Science and Technology Facilities Council

Conference Proceedings
DL-CONF-2021-002

Domain Ontologies for Research Data Management in Industry Commons of Materials and Manufacturing (DORIC-MM 2021)

Online, 7th June 2021

S Chiacchiera, MT Horsch, J Francisco Morgado, G Goldbeck

August 2021

Enquiries concerning this report should be addressed to:

Chadwick Library
STFC Daresbury Laboratory
Sci-Tech Daresbury
Keckwick Lane
Warrington
WA4 4AD

Tel: +44(0)1925 603397
Fax: +44(0)1925 603779
email: librarydl@stfc.ac.uk

Science and Technology Facilities Council reports are available online at:
https://epubs.stfc.ac.uk

ISSN 1362-0223

Contents

Preface

This volume collects all the contributed papers that have been presented in the *Domain Ontologies for Research Data Management in Industry Commons of Materials and Manufacturing* (DORIC-MM 2021) workshop. The event was held on-line on the 7th of June 2021, co-located with the 18th European Semantic Web Conference (ESWC), and was preceded by an on-line kick-off, on the 15th of March.

The purpose of this activity was to support the semantic landscape analysis in the field of materials and manufacturing (MM), and it aimed to gather all interested parties, including MM domain experts and ontologists.

DORIC-MM has been organized in the framework of the OntoCommons H2020 project[1]: for more information, we point the reader to the event website[2] and to the dedicated project deliverable[3]. In particular, the latter contains summaries of keynotes and invited contributions, highlights from the participants' input, and the workshop conclusions.

The Programme Committee for the event was constituted as follows (where organizers' names are underlined):

- Stefano Borgo (CNR, Italy)
- Welchy Leite Cavalcanti (Fraunhofer IFAM, Germany)
- Silvia Chiacchiera (STFC/UKRI, United Kingdom)
- Fabien Duchateau (University of Lyon 1, France)
- Iker Esnaola González (Tekniker, Spain)
- Anna Fensel (University of Innsbruck, Austria)
- Joana Francisco Morgado (Fraunhofer IWM, Germany)
- Gerhard Goldbeck (Goldbeck Consulting Ltd & EMMC ASBL, UK & Belgium)
- Martin Thomas Horsch (HLRS, Germany)
- Dimitrios Kyritsis (EPFL, Switzerland & UiO, Norway)
- María Poveda Villalón (Universidad Politécnica de Madrid, Spain)
- Umutcan Şimşek (University of Innsbruck, Austria)

Daresbury, Stuttgart,
Freiburg, and Cambridge
August 2021

Silvia Chiacchiera, Martin T. Horsch,
Joana Francisco Morgado and Gerhard Goldbeck

[1] `https://ontocommons.eu/`
[2] `https://ontocommons.eu/doric-mm-2021`
[3] OntoCommons Deliverable 3.9, to appear on Zenodo.

Acknowledgements

We would like to thank all authors, reviewers, participants, and all the colleagues, mainly from OntoCommons WorkPackage 3 ("Industrial domain ontologies"), but not only, who supported the event preparation and running.

OntoCommons - Ontology-driven data documentation for Industry Commons, has received funding from the European Unions Horizon 2020 research and innovation programme under Grant Agreement no. 958371.

DORIC-MM event program

Kick-off of DORIC-MM, 15th March (14:30-17:30 CET)

- **14:30-15:30: Initial plenary session. Introduction, interactive session. Input from the survey and recent events within the field of semantics applied to materials and manufacturing**
  - Welcome by the Organizers
  - Introduction to OntoCommons and WP3 (Industrial domain ontologies) [by Hedi Karray, ENIT, France WP3 Leader and Project Technical Manager]
  - Landscape analysis [by Yann Le Franc, e-Science Data Factory, France]
  - Introduction to the Industrial Ontologies Foundry [Dimitris Kiritsis, EPFL, Switzerland & UiO, Norway]
  - Input from the EMMC 2021 International Workshop [Gerhard Goldbeck, Goldbeck Consulting Ltd & EMMC ASBL, UK & Belgium]
  - Interactive presentation [Silvia Chiacchiera, UKRI, UK]
- **15:30-16:30: Domain-specific interactive parallel sessions (D1, D2, D3, D4)**
  - D1: Physics and Chemistry [Moderator: Gerhard Goldbeck]
  - D2: Mechanical and Industrial Engineering [Moderator: Hedi Karray]
  - D3: Thermal Engineering/Process Engineering [Moderator: Martin Thomas Horsch]
  - D4: Material Science and Engineering [Moderator: Yann le Franc]
- **16:45 - 17:30: Final plenary session - Joining, analyzing and wrapping up**
  - 16:45-16:50: General intro and exchange
  - 16:50-17:10: Reports from each of the domain parallel sessions D1, D2, D3, D4
  - 17:10-17:25: Panel discussion
  - 17:25-17:30: Closing

DORIC-MM Workshop, 7th June (full day)

- **Morning session (10:30-13:45 CEST, 7th June): Introduction, 1 keynote, 4 contributions (3 papers + 1 invited) on "Materials & modelling" and discussion.**
  - 10:30-10:40 Welcome and introduction
  - 10:40-11:05 [20+5 min] Hedi Karray, "Ontologies Interoperability: concerns and perspectives"
  - 11:05-12:00 "Materials & modelling" session

    - 11:05-11.20 [10+5 min] M. Abd Nikooie Pour et al, “A First Step towards Extending the Materials Design Ontology”
    - 11:20-11:35 [10+5 min] M. T. Horsch et al, “Domain-specific metadata standardization in materials modelling”
    - 11:35-11:50 [10+5 min] F. Le Piane et al, “Introducing MAMBO: Materials And Molecules Basic Ontology”
    - 11:50-12:05 [10+5 min] J. Friis and E. Ghedini, “Domain-level ontologies and the methodology to connect them to a Top-level/Middle-level ontology”
  - 12:05-12:20 Break
  - 12:20-13:45 Discussion (Panel + all, interactively). Panel members: Alexander Behr (Dortmund Univ., Germany), Jesper Friis (SINTEF, Norway), David Leal (CAESAR Systems Ltd, UK), Heinz Preisig (NTNU, Norway). Moderator: Gerhard Goldbeck (Goldbeck Consulting Ltd & EMMC ASBL, UK & Belgium). Initial interactive presentation by Silvia Chiacchiera (UKRI, UK).

- **13:45-15:00 Lunch break**

- **Afternoon session (15:00-18.15 CEST, 7th June): Highlights from material gathered during the 15/03 preparatory event, 1 keynote, 5 contributions (3 papers + 2 invited) on “Industry & engineering” and discussion.**
  - 15:00-15:15 Highlights from the material gathered during the 15/03 preparatory event
  - 15:15-15:40 [20+5 min] Evgeny Kharlamov: “Industrial ontologies for manufacturing”
  - 15:40-16:55 “Industry & engineering” session
    - 15:40-15:55 [10+5 min] M. M. Vegetti et al, “SCONTO: A Modular Ontology for Supply Chain Representation”
    - 15:55-16:10 [10+5 min] S. Borgo, F. Compagno et al, “An overview of some ontological challenges in engineering maintenance”
    - 16:10-16:25 [10+5 min] I. Esnaola-Gonzalez and I. Fernandez, “Materials Tribological Characterisation: an OntoCommons Use Case”
    - 16:25-16:40 [10+5 min] Johan Wilhelm Klüwer, “READI: Ontology-based requirements management for industry”
    - 16:40-16:55 [10+5 min] Maja Milicic Brandt, “Industrial Ontology Library at Siemens”
  - 16:55-17:05 Break
  - 17:05-18:05 Discussion (Panel + all, interactively). Panel members: Mehwish Alam (KIT, Germany), Gianmaria Bullegas (Perpetual Labs Ltd, UK), David Cameron (Univ. of Oslo, Norway), Irlan Grangel-Gonzalez (Bosch, Germany), Johan Klüwer (DNV, Norway), Boonserm Kulvatunyou (NIST, Usa), Maja Milicic Brandt (Siemens AG, Germany), Robert Young (Loughborough Univ., UK). Moderator: Martin T. Horsch (HLRS, Germany).
  - 18:05-18:15 Wrapping up and closing

A First Step towards Extending the Materials Design Ontology

Mina Abd Nikooie Pour[1], Huanyu Li[1,3],
Rickard Armiento[2,3], and Patrick Lambrix[1,3]

[1] Department of Computer and Information Science,
Linköping University, 581 83 Linköping, Sweden
[2] Department of Physics, Chemistry and Biology,
Linköping University, 581 83 Linköping, Sweden
[3] The Swedish e-Science Research Centre, Linköping University,
581 83 Linköping, Sweden
`firstname.lastname@liu.se`

Abstract. Ontologies have been proposed as a means towards making data FAIR (Findable, Accessible, Interoperable, Reusable) and has recently attracted much interest in the materials science community. Ontologies for this domain are being developed and one such effort is the Materials Design Ontology. However, to obtain good results when using ontologies in semantically-enabled applications, the ontologies need to be of high quality. One of the quality aspects is that the ontologies should be as complete as possible. In this paper we show preliminary results regarding extending the Materials Design Ontology using a phrase-based topic model.

Keywords: ontology, ontology extension, materials design, topic model

1 Introduction

In many areas there is a recent interest in making data FAIR, i.e., Findable, Accessible, Interoperable, and Reusable [16]. Findable refers to the fact that data and metadata should be easy to find, accessible to the fact that it should be clear how to access the data, interoperable to the fact that the data needs to be integrated with other data and be usable by applications and workflows, and reusable to the fact that data and metadata are well described such that the data can be replicated or combined in different settings. Ontologies have been proposed as a means towards making data FAIR. Also in the materials science domain there is an awareness regarding the importance of the FAIR principles [4] and efforts are on the way to develop upper ontologies such as EMMO (European Materials & Modelling Ontology), and domain ontologies regarding different sub-domains of materials science such as Mat-Onto [2], Materials Ontology [1], NanoParticle Ontology [14], eNanoMapper ontology [6], ontologies related to computational molecular engineering [7], Materials Design Ontology (MDO) [11], and Materials Graph Ontology [15].

However, to obtain good results when using ontologies for semantically-enabled applications, the ontologies need to be of high quality. One of the quality aspects is that the ontologies should be as complete as possible which relates to the requirement of domain coverage in [12].[1] Many techniques exist for finding missing information in ontologies (see overview in [8]) and extending them. In this paper we show preliminary results of using a variant of the method for extending ontologies that we developed in [10] on MDO.

The remainder of the paper is organized as follows. In section 2 we describe MDO, while in section 3 we describe the method for extending ontologies. In section 4 we show preliminary results of applying the method to MDO. The paper concludes in section 5.

2 The Materials Design Ontology (MDO)

MDO [11] was developed using the NeOn ontology engineering methodology [13], as an answer to the need for an ontology to represent concepts which are the basis for materials design, such as structures of materials, properties of materials, materials calculations and relationships among them. The development was guided by the schemas of the Open Databases Integration for Materials Design (OPTIMADE[2]) project which aims at making materials databases interoperable by developing a common API. The OPTIMADE schemas are based on a consensus reached by several of the materials database providers in the field.

The current version of MDO is publicly available at w3id.org[3] and consists of four modules (Figure 1) [11]. The *Core* module consists of the top-level concepts and relationships of MDO that are reused in other modules. The *Structure* module represents the structural information of materials. The *Calculation* module represents a classification of different computational methods. The *Provenance* module represents provenance information of materials data and calculations. The OWL2 DL representation of the ontology contains 37 classes, 32 object properties, and 32 data properties.

3 Method for extending ontologies

In [10] we presented a general approach for extending ontologies, shown in Figure 2, and showed its use by extending two ontologies in the nanotechnology field. In this paper we use a variant of the approach. We mention the changes from the approach in [10] while describing how we extend MDO in section 4.

Our approach contains two steps. In the first step a phrase-based topic model is created using the ToPMine system [5]. Given a corpus of documents related

[1] In practice, it is difficult to know when an ontology is complete according to the domain, but it is possible to define an 'is more complete than' relation between ontologies which can be used for comparing completeness [8].

[2] `https://www.optimade.org/`

[3] `https://w3id.org/mdo/full/1.0/`

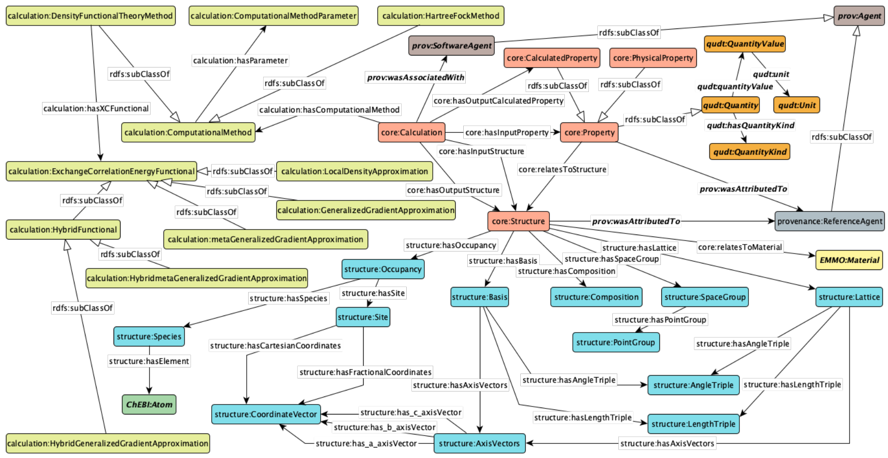


Fig. 1. The Materials Design Ontology [11].

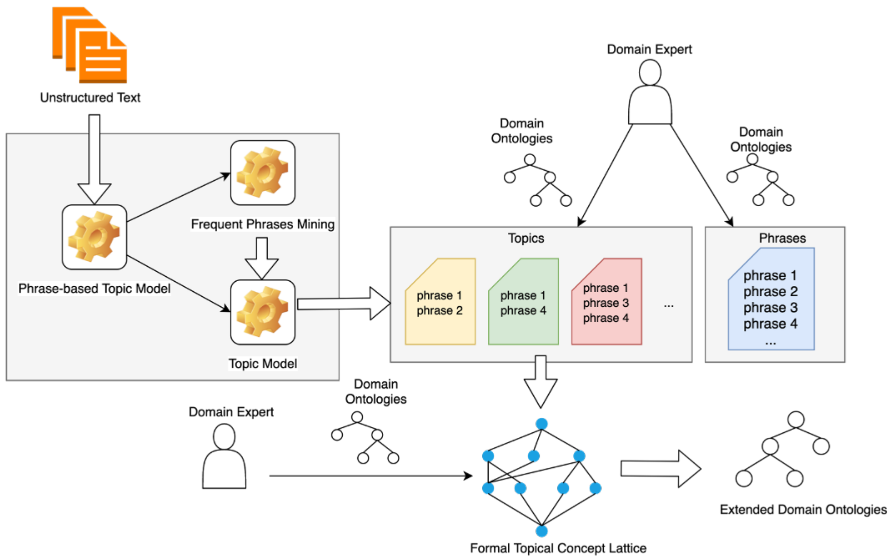


Fig. 2. Approach: The upper part of the figure shows the creation of a phrase-based topic model with unstructured text as input and phrases and topics as output. The lower part shows the formal topical concept analysis with as input topics and as output a topical concept lattice. In both parts a domain expert validates and interprets the results. [10]

to the domain of interest and the number of requested topics, representations of latent topics in the documents are computed. The phrases as well as the topics are suggestions that a domain expert should validate or interpret and relate to concepts in the ontology.

