Domain-specific metadata standardization in materials modelling

M. Horsch,[1] J. Francisco Morgado,[2] G. Goldbeck,[3] D. Iglezakis,[4] N. Konchakova,[5] B. Schembera,[1] *High Performance Computing Center Stuttgart,[1] Fraunhofer IWM,[2] Goldbeck Consulting Ltd,[3] University of Stuttgart,[4] Helmholtz-Zentrum Hereon[5]*

DORIC-MM 2021 Workshop, European Semantic Web Conference, 7th June 2021

Hype or realistic objectives?

Example from a Horizon 2020 call[1] on „Open Science“ and „Open Innovation“

The Commission considers that proposals requesting a contribution from the EU around EUR 5 million would allow this specific challenge to be addressed appropriately. Nonetheless, this does not preclude submission and selection of proposals requesting other amounts.

Expected impact:

- Remove barriers to the use of materials models by lowering the learning curve, increase the knowledge-base of European industry and the total cost of ownership leading to an industrial user base of companies increased by a factor 2;
- ***Increased speed of material and/or product development time and rapid design from concept to market by factor 5*** and allow industry to react to changing market and regulatory demands;
- Change the operational practice of companies by making them more data driven, agile, light and competitive and thus support the Digital Single Market (DSM) objectives and thus ***drastically reduced development costs for industry by a factor 2***;

[1]Horizon 2020 call no. DT-NMBP-10-2019

Hype or realistic objectives?

Proposal for community recommendations: The DORIC principles

Focus on five lines of development as priorities for FAIR research data management in domains of knowledge involving modelling and simulation:

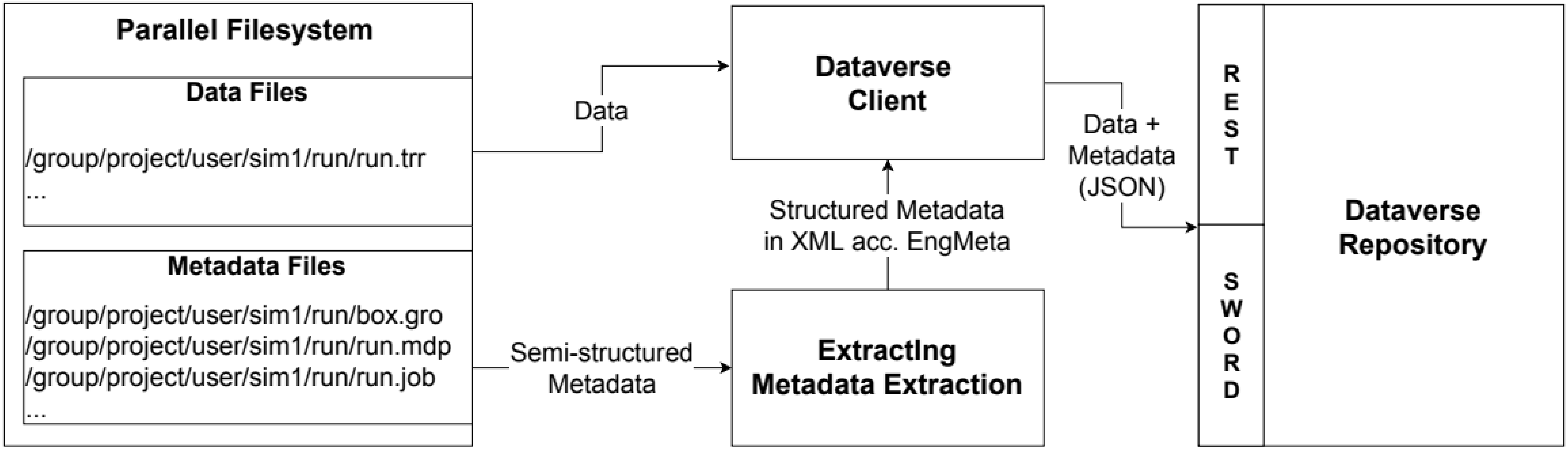
Priorities ("DORIC principles") following doi:10.5281/zenodo.4571052

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

reduced development costs for industry by a factor 2;

Reducing dark data by automated metadata extraction

Dark data[1] are the opposite of FAIR data: On HPC systems, in filesharing environments, and on computers of individual researchers, data become worthless unless they are sufficiently annotated, *e.g.*, supported by an automated workflow.[2]



Above: ExtractIng workflow,[2] yielding provenance metadata, quantity values, *etc.*

[1]B. Schembera, J. Durán, *Philosophy & Technology* 33, 93–115, doi:10.1007/s13347-019-00346-x, **2019**.
[2]B. Schembera, *Journal of Supercomputing*, doi:10.1007/s11227-020-03602-6, **2021**.

