Formal Ontology in Information Systems
N. Aussenac-Gilles et al. (Eds.)

doi:10.3233/FAIA231136

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] and 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, 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 a priority for innovation in the engineering sciences. The digital transformation requires making the knowledge claims from scientific research data machine-actionable, so that they can be integrated and analysed with minimal human intervention. Up until now, the depth of digitalization is often 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 demand too much effort from researchers, much of which goes into metadata that contribute nothing to an improved reuse of knowledge. These shortcomings are related. Data documentation and annotation are complicated *and* of little use whenever the metadata that make knowledge reusable are not prioritized. Addressing this gap, we discuss metadata standardization efforts targeted at documenting the knowledge status of data; we refer to such an annotation as *epistemic metadata*. We propose a schema for epistemic metadata, with a focus on knowledge and reproducibility claims, that is designed to be user-friendly and flexible enough to apply to a spectrum of circumstances and validity assessments. These developments are implemented as part of the PIMS-II ontology. They were conducted in line with requirements procured through a case study on papers and claims from molecular modelling and simulation.

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

[1]Corresponding Author: Martin Thomas Horsch, e-mail: martin.thomas.horsch@nmbu.no.

1. Introduction

The creation of Industry Commons, driven by effective data documentation practices and standards, makes it easier to *produce new knowledge from existing knowledge*. Metadata thereby become constitutive to the mode of production of knowledge [1,2,3].[2] Much of the economic and societal benefits expected from this transition will hinge on the possibility of deploying data-driven research and innovation in business- and security-critical environments and fields of work. To facilitate the reliability and trust required by such use cases, the research data must become and remain *explainable AI ready* (XAIR), for which they need to be stored and exchanged jointly with metadata that characterize their knowledge status [4]. This work is about such metadata, which we call *epistemic metadata*, and their nature and use in computational engineering; it is driven by use-case considerations from molecular modelling and simulation, and underlying to it, there has been a domain-specific requirements analysis and community discussion [5,6,7].

Inappropriate data management results in *dark data* [8,9], which poses an increasing challenge due to the overall volume of data that are being produced. Dark data are not FAIR (findable, accessible, interoperable, reusable [10]), with many undesirable consequences. In particular, only FAIR data can be XAIR. Where the metadata are inadequate, incomplete, or missing [9], any knowledge that has been obtained from the data (or might be obtained in the future by reusing the data) is blurred, or *epistemically opaque* [8,11,12]. Applications of machine learning in molecular modelling are increasingly powerful and diverse [13,14,15], but their potential for a reliable deployment depends on XAIR data. Experimental or simulation data used for training a surrogate model need to be made available jointly with metadata that certify their high quality (*e.g.*, low noise). Where that cannot be guaranteed, the "inductive risk" increases [12]: The model may be biased, and its predictions may become quantitatively and qualitatively inaccurate.

Reproducibility is known to be one of the major challenges to open science and the scientific process at large [16]; *e.g.*, examining applied mathematics papers, Riedel *et al.* [17] find the reported results to be completely reproducible in only four out of 108 cases, and partially reproducible in only one other case. To advance on this challenge, *reproducibility networks* are being formed, for which the RIOT principles (reproducibility, interpretability, openness, transparency [18]) were formulated as guidelines, complementing the FAIR principles. Unfortunately, when working toward an improved reproducibility and compliance with the FAIR principles, communities risk overshooting by formulating excessive demands on the data provenance (*i.e.*, origin and genesis) documentation. This is a shortcoming that has also been observed in attempts at building computational engineering information systems, which led community stakeholder discussions within the OntoCommons project to formulate the DORIC principles [19] in an effort to make future attempts more targeted and successful: While there is no doubt that recording the data provenance can bring some benefit to information systems, this must be balanced against the effort required to provide these metadata.

[2] "Labour is tool-mediated. In case of knowledge-producing labour, these tools are *formalizing means* such as *sign systems*" (Azeri [2], emphases added by us).

This work starts from the basic understanding that there is an overarching need for an improved AI-readiness and machine-actionability of data [20], particularly in view of explainable AI applications, and that a key performance indicator for meeting this need is the *epistemic FAIRness* of the data, *i.e.*, to what extent FAIR data are accompanied by FAIR epistemic metadata. It addresses this challenge as an application of formal ontology development, informed by the requirements for data and metadata documentation for computational engineering information systems gathered from a multiple disciplinary case studies [5,6,7,19,21].

2. Epistemic Metadata

For a research outcome, based on or including research data δ, we can distinguish:

- The *knowledge claim* φ, *i.e.*, something that, it is claimed, δ shows us;
- their *provenance* ("δ and/or φ come from the cognitive process κ");
- a proposed *epistemic grounding* ("accept the claim φ by virtue of γ");
- any *validity claim(s)* ψ supporting or opposing the claim φ.

The above are what we consider to be epistemic metadata [22]; they can be combined with the data δ in the form of a FAIR digital object [23,24].

This conceptualization of the information to be dealt with by digital platforms for the engineering sciences is based on ingesting *claims*, especially knowledge claims, and not *knowledge* as such. This is done because even carefully curated collections of engineering data are to some extent contradictory internally. More generally, conflicting claims are part of the scientific process as such, and we do not want to force the data infrastructure to favour one of the sides each time that there is such a dispute. But an information system becomes inconsistent if it holds that "(it is known to us that) a knows φ, while b knows $\neg\varphi$," or else it would be necessary to abandon the widespread notion that only a true belief can be knowledge. Otherwise, "a knows φ" entails φ, and "b knows $\neg\varphi$" entails $\neg\varphi$, yielding a contradiction. However, there is no danger to logical consistency from holding simultaneously that "(it is known to us that) a *claims* to know φ" and "(it is known to us that) b *claims* to know $\neg\varphi$." We therefore prefer this.

This article is *not* about provenance documentation. There is no urgent need to discuss it in depth: First, because this work aims at making research data and scientific claims machine-actionable and AI-ready, not their provenance documentation. Second, in practice, there is already an excessive focus on provenance – more care is taken to state *where the data come from* as opposed to *what the data mean* [22]. And even where rather comprehensive provenance documentation requirements are met, such as with CHADA [25,26] or MODA [27], this in and of itself hardly contributes anything to making the research outcomes AI-ready [4,28]. Third, actionable workflow descriptions from the domain level in computational engineering (*e.g.*, AiiDA [29], more recently ModGra [30,31]) up to the general level (*e.g.*, represented by BPMN [32,33]) are already in place. We have discussed provenance documentation elsewhere – at the computational engineering domain level by OSMO [34], the ontologization of MODA, and as cognitive processes by using the PIMS-II mid-level ontology [22,35]. These discussions continue to apply, and the proposed solutions continue to remain viable for their respective purpose.

