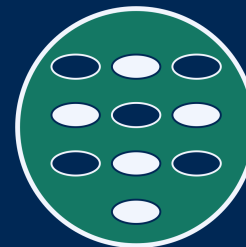


Norges miljø- og
biovitenskapelige
universitet

Materialteori og -informatikk



Digitalisering på Ås

Epistemic metadata

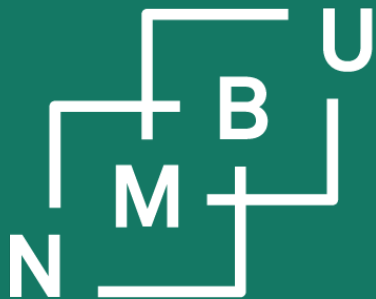
Annual Digital Catalysis & Catalysis-Related Sciences Conference

2nd November 2023

Frankfurt am Main

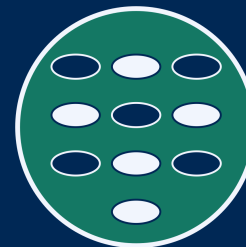
Fakultet for realfag og teknologi

Forskergruppe materialteori og -informatikk



Norges miljø- og
biovitenskapelige
universitet

Materialteori og -informatikk



Digitalisering på Ås

Epistemic metadata

joint work with Silvia Chiacchiera, Heinz A. Preisig, and Björn Schembera

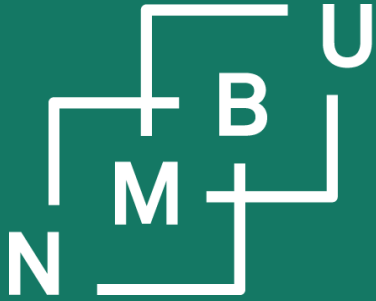
Annual Digital Catalysis & Catalysis-Related Sciences Conference

2nd November 2023

Frankfurt am Main

Fakultet for realfag og teknologi

Forskergruppe materialteori og -informatikk



Noregs miljø- og
biovitenskaplege
universitet

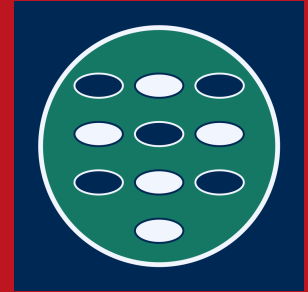
The problem

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**".