Design of disciplinary interoperability architectures

Strategies for reusing pre-existing semantic artefacts

1) Total *design from scratch*: Coherent system and philosophy. This approach risks failing to create sufficiently well-designed schemas or patterns for all aspects.

2) Middle path: Particularly *important patterns and branches are designed for a dedicated purpose*, new concepts and relations for domains are created where needed, while pre-existing ontologies are referenced where that is beneficial.

3) Reuse *pre-existing semantic artefacts wherever possible* for a *maximum of immediate interoperability* with external data and platforms. Specific new concepts and relations are introduced at the domain level, to a restricted extent.

Challenge: Pre-existing ontologies can contradict each other. If strategy 2 or 3 is followed, the implementation must allow knowledge bases to remain consistent.

Design of disciplinary interoperability architectures

Strategies for integrating dedicated with pre-existing semantic artefacts

1) Explicit *import statements* in the TTL file: Some or all reused ontologies are loaded automatically, and the reasoner is applied to the whole framework.

2) References to external namespaces are treated *analogous to hyperlinks*, where supporting information is reproduced (*e.g.*, "schema:Action a owl:Class"), using

   a) redeclaration (*e.g.*, "m4i:Action owl:equivalentClass schema:Action")
   b) no redeclarations, but subsumption statements using rdfs:subClassOf;
   c) annotation properties or comments about ontology alignment.

3) No external information is reproduced; users load these ontologies if needed.

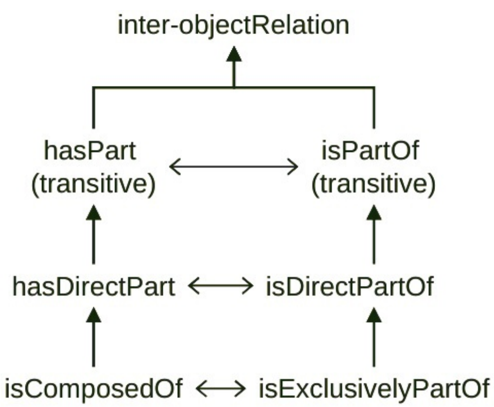
Combining mutually inconsistent ontologies is limited to **weak alignment** (no. 2).

Metadata standardization based on top-level ontologies

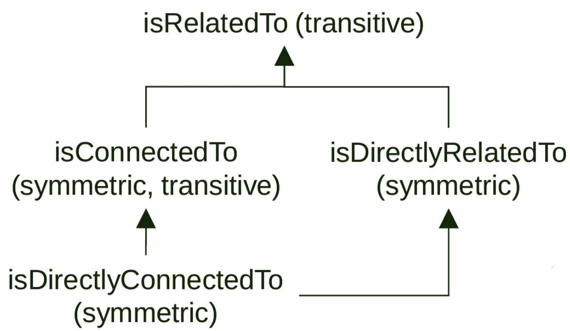
NFDI4Cat explores using OntoCAPE, developed by the Marquardt group[1, 2, 3] for

CAPE: Computer-Aided Process Engineering.

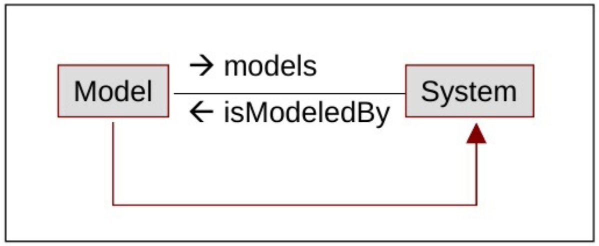
OntoCAPE combines domain-specific and top-level conceptualizations.



mereology



topology of systems

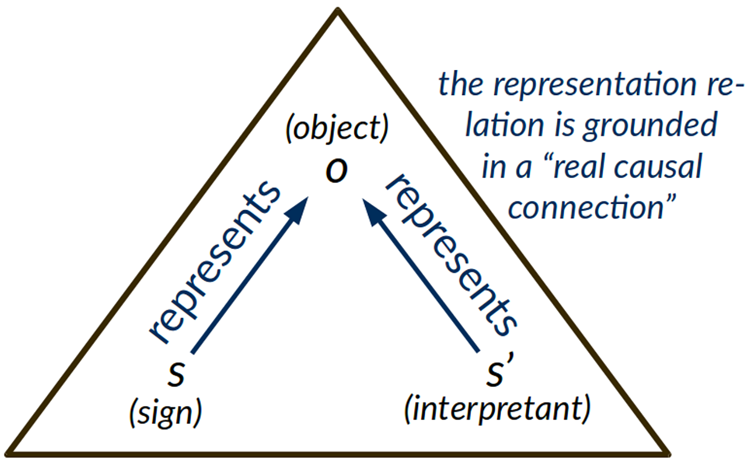


representation by **models**

[1]J. Morbach *et al.*, technical reports LPT-2008-24 & LPT-2008-25, RWTH Aachen University, **2008**.
[2]J. Morbach, A. Wiesner, W. Marquardt, *Computer Aided Chemical Engineering* 25, 991–996, **2008**.
[3]A. Wiesner, J. Morbach, W. Marquardt, *Computers & Chemical Engineering* 35(4), 692–708, **2011**.

