INPRODAT Innovationszentrum für Prozessdatentechnik e.V.

Technical report no. 2023–B
first revised version

Epistemic metadata in molecular modelling: Second-stage case-study report (12 claims)

Date: 6th February 2023 (original version: 5th February 2023)

Authors: Horsch, M.; Chiacchiera, S.; Guevara, G.; Kohns, M.; Müller, E.; Šarić, D.; Stephan, S.; Todorov, I.; Vrabec, J.; Schembera, B.

Funding information:

- BMBF, WindHPC, grant identifier 16ME0613
- DFG, MaRDI, project no. 460135501
- EC H2020, DOME 4.0, grant agreement no. 953163
- EC H2020, OntoCommons, grant agreement no. 958371

Accessibility:

- doi:10.5281/zenodo.7608074
- https://zenodo.org/communities/inprodat/

Epistemic metadata in molecular modelling: Second-stage case-study report (12 claims)

M. Horsch, S. Chiacchiera, G. Guevara, M. Kohns, E. Müller, D. Šarić, S. Stephan, I. Todorov, J. Vrabec, and B. Schembera

6th February 2023

About this document

This document reports on the outcomes from the second stage of our *case study on epistemic metadata in molecular modelling.* It builds on the outcomes from the first stage as summarized in the first-stage report [1].

Metadata are *data about data*, and epistemic metadata are *metadata that help establish the knowledge status of data.* There are various kinds of epistemic metadata; here, we are most concerned with *knowledge claims.* Specifically, we considered six journal articles from 2020 within the domain of molecular modelling, describing and discussing two knowledge claims from each of the papers. The aim was to approach the subject from an angle as indicated by the following guiding questions:

- What do the author(s) claim to know?
- Why should we accept the result as knowledge? *(epistemic grounding)*
- Is there any validation/verification being done in the paper itself?
- To what extent do the author(s) claim that the result can be reproduced?

The authors are preparing a manuscript, titled "Epistemic metadata for computational engineering information systems" [2], on the basis of the present report.

Acknowledgment. Silvia Chiacchiera, Martin Horsch, and Ilian Todorov acknowledge DOME 4.0 and OntoCommons, EC H2020 grant agreements no. 953163 and 958371. Simon Stephan and Jadran Vrabec acknowledge WindHPC, BMBF grant ID 16ME0613. Björn Schembera acknowledges MaRDI, DFG project no. 460135501.

Chapter 1

Technical remarks on epistemic metadata for computational engineering information systems

1.1 Documentation of claims using PIMS-II

The PIMS-II mid-level ontology[1] includes a hierarchy of claims, worked out during the first stage of the present case study, *cf.* Section 2 of the first-stage report [1]. Knowledge claims (KCs) are distinguished from validity claims (VCs),[2] with the reproducibility claim (RC) concept as a subclass of VC, *cf.* Fig. 1.1.

Advice: Do **confer the Appendix** when reading the statements below.[3] A knowledge claim φ from data δ can be documented in accordance with the following schema, *i.e.*, knowledge-graph pattern (*cf.* Fig. 1.2, top right):

$$\begin{aligned}
&\mathsf{KC}\varphi, && \varphi \text{ is a knowledge claim,}\\
&\mathsf{A}\varphi q, && \text{and its subject matter is } q,\\
&\mathsf{RQN}q, && \text{which is a research question.}\\
&\dot{\mathsf{B}}\varphi a, && \varphi \text{ is (or has been) asserted by } a,\\
&\mathsf{IO}a, && \text{an interlocutor (namely, the one who made the claim),}\\
&\ddot{\mathsf{P}}_{\iota} a\iota, && \text{who acted in the "interpreter" role in the cognition } \iota,\\
&3\iota\delta q\varphi, && \text{which has Peircean triadic form } \iota : \delta\text{—}q\text{—}\varphi, \qquad (1)
\end{aligned}$$

[1] See `http://www.molmod.info/semantics/pims-ii.ttl` for the OWL ontology TTL file.

[2] See the **Appendix** for **lists and explanations of the PIMS-II symbols** employed in this document. Moreoever, these symbols and abbreviations are also given in the PIMS-II OWL ontology TTL file (using skos:altLabel from SKOS [3]) and additionally in the LaTeX include file available under `http://www.molmod.info/semantics/pims-ii-latex-symbols.tex`.

[3] Notation: Predicates are followed by their arguments.

Example – instantiation of a concept: KC denotes the concept pims-ii:KnowledgeClaim, a unary predicate; in "KCφ," it is applied to φ in order to assert that "φ is a knowledge claim."

Example – instantiation of a relation: A denotes the relation pims-ii:hasSubjectMatter, a binary predicate; in "Aφq," the predicate A is applied to the two arguments φ, in subject position, and q, in object position, to assert that "φ has the subject matter q."

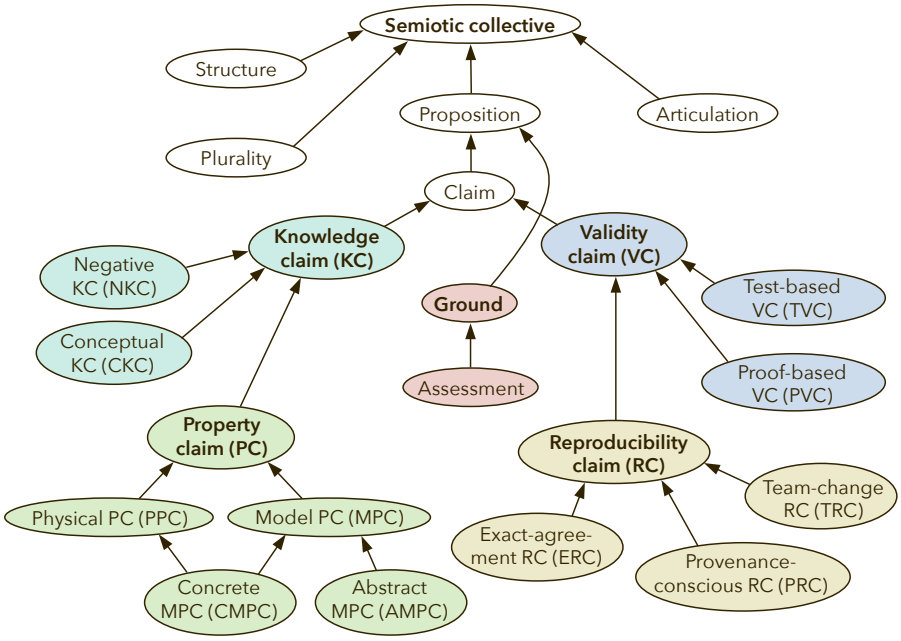

Figure 1.1: PIMS-II taxonomy of claims, from our CAOS 2022 paper [4].

and so forth as shown in the upper half of Fig. 1.2. Peircean triadic form for the semiosis above, δ—q—φ, is to be read such that the first element δ is the *sign*, the second element q is the *object*, and the third element φ is the *interpretant* [5].