3. Knowledge Claims and Validity Claims

It would be a mistake to understand knowledge as a set of propositions that happen to correctly describe a state of affairs and otherwise are detached from the world. Scientific knowledge is the output of scientific labour and realizes itself only when used in research and development practice; "there is no knowledge independent of the knowing activity," as Azeri argues [2]. Our approach to conceptualizing knowledge and validity claims is anchored in the mereosemiotics paradigm [22,35,36] and Peircean semiotics [37]. Consequently, knowledge is analysed as something that is inherently dynamic, arising from and realizing itself in cognitive processes as the *action of signs*, through the use of *signs in action* [38].

Within this specific framework and using the PIMS-II ontology, a typical scenario with a knowledge claim (KC) is described and documented as follows:

Free variables:
DigitalArticulation(δ), InformationProcessing(ι), Interlocutor(a), KnowledgeClaim(φ), Topical(q).

Knowledge graph pattern:
isAssertedBy(φ, a).
isAbout(δ, q), hasSubjectMatter(φ, q).
isSignIn(δ, ι), isObjectIn(q, ι), isInterpretantIn(φ, ι), isInterpreterIn(a, ι).

(Read: "φ is asserted by a, δ is about q, φ has the subject matter q," *etc.*)

In Fig. 1 (top), this schema is visualized as a knowledge graph pattern (top right) in combination with a Peircean semiotic triad (top left). See also Tables 1 and 2; for more detail, *cf.* the PIMS-II OWL ontology TTL file.[3] In the same way, the example for validity claims (VCs) shown in Fig. 1 (bottom) can be denoted by:

Free variables:
Claim(φ), Cognition(κ), GoalDirectedAgent(b), Intention(t), Validation(τ), ValidityClaim(ψ).

Knowledge graph pattern:
isAbout(ψ, φ), isResultOf(φ, κ).
isGoalFor(t, b), isRepresentamenFor(t, κ).
isAssertedBy(ψ, b), isRepresentamenFor(ψ, κ).
isSignIn(t, τ), isObjectIn(κ, τ), isInterpretantIn(ψ, τ), isInterpreterIn(b, τ).

Accordingly, a VC is a claim that has another claim as its referent; it helps establish to what degree that claim is accurate or inaccurate, to what extent it should be trusted or distrusted. This includes reproducibility claims (RCs), *cf.* Section 5.

The taxonomy of claims in the PIMS-II ontology was finalized during the first stage of our case study – *cf.* the first-stage report from the case study [5] for more detail and a previous publication [22] for a concise summary of that work.

[3] http://www.molmod.info/semantics/pims-ii.ttl

Table 1. Selected unary predicates (owl:Class) from the PIMS-II ontology [22,35].

concept	**explanation**, including selected rules from the ontology
Articulation(δ)	collective of concrete realizations with the same semiotic role $\models$ SemioticCollective(δ) $\wedge\ \forall s\,(\mathsf{isSemioticMemberOf}(s,\delta) \rightarrow \neg\mathsf{SemioticCollective}(s))$ *e.g., two realizations (members) of the same articulation could be two different copies of the text or string "Hello world!"*
Claim(φ)	proposition asserted by someone or held by an intelligent agent $\models$ Proposition(φ)
Cognition(κ)	process within which objects are being represented by signs
DigitalArticulation(δ)	collective of members that are copies of the same digital content $\models$ Articulation(δ)
GoalDirectedAgent(a)	agent that has an internal representation of own goals [39] $\models\ \exists t\ (\mathsf{Intention}(t) \wedge \mathsf{isGoalFor}(t,a))$
InformationProcessing(ι)	cognitive step in which information is handled and/or revised $\models$ Semiosis(ι)
Intention(t)	proposition that constitutes an aim or goal $\models$ Proposition(t)
Interlocutor(a)	agent that can be addressed and that can address others
KnowledgeClaim(φ)	claim about the knowledge status of something (*e.g.*, data) $\models$ Claim(φ) $\wedge\ \exists q\ (\mathsf{hasSubjectMatter}(\varphi,q) \wedge \mathsf{Topical}(q))$
PropertyClaim(φ)	knowledge claim concerning a property $\models$ KnowledgeClaim(φ)
Proposition(φ)	collective of members to which the same semantics are ascribed $\models$ SemioticCollective(φ) $\wedge\ \exists q\,\mathsf{isAbout}(\varphi,q)$ $\wedge\ \forall\delta\,(\mathsf{isSemioticMemberOf}(\delta,\varphi) \rightarrow \mathsf{Articulation}(\delta))$ *e.g., two articulations (members) of the same proposition could be* Articulation *objects for "number 5 exists" and "*$\exists x\ x = 5$*."*
ReproducibilityClaim(ψ)	validity claim addressing the reproducibility of another claim $\models$ ValidityClaim(ψ) $\wedge\ \exists\delta\,\mathsf{isOrthodataWithin}(\delta,\psi)$
Semiosis(σ)	process of using a sign; concept introduced by Peirce [37] $\models$ Cognition(σ) $\wedge\ \exists asos'\ \big(\mathsf{isInterpreterIn}(a,\sigma)$ $\wedge\ \mathsf{isSignIn}(s,\sigma) \wedge \mathsf{isObjectIn}(o,\sigma) \wedge \mathsf{isInterpretantIn}(s',\sigma)\big)$
SemioticCollective(s')	collective of members partaking in a sign-object relation jointly $\models\ \exists s\,\mathsf{isSemioticMemberOf}(s,s')$
Topical(q)	subject matter, *e.g.*, proposition with free information slots [40] $\models$ Proposition(q)
TopicalSum(q')	plurality of independent topicals, *i.e.*, summands $\models\ \neg\mathsf{Proposition}(q') \wedge \exists q\,\mathsf{isTopicalSummandIn}(q,q')$
Validation(τ)	evaluation of a cognition, yielding a validity claim $\models\ \exists t\kappa\psi\ \big(\mathsf{Intention}(t) \wedge \mathsf{Cognition}(\kappa) \wedge \mathsf{ValidityClaim}(\psi)$ $\wedge\ \mathsf{isSignIn}(t,\tau) \wedge \mathsf{isObjectIn}(\kappa,\tau) \wedge \mathsf{isInterpretantIn}(\psi,\tau)\big)$
ValidityClaim(ψ)	claim that supports or opposes the validity of another claim $\models$ Claim(ψ) $\wedge\ \exists\varphi\ (\mathsf{Claim}(\varphi) \wedge \mathsf{isAbout}(\psi,\varphi))$

