Domain-specific metadata standardization in materials modelling

Martin Thomas Horsch[*]
High Performance Computing Center Stuttgart, Stuttgart, Germany,
UK Research and Innovation, STFC Daresbury Laboratory, Daresbury, United Kingdom

Joana Francisco Morgado
Fraunhofer Institute for Mechanics of Materials, Freiburg, Germany

Gerhard Goldbeck
Goldbeck Consulting Ltd, Cambridge, United Kingdom

Dorothea Iglezakis
University of Stuttgart, University Library, Stuttgart, Germany

Natalia A. Konchakova
Helmholtz-Zentrum Hereon, Institute of Surface Science, Geesthacht, Germany

Björn Schembera
High Performance Computing Center Stuttgart, Stuttgart, Germany

Abstract

Domain-specific metadata standards, including ontologies, markup languages, and technical interface specifications, are a necessary component of solutions for FAIR research data management with industrial applications. The Workshop on Domain Ontologies for Research Data Management in Industry Commons of Materials and Manufacturing (DORIC-MM 2021) discusses the state of the art, challenges, and perspectives for continuing innovation in this field. The present work comments on the landscape of semantic assets in the field of materials modelling, covering electronic, atomistic, mesoscopic, and continuum methods. Summaries are given of particularly promising lines of work, including the CAPE-OPEN interface standard, the XML schemas EngMeta, CML, and ThermoML, and the ontologies OntoCAPE, Metadata4Ing/Metadata4HPC, OSMO (the ontology version of MODA) and the VIMMP system of ontologies, and the domain-level modules of the European Materials and Modelling Ontology (EMMO). For future work, it is recommended to emphasize advancing in accordance with five principles: 1. Diversification of technologies; 2. Observation of practices; 3. Realistic objectives; 4. Incentives for providing citable data and software; 5. Co-design of simulation and data technology.

[*]Corresponding author: M. T. Horsch (martin.horsch@hlrs.de).

1 Introduction

Metadata standardization can be implemented in a wide variety of ways. It is therefore unsurprising that in materials modelling, similar to other fields, many different approaches have been applied to support *findability, accessibility, interoperability, and reusability*, *i.e.*, the FAIR principles of data management [1, 2]. What these approaches have in common is that they are applications of semantic technology in that they need to go beyond expressing formal, syntactic requirements on input/output conventions and formats (*e.g.*, "a configuration input for code X consists of an integer number N followed by $6N$ floating-point values") by giving an indication on the meaning of the communicated data and metadata; by annotating data (*i.e.*, by providing metadata), data become information, and in a semantic-web based approach, ontologies are used to associate data and metadata with an agreed meaning.

Whenever a collection of codes or platforms interact systematically or on a regular basis, interoperability is required. This implies semantic interoperability, *i.e.*, agreement on the meaning of the exchanged information, since the output of one workflow element needs to be understood correctly when it acts as input for the next element. In this sense, any thoroughly documented serialization, graphical notation, or other syntactic standard can act as a metadata standard; substantial efforts have been dedicated to this sort of documentation by which guidelines on the structure, content, and use of databases [3, 4], interfaces [5, 6], or workflow management systems [7, 8, 9, 10, 11, 12, 13] can play this role.

Most, if not all, metadata standardization from this kind of work is intended for human readers, *e.g.*, as support for programmers who aim at coupling or linking two or more codes correctly. For compendia such as the Review of Materials Modelling (RoMM), *cf.* de Baas [14], or documentation forms such as MODA [15] (abbreviation of "Materials Modelling Data") and the EMMC Translation Case Template [16] (ETCT), which consist of sets of tables with text content to be filled in by a user [17], the situation is similar: Such metadata standards, which instruct members of a community on a recommended way of annotating their data, are human-readable, but not machine-processable. However, metadata standards can only fully exploit the capabilities of semantic technology if they are machine-processable, supporting (at least in principle) computational tasks such as automated reasoning, validity checks, the formulation and processing of queries, and the transformation or mapping from one representation to another [18]. The two main technologies [19] that fulfill these requirements are, first, markup languages specified by XML schema definitions (XSD) and, second, the semantic web based on the resource description framework (RDF). The main ordering feature in markup language technology is containment, *i.e.*, one XML tag (or an object in JSON) contains others, yielding a structural inclusion hierarchy.

Applications of this approach to materials modelling include CML [20, 21, 22], CSX [23], EngMeta [24, 25, 26, 27, 28], MSML [29], and UDLS [30]. In semantic web technology, employing RDF schemas and the web ontology language OWL, concepts are structured taxonomically by a subsumption hierarchy, while the information content itself takes the non-hierarchical form of a knowledge graph. Many existing domain ontologies are relevant to the domain of knowledge discussed here; this includes ChemAxiom [31], OntoCAPE [32, 33, 34], OntoCompChem [35, 36], OntoKin [36, 37], PHYSSYS [38], the PSO [39], the simulation intent ontology [40], multiple domain ontologies from the Virtual Materials Marketplace (VIMMP) project [12, 28, 41, 42], and some of the domain-level modules of the European Materials and Modelling Ontology (EMMO), *cf.* Goldbeck *et al.* [43], Francisco Morgado *et al.* [44], and Ghedini *et al.* [45].

The present group of authors comprises both *developers* and *end users* of domain-specific metadata standards in materials modelling. Most of us are affiliated with organizations that act as *translators* in the sense given to the term by the EMMC ASBL community: The Fraunhofer Institute for Mechanics of Materials (IWM) is an institution with the explicit purpose of facilitating industry uptake of new technologies, Goldbeck Consulting Ltd. is an independent consultancy, and the High Performance Computing Center Stuttgart (HLRS) is a facility that provides services to both academic and industrial users. Helmholtz-Zentrum Geesthacht is a scientific institution developing and implementing industrially relevant research topics, innovation platforms, and knowledge transfer systems. Below, we comment on specific lines of work, all of which are promising in our view; however, they are also disparate efforts, and the attempts to create robust connections (or any connections at all) between them have so far been insufficient. Working toward a convergence or an alignment between existing standards will create significant synergies. It will permit integrating more diverse software components into materials modelling workflows and facilitate an interaction between a greater number of digital infrastructures.