For a validity claim, the schema in Fig. 1.2 (bottom right) is spelled out as:

$$
\begin{aligned}
\ddot{\mathsf{P}}_{\mathsf{e}} b\tau, &\quad \text{agent } b \text{ participates in the role of the “evaluator” in } \tau,\\
\mathsf{VAL}\tau, &\quad \text{a validation action,}\\
3\tau t\kappa\psi, &\quad \text{which has triadic form } \tau : t\text{—}\kappa\text{—}\psi, \text{ and therein,}\\
\mathsf{VC}\psi, &\quad \psi \text{ is a validity claim,}\\
\dot{\mathsf{B}}\psi b, &\quad \text{asserted by } b,\\
\hat{\mathsf{A}}\psi\varphi, &\quad \text{about the claim } \varphi,\\
\bar{\mathsf{R}}\varphi\kappa, &\quad \text{which was obtained as an outcome of cognition } \kappa,
\end{aligned}
\tag{2}
$$

etc., as visualized in the lower half of Fig. 1.2.

Naturally, users are not tied to using any such shape constraints. They can structure their knowledge graphs in any way that is compliant with applying the open world assumption to the rules from the PIMS-II ontology.

1.2 Use of m4i and D-SI for low-level KCs

The ontologies PIMS-II, developed mainly within Inprodat e.V., and Metadata4Ing (m4i), developed mainly within NFDI4Ing [6], were co-designed in alignment with each other and with the pre-existing metadata schema D-SI de-

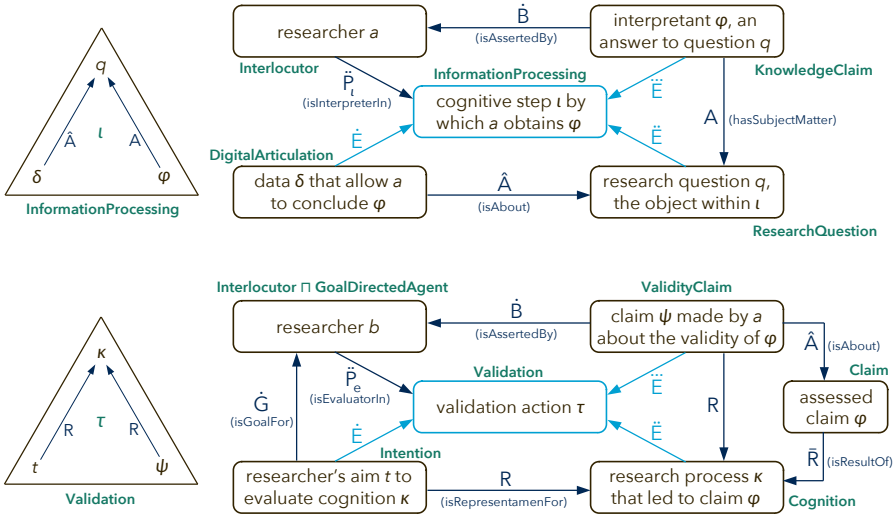

Figure 1.2: Knowledge and validity claim schemas, from our manuscript [2].

veloped by PTB.[4] This is particularly relevant to the documentation of low-level KCs, *e.g.*, where under specified boundary conditions, a property was found to have a certain value within some margin of error. Such property claims (PCs) are those for which we expect the benefit from machine-actionability to be greatest.

The main concept accessed from within D-SI is si:Real, a quantity with a numerical value, an uncertainty, and a unit. An si:Real value is associated with a pims-ii:Variable through an m4i:NumericalAssignment, subsumed taxonomically under pims-ii:Assignment. Through m4i, the "kind of quantity" is connected to the "quantity type" concept from QUDT [7]; materials can be identified via the EMMO [8, 9] or preferably by a canonical TUCAN [10] from NFDI4Chem [11].

Accordingly, to construct NumericalAssignment individuals expressing about water that "a vapour pressure is 88.9(4) kPa and a temperature is 369.5(1) K:"

```
@prefix m4i: <http://w3id.org/nfdi4ing/metadata4ing#>.
@prefix pims-ii: <http://www.molmod.info/semantics/pims-ii.ttl#>.
@prefix qudt-vocab-quantitykind: <http://qudt.org/vocab/quantitykind/>.
@prefix qudt-vocab-unit: <http://qudt.org/vocab/unit/>.

:water a pims-ii:CanonicalTUCAN;
   pims-ii:isLiterally "H2O/(1-3)(2-3)"^^xs:string.

:p_asg a m4i:NumericalAssignment;
   pims-ii:isAssignmentFor :water.
:p_var a m4i:Property;
   pims-ii:isVariableInAssignment :p_asg;
   m4i:hasKindOfQuantity qudt-vocab-quantitykind:VaporPressure.
:p_val a pims-ii:QuantityValue, si:Real;
   pims-ii:isValueInAssignment :p_asg;
   m4i:hasUnit qudt-vocab-unit:KiloPA;
   si:hasNumericalValue "88.9"^^xs:decimal;
```

[4] To be found on the `https://gitlab1.ptb.de/d-ptb/d-si/xsd-d-si` gitlab.

```
   m4i:hasExpandedUnc [
      a si:ExpandedUnc;
      si:hasUncertainty "0.1"^^xs:decimal;
      si:hasCoverageFactor "2"^^xs:decimal
   ].

:T_asg a m4i:NumericalAssignment;
   pims-ii:isAssignmentFor :water;
   pims-ii:isReferenceFrameFor :p_asg.
:T_var a m4i:Property;
   pims-ii:isVariableInAssignment :T_asg;
   m4i:hasKindOfQuantity qudt-vocab-quantitykind:ThermodynamicTemperature.
:T_val a pims-ii:QuantityValue, si:Real;
   pims-ii:isValueInAssignment :T_asg;
   m4i:hasUnit qudt-vocab-unit:K;
   si:hasNumericalValue "369.5"^^xs:decimal;
   m4i:hasExpandedUnc [
      a si:ExpandedUnc;
      si:hasUncertainty "0.1"^^xs:decimal;
      si:hasCoverageFactor "2"^^xs:decimal
   ].
```

This assumes a coverage factor $k = 2$, *i.e.*, that the error given is based on, or equivalent to, two times the standard deviation. In predicate notation,[5]

$$\begin{aligned}
\lambda\nu\ \text{“H2O/(1-3)(2-3),”} \quad & \nu \text{ is the TUCAN for water [10]},\\
\mathsf{ASG}\{d_p, d_T\}, \quad & d_p \text{ and } d_T \text{ are assignments},\\
\hat{\mathsf{A}}_{\mathsf{D}}\{d_p, d_T\}\nu, \quad & \text{both of which are assignments } \textit{for} \text{ water},\\
\ddot{\mathsf{R}} d_T d_p, \quad & \text{such that } d_T \text{ is a context or precondition for } d_p,\\
\ddot{\mathsf{D}} w_p d_p, \quad & \text{the assignment of the value } w_p,\\
\check{\mathsf{C}}_{\mathsf{u}} u_{\mathrm{kPa}} w_p, \quad & \text{which has the unit } u_{\mathrm{kPa}},\\
\dot{\mathsf{D}} v_p d_p, \quad & \text{to the variable } v_p. \qquad (3)
\end{aligned}$$

etc., restricting ourselves to concepts and relations taken from PIMS-II only. Continuing from there, multiple assignments can be encapsulated as follows:

```
:data_point a pims-ii:DataItem;
   pims-ii:isAbout :water.
:p_asg pims-ii:isSemioticallyConstitutiveOf :data_point.
:T_asg pims-ii:isSemioticallyConstitutiveOf :data_point.
```