Metadata standardization based on top-level ontologies

Peircean semiotics



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

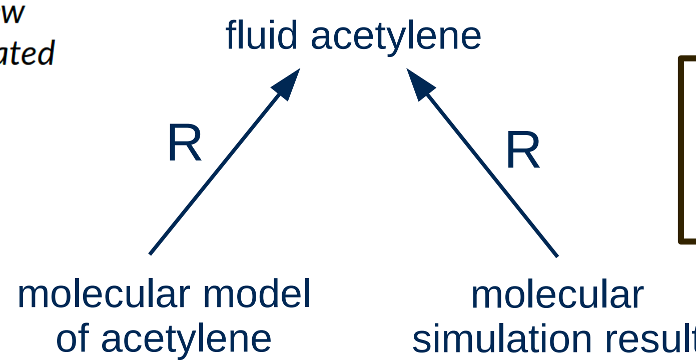

C. S. Peirce

European Materials and Modelling Ontology[1]

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

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

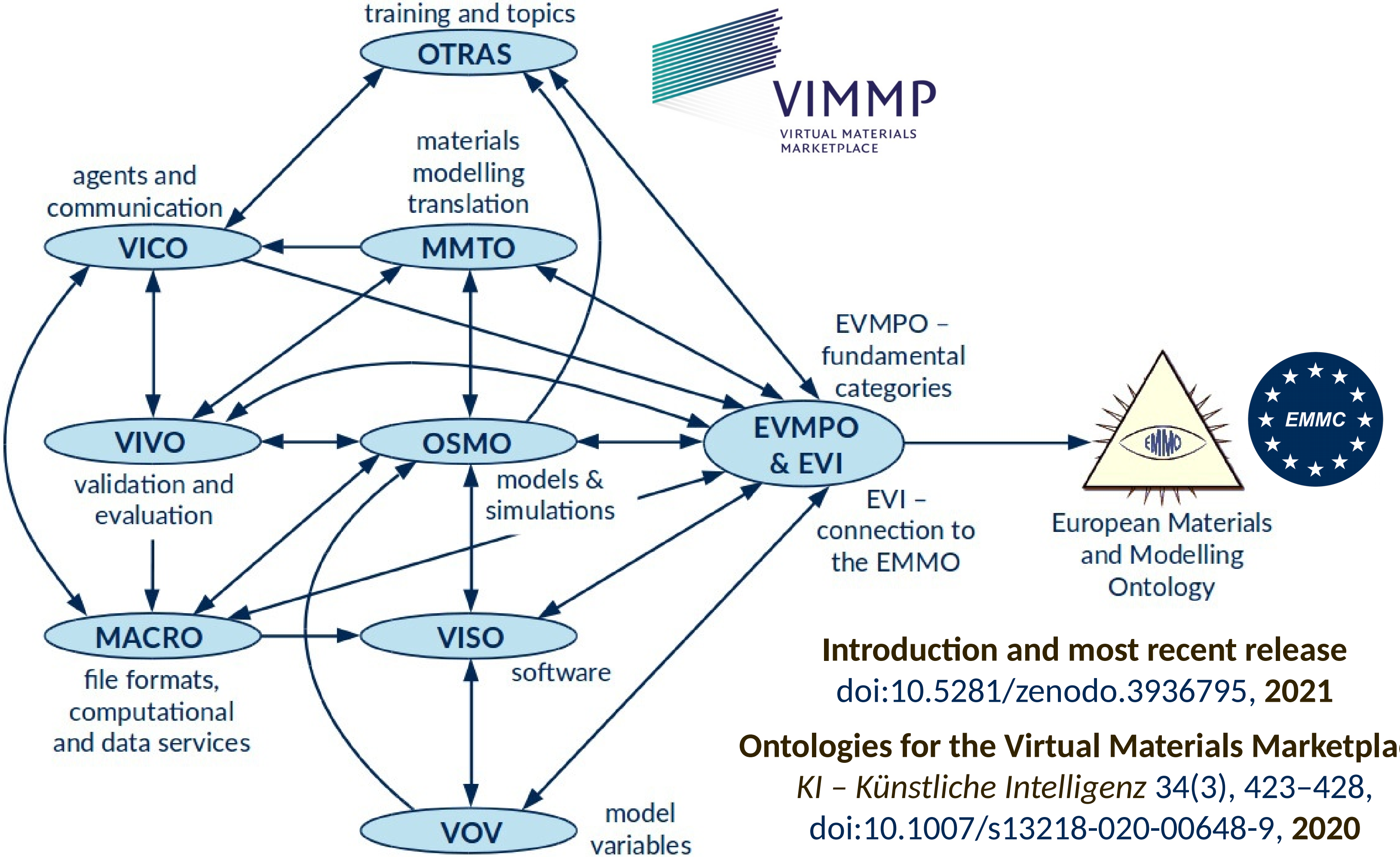
3) **Semiotics**:
Representation of physical entities by signs



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

[1] G. Goldbeck *et al.*, *Proc. NAFEMS World Congress*, NWC_19_86, **2019**.

