European standardization efforts from FAIR toward explainable-AI-ready data documentation in materials modelling

Martin Thomas Horsch
Department of Data Science, Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway

Björn Schembera
Institute of Applied Analysis and Numerical Simulation, University of Stuttgart
Stuttgart, Germany

Heinz A. Preisig
Department of Chemical Engineering, Norwegian University of Science and Technology,
Trondheim, Norway

***Abstract*—Security critical AI applications require a standardized and interoperable data and metadata documentation that makes the source data explainable-AI ready (XAIR). Within the domain of materials modelling and characterization, European initiatives have proposed a series of metadata standards and procedural recommendations that were accepted as CEN workshop agreements (CWAs): CWA 17284 MODA, CWA 17815 CHADA, and CWA 17960 ModGra. It is discussed how these standards have been ontologized, and gaps are identified as regards the epistemic grounding metadata, *i.e.*, an annotation of data and claims by something that substantiates whether, why, and to what extent they are indeed knowledge and can be relied upon.**

***Index Terms*—Applied ontology, data and metadata quality, explainable AI, pragmatic interoperability, process data technology.**

I. Introduction

The FAIR principles [1]–[3] formulate a minimum standard in data documentation, but they are not enough to ensure that data and claims are ready for deployment in security-critical AI applications. In other words: Data can be FAIR without being *explainable-AI ready* (XAIR). The focus of much of the efforts at progressing from FAIR to XAIR data has so far been on the annotation of data with a provenance documentation that is ideally machine-intelligible or even machine-actionable.

A variety of provenance documentation standards have been advanced in materials modelling and beyond. AiiDA, in particular, provides provenance descriptions that can simultaneously serve as actionable workflow representations [4], [5]. The OMG's business process model and notation (BPMN) can serve both as a specification/documentation of workflows and as an input to process automation platforms [6]; it is not particularly tailored to molecular methods, but has been put into practice in materials modelling for the COMPOSELECTOR business decision support system [7], [8]. The ontologization of BPMN is an active field of development with multiple specialized solutions already in existence [9]–[13]. European initiatives, including Horizon projects related to the European Materials Modelling Council (EMMC) and the European Materials Characterization Council (EMCC), have been working toward a standardized data and metadata documentation, resulting in three successive CEN workshop agreement (CWA) processes: First, MODA [14], for a human-readable (but not machine-intelligible) documentation of modelling and simulation workflows; second, CHADA [15], transferring the approach from MODA to materials characterization; third, ModGra [16], where modelling and simulation workflows are provided with a potentially machine-actionable graphical documentation in terms of process topologies.

However, such efforts are bound to remain constrained in terms of digitalization breadth and depth, which they do not always make explicit. There is only so much that *can* be made machine-actionable with a reasonable effort, whereas the text of a journal paper is sufficient to document the provenance intelligibly to a human. As Vogt analyses it, there is a "conflict between machine-actionability and human-actionability" [17]. Humans tend to rely on human language with all its gaps and ambiguities for a reason: They have mastered pragmatics and do not need to spell out every detail, making communication much more efficient. Databases, triple stores, and their interfaces have not yet mastered pragmatics (even though this now seems within reach). "Therefore, when human-readable statements are being translated into machine-actionable and easy to query representations that follow, for example, the RDF triple syntax paradigm of *Subject-Predicate-Object*, these graph-based representations are often much more complex than human readers like them to be" [17]. The *cognititive interoperability* of knowledge graphs remains limited [17].

This paper is structured as follows: In Section II, it summarizes the CWAs from the EMMC community (MODA, CHADA, ModGra). It discusses the shortcomings that will become limiting factors to the uptake of CHADA and MODA if it is ever seriously tried to disseminate them beyond the very core of the community they originate from. In parallel, the same community has also been endorsing Peircean semiotics

E-mail: martin.thomas.horsch@nmbu.no, bjoern.schembera@mathematik.uni-stuttgart.de, heinz.preisig@chemeng.ntnu.no. *Acknowledgment:* B.S. acknowledges DFG project no. 460135501 (MaRDI). H.A.P. acknowledges H2020 grant agreements no. 952903 (VIPCOAT) and NFR project no. 299363, "Nano-selective, bio-mimetic membranes for integrated biotechnological production of biofuels from lignocellulosic biomass."

as a way to analyse research processes. Despite this, no viable bridge has been built between the semiotics-based ontologies and the CHADA/MODA tables – they seem to counteract each other, rather than informing each other. Section III summarizes the semiotics-inspired understanding of scientific practice as a collective cognitive process. Section IV outlines how it is the latter approach, and not CHADA/MODA, that has the potential to be developed into a framework for making materials modelling data XAIR. The semiotics-based view is more general and can be adapted to diverse processes and practices, including new developments – by construction, its epistemic space is open. This opposes it squarely to MODA's closed epistemic space[1] where all of materials modelling is once and for all categorized into a small, enumerated set of predefined physical equations. Moreover, the Peircean approach resolves complex communication and reasoning processes into a simple process model and notation in terms of triadic steps (sign – object – interpretant); this directly helps formulate *explanations* of how the eventual outcome has come about and why it should be accepted as knowledge (epistemic grounding). Finally, a conclusion is given in Section V.