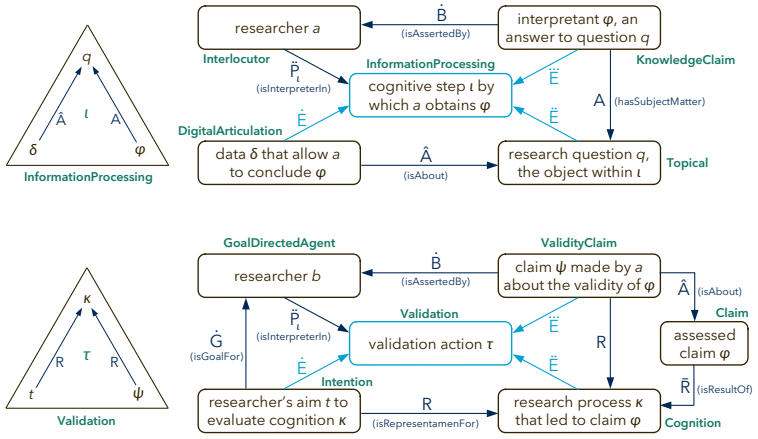

Figure 1. Knowledge claim schema (top) and validation claim schema (bottom). Left: Peircean semioses [37], using notation from previous work [22,35]; right: Knowledge graph patterns.

4. Research Questions

Information systems in computational engineering, specially research data infrastructures, will increasingly need to deal not only with increasing quantities of data and records, but also with more diverse data and records. Findability of entries is the most basic criterion with which any effort toward compliance with the FAIR principles ought to begin. Both human users and interoperating digital platforms and tools will best make sense of a lookup functionality that allows them to query, "what do you have/know about q?" This turns topicality, or "subject matter" and "aboutness" of information content, into an annotation that is both important and fundamental. We are concerned with epistemic metadata, for which such annotations are needed: As shown in Fig. 1, a validity claim is about a knowledge claim, while a knowledge claim has a research question as its subject matter. But we cannot restrict ourselves to these items only; from the level of a single data item, over FAIR digital objects containing a series of claims as annotation of data, up to whole journal articles or data publications and even collections of publications – all will benefit from an improved findability by having a topic. Naturally, work in foundational and applied ontology has been approaching this semantic field from multiple angles [40,41,42,43,44]; *e.g.*, the OBO's information artifact ontology (IAO) documents its relation IAO_0000136 (is_about) with the "definition source: Smith, Ceusters, Ruttenberg, 2000 years of philosophy."[4]

One foundational issue is that of the relationship between aboutness and designation or representation. Building upon the aforementioned 2000 years, Ruttenberg in the same IAO entry[4] remarks that information "can be topical without explicitly mentioning the topic." This is a feature that aboutness shares with signification and representation. In Augustine's example, smoke signifies fire, but the smoke does not "mention" the fire. In Peirce's example, a sunflower represents the sun, but it does not "mention" the sun. Here we concur: Data can be about

[4] http://purl.obolibrary.org/obo/IAO_0000136

Table 2. Selected binary predicates (owl:ObjectProperty) from the PIMS-II ontology [22,35].

relation	**explanation**, including selected rules from the ontology
$\mathsf{hasSubjectMatter}(s, q)$	q is the unique subject matter of s
	$\models\ \mathsf{isAbout}(s, q)\ \wedge\ (\mathsf{Topical}(q)\ \vee\ \mathsf{TopicalSum}(q))$
	$\wedge\ \forall q'\ (\mathsf{hasSubjectMatter}(s, q')\ \rightarrow\ q = q')$
$\mathsf{isAbout}(s, q)$	s is about q, and might be about other things as well
	$\models\ \mathsf{isRepresentamenFor}(s, q)$
	$\wedge\ (\mathsf{Articulation}(s)\ \vee\ \mathsf{Proposition}(s))$
$\mathsf{isAssertedBy}(\varphi, a)$	a affirmatively claims φ
	$\models\ \mathsf{Claim}(\varphi)\ \wedge\ \mathsf{Interlocutor}(a)$
$\mathsf{isConstitutiveOf}(o, o')$	o contributes causally to o' fulfilling a certain function
$\mathsf{isGoalFor}(t, a)$	agent a regards t as something that is to be reached
	$\models\ \mathsf{Intention}(t)\ \wedge\ \mathsf{GoalDirectedAgent}(a)$
$\mathsf{isInterpretantIn}(s, \sigma)$	s is the third element, the *interpretant*, in semiosis σ [37]
	$\models\ \mathsf{isParticipantInCognition}(s, \sigma)\ \wedge\ \mathsf{Semiosis}(\sigma)$
$\mathsf{isInterpreterIn}(a, \kappa)$	cognitive action κ has the agent a
	$\models\ \mathsf{isParticipantInCognition}(a, \kappa)$
$\mathsf{isParticipantInCognition}(o, \kappa)$	cognition κ requires the physical presence of o
	$\models\ \mathsf{Cognition}(\kappa)$
$\mathsf{isObjectIn}(o, \sigma)$	o is the second element, the *object*, in semiosis σ [37]
	$\models\ \mathsf{Semiosis}(\sigma)$
$\mathsf{isOrthodataWithin}(\delta, \psi)$	δ articulates something that is substantial to ψ
	$\models\ \mathsf{isConstitutiveOf}(\delta, \psi)\ \wedge\ \mathsf{DigitalArticulation}(\delta)$
$\mathsf{isRepresentamenFor}(s, o)$	sign-object relation between representamen s and referent o
$\mathsf{isResultOf}(\varphi, \kappa)$	cognition κ has the outcome φ
	$\models\ \mathsf{isParticipantInCognition}(\varphi, \kappa)$
$\mathsf{isSemioticMemberOf}(s, s')$	part-to-whole relation for a semiotic collective
	$\models\ \mathsf{isConstitutiveOf}(s, s')\ \wedge\ \mathsf{SemioticCollective}(s')$
$\mathsf{isSignIn}(s, \sigma)$	s is the first element, the *sign*, in semiosis σ [37]
	$\models\ \mathsf{isParticipantInCognition}(s, \sigma)\ \wedge\ \mathsf{Semiosis}(\sigma)$
$\mathsf{isTopicalFactorIn}(q, q')$	q is a factor in the topical product q'
	$\models\ \mathsf{isConstitutiveOf}(q, q')$
	$\wedge\ (\mathsf{Topical}(q)\ \vee\ \mathsf{TopicalSum}(q))\ \wedge\ \mathsf{Topical}(q')$
$\mathsf{isTopicalSummandIn}(q, q')$	q is a summand in the topical sum q'
	$\models\ \mathsf{isSemioticMemberOf}(q, q')\ \wedge\ \mathsf{Topical}(q)\ \wedge\ \mathsf{TopicalSum}(q')$