The second step generates suggestions to the domain expert regarding relations between topics based on formal topical concept analysis [10].

Based on the validations and interpretations of the domain expert, concepts and axioms are added to the ontology.

4 Extending the Materials Design Ontology

4.1 Data

A first step is to collect the corpus that is used as input. The approach in [10] does not specify how the corpus should be collected. In that paper we used an existing library of documents related to the field. In this paper we use MDO as a seed for querying journal databases. We use two journals in the field of materials design: NPJ Computational Materials[4] and Computational Materials Science[5]. We use the 37 concepts of MDO as search phrases in the two journals to find relevant articles and retrieve titles and abstracts of the returned articles. The corpus contains titles and abstracts from 403 articles of NPJ Computational Materials and 8,193 from Computational Materials Science.

In the preprocessing step characters are set to lower case and punctuations are removed. Further, we remove words of length one or two. After preprocessing there are 21,548 distinct words which together occur 808,862 times. An overview of the frequency of the words is presented in Table 1. Most of the words (72.27%) occur less than 10 times, while there are 17 words that occur more than 3000 times. These are 'based', 'properties', 'method', 'calculations', 'phase', 'materials', 'study', 'structure', 'temperature', 'density', 'results', 'energy', 'electronic', 'model', 'molecular', 'simulations', 'surface'.

Table 1. The distribution of word frequency after preprocessing.

Frequency	Percentage of words
less than 10	72.27
10-30	13.25
31-100	7.76
101-500	5.25
501-1000	0.83
1001-2000	0.44
2001-3000	0.12
More than 3000	0.08

[4] https://www.sciencedirect.com/journal/computational-materials-science
[5] https://www.nature.com/npjcompumats/

4.2 Frequent phrases

Given a minimum support threshold *min_support*, we say that phrases that occur at least *min_support* times are *frequent phrases.* ToPMine generates frequent phrases of a length up to a maximum length that is given as an input parameter. In our experiments this was set to 10. Further, ToPMine does not generate all frequent phrases but uses a method based on partitioning documents and using a significance score for deciding which words likely belong together, to produce high-quality frequent phrases [5].

The second column of Table 2 shows the number of frequent phrases that ToPMine generates for different values of *min_support.* The higher the *min_support*, the fewer frequent phrases are generated.

Table 2. Number of frequent phrases for *min_support* 10, 15, 20, 25 and 30 respectively, and three different versions of the ToPMine algorithm.

min_support	original TopMine	New ToPMine without stemming	New ToPMine with stemming
10	6901	6,478	5,452
15	3826	3,578	3,022
20	2542	2,402	2,046
25	1816	1,722	1,477
30	1375	1,298	1,119

In addition, in this paper we also define a maximum support threshold *max_support_word.* Words that occur more than *max_support_word* times are removed. These words are usually very general terms that are not interesting for an ontology or that would not be interesting for a domain ontology, but possibly for an upper ontology. We do note, however, that some of these words could be useful such as 'method', 'electronic', 'model', and 'molecular'. In the remainder we call 'New ToPMine' the algorithm that adds *max_support_word* as well as the preprocessing step. The second column in Table 3 shows how *max_support_word* influences the number of generated frequent phrases with a constant *min_support* of 10. The higher *max_support_word*, the more frequent phrases are generated. Note that no word occurs more than 8000 times in our corpus, so setting *max_support_word* to 8000 allows all words (or, in other words, *max_support_word* is not used).

Another way to look at the influence of *min_support* and *max_support_word* is to compare how many of the frequent phrases are the same and different for different settings. In Figure 3 we show this comparison of different settings to the base setting where *min_support* is 10 and *max_support_word* is 8000 (i.e., *max_support_word* is not used) which is shown in the middle of the figure. The 'Same' bars show how many generated phrases occur both in the base setting and the compared setting. The 'Removed' bars show how many frequent phrases occur in the base setting, but not in the compared setting. For the cases where

Table 3. Number of frequent phrases for *min_support* to 10 and for *max_support_word* 500, 1000, 3000, 5000, and 8000, respectively for two different versions of the ToPMine algorithm.

max_support_word	New ToPMine without stemming	New ToPMine with stemming
8,000	6,478	5,452
5,000	5,947	5,023
3,000	4,692	4,090
1,000	1,878	1,692
500	932	866

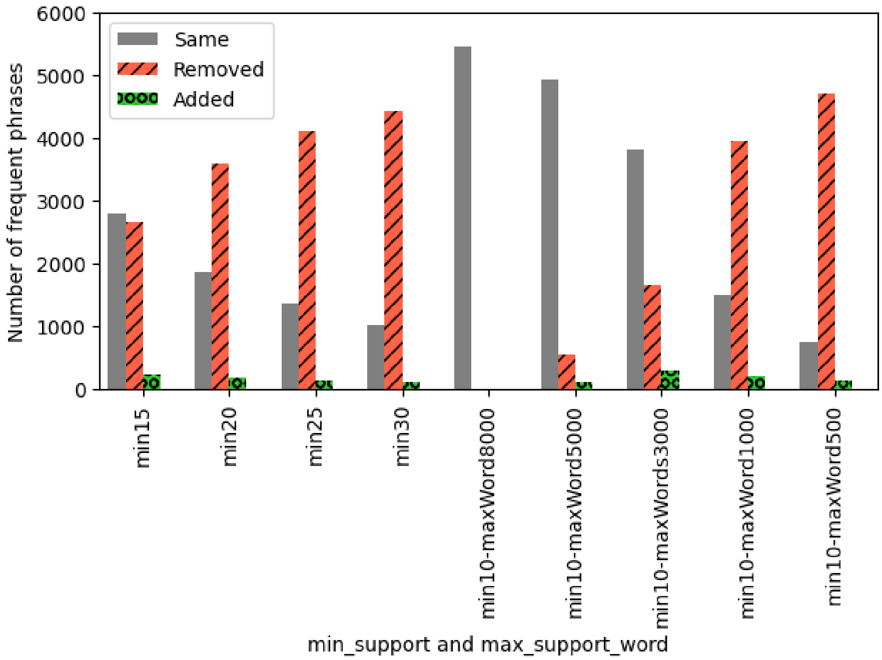


Fig. 3. Comparison of the frequent phrases of new ToPMine algorithm with *min_support* 10 (and *max_support_word* 8000) to settings with *min_support* in 15, 20, 25 and 30, respectively, and settings with *min_support* 10 and *max_support_word* 500, 1000, 3000, 5000, respectively.

we change *min_support*, these would be phrases that are frequent phrases for *min_support* 10, but not for the higher *min_support* in the compared setting. For example 'computational screening' is removed for *min_support* 15. For the cases where we change the *max_support_word*, these would be phrases with words that occur more often than the *max_support_word* in the compared setting. For instance, 'sheet metal forming' contains the word 'metal' with frequency 3457 and would be removed for *max_support_word* 1000. The 'Added' bars show which frequent phrases occur newly in the compared settings. This happens, as stated before, because ToPMine does not generate all frequent phrases, but focuses on high-quality frequent phrases. As an example, 'exchange correlation potential' appears at least 10 times and less than 30 times and 'exchange correlation' appears at least 30 times. Both are frequent phrases for *min_support* 10. However, ToPMine does not generate 'exchange correlation' for *min_support* 10, but it does generate 'exchange correlation potential'. For *min_support* 30 'exchange correlation potential' is not a frequent phrase, while 'exchange correlation' is, and ToPMine does generate 'exchange correlation' as a frequent phrase.

Further, in this paper we also investigate using stemming on the frequent phrases. As an example, the phrases 'molecular dynamics simulations', 'molecular dynamics simulation', 'molecular dynamic simulations' and 'molecular dynamic simulation' have the same stem 'molecular dynam simul'. Stemming allows for removing redundant phrases and thus reduces the work of the domain expert. The influence on the number of generated phrases can be seen by comparing the last two columns in Tables 2 and 3. A disadvantage is that in some cases possible concept candidates may be removed. To alleviate this problem we show the domain expert for each of the stemmed frequent phrases the list of corresponding original phrases. This also helps the domain expert to choose terms to be added to the ontology.

In Table 4, we show the candidate concepts based on the validation of a domain expert on the frequent phrases from the experiment with *min_support* 30 and *max_support_word* 500. In total, 88 candidate concepts are suggested based on 81 out of 131 frequent phrases generated by the experiment. Some candidate concepts can be added into MDO as sub-concepts of existing concepts. For instance, 'Linearized Augmented Plane Wave Method' is a sub-concept of 'Density Functional Theory Method'. Some candidate concepts are relevant to materials design domain but may be not interesting for data access or data integration over materials design databases. For instance, 'Covalent Bond' is a bonding type that can be used to describe materials structures.

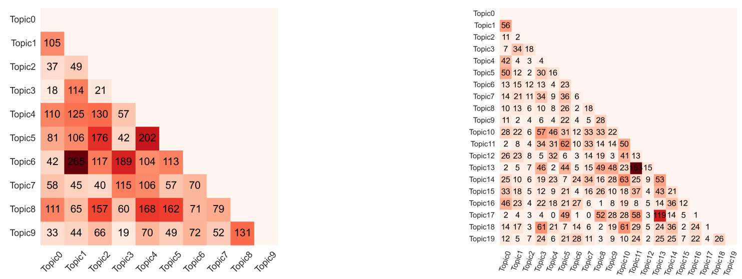
4.3 Topics

Using the frequent phrases, PhraseLDA, a variant of Latent Dirichlet Allocation, is used to generate topics. The number of topics (*num_topic*) is an input parameter to ToPMine. Each topic contains a set of phrases and these sets do not have to be disjoint. For instance, Figure 4 shows the overlap of phrases between topics for different settings of input parameters. In general, when we increase the number of topics, the number of frequent phrases in each topic decreases and the overlap between topics decreases as well.

The domain expert validates these topics and if possible, labels them to generate concepts for the ontology. In Table 5, we show the domain expert validation on 10 topics generated by the New ToPMine with stemming, *min_support* 30 and *max_support_word* 500. Among these topics, there are two topics (topics 0 and 9) that are interpreted with multiples labels, i.e., the domain expert divided the

Table 4. Candidate concepts based on domain expert validation on the experiment with *min_support* 30 and *max_support_word* 500.

Stacking Fault	Stone-wales Defect	Cement Paste
Van der Waals Force	Covalent Bond	Perdew-Burke-Ernzerhof (PBE) Exchange-Correlation Functional
Functionally Graded Material	Symmetric Tilt Grain Boundary Structure	Fatigue Limit
Linearized Augmented Plane Wave Method	Asymmetric Tilt Grain Boundary Structure	Edurance Limit
Face Centered Cubic	Rock Salt Structure	Porous Media
Boron Nitride	Rock Salt	Microstructural Features
Nearest Neighbor	Projector Augmented Wave Method	Hall-Petch Relation
Body Centered Cubic	Iron	Conduction Band
Coarse Grained Model	Cahn–Hilliard Equation	Slip Plane
Fiber Reinforced	Cauchy-Born Rule	Vapor Deposition
Zinc Blende	Domain Wall	Spinodal Decomposition
Core Shell	Armchair	Spontaneous Polarization
Rare Earth	Zigzag	Absorption Spectrum
Refractive Index	Double Walled Nanotube	Charpy Impact Test
Half metallicity	Power Factor	Alkaline Earth Metal
X-ray diffration	Carbon Nanotube (cnt)	Contact Angle
Modified Embedded Atom Method	Mixed Mode Fracture	Vickers Hardness
Unit Cell	Homo-lumo Energy Gap	Rutile Titanium Dioxide (TiO_2)
Absorption Spectra	Stainless Steels	Kinematic Hardening
Glass Formation	Vibrational Modes	Hexagonal Close Packed (hcp)
Brillouin Zone	Domain Switching	Anomalous Hall Effect
Lennard Jones	Sound Velocity	Valence Band
Dispersion Curves	Anatase (TiO_2)	Voight Model
Cohesive Zone Model	Austenitic Stainless Steel	Reuss Model
Quasi-harmonic Debye Model	Crystallographic Orientation	Solute Segregation
Additive Manufacturing	Brittle Transition	Directional Solidification
Real Space Methods	Ductile Transition	Muffin-tin Orbital method
Quasi-harmonic Model	Brittle-Ductile Transition	Muffin-tin Orbital Approximation
Quantum Dot	Modified Becke-Johnson Exchange-Correlation Functional	
Hexagonal Boron Nitride	Kohn-Sham	



(a) *min_support* 10, *num_topic* 10 (b) *min_support* 10, *num_topic* 20

Fig. 4. Number of common phrases between pairs of topics.

topic in different parts. The other topics received one label. Further, representative phrases are given for each topic. The labels and the representative phrases can all lead to new concepts.

Table 5. Topic labelling based on domain expert validation on the experiment with *min_support* 30 and *max_support_word* 500 (Up to five representative phrases are selected for each label).

Topic NO.	Topic Labels	Representative Phrases
0	Computational Method Categories	Linearized Augmented Plane Wave Method
		Hartree-Fock Method
		Perdew-Burke-Ernzerhof (PBE) Exchange-Correlation Functional
		Modified Becke-Johnson Exchange Correlation Functional
		Kohn-Sham
	Materials Properties and Features	Absorption Spectrum
		Refractive Index
		Homo-lumo Energy Gap
		Alkaline Earth Metal
		Dispersion curves
	Electronic Structure Features	Conduction Band
		Valence Band
	Materials Categorizations	Half metallicity
		Rare Earth
	Experimental Method Categories	X-ray Diffraction
	Specific Materials	Zinc Blende
	Applications	Optoelectronic Devices
1	Hardness-related Materials Concepts	Quasi-harmonic Debye Model
		Quasi-harmonic Model
		Rock Salt
		Sound Velocity
		Zinc Blende
2	Materials Strength-related Concepts	Stacking Fault
		Van der Waals Force
		Tension Compression
		Uniaxial Tension
		Symmetric Tilt Grain Boundary Structure
3	Materials Fatigue/Fracture-related Concepts	Functionally Graded Material
		Fiber Reinforced
		Cohesive Zone Model
		Unit Cell
		Cement Paste
4	Materials Synthesis Concepts	Additive Manufacturing
		Vapor Deposition
		Directional Solidification
		Microstructural Features
		Crystallographic Orientations
5	Battery-related Materials Concepts	Ion Batteries
		Anatase (TiO_2)
		Lithium Ion Batteries
		Rutile Titanium Dioxide (TiO_2)
		Boron Nitride
6	Materials Structural Categorizations	Face Centered Cubic
		Body Centered Cubic
		Coarse Grained Model
		Hexagonal Close Packed (hcp)
		Iron
7	Nanotube-related Concepts	Armchair
		Boron Nitride
		Hexagonal Boron Nitride
		Carbon Nanotube (cnt)
		Cross Section
8	Artificial Intelligence-Methods (NO)	Artificial Neural
		Neural Networks
		Open Source
		Degrees Freedom
		Artificial Neural Networks
9	Materials Concepts for Solar-cells	Solar Cells
		Quantum Dots
		Domain Wall
		Power Factor
		Electric Fields
	Materials Magnetism Concepts	Domain Switching
		Anomalous Hall Effect
	Materials Polarization Concepts	Spontaneous Polarization

5 Conclusion

In this paper we started our work on extending MDO using a topic model-based approach that relies on domain experts to validate whether candidate concepts should be added to the ontology. We investigated the influence of different settings on the number of frequent phrases that are generated. This is important as it influences the amount of work for the domain expert. Further, we have shown preliminary results on candidate concepts that are generated in the frequent phrases phase and the topics generation phase.