II. CEN STANDARDS FOR WORKFLOW DOCUMENTATION

A. CWA 17284:2018 E: Model data (MODA)

MODA first and foremost is a set of tables or forms that a user is supposed to fill in for documenting four kinds of elements, also called sections: Use cases, models, solvers, and any other processing steps. The standardization process for MODA was based on a previous community surveying exercise, the *Review of Materials Modelling* (RoMM) [19]. RoMM and MODA successfully accomplished something very important: They defined basic (fundamental) entities and provided a shared conceptualization of materials modelling that has since been in use within the EMMC community, across research groups that had not previously shared a common terminology and perspective. As a more problematic outcome of this process, however, MODA is by construction narrowly limited to a static set of physical equations [19]; theoretical or methodological work that goes beyond these pre-established categories either needs to be artificially mapped to the one that seems closest, or it cannot be accommodated at all. In addition to the tables, work toward providing a graphical notation for provenance in molecular and multiscale modelling has led to MODA graphs, which are part of the CWA standard. The MODA graphs are related to the MODA tables in that some of the sections can show up as nodes [14]; however, this is not done consistently, and data/information entities often occur as nodes instead. Where that is the case, it is hard to see how the MODA tables and graphs relate to each other conceptually at all. RoMM includes a corpus of MODA based simulation workflow documentations [19] from projects which were by EC policy required to supply such material; overall, this material mainly shows that the notational elements (arrows, nodes, *etc.*) are highly ambiguous and that there even are contradictory opinions among users as regards their purpose and meaning. Addressing this problem, a less ambiguous graph notation was developed, yielding logical data transfer (LDT) graphs [20], while the MODA tables and the LDT graphs were developed into the ontology for simulation, modelling, and optimization (OSMO) [20], [21]. However, the work on LDT and OSMO is not part of the CWA and has seen little uptake.

B. CWA 17815:2021 E: Characterization data (CHADA)

CHADA follows the same approach as MODA, and has the same structure: Use cases, experiments, raw data, and data processing steps are documented in tables; these sections can then be arranged into graphs [15]. Same as for MODA, the result is a human-intelligible, but often highly ambiguous representation that will usually tell the reader less than looking into a paper that describes how the experiments were conducted. Where OSMO is the ontology version of MODA, similarly, CHAMEO is the ontology version of CHADA [22]. CHAMEO is not part of the CWA and has seen little uptake.

C. Can there be a future for CHADA and MODA?

These two CWAs have never been in use outside the core community consisting of projects where they were made mandatory by the EC and its responsible project officers. Within that core community, the assessment of this approach has become more and more sceptical. Adamovic *et al.* [23] in a 2021 report summarized that the "topics that should be addressed by both EMMC and EMCC" first and foremost included: "Make MODA and CHADA interoperable. Make MODA and CHADA machine-readable." From this it can only be concluded that they were seen as neither interoperable nor machine-readable. Clark *et al.* [24] express this concern in a more clear and intelligible way: "Although these EMMC initiatives standardizing the metadata have been successful in improving communication and interoperability among humans, they have some shortcomings regarding machine-readable semantic interoperability. Descriptors in MODA are plain-text labels and [. . .] connections between steps in a modelling workflow are shown using blue arrows, which are ambiguous [. . .]. As in the case of the MODA templates, the lack of semantic descriptors in the CHADA template limits its direct use in the creation of Linked Data." Most recently, a memorandum on knowledge management for Industry Commons promotes FAIR data and ontology-based knowledge representation as crucial elements to realizing the "huge untapped value in data, which currently is lost or insufficiently valorised" [25]; it also gives a detailed discussion of the "translation" theme as it has been interpreted and developed by the EMMC. However, in this document, there are zero mentions of CHADA, MODA, or RoMM [25]. All indicates that CHADA and MODA are in the process of being superseded by the more recent ontology development based on Peircean semiotics (*cf.* Section III).

It will be unavoidable to go beyond CHADA/MODA. Their limitations include an idiosyncratic arrangement of data items that makes it hard to align them with other semantic artefacts [26], including the EMMO foundational ontology [24],

[1]See also the discussion of *epistemic spaces* by Tulatz [18].

[27]. The pro forma alignment of CHADA/MODA with the EMMO through the CHAMEO and OSMO ontologies only highlights how different the two approaches are. However, the most salient issue surrounding CHADA/MODA is that researchers are expected to go great lengths to document their research outcomes' provenance in detail,[2] but despite that, the documentation does not become meaningfully machine-actionable. It is in principle possible to imagine that, using AI-based text processing, at some point in the future, an automated simulation or lab environment could process a CHADA or MODA documentation and set up a similar simulation or experimental workflow for replication. However, just like humans, such AI-driven text processing could also directly work with a scientific paper or technical report.