something, especially a research question, without mentioning that thing.[5] Indeed we are separating data from claims because the same data can be reused over and over to help address many different scientific problems, through a research step classified as InformationProcessing within PIMS-II, *cf.* Fig. 1 (top left) and Table 1. When a user looks for entries about these issues, these data should be retrieved, *i.e.*, the data should be understood as being about research questions

[5]Reading *mention* here as some sort of containment, constitutivity, or abstract parthood relation expressing the inclusion of a designator of the mentioned object or at least "revolving around it," corresponding to the definition of *about* proposed by Ryle [41]. In our conceptualization, data can *become* about some phenomenon once some researcher has used them to analyse that phenomenon; this goes substantially beyond what Ryle would have admitted as aboutness.

without having mentioned or contained these questions beforehand. If we accept this, it follows that a single data item can be about many different things at once; isAbout is not an owl:FunctionalProperty. We reserve hasSubjectMatter (*cf.* Table 2) for the use case where, following Yablo, the topical object is identified uniquely, *i.e.*, "as *the* subject matter of sentence S – the one it is *exactly* about" [43, p. 44].

Yablo proposes to think 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 to another just if they have equally many stars" [43, p. 26]. This looks practical in the context of knowledge-graph based technologies: If the subject matter of a knowledge claim is given by the question that it answers, it might be expressed in SPARQL, as a graph pattern using wildcards, in a technical implementation. The remaining challenge consists in progressing from topics of data items and claims with little internal structure, along the lines of Yablo's "how many stars are there?," to a topic annotation that would adequately describe any of the more heterogeneous sorts of records that we must prepare to manage on a research data infrastructure. Barton *et al.* [40] call a wildcard an information slot, and an admissible valuation an information filler. These information entities can be combined into a hierarchical structure by which, *e.g.*, it is the definition of the second level that "a 2nd-level slot of s is a slot of a slot of s that is not a slot of a slot of a slot of s" [40, Section 5.2].

Like Barton *et al.*'s, the way of combining partial topics into a whole proposed in the present work is also capable of forming hierarchical structures; its implementation is less complicated, or at least it appears so to us, and is designed to integrate itself into the system of semiotic collectives from PIMS-II. It is based on two kinds of operations: First, a weak association of topical entities into a TopicalSum collective, carried out by using the relation isTopicalSummandIn. Second, a strong association, forming a TopicalProduct (subclass of Topical). The part-to-whole relation for the product is given by "isTopicalFactorIn," *cf.* Table 2,

$$\begin{aligned} q = q_1 + q_2 + \ldots + q_n \iff & \quad (\mathsf{isTopicalSummandIn}(s,q) \leftrightarrow s \in \{q_1, \ldots, q_n\}) \,, \\ q = q_1 q_2 \cdots q_n \iff & \quad (\mathsf{isTopicalFactorIn}(s,q) \leftrightarrow s \in \{q_1, \ldots, q_n\}) \,, \end{aligned}$$

While the concept TopicalProduct is subsumed under Topical, the concept TopicalSum is disjoint with Topical, *cf.* Table 1.

The topical product is meant for combining closely related subtopics when there is a non-trivial correlation or interaction; the topical sum is meant for collating subtopics that just stand beside each other [6, therein, Section 1.4].

5. Reproducibility Claims

Research that attempts to reproduce others' work will, if it is published, typically result in a passage or remark in a journal article that reports on the success of that attempt *explicitly*. Thereby, the original research outcome is corroborated or contradicted. However, reproducibility claims can be either explicit or implicit, in line *e.g.* with Grice's distinction between *what is said* and *what is implicated*.

In particular, it is not common practice today to make any explicit claims about the expected reproducibility of the own research outcomes, *i.e.*, those reported in the work itself. But does this really mean that research articles do not convey an understanding that their own outcomes should be reproducible? If that was the case, a refutation of research outcomes would not affect the standing of the paper reporting them. After all, the scenario at face value then looks like this:

1. Paper 1 reports that researcher a did κ and found φ.
2. Paper 2 reports that to assess Paper 1, researcher b did something rather similar to κ, from which they found something rather different from φ.
3. Nobody disputes that "a did κ and found φ." Therefore, the claim made by b in Paper 2 has no bearing whatsoever on a's claim from Paper 1.

From practice we know that this scenario cannot be taken at face value. The opposite holds: By publishing a scientific article, its authors do claim that the results they report are reproducible, they are just not making it explicit. Within a pragmatics framework, we may here be dealing with a case of *conventional implicature* in line with Grice's analysis of the word "therefore" [45, p. 25f.]. In a scientific context, "we obtained data δ, therefore we know φ," or a more elaborate passage to that effect, implicates that the step of concluding φ from δ is legitimate according to established disciplinary good practice. For simulation or experimental data, it thereby almost always implicates compliance with a basic reproducibility expectation. If they could be made explicit and documented, these reproducibility claims that a work makes about its own outcomes could have an important role to play in information systems for open science. As the scenario above shows, what a reproduction attempt strictly speaking really does is not to corroborate or to contradict the original knowledge claim; instead, it supports or opposes the original reproducibility claim. But as of now, following usual practice, this means that it supports or opposes a claim that was never made explicit and is only imputed to the original work. It is not AI-ready or machine-actionable.