For future work we continue to validate the results of the different variants and settings of the approach for generating frequent phrases and topics. We will also decide which of the candidate concepts should be added to MDO. Then, we will perform formal concept analysis to produce relations between the added concepts. Further, we will use complementary approaches such as Text2Onto [3] and RepOSE [9] to find more concepts and relations.

As a side effect of the validation work by the domain expert we found that in addition to a validation protocol, it would be valuable for the domain expert if there would be a system that helps the expert, e.g., by recommending validations, by allowing for easy search in the results and by clustering similar results together. Further, the system would allow for easy validation, notify when concepts with the same or similar names already exist in the ontology and generate OWL statements for the ontology extension. Developing such a system is one of our current priorities.

Acknowledgements. This work has been financially supported by the Swedish e-Science Research Centre (SeRC), the Swedish National Graduate School in Computer Science (CUGS), and the Swedish Research Council (Vetenskapsrådet, dnr 2018-04147).

References

1. Ashino, T.: Materials Ontology: An Infrastructure for Exchanging Materials Information and Knowledge. Data Science Journal **9**, 54–61 (2010). https://doi.org/10.2481/dsj.008-041
2. Cheung, K., Drennan, J., Hunter, J.: Towards an Ontology for Data-driven Discovery of New Materials. In: Semantic Scientific Knowledge Integration AAAI/SSS Workshop. pp. 9–14 (2008)
3. Cimiano, P., Völker, J.: Text2Onto. In: Montoyo, A., Muñoz, R., Métais, E. (eds.) Natural Language Processing and Information Systems, 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005, Alicante, Spain, June 15-17, 2005, Proceedings. pp. 227–238 (2005). https://doi.org/10.1007/11428817_21
4. Draxl, C., Scheffler, M.: Nomad: The fair concept for big data-driven materials science. MRS Bulletin **43**(9), 676–682 (2018). https://doi.org/10.1557/mrs.2018.208
5. El-Kishky, A., Song, Y., Wang, C., Voss, C.R., Han, J.: Scalable topical phrase mining from text corpora. Proceedings of the VLDB Endowment **8**(3), 305–316 (2014). https://doi.org/10.14778/2735508.2735519

6. Hastings, J., Jeliazkova, N., Owen, G., Tsiliki, G., Munteanu, C.R., Steinbeck, C., Willighagen, E.: eNanoMapper: harnessing ontologies to enable data integration for nanomaterial risk assessment. Journal of Biomedical Semantics **6**, 10:1–15 (2015). https://doi.org/10.1186/s13326-015-0005-5
7. Horsch, M.T., Niethammer, C., Boccardo, G., Carbone, P., Chiacchiera, S., Chiricotto, M., Elliott, J.D., Lobaskin, V., Neumann, P., Schiffels, P., Seaton, M.A., Todorov, I.T., Vrabec, J., Cavalcanti, W.L.: Semantic interoperability and characterization of data provenance in computational molecular engineering. Journal of Chemical & Engineering Data **65**(3), 1313–1329 (2020). https://doi.org/10.1021/acs.jced.9b00739
8. Lambrix, P.: Completing and debugging ontologies: state of the art and challenges (2020), arXiv:1908.03171
9. Lambrix, P., Ivanova, V.: A unified approach for debugging is-a structure and mappings in networked taxonomies. Journal of Biomedical Semantics **4**, 10 (2013). https://doi.org/10.1186/2041-1480-4-10
10. Li, H., Armiento, R., Lambrix, P.: A method for extending ontologies with application to the materials science domain. Data Science Journal **18**(1) (2019). https://doi.org/10.5334/dsj-2019-050
11. Li, H., Armiento, R., Lambrix, P.: An ontology for the materials design domain. In: Pan, J.Z., Tamma, V.A.M., d'Amato, C., Janowicz, K., Fu, B., Polleres, A., Seneviratne, O., Kagal, L. (eds.) The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II. Lecture Notes in Computer Science, vol. 12507, pp. 212–227. Springer (2020). https://doi.org/10.1007/978-3-030-62466-8_14
12. Sabou, M., Fernandez, M.: Ontology (network) evaluation. In: Suárez-Figueroa, M.C., Gómez-Pérez, A., Motta, E., Gangemi, A. (eds.) Ontology engineering in a networked world, pp. 193–212. Springer (2012). https://doi.org/10.1007/978-3-642-24794-1_9
13. Suárez-Figueroa, M.C., Gómez-Pérez, A., Fernández-López, M.: The NeOn methodology for ontology engineering. In: Suárez-Figueroa, M.C., Gómez-Pérez, A., Motta, E., Gangemi, A. (eds.) Ontology engineering in a networked world, pp. 9–34. Springer (2012). https://doi.org/10.1007/978-3-642-24794-1_2
14. Thomas, D.G., Pappu, R.V., Baker, N.A.: Nanoparticle ontology for cancer nanotechnology research. Journal of Biomedical Informatics **44**(1), 59–74 (2011). https://doi.org/10.1016/j.jbi.2010.03.001
15. Voigt, S., Kalidindi, S.: Materials Graph Ontology. Materials Letters (2021). https://doi.org/10.1016/j.matlet.2021.129836
16. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: The FAIR guiding principles for scientific data management and stewardship. Scientific data **3**, 160018:1–9 (2016). https://doi.org/10.1038/sdata.2016.18

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].

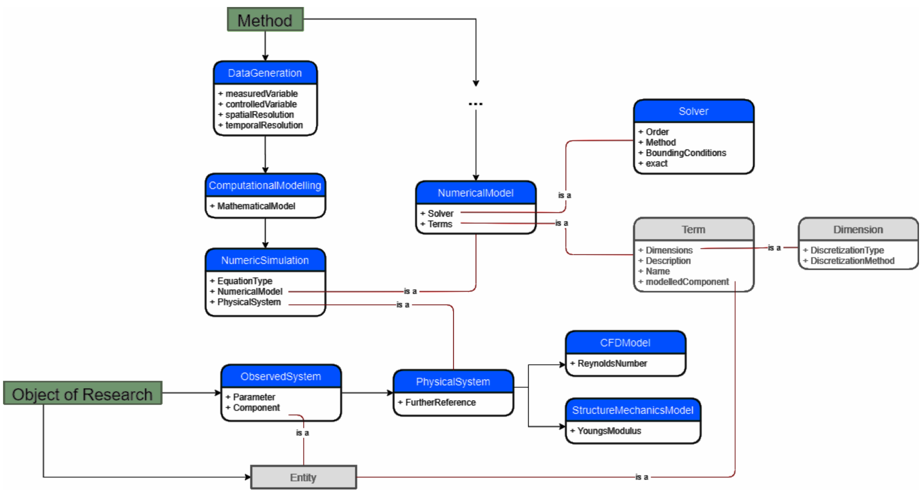


Figure 1: Classes and relations used to describe a CFD simulation.

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 CMLComp [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].

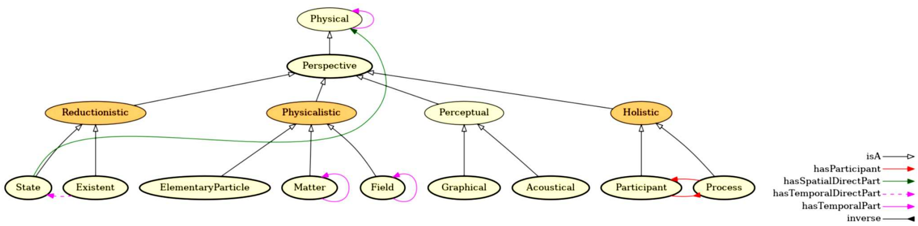


Figure 2: EMMO perspectives (from Ghedini *et al.* [45]).

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.

Introducing MAMBO: Materials And Molecules Basic Ontology

Fabio Le Piane
Istituto per lo Studio dei Materiali Nanostrutturati (ISMN)
Consiglio Nazionale delle Ricerche (CNR)
Bologna, Italy
fabio.lepiane@ismn.cnr.it

Matteo Baldoni
Istituto per lo Studio dei Materiali Nanostrutturati (ISMN)
Consiglio Nazionale delle Ricerche (CNR)
Bologna, Italy
matteo.baldoni@ismn.cnr.it

Mauro Gaspari
Alma Mater Studiorum
Università di Bologna
Bologna, Italy
mauro.gaspari@unibo.it

Francesco Mercuri*
Istituto per lo Studio dei Materiali Nanostrutturati (ISMN)
Consiglio Nazionale delle Ricerche (CNR)
Bologna, Italy
francesco.mercuri@cnr.it

Recent advances in computational and experimental technologies applied to the design and development of novel materials have brought out the need for systematic, rational and efficient methods for the organization of knowledge in the field. In this work, we present the initial steps carried out in the development of MAMBO - an ontology focused on the organization of concepts and knowledge in the field of materials based on molecules and targeted to applications. Our approach is guided by the needs of the communities involved in the development of novel molecular materials with functional properties at the nanoscale. As such, MAMBO aims at bridging the gaps of ongoing efforts in the development of ontologies in the materials science domain. By extending current work in the field, the modular nature of MAMBO also allows straightforward extension of concepts and relations to neighboring domains. Our work is expected to enable the systematic integration of computational and experimental data in specific domains of interest (nanomaterials, molecular materials, organic an polymeric materials, supramolecular and bio-organic systems, etc.). Moreover, MAMBO can be applied to the development of data-driven integrated predictive frameworks for the design of novel materials with tailored functional properties.

Keywords: Ontology; Materials Science; Nanomaterials; Molecular Materials; Knowledge Representation; Machine Learning

1 Introduction

The progress of several relevant fields in science and technology is often related to the design and development of novel materials. Accordingly, advancements in the materials development domain are considered key enablers for innovation in application fields of great technological and socio-economic relevance[1]. Recent developments of data-centric technologies have empowered significant progress in a very broad range of application sectors[2, 3]. As for many other fields, these advancements have had a significant impact also on research and innovation for materials[4, 5, 6].

Submitted to:
DORIC-MM 2021

This trend has also been boosted by the outstanding breakthroughs in multiscale materials modelling and in applied data-intensive approaches[7, 8]. In particular, advances in high-performance and high-throughput computing (HPC/HTC) and data-driven technologies, including artificial intelligence, have further accelerated the process.
The approaches currently pursued by the community involved in the process of design and development of materials leverage the integration of both computational and experimental methods. Computational techniques in the materials development domain cover a broad range of methods and approaches, from electronic structure calculations to continuum (full-scale) simulations[9]. Multi-scale techniques are generally applied to bridge the knowledge about materials across a wide range of spatial and temporal scales. From the experimental point of view, a huge variety of methodologies is commonly used to gather information about materials throughout the development process. These approaches lead typically to a very large amount of unstructured and uncorrelated information on materials.
The tremendous increase of the growth rate of data related to materials development has therefore led to the need for an organization and structuration in the field. Moreover, the application of FAIR (Findable, Accessible, Interoperable, Reusable) concepts to data that are relevant to the domain of materials development will trigger new paradigms[10]. Recent work has demonstrated the relevance of FAIR principles for the automatic retrieval of information or extraction of knowledge from materials data. These efforts are expected to put forward cutting-edge and efficient technologies for using/reusing data on materials and to support the development of data-centric predictive models for innovation[11, 12].
Ontologies constitute a valuable and powerful tool to address the issues related to the organization of knowledge within a given domain. Although still in its early stages, the development and application of ontologies in the materials domain is already displaying the tremendous potential of this approach[13, 14]. In consequence of this emerging interest, recent research and cooperation activities have addressed the development of ontologies targeted to materials. Cooperative efforts in the development of ontologies and semantic technologies for materials are expected to enable the implementation of efficient platforms for the organization of materials data or the realization of complex workflows in the research and innovation process[15]. These steps can be considered key enablers for the digitalisation of the materials development process at several levels.
A significant amount of work focused on the development of top- and middle-level ontologies for materials. The most relevant contribution to the field has been provided by the development of the European Materials Modelling Ontology (EMMO)[16], which is still in progress. Beside that, domain ontologies for specific use cases related to materials also began to be developed[17]. Indeed, the broad scope of the research on materials requires an extensive work covering a manifold of different aspects and knowledge. In this context, recent work has addressed specific application domains, such as simulations of crystalline materials or single-molecule systems[17, 5, 18]. However, attempts in the organization of knowledge focused on materials where aggregation properties at the molecular level are relevant seem to still be lacking.
In this work, we introduce MAMBO - the Materials and Molecules Basic Ontology. MAMBO is focused on concepts and relations emerging on materials where the relationship between individual molecules and molecular aggregation is relevant to the properties of the system. This is the case, for example, of molecular materials, nanomaterials, supramolecular materials, molecular thin-films and other similar systems of interest. These materials play a crucial role in several application fields and technologies, including organic electronics and optoelectronics (OLEDs, organic thin-film transistors), organic and hybrid photovoltaics (organic and perovskite solar cells), bioelectronics (neural and brain interfaces), molecular biomaterials, and several others.
In an essentially modular structure, MAMBO can easily be extended to cover other aspects of the do-

main considered (for example, introducing new computational methods for molecular materials) or to neighboring domains. Concepts and relations introduced in MAMBO could also possibly be used to develop new ontologies for modelling specific tasks, integrating them with top-level ontologies in order to provide full-level interoperability between different applications and computational methods. The development of MAMBO can lead to efficient tools for retrieving and analysing computational and experimental information in the development of materials based on molecular systems. Moreover, MAMBO can provide the basis for the implementation of data-driven technologies and workflows, for example based on machine learning, for the design and development of novel functional materials.

2 Related Work

Previous work has focused on the development of ontologies in the materials domain, focusing on different aspects of the problem and at different levels of details. As stated before, EMMO constitutes one of the most significant efforts in the field. EMMO aims at developing a general ontology for materials and modelling[19]. Other ontologies related to MAMBO address more specialized sub-topics in the domain of materials. ChEBI (Chemical Entities of Biological Interest) is an ontology focused mostly on chemical systems[18] . Despite the very focused scope, several concepts introduced in ChEBI can be related to the domain covered by MAMBO. A very recent effort led to the Materials Design Ontology (MDO), which defines concepts and relations to cover knowledge in the field of materials design and especially in solid-state physics[17]. Moreover, other ontologies developed in the framework of digitalisation and virtualisation can be related to MAMBO. These include OSMO (ontology for simulation, modelling, and optimization), and ontologies developed within the European project VIMMP (Virtual Materials Marketplace Project)[15, 20]. In addition, MAMBO also aims at connecting with other efforts focused on the development of materials databases (OPTIMADE, NOMAD)[21, 22, 12].