2 State of the art

2.1 CAPE-OPEN interoperability

CAPE-OPEN, wherein CAPE stands for *computer-aided process engineering* (and OPEN stands for "open"), has long been a widespread technical interoperability standard for flowsheet-based process simulation; developed from 1997 onward as a community-driven effort coordinated by the CAPE-OPEN Laboratories Network (CO-LaN), *cf.* Belaud and Pons [5, 6], it is presently supported by a multitude of process simulation packages, referred to as process modelling environments (PMEs) in CAPE-OPEN nomenclature, including leading commercial solvers such as Aspen (*cf.* Hillestad *et al.* [46]), COMSOL (*cf.* von Schenck *et al.* [47]), and gPROMS (*cf.* Moreira *et al.* [48]) as well as a dedicated free implementation by van Baten and Szczepanski [49] called COCO ("CAPE-OPEN to CAPE-OPEN"). Process modelling components (PMCs) that form part of a PME can exchange information on thermodynamic quantities; in this way, any code that provides predictions for thermodynamic data, including but not limited to fluid phase equilibria, can be connected to process simulation software if both components interoperate through CAPE-OPEN interfaces [50]. Popular thermodynamic property packages that can function as PMCs include gSAFT, MultiFlash [51], REFPROP [48], and Simulis Thermodynamics [50].

At a comparably early stage of development of CAPE-OPEN, Morbach *et al.* [32] introduced OntoCAPE as a recommended ontologization, aiming at connecting the COM based (and more recently .NET based) technical-level interoperability with data integration solutions grounded in semantic interoperability [34]; a detailed discussion of OntoCAPE is given in a reference manual by Marquardt *et al.* [33]. On the basis of OntoCAPE, Farazi *et al.* [36, 37] developed OntoKin, which specifically addresses continuum-level models of chemical reaction kinetics; their solution [37] includes an *ABox converter* that imports/exports description logic ABoxes (assertional boxes, *i.e.*, knowledge graphs) from and into the widespread file format used by the CHEMKIN-III reaction kinetics solver [52] and other interoperable packages [53], *e.g.*, for coupling reaction kinetics with CFD simulations [54]. A more recent attempt to combine CAPE-OPEN with semantic technology was made by Tolksdorf *et al.* [30] who introduced User-Defined Language Specificators (UDLS), based on the metadata standard for equations MathML [55], to support automated code generation.

2.2 EngMeta and Metadata4Ing

Within the project DIPL-ING, a metadata model for Engineering Metadata (EngMeta) was developed on the basis of requirements and use cases from thermodynamics and aerodynamics [24, 27]. EngMeta is a hierarchical metadata model, formalized in XSD, that serves as a convention on semantics in computational engineering [25, 26]; it is data-centric and permits including information on the underlying research processes (*i.e.*, the data provenance), which is crucial to data reusability. Beside process metadata, also technical, descriptive, and subject-specific metadata information from computational engineering can be stored. EngMeta is based on pre-existing metadata standards such as CodeMeta [56], DataCite [57], ExptML, and PREMIS [58]. It covers information on computational engineering research data and processes; *e.g.*, methods with their parameters, (computational) environments, and tools (hard- and software), the observed systems/research objects with their components and variables, the temporal and spatial resolution, and boundary conditions, among other data and metadata items. Metadata blocks based on EngMeta were integrated into the data repository of the University of Stuttgart and are widely used to describe research assets [59].

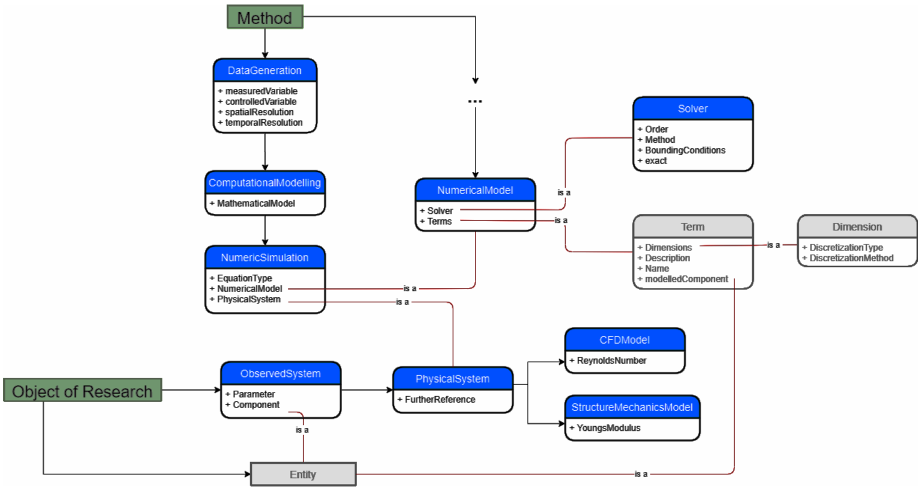

Figure 1: Classes and relations used to describe a CFD simulation.

EngMeta undergoes a process of continuous improvement and extension and should therefore be understood as a form of *scientific communication* following Edwards *et al.* [60, p. 667] rather than as a finalized outcome or product. Facilitating the scale-up from the level of a single university to academic activities at the national level, EngMeta serves as one of the starting points to metadata standardization within the German national research data infrastructure (NFDI) programme, in particular concerning the engineering sciences and the NFDI4Ing project, aiming at developing a system of ontologies for engineering and high performance computing (referred to as Metadata4Ing and Metadata4HPC). Within Metadata4Ing, the basic model of EngMeta, *cf.* Selent *et al.* [27], is combined with a hierarchical and modular approach. The main subject-specific building blocks (*i.e.*, Method, Tool, ObjectOfResearch, and Environment) are specified in more detail with the help of ontology branches; *e.g.*, Fig. 1 illustrates how the class NumericSimulation and its data and object properties can be used to annotate a CFD simulation. Apart from direct relations between the object of research and methods at a conceptual level, processing steps allow to describe specific research processes in a fine-grained way, specifying information on the employed methods, tools, and environments as well as input and output assets. This enables the provision of detailed provenance information associated with each research result. Metadata4Ing makes use of references to pre-existing semantic assets such as the Data Catalog Vocabulary [61] (DCAT), wikidata [62], and schema.org.