Horizon Europe
ID 101138510

Materialteori og -informatikk



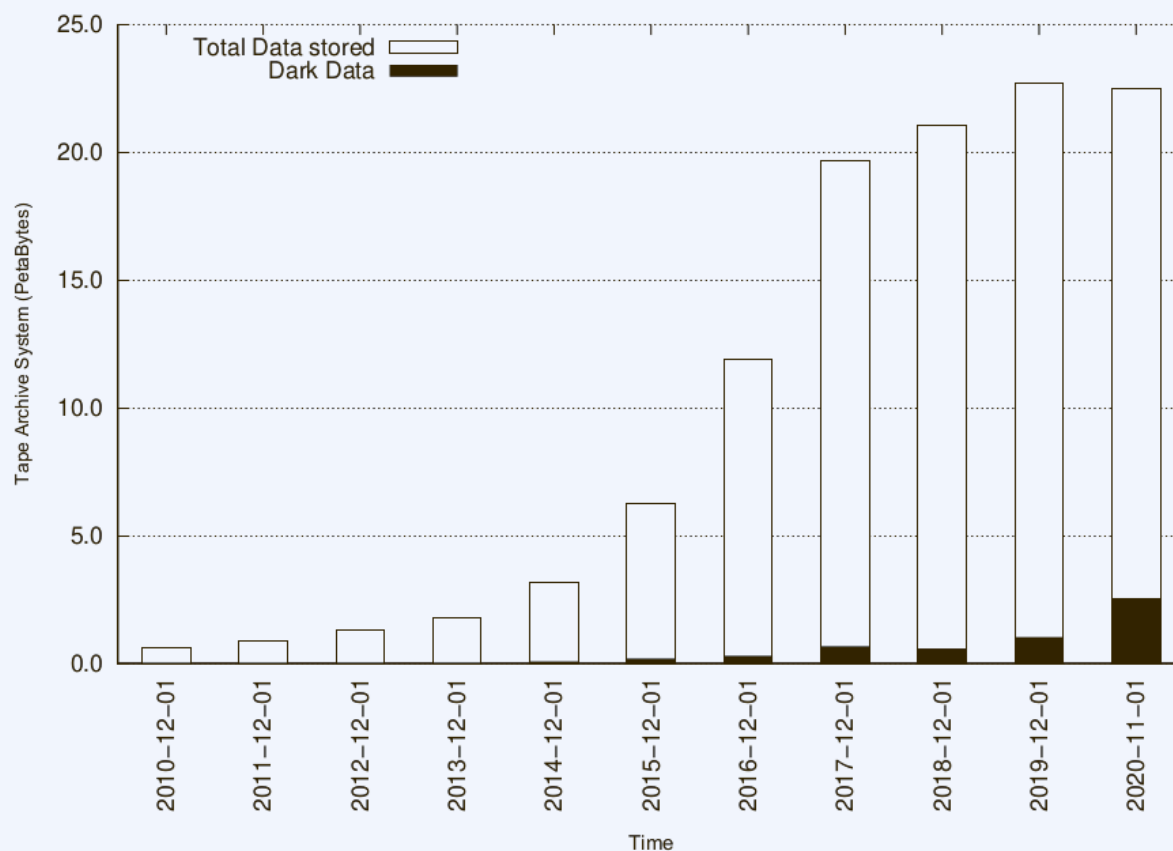
Digitalisering på Ås

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."

Dark data

Dark data are data with an uncharacterized knowledge status.

In other words: *We don't know what we know from and about the data.*



Flood of dark data:
More and more data are accumulated, but are dark - and useless.

Source: Work by Juan Durán and Björn Schembera.

See also: B. Schembera, J. Durán, *Philos. Technol.* **33**: 93-115, doi:10.1007/s13347-019-00346-x, **2019**.

European regulations

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**”.

Beginning with the EC’s **Battery Regulation**, **digital product passports (DPPs)** will become mandatory; first for batteries, later textiles, electronics, and for more and more products.



Epistemic opacity (Humphreys, 2011): A «process is **epistemically opaque** relative to a cognitive agent X at time t [... if ...] X does not know at t all of the **epistemically relevant elements**»²

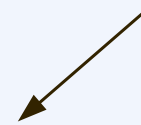
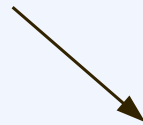
¹Systems with “**high risk**” include all “safety components” related to “water, gas, heating, and electricity.”

²P. Humphreys, in M. Carrier, A. Nordmann, *Science in the Context of Application*, pp. 131-142, Springer, **2011**.

"FAIR and XAIR data"

XAI: Explainable artificial intelligence

AIR: Artificial-intelligence ready



XAIR: Explainable AI-ready

Beginning with the EC's **Battery Regulation**, **digital product passports (DPPs)** will become mandatory; first for batteries, later textiles, electronics, and for more and more products.



Tendency: Data must become explainable-AI-ready (XAIR). **Making data trustworthy through explanations** will increasingly become a legal requirement.

Slogan: "FAIR and XAIR data." (Sounds similar to the idiom "fair and square.")

Case study in molecular thermodynamics

Epistemic metadata and their **documentation** were explored in molecular thermodynamics:

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.

Case study in molecular thermodynamics

Epistemic metadata and their **documentation** were explored in molecular thermodynamics:

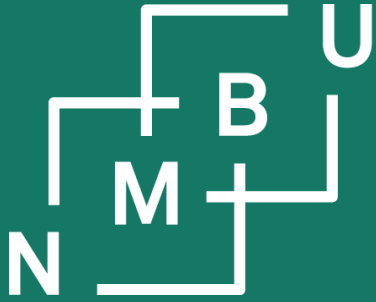
First stage report (10 cases), doi:10.5281/zenodo.7516532, **2023**.

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.

Second stage report (12 claims), doi:10.5281/zenodo.7608074, **2023**.

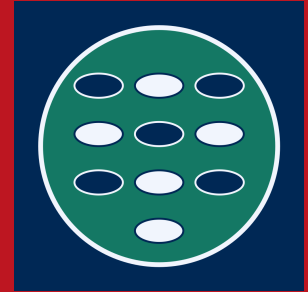
Discussion of *two claims each* from *six papers*, with two papers each from three research groups (Berlin, Kaiserslautern, London), involving the papers' authors. **Ontology of epistemic metadata** implemented.

Good data documentation standards give researchers freedom to say what they want. They **provide a language**, but **don't micromanage** researchers' self-expression. (Morton: Ontology should not be "object police.")



Noregs miljø- og
biovitenskaplege
universitet

Materialteori og -informatikk



Digitalisering på Ås

Key concepts

The epistemic grounding of a research outcome is an explanation for why the scientific community accepts that result as knowledge; i.e., a rationale for why it should be accepted as knowledge.

Key epistemic metadata items are the **knowledge claims** made based on data, their **provenance**, **validation** and **reproducibility**, and **epistemic grounding**.

What constitutes the knowledge status of data?

Epistemic metadata are the information that **establishes the knowledge status** of data or digital objects.¹

Questions we must answer to establish the knowledge status:

- 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**).

Key epistemic metadata items are the **knowledge claims** made based on data, their **provenance**, **validation and reproducibility**, and **epistemic grounding**.

¹«Documentation of epistemic metadata by a mid-level ontology of cognitive processes», in *Proc. JOWO 2022*, CEUR vol. **3249**: p. 2 (CAOS), CEUR-WS, **2022**.

