Automated metadata extraction and epistemic FAIRness in the engineering sciences

Martin Thomas Horsch, Taras Petrenko, and Björn Schembera

High Performance Computing Center Stuttgart (HLRS)

Data, Society, and Open Science III, TU Delft (digital), 30th March 2021

Research Data Management and Dark Data

- Typically, data organized in filesystems is not FAIR
- This may leave lots of dark data
- However, 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)

```
[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
```

Fig: Data organization in directory structures on filesystems. Sample from GROMACS

Research Data Management and Dark Data

- Typically, data organized in filesystems is not FAIR
- This may leave lots of dark data
- However, 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)

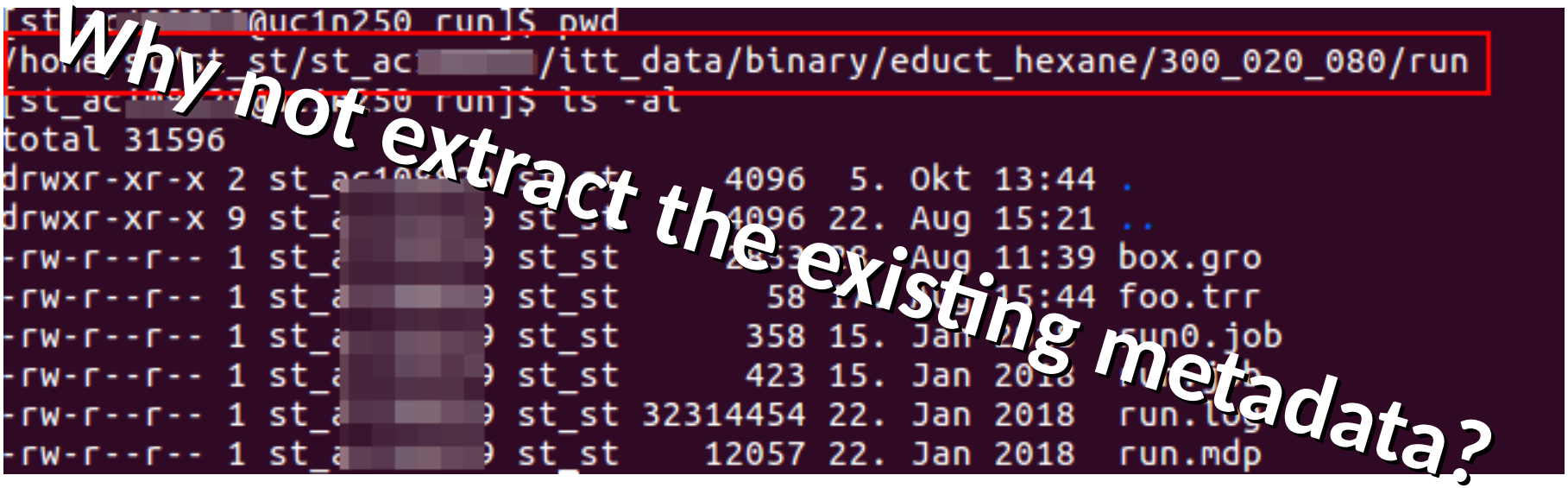


Fig: Data organization in directory structures on filesystems. Sample from GROMACS

ExtractIng metadata extraction tool

- Implemented in Java, published on GitHub (still a prototype though)[1]
- Native (Scanner API) and parallel (Apache Spark) version
- Generic: External configuration file based on the EngMeta convention
- Run of ExtractIng refers to a directory
- Data + metadata can then be ingested to a repository