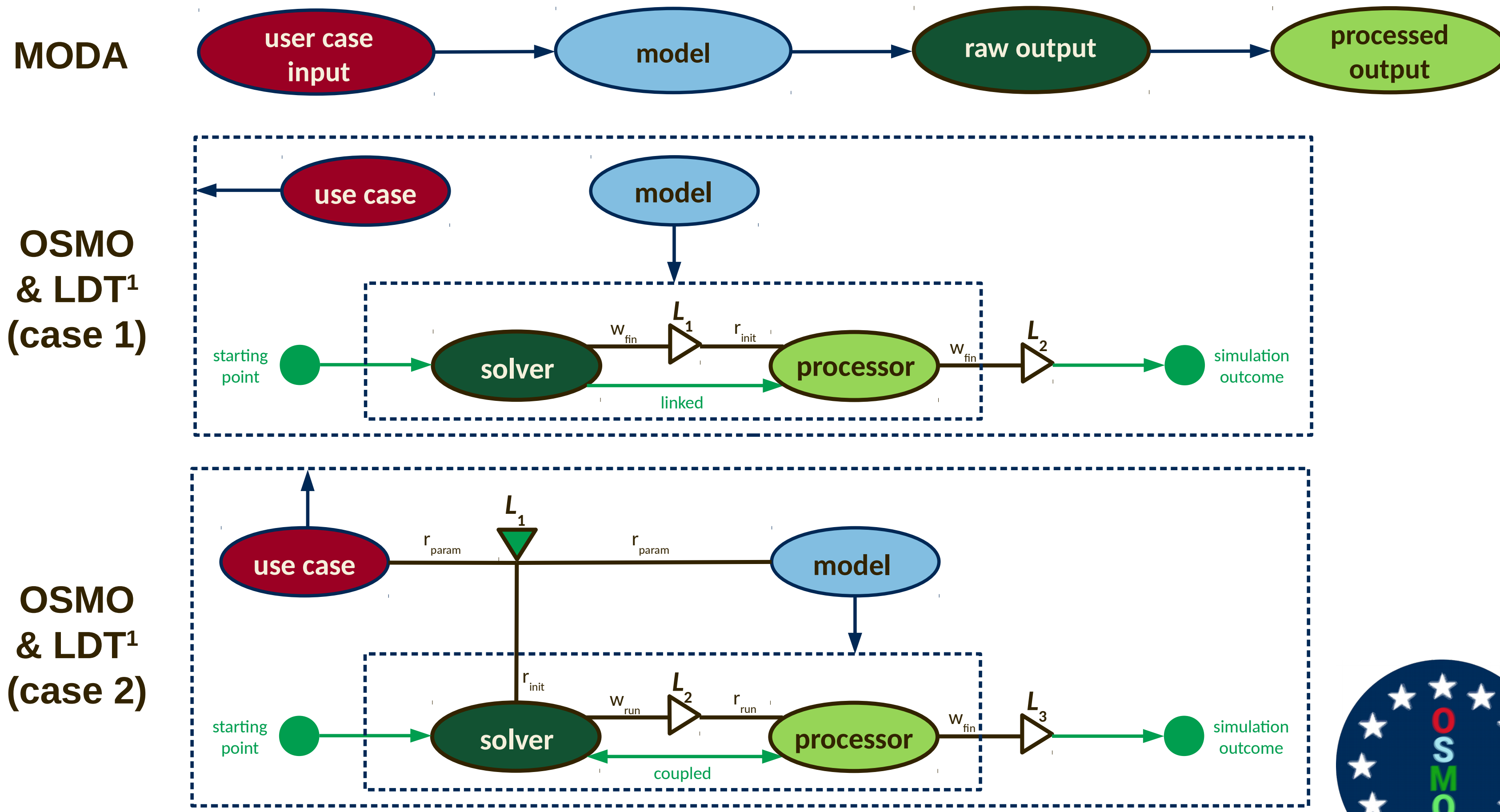
EMMO-based design of ontologies in materials modelling