Epistemic grounding

Distinction between Type-1 and Type-2 grounding inspired by Marr.^{1, 2}


Type-1 The results explain (or are presented in a way to explain) why they are valid.	<i>Example: Mathematical proof in statistical mechanics for a theoretical framework with widely accepted definitions and axioms.</i>
Type-2 The provenance of the results tells that they are valid.	<i>Example: We used a model, method, and simulation code validated in the past and - usually - very accurate.</i>

¹D. Marr, *Artificial Intelligence* 9(1): 37–48, doi:10.1016/0004-3702(77)90013-3, **1977**.

²«Documentation of epistemic metadata by a mid-level ontology of cognitive processes», in *Proc. JOWO 2022*, CEUR vol. **3249**: p. 2 (CAOS), CEUR-WS, **2022**.

Epistemic grounding

Distinction between Type-1 and Type-2 grounding inspired by Marr.^{1, 2}

Type-1 The results explain (or are presented in a way to explain) why they are valid.	<i>Example:</i> Mathematical proof in statistical mechanics for a theoretical framework with widely accepted definitions and axioms.
Type-2 The provenance of the results tells that they are valid.	Reliability of process <i>m</i> means that «If <i>S</i> 's believing <i>p</i> at <i>t</i> results from <i>m</i> , then <i>S</i> 's belief in <i>p</i> at <i>t</i> is justified ». ³  (process reliabilism) <i>Example:</i> We used a model , method , and simulation code validated in the past and - usually - very accurate.

¹D. Marr, *Artificial Intelligence* **9**(1): 37–48, doi:10.1016/0004-3702(77)90013-3, **1977**.

²«Documentation of epistemic metadata by a mid-level ontology of cognitive processes», in *Proc. JOWO*, **2022**.

³J. M. Durán, N. Formanek, *Minds and Machines* **28**(4): 645–666, doi:10.1007/s11023-018-9481-6, **2018**. 12

Epistemic grounding

Distinction between Type-1 and Type-2 grounding inspired by Marr.¹

	authority or trust	reliabilism
Type-1 The results explain (or are presented in a way to explain) why they are valid.	<i>Example:</i> Mathematical proof in statistical mechanics for a theoretical framework with widely accepted definitions and axioms.	<i>Example:</i> The new theory is better because it is simpler, has fewer parameters, or “looks more” like reality. (virtue reliabilism)
Type-2 The provenance of the results tells that they are valid.	<i>Example:</i> We validated the artificial neural network as specified by the ISO 24029 norm, and established its prediction error accordingly.	<i>Example:</i> We used a model, method, and simulation code validated in the past and - usually - very accurate. (process reliabilism)

Distinction between “**moral grounds**” and grounding by **appeal to reliability**.

¹D. Marr, *Artificial Intelligence* 9(1): 37-48, doi:10.1016/0004-3702(77)90013-3, 1977.

Subject matter (aboutness) and logical subtraction

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.

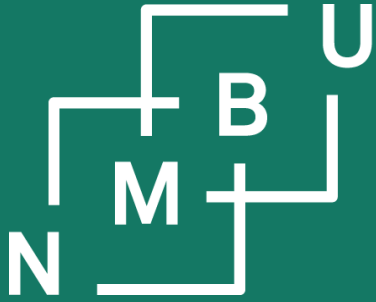
Our **simulation of object o** confirms theory s.

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.

¹R. A. Jaeger, *Philos. Rev.* **82**(3): 320–329, doi:10.2307/2183898, **1973**.

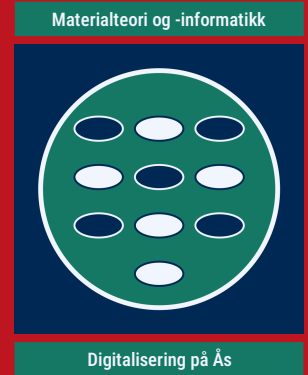
²S. Yablo, *Aboutness*, Princeton Univ. Press (ISBN 978-0-691-14495-5), **2014**.

³K. Fine, *J. Philos. Log.* **46**: 675–702, doi:10.1007/s10992-016-9419-5, **2017**.



Noregs miljø- og
biovitenskapelige
universitet

Research and publication practices



Reproducibility claim

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

Validation and reproducibility claims

reproducibility

There are many definitions of reproducibility and replicability; see the review by Plesser.¹

- 1) Researcher a did κ and found φ .
- 2) Researcher b did γ (similar enough to κ) and found ζ (not similar enough to φ).

Reproducibility claim

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

Validation and reproducibility claims



If this is a falsification, what is it technically that was shown to be false?

- It is not the knowledge claim φ or even the underlying data.
- It is not the literal claim by the researcher on what was done and found.

1) Researcher a did κ and found φ .

2) Researcher b did γ (similar enough to κ) and found ζ (not similar enough to φ).

3) If ζ was similar enough to φ , it would still contradict φ , but not be a falsification.