CHADA and MODA do not really target and therefore also do not achieve what Vogt calls cognitive interoperability, *i.e.*, usefulness to humans "not only in terms of how humans prefer to interact with technology (human-computer interaction) but also in terms of how they interact with information (human-information interaction)" [17] – everything about these tables and diagrams is unnecessarily complicated, particularly of course for the creator, but also for the reader.[2] However, they do not really target and therefore do not achieve machine-actionability either, resulting in a lose-lose answer to the conflict between machine- and human-actionability.

D. CWA 17960:2022 E: Model graphs (ModGra)

Against this background, the novel European CWA standard ModGra [16] promises to deliver the required paradigm shift, replacing MODA with a more flexible and meaningful way of denoting how physical quantities at multiple levels relate to each other, and how one simulation step can feed into another. ModGra is the community-agreed version of a notation developed over years, with a variety of use cases in computer aided process engineering [29]–[32]. The core concept in ModGra is that of the *process model topology*, defined in analogy with Petri nets as "a directed graph, a network of nodes connected by directed arcs. The nodes are token capacities, and the edges are arcs transporting tokens" [16]. Different sorts of nodes can contain different sorts of tokens: In the case of a control capacity, representing the logics of the simulation workflow, the tokens are information items, whereas the tokens contained by physical capacities are extensive physical quantities, representing the logics of the model of a physical system. The examples given in the ModGra document [16] also show that these diagrams can become very complicated, requiring a major effort from the side of the researcher and even the reader – a challenge that ModGra shares with MODA. Naturally, when attempting to completely describe materials modelling scenarios in detail, this challenge cannot be avoided.

III. Cognitive process models

A. Semiotics

Following Peirce's approach to semiotics, an elementary cognitive step is a process which is conceptualized as a *triad* and starts from the *representation relation* between a sign s (*i.e.*, a representamen) and an object o (*i.e.*, the referent of the representamen); the cognitive step adds a third element to the sign and the object, by which a new representation relation is created [33]. Peirce's conceptualization can thus be summarized as that of a sign in action [34]. In the case of a semiosis σ, the third element is the interpretant s', a newly created sign. As such the triadic cognition can be denoted by

$$\sigma : \quad s — o — s' \ [\kappa]; \tag{1}$$

for instance, σ can be a simulation of a system o, where s is the relevant part of the simulation input (namely, the model of o), and s' is the relevant part of the simulation output (*e.g.*, a property of o), such that o is the common referent of both s and s'. Above, κ is the preceding cognitive step from which the representation relation between s and o is carried over.

Symmetrically complementing this, in a Metonymization μ the sign is assigned a new referent o'

$$\mu : \quad o — s — o' \ [\kappa], \tag{2}$$

such that a subsequent cognitive step can build on the representation relation between s and o'. For instance, the sign s could be a molecular model of methane: On the one hand, it represents the interaction between pairs of molecules o. On the other hand, we practically use it as a model for a macroscopic amount of methane o' and its phenomenological thermodynamic properties, when used in a simulation to compute such properties. From one to the other, a metonymization occurs; this often takes the form of a synecdoche, going over from a part to the whole, or vice versa (*cf.* Fig. 1) [35], [36].

The above has been implemented in the *physicalistic interpretation of modelling and simulation interoperability infrastructure* (PIMS-II) mid-level ontology [37], [38].

B. Mereosemiotics: Classification of cognitive steps

Mereosemiotics, the combination of cognitive process models and the ontology of spatiotemporal parthood and contact [37], [39], is implemented in PIMS-II through a three-way case distinction between branches of a cognitive step taxonomy: Perception, interpretation, metonymization, *cf.* Fig. 1.

[2]See *e.g.* the diagrams from ReaxPro [28], particularly the one on *p.* 17.

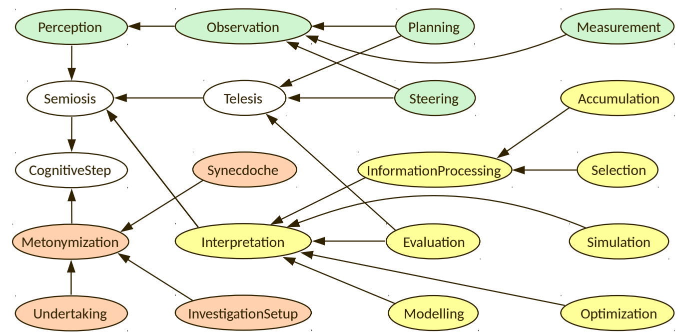

Fig. 1. Fragment of the cognitive step taxonomy from PIMS-II [38].