Not only are the original reproducibility claims established by implicature; further complicating the issue, the extent of such claims is not agreed upon, but based on unspoken rules. Consider another possible mechanism, *conversational implicature*: If the researchers did not believe their results to be sufficiently reproducible, it would have been inappropriate[6] to submit them for publication as a journal article to begin with. A paper *was* submitted to an engineering or physics journal, however. It follows that a claim for the results to be ordinarily reproducible is implicated. But what does "ordinarily reproducible" mean exactly? Since by the very nature of implicature it is not usually spoken about, different people in the field have different expectations. The original authors could even recede to the position "we just described what we did and what we saw happening, no more, no less." This opacity of reproducibility claims becomes a barrier once we need to make them machine-actionable [20] to support the validation of research outcomes through base services on an open science information system.

[6]That is, it would violate Grice's conversational maxim of "*Quality.* I expect your contributions to be genuine and not spurious. If I need sugar, [...] I dot not expect you to hand me salt; if I need a spoon, I do not expect a trick spoon made of rubber" [45, p. 28]. When we need research outcomes, we do not expect 'trick research outcomes' made of irreproducible claims.

But why is it that researchers do not write down any reproducibility expectations for their own results? After all, reproducibility expectations about own research outcomes are made explicit in works *about* reproducibility [16,46], so in principle, it would be possible to do it regularly. Plausible explanations include:

1. It is taken for granted that the results are ordinarily reproducible, as usual.
2. The authors do not want to attract any attention to the issue of reproducibility. They do not know how reproducible their results are, and they are content with deceiving the community into just assuming they are.
3. A statement on how the outcome could/should be reproduced is seen as superfluous. After all, it has already been stated how the original work was done in the first place – just do it again. It is seen as other people's job to figure out what parts of the described research process are essential (orthodata, in our terminology) and which are circumstantial (paradata).
4. They have little practice in documenting or writing about reproducibility, and there are no clear community rules and conventions, other than unwritten rules that are not even really agreed upon by all in the field.

The first three explanations do not provide a good excuse, they are rather indicative of careless or egoistic behaviour – scientific communities can justifiably take action against them: No. 1 does not work, since there is so little agreement on the ordinary level of reproducibility that we cannot just take its understanding for granted; researchers may be unaware of this, but institutions, organizations, projects, and bottom-up initiatives are already raising that awareness [18] and will need to continue to do so. No. 2 can be overcome culturally, in the same way as numerical values without error bars are seen as incomplete data points today, whereas they used to be acceptable in the past; journals and data infra-

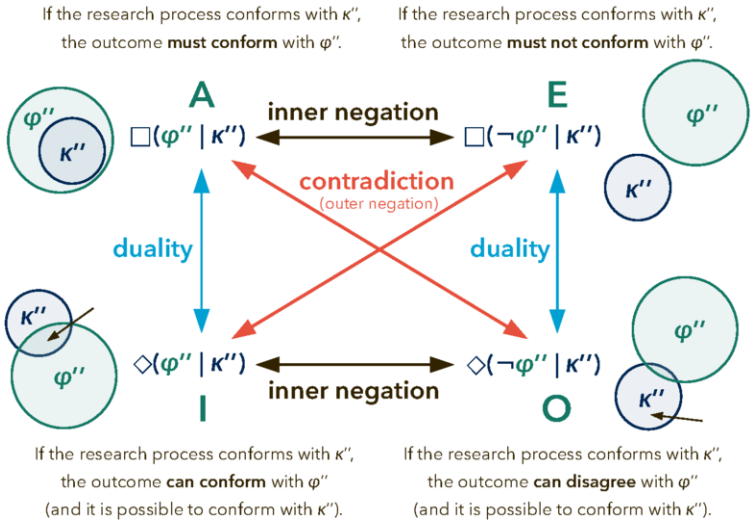

Figure 2. Modal square of opposition for RCs. Therein, φ'' are the *outcome orthodata* (part of the outcome that is covered by the reproducibility claim), and κ'' are the *provenance orthodata* (part of the provenance documentation that a reproduction attempt needs to conform with).

structures can impose this legitimately as a requirement. No. 3 is laziness; the original researchers know their own process best. They should say what part of their described procedures others will need to keep as they were (orthodata), as opposed to information that they were providing more as a side note (paradata).

No. 4 is a valid point, on the other hand. Researchers experience little assistance toward an interoperable documentation of their own reproducibility expectations with regard to outcomes of their own work, and there has been no metadata standard supporting such an interoperable documentation, until now. PIMS-II introduces such a metadata standard, for which this work also proposes a formal notation in terms of dedicated modal operators, *cf.* Fig. 2. However, while the gap mentioned above is genuine, there is no shortage of literature on reproducibility and conceptualizations of reproducibility as such. Leonelli [47] distinguishes five kinds of reproducibility studies (computational, direct, scoping, expertise, and observation) plus the related sixth category of "irreproducible observation." Fineberg *et al.* [48] define computational reproducibility as obtaining results that are in line with previous results while retaining some of the provenance information (input data, computational steps, methods, code, and conditions) identically; replicability is defined as getting consistent results with other than the original data and with similar methods. Such reproducibility and replicability definitions are reviewed by Plesser [49]. Beside showing that there is a substantial semantic heterogeneity between and even within disciplines, as it should be expected, Plesser [49] also points to the particularly promising approach toward an analytical formalization proposed by Patil *et al.* [50]. Their work introduces a highly expressive notation based on conditional probabilities. Most of the time, however, RCs do not involve any such detailed probability statements, and a less expressive, but more accessible and tractable framework could be appropriate.

We reduce Patil *et al.*'s idea from conditional probabilities to conditional modal operators, which is roughly equivalent to only allowing the conditional probabilities $p = 0$, $p > 0$, $p < 1$, and $p = 1$. The notation in Fig. 2 is also inspired by conditional probabilities, where the probability of A given B is denoted by $p(A \mid B)$. Accordingly, we write $\Box(\varphi'' \mid \kappa'')$ for the necessity of φ'' given κ'', and so on for the square's other three corners. Applied to RCs, 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 its outcome.