The literal claim of a , "I carried out process κ and found φ " is not contradicted.

Reproducibility claim

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

Validation and reproducibility claims



Common formulation and schema for reproducibility claims (RCs):

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

1) Researcher a did κ and found φ .

Here, a also made a **positive reproducibility claim** ψ .

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

Here, b made the **negative reproducibility claim** $\neg\psi$.

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'

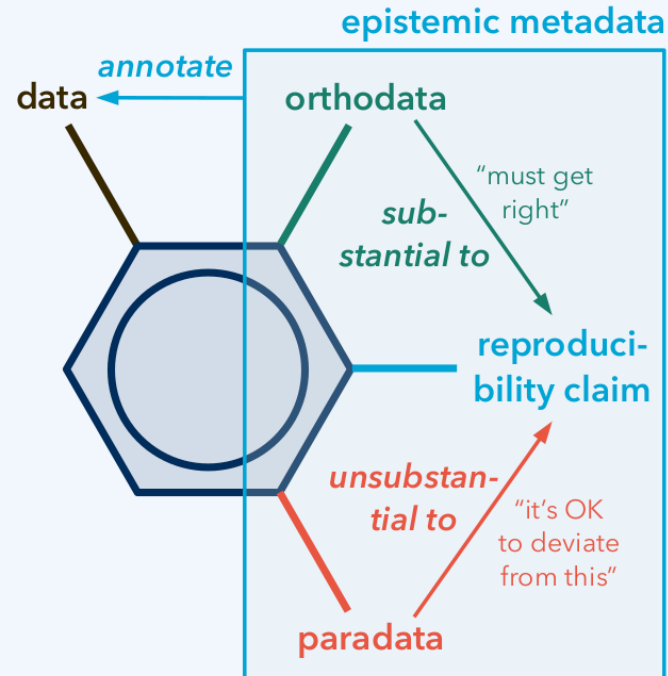
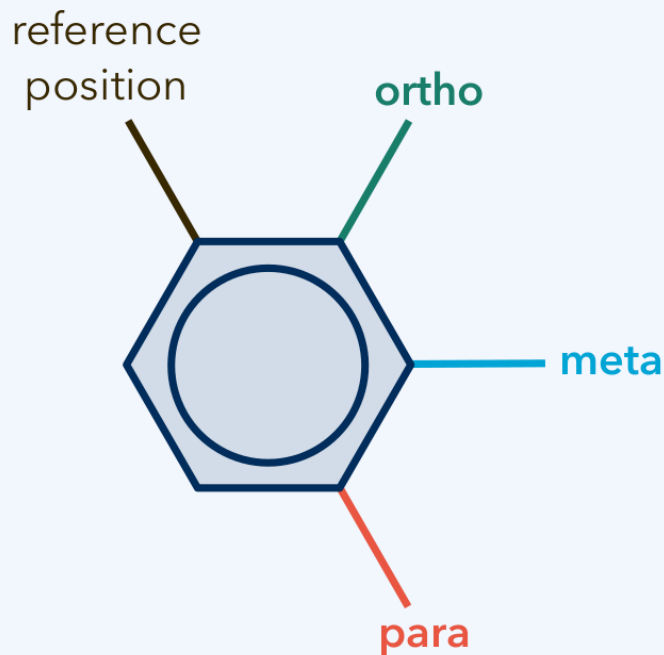
knowledge claim orthodata $\varphi'' = \varphi - \varphi'$

«obtain φ again, except for φ' maybe»

Reproducibility claims: Make them explicit

The scientific process can benefit from making reproducibility claims explicit.

In this way, other researchers know what exactly they need to comply with when attempting to replicate and validate or falsify others' work.



«repeat κ , but no need to retain κ' »

«obtain φ again, except for φ' maybe»

Reproducibility claims: Make them explicit

The scientific process can benefit from making reproducibility claims explicit.

Blue: **Orthodata**. Red: **Paradata**.

If you use the same model,
method, **solver** code (& version),
and **execution environment**, ...

If you use the same model and
method (but any code and
execution environment), ...

If you apply any method (including
experiment) to the problem ...

«repeat κ , but no need to retain κ' »

You will find the same value for the
property, within a margin of **2%**.
The **runtime** will be the same within **40%**.

You will find the same value, within **5%**,
except for errors due to *your code*, etc.
(No claim on **computational resources**.)

You will find the same value, within **20%**,
except for errors due to *your methods*.
(No claim on **computational resources**.)

«obtain φ again, except for φ' maybe»

Scientific communication is human communication

Typical ambiguous situation observed in the case study.

Epistemic grounding from paper or discussions: Results are scientific (knowledge), because methods and models are well established.

Is this an **appeal to authority**? The scientific community has long accepted these methods, many previous works have used them.

Is it **reliabilism**? Good predictions were made in similar previous works.

