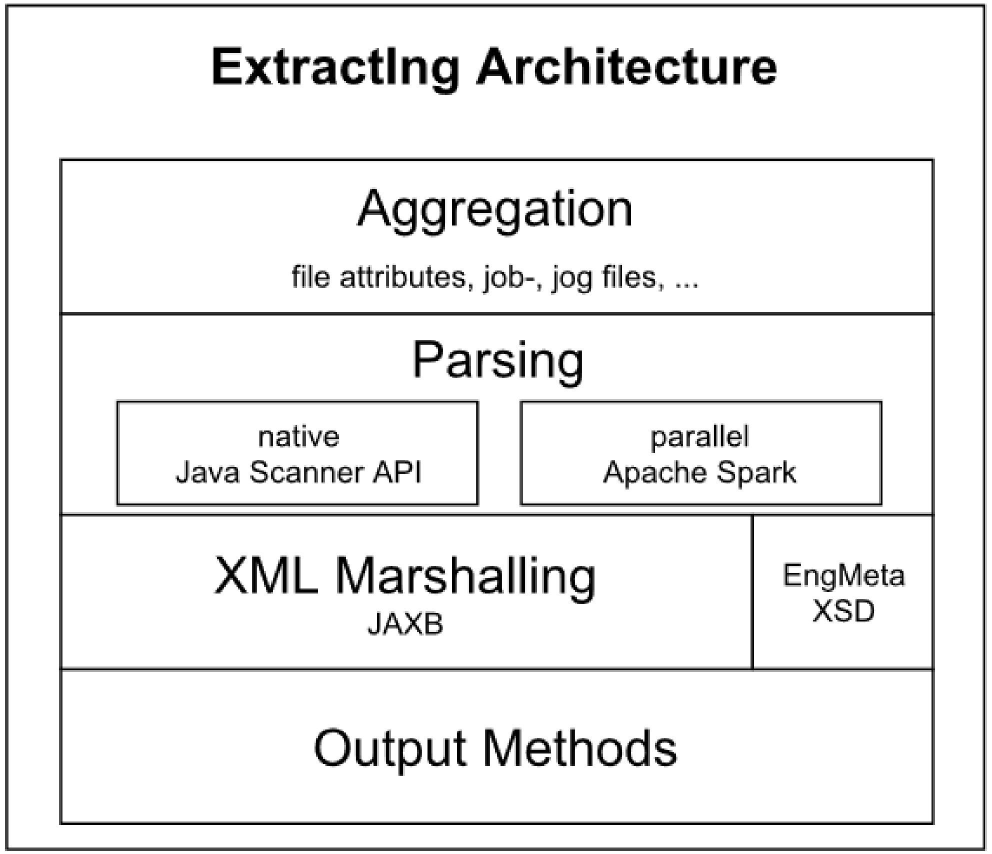


Figure: Darstellung der Metadatenblöcke. Daraus Serialisierung für XSD. [EngMeta 2018]

Automated metadata extraction

Metadata category	Extractability
Technical metadata	easy to extract automatically
Process metadata	partly available and extractable
Domain-specific metadata	partly available and extractable
Descriptive metadata	**not available/not extractable**

Automated metadata extraction - Extractor


ExtractIng Architecture

Aggregation

file attributes, job-, jog files, ...