In a Perception, a subclass of Semiosis, all the three elements of the triad, s (sign), o (object), and s' (interpretant), need to participate physically, which is characterized by spatiotemporal overlap of each element with the cognitive step.

In an Interpretation, another subclass of Semiosis, s and s' must be present, while this is not required for o. Example: We can simulate an explosion even if that explosion does not actually occur there and then (or ever). The object, the explosion, does not need to be present. This distinguishes the interpretation (simulation) from a perception of the explosion.

In a Metonymization, the sign s needs to participate physically, whereas the old referent o and the new referent o' need not. Example: An equation of state s that describes a control volume o at equilibrium can be metonymized to apply to a reactor o' which is not at equilibrium. This is a simple modelling decision. Neither the reactor nor the control volume need to be physically present for us to make that decision.

C. Documenting and annotating the research outcomes

Eventually, outcomes of the research workflow are obtained in the role of a Peircean representamen; research outcomes can be data items or datasets, but in particular, they include the interpretation that the data have been given, *i.e.*, the *knowledge claims* [38], [40] formulated by the researcher(s); these are the elements that require the most careful annotation.

PIMS-II based documentations can be transformed to EMMO based documentations, since the EMMO is similarly designed around mereotopology and semiotics [24], [27], [39], [41]. Moreover, PIMS-II is closely co-designed with the Metadata4Ing ontology [42] from NFDI4Ing, the German NFDI's disciplinary research data infrastructure consortium for the engineering sciences [43], [44]. As a consequence, when documenting a *property claim*, *i.e.*, a knowledge claim including a statement to the effect that some property has some value, the corresponding PIMS-II and Metadata4Ing based annotations are aligned 1:1 by construction [45]. It is recommended to take units and quantity kinds from QUDT [46], and to use canonical TUCAN identifiers for molecules [47].

Example: Property change attributed to an object

Consider this example scenario from Borgo *et al.* [48]:[3]

> *"at first, a component surface is red. As time passes, its colour changes. A few months later it has become brown."*

We define the individuals

DataItem(δ_1), DataItem(δ_2), Item(o), Item(ρ_1), Item(ρ_2), Observation(π_1), Observation(π_2), Property(p), Value(ℓ_1), Value(ℓ_2),

where o is the component surface, p is the property *colour*, $\pi_{1,2}$ are the observation processes, and $\rho_{1,2}$ are spacetime regions restricting the time intervals in which these observations occur (*e.g.*, first in summer, then in autumn [48]). The observation outcomes are given by $\delta_{1,2}$.

Accordingly, the colour is observed (*e.g.*, measured) twice,

isTargetPropertyIn(p,π_1), isObservedIn(o,π_1), isObservationOutcomeIn(δ_1,π_1), isTargetPropertyIn(p,π_2), isObservedIn(o,π_2), isObservationOutcomeIn(δ_2,π_2), temporallyPrecedes(ρ_1, ρ_2), isTemporallyIncludedIn(π_1, ρ_1), isTemporallyIncludedIn(π_2, ρ_2).

First it is found to be red (colour value ℓ_1), and then it is found to be brown (colour value ℓ_2),

isAssignmentFor(δ_1, o), isAssignmentFor(δ_2, o), isValueInAssignment(ℓ_1, δ_1), isValueInAssignment(ℓ_2, δ_2), isVariableInAssignment(p,δ_1), isVariableInAssignment(p,δ_2), isAdmissibleValueFor(ℓ_1, p), isAdmissibleValueFor(ℓ_2, p).

IV. Toward explainable AI-readiness

Taking the progress that was already made as a background, as summarized above, improvements are needed specifically at communicating reliability and motivating an appropriate level of trust in research data. The annotation directly addressing for these communication requirements can be referred to as *epistemic grounding metadata*. Supplying sufficient epistemic grounding metadata in a standardized and interoperable way would support making the research data explainable-AI ready. There, in particular, it would be beneficial to provide knowledge claims (jointly with any associated data) with an annotation that is meant to motivate its acceptance by others.

The employed rationales vary between academic communities. Two categories of epistemic grounding can be distinguished [38]; *type-1 grounding*, which is part of the outcome at large (*e.g.*, a mathematical proof that motivates acceptance of a new theorem), and *type-2 grounding*, where the provenance documentation itself motivates why the outcome should be accepted [49]. Often, as Newman [50] argues, acceptance of scientific evidence *as knowledge* is obtained on the basis of an insight into all the individual internal processing steps. But it can just as well be that a detailed step-by-step analysis fails to substantiate trust in the outcome. In molecular modelling, due to the "epistemic schism" [51] of phenomenological and statistical methods, this is the rule rather than the exception, so that "transparency falls short of offering the right epistemic reasons for trusting the outcome of the algorithms" [52]. Here, process reliabilism[4] [49], [52], [53] comes into play as an option: The outcome of a cognitive process becomes acceptable as knowledge by instantiation of the schema [53]

> "if S's believing p at t results from m, then S's belief in p at t is justified,"

where the provenance documentation is sufficient to motivate acceptance of m as "a reliable process" [53]. Both the step-by-step verification-based approach and computational process reliabilism are type-2 grounding methods [38]: They are directly based on the provenance documentation.