- **Epistemic grounding** is **usually not spelled out** in detail (or at all).
We often need to “make up” an interpretation on behalf of the authors.

Scientific communication is human communication

When digitalizing research data, we should respect that:

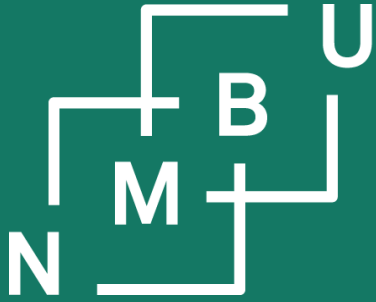
- Research is a **social process** among humans;
- scientific communication is **human communication**;
- it can rely on **pragmatics** – no need say every small thing explicitly.

This is not a bug, it's a feature!

But as a consequence:

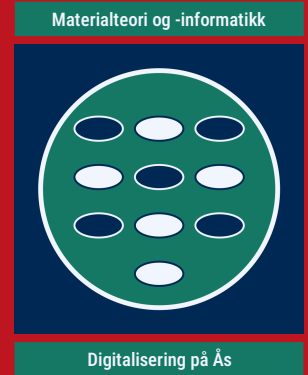
- **Reproducibility claims** are **not usually made explicit**.
It is left up to the reusing party to think what is appropriate.
- **Epistemic grounding** is **usually not spelled out** in detail (or at all).
We often need to “make up” an interpretation on behalf of the authors.

Our aim should be to help people make such explicit statements **if they want**.



Noregs miljø- og
biovitenskaplege
universitet

Development work



Early or naive attempts at metadata standardization can **fail to meet researchers' needs** by making far too much annotation mandatory, where it is not really needed in practice.

- MODA was a closed semantic and epistemic space: Modelling methods had to be chosen from a small list.^{1,2}
- MODA imposed a **given level of detail** in workflow documentation.¹
- MODA documentations were **complicated**.³

³ReaxPro project deliverable D2.1, «ReaxPro MODA diagrams», **2020**.

The first attempts (e.g., MODA¹ and RoMM²)

Naive metadata standards often **focus(ed) on provenance documentation only**.¹

Early or naive attempts at metadata standardization can **fail to meet researchers' needs** by making far too much annotation mandatory, where it is not really needed in practice.

3.1 SOURCE AND CONVENTIONAL TRANSLATION OF THE SIMULATIONS	
3.1.1 NUMERICAL SOURCE	Please give name and type of the solver e.g. Nethe Carlo, SPH, FE, ... (optional: multi-grid, adaptive...)
3.1.2 SOFTWARE TOOL	Please give the name and if this is your own code, please specify if it can be shared with an external link to website/publication
3.1.3 TIME STEP	If applicable, please give the time step used in the solving operations. This is the numerical time step and this is not the same as the time step of the case to be simulated (see 3.4.1)
3.1.4 COMPUTATIONAL REPRESENTATION	<p>PHYSICS Equation, Material, Relations, Material.</p> <p>Computational representation of the physics equation, material, relation and material.</p> <p>There is no need to repeat over case (file).</p> <p>"Computational" means that this only needs to be filled in when your computational solver represents the material, properties, equation variables, in a specific way.</p>
3.1.5 COMPUTATIONAL BOUNDARY CONDITION	If applicable. Please note that these can be translations of the physical boundary conditions set in the case case or they can be pure computational (e.g. a unit cell with mirror S.C. to simulate an infinite domain).
3.1.6 ADDITIONAL SOURCE PARAMETERS	Please specify pure internal numerical other details (if applicable), like <ul style="list-style-type: none"> • Specific tolerances • Cell size, convergence criteria • Integration options

MODA Modelling Data providing a description for simulation data simulated in project «acronym» Data Owner [name, organisation, e-mail]	
DESCRIPTION OF THE SIMULATION	
1	<p>Case ID Please identify the first model name that is assumed to be physics based (model name) or a specific identifier.</p> <p>Please identify the first model name that is assumed to be physics based (model name) or a specific identifier.</p> <p>Please identify the first model name that is assumed to be physics based (model name) or a specific identifier.</p>
2	<p>Case ID (Model) Please identify the first model name that is assumed to be physics based (model name) or a specific identifier.</p> <p>Please identify the first model name that is assumed to be physics based (model name) or a specific identifier.</p> <p>Please identify the first model name that is assumed to be physics based (model name) or a specific identifier.</p>
3	<p>Reference for the model Please give the reference which documents the data of this case simulation.</p> <p>Please give the reference which documents the data of this case simulation.</p> <p>Please give the reference which documents the data of this case simulation.</p>
4	<p>Access location Please give a textual reference of only you as a modeler have chosen these models and this reference.</p> <p>Please give a textual reference of only you as a modeler have chosen these models and this reference.</p> <p>Please give a textual reference of only you as a modeler have chosen these models and this reference.</p>
5	<p>Workflow name Please give a textual reference of only you as a modeler have chosen these models and this reference.</p> <p>Please give a textual reference of only you as a modeler have chosen these models and this reference.</p> <p>Please give a textual reference of only you as a modeler have chosen these models and this reference.</p>
<p>Workflow picture Please insert your workflow picture in</p>	