Or denoting the predicates by their skos:altLabel symbol representations,

$$\begin{aligned}
\mathsf{DI}\delta, \quad & \delta \text{ is a data item},\\
\check{\mathsf{C}}\{d_p, d_T\}\delta, \quad & \text{it is constituted by } d_p \text{ and } d_T,\\
\hat{\mathsf{A}}\delta\nu, \quad & \text{and it is about water}. \qquad (4)
\end{aligned}$$

This now asserts about water that, for a single data point, “*the* vapour pressure is 88.9(4) kPa and *the* temperature is 369.5(1) K.” If we had only wanted to

[5] Here and below, predicate notation is combined with set notation as an abbreviation. In this way, for example, $\hat{\mathsf{A}}_{\mathsf{D}}\{d_p, d_T\}\nu$ means that we assert both $\hat{\mathsf{A}}_{\mathsf{D}} d_p \nu$ and $\hat{\mathsf{A}}_{\mathsf{D}} d_T \nu$.

say that the temperature assignment is a reference frame (*e.g.*, a precondition[6]) for the vapour pressure, without combining the two into a single data item, the triple from above stating $\dddot{\mathsf{R}} d_T d_p$ would already have been enough:

```
:T_asg pims-ii:isReferenceFrameFor :p_asg.        (5)
```

Multiple similarly structured data items can in turn be packaged into a dataset, using PIMS-II or following the connection of m4i to DCAT version 3.[7] The above summarizes how a typical case of a low-level claim from the PIMS-II taxonomy (Fig. 1.1) such as a PC can be documented in terms of triples.

In addition to the combination of ontologies for KCs realized here, it could also be of interest to further look into connecting the present VC documentation to pre-existing semantic artefacts; in particular, to the VIMMP Validation Ontology (VIVO) [12], *cf.* Fig. 1.3, and the Citation Typing Ontology (CITO) [13]. "Enriched cited references" by Clarivate[8] might also develop in this direction.

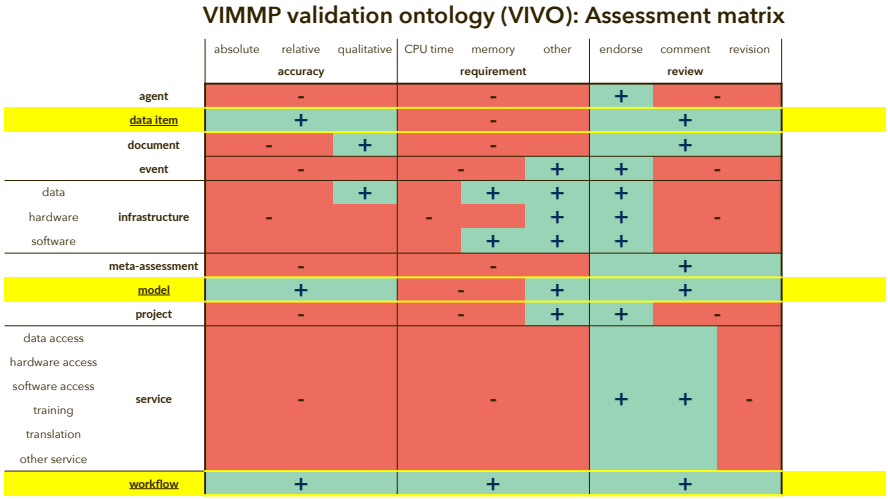

Figure 1.3: VIVO assessment matrix [14]; highlighted: Categories of assessments that are directly applicable to knowledge, validity, and reproducibility claims.

1.3 Reproducibility claims and ortho-/paradata

Say that researcher a carried out research process κ, which is a Cognition, and obtained[9] the research outcome φ, for example as in Fig. 1.2. Presumably even

[6] This is all assuming that T was given or pre-specified in some way, and $p_{\mathrm{sat}}(T)$ was measured, calculated, or looked up using the given temperature as a boundary condition or "reference frame." It could naturally also be the other way around, in which case the obverse can be stated, $\dddot{\mathsf{R}} d_p d_T$, or neither of the assignments could be a reference frame for the other.

[7] See in particular `https://www.w3.org/TR/vocab-dcat-3/#Class:Dataset`.

[8] `https://clarivate.com/webofsciencegroup/release-notes/wos/new-wos-april-29-release-notes/`

[9] The outcomes of a research process, understood as a social cognitive process, might include KCs, datasets, data items, or other elements that occur in the role of the interpretant [4, 5] within the cognition. Without loss of generality let us here assume the outcome to be a KC.

if κ is a fairly detailed documentation of the provenance, there will be gaps in it, or there may have been an element of chance involved in finding exactly φ as the outcome. But as a minimum, if we accept the above, a has at least succeeded at showing that φ can be the outcome of a process compliant with κ. We can then say, "*given* the cognitive process κ, it is *possible* to obtain φ." A compact notation for this can build on the established way of writing $p(A \mid B)$ for the conditional probability of "A given B," where the vertical bar means "given." Sure enough, also $\Diamond$ means "possibly," and $\Box$ "necessarily." We thus denote,

$$\Diamond(\varphi \mid \kappa), \quad \text{doing } \kappa \text{ can possibly yield an outcome consistent with } \varphi. \tag{6}$$

This schema may be understood as specifying the minimum semantics of a knowledge claim in combination with its provenance documentation.

Similarly, a reproducibility claim (RC) states that doing something specific will *necessarily* result in something specific. First, however, these claims do not require the reproducing researcher b to comply with the entire provenance documentation completely and exactly; *e.g.*, if $a \neq b$, changing the researcher, this alone is already a part of the provenance that has changed. Second, it is also not usually a complete and exact agreement with the original outcome that is demanded; *e.g.*, if a found $p_{\text{sat}}(T) = 88.9(4)$ kPa, b finding $p_{\text{sat}}(T) = 88.9(2)$ kPa will certainly be in order even though that is not the same si:Real value, and normally so will $p_{\text{sat}}(T) = 89.4(7)$ kPa. This means that for an RC we need to be ready to deal with *weaker* versions κ'' and φ'' of both cause and effect,

$$\psi : \Box(\varphi'' \mid \kappa''), \quad \text{doing } \kappa'' \text{ always yields an outcome consistent with } \varphi''. \tag{7}$$

We call κ'' and φ'' the *orthodata* [15] associated with the RC ψ, *cf.* Fig. 1.4,

$$\begin{aligned} &\mathsf{RC}\psi, && \psi \text{ is a reproducibility claim,} \\ &\check{\mathsf{C}}_{\perp}\{\kappa'', \varphi''\}\psi, && \text{and } \kappa'' \text{ and } \varphi'' \text{ are its orthodata.} \end{aligned} \tag{8}$$

The rationale underlying the term "orthodata," from Gr. *ορθός*, "right," is that these are the elements that the reproducing researcher attempts to "get right."