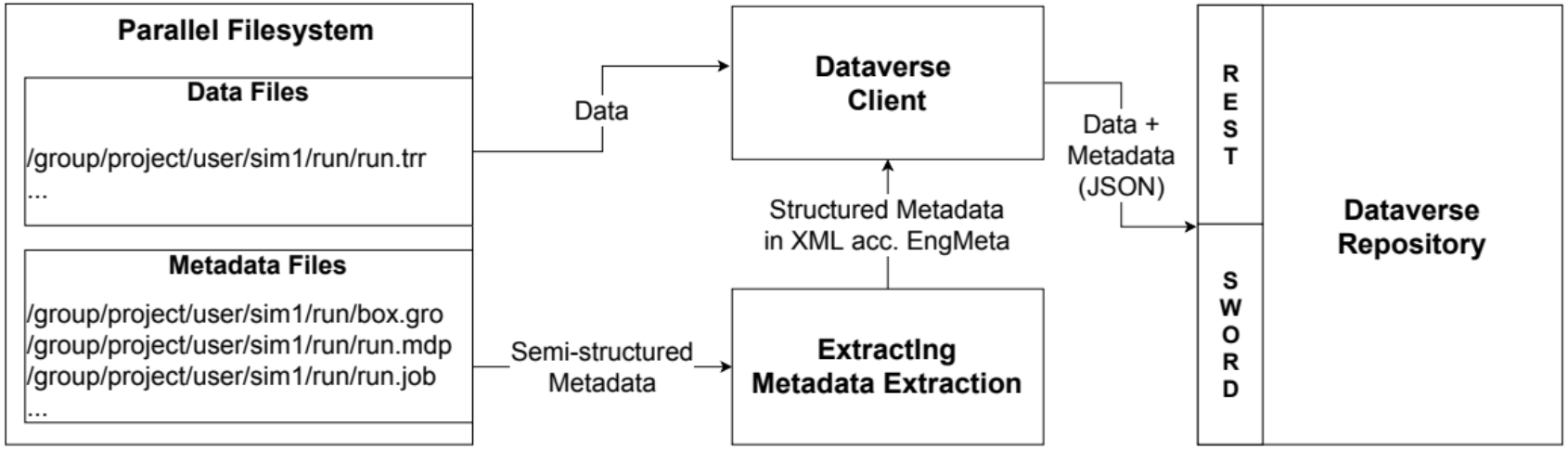


Fig: Workflow of the Metadata Extraction

[1] https://github.com/bjschembera/ExtractIng

Outcome: Extractable metadata

- Extractability of the metadata strongly related to
  - The type of metadata (technical, process, ...)
  - Also related to the simulation code output
- Evaluation with the following simulation codes output
  - GROMACS (molecular dynamics)
  - EAS3 (aerodynamics)
  - CCSM (climate modelling, in NetCDF conventions)

Type of metadata	Extractability
Technical metadata	high, as available via file attributes
Process metadata	medium, as available in log-, job- or system files
Domain-specific metadata	medium, as available in log- or output files
Descriptive metadata	poor, as it's a description from a higher level

Table: Extractability of the different metadata categories. It is strongly dependent on the field of science.

Schembera, B. Like a rainbow in the dark: metadata annotation for HPC applications in the age of dark data. J Supercomput (2021). https://doi.org/10.1007/s11227-020-03602-6

Automated extraction of data provenance information

- Provenance information/metadata is key for FAIR data.
- For GROMACS, the tool can extract lots of this provenance information:

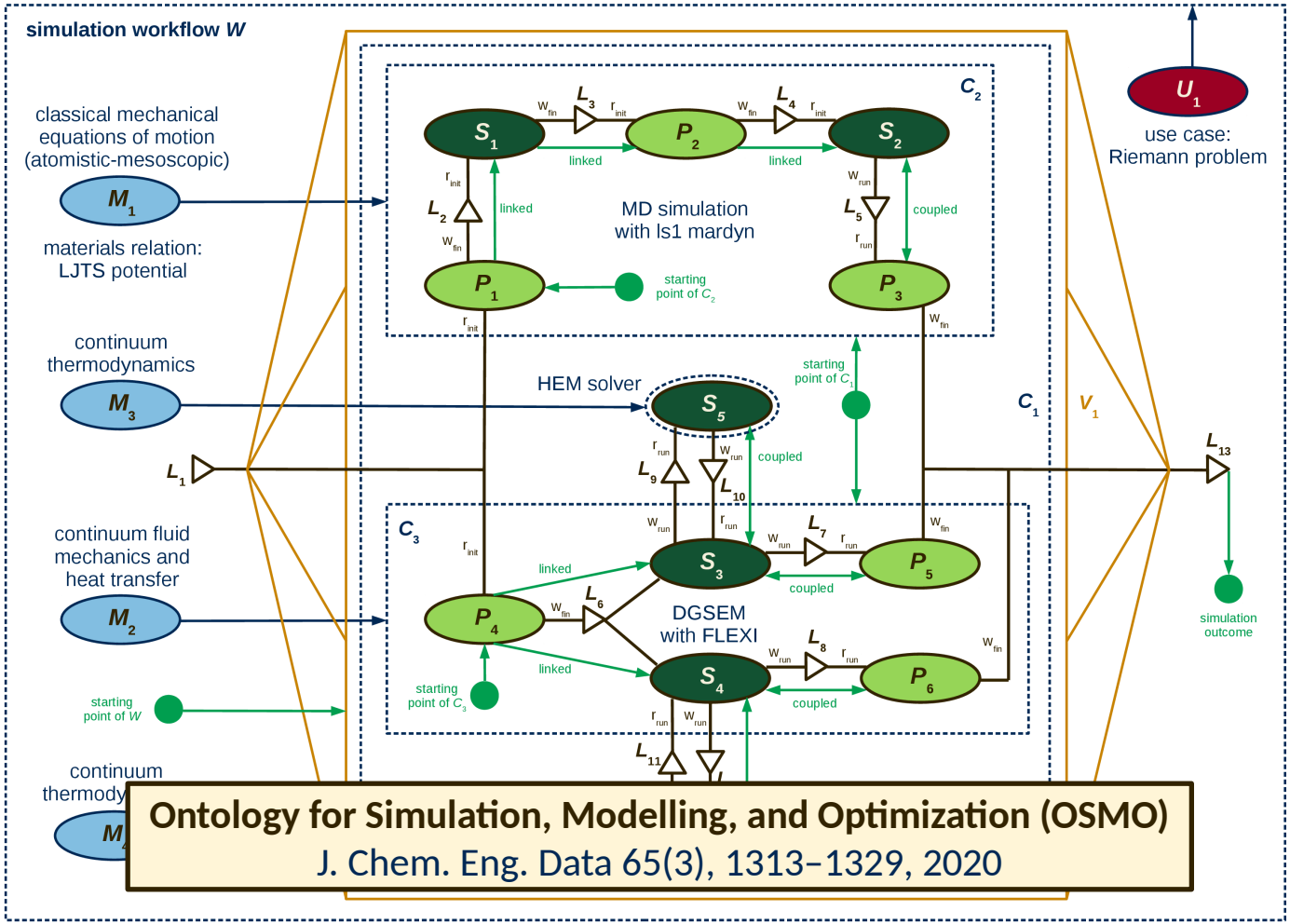
processingStep.tool.name	*.log	GROMACS
processingStep.tool.softwareVersion	*.log	GROMACS version
processingStep.tool.operatingSystem	*.log	Build OS/arch
processingStep.executionCommand	*.log	gmx_mpi mdrun
processingStep.executionCommand	*.log	gmx_mpi grompp
processingStep.environment.compiler.name	*.log	C++ compiler
processingStep.environment.compiler.flags	*.log	C++ compiler flags
processingStep.environment.compiler.name	*.log	C compiler
processingStep.environment.compiler.flags	*.log	C compiler flags
processingStep.environment.nodes	*.job	nodes
processingStep.environment.ppn	*.job	ppn
processingStep.environment.cpu	*.log	Build CPU brand

Automated extraction and metadata standardization of data provenance information

After the extraction run, provenance information is extracted and structured according to the EngMeta metadata standard (in a subdirectory *.metadata* as XML):

```
[[redacted]@nid00030 .metadata]$ pwd
/mnt/lustre/[redacted]/itt_data/binary/educt_hexane/300_020_080/run/.metadata
[[redacted]@nid00030 .metadata]$ ls -alrt
total 20
drwxr-xr-x 2 [redacted] s29931 4096 Jan 29 15:39 .
-rw-r--r-- 1 [redacted] s29931 1520 Feb  6 11:46 metadata.txt
-rw-r--r-- 1 [redacted] s29931 2717 Feb  6 11:46 engMeta.xml
-rw-r--r-- 1 [redacted] s29931  630 Feb  6 11:46 atom.xml
drwxr-xr-x 3 [redacted] s29931 4096 Feb 13 11:49 ..
[[redacted]@nid00030 .metadata]$ tail engMeta.xml
                    <flags>-mavx     -O3 -DNDEBUG -funroll-all-loops -fexcess-precision=fast</flags>
                </compiler>
                <nodes>1</nodes>
                <ppn>8</ppn>
                <cpu>Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz</cpu>
            </environment>
        </step>
    </provenance>
    <size>58</size>
</dataset>
[[redacted]@nid00030 .metadata]$
```

Metadata standardization for research data provenance



Ontology for Simulation, Modelling, and Optimization (OSMO)
J. Chem. Eng. Data 65(3), 1313–1329, 2020

OSMO-based **provenance description** as an extension of the MODA workflow meta-data standard:

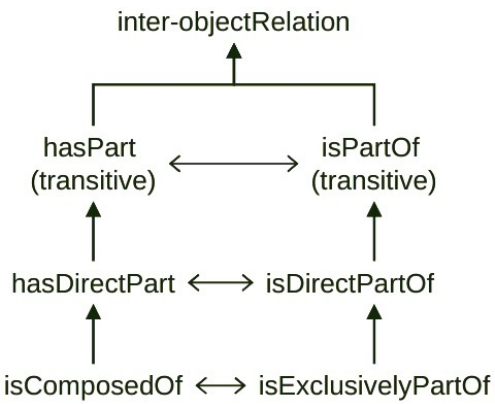
For all elements of the graph notation, there are corresponding concepts and relations from the ontology OSMO.

