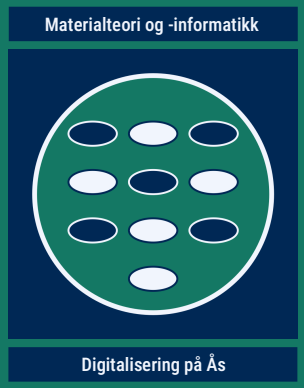
Norges miljø- og biovitenskapelige universitet



Epistemic metadata for computational engineering information systems

Martin Horsch,[1,2] Silvia Chiacchiera,[2] Gabriela Guevara,[3] Max Kohns,[4] Erich Müller,[5] Denis Šarić,[3] Simon Stephan,[4] Ilian Todorov,[2] Jadran Vrabec,[3] and Björn Schembera[6]

[1]Norwegian University of Life Sciences, [2]UK Research and Innovation, [3]TU Berlin, [4]RPTU Kaiserslautern-Landau, [5]Imperial College London, [6]University of Stuttgart

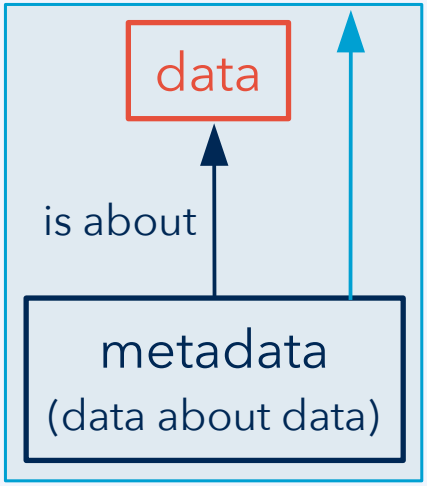
FOIS 2023, Sherbrooke, 17th July 2023

Fakultet for realfag og teknologi

Forskergruppe materialteori og -informatikk

The XAIR challenge

Challenge: Data and metadata need to become explainable-AI-ready (XAIR).



Leiden 2022 Declaration for **FAIR digital objects**:

https://www.fdo2022.org/site/fdo/programme/leiden-declaration

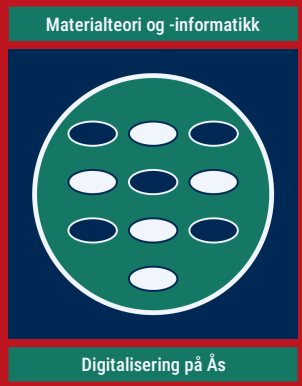
The librarian:

- Focus on **archival** and **curation**
- **Help humans** use digital artefacts
- Focus on **provenance**, like for artefacts in a museum, so humans understand where they come from

The engineer:

- **Computers** must **understand** what the digital artefacts mean
- **Focus on knowledge**/meaning
- FAIR digital objects

Noregs miljø- og biovitskaplege universitet



1. The setting
2. Mid-level ontology
3. Reproducibility and topics

Epistemic opacity (Humphreys, 2011): A cognitive "process is **epistemically opaque** relative to a cognitive agent *X* at time *t* just in case *X* does not know at *t* all of the **epistemically relevant elements** of the process."

European AI Act proposal: "To address the **opacity** that may make certain AI systems **incomprehensible to or too complex for natural persons**, a certain degree of transparency should be required for high-risk AI systems. […] High-risk AI systems should therefore be accompanied by **relevant documentation**".

Norwegian University of Life Sciences

Epistemic opacity

Epistemic opacity (Humphreys, 2011): A cognitive "process is **epistemically opaque** relative to a cognitive agent *X* at time *t* just in case *X* does not know at *t* all of the **epistemically relevant elements** of the process."

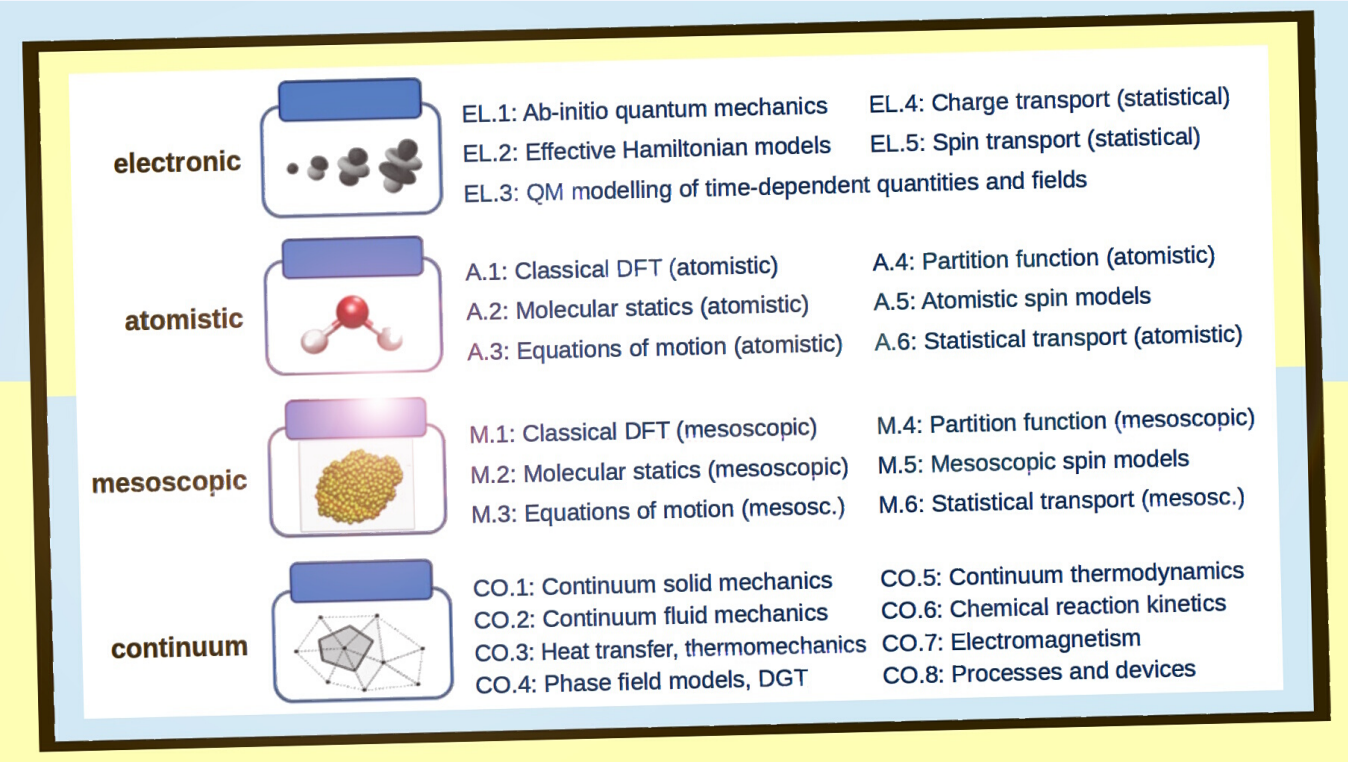
Epistemic metadata: Information that should be included in an adequate response to the queries "what **knowledge claims** have been formulated on the basis of the given data?" and "what exactly is the relation between the knowledge claims, their proponents, and the data?"