A positive reproducibility claim ψ, asserting the reproducibility of φ (obtained from κ), takes the form $\psi \equiv \Box(\varphi'' \mid \kappa'')$ with orthodata φ'' and κ'' such that $\varphi \models \varphi''$ and $\kappa \models \kappa''$. The corresponding negative RC, asserted when an attempt at reproducing the outcome has failed, is $\Diamond(\neg\varphi'' \mid \kappa'') \equiv \neg\Box(\varphi'' \mid \kappa'') \equiv \neg\psi$.

6. Evaluation

A requirement that many ontologies struggle to meet is that semantic artefacts for FAIR data must also be FAIR themselves [51]. The PIMS-II ontology was evaluated for FAIRness using the FOOPS! validator [52]. It was scored at 81%

compliance,[7] with 6/9 points from the category "findable," 2/3 from "accessible," 3/3 from "interoperable," and $8\frac{1}{3}/9$ from "reusable." This is a higher score than that of the good-practice example included in FOOPS! itself as an orientation (`https://w3id.org/example`), which is scored as having 78% FAIR compliance.

Similarly, for the technical side of the implementation, PIMS-II was evaluated using OOPS! [53], with good results: Three out of 41 pitfalls are highlighted as relevant (P11, P13, P36), but none of them as critical;[8] upon inspection these three are found to refer to legitimate design choices. Additionally, the case study [5,6], which will be carried further, is our means of continuously evaluating whether PIMS-II meets the metadata documentation needs for epistemic metadata.

7. Conclusion

By providing FAIR data and epistemic metadata, the knowledge status of data becomes both actionable to human users and machine-actionable to digital information systems. Research data management and the development of semantic artefacts for such platforms therefore requires a focus on capturing epistemic metadata: It should be made as easy as possible to provide an epistemic characterization. Further work in this direction will be done within the Mathematical Research Data Initiative (MaRDI) [54,55], one of the consortia of the German national research data infrastructure (NFDI). This will include building a knowledge graph of mathematical models that will make it possible for researchers from computational engineering to document their basic epistemic modelling decisions.

In addition to the co-development with the Metadata4Ing ontology [56] realized for documenting property claims [6, Section 1.2], it could be of interest to connect the VC documentation to other semantic artefacts; in particular, to the Citation Typing Ontology (CiTO) [57,58] and the VIMMP Validation Ontology (VIVO) [59]. The "enriched cited references," which are work in progress by Clarivate within the ISI Web of Knowledge, might also develop in this direction.

Before all else, it must be digitalized what the data mean, or *what the data are taken to mean*, *i.e.*, what knowledge claims have been formulated. As regards provenance, we recommend a clear focus on a documentation that is relevant epistemically, keeping it both useful and brief. In DataverseNO,[9] *e.g.*, "all observations and variables relevant for a replication of the study" are requested, but complicated diagrams such as MODA, CHADA, or ModGra are not required, nor are any excessive accounts of the provenance; the same holds for DaRUS [60]. In this way, the uptake of epistemic FAIRness will improve, so that more research data can become XAIR [4]. This is a prerequisite for arranging computational engineering information systems into an open ecosystem for interoperable and explainable data-driven tools and services [3,7], creating conditions that permit the reuse of domain-specific "knowledge as the mode of its own production" [1].

[7] To reproduce this, enter `http://www.molmod.info/semantics/pims-ii.ttl` on the FOOPS! validator front end at `https://foops.linkeddata.es/FAIR_validator.html` [52].

[8] Reproduce this by entering `http://www.molmod.info/semantics/pims-ii.ttl` on the OOPS! front end site at `https://oops.linkeddata.es/` [53].

[9] `https://dataverse.no/`.

Acknowledgment

M.T.H., S.C., and I.T.T. acknowledge DOME 4.0 and OntoCommons, Horizon 2020 grant agreements no. 953163 and 958371; S.S. and J.V. acknowledge WindHPC, BMBF grant identifier 16ME0613; B.S. acknowledges MaRDI, DFG project no. 460135501, within the NFDI programme of the Joint Science Conference (GWK). Creston Davis is acknowledged for suggesting the working hypothesis [1].

Supplementary information: https://dx.doi.org/10.5281/zenodo.7608074 [6].

References

[1] Davis C. "Knowledge as the mode of its own production." Working hypothesis; 2023.
[2] Azeri S. Activity, labour, and praxis: An outline for a critique of epistemology. Critique. 2019;47:585. Available from: https://dx.doi.org/10.1080/03017605.2019.1678267.
[3] Magas M, Kiritsis D. Industry Commons: An ecosystem approach to horizontal enablers for sustainable cross-domain industrial innovation. Internat J Product Res. 2022;60(2):479-92. Available from: https://dx.doi.org/10.1080/00207543.2021.1989514.
[4] Horsch MT, Schembera B, Preisig HA. European standardization efforts from FAIR toward explainable-AI-ready data documentation in materials modelling. In: Nichele S, Misra S, Molder A, editors. Proc. ICAPAI 2023. Piscataway: IEEE; 2023. To appear.
[5] Horsch MT, Schembera B. Epistemic metadata in molecular modelling: First-stage case-study report (10 cases). Kaiserslautern: Inprodat; 2023. Technical report no. 2023–A. Available from: https://dx.doi.org/10.5281/zenodo.7516532.
[6] Horsch MT, Chiacchiera S, Kohns M, Müller EA, Stephan S, Todorov IT, et al. Epistemic metadata in molecular modelling: Second-stage case-study report (12 claims). Kaiserslautern: Inprodat; 2023. Available from: https://dx.doi.org/10.5281/zenodo.7608074.
[7] Chiacchiera S, et al. Review of Domain Interoperability; 2023. Work in progress.
[8] Schembera B, Durán JM. Dark data as the new challenge for big data science and the introduction of the scientific data officer. Philos Technol. 2020;33(1):93-115. Available from: https://dx.doi.org/10.1007/s13347-019-00346-x.
[9] Schembera B. Like a rainbow in the dark: Metadata annotation for HPC applications in the age of dark data. J Supercomput. 2021;77:8946-66. Available from: https://dx.doi.org/10.1007/s11227-020-03602-6.
[10] Wilkinson MD, Dumontier M, Sansone SA, Bonino da Silva Santis LO, et al. Evaluating FAIR maturity through a scalable, automated, community-governed framework. Sci Data. 2019;9:174. Available from: https://dx.doi.org/10.1038/s41597-019-0184-5.
[11] Resch M, Kaminski A. The epistemic importance of technology in computer simulation and machine learning. Minds Machin. 2019;29(1):9-17. Available from: https://dx.doi.org/10.1007/s11023-019-09496-5.
[12] Sullivan E. Inductive risk, understanding, and opaque machine learning models. Philos Sci. 2022;89(5):1065-74. Available from: https://dx.doi.org/10.1017/psa.2022.62.
[13] Gebhardt J, Kiesel M, et al. Combining molecular dynamics and machine learning to predict self-solvation free energies and limiting activity coefficients. J Chem Inform Model. 2020;60(11):5319-30. Available from: https://dx.doi.org/10.1021/acs.jcim.0c00479.
[14] Zhu K, Müller EA. Generating a machine-learned equation of state for fluid properties. J Phys Chem B. 2020;124(39):8628-39. Available from: https://dx.doi.org/10.1021/acs.jpcb.0c05806.
[15] Jirasek F, Hasse H. Machine learning of thermophysical properties. Fluid Phase Equilib. 2021;549:113206. Available from: https://dx.doi.org/10.1016/j.fluid.2021.113206.