Introduction and most recent release
doi:10.5281/zenodo.3936795, **2021**

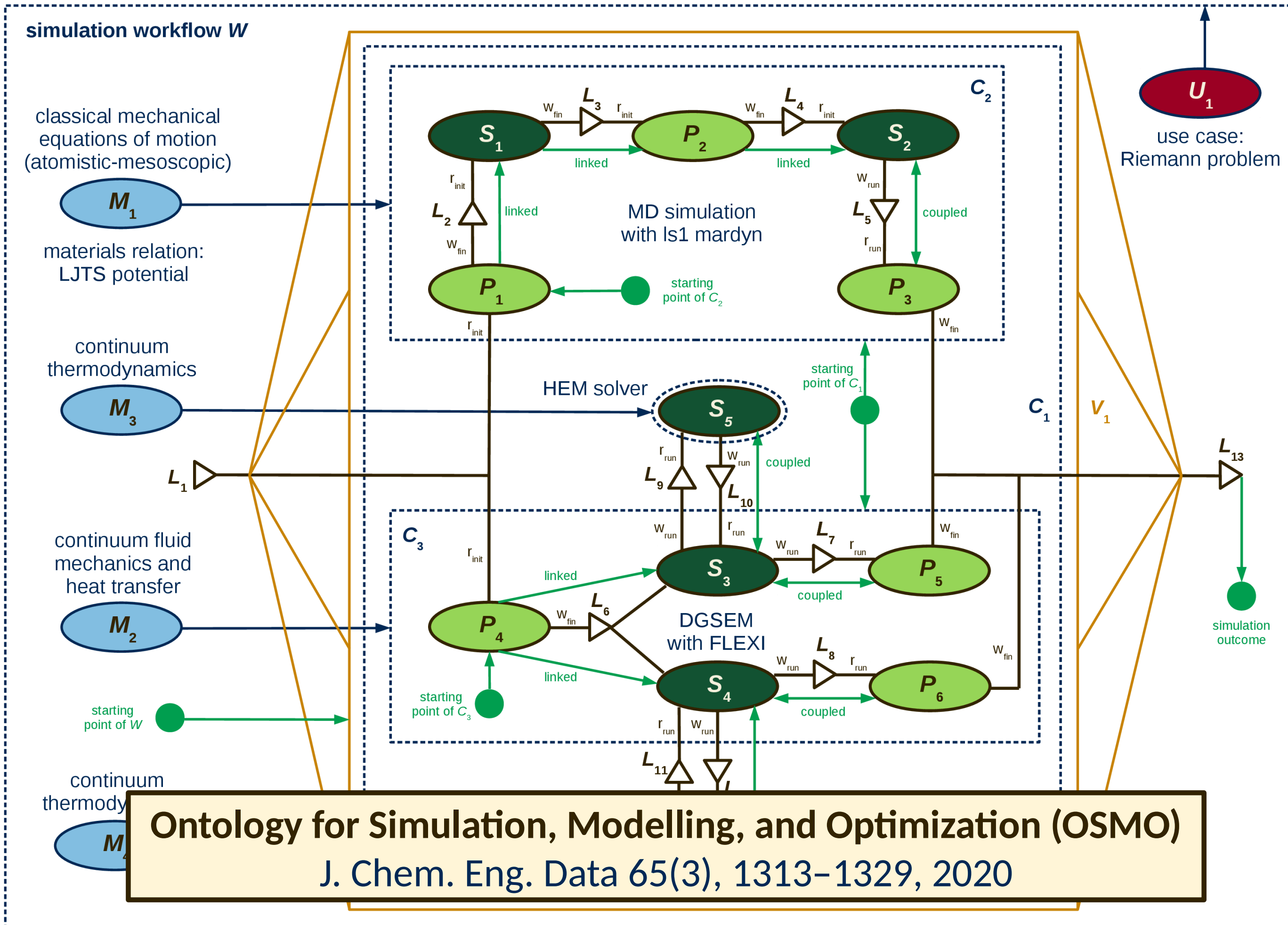
Ontologies for the Virtual Materials Marketplace
KI – Künstliche Intelligenz 34(3), 423–428,
doi:10.1007/s13218-020-00648-9, **2020**

Connecting MODA provenance metadata to the EMMO



[1]M. T. Horsch *et al.*, *Journal of Chemical & Engineering Data* 65(3), 1313–1329, **2020**.