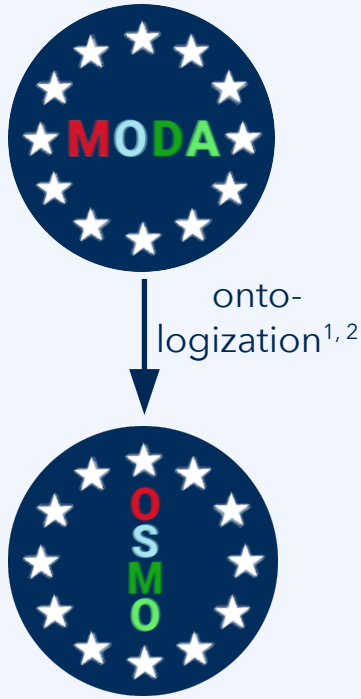
European AI Act proposal: "To address the **opacity** that may make certain AI systems **incomprehensible to or too complex for natural persons**, a certain degree of transparency should be required for high-risk AI systems.[1] Users should be able to interpret the system output and use it appropriately. High-risk AI systems should therefore be accompanied by **relevant documentation**".

[1]Systems with "high risk" include all "safety components" related to "water, gas, heating, and electricity."

Practices in materials modelling

As an attempt at metadata standardization, RoMM/MODA resulted in a closed epistemic space with a rigid categorization of modelling methodologies. MODA/CHADA documentations are hard to create and **hard to use by humans, but not machine-actionable**.





[1]M. T. Horsch *et al.*, *J. Chem. Eng. Data* **65(3)**: 1313–1329, doi:10.1021/acs.jced.9b00739, **2020**.

[2]M. T. Horsch *et al.*, in *Proc. JOWO 2021*, CEUR *vol.* **2969**: *p.* 47 (FOIS ontology showcase), **2021**.

Practices in materials modelling

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

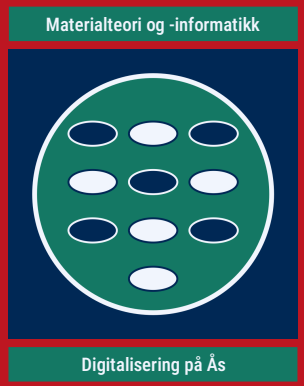
diversify technology | **observe** practices | have **realistic** objectives | **incentivize** open data | **co-design** data and workflows

The aim of the present work is to permit **communicating epistemic metadata** by developing a semantic artefact that fits into the pre-existing environment.

Cognitive processes are a very broad category[1] by which semantics about research practices and workflows can be formalized with a **mid-level ontology**.

[1]See for example a recent review by Elkobaisi *et al.* on ontologization of human emotional responses, *SN Computer Science* 3, 282, doi:10.1007/s42979-022-01116-x, **2022**.

Noregs miljø- og biovitskaplege universitet



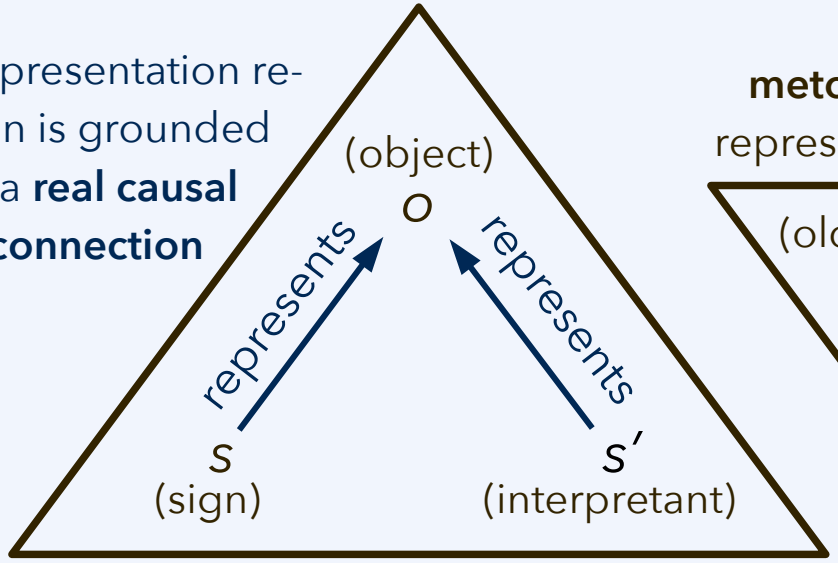
1. The setting
2. Mid-level ontology
3. Reproducibility and topics

The **PIMS-II mid-level ontology** implements a data documentation strategy based on **epistemic metadata** building on Peircean semiotics. Our present work has its focus on **knowledge claims** (what we know from data) and their assessment through validity claims, including **reproducibility claims**.