Metadata standardization based on top-level ontologies

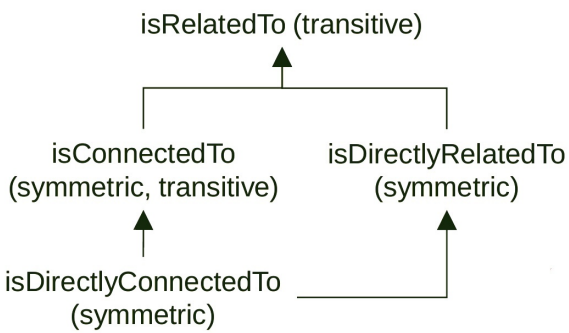
NFDI4Cat uses OntoCAPE,[1] the ontology for the CAPE-OPEN interface standard.

„CAPE“ = „computer-aided process engineering“

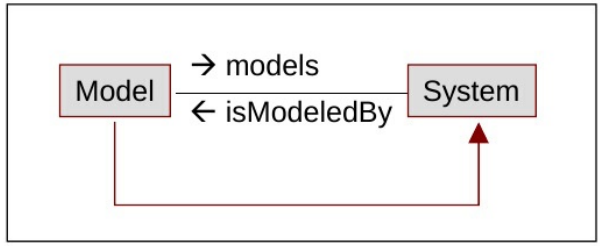
OntoCAPE combines domain-specific and **top-level conceptualizations**.



mereology



topology of systems

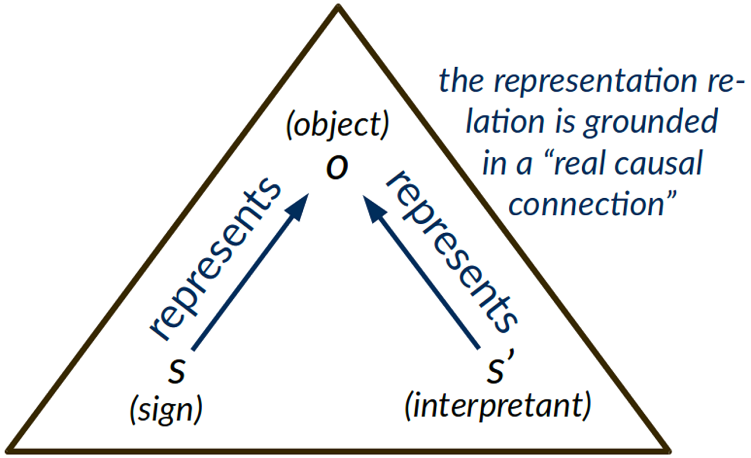


representation by **models**

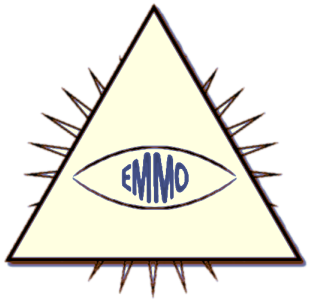
[1]Morbach *et al.*, Technical Reports LPT-2008-24 & LPT-2008-25, RWTH Aachen, **2008**.

Metadata standardization based on top-level ontologies

Peircean semiotics



the semiosis, a process by which a new representamen, the interpretant, is created



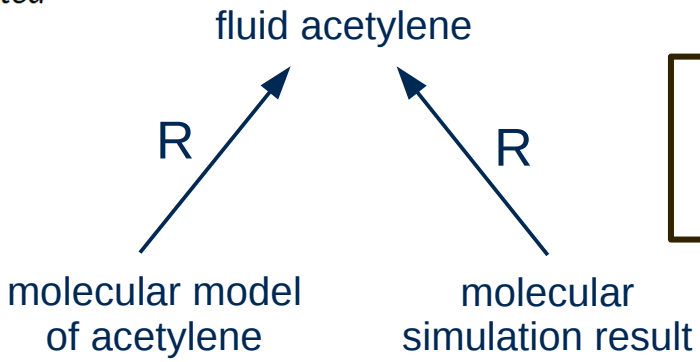
C. S. Peirce

European Materials and Modelling Ontology

1) **Taxonomy**:
   Conceptual hierarchy (subclass relation)