Connecting MODA provenance metadata to the EMMO



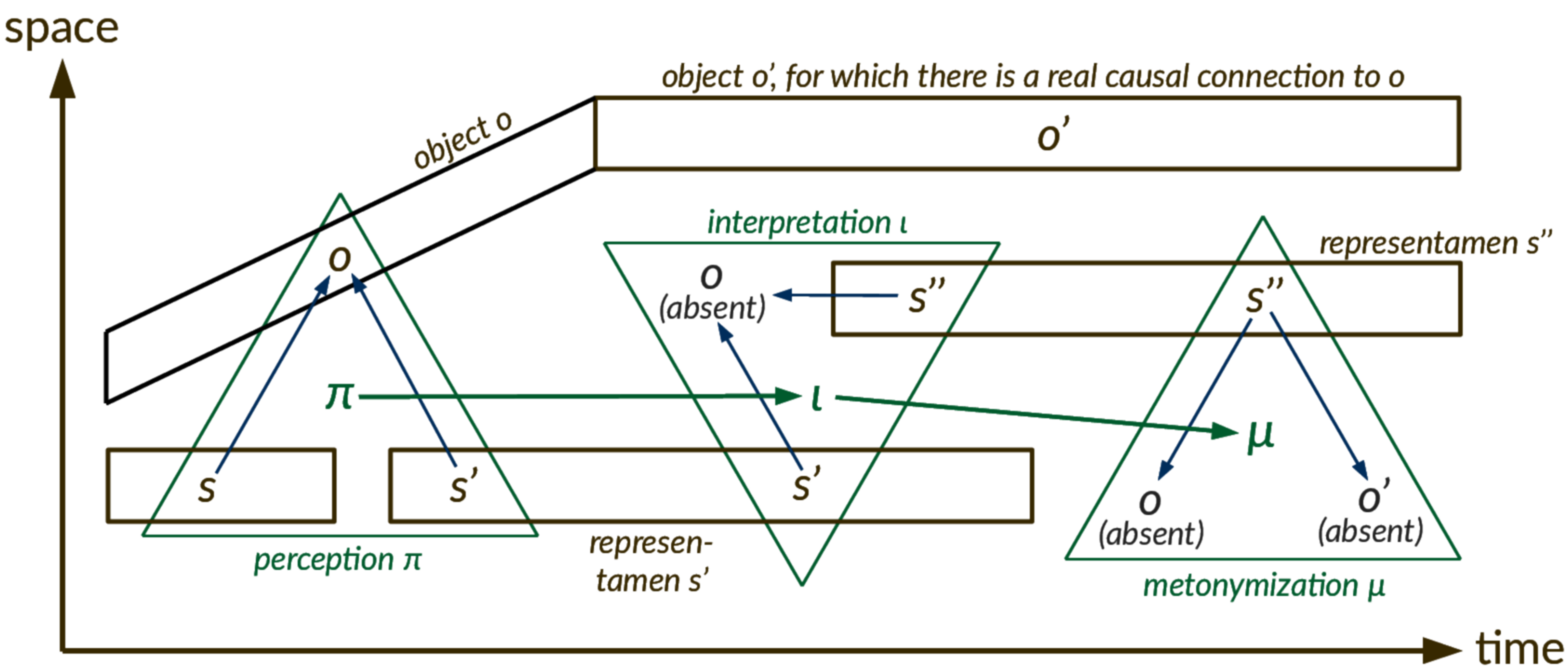
Ontology for Simulation, Modelling, and Optimization (OSMO)
J. Chem. Eng. Data 65(3), 1313–1329, 2020

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.

Data provenance description following Peircean semiotics

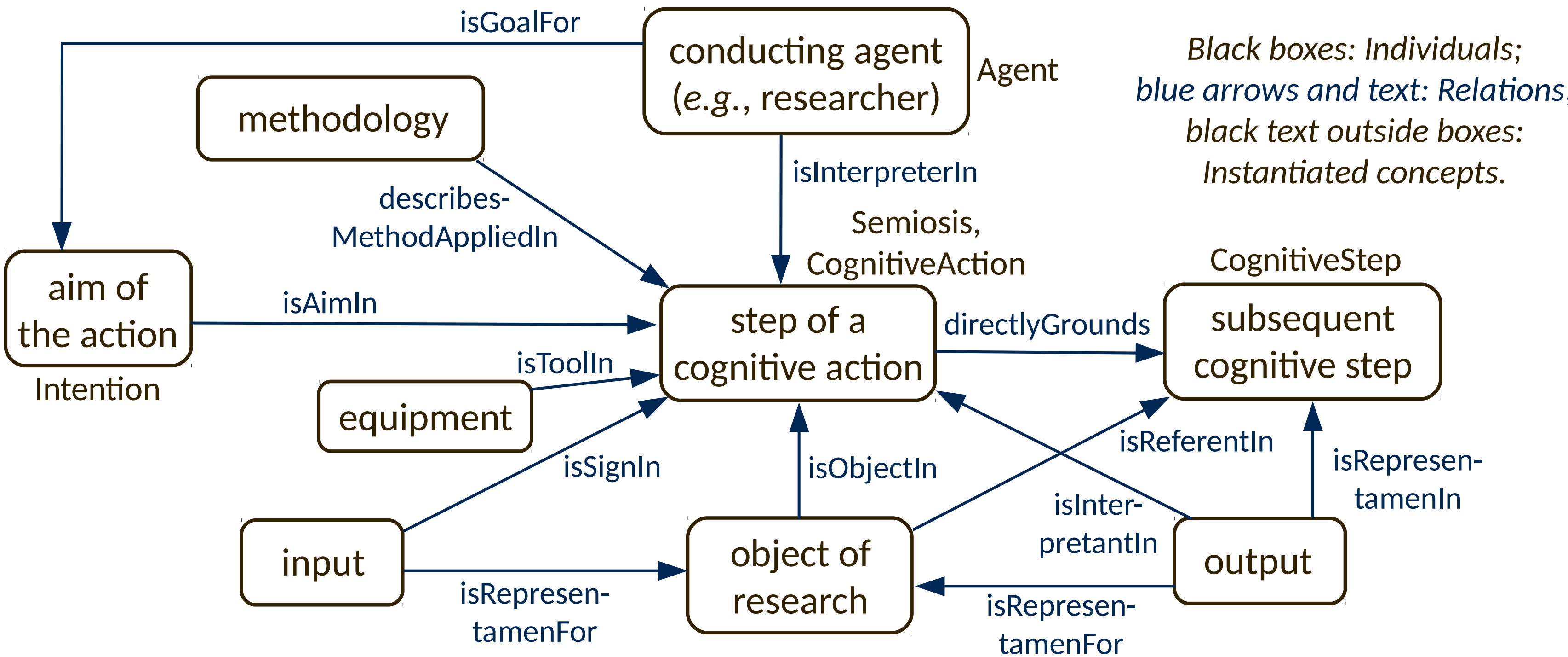
Mereosemiotics:[1, 2, 3] **Combination of mereotopology and Peircean semiotics**