Not all results in molecular modelling are best grounded epistemically by a detailed discussion of their procedural

[3] Instead of a component surface, Borgo *et al.* [48] mention a "flower," which unfortunately is too romantic and not technical enough for our purpose.

[4] As formulated by Williams: "Epistemic grounding is a matter of *reliability*. A belief is epistemically grounded [...] iff it is formed via a process that in fact makes it likely to be true" [49].

provenance. Sometimes we would be outright mistaken to discuss all the details, namely, where the provenance is completely immaterial to the research outcome (*e.g.*, "we held weekly discussion meetings"). Researchers often decide not to expand on the provenance of some of their scientific results in detail; this is generally the case for theoretical results or for the strategy and design choices underlying the methodology itself. A purely theoretical result, *e.g.*, in statistical mechanics, is mathematical in nature and requires a validity documentation, usually in the form of a mathematical proof-based validity claim; but it does not matter at all how that proof was obtained. Similarly, for new methods or new solvers implementing a method, it may but need not be documented in detail how they were developed; what matters is to what extent their validity can be documented. Such grounding mechanisms, which only have in common that the logical steps by which their reasoning proceeds do not coincide with a description of the research process, are referred to as type-1 grounding [38].

Resch and Kaminski propose that whenever "a method is partially opaque, this precludes internal justification [...] such as evidence and arguments [...]. This means that justifications must be external" [54]; they consider type-2 grounding by reliabilism as a way around this [54]. Symons and Alvarado [55] argue, however, that type-1 grounding is ultimately foundational even to the very rules that facilitate type-2 grounding: "We should not, and (for the most part) do not treat simulations as oracles. [...] warranted trust in the results of simulations is grounded in much more complex practices than simply deferring to a record of predictive success." For a process-reliabilism schema to apply to materials modelling methods, their derivation must be sound mathematically: By disciplinary standards, it is only jointly with this, not alone, that holistic validation [56], [57] might establish using its implementation in a simulation code as characteristic of a reliable process.

V. Conclusion

Epistemically opaque claims [53] can become XAIR (and no longer opaque) as soon as the "epistemically relevant elements" [58] underlying the claims become FAIR. The epistemic metadata, *i.e.*, the annotation required for this purpose, can be provided using the PIMS-II mid-level ontology [59]. As a conclusion from the above, *we propose to develop epistemic FAIRness – i.e.*, the degree to which an annotation is provided that permits a reusing entity to form an adequate level of trust in the communicated claims – *as a new metadata quality metric*. Work on compliance testing [60] suggests that an automated assessment of epistemic FAIRness is feasible.

Unfortunately, the CHADA/MODA approach contributes nothing to this. When it comes to making knowledge machine-actionable, CHADA/MODA is useless, since only humans can understand it. For humans, on the other hand, it is cumbersome to use, and it only provides a provenance documentation that is redundant with the content of papers or reports where the same information is already conveyed to the same human readers in a much more accessible way. It will be a matter of future work as well as disciplinary decision making, self-government, and everyday practice to make materials modelling data XAIR, fighting back on the phenomenon of dark data [61]–[63]. Cognitive process models based on semiotics and process model topologies following CWA ModGra are steps in this direction. To the extent that these efforts succeed, legal requirements for documentation of security-critical AI systems can be better addressed, and a barrier to digitalization in process engineering and materials research will be moved out of the way.

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Sci. Data*, vol. 3, p. 160018, 2016, doi:10.1038/sdata.2016.18.

[2] J. Bicarregui, "Building and sustaining data infrastructures: Putting policy into practice," Wellcome Trust, Tech. Rep., 2016, doi:10.6084/m9.figshare.4055538.v2.

[3] I. Anders, C. Blanchi, D. Broeder, M. Hellström *et al.*, "FAIR digital object specifications," FDO Forum, Proposed recommendation, 2022.

[4] A. Merkys, N. Mounet, A. Cepellotti *et al.*, "A posteriori metadata from automated provenance tracking: Integration of AiiDA and TCOD," *J. Cheminform.*, vol. 9, p. 56, 2017, doi:10.1186/s13321-017-0242-y.

[5] S. P. Huber, "Automated reproducible workflows and data provenance with AiiDA," *Nat. Rev. Phys.*, vol. 4, p. 431, 2022, doi:10.1038/s42254-022-00463-1.

[6] B. Ruecker, *Practical Process Automation*. Sebastopol, California: O'Reilly (ISBN 978-1-4920-6145-8), 2021.

[7] C. Kavka, D. Campagna, M. Milleri, A. Segatto *et al.*, "Business decisions modelling in a multi-scale composite material selection framework," in *Proc. ISSE*. IEEE (ISBN 978-1-5386-4446-1), 2018, doi:10.1109/SysEng.2018.8544386.