Mereosemiotics

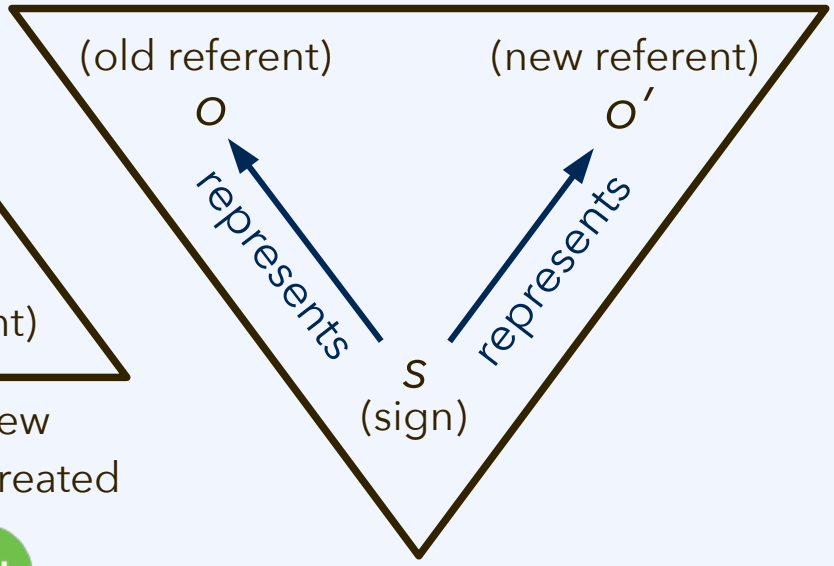
Peircean semiotics: By using a sign (1st) for an object (2nd), a "Third" is created.

the representation relation is grounded in a **real causal connection**



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

metonymization, a process by which a representamen is assigned a new referent



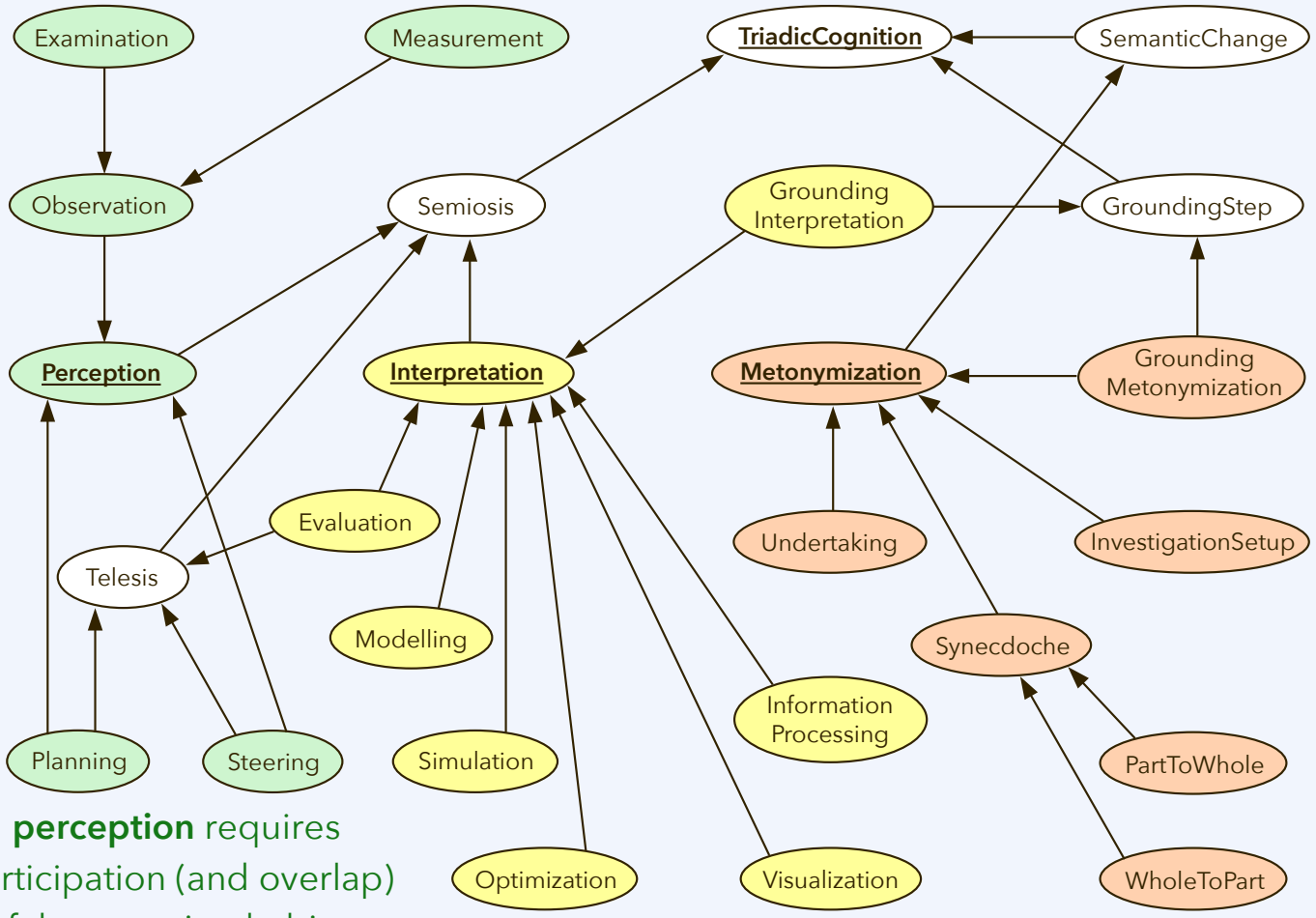
C. S. Peirce

Elementary Multi-perspective Material Ontology (EMMO)

The EMMO[1] combines this with mereocausality – foundational ontology as **mereosemiotics**.

[1]The work on the EMMO (2017 – present) is coordinated by Emanuele Ghedini.