3 Typical application scenarios

The application scenarios addressed by MAMBO focus on frameworks occurring in the development of the class of molecular materials and related systems. The main application scenarios we have selected, stemming from potential use cases, are:

- The retrieval of structured information from databases on molecular materials.
- The definition of complex workflows for the modelling of systems based on molecular materials.

These scenarios were defined on the basis of the analysis of current work in the field of ontology development in the materials science domain and discussions with domain experts. For example, MAMBO will enable semantic searches in databases on multi-scale modelling and characterization data on OLEDs[23, 24]. Here, data can include information about the basic chemico-physical entities constituting parts of the active systems (for example molecules, polymers, etc.), aggregated systems and full-scale devices. In other applications, MAMBO can be used to model the steps of a complex computational workflow which addresses a specific scientific/technological problem in the framework of molecular materials. For example, the modelled workflows can lead to the implementation of efficient computational approaches for the screening of properties of a given class of molecular systems. The data obtained by simulations can further be used to implement predictive data-driven models, for example for designing novel materials. The semantic interoperability provided by MAMBO will enable the integration between simulation data, possible integration with data stemming from other sources (experiments, characterization, etc.) and the application of data-centric approaches.

4 Basic principles, development process and methods

MAMBO was designed and developed to address the challenges in material science we have highlighted. We started with in-depth discussions and meetings between the MAMBO development team and domain experts to define a set of possible application scenarios and use cases. These steps led us to the definition of:

- A set of competency questions, that is a set of typical questions for which the information in MAMBO should provide answers.
- A set of typical tasks, which will be supported by the MAMBO ontology.
- A set of typical use cases, as in the examples outlined above.

Due to the peculiar nature of the typical development approaches pursued in the considered application area, we modelled the main concepts of the ontology associating them to specific problem solving methods (PSMs)[25]. PSMs can be used to define operations to be performed to accomplish a given goal, related to a task. According to standard PSMs development approaches, complex tasks, defined by specific use-cases, were decomposed into subtasks. In each task and subtask, the required set of pre- and post-conditions was also defined. This approach helped us to individuate relevant terms and connections between concepts in the considered application scenarios and use cases. A wide set of different techniques was used, aimed at catching relevant terms and concepts on the basis of the textual content of the discussions. In this way, we have individuated an initial set of terms to be used as a ground basis for MAMBO.

We interviewed experts in many different sub-domains of the general fields related to the main MAMBO topics (researchers and professionals working in the field of molecular and nanostructured materials and their applications). The collection of information involved asking the experts a general description of their research work, also identifying the most crucial terms and concepts without which they are unable to describe or define their activities. This step allowed us to identify the main common concepts used and to annotate an initial list of relevant terms. An initial group of 5 experts was initially involved with in-depth interview taking place over the course of several days. This work is still in progress, involving a much larger group of experts. Namely, about 80 experts have been contacted for a survey to be used in the next development stage. From the terms obtained in these first steps, an initial representation of the concepts and relations was drafted. In more detail, a "hybrid" approach (bottom-up and top-down) was used, to better represent the different nature of concepts involved in the development of the MAMBO ontology. A tentative set of relationships among terms was initially built and adjusted iteratively. Further details about the development process of MAMBO will be provided in a future work.

AMB

4.1 Integration with other ontologies

Beside the interaction with domain experts and knowledge engineers, one of the steps considered in the development of MAMBO concerns the possible integration with other ontologies. The integration will also involve reuse of concepts and terms from other ontologies. In particular, the specific domain of application of MAMBO suggests connections with the following ontologies:

- EMMO - a reference ontology with possibile links at the upper level
- ChEBI - connections with MAMBO on concepts related to individual molecules

- MDO - integration between crystalline (typically inorganic) systems and molecular (organic) materials.

To approach this problem, we started a conceptualization process based on the efforts made in the development of these ontologies and, when needed, we partly redefined some of the concepts on the basis of the specific target domain. We stress that, at the current stage of development, we are focusing on the conceptualization of entities and relationships in the domain of interest, trying to capture relevant concepts in a very complex field. The full formalization of MAMBO will follow this step, and will be the subject of a forthcoming paper. Acoordingly, the formalization of the integration with other ontologies and of the reuse in terms of concepts and relationships in MAMBO is still in progress.

5 Realization of MAMBO

To address the design principles outlined above, we started the design of MAMBO with an essentially modular structure in mind. According to this set-up, we defined a set of core concepts and basic relationships, which are connected to other lower-level hierarchies, as it will be shown below.

5.1 Main concepts defined in MAMBO

The basic initial structure of MAMBO provides a generalisation of terms and relationships emerged throughout the development process. The main concepts will be represented by classes in the formalized ontology. In a nutshell, the output of this process is the identification of the MAMBO "entities" (more specifically, materials entities), which are the objects of our investigations and can have a structure and/or a property. Structure and Material Property[1] can therefore be defined as concepts related to the main class Material. Other relevant concepts are Calculation and Measurement. The scheme of the main concepts defined in MAMBO is shown in Fig. 1.

5.2 Drafting the "Structure" class

As an example of the strategy pursued in the development of MAMBO, we briefly discuss the initial draft of the `Structure` class. Some of the concepts and relationships identified are shown in Fig. 2. This schema aims at providing a semantic asset for the organisation of knowledge concerning the structure of systems based on molecular materials. Generally speaking, the Structure class relates to the structural property (in 3D space and time) of an object (Material) that can be measured and/or simulated. In the context of molecular materials, we found it useful to consider a structure as made by "structural entities", which can have different features. A structural entity can be a molecular system (for example, a molecule), a molecular structural unit (for example, a functional group/subgroup), a particle or an atom. On the basis of these definitions and of the analysis performed in the previous steps, we started to define lower-level concepts and attributes. In particular, our effort aims at generalising and extending some of the cases considered in other similar ontologies. For example, the position in space of a molecular system can be defined through the concept of "orientation", which can have a rotation matrix or a quaternion vector as properties defining the actual molecular orientation. The `Structure` class can have properties that are related to the object (simulated or measured) as a whole, for example defining the periodicity of the system. Clearly, in Fig. 2 only a subset of the full set of required concepts and relationships is shown.

[1] It must be noted that `Material Property` should not be intended as a concept taken from particular ontology languages but it is a proprietary definition of MAMBO, opt to represent the concept of "property" in the chemical/materials science realm.

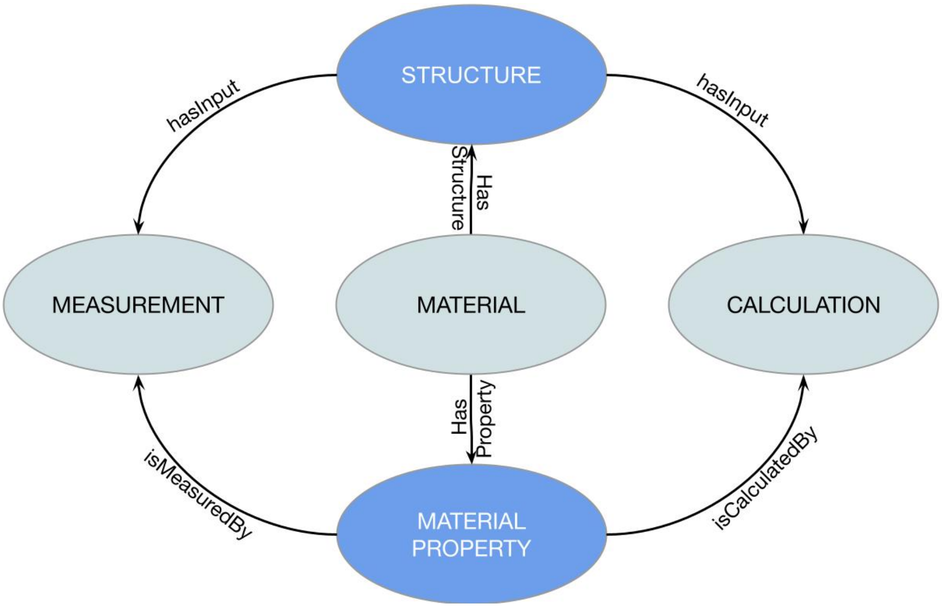


Figure 1: Core concepts of MAMBO: the ontology revolves around the concepts of Material, Calculation and Measurement. An object (Material) is represented by its structural features and properties, while Computational and experimental (Measurement) workflows are connected through a common interface to `Material Property`.

Initial instantiation tests on different use cases provided encouraging results. For example, let us consider the case of a simulation workflow involving a liposome in a water solution. One of the main objects of our investigation will be the liposome structure, which can be considered as a particular shape of a lipid bilayer. Therefore, the liposome will be the instance of the `Structure` class. The particular phospholipid constituting the liposome bilayer (for example, dipalmitoylphosphatidylcholine or DPPC) will be an instance of the `Molecular System` class. A phosphate group is a possible instance of `Structural Unit`. Information about this class can for example be useful in some simulation methods (molecular dynamics, etc.). A phosphorus atom can be an instance of the `Atom` class. In the same example, the water solution surrounding the liposome (and contained within the liposome cavity) can be considered as another instance of the `Structure` class. The analysis of use cases is currently in progress to extend the scope of the `Structure` class and of other classes shown in Fig. 1.

5.3 Drafting the "Property", "Measurement" and "Calculation" classes

We then proceeded defining the bigger picture, focusing in particular on three classes, `Property`, `Measurement` and `Calculation`, investigating their mutual relationships. These three classes are deeply linked: the latter two act as a middle ground that let us treat experimental and computational results in all their specificity, merging then the results into `Property`. This allow us to convey the general characteristics (i.e. both computationally and experimentally accessible) of a system into the more general `Property` class, while keeping more specific attributes in the `Measurement` and `Calculation` classes. In turn, `Measurement` and `Calculation` classes are connected to a specific class for their corresponding method (`Experimental Method` and `Computational Method`, respectively) which represent the set of differ-

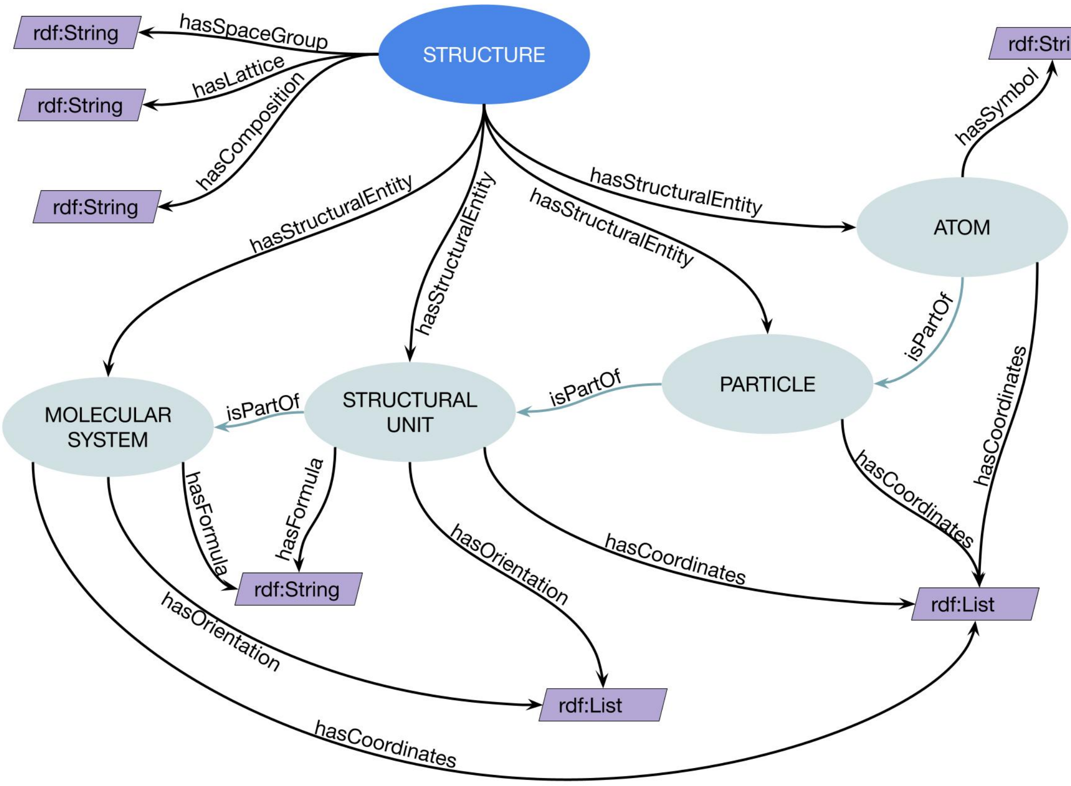


Figure 2: Draft scheme of the Structure class. The main concepts and relationships used in the Structure class are related to the analysis of actual workflows emerging from typical problem solving tasks involving molecular materials. Terms and relationships are connected to both computational and experimental techniques and methods.

ent methodologies and their respective parameters.
These relationships are represented in Fig. 3 and Fig. 4.

5.4 Formalization and implementation procedures

Current steps in the development of MAMBO concern the initial formalisation of relationships between concepts. This strategy is initially applied to the classes shown in Fig. 1, and progressively extended to include a more structured representation. To this end, we used the OWL 2 language[26], while we are evaluating the possibility of re-implementing MAMBO with the OWLReady framework and library[27]. At the moment, we are using the RDF/XML syntax. First of all, we drew the informal representation of a module of the ontology, trying to define the relationships between different concepts and trying to identify the main properties of each class and subclass. This step also involved extracting the main hierarchical relationships between classes, identified according to the hybrid (top-down/bottom-up) approach mentioned above. Once we reached a results that we felt to be consistent, we then shifted to the formal implementation in OWL to see if the proposed schema can be effectively modelled in a formal way. In this initial implementation stage, we applied OWL constructs to define the relationships between classes. Moreover, instantiation tests allowed us to validate the relationships between the defined classes. As

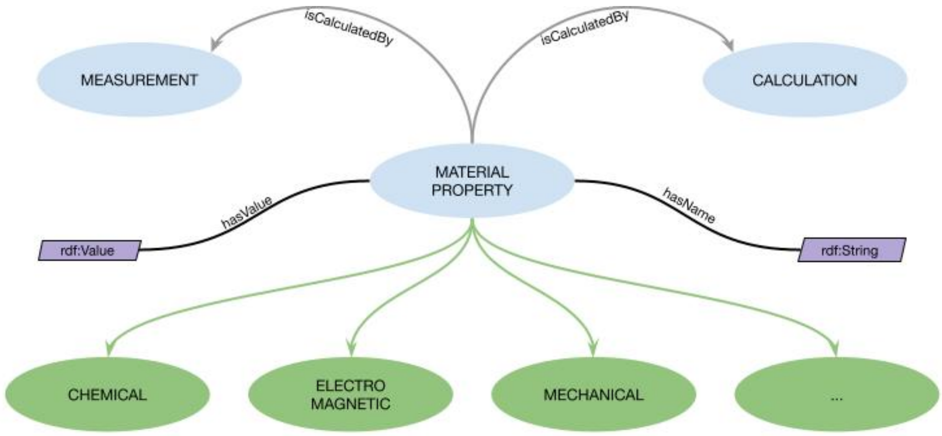


Figure 3: Scheme of the `Material Property` class. This class is mainly described by name and value attributes. The `Material Property` class is also symmetrically linked to `Measurement` and `Calculation` classes.