2.3 Chemical Markup Language

The Chemical Markup Language (CML) is a metadata standard for the chemical sciences, going back to the late 1990s [20], that is formalized as an XML schema [22]. While it was originally mainly employed to represent chemical formulas, its scope has in the meantime been generalized, covering computational chemistry and molecular dynamics simulation in general [21]; its use for data integration in molecular modelling includes the Simulation Foundry by Gygli and Pleiss [11]. An extension covering these domains is called CML-Comp [63] and was developed until 2012. In this branch of CML, information on the machine configuration and computational environment, control parameters, computational methods, thermodynamic properties, and the employed algorithms can be represented, allowing for a high level of detail, including the representation of molecules. Another standard that envolved out of CML is CompChem2 [64], which enriches CML with semantics for computational chemistry [65]. Krdzavac *et al.* [35] use concepts from CompChem2 as the foundation for OntoCompChem, an ontology for quantum chemistry, which has mainly been applied to the Gaussian code by its creators so far [35, 36].

The Molecular Simulations Markup Language (MSML) is a variant of CML adapted to the Molecular Simulation Grid (MoSGrid) platform [29]. Typically, in a first step, MSML can be used by researchers to document their workflows, providing a high-level logical (*i.e.*, non-technical) provenance description that is simulation-code agnostic. MoSGrid then uses this information to generate the simulation-code specific input data (*e.g.*, job files). After the simulation run, the MSML document is complemented by parsing the output information, *e.g.*, concerning the simulated compounds, the employed force fields and thermodynamic boundary conditions, and the computational environment. Thereby, MSML takes a role as an information broker for the simulation itself and, beyond this, for a subsequent metadata-extraction step that transforms all information from the MSML document to JSON and registers the data and metadata in a central repository service. MSML is strongly tied to the MoSGrid platform for defining workflows and extracting information, where acts as a mediator, not as the final metadata document itself. The XML schema CSX (Common Standard for eXchange), *cf.* Wang *et al.* [23], is an alternative to CompChem2 and MSML that is based on a similar choice of technology and targets roughly the same domain of knowledge, *i.e.*, MD simulation and quantum mechanics, at present mainly for GAMESS.

2.4 Thermodynamics Markup Language

The Thermodynamics Markup Language (ThermoML), an XML-based hierarchical metadata schema following a similar technological approach as EngMeta or CML, is developed by NIST and endorsed by IUPAC [66] to facilitate the annotation of thermodynamic data published in journals [67, 68]; so far, practices supporting the availability of data and metadata in ThermoML XML and JSON formats have been implemented by five journals: *Fluid Phase Equilib.*, *Int. J. Thermophys.*, *J. Chem. Eng. Data*, *J. Chem. Thermodyn.*, and *Thermochim. Acta*, *i.e.*, journals covering a significant research output which, as discussed by Frenkel [69, 70], has been growing by "more than a factor of 2 every 10 years" [70]. The aim of this effort consists in advancing research data infrastructures such as the NIST/TRC SOURCE data archival system [71], eventually yielding a "Global Information System in Thermodynamics" [69, 70]; at present, the annotated data, including over six million thermodynamic data points, are ingested into the ThermoData Engine (TDE) expert system at NIST [72]. By means of the TDE, data can be assessed for mutual consistency [73], and the accessible amount of thermodynamic data permits conducting comparably complex uncertainty analyses for models, *e.g.*, as applied by Cheung *et al.* [74] to phenomenological equations of state. ThermoML has so far only been used to annotate experimental data; however, the journals *Fluid Phase Equilib.* and *J. Chem. Eng. Data*, both with a strong focus on quantitatively characterizing the behaviour of concrete thermodynamic systems (previously, experimentally only), have in the meantime expanded their scope to include molecular modelling and simulation; the other three journals traditionally address both experimental and theoretical methods. From the ThermoML Archive [75] it is evident that nonetheless, the present implementation of the approach simply ignores simulation-based data published in these journals: Where combined experimental and simulation work has been published, the ThermoML annotation covers the experimental data only; for articles that exclusively report on molecular modelling and simulation, no XML and JSON files are generated at all. Revising ThermoML and appropriately adjusting editorial policies might provide the community with a substantial corpus of published molecular simulation results annotated in ThermoML and, thereby, advance efforts toward coherently integrating experimental and simulation data in research data infrastructures. Alternatively, European funded repositories could take over the NIST data and supplement them by simulation results; for this purpose, the NOMAD centre of exellence could be a promising candidate [76].

2.5 MODA-OSMO provenance descriptions

Interoperability between data and simulation tools, including thermodynamic property and model databases, data analysis software, and LIMS/ELN systems, and solvers for materials modelling at different granularity levels, is a significant challenge for implementing multiphysics approaches that rely on complex data processing and simulation workflows [77]. Additionally, including in the analysis a business-relevant data component increases the complexity of the problem, in particular by connecting modelling and simulation to decision-making in materials design and the application of new functional materials in industry. Moreover, the interdisciplinarity of the problem requires the collaboration of multiple scientific or industrial communities participating in the development of new products. Usually, these communities rely on their own terminologies that deviate from each other. Hence, there is a strong need for standardization of model and provenance descriptions and for the development of translation services for industry, so that partners from industry and academia can exchange information reliably. An initial effort in this direction has been undertaken by the European Materials Modelling Council (EMMC ASBL), which created a set of recommendations concerning good practices in *materials modelling translation* [41, 78] as well as business decision support systems (BDSS), *cf.* Dykeman *et al.* [79]; as a prerequisite for these developments, the EMMC coordinated efforts that led to a CEN Workshop Agreement (CWA 17284) on modelling terminology, classification, and metadata for materials modelling [15]; this CWA provides a standardized template for describing materials modelling data (*i.e.*, MODA), accounting for multiphysics approaches in terms of a uniform vocabulary [14, 15].