[1]M. T. Horsch, S. Chiacchiera, B. Schembera, M. A. Seaton, I. T. Todorov, *Proc. WCCM-ECCOMAS 2020*, **2021**.
[2]First-order logic implementation, doi:10.5281/zenodo.4849611; examples, doi:10.5281/zenodo.4679522.
[3]Introduction and relation to EMMO and MODA discussed by H. A. Preisig *et al.*, DAMDID 2021, submitted.

Data provenance description following Peircean semiotics

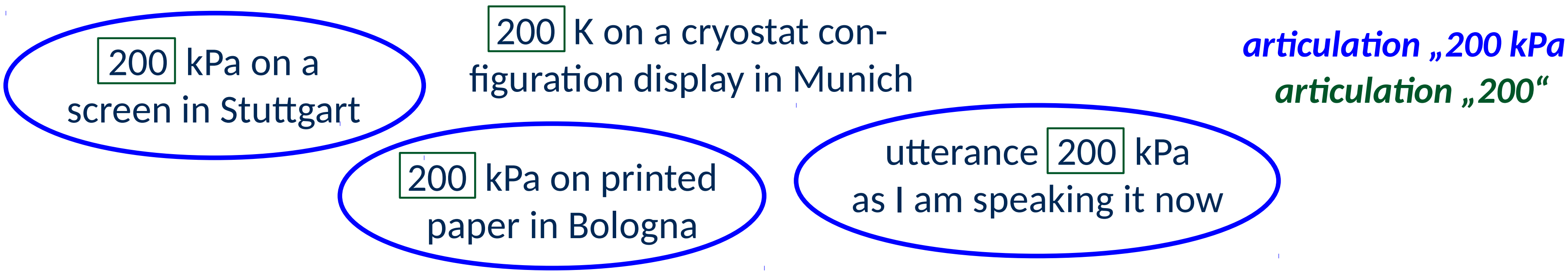
Schema from the PIMS interoperability infrastructure[1, 2, 3] (PIMS-II)



[1]Introduction and relation to EMMO and MODA discussed by H. A. Preisig *et al.*, DAMDID, submitted, **2021**.
[2]PIMS-II OWL ontology for cognitive processes accessible at http://www.molmod.info/semantics/pims-ii.ttl.
[3]Modal first-order logic ontology at 10.5281/zenodo.4849611; examples at doi:10.5281/zenodo.4679522.