Cognitive steps in the PIMS-II ontology



perception requires participation (and overlap) of the perceived object

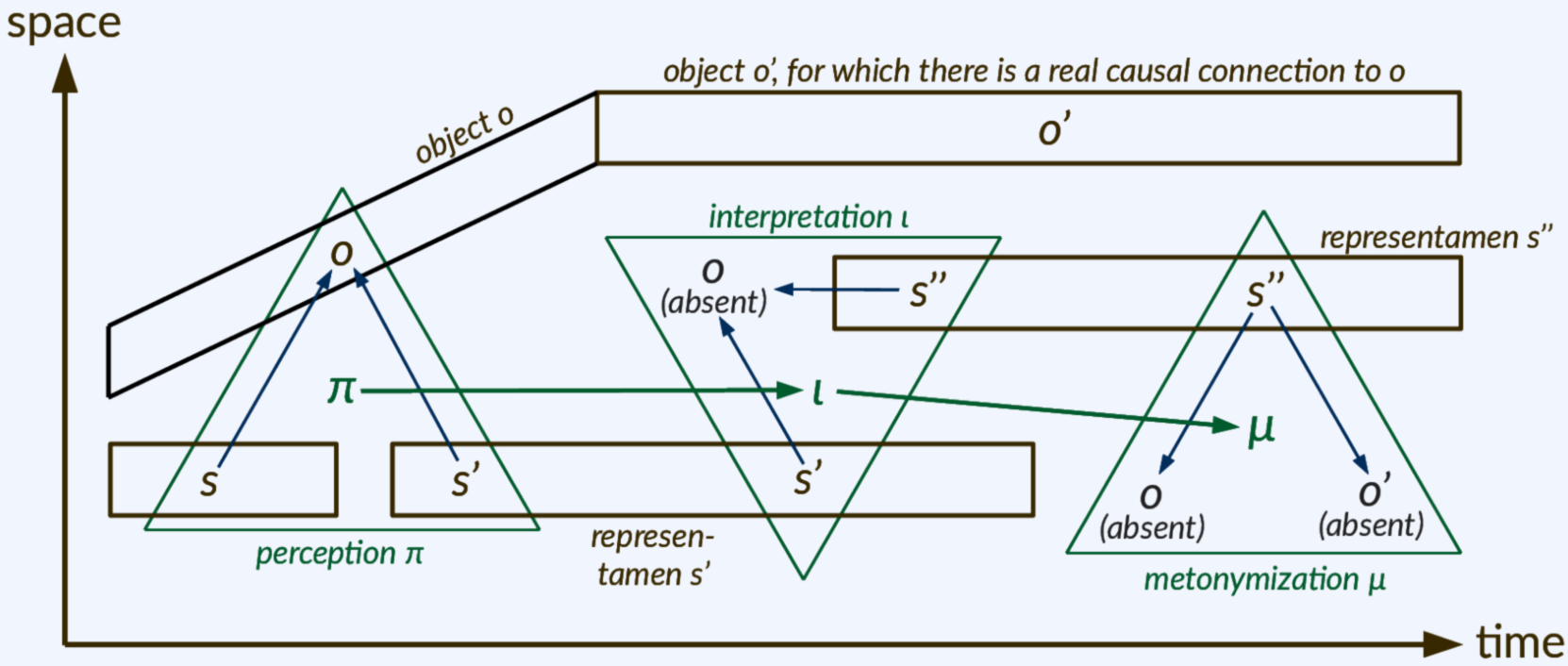
interpretation and **metonymization** do not entail physical participation of the referents

metonymization preserves the „real causal connection“ (Peirce) between the sign and its old & new referents

Cognitive steps in the PIMS-II ontology

PIMS-II mid-level ontology:[1, 2] http://www.molmod.info/semantics/pims-ii.ttl

Mereosemiotics:[1-3] Combination of mereotopology and Peircean semiotics



[1]M. T. Horsch, no. 3 in *Proc. JOWO 2021*, **2021**. [2]P. Klein *et al.*, no. 26 in *Proc. JOWO 2021*, **2021**.
[3]M. T. Horsch, S. Chiacchiera, B. Schembera, M. Seaton, I. T. Todorov, in *Proc. ECCOMAS 2020*, **2021**.

Epistemic metadata in the PIMS-II ontology

Metadata are "descriptive data about an object" (ISO 11179).

Epistemic metadata are those that help establish the knowledge status of data.[1]

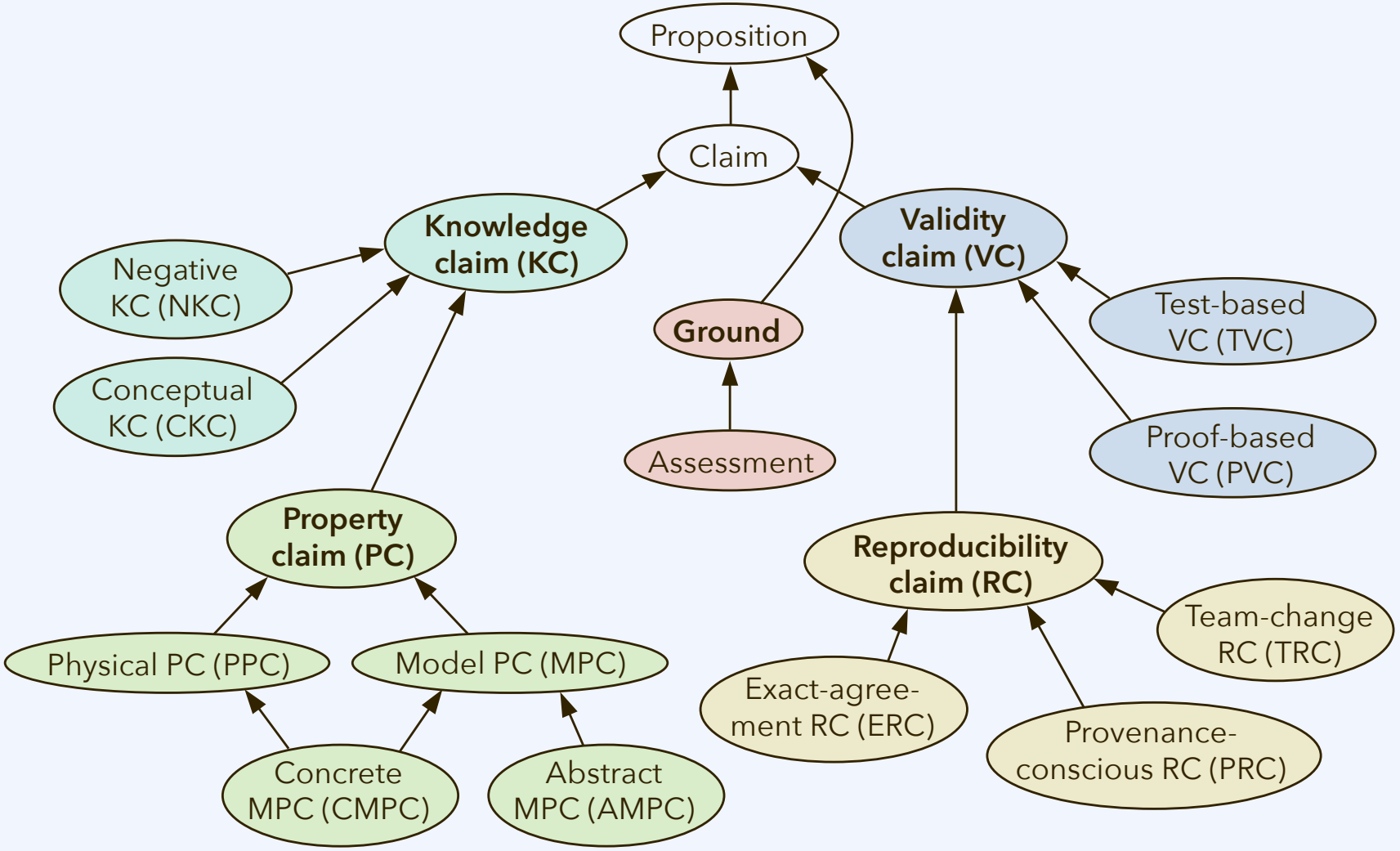
> **Epistemic metadata in the PIMS-II mid-level ontology:**
>
> **a)** "what **knowledge claim *φ*** has been formulated?,"
> **b)** "where do the data and the claim come from?" (**provenance**),
> **c)** "what **validity claim** was made about *φ*?,"
> **d)** "why should we accept any of this?" (**grounding**).