[8] C. Kavka, D. Campagna, and H. Koelman, "A business decision support system supporting early stage composites part design," in *Advances in Computational Methods and Technologies in Aeronautics and Industry*, ser. Computational Methods in Applied Sciences, D. Knoerzer, J. Periaux, and T. Tuovinen, Eds. Cham: Springer (ISBN 978-3-031-12018-3), 2022, no. 57, pp. 263–279, doi:10.1007/978-3-031-12019-0_19.

[9] M. Rospocher, C. Ghidini, and L. Serafini, "An ontology for the business process modelling notation," in *Proc. FOIS 2014*, P. Garbacz and O. Kutz, Eds. IOS (ISBN 978-1-61499-437-4), 2014, pp. 133–146.

[10] L. Riehl Figueiredo and H. Carvalho de Oliveira, "Automatic generation of ontologies from business process models," in *Proc. ICEIS*. SciTePress (ISBN 978-989-758-298-1), pp. 81–91, doi:10.5220/0006709100810091.

[11] A. Annane, N. Aussenac-Gilles, and M. Kamel, "Une ontologie des processus métier (BBO) pour guider un agent virtuel," in *Proc. IC 2019*. AFIA, 2019, pp. 183–198.

[12] M. Kchaou, W. Khlif, F. Gargouri, and M. Mahfoudh, "Transformation of BPMN model into an OWL2 ontology," in *Proc. ENASE 2021*. SciTePress (ISBN 978-989-758-508-1), 2021, pp. 380–388, doi:10.5220/0010479603800388.

[13] B. Di Martino, M. Graziano, L. Colucci Cante, A. Esposito, and M. Epifania, "Application of business process semantic annotation techniques to perform pattern recognition activities applied to the generalized civic access," in *Proc. CISIS-2022*. Springer (ISBN 978-3-031-08811-7), 2022, pp. 404–413, doi:10.1007/978-3-031-08812-4_39.

[14] CEN-CENELEC Management Centre, "Materials modelling: Terminology, classification and metadata," CEN, Brussels, CEN workshop agreement 17284:2018 E, 2018.

[15] ——, "Materials characterization: Terminology, metadata and classification," CEN, Brussels, CEN workshop agreement 17815:2021 E, 2021.

[16] ——, "ModGra: A graphical representation of physical process models," CEN, Brussels, CEN workshop agreement 17960:2022 E, 2022.

[17] L. Vogt, "Extending FAIR to FAIREr: Cognitive interoperability and the human explorability of data and metadata," Tech. Rep. arXiv:2301.04202 [cs.DB], 2023, doi:10.48550/arXiv.2301.04202.

[18] K. Tulatz, *Epistemologie als Reflexion wissenschaftlicher Praxen*. transcript (ISBN 978-3-8376-4212-4), 2018, doi:10.1515/9783839442128.

[19] A. F. de Baas, Ed., *What Makes a Material Function?* Luxembourg: EU Publications (ISBN 978-92-79-63185-6), 2017, doi:10.2777/417118.

[20] M. T. Horsch, C. Niethammer, G. Boccardo, P. Carbone *et al.*, "Semantic interoperability and characterization of data provenance in computational molecular engineering," *J. Chem. Eng. Data*, vol. 65, no. 3, pp. 1313–1329, 2020, doi:10.1021/acs.jced.9b00739.

[21] M. T. Horsch, D. Toti, S. Chiacchiera, M. A. Seaton *et al.*, "OSMO: Ontology for simulation, modelling, and optimization," in *Proc. JOWO 2021*, ser. CEUR Worksh. Proc., no. 2969. CEUR-WS, 2021.

[22] P. Del Nostro, G. Goldbeck, and D. Toti, "CHAMEO: An ontology for the harmonisation of materials characterisation methodologies," *Appl. Ontol.*, vol. 17, no. 3, pp. 401–421, 2022, doi:10.3233/ao-220271.

[23] N. Adamovic, B. Boskovic, M. Celuch, C. Charitidis *et al.*, "Report on advanced materials modelling and characterisation: Strategies for integration and interoperability," EMMC and EMCC, Tech. Rep., 2021, doi:10.5281/zenodo.4912683.

[24] S. Clark, F. L. Bleken, S. Stier, E. Flores *et al.*, "Toward a unified description of battery data," *Adv. Energy Mater.*, vol. 12, no. 17, p. 2102702, 2022, doi:10.1002/aenm.202102702.

[25] G. Goldbeck, A. Simperler, L. Bull, D. Gao *et al.*, "The translator in knowledge management for innovation: Towards Industry Commons," OntoCommons, Tech. Rep., 2022, doi:10.5281/zenodo.7041696.