MODA serves as an instrument for documenting complex modelling and simulation approaches; MODA provenance descriptions facilitate the provision of metadata concerning the general modelling workflow, specifying qualitatively in what way multiple models, solvers, and data-processing operations are combined in order to obtain the final simulation outcome. At present, MODA is used mainly within the EMMC community, including many projects from the LEIT NMBP part of the EU's Horizon 2020 research and innovation programme [80]. MODA contains a use-case description that is separate and independent of any modelling information, allowing benchmarking of different simulation and experimental approaches [15]. In combination with the use-case description and a general overview, a materials simulation is described at a logical level, *i.e.*, it is stated between what elements of a workflow there is a transfer of information. This graphical representation is targeted at human readers and aims to support them at understanding the basic reasoning underlying the implemented approach; for a set of examples, the reader is referred to de Baas [14]. Beyond the CWA, the MODA metadata schema has in the meantime been extended to cover BDSS and bespoke model design for specific business cases [79]. An ontologization of these standards is provided by two components of the VIMMP system of domain ontologies [28, 42]: The Ontology for Simulation, Modelling, and Optimization [12] (OSMO) in combination with the Materials Modelling Translation Ontology [41] (MMTO). In this way, data annotated according to MODA [15], the ETCT [16, 17], and the EMMC Translators' Guide [78] can be integrated into a semantic-web framework.

Sections constitute the basic elements of a MODA-OSMO workflow. They can be of the following types:

- A *simulation overview* (summary, rationale, access conditions, *etc.*), corresponding to a MODA cover sheet.

- An *application case* describes the real phenomenon under consideration; this can be a use case (following MODA), referring to a simulated physical system, or a business, industrial, or translation case (following the ETCT).

- A *materials model* represents a physical entity by similarity and through a mathematical formalism, constituted by its governing equations (GEs); following de Baas [14], depending on the way in which the considered physical system is represented, a model is categorized as being at the electronic, atomistic, mesoscopic, or continuum granularity level.

- A *solver* provides a computational representation for the GEs and is employed to solve these equations numerically. Its scope is strictly limited to the GEs and the variables explicitly occurring in these equations.

- A *processor* represents any software carrying out computational operations that go beyond solving the GEs of a model. Usually, a simulation code plays the role of a solver and a processor; these roles are split in the logical workflow representation. OSMO distinguishes preprocessors (run before a simulation), coupled processors (synchronous with it), postprocessors (succeeding it), and data processors (independent of solver execution).

Metadata associated with these sections according to MODA (and the ETCT, respectively, in the case of business, industrial, and translation cases) are referred to as their aspects. Concepts and relations from OSMO and the MMTO cover *a)* sections and their aspects, *b)* coupling and linking of sections within a workflow, and *c)* the exchanged logical variables and key performance indicators (KPIs); in particular, every MODA workflow description has a canonical mapping into OSMO, which thereby functions as the ontology version of MODA. The logical data transfer (LDT) representation of workflows associated with OSMO includes a graphical notation that eliminates ambiguities, present in the graphical notation from MODA, concerning the precise way in which multiple elements are connected [12]. More detailed illustrations of OSMO and the MMTO are to be found in previous work [12, 28, 41, 81].

2.6 European Materials and Modelling Ontology

The EMMO is a community effort towards unifying the nomenclature within the materials science field that is led by the European Materials Modelling Council and applied in various EU projects (VIMMP, MarketPlace, H2020 DT-NMBP-09-2018 projects, *etc.* [80]). As a top-middle-level ontology, the EMMO provides a common semantic framework for representing the complex and multidisciplinary domain of materials science (including materials, models, characterization, and data) with the possibility of addressing any domain of knowledge within the applied sciences [43, 44, 45]. Its foundations in physical sciences, analytical philosophy (*i.e.*, mereotopology and semiotics [43, 82, 83, 84, 85, 86]) as well as information and communication technologies offer a representational approach to describing the real physical world and ultimately facilitate data integration and interoperability. The EMMO framework is structured into levels

– the top, middle, and domain levels – that consist of modules describing fundamental concepts (at the top) followed by generic cross-domain concepts (middle) down to application-specific representations (domain). At its middle level, the EMMO provides different options to categorize real-world objects through multiple *perspectives*, *cf.* Fig. 2, that are used as a root for the development of EMMO-compliant domain ontologies [28, 44]:

- The Reductionistic perspective provides classes, relationships, and axioms to describe real-world objects by a hierarchy of direct parts (temporal and spatial) down to its fundamental elementary level. This strict hierarchy of parts is achieved through non-transitive direct parthood relations.
- The Holistic perspective enables the description of objects as a whole. This perspective supports describing processes in terms of their participants. In particular, this is applied to represent a semiotic process (*i.e.*, a semiosis) following the theory by Charles S. Peirce [82, 84]; accordingly, a semiosis is an elementary cognitive process that involves a sign, an object, and an interpretant [45, 82, 86]. In the EMMO, semiosis is fundamental to describing models, formal languages, and properties, including thermodynamic and mechanical properties of physical systems [43, 45].
- The Perceptual perspective concerns symbolic objects; it provides a conceptualization of formal languages, pictures, geometry, and mathematics.
- The Physicalistic perspective represents real-world objects based on applied physics. This branch categorizes the physical objects into matter, fields, and elementary particles following the standard model of particle physics.

Combining multiple EMMO perspectives can facilitate bridging the gap between different domains [44, 86, 87, 88, 89].

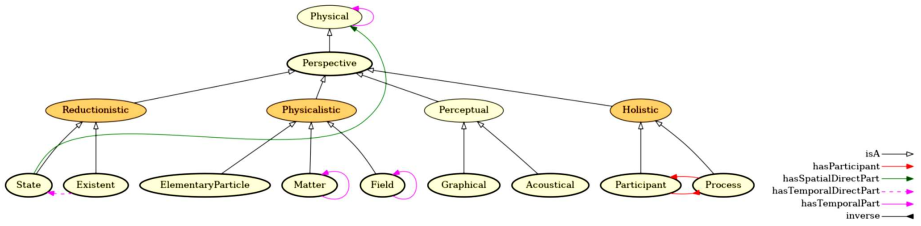

Figure 2: EMMO perspectives (from Ghedini *et al.* [45]).