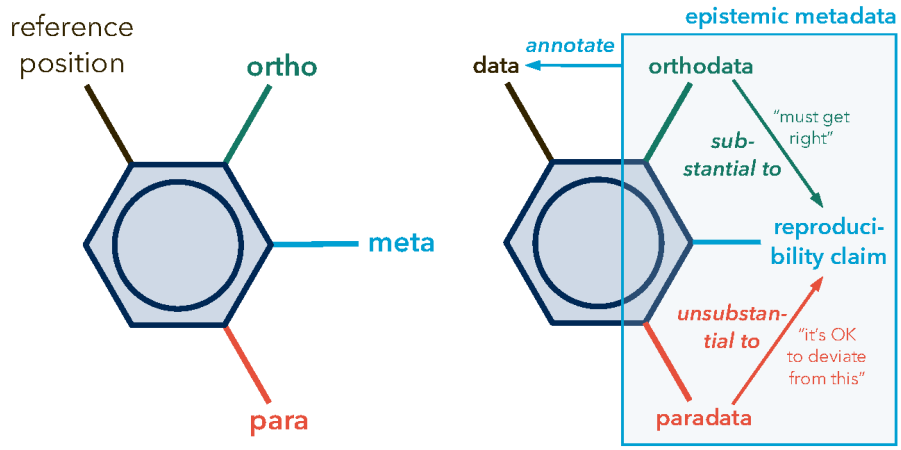

Figure 1.4: Idea behind the ortho-/paradata split, from our book chapter [15].

We have thus introduced a notation for conditional necessity or possibility, inspired by the notation for conditional probabilities, such that the right-hand side or antecedent κ'' is a partial documentation of the research process, while the left-hand side or consequence φ'' is a partial documentation of the research outcome. This is visualized in Fig. 1.5 using a modal square of opposition; technically, this is a "modern" or "non-classical" square of opposition: $\Box(\varphi'' \mid \kappa'')$ does not entail $\Diamond(\varphi'' \mid \kappa'')$, as opposed to the classical Aristotelian square where there would be such an entailment by subalternation [16]. Necessity does not entail possibility because it could be the case that it is impossible to satisfy the premise. If compliance with κ'' is an impossibility, φ'' given κ'' is "necessary," but not "possible." This technicality simply allows the operators to work just as one would intuitively expect it. In particular, they satisfy De Morgan's law,

$$\Diamond(\neg\varphi'' \mid \kappa'') \;\Leftrightarrow\; \neg\Box(\varphi'' \mid \kappa''). \tag{9}$$

The above is also the schema for contradicting or falsifying a reproducibility claim. If it is found through a reproduction attempt that doing κ'' can possibly yield $\neg\varphi''$, then this opposes the RC that doing κ'' must necessarily yield φ''.

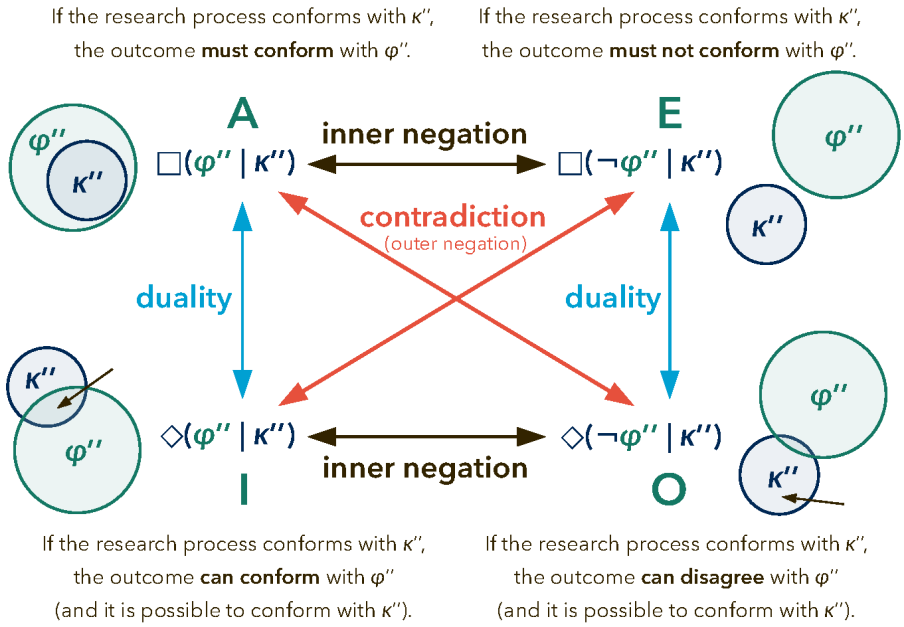

Figure 1.5: Square of opposition for conditional claims, from our manuscript [2].

1.4 Research questions and composite subjects

Yablo [17] proposes to conceive of a sentence's subject matter m in terms of "an equivalence relation on logical space: Worlds are equivalent, or cell-mates, just in case they are indiscernible where that subject matter is concerned. If m is the number of stars, $\equiv_m$ is the relation one world bears just if they have equally many stars" [17, p. 24]. Instead of saying that m is the number of stars,

the topic can be equivalently expressed in question form, "how many stars are there?" or "what is the number of stars?" The viability of an approach to identifying aboutness and subject matter with a partitioning of epistemic space in alignment with one or multiple *information slots* such as the above becomes clear in the treatment by Barton *et al.* [18]. This naturally looks practical in the context of knowledge-graph based technology: If the subject matter of a KC is simply the question it answers, it could potentially be expressed in SPARQL, as a knowledge graph pattern using wildcards, in a technical implementation.

In other words, the topic or subject matter of a KC, the associated data, or the paper section containing that KC, could be research questions such as [19]

$$\begin{aligned} q_1 &= \text{"What is the } \mathbf{D} \text{ matrix of liquid } M \text{ as a function of } \mathbf{x}, p, T?\text{,"} \\ q_2 &= \text{"What is the } \mathbf{\Gamma} \text{ matrix of liquid } M \text{ as a function of } \mathbf{x}, p, T?\text{,"} \end{aligned} \tag{10}$$

where $\mathbf{D}$ are the Fick diffusion coefficients and $\mathbf{\Gamma}$ are the thermodynamic correction factors relating them to Maxwell-Stefan coefficients. M is the quaternary mixture considered by Guevara *et al.* [19]. Naturally these can be combined as

$$q_1 q_2 = \text{"What are the } \mathbf{D} \text{ and } \mathbf{\Gamma} \text{ matrices of liquid } M \text{ as a function of } \mathbf{x}, p, T?\text{,"}$$

such that, as the notation suggests, the set of equivalence classes "on logical space" with respect to $\equiv_{q_1 q_2}$ is the product set out of those for $\equiv_{q_1}$ and $\equiv_{q_2}$.