- MODA was a closed semantic and epistemic space: Modelling methods had to be chosen from a small list.
- MODA imposed a **given level of detail** in workflow documentation.
- MODA documentations were **complicated**.²

CWA 17284:
2018 E
«MODA»

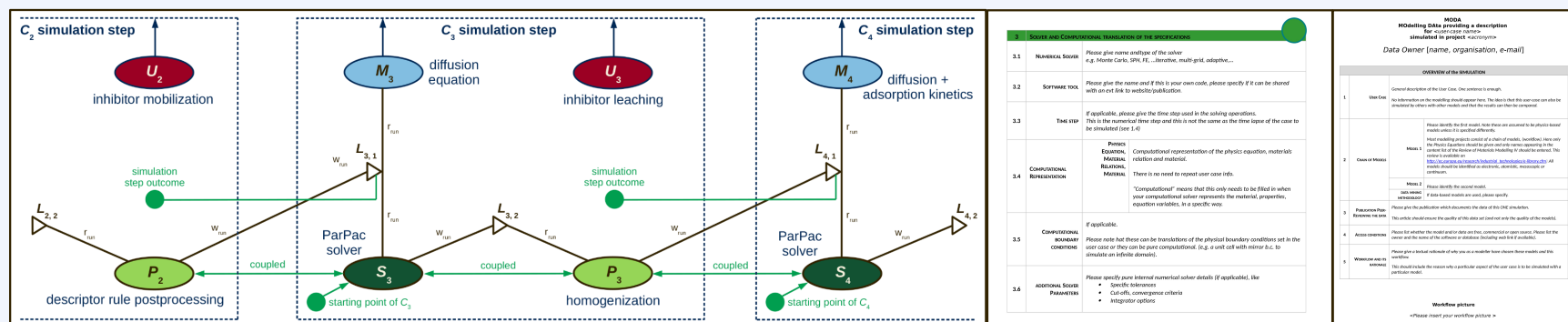


¹«Semantic interoperability and characterization of data provenance in computational molecular engineering», *J. Chem. Eng. Data* **65**(3): 1313–1329, doi:10.1021/acs.jced.9b00739, **2020**.

²ReaxPro project deliverable D2.1, «ReaxPro MODA diagrams», **2020**.

The first attempts (e.g., MODA): What went wrong?

Naive metadata standards often **focus(ed) on provenance documentation** only.¹



- MODA was a **closed semantic and epistemic space**: Modelling methods had to be chosen from a small list.
- MODA imposed a **given level of detail** in workflow documentation; namely, **unrealistically detailed**.
- MODA documentations were **complicated**,² and of **limited use** to all,³ including to humans.

CWA 17284:
2018 E
«MODA»



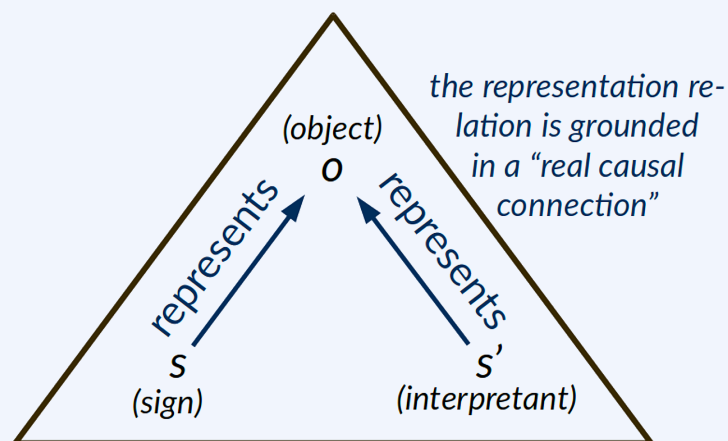
¹«Semantic interoperability and characterization of data provenance in computational molecular engineering», *J. Chem. Eng. Data* **65**(3): 1313–1329, doi:10.1021/acs.jced.9b00739, **2020**.

²ReaxPro project deliverable D2.1, «ReaxPro MODA diagrams», **2020**.

³«European standardization efforts from FAIR toward explainable-AI-ready data documentation in materials modelling», in *Proc. ICAPAI 2023*, doi:10.1109/icapai58366.2023.10193944, IEEE, **2023**.