Case study from molecular thermodynamics

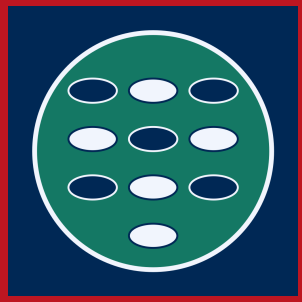
- First stage, evaluating ten journal articles, doi:10.5281/zenodo.7516532.
- Second stage, discussing twelve claims, doi:10.5281/zenodo.7608074.

[1]M. T. Horsch, B. Schembera, in *Proc. JOWO 2022*, CEUR *vol.* **3249**: *p.* 2 (CAOS), **2022**.

Epistemic metadata in the PIMS-II ontology



Noregs miljø- og biovitskaplege universitet



1. The setting
2. Mid-level ontology
3. Reproducibility and topics

Reproducibility claim (RC)

«Whenever the research process κ'' is carried out, it **must** lead to the outcome φ''.»

Reproducibility and falsification[1, 2]

Research data infrastructures must accommodate mutually contradicting claims. They should also assist researchers at validating/falsifying each other's work.

Let us look into a "falsification" or "unsuccessful reproduction" of *a*'s work by *b*:

Knowledge claim (KC), including the provenance

«Researcher a did κ and found φ (and thus claims to know φ).»
→ Therefore, when research process κ is carried out, it **can** lead to the outcome φ.

1) Reseacher *a* did κ and found φ.

2) Reseacher *b* did γ, which is very similar to κ, and found ζ, not very similar to φ.

3) Now *b*'s work is regarded as a refutation/falsification, going against *a*'s work.

Knowledge and reproducibility claims both use **conditional modal statements**.

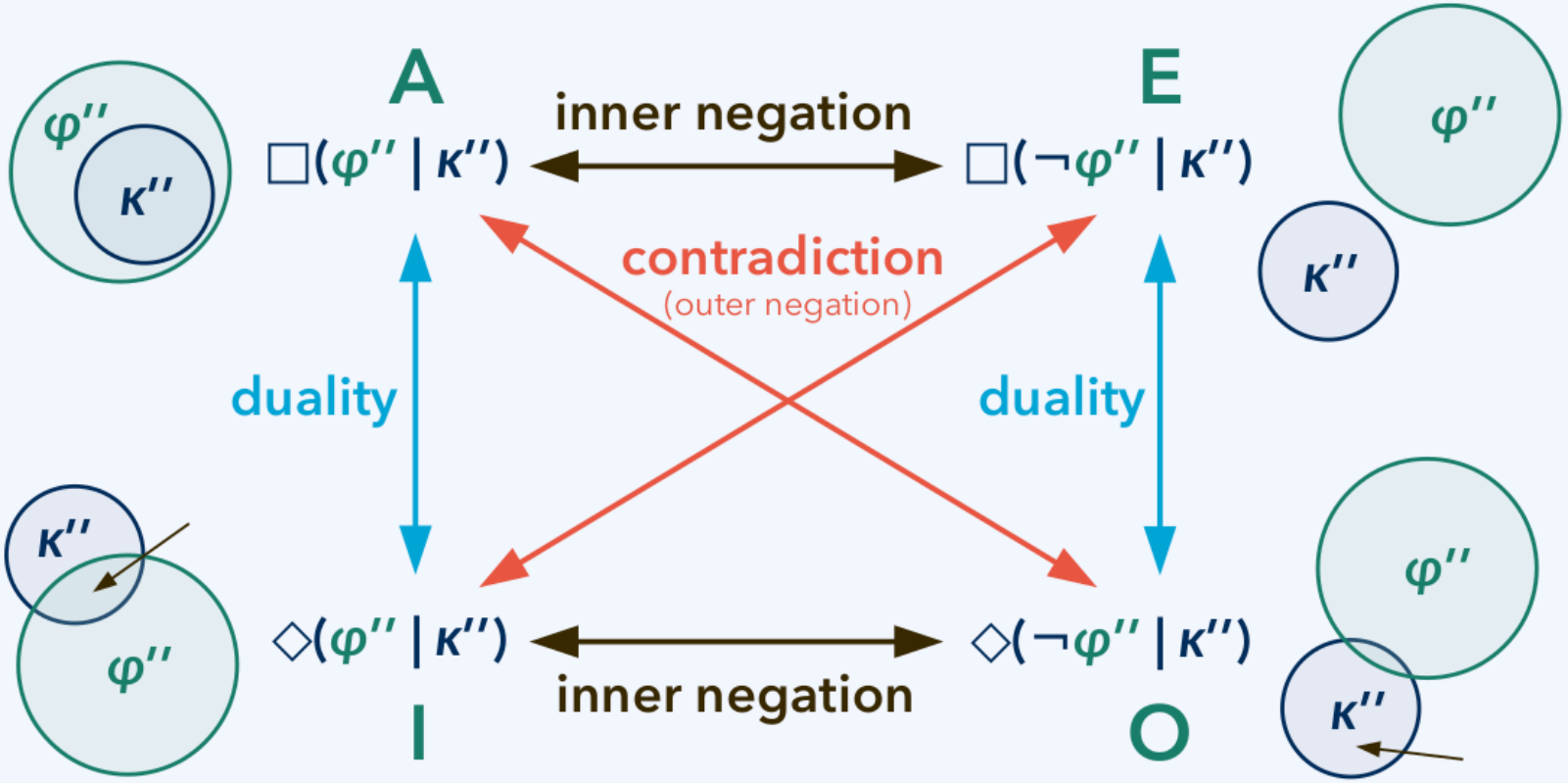
[1]M. T. Horsch, S. Chiacchiera, G. Guevara, M. Kohns, *et al.*, in *Proc. FOIS 2023*, to appear, **2023**.