Parsing

native
Java Scanner API

parallel
Apache Spark

XML Marshalling
JAXB

EngMeta
XSD

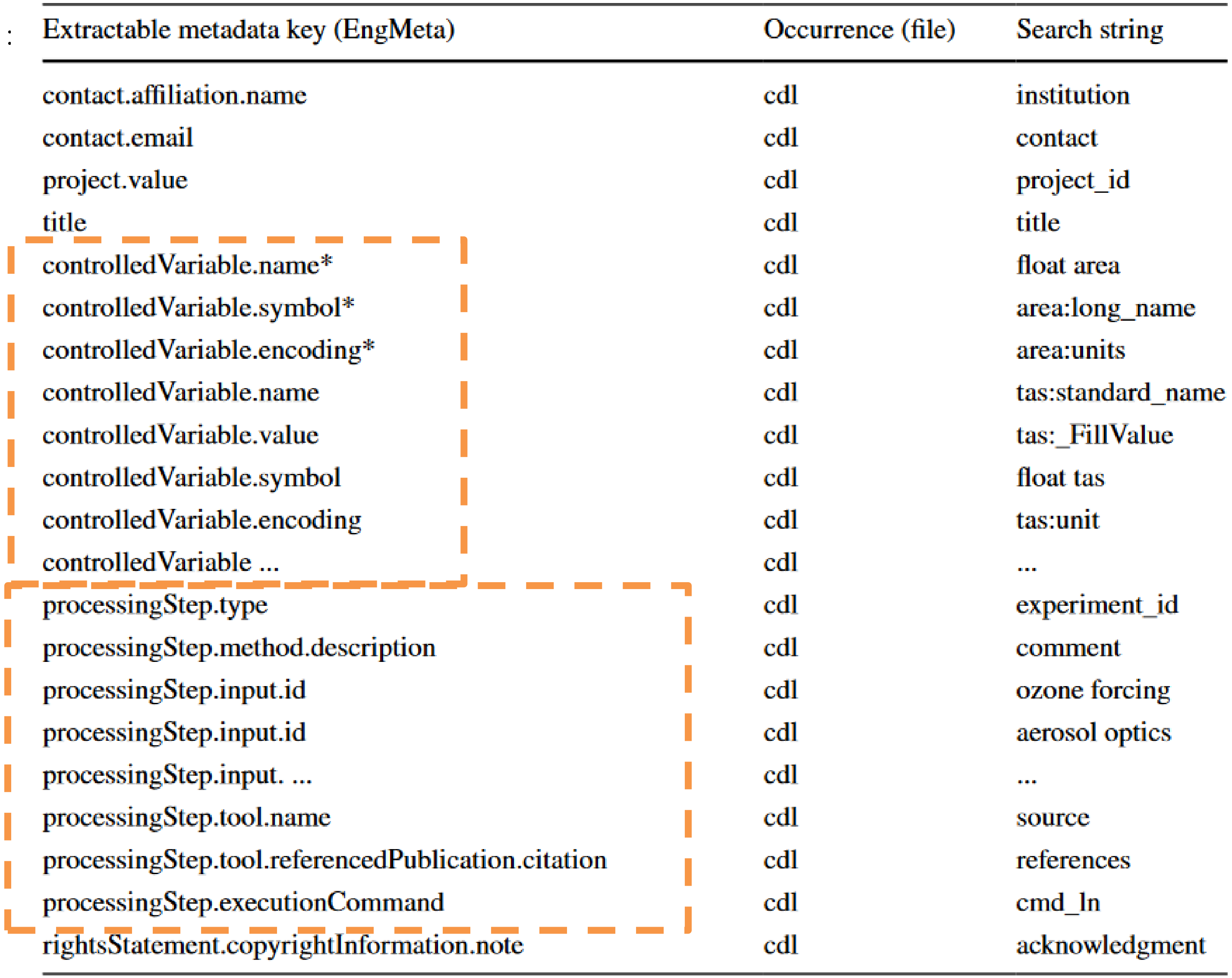
Output Methods

Figure 2. Architecture and data flow of the automated metadata extraction

Architectural design:

- automated metadata extraction is implemented in Java
- can be run on any system which offers a Java Runtime Environment
- can be integrated in any research workflow

https://doi.org/10.1007/s11227-020-03602



Extractable metadata key (EngMeta)	Occurrence (file)	Search string
contact.affiliation.name	cdl	institution
contact.email	cdl	contact
project.value	cdl	project_id
title	cdl	title
controlledVariable.name*	cdl	float area
controlledVariable.symbol*	cdl	area:long_name
controlledVariable.encoding*	cdl	area:units
controlledVariable.name	cdl	tas:standard_name
controlledVariable.value	cdl	tas:_FillValue
controlledVariable.symbol	cdl	float tas
controlledVariable.encoding	cdl	tas:unit
controlledVariable ...	cdl	...
processingStep.type	cdl	experiment_id
processingStep.method.description	cdl	comment
processingStep.input.id	cdl	ozone forcing
processingStep.input.id	cdl	aerosol optics
processingStep.input. ...	cdl	...
processingStep.tool.name	cdl	source
processingStep.tool.referencedPublication.citation	cdl	references
processingStep.executionCommand	cdl	cmd_ln
rightsStatement.copyrightInformation.note	cdl	acknowledgment

Domain-specific metadata:

- information on the controlled variables

Process metadata:

- information on the simulation code tools

  the compiler information

Outlook

- Provided tool for metadata extraction can be easily integrated into the research process;
- No need to change research workflows or simulation code;
- Designed tool can be applied to the various fields of science;
- In related works the tool was evaluated for GROMACS, EAS3 and CCSM employing the NetCDF data format.

Thank you