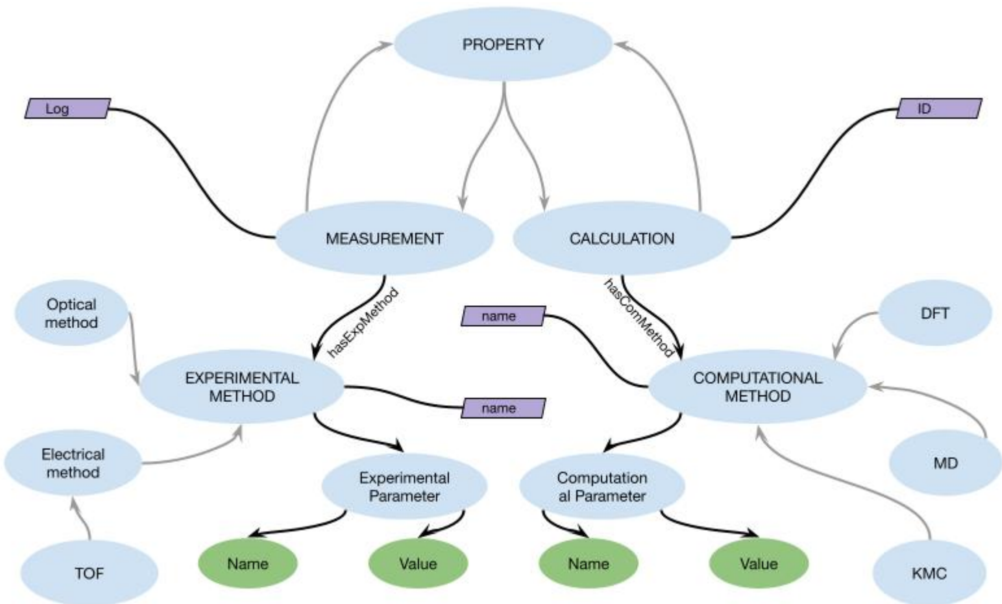


Figure 4: Scheme of the `Measurement` and `Calculation` classes. They both have their respective "method" class (`Experimental Method` and `Computational Method`, respectively) which lead to the different experimental and computational methods, while gathering their parameters.

expected, our reasoning turned out to be slightly imperfect but the conceptual structure of the actual implementation was almost identical to that of the informal scheme. In particular, we used the *WebProtégé* editor[28] for editing the project interactively and collaboratively.

At the time of this writing, we implemented a testing version of the Core and Structure hierarchies shown in Fig. 1 and Fig. 2, respectively. However, we stress that the formal implementation of MAMBO is still at its early stages, and the relationships and classes shown in the aforementioned Figures are susceptible to changes. Also, we are currently setting up a consistent naming standard, able to support and express the semantic relationships involved in the ontology modelling. An initial version of the MAMBO ontology is available on GitHub[2].

6 Future steps

The development of MAMBO is still in progress and will require several additional steps. As mentioned above, we are currently working at the formalization and implementation of the basic structure of the ontology, encoding the main concepts and hierarchies. This initial implementation step will also guide us in defining a set of useful relationships between classes. These relationships will possibly be reused for other hierarchies in MAMBO. This work will also allow us to assess the possibilities offered by different implementation strategies. Another development step will concern the extension of MAMBO to cover specialized domain areas. This extension process will be focused on the classes shown in Fig. 1. One of the ambitions of MAMBO is to organize in a comprehensive formal way the computational and experimental knowledge emerging from research on molecular materials. As such, MAMBO will target a broad range of concepts and relationships in the domain, ranging from multiscale computational methods to experimental characterization of the specific class of materials considered. Tho address this general goal, however, we plan to reuse concepts and terms from other ontologies, defining new relationships targeted to the specific use cases and tasks in a functional way. Finally, we will carry out a thorough evaluation of MAMBO, targeted especially to the assessment of the performance in terms of tasks defined by specific use-cases.

7 Conclusions

In summary, we introduced MAMBO - an ontology focused on the representation of the terminological knowledge relevant for applications and computational processes involving materials based on molecules and similar systems. The knowledge modelled in the proposed ontology relates to both computational and experimental information about molecular materials and related systems, providing a fully interoperable platform. The ambition of MAMBO is also to model a broad range of concepts and relationships of common use in the field. These include for example methods and approaches involved in the multiscale modelling of molecular materials. Treating empirical and computational information on materials on the same footing will also enable a full integration of data. Beside the realization of a platform for the organization of pre-evaluated data, MAMBO will also find application in the definition of computational, experimental or integrated workflows targeted to specific tasks. The approach pursued in the development of MAMBO will allow the extension of the semantic asset towards other similar fields of interest in the domain of molecular materials. Moreover, the concepts and relationships defined within MAMBO can also be structured in the framework of other top-level ontologies. Although still in an early development stage, our initial assessment and instantiation tests demonstrate the potential of MAMBO in the specific field of molecular materials and nanostructures. Work is in progress to implement the formal structure of MAMBO, to extend the scope of classes and to test performance and use in applications.

[2]https://github.com/egolep/MAMBO

References

[1] *Key enabling technologies policy*. URL: https://ec.europa.eu/info/research-and-innovation/research-area/industrial-research-and-innovation/key-enabling-technologies_en.

[2] Weidong Li, Yuchen Liang, and Sheng Wang, eds. *Data Driven Smart Manufacturing Technologies and Applications*. Chambridge: Springer, 2021. ISBN: 978-3-030-66851-8. DOI: https://doi.org/10.1007/978-3-030-66849-5.

[3] S Joe Qin. “Survey on data-driven industrial process monitoring and diagnosis”. In: *Annual Reviews in Control* 36.2 (2012), pp. 220–234. ISSN: 1367-5788. DOI: https://doi.org/10.1016/j.arcontrol.2012.09.004. URL: https://www.sciencedirect.com/science/article/pii/S1367578812000399.

[4] Lauri Himanen et al. “Data-Driven Materials Science: Status, Challenges, and Perspectives”. In: *Advanced Science* 6.21 (2019). ISSN: 21983844. DOI: 10.1002/advs.201900808.

[5] Huanyu Li, Rickard Armiento, and Patrick Lambrix. “A method for extending ontologies with application to the materials science domain”. In: *Data Science Journal* 18.1 (2019), pp. 1–21. ISSN: 16831470. DOI: 10.5334/dsj-2019-050.

[6] Robert Pollice et al. “Data-Driven Strategies for Accelerated Materials Design”. In: *Accounts of Chemical Research* 54.4 (2021), pp. 849–860. DOI: 10.1021/acs.accounts.0c00785. URL: https://doi.org/10.1021/acs.accounts.0c00785.

[7] Ankit Agrawal and Alok Choudhary. “Perspective: Materials informatics and big data: Realization of the ”fourth paradigm” of science in materials science”. In: *APL Materials* 4.5 (2016), pp. 1–10. ISSN: 2166532X. DOI: 10.1063/1.4946894. URL: http://dx.doi.org/10.1063/1.4946894.

[8] Fabio Le Piane, Matteo Baldoni, and Francesco Mercuri. “Predicting the properties of molecular materials: Multiscale simulation workflows meet machine learning”. In: *arXiv* (2020), pp. 1–14. URL: https://arxiv.org/abs/2007.14832.

[9] Lula Rosso and Anne F. de Baas. *What makes a material function? Let me compute the ways... (Short Version)*. Ed. by Anne F. de Baas. 2017, p. 264. ISBN: 9789279265976. URL: http://ec.europa.eu/research/industrial_technologies/pdf/modelling-brochure_en.pdf.

[10] Mark D. Wilkinson et al. “Comment: The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific Data* 3 (2016), pp. 1–9. ISSN: 20524463. DOI: 10.1038/sdata.2016.18.

[11] Gudrun Gygli and Juergen Pleiss. “Simulation Foundry: Automated and F.A.I.R. Molecular Modeling”. In: *Journal of chemical information and modeling* 60.4 (2020), pp. 1922–1927. ISSN: 1549960X. DOI: 10.1021/acs.jcim.0c00018.

[12] Claudia Draxl and Matthias Scheffler. “NOMAD: The FAIR concept for big data-driven materials science”. In: *MRS Bulletin* 43.9 (Sept. 2018), pp. 676–682. ISSN: 08837694. DOI: 10.1557/mrs.2018.208. URL: http://link.springer.com/10.1557/mrs.2018.208.

[13] Toshihiro Ashino. “Materials Ontology: an Infrastructure for Exchanging Materials Information and Knowledge”. In: *Data Science Journal* 9.July (2010), pp. 54–61.

[14] Kwok Cheung, John Drennan, and Jane Hunter. “Towards an ontology for data-driven discovery of new materials”. In: *AAAI Spring Symposium - Technical Report* SS-08-05 (2008), pp. 9–14.

[15] Martin Thomas Horsch et al. “Ontologies for the Virtual Materials Marketplace”. In: *KI - Kunstliche Intelligenz* 34.3 (2020), pp. 423–428. ISSN: 16101987. DOI: 10.1007/s13218-020-00648-9. URL: https://doi.org/10.1007/s13218-020-00648-9.

[16] Martin Thomas Horsch et al. “Reliable and interoperable computational molecular engineering: 2. Semantic interoperability based on the European Materials and Modelling Ontology”. In: *arXiv* (2020), pp. 1–25. ISSN: 23318422. URL: https://arxiv.org/abs/2001.04175.

[17] Huanyu Li, Rickard Armiento, and Patrick Lambrix. “An Ontology for the Materials Design Domain”. In: *The Semantic Web – ISWC 2020*. Ed. by Jeff Z Pan et al. Vol. 12507 LNCS. Cham: Springer International Publishing, 2020, pp. 212–227. ISBN: 9783030624651. DOI: 10.1007/978-3-030-62466-8{_}14. URL: http://dx.doi.org/10.1007/978-3-030-62466-8_14.

[18] Kirill Degtyarenko et al. “ChEBI: A database and ontology for chemical entities of biological interest”. In: *Nucleic Acids Research* 36.SUPPL. 1 (2008), pp. 344–350. ISSN: 03051048. DOI: 10.1093/nar/gkm791.

[19] Emanuele Ghedini and Georg Schmitz. “EMMO the EUROPEAN MATERIALS MODELLING ONTOLOGY”. In: *EMMC Workshop on Interoperability in Materials Modelling* November (2017), pp. 7–8. URL: https://emmc.info/wp-content/uploads/2017/12/EMMC_IntOp2017-Cambridge_Ghedini_Bologna.pdf.

[20] *VIMMP Project Consortium Virtual Materials Marketplace*. 2020. URL: https://www.vimmp.eu.

[21] *The Open Databases Integration for Materials Design (OPTIMADE)*. 2021. URL: https://www.optimade.org.

[22] *The Novel Materials Discovery (NOMAD)*. 2021. URL: https://www.nomad-coe.eu.

[23] Denis Andrienko. “Multiscale Concepts in Simulations of Organic Semiconductors”. In: *Handbook of Materials Modeling: Methods: Theory and Modeling*. Ed. by Wanda Andreoni and Sidney Yip. Cham: Springer International Publishing, 2020, pp. 1431–1442. ISBN: 978-3-319-44677-6. DOI: 10.1007/978-3-319-44677-6{_}39. URL: https://doi.org/10.1007/978-3-319-44677-6_39.

[24] Matteo Baldoni et al. “Spatial and orientational dependence of electron transfer parameters in aggregates of iridium-containing host materials for OLEDs: Coupling constrained density functional theory with molecular dynamics”. In: *Physical Chemistry Chemical Physics* 20.45 (2018), pp. 28393–28399. ISSN: 14639076. DOI: 10.1039/c8cp04618b. URL: https://pubs.rsc.org/en/content/articlehtml/2018/cp/c8cp04618b.

[25] Dieter Fensel et al. “The Unified Problem-Solving Method Development Language UPML”. In: *Knowledge and Information Systems* 5.1 (2003), pp. 83–131. ISSN: 0219-1377. DOI: 10.1007/s10115-002-0074-5.

[26] W3C OWL Working Group. “OWL 2 Web Ontology Language Document Overview”. In: *OWL 2 Web Ontology Language* December (2012), pp. 1–7. URL: http://www.w3.org/TR/owl2-overview/.

[27] Jean Baptiste Lamy. “Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies”. In: *Artificial Intelligence in Medicine* 80 (2017), pp. 11–28. ISSN: 18732860. DOI: 10.1016/j.artmed.2017.07.002.

[28] Tania Tudorache et al. "WebProtégé: A collaborative ontology editor and knowledge acquisition tool for the web". In: *Semantic Web* 4.1 (2013), pp. 89–99. ISSN: 15700844. DOI: 10.3233/SW-2012-0057.

SCONTO: A Modular Ontology for Supply Chain Representation

Marcela Vegetti[1], Alicia Böhm[1], Horacio Leone[1], Gabriela Henning[2]
[1] INGAR (CONICET-UTN), Avellaneda 3657 – Santa Fe, Argentina
[2] INTEC (CONICET-UNL), Ruta Nacional N° 168, Km 0 – Santa Fe, Argentina

Abstract. Supply Chain Management (SCM) involves coordinating and integrating material, information and money flows, both within and across several companies. The information associated with these flows is perceived differently by distinct companies, raising semantics-related problems. Industry 4.0 and the Digital Supply Chain initiative request a seamless integration. This implies achieving technical, syntactic, semantic and organizational interoperability. For several years, ontologies have been considered the key technology to achieve semantic integration. Therefore, this contribution presents SCOPRO, which is the main module of SCONTO, supply chain ontology. SCOPRO is an ontology that formally describes a supply chain (SC) at various abstraction levels, by specifying its associated business processes based on the SCOR de facto standard and by sharing a precise meaning of the information exchanged among the many stakeholders involved in the SC.

Keywords: SCOR model, Supply Chain, Ontology, Interoperability

1. Introduction

To efficiently operate a Supply Chain (SC), all its participants (suppliers, manufacturers, distributors, customers, third and fourth party logistics) must have an enhanced and common understanding of it. Reaching a shared vocabulary allows SC partners to communicate more efficiently, achieving a genuine integration of the activities executed by the different actors. This challenge has motivated several research efforts that addressed the development of models in two different directions. One of them has proposed models that describe the elements and processes associated with a supply chain. The other one has tackled specific SC integration problems.