Variables and quantity values: Semiotic collectives

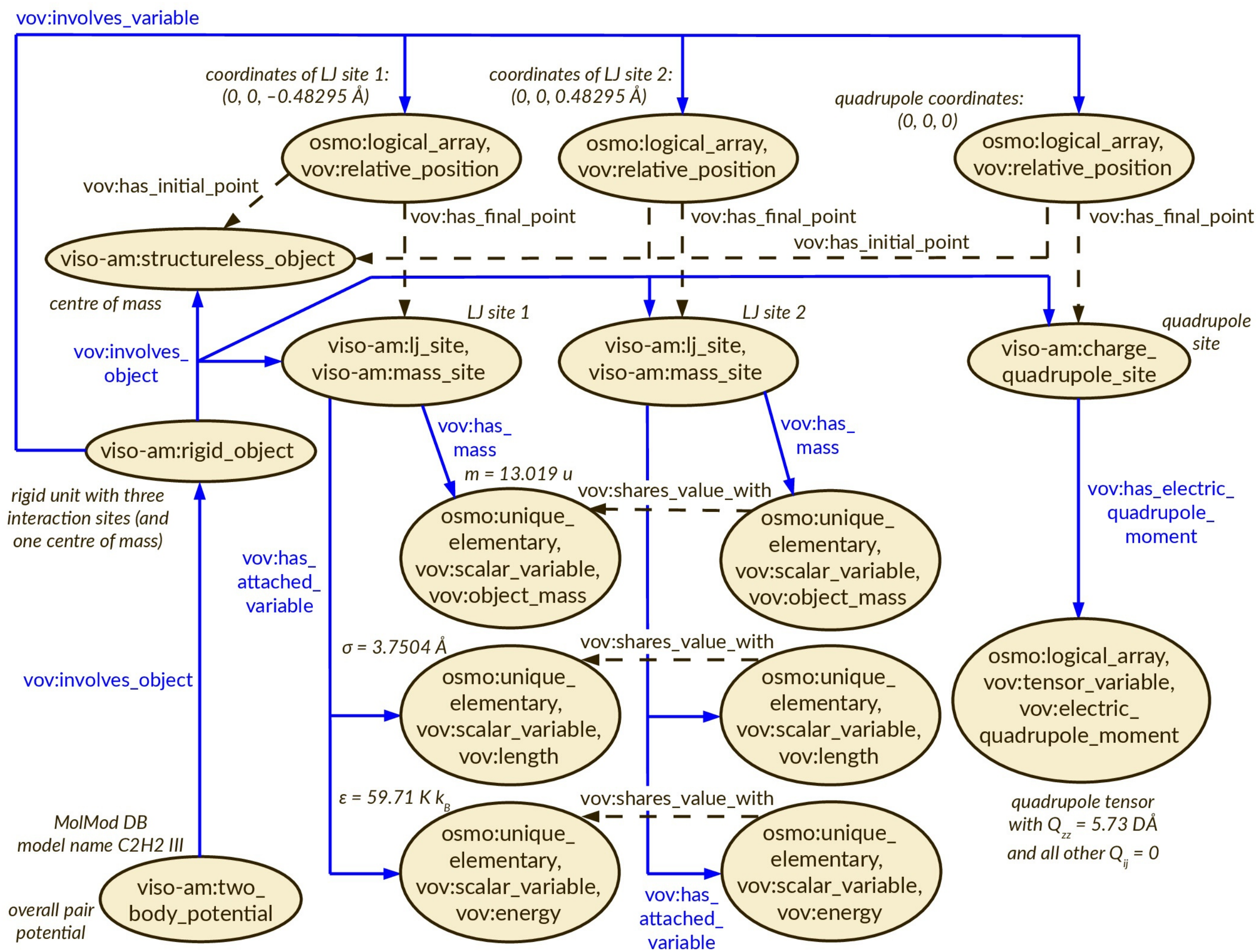
„200 kPa“ should be **the same value**, no matter where it appears as 4D object. Implementing this with the EMMO is challenging due to nominalism. Solution:

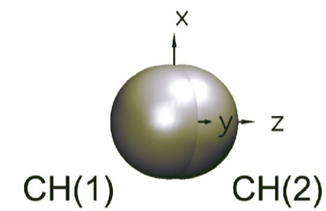


Previous work by Masolo *et al.*[1] proposes three categories: Pluralities, proper collectives, and composites. The PIMS interoperability infrastructure[2, 3] (PIMS-II) ontology for cognitive processes defines **semiotic collectives** as entities that appear jointly as a representational element, *i.e.*, as a representamen or re-ferent. Four categories: Pluralities, structures, articulations, and propositions.[3]

[1]C. Masolo, L. Vieu, R. Ferrario, S. Borgo, D. Porrello, *Proc. FOIS 2020*, pp. 186–200, **2020**.
[2]H. A. Preisig, P. Klein, N. Konchakova, M. T. Horsch, *DAMDID 2021*, submitted, **2021**.
[3]M. T. Horsch, *FOUST 2021*, submitted, manuscript at doi:10.5281/zenodo.4849611, **2021**.

Variables and quantity values (example: molecular models)





MolMod DB (Molecular Model Database)

http://molmod.boltzmann-zuse.de/

pair potentials for over 150 molecular fluids

Domain-specific metadata standardization in materials modelling

M. Horsch,[1] J. Francisco Morgado,[2] G. Goldbeck,[3] D. Iglezakis,[4] N. Konchakova,[5] B. Schembera,[1] *High Performance Computing Center Stuttgart,[1] Fraunhofer IWM,[2] Goldbeck Consulting,[3] University of Stuttgart,[4] Helmholtz-Zentrum Hereon[5]*

DORIC-MM 2021 Workshop, European Semantic Web Conference, 7th June 2021