3 Discussion and conclusion

The ongoing and pre-existing work discussed above shows that, on the one hand, metadata standardization and interoperability are of great concern to developers, and that both developers and users in materials modelling increasingly prioritize adherence to the FAIR principles in dealing with data; this reflects trends in scientific research and development at large. On the other hand, the existing approaches to data technology in materials modelling are poorly integrated or aligned with each other so far, and there is little common understanding of best and good practices (even "FAIR" is typically only used as a fashionable label), while promises and expectations associated with ontologies and semantic interoperability are often exaggerated irresponsibly. All this is characteristic of technology uptake in its early stages, particularly when there is a hype surrounding it. To suggest a way forward based on this assessment of the situation, we recommend to focus on five lines of development as a priority for realizing FAIR data management at the domain-specific level. We thereby limit ourselves to domains of knowledge that involve modelling and simulation. The recommended strategy can be summarized as follows:

(D) Diversification

Interoperability is only present to the extent that *different* approaches and solutions are combined with each other. Despite obverse claims, the widespread tendency among developers to say that their specific platform or tool will integrate all that exists in the field does not endorse, but counteract interoperability; "everybody will be using X in the future" not only aims at a situation where interoperability is not needed and therefore absent, it also creates conflict rather than cooperation in the predictable case of there being multiple X's. Beside addressing semantic heterogeneity as such, which can be done by implementing existing alignment techniques [90, 91, 92], the main perspective for advancing interoperability consists in *technological diversification*, since metadata standards that are given as interface specifications (Section 2.1), hierarchical XML schemas (Sections 2.2 to 2.4), and ontologies (Sections 2.5 and 2.6) represent different paradigms that need to be reconciled with each other [19] to properly communicate information on materials and processes. This includes the problem of specifying non-heuristic, canonical ways of eliminating cycles from knowledge graphs to obtain a tree-like structure that can be given a hierarchical representation [28, Chapter 5].

(O) Observation

The annotation of data must occur where the data are generated: In actual research practice. However, in metadata standardization efforts, the *observation of research practices* too often limits itself to "doing a survey," *i.e.*, encouraging or requesting prospective users to fill in a multitude of complicated forms. This cannot replace listening and actual engagement. We suggest to proceed to more interactive forms of community involvement, *e.g.*, as outlined in previous work [24]. As an outcome, agreed semantics and pragmatics must go hand in hand, such that user rights and roles as well as good (or minimally required) and best practices are specified, facilitating pragmatic interoperability [41, 93].

(R) Realistic objectives

Ontology engineering is among the fields that have in recent years been surrounded by a considerable hype, though to a lesser extent than other fields such as quantum computing or artificial intelligence which, however, is often taken to include automated reasoning and knowledge representation. In such situations, it is common for people who are superficially acquainted with a certain technology (including, but not limited to politicians) to formulate exaggerated expectations of what can be immediately accomplished to improve certain systems or entire industries. It is the *responsibility of practitioners* to correct them; nobody else can do it. Where a call for proposals is formulated along the lines of "apply for the sum of money X for a project that will reduce the cost of the industrial process Y by a factor Z," the relation between X, Y, and Z needs to be appropriate. It should not be wildly unrealistic; otherwise, project consortia will be encouraged to fuel the hype. In the worst case, this will even promote unacademic behaviour. We refrain from giving concrete examples, since this is not intended as a criticism of any institution or organization (or even any particular project or person), and doing justice to the topic would require a dedicated work of its own. This is a common challenge to technological innovation that has historically affected many new disciplines; it either ends in disappointment or, if practitioners succeed at educating decision makers and potential users, in a successful technology uptake. Under the presently predominant paradigm of organizing research work, this challenge is further exacerbated by the fact that according to conventional practices of project management, the desired outcomes are specified in advance – sometimes down to the level of detailed KPIs. That makes it even more important for such objectives to be actually realistic.

(I) Incentivization

Incentives must be in place for researchers to provide citable software [94, 95, 96] and citable open data [97]. This requires a revision of the system of metrics by which academics are evaluated, where the Hirsch index and the total number of journal-article citations are presently of major importance, whereas other modes of propagating research outcomes do not count; this creates a situation where authors are indirectly discouraged from making data and software citable in any other way than by referencing a journal article. We further refer to Mons [97] for an analysis of challenges related to incentivizing open data and to Katz *et al.* [96] for a discussion of *Software Citation Implementation Challenges*.

(C) Co-design

To ensure that the bulk of the research output in molecular and multiscale modelling and simulation is appropriately annotated and made available to all through an ingest into FAIR research data infrastructures, it is essential for solver development to go hand in hand with the development of the targeted digital platforms. Since many different solvers produce data that need to be processed by many digital infrastructures, this is a n:n communication problem that requires genuine interoperability, both at the semantic and at the technical level. As Gygli and Pleiss [11] observe, interoperability in molecular modelling and simulation can only be achieved when simulation deployment is linked to automated annotation in accordance with metadata standards that enjoy widespread recognition. The required co-design of data technology and simulation technology can be mediated by a workflow management system that ensures technical interoperability with respect to multiple solvers and processing elements, while ensuring semantic interoperability in its interactions with digital platforms. In this respect, best practice in the field is represented by SimPhoNy [10], a workflow management system that is co-designed with the EMMO through EMMO-CUDS, a semantic data structure; other promising developments include AiiDA [8, 13], from which provenance descriptions can be obtained [9, 13], and the Salome/YACS workflow management system [7] which is connected to the VIMMP ontologies by an XSD-based common data model.