Foundational ontology: EMMO

Peircean semiotics



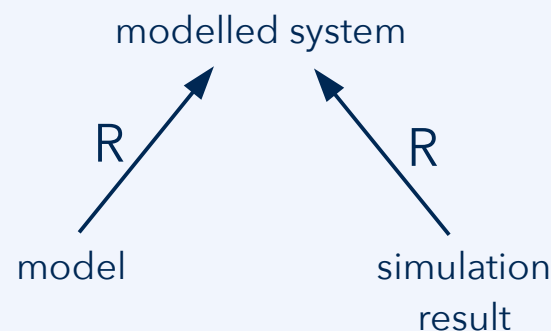
the semiosis, a process by which a new representamen, the interpretant, is created



C. S. Peirce

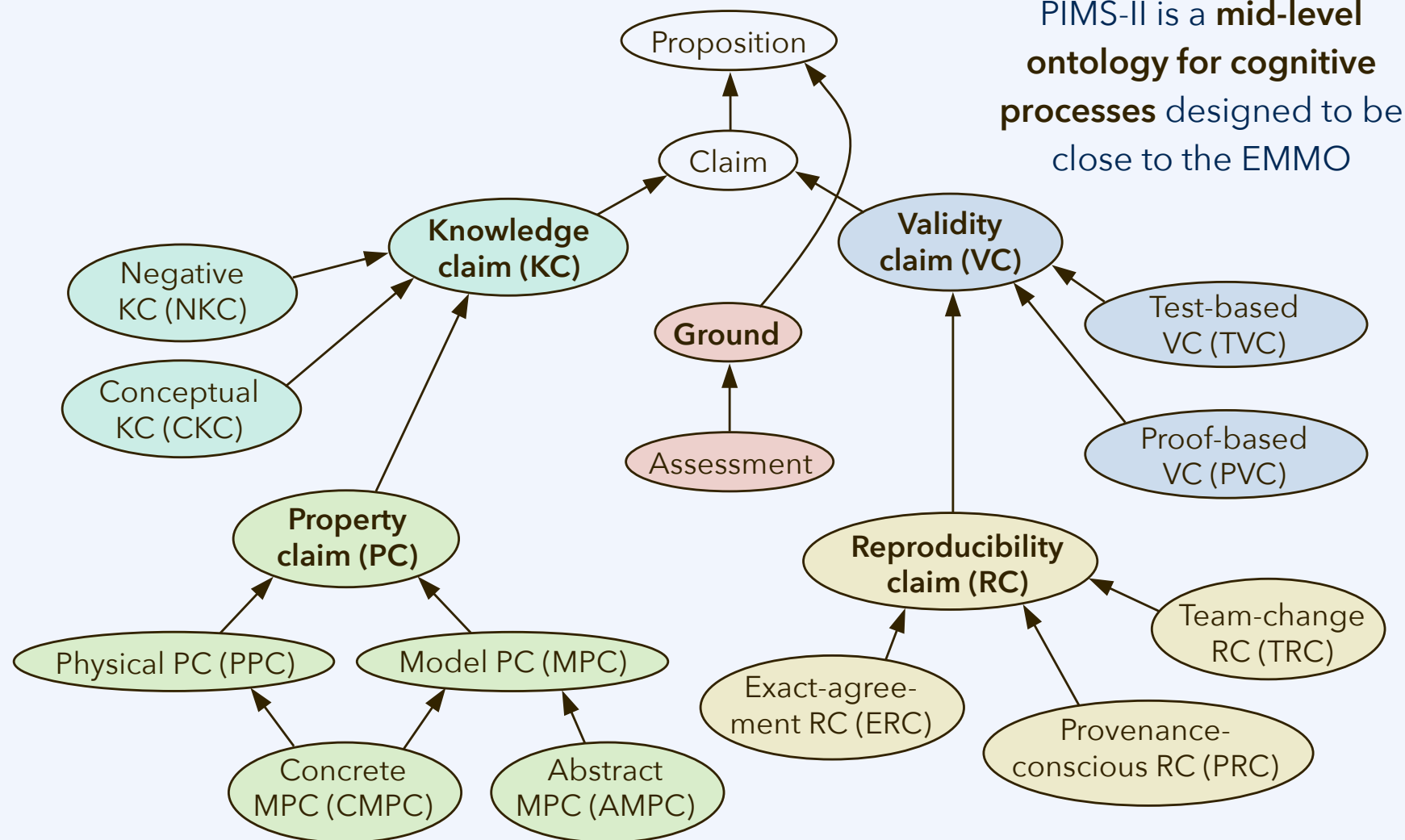
Elementary Multiperspective Material Ontology¹

- **Mereocausality:**
Physical parthood and causal connectedness
- **Peircean semiotics:**
Representation of physical entities by signs
(under development by E. Ghedini *et al.*)