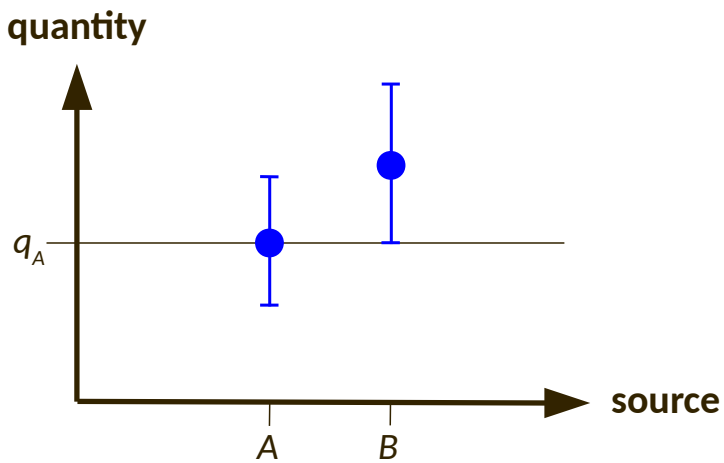
2) **Mereotopology**:
   Spatiotemporal parthood and connectivity

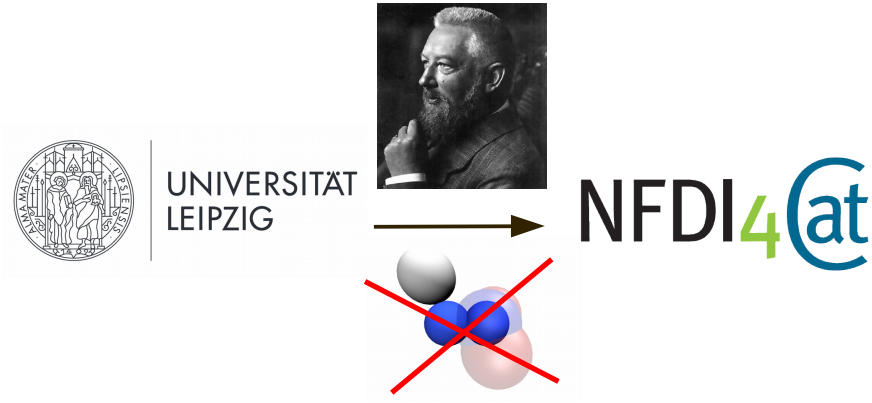
3) **Semiotics**:
   Representation of physical entities by signs



"represents" or **"is sign for"** is here abbreviated by **R**

Research data infrastructures and scientific knowledge





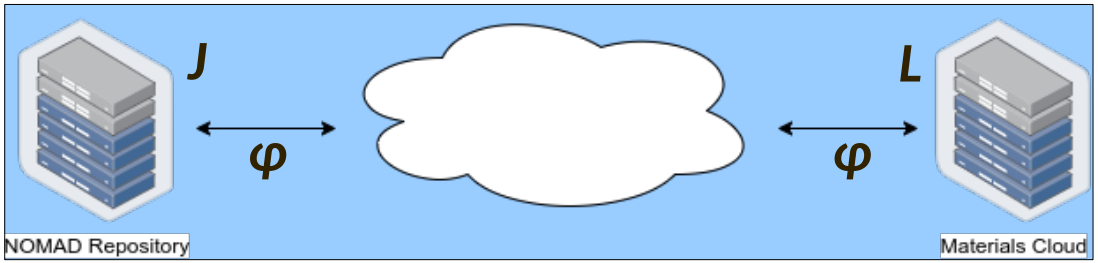
Averroes[1]
(Bidayat al-Mujtahid)
al-Sulami
(*Book of the Jihad*, 1106)

Pope Urban II
(in *Proc. Council Clermont*, 1059)
Augustine (*Civitas Dei*, 426)

[1]Research data infrastructure on Averroes' works: https://averroes.uni-koeln.de/

Digital infrastructures and communication of knowledge

Scientific knowledge is a kind of knowledge (or else, little will qualify as knowledge). Research data infrastructures store and exchange scientific knowledge.

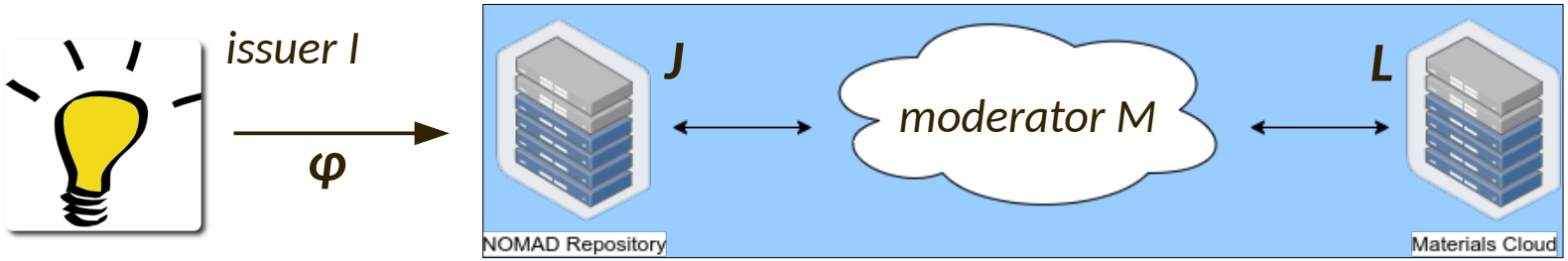


Scenario requiring epistemological formalization:

- "The scientific knowledge φ is communicated by knowledge base J to L."
- φ is a justified tenable assertion, by standards applied to its source. But it would be inappropriate to require every φ to be a justified true belief.

Digital infrastructures and communication of knowledge

Scientific knowledge is a kind of knowledge (or else, little will qualify as knowledge). Research data infrastructures store and exchange scientific knowledge.



Scenario requiring epistemological formalization:

- "*M* asserts and approves *φ*'(*I*, *J*, *L*, *φ*)," where *φ*'(*I*, *J*, *L*, *φ*) is given by:
- "The scientific knowledge *φ*, previously issued by a source *I*, has been communicated by the knowledge base *J* to the knowledge base *L*."
- *J*, *L*, and *M* have a justified true belief in *φ*'.
- *φ* is a justified tenable assertion, by the standards applied to *I* by *M*.

User stories: Representative research workflows

user story 3: CO_2 methanation

user story 2: synchrotron QEXAFS experiments

user story 4: biocatalysis & multiphase flow

user story 1: photocatalytic water splitting

user story 5: CO_2 electrolysis

user story 6: metal-organic frameworks

user story 7: nanoparticle & SILP catalysts

user story 8: selective NOx reduction

User stories: Representative research workflows

user story 11: coupled in-situ and offline analytics

user story 10: Fischer-Tropsch process

user story 12: syngas-to-ethanol catalyst characterization

TECHNISCHE UNIVERSITÄT DARMSTADT

Catalysis Leibniz-Institut für Katalyse e.V.

user story 9: photocatalytic CO_2 reduction

user story 13: syngas-to-ethanol catalyst performance

UNIVERSITÄT GREIFSWALD Wissen lockt. Seit 1456

FAU

user story 16: water-gas shift reaction

user story 14: RDM for enzyme activity data

user story 15: modelling SILP catalysed reactions

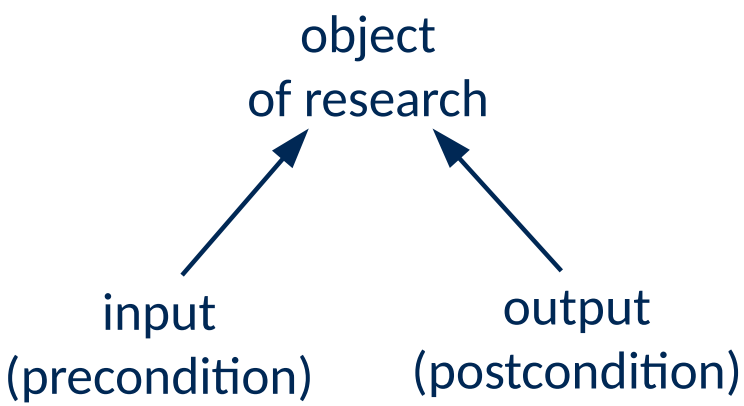
User stories: Representative research workflows

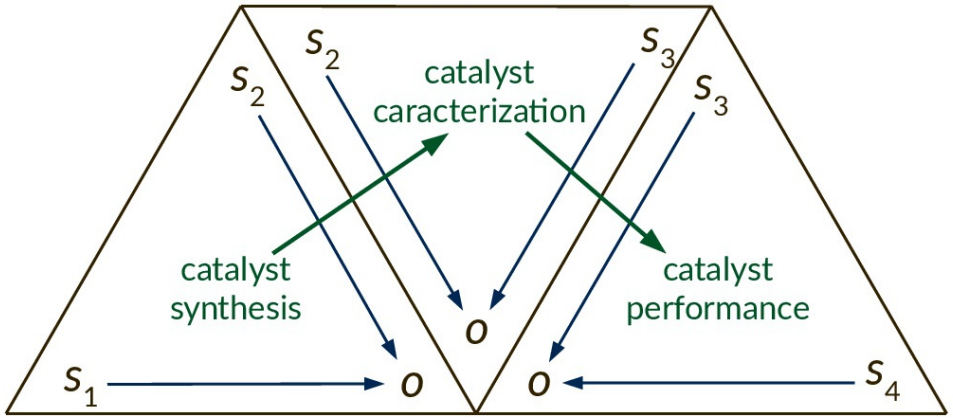
Interviews of 30 minutes each are conducted with internal prospective users.