[26] P. Klein, H. A. Preisig, M. T. Horsch, and N. Konchakova, "Application of an ontology based process model construction tool for active protective coatings: Corrosion inhibitor release," in *Proc. JOWO 2021*, ser. CEUR Worksh. Proc., no. 2969. CEUR-WS, 2021.

[27] J. Francisco Morgado, E. Ghedini, G. Goldbeck, A. Hashibon *et al.*, "Mechanical testing ontology for digital-twins: A roadmap based on EMMO," in *Proc. SeDiT 2020*, ser. CEUR Worksh. Proc., no. 2615. CEUR-WS, 2020.

[28] J. Francisco Morgado, A. Hashibon, M. Hus, A. Bjelić *et al.*, "ReaxPro MODA diagrams," ReaxPro, project deliverable D2.1, 2020.

[29] H. A. Preisig, "ProMo – a multi-disciplinary process modelling suite," in *Proc. ESCAPE30*, ser. Comp. Aid. Chem. Eng., no. 48. Elsevier, 2020, pp. 571–576, doi:10.1016/b978-0-12-823377-1.50096-3.

[30] ——, "Ontology-based process modelling – with examples of physical topologies," *Processes*, vol. 9, no. 4, p. 592, 2021, doi:10.3390/pr9040592.

[31] ——, "Modelling – mechanistic reductionism extended with holistic empirical components," in *Proc. ESCAPE31*, ser. Comp. Aid. Chem. Eng., no. 50. Elsevier, 2021, pp. 689–694, doi:10.1016/b978-0-323-88506-5.50109-1.

[32] ——, "Documenting models comprehensively using a minimal graphical language," in *Proc. PSE 2021+*, ser. Comp. Aid. Chem. Eng., vol. 49. Elsevier, 2022, p. 1021, doi:10.1016/b978-0-323-85159-6.50170-6.

[33] C. S. Peirce, *Peirce on Signs*, J. Hoopes, Ed. Chapel Hill: Univ. North Carolina Press (ISBN 978-0-80784342-0), 1991.

[34] P. Atã and J. Queiroz, "O externalismo semiótico ativo de C. S. Peirce e a cantoria de viola como signo em ação," *Trans/Form/Ação*, vol. 44, no. 3, pp. 177–204, 2021, doi:10.1590/0101-3173.2021.v44n3.15.p177.

[35] R. Benczes, A. Barcelona, and F. J. Ruiz de Mendoza Ibáñez, Eds., *Defining Metonymy in Cognitive Linguistics: Towards a Consensus View*. Amsterdam: John Benjamins (ISBN 978-9-02728676-5), 2011.

[36] C. Denroche, *Metonymy and Language: A New Theory of Linguistic Processing*. Abingdon: Routledge (ISBN 978-1-138-81062-4), 2015.

[37] M. T. Horsch, "Mereosemiotics: Parts and signs," in *Proc. JOWO 2021*, ser. CEUR Worksh. Proc., no. 2969. CEUR-WS, 2021.

[38] M. T. Horsch and B. Schembera, "Documentation of epistemic metadata by a mid-level ontology of cognitive processes," in *Proc. JOWO 2022*, ser. CEUR Worksh. Proc., no. 3249. CEUR-WS, 2022.

[39] M. T. Horsch, S. Chiacchiera, B. Schembera, M. A. Seaton, and I. T. Todorov, "Semantic interoperability based on the European Materials and Modelling Ontology and its ontological paradigm: Mereosemiotics," in *Proc. WCCM-ECCOMAS 2020*. Scipedia, 2021, doi:10.23967/wccm-eccomas.2020.297.

[40] M. T. Horsch and B. Schembera, "Epistemic metadata in molecular modelling: First-stage case-study report (10 cases)," Inprodat e.V., Kaiserslautern, Tech. Rep. 2023–A, 2023, doi:10.5281/zenodo.7516532.

[41] H. A. Preisig, P. Klein, N. Konchakova, T. F. Hagelien *et al.*, "Topology-based construction of business-integrated material modelling workflows," in *Proc. ESCAPE32*, ser. Comp. Aid. Chem. Eng., no. 51. Elsevier, 2022, p. 1237, doi:10.1016/b978-0-323-95879-0.50207-1.

[42] A. Karmacharya, B. Farnbacher, C. Wiljes, D. Iglezakis *et al.*, "Metadata4Ing: An ontology for describing the generation of research data within a scientific activity," NFDI4Ing, documentation, 2022.

[43] R. H. Schmitt, V. Anthofer, S. Auer, S. Başkaya *et al.*, "NFDI4Ing: The national research data infrastructure for engineering sciences," NFDI4Ing, Tech. Rep., 2020, doi:10.5281/zenodo.4015201.

[44] T. Bronger, É. Demandt, I. Heine, A. Kraft *et al.*, "Die Nationale Forschungsdateninfrastruktur für die Ingenieurwissenschaften (NFDI4Ing)," *Baust. FDM*, vol. 2021, no. 2, pp. 110–123, 2021, doi:10.17192/bfdm.2021.2.8329.