These five recommendations or principles, the DORIC principles, are proposed to the community for a thorough critique and discussion at the DORIC-

MM 2021 workshop so that they can become a part of the associated OntoCommons project deliverable. Where appropriate, we suggest that they be implemented into work programmes of EMMC and RDA task groups as well as collaborative projects, *e.g.*, within Horizon Europe.[1]

Acknowledgment

This work was supported by activities of the Innovation Centre for Process Data Technology (Inprodat e.V.), Kaiserslautern. The co-authors M.T.H. and B.S. acknowledge funding from the German Research Foundation (DFG) through the National Research Data Infrastructure for Catalysis-Related Sciences (NFDI4Cat), DFG project no. 441926934, and the co-author D.I. through the National Research Data Infrastructure for Engineering Sciences (NFDI4Ing), DFG project no. 442146713, both within the National Research Data Infrastructure (NFDI) programme of the Joint Science Conference (GWK). The co-authors G.G. and N.A.K. acknowledge funding from the European Union's Horizon 2020 research innovation programme (H2020) under grant agreements no. 723867 (EMMC-CSA) and 862136 (OntoTrans), N.A.K. also acknowledges funding from H2020 under grant agreement no. 952903 (VIPCOAT), and the co-authors J.F.M. and M.T.H. acknowledge funding from H2020 under grant agreement no. 958371 (OntoCommons). Simon Stephan is acknowledged for bringing up the suggestion to write a manuscript along the lines of the present work.

References

[1] J. Bicarregui: 2016. Technical report, Wellcome Trust. doi:10.6084/m9.figshare.4055538.v2.

[2] G. Guizzardi: 2020. *Data Intelligence* **2**(1–2), 181–191. doi:10.1162/dint_a_00040.

[3] Å. Ervik, A. Mejía, and E. A. Müller: 2016. *Journal of Chemical Information and Modeling* **56**(9), 1609–1614. doi:10.1021/acs.jcim.6b00149.

[4] S. Stephan, M. T. Horsch, J. Vrabec, and H. Hasse: 2019. *Molecular Simulation* **45**(10), 806–814. doi:10.1080/08927022.2019.1601191.

[5] J.-P. Belaud and M. Pons: 2002. *Computer Aided Chemical Engineering* **10**, 847–852. doi:10.1016/S1570-7946(02)80169-9.

[6] J.-P. Belaud and M. Pons: 2014. *Chemie Ingenieur Technik* **86**(7), 1052–1064. doi:10.1002/cite.201400009.

[7] A. Ribes and C. Caremoli: 2007. In: *Proc. COMPSAC 2007*, Vol. 2. Los Alamitos: IEEE, pp. 553–564. doi:10.1109/COMPSAC.2007.185.

[8] G. Pizzi and others: 2016. *Computational Materials Science* **111**, 218–230. doi:10.1016/j.commatsci.2015.09.013.

[9] A. Merkys and others: 2017. *Journal of Cheminformatics* **9**, 56. doi:10.1186/s13321-017-0242-y.

[10] J. Adler and others: 2018. *Computer Physics Communications* **231**, 45–61. doi:10.1016/j.cpc.2018.05.005.

[11] G. Gygli and J. Pleiss: 2020. *Journal of Chemical Information and Modeling* **60**(4), 1922–1927. doi:10.1021/acs.jcim.0c00018.

[1] The reader is referred to the proceedings from a recent WCCM-ECCOMAS minisymposium organized by Konchakova and Klein where a series of topics related to the present work were discussed [86, 87, 88, 89].

[12] M. T. Horsch and others: 2020. *Journal of Chemical & Engineering Data* **65**(3), 1313–1329. doi:10.1021/acs.jced.9b00739.

[13] S. P. Huber and others: 2020. *Scientific Data* **7**, 300. doi:10.1038/s41597-020-00638-4.

[14] A. F. De Baas (ed.): 2017, *What Makes a Material Function? Let Me Compute the Ways...*. Luxembourg: EU Publications Office. ISBN 978-92-79-63185-6.

[15] CEN workshop agreement (CWA) 17284:2018 (E). ftp://ftp.cencenelec.eu/CEN/Sectors/TCandWorkshops/Workshops/WS%20MODA/CWA_17284.pdf; date of access: 1st March 2021.

[16] https://emmc.info/emmc-translation-case-template/; date of access: 23rd January 2021.

[17] M. Pezzotta and others: 2021. Technical report, EMMC. doi:10.5281/zenodo.4457849.

[18] S. Zhao and Q. Qian: 2017. *AIP Advances* **7**, 105325. doi:10.1063/1.4999209.

[19] T. Kramer and others: 2015. Technical Report NISTIR 8068, NIST. doi:10.6028/NIST.IR.8068.

[20] P. Murray-Rust and H. S. Rzepa: 1999. *Journal of Chemical Information and Computer Sciences* **39**(6), 928–942. doi:10.1021/ci990052b.

[21] T. O. H. White and others: 2006, 'Application and uses of CML within the eMinerals project'. In: *Proc. UK e-Science All Hands Meeting 2006*. Edinburgh: National e-Science Centre, pp. 606–613. ISBN 978-0-95539881-0.

[22] P. Murray-Rust and H. S. Rzepa: 2011. *Journal of Cheminformatics* **3**, 44. doi:10.1186/1758-2946-3-44.

[23] B. Wang and others: 2017. *Journal of Physical Chemistry A* **121**(1), 298–307. doi:10.1021/acs.jpca.6b10489.

[24] D. Iglezakis and B. Schembera: 2018. *o-bib. Das offene Bibliotheksjournal* **5**(3), 46–60. doi:10.5282/o-bib/2018H3S46-60.

[25] B. Schembera and D. Iglezakis: 2019, 'The genesis of EngMeta: A metadata model for research data in computational engineering'. In: *Metadata and Semantic Research*. Cham: Springer, pp. 127–132. ISBN 978-3-030-14400-5.

[26] B. Schembera and D. Iglezakis: 2020. *International Journal of Metadata, Semantics and Ontologies* **14**(1), 26–38. doi:10.1504/IJMSO.2020.107792.

[27] B. Selent and others: 2020, 'Management of research data in computational fluid dynamics and thermodynamics'. In: *Proc. E-Science-Tage 2019*. Heidelberg: heiBOOKS, pp. 128–139. ISBN 978-3-948083-15-1.