[16] Schappals M, Mecklenfeld A, et al. Round robin study: Molecular simulation of thermodynamic properties from models with internal degrees of freedom. J Chem Theory Comput. 2017;13:4270. Available from: https://dx.doi.org/10.1080/10.1021/acs.jctc.7b00489.
[17] Riedel C, Geßner H, Seegebrecht A, Ayon SI, et al. Including data management in research culture increases the reproducibility of scientific results. In: Proc. INFORMATIK 2022. GI; 2022. p. 1341. Available from: https://dx.doi.org/10.18420/inf2022_114.
[18] Ganley E, Coriat AM, Shenow S, Prosser D. Systemic problems require systemic solutions: The need for coordination and cooperation to improve research quality. BMC Res Notes. 2022;15:51. Available from: https://dx.doi.org/10.1186/s13104-022-05932-5.
[19] Horsch MT, Francisco Morgado J, Goldbeck G, Iglezakis D, et al. Domain-specific metadata standardization in materials modelling. In: Proc. DORIC-MM 2021. UKRI; 2021. p. 12. Available from: http://purl.org/net/epubs/work/50300311.
[20] Weiland C, Islam S, Broeder D, Anders I, Wittenburg P. FDO machine actionability. FDO Forum; 2022. Available from: https://dx.doi.org/10.5281/zenodo.7825649.
[21] Horsch M, Petrenko T, Kushnarenko V, Schembera B, Wentzel B, Behr A, et al. Interoperability and architecture requirements analysis and metadata standardization for a research data infrastructure in catalysis. In: Proc. DAMDID 2021. Springer; 2022. p. 166-77. Available from: https://dx.doi.org/10.1007/978-3-031-12285-9_10.
[22] Horsch MT, Schembera B. Documentation of epistemic metadata by a mid-level ontology of cognitive processes. In: Sales TP, et al., editors. Proc. JOWO 2022. Aachen: CEUR-WS; 2022. Available from: http://ceur-ws.org/Vol-3249/paper2-CAOS.pdf.
[23] Anders I, Blanchi C, Broeder D, Hellström M, et al. FDO requirement specifications. FDO Forum; 2023. Available from: https://dx.doi.org/10.5281/zenodo.7781925.
[24] Anders I, Blanchi C, Broeder D, Hellström M, et al. FAIR digital object technical overview. FDO Forum; 2023. Available from: https://dx.doi.org/10.5281/zenodo.7824713.
[25] Romanos N, Kalogerini M, Koumoulos EP, Morozinis K, et al. Innovative data management in advanced characterization: Implications for materials design. Mater Today Comm. 2019;20:100541. Available from: https://dx.doi.org/10.1016/j.mtcomm.2019.100541.
[26] CEN-CENELEC. Brussels: CEN (CWA 17815:2021 E); 2021. Available from: https://www.cencenelec.eu/media/CEN-CENELEC/CWAs/ICT/cwa17815.pdf.
[27] CEN-CENELEC. Brussels: CEN (CWA 17284:2018 E); 2018. Available from: https://www.cencenelec.eu/media/CEN-CENELEC/CWAs/RI/cwa17284_2018.pdf.
[28] Del Nostro P, Goldbeck G, Toti D. CHAMEO: An ontology for the harmonisation of materials characterisation methodologies. Appl Ontol. 2022;17(3):401-21. Available from: https://dx.doi.org/10.3233/AO-220271.
[29] Huber SP. Automated reproducible workflows and data provenance with AiiDA. Nature Rev Phys. 2022;4:431. Available from: https://dx.doi.org/10.1038/s42254-022-00463-1.
[30] Preisig HA. Documenting models comprehensively using a minimal graphical language. In: Yamashita Y, Kano M, editors. Proc. PSE 2021+. Amsterdam: Elsevier; 2022. p. 1021-6. Available from: https://dx.doi.org/10.1016/b978-0-323-85159-6.50170-6.
[31] CEN-CENELEC. Brussels: CEN (CWA 17960:2022 E); 2022. Available from: https://www.cencenelec.eu/media/CEN-CENELEC/CWAs/RI/cwa17960_2022.pdf.
[32] Rospocher M, Ghidini C, Serafini L. An ontology for the business process modelling notation. In: Garbacz P, Kutz O, editors. Proc. FOIS 2014. Amsterdam: IOS; 2014. p. 133-46. Available from: https://dx.doi.org/10.3233/978-1-61499-438-1-133.
[33] Ruecker B. Practical Process Automation. O'Reilly (ISBN 978-1-4920-6145-8); 2021.
[34] Horsch MT, Niethammer C, Boccardo G, et al. Semantic interoperability and characterization of data provenance in computational molecular engineering. J Chem Eng Data. 2020;65:1313. Available from: https://dx.doi.org/10.1021/acs.jced.9b00739.
[35] Horsch MT. Mereosemiotics: Parts and signs. In: Sanfilippo EM, Kutz O, Troquard N, Hahmann T, Masolo C, Hoehndorf R, et al., editors. Proc. JOWO 2021. Aachen: CEUR-WS; 2021. Available from: http://ceur-ws.org/Vol-2969/paper3-FOUST.pdf.
[36] Horsch MT, Chiacchiera S, Schembera B, Seaton MA, Todorov IT. Semantic interoperability based on the European Materials and Modelling Ontology and its ontological paradigm: Mereosemiotics. In: Chinesta F, Abgrall R, et al., editors. Proc. ECCOMAS 2020. Sci-