[45] M. T. Horsch, S. Chiacchiera, M. Kohns, E. A. Müller *et al.*, "Epistemic metadata in molecular modelling: Second-stage case-study report (12 claims)," Inprodat e.V., Kaiserslautern, Tech. Rep. 2023–B, 2023, doi:10.5281/zenodo.7608074.

[46] X. Zhang, K. Li, C. Zhao, and D. Pan, "A survey on units ontologies: Architecture, comparison and reuse," *Progr. ELIS*, vol. 51, no. 2, pp. 193–213, 2017, doi:10.1108/prog-08-2015-0056.

[47] J. C. Brammer, G. Blanke, C. Kellner, A. Hoffmann *et al.*, "TUCAN: A molecular identifier and descriptor applicable to the whole periodic table from hydrogen to oganesson," *J. Cheminform.*, vol. 14, p. 66, 2022, doi:10.1186/s13321-022-00640-5.

[48] S. Borgo, A. Galton, and O. Kutz, "Foundational ontologies in action: Understanding foundational ontology through examples," *Appl. Ontol.*, vol. 17, no. 1, pp. 1–16, 2022, doi:10.3233/ao-220265.

[49] M. Williams, "Dretske on epistemic entitlement," *Philos. Phenomenol. Res.*, vol. 60, no. 3, pp. 607–612, 2000, doi:10.2307/2653818.

[50] J. Newman, "Epistemic opacity, confirmation holism and technical debt: Computer simulation in the light of empirical software engineering," in *Proc. HaPoC 2015*. Springer (ISBN 978-3-319-47285-0), 2016, pp. 256–272, doi:10.1007/978-3-319-47286-7_18.

[51] J. Anta, "The epistemic schism of statistical mechanics," *Theoria*, vol. 36, no. 3, pp. 399–419, 2021, doi:10.1387/theoria.22134.

[52] J. M. Durán and K. R. Jongsma, "Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical AI," *J. Med. Ethics*, vol. 47, no. 5, pp. 329–335, 2021, doi:10.1136/medethics-2020-106820.

[53] J. M. Durán and N. Formanek, "Grounds for trust: Essential epistemic opacity and computational reliabilism," *Minds Machin.*, vol. 28, no. 4, pp. 645–666, 2018, doi:10.1007/s11023-018-9481-6.

[54] M. Resch and A. Kaminski, "The epistemic importance of technology in computer simulation and machine learning," *Minds Machin.*, vol. 29, no. 1, pp. 9–17, 2019, doi:10.1007/s11023-019-09496-5.

[55] J. Symons and R. Alvarado, "Epistemic entitlements and the practice of computer simulation," *Minds Machin.*, vol. 29, pp. 37–60, 2019, doi:10.1007/s11023-018-9487-0.

[56] J. Lenhard, "Holism, or the erosion of modularity: A methodological challenge for validation," *Philos. Sci.*, vol. 85, no. 5, pp. 832–844, 2018, doi:10.1086/699675.

[57] ——, "How does holism challenge the validation of computer simulation?" in *Computer Simulation Validation*, C. Beisbard and N. J. Saam, Eds. Cham: Springer (ISBN 978-3-319-70766-2), 2019, pp. 943–960, doi:10.1007/978-3-319-70766-2_39.

[58] P. Humphreys, "Computational science and its effects," in *Science in the Context of Application*, M. Carrier and A. Nordmann, Eds. Heidelberg: Springer (ISBN 978-90-481-9051-5), 2011, pp. 131–142, doi:10.1007/978-90-481-9051-5_9.

[59] M. T. Horsch, S. Chiacchiera, M. Kohns, E. A. Müller *et al.*, "Epistemic metadata for computational engineering information systems," Manuscript, to appear in *Proc. FOIS 2023*, 2023.

[60] M. D. Wilkinson, M. Dumontier, S.-A. Sansone, L. O. Bonino da Silva Santis *et al.*, "Evaluating FAIR maturity through a scalable, automated, community-governed framework," *Sci. Data*, vol. 9, p. 174, 2019, doi:10.1038/s41597-019-0184-5.

[61] P. B. Heidorn, "Shedding light on the dark data in the long tail of science," *Libr. Trends*, vol. 57, no. 2, pp. 280–299, 2008, doi:10.1353/lib.0.0036.

[62] B. Schembera, "Like a rainbow in the dark: Metadata annotation for HPC applications in the age of dark data," *J. Supercomput.*, vol. 77, pp. 8946–8966, 2021, doi:10.1007/s11227-020-03602-6.

[63] A. Corallo, A. M. Crespino, V. Del Vecchio, M. Lazoi, and M. Marra, "Understanding and defining dark data for the manufacturing industry," *IEEE Transactions on Engineering Management*, vol. 70, no. 2, pp. 700–712, 2022, doi:10.1109/tem.2021.3051981.