Despite several models exist, the only de facto standard is the Supply Chain Operations Reference (SCOR) model (Supply-chain Council, 2012). It is a good starting point for communication among SC stakeholders because it provides slender modeling of business processes. However, neither the syntax nor the semantics of SCOR are well defined. In addition, resources and their relationships with processes are not explicitly captured. Therefore, the formalization of SCOR becomes a requirement for a more comprehensive usage of the model (Böhm et al. 2008).

In the last decades, ontologies appear as a tool to reach a semantic agreement. According to Gruber (1993), an ontology is a formal, explicit specification of a shared conceptualization. It captures the knowledge accepted by a community about a specific domain. In general, ontologies are expressed using not ambiguous and formal languages. Such formality allows a computer to interpret them and infer new knowledge.

After the rise of the Semantic Web, several SC ontologies implemented with these technologies appeared (Grubic and Fan, 2010). Some use linked data principles to represent traceability-specific domain knowledge in supply chains (Solanki and Brewster, 2016), as well as to improve and facilitate knowledge management among SC partners (Rodríguez-Enríquez et al., 2015). Other proposals formalize the SCOR model using OWL (Ontology Web Language), but in a partial fashion. Ontologies have been developed to support the knowledge management of supply chain operation (Zdravkovic et al., 2010), the alignment of strategic knowledge (Sakka et al., 2011), the modeling of processes (Grubic et al., 2011), or the simulation of SC processes (Fayez et al., 2005). In addition, there are proposals with the aim of describing SC partners and their relationships (Chi 2010), managing product data flows in the SC (Lu et al., 2013), and providing a core vocabulary for logistics. To infer or validate knowledge, some authors add rules and queries using SWRL (Semantic Web Rule Language) and SPARQL to their ontologies. Leukel and Sugumaran (2013) focus on the development of SWRL rules for the correct construction of threads, based on the articulation of SCOR's "Make", "Source" and "Deliver" processes. In turn, Petersen et al. (2016) propose the SCORVoc lightweight ontology, which provides a set of SPARQL queries that enable the evaluation of metrics and key performance indicators (KPIs) defined by SCOR.

In conclusion, several contributions whose aim is SC modeling exist, but none is complete and expressive enough to represent all the SC related concepts, such as its structure, involved organizations, business processes, resources and the roles they play in each process, as well as SC performance measurement and benchmarking notions. In addition, the existing proposals lack a semantic well-defined domain vocabulary. In consequence, there is a need to provide more comprehensive and formal definitions of the SC domain and its performance measurement/benchmarking concepts. This will lead to a common understanding of the field, as well as a proper interpretation of the shared information.

This proposal aims at contributing towards the formalization of the SC domain. The goal of this paper is to present SCOPRO, a SC sub-ontology, which is part of SCONTO, a comprehensive ontology based on the SCOR model. In the next section, SCONTO is concisely described to give place to a more detailed presentation of SCOPRO. Section 3 briefly illustrates SCOPRO´s application to a case study and Section 4 describes its evaluation from structural, content, consistency, and complexity perspectives. In section 5, conclusions are drawn.

2. Supply Chain Ontology (SCONTO)

The Supply Chain ONTOlogy (SCONTO) aims at describing supply chains having different features through a generic vocabulary. SCONTO is organized in three complementary sub-ontologies: SC processes (SCOPRO), performance evaluation (SCOBE) and benchmarking (SCOME), which are shown in Fig. 1. Despite its generality, SCONTO can be specialized to include more specific concepts. A similar approach has been followed regarding SC performance evaluation and benchmarking. The proposed ontology specifies the necessary terms to define an evaluation system that provides a common understanding to all the SC participants. Although SCONTO is defined in terms of the three modules mentioned above, this article focuses on the first one.

SCOPRO specifies the main concepts that are needed to capture the essence of a supply chain by formalizing and extending the SCOR reference model (Supply Chain Council, 2012). SCOPRO includes the basic terms that represent the SC structure, its processes and associated resources, the resource roles, as well as the relations among these concepts. In addition, SCOPRO makes explicit the organizational units that participate in a SC, their organizational components and roles, as well as the way these units are linked to processes.

An ad-hoc methodology based on well accepted principles has been proposed for the development of SCONTO. It has the following main stages:

1. *Requirements specification*; this stage identifies the scope and purpose of the ontology.
2. *Conceptualization;* which organizes and converts an informally perceived view of the domain into a semi-formal specification using UML diagrams and OCL specifications.
3. *Implementation;* stage that implies the codification of the ontology using a formal language.
4. *Evaluation,* step at which a technical judgment of the ontology quality and usefulness with respect to the requirements specification, competency questions, and/or the real world is made.

It is worth mentioning that these stages are not truly sequential. In fact, ontology development is an iterative and incremental process. If deficiencies are detected at any stage in the process, it is possible to return to any of the previous steps to make modifications and/or refinements.

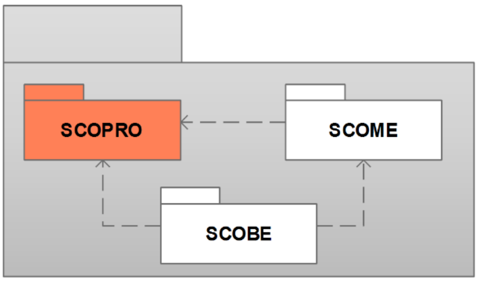


Fig. 1. SCONTO organization

2.1 Supply Chain Process Ontology

As already mentioned, SCOPRO integrates concepts representing the SC structure, as well as the business processes and resources involved in its operation. This ontology is presented in the following paragraphs using textual definitions, UML diagrams, and tables. The tables introduce concepts, relationships, and constraints that are expressed in natural language. Due to space limitations, the OCL specifications (OMG, 2014) of the textual constraints are not included in this article. The conceptualization of SCOPRO could be implemented in several languages. In particular, an OWL2 (W3C, 2004) implementation was made, which can be found in https://industrialonto.github.io/SCOPRO/OnToology/SCOPRO.owl/documentation/index-en.html.

Figure 2 presents an UML class diagram showing the SCOPRO main concepts, which are organized around three perspectives, labeled as structure, process and resource dimensions. The model includes the "SC Entity" concept, which is a generic and abstract notion representing every entity of a SC.

2.1.1. SCOPRO - Structure Dimension

This perspective comprises the terms that are identified when conceiving the SC as an extended organizational network, including the roles that the organizations play in it. The main concepts belonging to this dimension are: *Supply Chain*, *Organizational Unit*, *Organizational Unit Role* and *Functional Area*.

Definition 1: An Organizational Unit (OU) represents the enterprises or enterprise components that participate in the SC operation.

Each *Organizational Unit* can take part in more than one supply chain playing different roles (producer, supplier, distributor, etc.). Therefore, SCOPRO represents the role that is played by an OU in a specific SC through the concept *Organizational Unit Role*. Optionally, an OU can be related to one or more *Resource Role* and to one or more Functional Areas by the associations that are shown in Table 1. Regarding this table and the following ones, it is worth noting that each association is only included in the table of one of its end concepts. In addition, the constraints appearing in some of the tables are described in natural language.

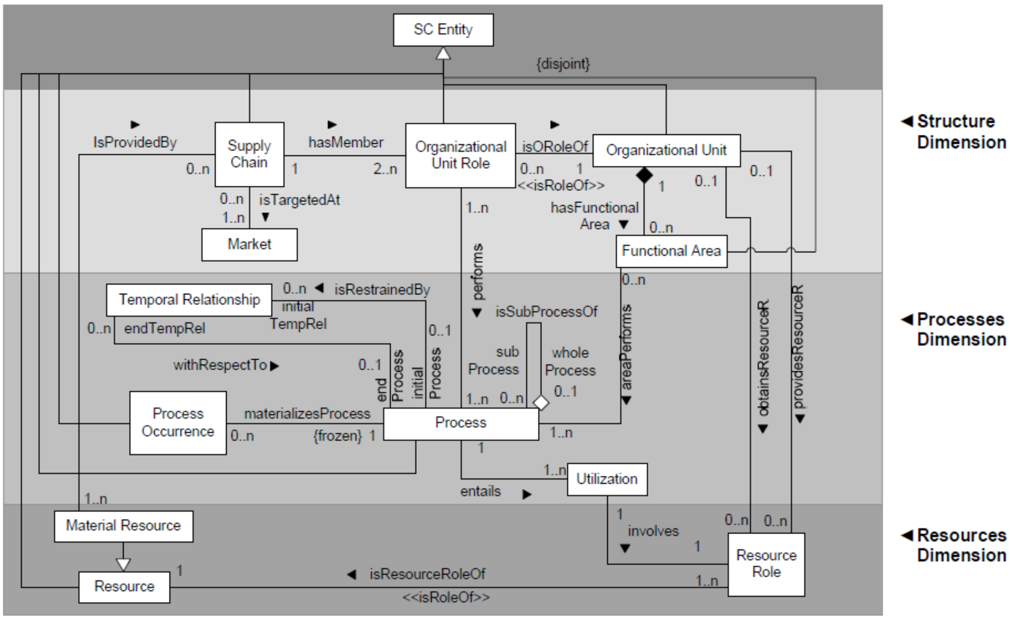


Fig. 2. SCOPRO main concepts

In order to specify the internal organization of the SC participants, the OU class is specialized into:

- *Dependent Organizational Unit*: is an OU that is a subordinate of another one, without losing its administration and functions. This concept is further specialized into any business entity, branch, or subsidiary of an enterprise or company.
- *Independent Organizational Unit:* is an autonomous OU to which belong, if any, all divisions, subsidiaries, branches, or other organizational components of a business. An *Independent OU* may be a company, a society, or an enterprise.

Table 1. *Organizational Unit* class relationships

Relationships	
hasFunctionalArea	It is optional. If exists, it relates an OU with one or more Functional Areas.
obtainsResourceR	It is optional. It specifies which organizational unit has a resource available once the resource participation in a process ends
providesResourceR	It is optional. It indicates the organizational unit responsible for making available a resource for its use in a process

Definition 2. An *Organizational Unit Role* represents the functionality assumed by an organizational unit when participating in a given supply chain. The *isRoleOf* association, which is shown in Fig. 2 and Table 2, between *Organizational Unit* and *Organizational Unit Role* classes agrees with the *isRole* generic relationship defined by Olivé (2007). In consequence, an instance of *Organizational Unit Role* is always the role of the same individual belonging to *Organizational Unit.*

In addition, SCOPRO specifies the following role types:

- *Primary Producer Role*: it is assumed by an organizational unit that carries out activities related to the primary production, such as agriculture, livestock, fishing, mining, fossil fuels extraction, etc.
- *Secondary Producer Role*: role that is played by an organizational unit when performing activities where raw materials or intermediate products are transformed into higher-value products.

Customer Role: a role that is performed by an organizational unit that carries out activities related to the purchase of goods for use, consumption or exploitation.

- *Service Provider Role:* a role that is played by an organizational unit that only performs activities related to transportation, marketing, storage, etc., that add value to a product without transforming it in a physical and/or chemical way.

The functional organization of work within organizational units is captured through the *Functional Area* and *Functional SubArea* concepts.

Table 2. *Organizational Unit Role* class relationships and constraints

Relationships	
isRoleOf	Relates an *Organizational Unit* with one or more roles that the OU plays within a specific Supply Chain
performs	Links an *Organizational Unit Role* with one or more Processes that the OU carries out.
Constraints	
If an *Organizational Unit* performs through a certain role some part of a complex process, it will also execute such complex process	
All the roles that carry out a process have to belong to the same supply chain to which such process belongs.	
An OU playing a *Primary Producer* or *Secondary Producer* role can take part in any business process except in the *Deliver Retail Product* one.	
An OU playing the *Service Provider* role can perform any process except the *Make process*.	
Every organizational Unit playing as a Customer cannot participate in the *Make*, *Deliver* and *Deliver Return business processes.*	

Definition 3. The *Functional Area* class captures the function-based work division that often occurs within organizations. This concept is further specialized into Purchasing, Production. Logistics, Marketing, Sales, R&D and Finance, based on the functions defined by Lambert (2008a). Table 3 presents the relationships in which this concept participates and its restrictions.

Table 3. *Functional Area* class relationships and constraints

Relationships	
areaPerforms	Represents the link between a functional area of an organization and the SC process in which this area participates.
hasFunctionalArea	Specifies the link between a Functional Area with one or more Functional Subareas.
Constraints	
The range of an *areaPerforms* relationship is restricted to instances of the *Process Element* class.	
If a functional area belonging to an organizational unit executes a *Process Element*, then an Organizational Unit Role associated with such Process Element has to exist.	

Definition 4. A *Functional SubArea* is a grouping of activities that are part of a *Functional Area* and that deals with a type of common tasks within the functional area of which they are part. Table 4 presents the relationships in which the *Functional SubArea* concept participates and its restrictions.

Table 4. *Functional SubArea* class relationships and constraints

Relationships	
subAreaPerforms	Specifies the link between a *Functional SubArea* of an OU and one or more *Tasks* belonging to the process in which this subarea participates.
Constraints	
If a functional subarea belonging to an OU performs a task, then this task has to be linked to some role of such OU.	

Definition 5. The *Supply Chain* class represents a network of *Organizational Units* (OUs) that transforms or adds value to materials, ranging from raw materials sourcing, to the final product distribution in specific markets. The relations in which this class participates and its constraints are listed in Table 5.

Table 5. *Supply Chain* class relationships and constraints

Relationships	
isTargetedAt	Links a *Supply Chain* with at least one *Market*.
isProvidedBy	Relates a *Supply Chain* with at least one *Material Resource*
hasMember	Connects a *Supply Chain* with two or more organizational units playing a specific role
Constraints	
Each SC must have at least a member playing the primary producer or secondary producer role	

Definition 6. A *Market* is a set of actual and potential buyers that are grouped together because they share certain distinctive characteristics.

2.1.2. SCOPRO - Process Dimension

This dimension includes the concepts that are needed for a detailed description of processes. In particular, the terms required for defining processes and their sub-processes, the temporal relations among them, the resource participation in processes and the occurrences of the defined processes. Figure 3 presents the main classes of this dimension. In this figure the classes belonging to other dimensions are identified using their corresponding names.

Definition 7. A *process* represents an activity chain that is carried out in a SC to achieve certain results. A process execution implies the creation, modification, use or movement of several resources, which may be physical or conceptual ones.

In order to properly describe a SC, the business processes belonging to it should be modeled at different levels of abstraction. SCOPRO represents the process decomposition into more specific activities using the *isSubProcessOf* association., which is shown in Table 6. This table also presents the *materialize Process* relationship, which links a *Process* with one or more *Process Occurrences.*

Definition 8. The *Process Occurrence* concept represents a particular execution of a process in a given period.

Based on the SCOR model (Supply Chain Council, 2012), SCOPRO considers three different types of processes that are needed to model and analyze a SC. The three concepts that specialize the *Process* class are introduced in Fig. 4, and are labeled as *Business Process*, *Process Element* and *Task*.