Yablo states about his formalism that on the question "what subject matters are 'of,' on my account[,] I *think* the answer is sentences in context" [17, p. ix]. In most cases this should carry over to a single KC, since typical KCs can be expressed as a sentence. But what is the subject matter of a heterogeneous collection of datasets, or of a FAIR digital object containing multiple claims? It would be helpful if publications, and maybe even collections of publications, could be annotated with a subject matter as well; the topic or subject matter would usually be looked for during data retrieval on an information system.

It is for these more complex or heterogeneous items, going far beyond single sentences, that the above construction that we call the *topical product* becomes very inelegant; similar to the notion of "possible worlds," it suffers from state explosion as more and more features are brought to the table, partitioning epistemic space further and further, even where the questions that are being considered have little to nothing to do with each other. We therefore suggest a *topical sum* (TLS) construct by which independent topical elements, such as [20]

$$\begin{aligned} q_3 &= \text{"What qualitative relationship is there between interfacial properties} \\ &\quad \text{and dispersive long-range interactions in a molecular model?,"} \\ q_4 &= \text{"What is } \gamma^\star \text{ as a function of } x_2^\star \text{ and } T^\star \text{ for mixture } A?\text{,"} \end{aligned} \tag{11}$$

are combined into a plurality of topics, $q_3 + q_4$, instead of the product $q_3 q_4$.

To a technical implementation this might suggest not to package the two Topical (TL) individuals into a joint query against a knowledge base, but to consider them separately. With regard to the theoretical formalism, in this way, we refrain from considering the product space of equivalence classes with respect to $\equiv_{q_3}$ and $\equiv_{q_4}$ over epistemic space, avoiding state explosion. In PIMS-II, the relation isTopicalFactorIn (symbol $\check{\mathsf{C}}_\mathsf{T}$) is used to create a topical product, while the relation isTopicalSummandIn (symbol $\lhd_\mathsf{T}$) is used to create a topical sum.

Chapter 2

Selected knowledge claims

2.1 Guevara *et al.*: Fick diffusion coefficient matrix of a quaternary liquid mixture by molecular dynamics [19]

About the interview. This paper was discussed on Tuesday, 24th January 2023; its authors were represented by Gabriela Guevara and Jadran Vrabec.

Selected knowledge claims from the paper:

1. A novel finite-size correction methodology for the phenomenological diffusion coefficient matrix $\mathbf{L}$ based on linear extrapolation over $1/N^3$ to the limit $1/N^3 \to 0$ is proposed and successfully used to calculate $\mathbf{D}$.

2. The Fick diffusion coefficient matrix $\mathbf{D}$ of the considered mixture has the values given in Table 1 of the paper under the conditions specified there.

From the discussion of Claim 1:

a The main other finite-size correction method is the one by Yeh and Hummer [21]; this is the method that most researchers resort to by default.

b Whereas the Yeh-Hummer method (in a case considered here) yields a $+14\%$ correction, the novel method only yields a $+6\%$ correction. If the novel method is correct, that would mean that using the Yeh-Hummer correction is worse (more inaccurate) than no finite-size correction at all.

c Linear extrapolation over $1/N^3 \to 0$ is a common approach to finite-size corrections; it is underlying to both Yeh-Hummer and to the novel method.

d The novel method looks preferable or more plausible as it exhibits what is typically seen in the community as theoretical virtues: *First*, Yeh and Hummer [21] use a semiempirical correlation relying on multiple properties, while the novel method is formally much simpler, relying only on N. *Second*, the Yeh-Hummer method operates on the end result $\mathbf{D}$, whereas the novel method operates on the intermediate result $\mathbf{L}$ that directly experiences the finite-size limitation in the molecular simulation.

A validation of the novel method has not been done; its epistemic grounding here rests purely on its theoretical virtues.[1]

From the discussion of Claim 2:

e The numerical uncertainties are such as given in Table 1 of the paper, obtained as usual through Flyvbjerg-Petersen type [24] block averaging.

f Results for the quaternary system are validated against previous results for ternary subsystems.

g It would be very hard to measure diffusion coefficients in the considered quaternary mixture because the compounds are so similar to each other. More generally, such experiments become complicated for mixtures with a greater number of components, and they are rarely done.

h However, some of Pařez *et al.*'s simulation results on one of the ternary subsystems [25] have in the meantime been confirmed by experiment.

i Moreover, in the meantime, Peters *et al.* [26] succeeded at experimentally measuring quaternary diffusion coefficients, but not for the same mixture as the one considered by Guevara *et al.* [19]. Subsequently, one of their data points was confirmed by molecular simulation.

2.2 Haslam *et al.*: Expanding the applications of the SAFT-γ Mie group-contribution equation of state – prediction of thermodynamic properties and phase behavior of mixtures [27]

About the interview. This paper was discussed on Tuesday, 24th January 2023; its authors were represented by Max Kohns.

Selected knowledge claims from the paper:

3. The model accuracy for osmotic coefficients and mean ion activity coefficients has the value(s) as given in Table 8 of the paper.

4. LLE and overall phase behaviour predicted qualitatively correctly for binary mixtures of water with 1-butanol or 2-butanol.

From the discussion:

a The parameters were adjusted to LLE of water + 2-hexanol and water + 2-octanol, so the qualitative agreement (Claim 4) is unsurprising.

[1] This would be: *First*, the virtue of *simplicity*. Interestingly, communities differ in their views on simplicity as a virtue; from an empirical study, Schindler [22] finds that "simplicity is viewed as an epistemic virtue particularly by social scientists (but not by philosophers)." *Second*, by operating on $\mathbf{L}$ which is the direct simulation outcome, and hence most immediately influenced by the finite size of the simulated system, the Guevara *et al.* correction serves as a *mathematical explanation of physical phenomena* following Bangu [23], *i.e.*, it "reveals, or *explicates*, the relevant dependence relations" more clearly than the Yeh-Hummer method.

b What would we need to do in order to falsify or unsuccessfully attempt to reproduce the claims?

- Reimplement the EOS solver. (The present one is not published.)
- The EOS parameters all need to be taken over correctly.
- For Claim 3, look up all the experimental data from the cited references; calculate the EOS values and the deviations. They will probably not be exactly the same even in the successful case, due to numerical noise; expertise and sound judgment will be needed to distinguish numerical noise from an actual falsification of the results.
- For Claim 4, it is not necessary to look up experimental data (assuming the authors can be trusted on this); just reevaluate the EOS for the same system and check if the phase-diagram topology is retained.

2.3 Šarić *et al.*: Dielectric constant and density of aqueous alkali halide solutions by molecular dynamics – a force field assessment [28]

About the interviews. This paper was discussed on Tuesday, 24th January 2023; its authors were represented by Denis Šarić and Jadran Vrabec when discussing Claim 5 and by Max Kohns when discussing Claim 6.

Selected knowledge claims from the paper:

5. For eleven types of alkali halide salt solutions, a concrete recommendation for a force field is given; the expected deviation for the dielectric constant and the density is given (for numerical values *cf.* Table II of the paper).
6. Claim 5 must be checked against phase behaviour. For NaCl, *e.g.*, the HMN-S force field would be recommended as performing best for the dielectric constant, but in the simulation with HMN-S, there is a big salt crystal. In such cases, the recommendation from Claim 5 is not upheld.