[2]H. E. Plesser, *Frontiers Neuroinform* **11**: 76, doi:10.3389/fninf.2017.00076, **2018**.

Conditional necessity and possibility

If the research process conforms with κ'', the outcome **must conform** with φ''.

If the research process conforms with κ'', the outcome **must not conform** with φ''.



If the research process conforms with κ'', the outcome **can conform** with φ'' (and it is possible to conform with κ'').

If the research process conforms with κ'', the outcome **can disagree** with φ'' (and it is possible to conform with κ'').

Reproducibility claims

Common formulation and schema for reproducibility claims (RCs):

«Whenever research process κ'' is carried out, it must lead to the outcome φ''.»

1) Reseacher *a* did κ and found φ.

2) Reseacher *b* did γ, **consistent with κ''**, and found ζ, **inconsistent with φ''**.

3) Now *b*'s work is regarded as a refutation/falsification or going against *a*'s work.

Reproducibility claims

Common formulation and schema for reproducibility claims (RCs):

«Whenever research process κ'' is carried out, it must lead to the outcome φ''.»

we argue that there is a mechanism from pragmatics at work here[1]

1) Reseacher *a* did κ and found φ.

   Here, *a* also made the **positive reproducibility claim $\psi = \Box(\varphi'' \mid \kappa'')$.**

2) Reseacher *b* did γ, **consistent with κ''**, and found ζ, **inconsistent with φ''**.

   Here, *b* made the **negative reproducibility claim $\Diamond(\neg\varphi'' \mid \kappa'') \equiv \neg\Box(\varphi'' \mid \kappa'') \equiv \neg\psi$.**

3) What is relevant there is the **contradiction between ψ and $\neg\psi$.**

provenance metadata κ
provenance paradata κ'

provenance orthodata $\kappa'' = \kappa - \kappa'$

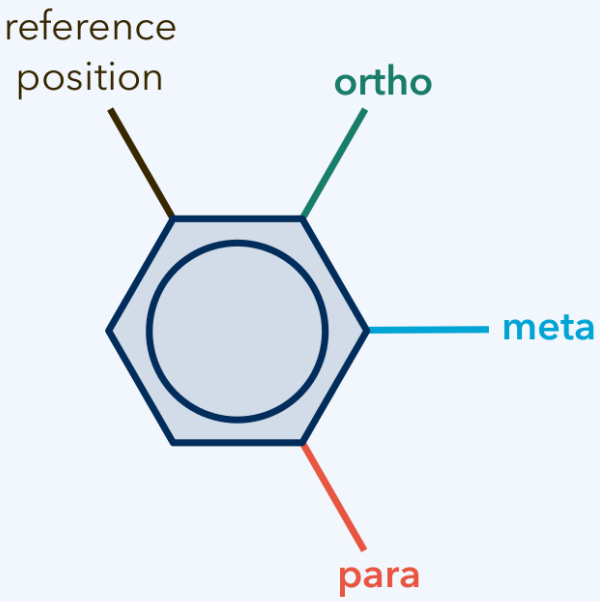
«repeat κ, but no need to retain κ'»

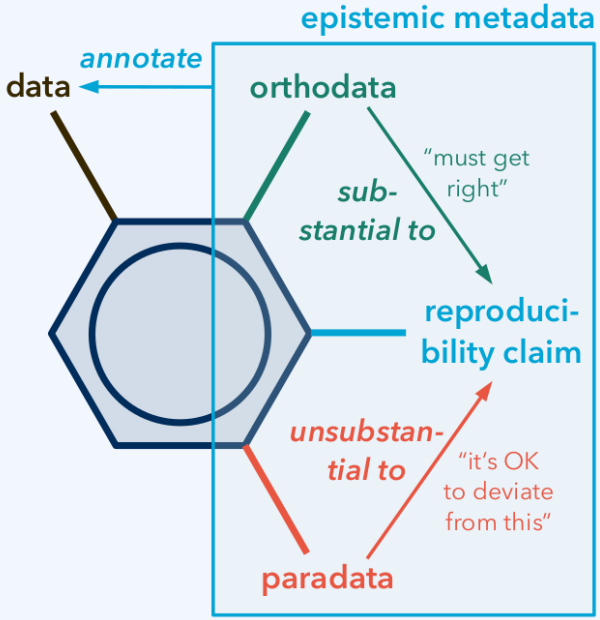
knowledge claim metadata φ
knowledge claim paradata φ'

knowledge claim orthodata $\varphi'' = \varphi - \varphi'$

«obtain φ again, except for φ' maybe»

Orthodata, paradata, and logical subtraction





provenance metadata κ
provenance paradata κ'

---

provenance orthodata $\kappa'' = \kappa - \kappa'$

«repeat κ, but no need to retain κ'»