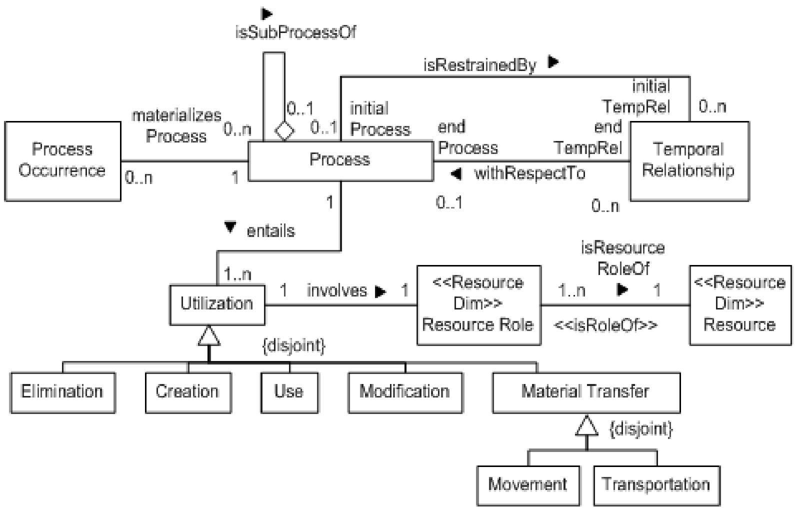


Fig. 3. Main classes of the *Process* Dimension

Definition 9. A *Business Process* is a type of process that is composed of value-added activities. It is designed for achieving a result having a significant impact on clients and in the efficient management of SC flows. Each *Business Process* is composed of *Process Elements.*

Table 6. *Process* class relationships and constraints

Relationships	
isSubProcessOf	This association represents a *Process* decomposition into more specific activities.
materialize Process	It connects a *Process* with one or more of its occurrences. The semantics of this relation is the same one proposed by Olivé (2007) for the materialize association.

isRestrainedBy	This association represents a temporal relationship between a process and another one that restrains its start time.
withRespectTo	It represents a temporal relationship between a process and another one that imposes a limit on its end time.
Constraints	
If two processes are linked, then there is only one temporal relationship (atomic or composite) that connects them.	
Temporal relationships can be defined between business processes, between process elements or between tasks, but never between processes of different type.	

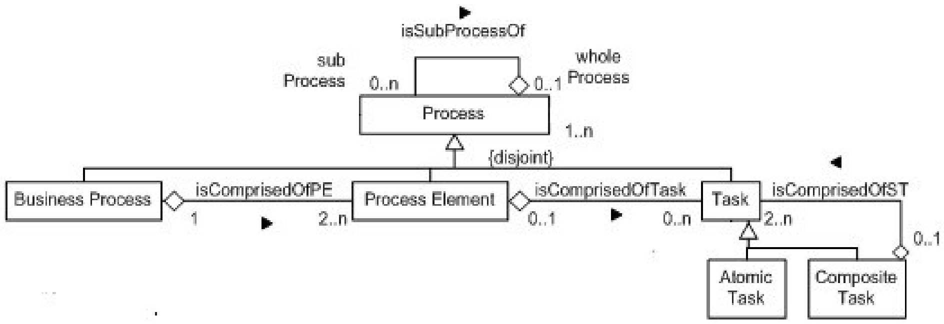


Fig. 4. SCOPRO - Process Dimension. Specialization of the *Process* concept

Definition 10. A *Process Element* is an activity or a logical structure of activities that is part of a *Business Process*. The detail level that is provided by the *Process Element* class allows the decomposition of a *Business Process* into specific operations but, at the same time, is general enough to describe activities that are valid for different types of supply chains.

Definition 11. A *Task* is an activity or a logical structure of activities that is part of a *Process Element* in a certain SC. For example, in a particular supply chain the process element labeled as "Schedule Product Deliveries" is decomposed into the following tasks: "Send material orders", "Coordinate delivery place and date." However, in other SC, in which the provider is responsible for monitoring and maintaining the client stock, the same process element is broken down into different tasks, like "Monitor stock inventory", "Send information to client" and "Propose delivery date and terms".

Table 7 presents the relationships associated with *Business Process, Process Element* and *Task*, as well as their constraints.

Table 7. Relationships and constraints of the *Business Process*, *Process Element* and *Task* classes

Relationships	
isComprisedOfPE	It relates a *Business Process* with the *Process Elements* belonging to it.
isComprisedOfTask	It links a *Process Element* with its *Tasks.*
isComprisedOfST	It connects a *Composite Task* with one or more *Tasks* that are part of it.
Constraints	
Each *Business Process* is only composed of *Process Elements*	
Each *Process Element* is only composed of *Tasks*	
Each *Task* must be part of a *Process Element* or a *Composite Task*, but not both at the same time.	
An *Atomic Task* cannot be further decomposed	

SCOPRO extends the *Business Process* and *Process Element* concepts using the vocabulary of the SCOR reference model (Supply Chain Council, 2010). This proposal recognizes as the main business processes of a SC the following ones: *Sources*, *Make* and *Deliver,* as well as the activities of material returns to providers (*Source Return*) and from clients (*Deliver Return*). The model also includes the planning of the operational activities related to material transformations and movements (*Plan*). As seen in Fig. 5, these processes are specializations of the *Business Process* class.

The SCOR reference model also includes a classification of the activities belonging to business processes. In SCOPRO, such activities specialize the *Process Element* class. As already mentioned, *Organizational Units* participate in supply chain processes playing several roles, which can be *Primary Producer*, *Secondary Producer*, *Service Provider* or *Client*. Each role implies that the organizational unit playing it performs a specific set of activities in a particular SC. If an *Organizational Unit* performs a *Process Element* or *Task*, then such OU also executes the *Business Process* from which the *Process Element* or *Task* belongs to. Therefore, constraining which business processes may be performed when adopting a specific organizational unit role, also restricts the process elements and tasks that such organizational unit can execute.

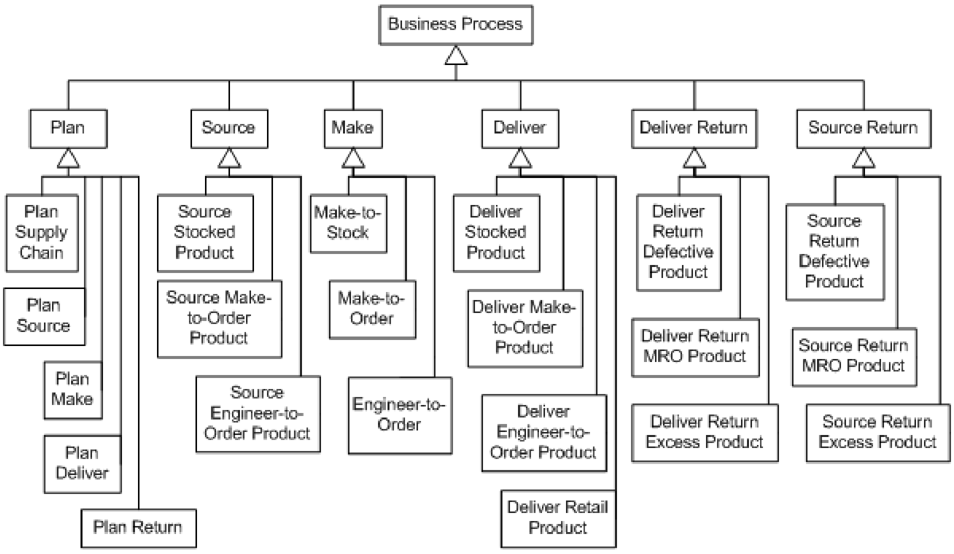


Fig. 5. *Business Process* class specialization

As seen in Fig. 3, SCOPRO introduces the *Temporal Relationship* class, which explicitly represents temporal links between two processes. In the following paragraphs, a refinement of such a concept is presented.

Definition 12. *Temporal Relationship* represents different constraints related to the partial order between the executions of two processes in a SC. To specify this class, this proposal uses the conceptualization developed by Allen (1983) that considers the time intervals in which the processes or activities are performed and the relations between them. Figure 6 introduces the different partial order relations associated with the execution of two processes.

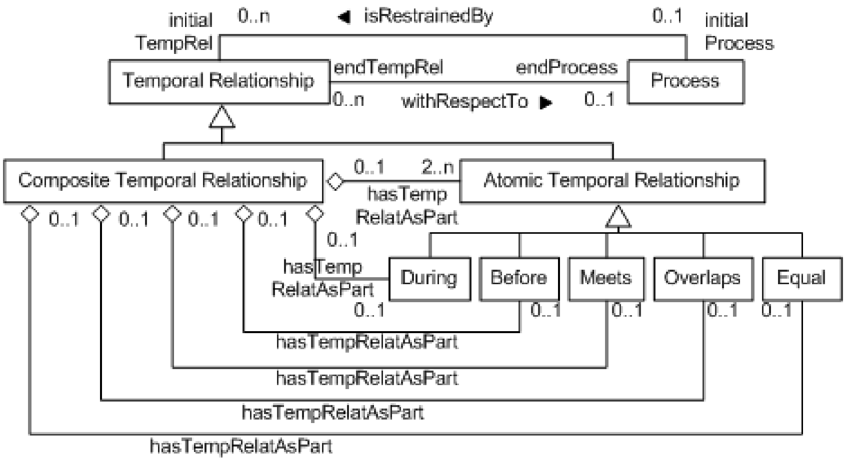


Fig. 6. *Temporal Relationships* in SCOPRO

As shown in Fig. 6, SCOPRO considers two types of temporal relation: atomic and composite ones. The former (Allen, 1983) includes the following subtypes:

- *Before than*: if the P1 process is executed before than P2, the time interval in which P1 takes place is previous to the P2 time interval. In other words, the end of P1 is prior to the P2 start time.
- *Meets:* if P1 meets P2, then the time interval in which P1 is executed ends at the same time the P2 activity begins. The end of P1 is equal to the start time of P2.

- *Overlaps*: If P1 overlaps P2, the start time of P1 takes place before the beginning of P2, but P1 ends while activity P2 is still executing. Therefore, the start time of P1 is prior to the beginning of P2 and the end time of P1 takes between the start and the end time of P2.
- *Equals:* If P1 equals P2, the time interval at which P1 is performed is the same as the one associated with the P2 activity. In other words, the start and end times of P1 and P2 coincide.
- *During*: If P1 is performed during P2, P1 begins after P2 starts and ends before P2 finishes.

The *Composite Temporal Relationship* combines different atomic temporal relationships as disjunctions. As seen in Fig. 6, a composite temporal relationship involves at least two atomic temporal relationships as parts.

Definition 13. The Utilization concept represents the way in which a process affects a resource that participates in a given component (subprocess, activity) of such process: *Creation*, *Elimination*, *Modification*, *Use,* and *Material Transfer* (See Fig. 3). This last type of utilization may occur inside a facility or between different ones. SCOPRO employs the *Movement* and *Transportation* classes, respectively, to represent these material transfer types. The *Transportation* class represents a material flow between two different geographic points. Table 8 presents the relationships associated with the *Utilization* class.

Table 8. *Utilization* class relationships

Relationships	
entails	This relation links a *Process* with one or more *Utilization* types
Involves	It associates a *Utilization* type with the role that a *Resource* plays in such utilization.

2.1.3. SCOPRO Resource Dimension

This dimension specifies the resources and the role they play when participating in processes. Figure 7 illustrates these concepts and their associations.

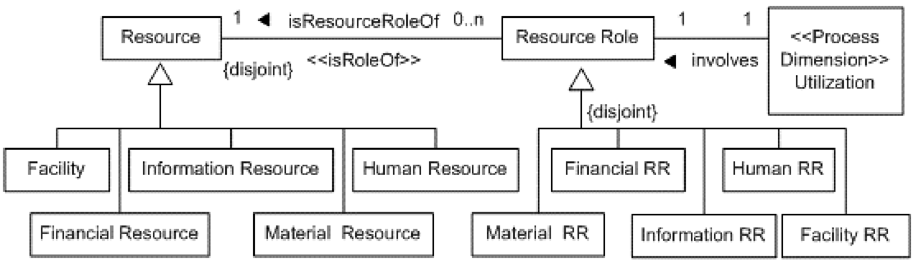


Fig. 7 Specialization of *Resource* and *Resource Role* classes

Definition 14. A Resource can be any type of physical or conceptual medium that participates in a process via one of its roles. This class represents resources that flow through the network (e.g., commercialized goods) and those that remain static (e.g., industrial plants). Supply Chain resources include buildings or facilities, material handling equipment, different types of products, information and financial resources, among others. Therefore, SCOPRO specializes the *Resource* class into the *Material Resource*, *Information Resource*, *Financial Resource*, *Human Resource* and *Facility* classes, which are further specialized.

Definition 15. A *Resource Role* reflects the participation type that a resource has in a given process. The same resource can play different roles in distinct processes. For example, a drill can be the final product of a production process and can be a tool in another one. SCOPRO specializes *Resource Role* into subclasses to represent more specific roles, and these subclasses are also further specialized. For example, *Payment* and *Proceeds* are a specialization of *Financial RR.* Similarly, *Responsible*, *Supervisor* and *Executor* are s*ubclasses of Human RR.* Table 9 presents the relationships associated with the *Resource Role* class, as well as its constraints.

Table 9. *ResourceRole* class relationships and constraints

Relationships	
providesResourcesR	It states which is the OU responsible for making available a *Resource* for its use in a *Process*
obtainsResourceR	It specifies the OU that has a *Resource* available after the end of the resource participation in a given process.
isResourceRoleOf	It links a *Resource* with its *Resource Roles*. The stereotype <<isRoleOf>> in the association constraints that an instance of *Resource Role* is always connected to the same instance of

	Resource. In contrast, an instance of *Resource* can be linked to several instances of the *Resource Role* class.
Constraints	
SCONTO includes several specifications that state which are the roles that the different resources can play, and which are the utilizations that may be associated with them. Due to space limitations, it is not possible to describe all of these constraints in this table.	

3. Case Study

This section illustrates the application of SCOPRO to represent an orange juice supply chain. This SC comprises three farms located in the Corrientes province of Argentina, which produce and pack oranges; the *CITRIX Company* that is the orange juice manufacturer, and the *ArgenTruck* company, which takes care of the material transportation between SC partners. Figure 8 illustrates this value chain in a very simplified way.

The *CITRIX Company* owns two production plants, the *BellaVista* plant *(FCojPlant)* and the *NaranUp* plant *(RojPlant)*, which are managed as separate business units. The *FcojPlant*, which is also located in Corrientes, manufactures frozen concentrated orange juice (*FCoj*). In turn, the *RojPlant,* sited nearby Buenos Aires, produces and bottles reconstituted orange juice (*Roj*). Finally, the *ArgenTruck* company provides transportation to and from the various farms, plants, resellers and retail stores.