For each research step, we jointly identify:

- **input**, *i.e.*, all that needs to be present in advance (incuding equipment)
- **output**, *i.e.*, that which is generated as an outcome of the research step

Pre- and postcondition are causally connected by participating in the same step of a research process, and they are applied to the same **object of research**.

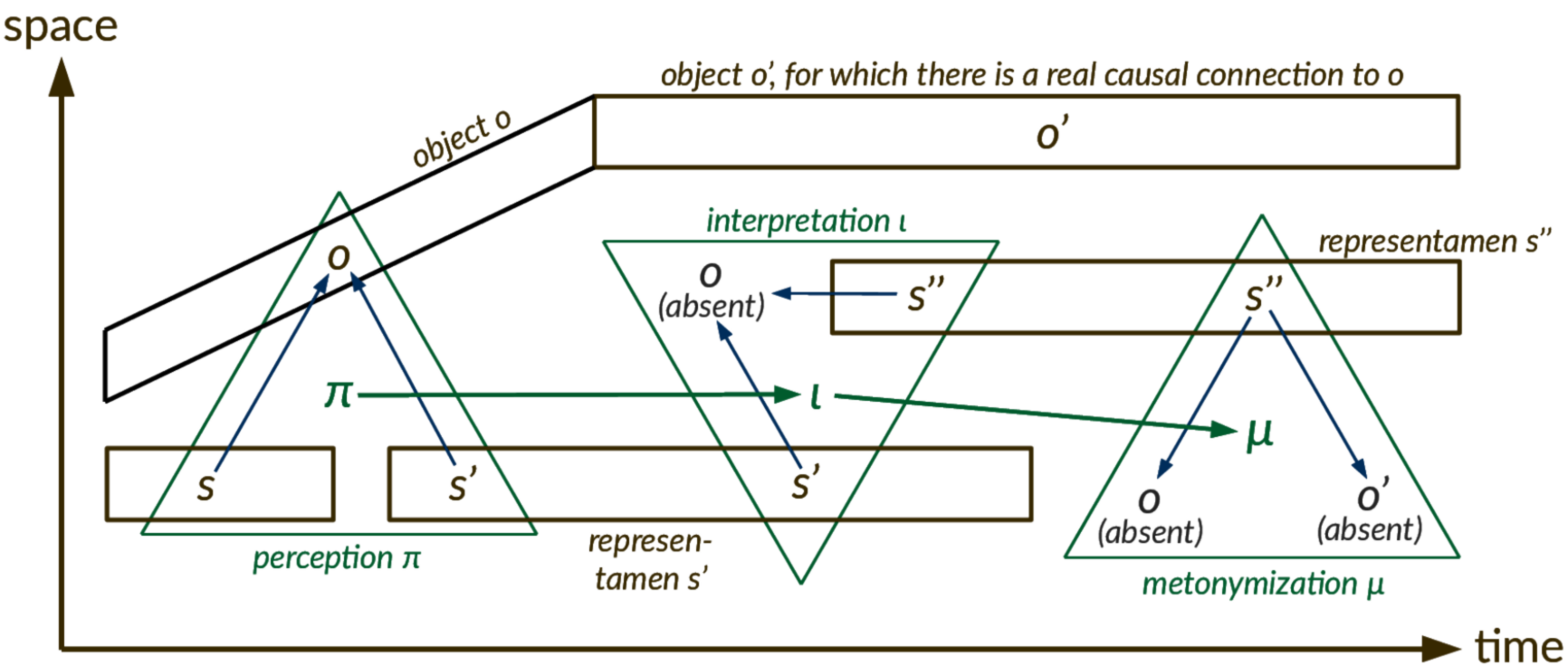




User story by A. Bordet, MPI CEC, Mülheim a. d. Ruhr

Cognitive processes following Peircean semiotics

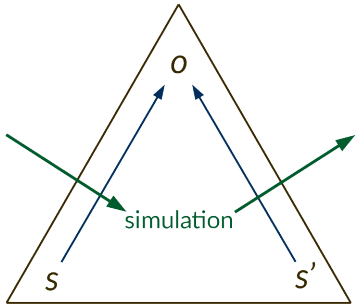
Mereosemiotics:[1] Combination of mereotopology and Peircean semiotics



[1]In *Proceedings of WCCM-ECCOMAS 2020*, doi:10.23967/wccm-eccomas.2020.297, **2021**.