[28] M. T. Horsch, S. Chiacchiera, W. L. Cavalcanti, and B. Schembera: 2021, *Data Technology in Materials Modelling*. Cham: Springer. ISBN 978-3-03068596-6.

[29] R. Grunzke and others: 2014. *Concurrency and Computation: Practice and Experience* **26**(10). doi:10.1002/cpe.3116.

[30] G. Tolksdorf, E. Esche, G. Wozny, and J.-U. Repke: 2019. *Computers & Chemical Engineering* **121**, 670–684. doi:10.1016/j.compchemeng.2018.12.006.

[31] N. Adams, E. Cannon, and P. Murray-Rust: 2009, 'ChemAxiom – An ontological framework for chemistry in science'. *Nature Precedings*. doi:10.1038/npre.2009.3714.1.

[32] J. Morbach, A. Wiesner, and W. Marquardt: 2008. *Computer Aided Chemical Engineering* **25**, 991–996. doi:10.1016/S1570-7946(08)80171-X.

[33] W. Marquardt, J. Morbach, A. Wiesner, and A. Yang: 2010, *OntoCAPE: A Re-Usable Ontology for Chemical Process Engineering.* Heidelberg: Springer. ISBN 978-3-642-04654-4.

[34] A. Wiesner, J. Morbach, and W. Marquardt: 2011. *Computers & Chemical Engineering* **35**(4), 692–708. doi:10.1016/j.compchemeng.2010.12.003.

[35] N. Krdzavac and others: 2019. *Journal of Chemical Information and Modeling* **59**(7), 3154–3165. doi:10.1021/acs.jcim.9b00227.

[36] F. Farazi and others: 2020. *Computers and Chemical Engineering* **137**, 106813. doi:10.1016/j.compchemeng.2020.106813.

[37] F. Farazi and others: 2020. *Journal of Chemical Information and Modeling* **60**(1), 108–120. doi:10.1021/acs.jcim.9b00960.

[38] P. Borst, H. Akkermans, and J. Top: 1997. *International Journal of Human-Computer Studies* **46**(2–3), 365–406. doi:10.1006/ijhc.1996.0096.

[39] H. Cheong and A. Butscher: 2019. *Journal of Engineering Design* **30**(10–12), 655–687. doi:10.1080/09544828.2019.1644301.

[40] F. Boussuge and others: 2019. *Journal of Engineering Design* **30**, 688–725. doi:10.1080/09544828.2019.1630806.

[41] M. T. Horsch and others: 2021. In: *Proc. DAMDID 2020.* Cham: Springer, pp. 45–59. ISBN 978-3-030-81199-0, doi:10.1007/978-3-030-81200-3_4.

[42] M. T. Horsch and others: 2020. *KI – Künstliche Intelligenz* **34**(3), 423–428. doi:10.1007/s13218-020-00648-9.

[43] G. Goldbeck and others: 2019, 'A reference language and ontology for materials modelling and interoperability'. In: *Proc. NAFEMS World Congress 2019.* Knutsford: NAFEMS, p. NWC_19_86.

[44] J. Francisco Morgado and others: 2020, 'Mechanical testing ontology for digital-twins: A roadmap based on EMMO'. In: *Proc. SeDiT 2020.* Aachen: CEUR-WS, p. 3.

[45] E. Ghedini and others: 2020. https://emmo-repo.github.io/versions/1.0.0-beta/; date of access: 8th February 2021.

[46] M. Hillestad and others: 2018. *Fuel* **234**, 1431–1451. doi:10.1016/j.fuel.2018.08.004.

[47] H. von Schenk, G. Andersson, J. van Baten, and E. Fontes: 2008, 'A COMSOL interface to CAPE-OPEN compliant physical and thermodynamic property packages'. In: *Proc. 2008 AIChE Annual Meeting.* ISBN 978-0-81691050-2.

[48] M. A. Moreira, A. M. Ribeiro, A. F. P. Ferreira, and A. E. Rodrigues: 2017. *Separation and Purification Technology* **173**, 339–356. doi:10.1016/j.seppur.2016.09.044.

[49] J. van Baten and R. Szczepanski: 2011. *Computers & Chemical Engineering* **35**(7), 1251–1256. doi:10.1016/j.compchemeng.2010.07.016.

[50] R. Morales Rodríguez and others: 2008. *Chemical Engineering Research and Design* **86**(7). doi:10.1016/j.cherd.2008.02.022.

[51] B. Edmonds and T. Moorwood: 2007, 'Multiflash: A Cape Open 1.1 thermodynamics package with multiphase capabilities'. In: *Proc. AIChE Annual Meeting 2007.* ISBN 978-081691022-9.

[52] R. J. Kee, F. M. Rupley, E. Meeks, and J. A. Miller: 1996. Technical Report SAND-96-8216, Sandia National Laboratories. doi:10.2172/481621.

[53] D. G. Goodwin, R. L. Speth, H. K. Moffat, and B. W. Weber: 2018. Technical report. doi:10.5281/zenodo.1174508.

[54] J. U. Eichmeier, R. Reitz, and C. Rutland: 2014. *SAE International Journal of Engines* **7**(1), 106–119. doi:10.4271/2014-01-1074.

[55] D. Carlisle, P. Ion, and R. Miner: 2014. W3C recommendation. https://www.w3.org/TR/2014/REC-MathML3-20140410/; date of access: 24th January 2021.

[56] https://codemeta.github.io/; date of access: 25th January 2021.

[57] DataCite Metadata Working Group: 2014. Technical report, DataCite e.V. doi:10.5438/0010.

[58] P. Caplan: 2017. Technical report, Library of Congress Network Development and MARC Standards Office. Revised by the PREMIS editorial committee. https://www.loc.gov/standards/premis/understanding-premis-rev2017.pdf; date of access: 1st March 2021.

[59] B. Schembera: 2021. *Journal of Supercomputing* **77**, 8946–8966. doi:10.1007/s11227-020-03602-6.

[60] P. N. Edwards and others: 2011. *Social Studies of Science* **41**(5), 667–690. doi:10.1177/0306312711413314.