About the relationship between the two claims:

a The recommended model is the one that performed best out of eight investigated force-field families.

b Claim 6 expresses/relates to a partly unsuccessful validation of Claim 5.

Why is it new and relevant knowledge?

c It is knowledge because established methods are employed, using established tools.

d It is important to have a recommendation; many people just start simulating with any model that they find.

e The dielectric constant is technically important, but experiments are hard to conduct; available data are surprisingly scarce.

What should be taken into consideration for a reproduction attempt?

f The same force fields need to be used, since this is about the force field.

g Everything else can be varied, including the solver, detailed steering parameters of the method, or the method as a whole.

2.4 Stephan and Hasse: Enrichment at vapour-liquid interfaces of mixtures – establishing a link between nanoscopic and macroscopic properties [29]

About the interview. This paper was discussed on Tuesday, 24th January 2023; its authors were represented by Simon Stephan.

Selected knowledge claims from the paper:

7. There is a relation between vapour-liquid equilibrium bulk data and nanoscopic interfacial enrichment in binary mixtures of chemically inert molecular fluids. (Based on literature data, packaged here as a database.)

8. A model with twelve empirical parameters is proposed for the low-boiling component enrichment E_2 of a mixture as a function of the temperature, the liquid-phase mole fraction, and the partition coefficient.

From the discussion:

a Is it important? The nanoscopic interfacial enrichment cannot be measured experimentally so far – but by simulation we can access it.

b Is it new? Yes it is.

c Why should we accept it as knowledge? It was tested *i)* on data for simple fluids used for the fit and *ii)* on all available real-substance model data.

d How about reproducibility? While the correlation applies to many data points well, a substantial deviation under some conditions does occur. Due to the presence of such deviating data points, ± 1 is expected as a characteristic deviation for E_2; this includes future reproduction attempts as well as the model accuracy for new applications.

2.5 Stephan and Hasse: Influence of dispersive long-range interactions on properties of vapour-liquid equilibria and interfaces of binary Lennard-Jones mixtures [20]

About the interview. This paper was discussed on Tuesday, 24th January 2023; its authors were represented by Simon Stephan.

Selected knowledge claims from the paper:

9. There is a qualitative influence of dispersive long-range interactions on vapour-liquid interfacial properties of mixtures; namely, the mixture surface tension is higher if dispersive long-range interactions are present.

10. For the binary mixture labelled "A," the reduced surface tension is $\gamma^\star = 0.672 \pm 0.003$ at $T^\star = 0.92$ and $x_2^\star = 0.01$ mol mol^{-1}.

From the discussion of Claim 9:

a Is it relevant? Yes: Understanding the relation of molecular interactions and macroscopic properties is important for the development of a model class that describes macroscopic properties based on molecular properties.

b Is it new? Yes, for mixtures. For pure components this was already known.

c Why should we accept it as knowledge? It is a finding from new primary data, based on well-established methods and tools (and models). The selected mixtures/thermodynamic conditions are representative.

d The difference between the LJ and LJTS results are significant, *i.e.*, they significantly exceed the statistical uncertainties; there is little doubt that this should be a reproducible observation.

Remark on validation:

e Data for the full LJ potential are compared to data for the LJTS potential; the latter are taken from the literature.

From the discussion of Claim 10:

f The numerical value of the surface tension is believed to be reproducible since well-established methods and tools were used that were also validated for pure component surface tension data by comparison to literature data.

2.6 Zhu and Müller: Generating a machine-learned equation of state for fluid properties [30]

About the interview. This paper was discussed on Monday, 16th January 2023; its authors were represented by Erich Müller.

Selected knowledge claims from the paper:

11. Artificial neural networks (ANNs) beat Gaussian process regression (GPR) at quantitative agreement with data from the SAFT-VR-Mie EOS.

12. ANNs have the capability of correlating thermophysical property data and can therefore be used as a surrogate model for a molecular EOS.

From the discussion:

a Upon superficial consideration, we could think that the GPR performs better, but this is due to overfitting. "With a combination of RBF and linear kernels, GPR can predict critical points with similar performance ($R^2 = 0.9999$) with only 300 data points."

b Once we look at the GPR beyond just numerical agreement with data, but also at the overall shape, the overfitting becomes clear: "A similar GPR model [...] albeit converging and providing acceptable statistical indicators fails to capture the VLE envelope shape even with the inclusion of critical points and employing over 2000 data points in the fitting process."

c Same training/validation set split (80% : 20%) applied to ANN and GPR. The observation on the shape of the binodal is also a validation statement.

d Related to the question of epistemic grounding, Erich asks further: What do we mean when we ask, why/how is something relevant knowledge? First, it must be important or valuable, second it must be right.

e Why is it important/valuable? It is testing out a novel methodology, establishing its potential as a useful method through a proof of concept.

f Why is it true? Let us ask: Could it be false?

As regards Claim 12, there could also be something simple or special about SAFT-like equations of state whereas for other, more complicated EOS the ANN would work less nicely, or the GPR could outperform the ANN.

Bibliography

[1] M. T. Horsch and B. Schembera: 2023, 'Epistemic metadata in molecular modelling: First-stage case-study report (10 cases)'. Technical Report 2023–A, Inprodat, Kaiserslautern. doi:10.5281/zenodo.7516532.

[2] M. Horsch, S. Chiacchiera, M. Kohns, E. Müller, S. Stephan, I. Todorov, J. Vrabec, and B. Schembera: 2023, 'Epistemic metadata for computational engineering information systems'. Manuscript.

[3] A. Isaac and E. Summers: 2009, 'SKOS Simple Knowledge Organization System primer'. Working group note, W3C. https://www.w3.org/TR/2009/NOTE-skos-primer-20090818/.

[4] M. T. Horsch and B. Schembera: 2022, 'Documentation of epistemic metadata by a mid-level ontology of cognitive processes'. In: T. P. Sales, M. Hedblom, and H. Tan (eds.): *Proceedings of JOWO 2022*. CEUR-WS, Aachen, pp. 2–CAOS. http://ceur-ws.org/Vol-3249/paper2-CAOS.pdf.

[5] C. S. Peirce: 1955, 'Logic as semiotic: The theory of signs'. In: J. Buchler (ed.): *Philosophical Writings of Peirce*. New York: Dover (ISBN 978-0-48620217-4), pp. 98–119.

[6] A. Karmacharya, B. Farnbacher, C. Wiljes, D. Iglezakis, D. Terzijska, G. Lanza, J. Hickmann, J. Theissen-Lipp, J. Munke, J. Windeck, M. Fuhrmans, M. T. Horsch, N. Hoppe, S. Leimer, S. Hachinger, and S. Arndt: 2022, 'Metadata4Ing: An ontology for describing the generation of research data within a scientific activity'. Documentation, NFDI4Ing. https://w3id.org/nfdi4ing/metadata4ing/1.0.0/.

[7] X. Zhang, K. Li, C. Zhao, and D. Pan: 2017, 'A survey on units ontologies: Architecture, comparison and reuse'. *Program: Electronic Library and Information Systems* **51**(2), 193–213. doi:10.1108/prog-08-2015-0056.