Epistemic opacity: The challenge

Issue raised by Humphreys:[1] **Justification of *φ*** appears (to some) to be **opaque.**[1, 2]



alleged sources of opacity

validation becomes „holistic“

no explicit justification is computed

stochastic elements (Monte Carlo)

Lyapunov instability (molecular dynamics)

However, experiments are not usually viewed as opaque.

predetermination[3]

e.g., formal software verification[3]

usually inapplicable

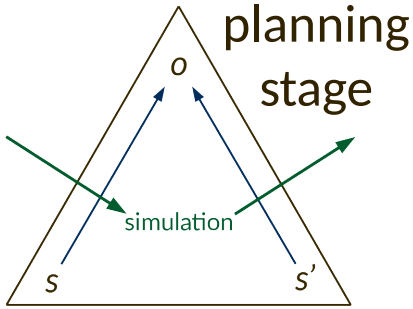
Underlying requirement: Provenance description delivering *scientia media* (system retains freedom).

L. de Molina

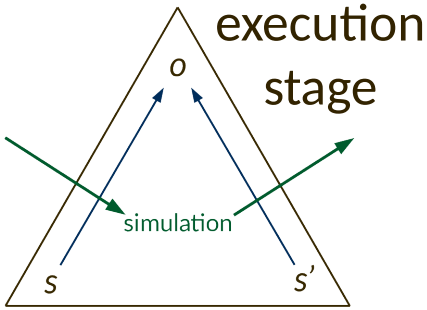
[1]Knowledge of *„all epistemically relevant elements“* cannot be attained (Humphreys, 2004, 2011).
[2]Durán and Formanek (2018): *„epistemically relevant elements“* = *„steps of the [...] justification“*.
[3]Required for non-opacity by Newman (2016), a requirement criticized by Durán & Formanek (2018).

Epistemic opacity as opposed to epistemic FAIRness

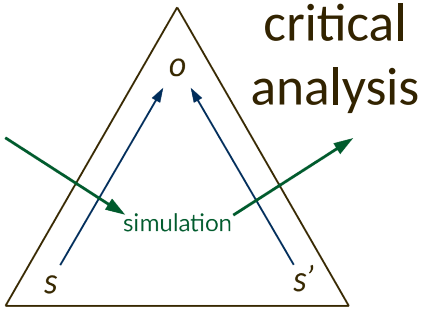
Three modes of justification by epistemic grounding:



ex ante
predetermination
(& model validation)



in actu
determination
„Reflexion im Vollzug“[1]



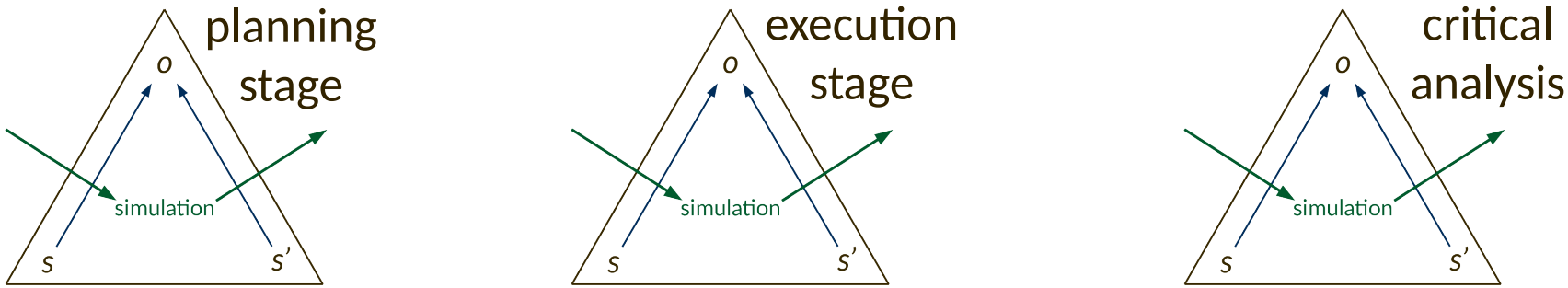
ex post
redetermination
„Reflexion des Vollzugs“[1]

Epistemic opacity is reduced by **epistemic FAIRness**, *i.e.*, the FAIR provision of a provenance description via a research data infrastructure that permits a reevaluation of the research workflow over an open epistemic space.

[1]Tulatz, *Epistemologie als Reflexion wissenschaftlicher Praxen*, **2018**.

Conclusion

To make scientific knowledge FAIR, research data infrastructures need to support the documentation, ingest, retrieval, and revision of data provenance.



Priorities ("DORIC principles") following doi:10.5281/zenodo.4571052

D	O	R	I	C
diversify technologies	**observe** practices	**realistic** objectives	**incentivize** open data	**co-design** data and workflows