[61] R. Albertoni and others: 2020. W3C recommendation. https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/; date of access: 26th January 2021.

[62] https://www.wikidata.org/wiki/Wikidata:Introduction; date of access: 25th January 2021.

[63] A. Walker: 2012. http://homepages.see.leeds.ac.uk/~earawa/CMLComp/; date of access: 14th January 2021.

[64] W. Phadungsukanan, S. Adams, J. Townsend, and J. Thomas: 2011. http://www.xml-cml.org/convention/compchem-20110524; date of access: 14th January 2021.

[65] W. Phadungsukanan, M. Kraft, J. A. Townsend, and P. Murray-Rust: 2012. *Journal of Cheminformatics* **4**, 15. doi:10.1186/1758-2946-4-15.

[66] IUPAC Committee on Printed and Electronic Publications. https://old.iupac.org/namespaces/ThermoML/; date of access: 27th January 2021.

[67] M. Frenkel and others: 2006. *Pure and Applied Chemistry* **78**(3), 541–612. doi:10.1351/pac200678030541.

[68] M. Frenkel and others: 2011. *Pure and Applied Chemistry* **83**(10), 1937–1969. doi:10.1351/PAC-REC-11-05-01.

[69] M. Frenkel: 2009. *Journal of Chemical & Engineering Data* **54**(9), 2411–2428. doi:10.1021/je800877f.

[70] M. Frenkel: 2015. *Journal of Chemical Thermodynamics* **84**, 18–40. doi:10.1016/j.jct.2014.12.016.

[71] A. Kazakov and others: 2012. *International Journal of Thermophysics* **33**, 22–33. doi:10.1007/s10765-011-1107-7.

[72] V. Diky and others: 2013. *Journal of Chemical Information and Modeling* **53**(12), 3418–3430. doi:10.1021/ci4005699.

[73] V. Diky and others: 2019. *Journal of Chemical Thermodynamics* **133**, 208–222. doi:10.1016/j.jct.2019.01.029.

[74] H. Cheung and others: 2020. *Journal of Chemical & Engineering Data* **65**(2), 503–522. doi:10.1021/acs.jced.9b00689.

[75] NIST Thermodynamics Research Center (TRC). https://trc.nist.gov/ThermoML.html; date of access: 30th January 2021.

[76] C. Draxl and M. Scheffler: 2018. *MRS Bulletin* **43**(9), 676–682. doi:10.1557/mrs.2018.208.

[77] Z. M. Mir and others: 2020. *Modelling and Simulation in Materials Science and Engineering* **28**, 025003. doi:10.1088/1361-651X/ab6209.

[78] D. Hristova-Bogaerds and others: 2019. Technical report. doi:10.5281/zenodo.3552260.

[79] D. Dykeman, A. Hashibon, P. Klein, and S. Belouettar: 2020. Technical report. doi:10.5281/zenodo.4054009.

[80] European Commission: 2020. https://ec.europa.eu/programmes/horizon2020/en/h2020-section/nanotechnologies-advanced-materials-advanced-manufacturing-and-processing-and/; date of access: 27th December 2020.

[81] M. Heinen and others: 2021. In: *High Performance Computing in Science and Engineering '20*. Cham: Springer. To appear. Preprint: doi:10.5281/zenodo.3932942.

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

[83] N. Asher and L. Vieu: 1995, 'Toward a geometry of common sense: A semantics and a complete axiomatization of mereotopology'. In: *Proc. 14th IJCAI*. San Mateo: Morgan Kaufmann, pp. 846–852. ISBN 978-1-55860-363-9.

[84] M. Keeler: 2000. In: *Proc. ICCS 2000*. Heidelberg: Springer, pp. 82–99. doi:10.1007/10722280_6.

[85] B. Smith and A. C. Varzi: 2000. *Philosophy and Phenomenological Research* **60**(2), 103–119. doi:10.2307/2653492.

[86] M. T. Horsch and others: 2021. In: *Proc. WCCM-ECCOMAS 2020*. Barcelona: Scipedia. doi:10.23967/wccm-eccomas.2020.297.

[87] D. Höche and others: 2021. In: *Proc. WCCM-ECCOMAS 2020*. Barcelona: Scipedia. doi:10.23967/wccm-eccomas.2020.263.

[88] T. F. Hagelien and others: 2021. In: *Proc. WCCM-ECCOMAS 2020*. Barcelona: Scipedia. doi:10.23967/wccm-eccomas.2020.035.

[89] H. A. Preisig and others: 2021. In: *Proc. WCCM-ECCOMAS 2020*. Barcelona: Scipedia. doi:10.23967/wccm-eccomas.2020.262.

[90] M. Koutraki, N. Preda, and D. Vodislav: 2017. In: *Proc. ESWC 2017*. Cham: Springer, pp. 152–168. ISBN 978-3-319-58067-8, doi:10.1007/978-3-319-58068-5_10.

[91] P. Ochieng and S. Kyanda: 2018. *Distributed and Parallel Databases* **36**, 195–217. doi:10.1007/s10619-017-7206-0.

[92] L. Zhou, M. Cheatham, and P. Hitzler: 2020. In: *Proc. JIST 2019*. Cham: Springer, pp. 287–303. doi:10.1007/978-3-030-41407-8_19.

[93] F. Weidt Neiva, J. M. N. David, R. Braga, and F. Campos: 2016. *Information and Software Technology* **72**, 137–150. doi:10.1016/j.infsof.2015.12.013.

[94] A. M. Smith, D. S. Katz, K. E. Niemeyer, and FORCE11 Software Citation WG: 2016. *PeerJ Computer Science* **2**, e86. doi:10.7717/peerj-cs.86.

[95] S. Druskat, N. C. Hong, R. Haines, and J. Baker: 2018. Technical report. doi:10.5281/zenodo.1405679.

[96] D. S. Katz and others: 2019, 'Software Citation Implementation Challenges'. Technical report arXiv:1905.08674 [cs.CY].

[97] B. Mons: 2018, *Data Stewardship for Open Science*. Boca Raton: CRC. ISBN 978-1-4987-5317-3.