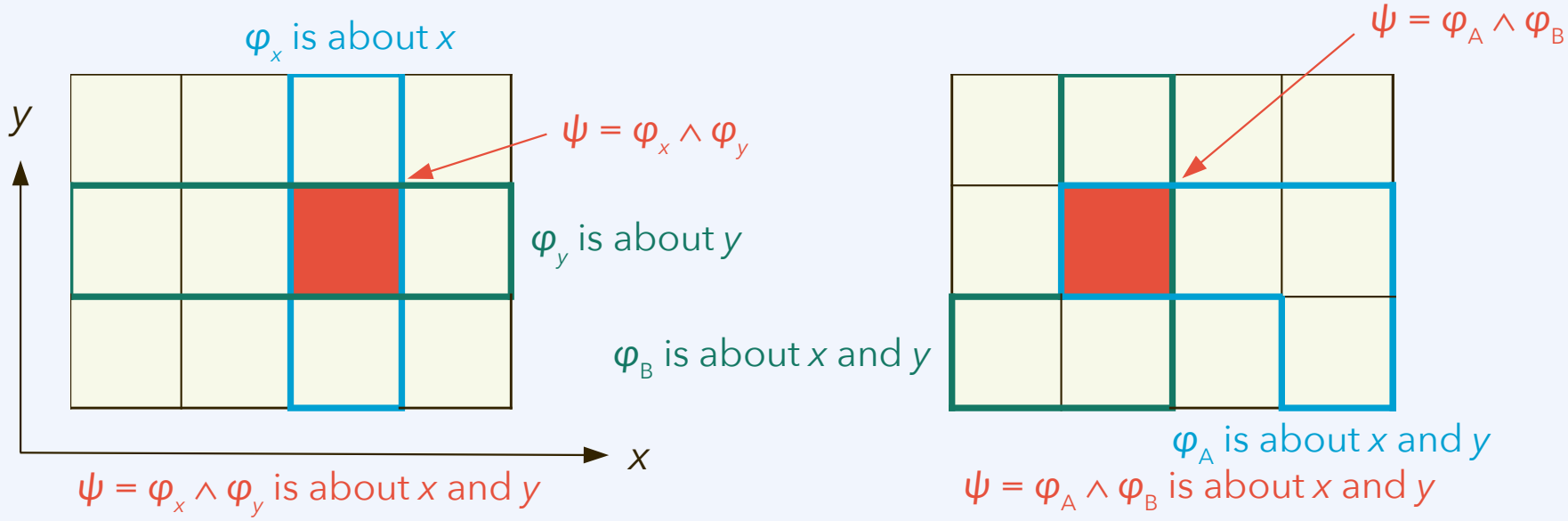
knowledge claim metadata φ
knowledge claim paradata φ'

---

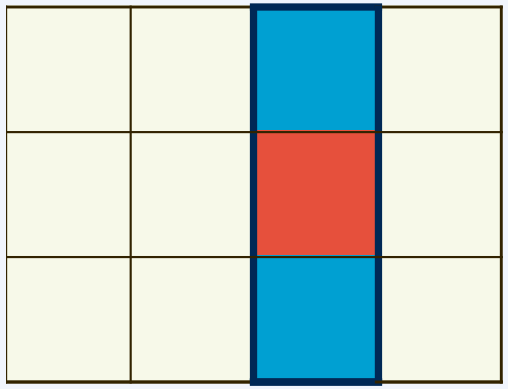
knowledge claim orthodata $\varphi'' = \varphi - \varphi'$

«obtain φ again, except for φ' maybe»

Orthodata, paradata, and logical subtraction

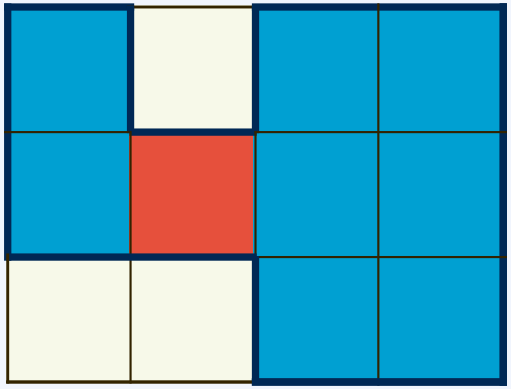


$\psi - \varphi_y = (\varphi_x \wedge \varphi_y) - \varphi_y \equiv \varphi_x$



φ_x is recovered by subtracting φ_y

$\psi - \varphi_B = (\varphi_A \wedge \varphi_B) - \varphi_B \not\equiv \varphi_A$



φ_A cannot be perfectly recovered

Norwegian University of Life Sciences

Subject matter of research outcomes

Logical subtraction is a concept from analytic philosophy.[1-3]

Its formalization is closely connected to the theory of **subject matter**.[2, 3]

Could you try to **replicate my old simulation result**? Just do the same as I did.
Except that you of course log in with your user account, not mine.
Your result was off by 0,5%? **Don't worry**, that is totally normal.

Our **simulation of object *o*** confirms theory *s*.
Except that theory *s* deals with physical reality, and *o* is so simplified
that **we know it cannot exist** or be built exactly in physical reality.

Example from Yablo:[2] Someone who rejects ontological commitment to the
existence of numbers is asked **how many prime numbers there are** greater than ten.
"Infinitely many, of course, **except that** numbers don't exist."

[1]R. A. Jaeger, *Philos. Rev.* **82**(3): 320–329, doi:10.2307/2183898, **1973**.
[2]S. Yablo, *Aboutness*, Princeton Univ. Press (ISBN 978-0-691-14495-5), **2014**.
[3]K. Fine, *J. Philos. Log.* **46**: 675–702, doi:10.1007/s10992-016-9419-5, **2017**.

Norwegian University of Life Sciences

Subject matter of research outcomes

We understand **subject matter** of a knowledge claim and/or the associated research data as given by the research **question that is being answered**, or by the "equivalence relation over logical space" with respect to that question.[1]

With respect to the research question[2]

q_1 = "What is the **D** matrix of liquid *M* as a function of **x**, *p*, and *T*?,"

two states of affairs are equivalent if their **D**(**x**, *p*, *T*) dependencies are the same.

Knowledge bases *K*, *L* are equivalent, $K \equiv_{q1} L$, if they return equivalent sets of valuations in response to a SPARQL query expressing the research question.

[1] S. Yablo, *Aboutness*, Princeton Univ. Press (ISBN 978-0-691-14495-5), **2014**.
[2] G. Guevara Carrión *et al.*, *J. Phys. Chem. B* **124**(22): 4527–4535, doi:10.1021/acs.jpcb.0c01625, **2020**.

Subject matter of research outcomes

Norwegian University of Life Sciences

For cataloguing and information retrieval, we distinguish two ways of **combining topics**. Two closely related, interacting topics q_1 and q_2 form a **topical product** q_1q_2 where the partitioning of logical space by $\equiv_{q1q2}$ is obtained from the *product of the sets of equivalence classes* with respect to $\equiv_{q1}$ and $\equiv_{q2}$.

q_1 = "What is the **D** matrix of liquid M as a function of **x**, p, and T?,"
q_2 = "What is the **Γ** matrix of liquid M as a function of **x**, p, and T?,"

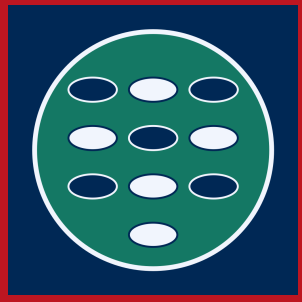
combine in this way to q_1q_2 = "What are the **D** and **Γ** matrices of […]?".