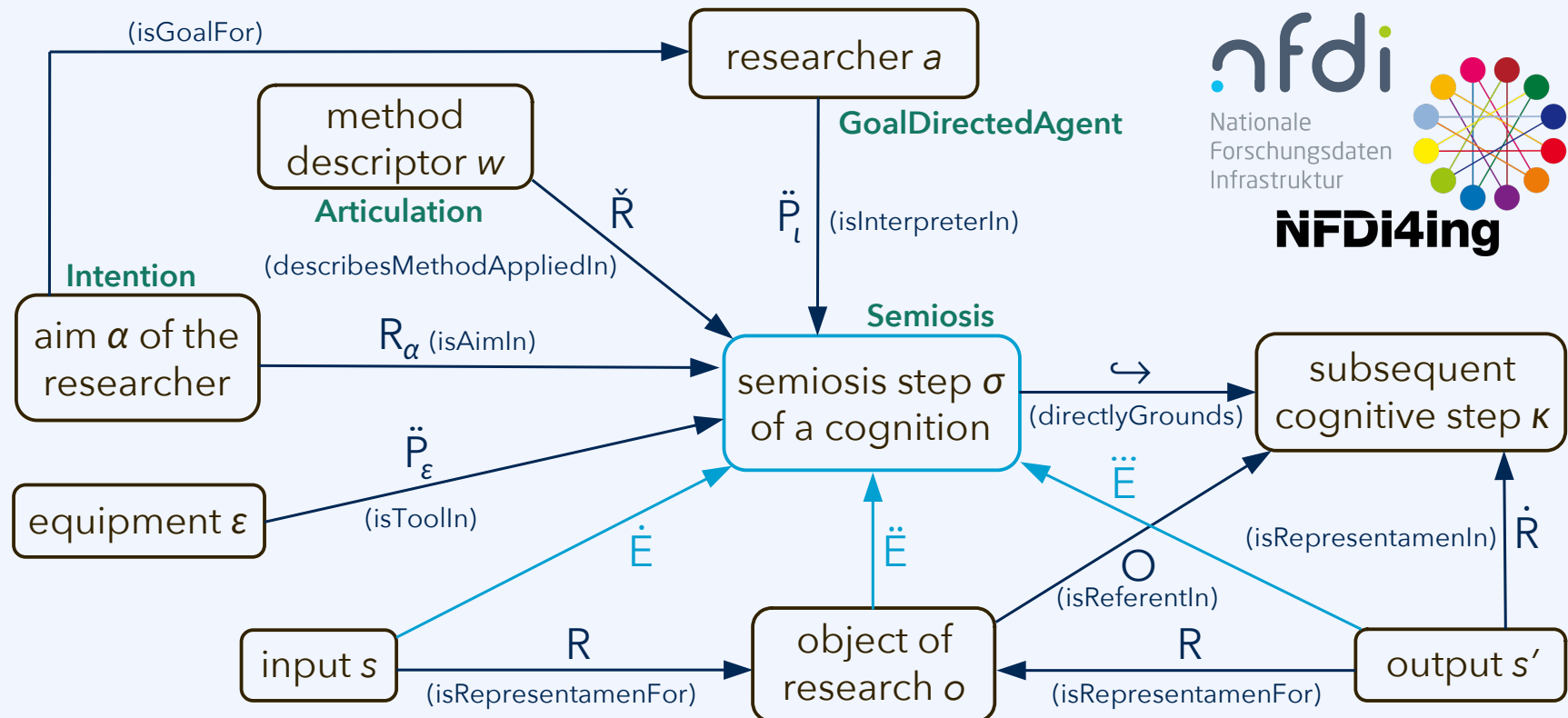
¹S. Clark *et al.*, *Adv. Energ. Mat.* **12**(17), 2102702, doi:10.1002/aenm.202102702, **2022**.

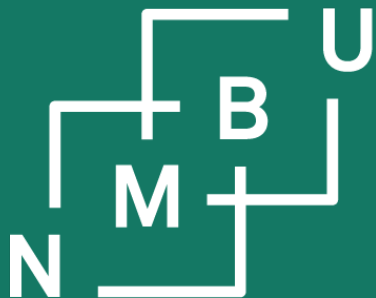
Ontology of epistemic metadata



Alignment with Metadata4Ing

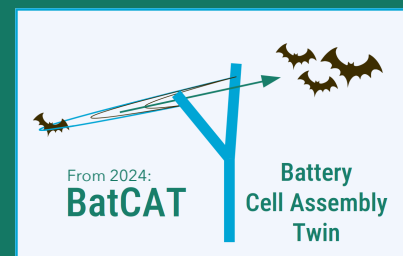
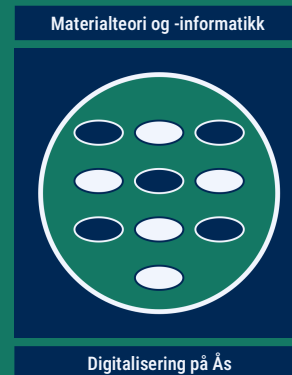
The core of Metadata4Ing development was the “processing step.”
This is 1:1 aligned between PIMS-II and Metadata4Ing.





Norges miljø- og
biovitenskapelige
universitet

Epistemic metadata



Horizon Europe ID 101137725



Horizon Europe ID 101138510

Annual Digital Catalysis & Catalysis-Related Sciences Conference

2nd November 2023

Frankfurt am Main