pedia; 2021. Available from: https://dx.doi.org/10.23967/wccm-eccomas.2020.297.
[37] Peirce CS. Logic as semiotic: The theory of signs. In: Buchler J, editor. Philosophical Writings of Peirce. New York: Dover (ISBN 978-0-48620217-4); 1955. p. 98-119.
[38] Atã P, Queiroz J. O externalismo semiótico ativo de C. S. Peirce e a cantoria de viola como signo em ação. Trans/Form/Ação. 2021;44(3):177-204. Available from: https://dx.doi.org/10.1590/0101-3173.2021.v44n3.15.p177.
[39] Conte R. Rational, goal-oriented agents. In: Computational Complexity. Springer; 2012. p. 2578-93. Available from: https://dx.doi.org/10.1007/978-1-4614-1800-9_158.
[40] Barton A, Toyoshima F, Vieu L, Fabry P, Ethier JF. The mereological structure of informational entities. In: Brodaric B, Neuhaus F, editors. Proc. FOIS 2020. IOS (ISBN 978-1-64368-128-3); 2020. p. 201-15. Available from: https://dx.doi.org/10.3233/faia200672.
[41] Ryle G. About. Analysis. 1933;1(1):10-2. Available from: https://dx.doi.org/10.1093/analys/1.1.10.
[42] Ceusters W. An information artifact ontology perspective on data collections and associated representational artifacts. In: Quality of Life through Quality of Information. IOS; 2012. p. 68-72. Available from: https://dx.doi.org/10.3233/978-1-61499-101-4-68.
[43] Yablo S. Aboutness. Princeton Univ. Press (ISBN 978-0-691-14495-5); 2014.
[44] Hawke P. Theories of aboutness. Australasian J Philos. 2018;96(4):697-723. Available from: https://dx.doi.org/10.1080/00048402.2017.1388826.
[45] Grice P. Studies in the Ways of Words. Cambridge, MA: Harvard Univ. Press; 1989.
[46] Stephan S, Dyga M, Alabd I, Lenhard J, Urbassek HM, Hasse H. Reproducibility of atomistic friction computer experiments: A molecular dynamics simulation study. Mol Sim. 2021;47:1509. Available from: https://dx.doi.org/10.1080/08927022.2021.1987430.
[47] Leonelli S. Rethinking reproducibility as a criterion for research quality. In: Fiorito L, et al., editors. Proc. Symposium on Mary Morgan. Emerald; 2018. p. 129-46. Available from: https://dx.doi.org/10.1108/s0743-41542018000036B009.
[48] Fineberg HV, Allison DB, Barba LA, Chong D, Freire J, Gabrielse G, et al. Reproducibility and Replicability in Science. Washington, DC: National Academies Press (ISBN 978-0-309-48616-3); 2019. Available from: https://dx.doi.org/10.17226/25303.
[49] Plesser HE. Reproducibility vs. replicability: A brief history of a confused terminology. Frontiers Neuroinform. 2018;11:76. Available from: https://dx.doi.org/10.3389/fninf.2017.00076.
[50] Patil P, Peng RD, Leek JT. A statistical definition for reproducibility and replicability; 2016. Available from: https://dx.doi.org/10.1101/066803.
[51] Poveda Villalón M, Espinoza Arias P, Garijo D, Corcho O. Coming to terms with FAIR ontologies. In: Proc. EKAW 2020. New York: ACM (ISBN 978-3-030-61243-6); 2020. p. 255-70. Available from: https://dx.doi.org/10.1007/978-3-030-61244-3_18.
[52] Garijo D, Corcho O, Poveda Villalón M. FOOPS!: An ontology pitfall scanner for the FAIR principles. In: Seneviratne O, et al., editors. Proc. ISWC 2021 Posters, Demos, Industry. CEUR-WS; 2021. Available from: http://ceur-ws.org/Vol-2980/paper321.pdf.
[53] Poveda Villalón M, Gómez Pérez A, Suárez Figueroa MC. OOPS! (Ontology pitfall scanner!): An on-line tool for ontology evaluation. Int J Semant Web Inform Sys. 2014;10:7-34.
[54] Görgen C, Sinn R. Mathematik in der Nationalen Forschungsdateninfrastruktur. Mitteil DMV. 2021;29(3):122-3. Available from: https://dx.doi.org/10.1515/dmvm-2021-0049.
[55] Boege T, Fritze R, Görgen C, Hanselman J, Iglezakis D, Kastner L, et al. Research-data management planning in the German mathematical community; 2022. arXiv:2211.12071 [math.HO]. Available from: https://dx.doi.org/10.48550/arXiv.2211.12071.
[56] Karmacharya A, Farnbacher B, Wiljes C, Iglezakis D, Terzijska D, Lanza G, et al. Metadata4Ing: An ontology for describing the generation of research data within a scientific activity. NFDI4Ing; 2022. Available from: https://w3id.org/nfdi4ing/metadata4ing/.
[57] Willighagen E. Two years of explicit CiTO annotations. J Cheminform. 2023;15:14. Available from: https://dx.doi.org/10.1186/s13321-023-00683-2.
[58] Guha R, Zdrazil B, Jeliazkova N, Martínez Mayorga K. A look back at a pilot of the citation typing ontology. J Cheminform. 2023;15:15. Available from: https://dx.doi.org/10.1186/s13321-023-00684-1.

[59] Horsch MT, Chiacchiera S, Seaton MA, Todorov IT, Šindelka K, Lísal M, et al. Ontologies for the Virtual Materials Marketplace. Künstl Intell. 2020;34(3):423-8. Available from: https://dx.doi.org/10.1007/s13218-020-00648-9.

[60] Schembera B, Iglezakis D. EngMeta: Metadata for computational engineering. Int J Metadat Semant Ontol. 2020;14(1):26-38. Available from: https://dx.doi.org/10.1504/ijmso.2020.107792.