[8] J. Francisco Morgado, E. Ghedini, G. Goldbeck, A. Hashibon, G. J. Schmitz, J. Friis, and A. de Baas: 2020, 'Mechanical testing ontology for digital-twins: A roadmap based on EMMO'. In: R. García Castro, J. Davies, G. Antoniou, and C. Fortuna (eds.): *Proceedings of SeDiT 2020*. CEUR-WS, Aachen, p. 3. http://ceur-ws.org/Vol-2615/paper3.pdf.

[9] H. A. Preisig, T. F. Hagelien, J. Friis, P. Klein, and N. Konchakova: 2021, 'Ontologies in computational engineering'. In: F. Chinesta, R. Abgrall, O. Allix, and M. Kaliske (eds.): *Proceedings of WCCM-ECCOMAS 2020*. Scipedia, Barcelona. doi:10.23967/wccm-eccomas.2020.262.

[10] J. C. Brammer, G. Blanke, C. Kellner, A. Hoffmann, S. Herres-Pawlis, and U. Schatzschneider: 2022, 'TUCAN: A molecular identifier and descriptor applicable to the whole periodic table from hydrogen to oganesson'. *Journal of Cheminformatics* **14**, 66. doi:10.1186/s13321-022-00640-5.

[11] S. Herres-Pawlis, O. Koepler, and C. Steinbeck: 2019, 'NFDI4Chem: Shaping a digital and cultural change in chemistry'. *Angewandte Chemie International Edition* **58**(32), 10766–10768. doi:10.1002/anie.201907260.

[12] M. T. Horsch, S. Chiacchiera, M. A. Seaton, I. T. Todorov, K. Šindelka, M. Lísal, B. Andreon, E. Bayro Kaiser, G. Mogni, G. Goldbeck, R. Kunze, G. Summer, A. Fiseni, H. Brüning, P. Schiffels, and W. L. Cavalcanti: 2020, 'Ontologies for the Virtual Materials Marketplace'. *Künstliche Intelligenz* **34**(3), 423–428. doi:10.1007/s13218-020-00648-9.

[13] E. Willighagen: 2020, 'Adoption of the Citation Typing Ontology by the Journal of Cheminformatics'. *Journal of Cheminformatics* **12**, 47. doi:10.1186/s13321-020-00448-1.

[14] M. T. Horsch, S. Chiacchiera, M. A. Seaton, I. T. Todorov, D. Toti, and G. Goldbeck: 2021, 'Introduction to the VIMMP ontologies'. Release, Virtual Materials Marketplace. doi:10.5281/zenodo.4411422.

[15] M. T. Horsch and B. Schembera: 2023, 'Molecular modelling and simulation reproducibility documentation by orthodata and paradata'. In: O. Sköld, L. Andersson, and I. Huvila (eds.): *Perspectives on Paradata*.

[16] D. Westerståhl: 2012, 'Classical vs. modern squares of opposition, and beyond'. In: J.-Y. Béziau and G. Payette (eds.): *The Square of Opposition*. Bern: Peter Lang (ISBN 978-3-03430537-2), pp. 195–229.

[17] S. Yablo: 2014, *Aboutness*. Princeton U. Press (ISBN 978-0-6911-4495-5).

[18] A. Barton, F. Toyoshima, L. Vieu, P. Fabry, and J.-F. Ethier: 2020, 'The mereological structure of informational entities'. In: B. Brodaric and F. Neuhaus (eds.): *Proceedings of FOIS 2020*. IOS (ISBN 978-1-64368-128-3), Amsterdam, pp. 201–215. doi:10.3233/faia200672.

[19] G. Guevara Carrión, R. Fingerhut, and J. Vrabec: 2020, 'Fick diffusion coefficient matrix of a quaternary liquid mixture by molecular dynamics'. *Journal of Physical Chemistry B* **124**(22), 4527–4535. doi:10.1021/acs.jpcb.0c01625.

[20] S. Stephan and H. Hasse: 2020, 'Influence of dispersive long-range interactions on properties of vapour-liquid equilibria and interfaces of binary Lennard-Jones mixtures'. *Molecular Physics* **118**(9–10), e1699185. doi:10.1080/00268976.2019.1699185.

[21] I.-C. Yeh and G. Hummer: 2004, 'System-size dependence of diffusion coefficients and viscosities from molecular dynamics simulations with periodic boundary conditions'. *Journal of Physical Chemistry B* **108**(40), 15873–15879. doi:10.1021/jp0477147.

[22] S. Schindler: 2022, 'Theoretical virtues: Do scientists think what philosophers think they ought to think?'. *Philosophy of Science* **89**(3), 542–564. doi:10.1017/psa.2021.40.

[23] S. Bangu: 2021, 'Mathematical explanations of physical phenomena'. *Australasian Journal of Philosophy* **99**(4), 669–682. doi:10.1080/00048402.2020.1822895.

[24] H. Flyvbjerg and H. G. Petersen: 1989, 'Error estimates on averages of correlated data'. *Journal of Chemical Physics* **91**(1), 461–466. doi:10.1063/1.457480.

[25] S. Pařez, G. Guevara Carrión, H. Hasse, and J. Vrabec: 2013, 'Mutual diffusion in the ternary mixture of water + methanol + ethanol and its binary subsystems'. *Physical Chemistry Chemical Physics* **15**(11), 3985–4001. doi:10.1039/C3CP43785J.

[26] C. Peters, J. Thien, L. Wolff, H.-J. Koss, and A. Bardow: 2020, 'Quaternary diffusion coefficients in liquids from microfluidics and Raman microspectroscopy: Cyclohexane + toluene + acetone + methanol'. *Journal of Chemical & Engineering Data* **65**(3), 1273–1288. doi:10.1021/acs.jced.9b00632.

[27] A. J. Haslam, A. González Pérez, S. Di Lecce, S. H. Khalit, F. A. Perdomo, S. Kournopoulos, M. Kohns, T. Lindeboom, M. Wehbe, S. Febra, G. Jackson, C. S. Adjiman, and A. Galindo: 2020, 'Expanding the applications of the SAFT-γ Mie group-contribution equation of state: Prediction of thermodynamic properties and phase behavior of mixtures'. *Journal of Chemical & Engineering Data* **65**(12), 5862–5890. doi:10.1021/acs.jced.0c00746.

[28] D. Šarić, M. Kohns, and J. Vrabec: 2020, 'Dielectric constant and density of aqueous alkali halide solutions by molecular dynamics: A force field assessment'. *Journal of Chemical Physics* **152**, 164502. doi:10.1063/1.5144991.

[29] S. Stephan and H. Hasse: 2020, 'Enrichment at vapour-liquid interfaces of mixtures: Establishing a link between nanoscopic and macroscopic properties'. *International Reviews in Physical Chemistry* **39**(3), 319–349. doi:10.1080/0144235x.2020.1777705.

[30] K. Zhu and E. A. Müller: 2020, 'Generating a machine-learned equation of state for fluid properties'. *Journal of Physical Chemistry B* **124**(39), 8628–8639. doi:10.1021/acs.jpcb.0c05806.