However, a long paper or book (or multiple books, etc.) can also be about a collection of topics that are effectively unrelated and only stand side by side. We call this *plurality of topics* a **topical sum** $q_1 + q_2$ such as from

q_1 = a theoretical research question from statistical mechanics,
q_2 = topic of a concrete series of simulations from the same paper.

Noregs miljø- og biovitskaplege universitet

Materialteori og -informatikk



Digitalisering på Ås

Evaluation

DOME 4.0
H2020 GA no. 953163

ONTO COMMONS
ONTOLOGY-DRIVEN DATA DOCUMENTATION FOR INDUSTRY COMMONS
H2020 GA no. 958371

Molecular modelling case study

The **case study** on **epistemic metadata in molecular modelling** proceeds in multiple stages. The plan is to conclude with the presently ongoing third stage.

1) First stage report (10 cases), doi:10.5281/zenodo.7516532, 2023.
   Discussion of *five papers each* from *two research groups* (Berlin, London) without involving the papers' authors. Obtained a tentative taxonomy for epistemic metadata and explored the patterns of epistemic grounding.

2) Second stage report (12 claims), doi:10.5281/zenodo.7608074, 2023.
   Discussion of *two claims each* from *six papers*, involving the papers' authors, some of whom became co-authors of the present work. Ontology of epistemic metadata, except for epistemic grounding, implemented in PIMS-II.

3) Annotation of examples, leading to a semantic artefact that is ready for use.

Norwegian University of Life Sciences

Oops! and Foops! tests

Ontology design and FAIRness were evaluated using **Oops!**[1] and **Foops!**[2]

For this purpose, we ingested the **PIMS-II** TTL file[3] URL into both engines.

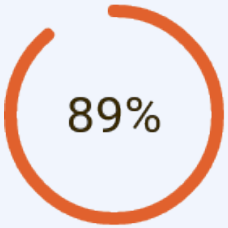
- **Critical** : It is crucial to correct the pitfall. Otherwise, it could affect the ontology consistency, reasoning, applicability, etc.
- **Important** : Though not critical for ontology function, it is important to correct this type of pitfall.
- **Minor** : It is not really a problem, but by correcting it we will make the ontology nicer.

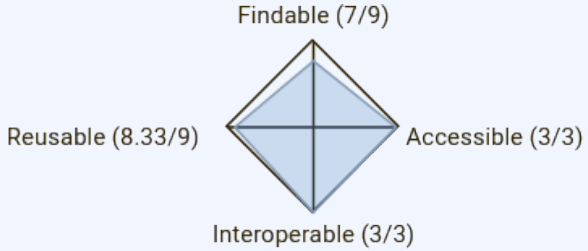
[Expand All] | [Collapse All]

Results for P11: Missing domain or range in properties.	**195 cases | Important**
Results for P13: Inverse relationships not explicitly declared.	**45 cases | Minor**
Results for P36: URI contains file extension.	**ontology* | Minor**
SUGGESTION: symmetric or transitive object properties.	**5 cases**

result from Oops!

result from Foops!





[1]M. Poveda *et al.*, *Int. J. Semant. Web Inform. Sys.* **10**(2): 7–34, doi:10.4018/ijswis.2014040102, **2014**.
[2]D. Garijo *et al.*, *Proc. ISWC 2021 Posters/Demos/Industry*, p. 321, **2021**.
[3]http://www.molmod.info/semantics/pims-ii.ttl

Norwegian University of Life Sciences

Epistemic Metadata for Computational Engineering Information Systems

Martin Thomas HORSCH [a,b,1] Silvia CHIACCHIERA [b]
Gabriela GUEVARA CARRIÓN [c] Maximilian KOHNS [d] Erich A. MÜLLER [e]
Denis ŠARIĆ [c] Simon STEPHAN [d] Ilian T. TODOROV [b] Jadran VRABEC [c]
Björn SCHEMBERA [f]

[a] *Norwegian University of Life Sciences, Faculty of Science and Technology, Department of Data Science, Drøbakveien 31, 1430 Ås, Norway*
[b] *UK Research and Innovation, STFC Daresbury Laboratory, Scientific Computing Department, Keckwick Ln, Daresbury WA4 4AD, UK*
[c] *Technische Universität Berlin, Thermodynamics, Ernst-Reuter-Platz 1, 10587 Berlin, Germany*
[d] *Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, Laboratory of Engineering Thermodynamics, Erwin-Schrödinger-Str. 44, 67663 Kaiserslautern, Germany*
[e] *Imperial College London, Department of Chemical Engineering, South Kensington Campus, London SW7 2AZ, UK*
[f] *University of Stuttgart, Institute of Applied Analysis and Numerical Simulation, Pfaffenwaldring 57, 70569 Stuttgart, Germany*

Abstract. Digitalization is the main priority for innovation in the engineering sciences at present. This includes making the knowledge from scientific research data machine-actionable so that it can be integrated and analysed with minimal human intervention. Computational engineering has been advancing on this path for some time; *e.g.*, FAIR digital objects are gaining momentum as a paradigm for communicating data and metadata. Despite this, the depth of digitalization often remains too shallow, with annotations that are only of use to a human reader. In addition, digital infrastructures and their metadata standards are tedious to use: They require too much effort from researchers; *e.g.*, for providing input that contributes nothing to an automated reuse of knowledge. These two shortcomings, lack in depth and excess in breadth, are related. Addressing these gaps, the present contribution discusses metadata standardization efforts targeted at documenting the knowledge status of data; the required annotation is referred to as epistemic metadata. It is discussed how a metadata schema for knowledge and reproducibility can be designed such as to be user-friendly and flexible enough to apply to a spectrum of circumstances and types of replicability and consistency checks. These developments are positioned in the context of a recent case study on a sample of journal articles and knowledge claims from the domain of molecular modelling and simulation.

Keywords. Applied ontology, epistemic metadata, process data technology.

INPRODAT
https://www.inprodat.de/

EMMC
https://emmc.eu/

ONTO COMMONS
ONTOLOGY-DRIVEN DATA DOCUMENTATION FOR INDUSTRY COMMONS
https://ontocommons.eu/

DOME 4.0
https://dome40.eu/