Appendix: PIMS-II symbols

concept identifier suffix	symbol, position in taxonomy, explanation
Assignment	ASG $\sqsubseteq$ EAN (EqualityArticulation) $\sqsubseteq$ Dyad *for* (about) an object, assigns a value to a variable
DataItem	DI $\sqsubseteq$ DAN (DigitalArticulation) $\sqsubseteq$ AN (Articulation) dyad, triple, or n-tuple of digital conventionals
Interlocutor	IO $\sqsubseteq$ Agent $\sqsubseteq$ Object agent that is capable of two-way communication
KnowledgeClaim	KC $\sqsubseteq$ Claim $\sqsubseteq$ PN (Proposition) states what was found from research data
PropertyClaim	PC $\sqsubseteq$ KC (KnowledgeClaim) $\sqsubseteq$ Claim knowledge claim concerning a property
QualifiedLaw	QL $\sqsubseteq$ Law $\sqsubseteq$ Rule law of entailment by qualified necessity
ReproducibilityClaim	RC $\sqsubseteq$ QL (QualifiedLaw) $\sqcap$ VC (ValidityClaim) states whether another claim or underlying data are (ir)reproducible or have (not) been reproduced
ResearchQuestion	RQN $\sqsubseteq$ QN (Question) $\sqsubseteq$ TL (Topical) question suitable as a scientific research topic
Topical	TL $\sqsubseteq$ PN $\sqsubseteq$ SCO (SemioticCollective) proposition that can be a topic or subject matter (*e.g.*, a question with free information slots [18])
TopicalSum	TLS $\sqsubseteq$ PL (Plurality) $\sqsubseteq$ SCO (SemioticCollective) plurality of independent topicals, *i.e.*, summands
Validation	VAL $\sqsubseteq$ EVA $\sqsubseteq$ IPR (Interpretation) $\sqcap$ TEL (Telesis) evaluation of a cognition, yielding a validity claim
ValidityClaim	VC $\sqsubseteq$ Claim $\sqsubseteq$ PN (Proposition) claims (in)accuracy or (dis)trust for another claim

relation identifier suffix	symbol, position in relational hierarchy, explanation
articulates	$\lhd_{\mathsf{a}} \sqsubseteq \dddot{\leq}$ (isSemioticMemberOf) $\sqsubseteq (\leq \sqcap \check{\mathsf{C}})$ $\lhd_{\mathsf{a}}\delta\varphi$ means that δ articulates the proposition φ
hasSubjectMatter	$\mathsf{A} \sqsubseteq \hat{\mathsf{A}}$ (isAbout) $\sqsubseteq \mathsf{R}$ (isRepresentamenFor) $\mathsf{A}\varphi q$ means q is the (unique) subject matter of φ
isAssertedBy	$\dot{\mathsf{B}} \sqsubseteq \mathsf{B}$ (isExpressedBy) $\sqsubseteq \dot{\mathsf{P}}^{-}\dot{\mathsf{P}}$ (overlapsWith) $\dot{\mathsf{B}}\varphi a$ means that interlocutor a claims/claimed φ
isAssignmentFor	$\hat{\mathsf{A}}_{\mathsf{D}} \sqsubseteq \hat{\mathsf{A}}$ (isAbout) $\sqsubseteq \mathsf{R}$ (isRepresentamenFor) $\hat{\mathsf{A}}_{\mathsf{D}}do$ means that d is an assignment referring to o
isEvaluatorIn	$\ddot{\mathsf{P}}_{\mathsf{e}} \sqsubseteq \ddot{\mathsf{P}}_{\iota}$ (isInterpreterIn) $\sqsubseteq (\ddot{\mathsf{P}}_{\mathsf{a}} \sqcap \ddot{\mathsf{P}}_{\kappa})$ $\ddot{\mathsf{P}}_{\mathsf{e}}a\tau$ means that evaluation τ is conducted by a
isInterpreterIn	$\ddot{\mathsf{P}}_{\iota} \sqsubseteq \ddot{\mathsf{P}}_{\mathsf{a}}$ (isAgentIn) $\sqcap \ddot{\mathsf{P}}_{\kappa}$ (isParticipantInCognition) $\ddot{\mathsf{P}}_{\iota}a\kappa$ means that cognitive action κ has the agent a
isLiterally	λ is a datatype property; $\lambda s\ell$ associates the OWL object s with the RDFS elementary data value ℓ
isOrthodataWithin	$\check{\mathsf{C}}_{\perp} \sqsubseteq \check{\mathsf{C}}$ (isSemioticallyConstitutiveOf) $\sqsubseteq \ddot{\mathsf{C}}$ $\check{\mathsf{C}}_{\perp}\delta\Lambda$ means: What δ articulates is substantial to Λ
isReferenceFrameFor	$\dddot{\mathsf{R}} \sqsubseteq \mathsf{R}\mathsf{R}^{-}$ (sharesReferentWith) $\dddot{\mathsf{R}}ss'$ means that s is a context or precondition of s'
isResultOf	$\bar{\mathsf{R}} \sqsubseteq \ddot{\mathsf{P}}_{\kappa} \sqsubseteq \ddot{\mathsf{P}}$ (isParticipantIn) $\bar{\mathsf{R}}s\kappa$ means that the cognition κ has the outcome s
isTopicalFactorIn	$\check{\mathsf{C}}_{\mathsf{T}} \sqsubseteq \check{\mathsf{C}} \sqsubseteq \ddot{\mathsf{C}}$ (isConstitutiveOf) $\check{\mathsf{C}}_{\mathsf{T}}qq'$ means q is a factor in the topical product q'
isTopicalSummandIn	$\lhd_{\mathsf{T}} \sqsubseteq \lhd_{\mathsf{p}}$ (isMemberOfPlurality) $\sqsubseteq \dddot{\leq}$ $\lhd_{\mathsf{T}}qq'$ means q is a summand in the topical sum q'
isTriadOf	3 is a quaternary predicate; for a Peircean triad [5], $3\kappa e_1e_2e_3$ means that κ has triadic form e_1—e_2—e_3
isUnitOf	$\check{\mathsf{C}}_{\mathsf{u}} \sqsubseteq \check{\mathsf{C}}$ (isSemioticallyConstitutiveOf) $\sqsubseteq \ddot{\mathsf{C}}$ $\check{\mathsf{C}}_{\mathsf{u}}uw$ means that the value w is given in units of u
isValueInAssignment	$\ddot{\mathsf{D}} \sqsubseteq \ddot{\mathsf{E}}_{=} \sqsubseteq \ddot{\mathsf{E}}_{\mathsf{D}}$ (isSecondInDyad) $\ddot{\mathsf{D}}wd$ means that d assigns the value w to a variable
isVariableInAssignment	$\dot{\mathsf{D}} \sqsubseteq \dot{\mathsf{E}}_{=}$ (isLeftHandSideIn) $\sqsubseteq \dot{\mathsf{E}}_{\mathsf{D}}$ (isFirstInDyad) $\dot{\mathsf{D}}vd$ means that d assigns a value to the variable v