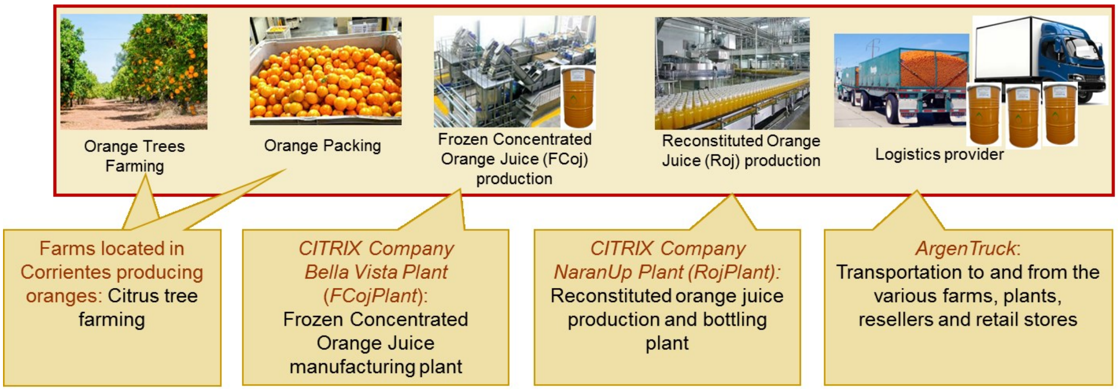


Fig. 8. Orange Juice Value Chain

Due to space limitations, only a small portion of the case study is addressed in this contribution. The delivery process associated with the distribution of frozen concentrated orange juice (*FCoj*) from the *Bella Vista* plant to the *NaranUp* one is represented. This process, which is carried out by both the *CITRIX Company* and *ArgenTruck*, is modelled through the *DFCoj* (Deliver frozen concentrated orange juice) business process, an instantiation of the *Deliver Make-to-Order Product* process that was presented in Fig. 5.

Figure 9 shows the decomposition of *DFCoj*, which is performed by the *CITRIXCompany Independent Organizational Unit* through its *FCoj production Primary Producer Role*. Similarly, *ArgenTruck* also participates through its *3PL Provider Role*. The figure depicts the seven *Process Elements* comprising the *DFCoj* business process; for instance, the *:Receive, Enter & Validate Order* one that is performed by the *CITRIX Company* and the *:Ship Product* process element, which is performed by *ArgenTruck*. The model also shows the various temporal relationships (e.g., *:Before*, *:Meets*) among the different process elements.

The *DFCoj* business process takes place because there is a sourcing need at the *NaranUp* plant, which employs frozen concentrated orange juice as one of its raw materials. In fact, there is a *SFCoj* (Source frozen concentrated orange juice) business process that is an instantiation of *Source Make-to-Order Product* (See Fig. 5), which is carried out by *NaranUp*. Both, the *SFCoj* and *DFCoj* processes represent, respectively, the customer and the provider views of the same business process. The *SFCoj* sourcing process includes ordering the frozen concentrated orange juice from the supplier, receiving the lot and transferring it to an area of the raw materials deposit of the *RojPlant*. In turn, *DFCoj* corresponds to the supplier vision for the same process, comprising the reception, management, and fulfilment of the frozen concentrated orange juice procurement order, as well as the product delivery to the *RojPlant.*

The *SFCoj* and *DFCoj* processes are shown side by side in Fig. 10, which captures the strong interactions that exist between these two processes. This figure presents the process elements belonging to both processes and the temporal relationships linking them. In addition, the model displays the resources that participate in each process element, the roles that the different resources play, and the way each process elements affects a given resource. For instance, it can be seen that the *:Procurement Order* resource participates with its *output* role in the *:Schedule Product Deliveries* process element, which is part of *SFCoj*. This process element affects the resource by creating it (*:Creation* in Fig. 10). It is important to note that the same resource also partakes through its *input* role in the

:Receive, Enter & Validate Order process element that comprises the *DFCoj* business process. However, in this case the *:Receive, Enter & Validate Order* process element uses the resource (*:Use* in Fig. 10).

The model depicted in Fig. 10 represents all types of resources that take part in the orange juice SC. For instance, it includes a portion of the flow of materials that occurs in such value chain. In particular, it shows the transport (*:Transportation* linked to the *:Ship Product* process element in Fig. 10) of the *FCoj material resource* from the *FCojPlant* to its destination, which is the *RojPlant*. It is also shown that this material resource plays the role of a load (*:Load* in Fig. 10) with regards to such transportation.

From the above description, it is possible to appreciate the close link that exists between certain supply chain business processes involving different organizations. The modeling of the process interactions is an essential step to achieve the semantic interoperability of the SC information, the design of information systems that support SC collaborative management, as well as the generation of information systems supporting materials traceability and allowing a comprehensive visibility of the SC information. In addition, this small example allows us to appreciate the important extensions that have been made on the SCOR model.

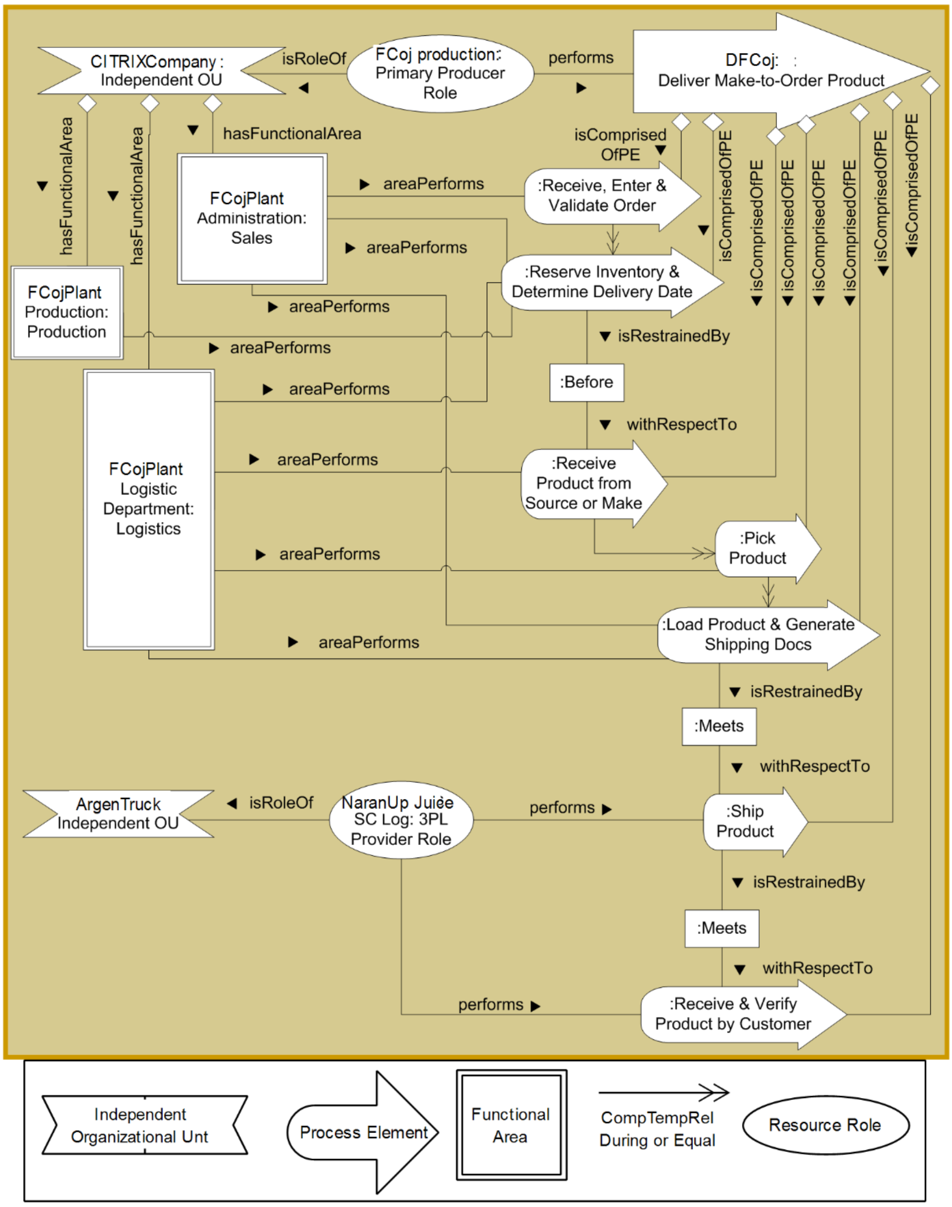


Fig. 9. Decomposition of the *DFCoj* Business Process

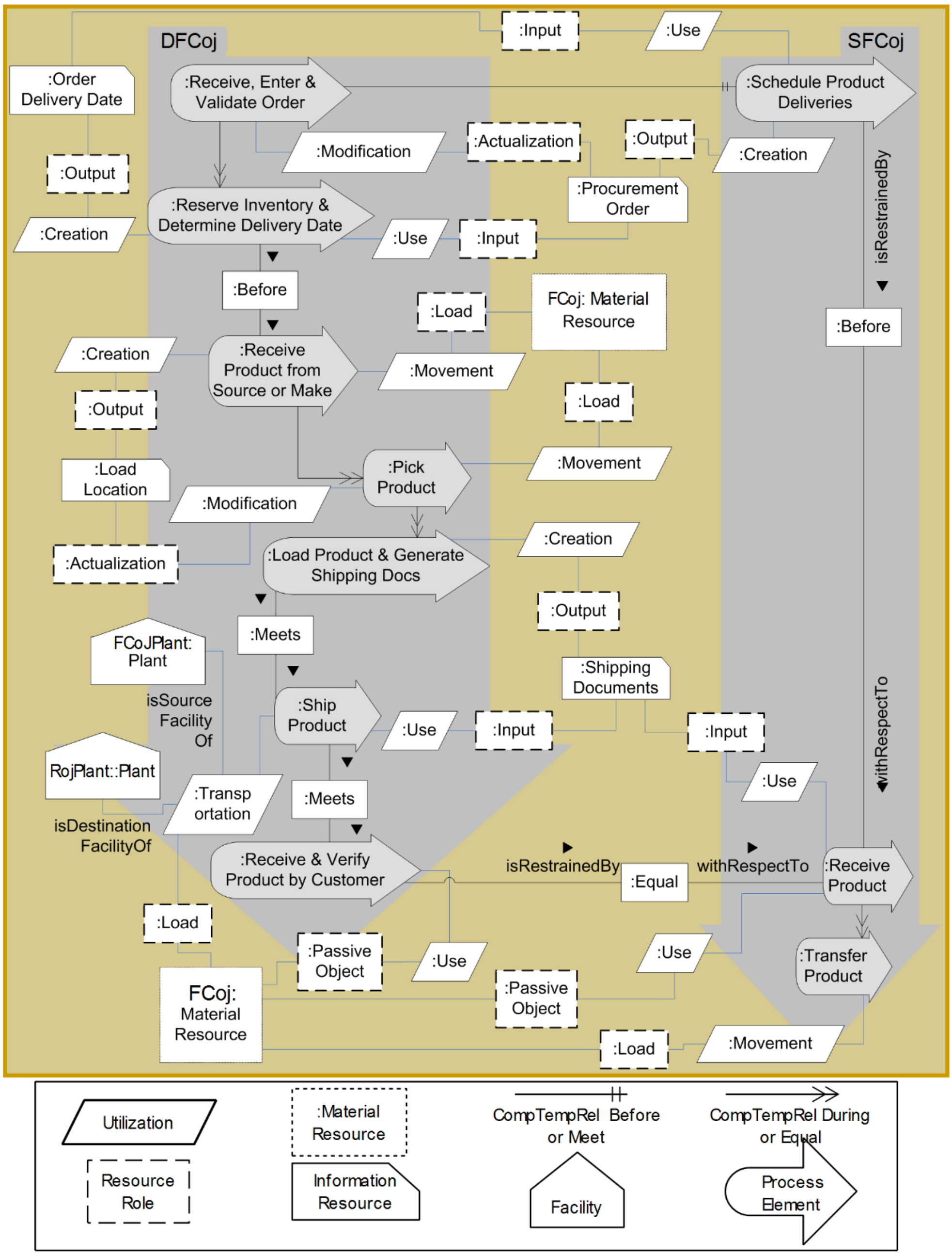


Fig. 10. Representation of *DFCoj* and *SFCoj* process elements, temporal relationships and resource flows

4. Ontology Evaluation

In order to evaluate SCONTO, aspects like content, structure, syntax and semantics have been analyzed. In such evaluation, the following methodologies and approaches have adopted:

- Ontoclean (Guarino and Welty, 2009) has been used to verify the correctness of the conceptual model structures
- A comparison with standards has been made to validate the ontology content depth and completeness.
- The tools available in the Protégé-OWL editor have been used to verify the consistency of the OWL implementation of the ontology.

In addition, a set of metrics have been adopted to obtain a quantitative description of the proposed ontology.

4.1. SCOPRO structure evaluation

The OntoClean methodology has been employed to validate the structure of SCOPRO. This methodology allows analyzing the ontology concepts and their hierarchical relations based on rigidity, identity, dependency and unity properties.

A concept is rigid (+R) if its instances are necessarily instances of it in all times. If all its instances can stop being instance of a concept in any time, this concept is anti-rigid (~R). A concept is not rigid if some instances can stop being instances of it.

The identity property qualifies the ability to distinguish between different individuals which instantiate a concept. A concept has an identity criterion (+I) if each of its instances can be distinguished. Otherwise, it has no identity (-I). The identity property can be inherited. Therefore, it is necessary to identify when the identity property is inherited or is owned by a concept. In this last case, the concept supplies identity (+O).

Regarding the dependency property, a concept is constantly dependent (+D) if each of its instances needs another individual to exist. In this situation does not hold, it is independent (-D).

An individual is considered a *whole* if its parts are linked among them by a relation R, and they are not associated with any other individual. The parts of a whole may also be a whole. The Unity property refers to the problem of describing the way the parts of an individual are bound together as a whole. A concept carries unity (+U) if all its instances exhibit a common unity criterion. A concept carries no unity (-U) if all its instances are wholes but with a different identity criterion. A concept carries anti-unity (~U) if all of its instances are not necessarily a whole.

Ontoclean proposes to analyse the ontology taxonomy using a set of rules that can help identifying problematic modelling choices. Given two concepts A and B, such that B subsumes A, the following rules must hold:

1. If B is anti-rigid, the A must be anti-rigid.
2. If B carries an identity criterion, then A must carry the same criterion.
3. If B carries a unity criterion, then A must carry the same criterion.
4. If B has anti-unity then A must also have anti-unity
5. If B is constantly dependent on a concept C, then A must also be dependent on C
6. Each individual belonging to A has to instantiate a unique class providing the identity criterion

The following paragraphs present a brief analysis of the SCOPRO concepts considering these properties. The classes that have been identified as rigid or anti-rigid, unity or anti-unity and the dependent ones are the only classes that may violate the OntoClean rules. Therefore, they are the only ones that are included in Table 10.

In the SCOPRO ontology, the *Independent OU*, *Dependent OU* classes and their subclasses are anti-rigid because the organizational units can be instances of one or the other at different times. For example, an independent organizational unit may become dependent when a company buys another one.

The SCOPRO classes that provide an identity criterion have also been identified. For each of them, their subclasses have been analyzed to check if there is no other class providing an incompatible identity. Besides, the verification that each "leaf" class has an identity (either own or inherited) has been done.

The *Supply Chain* class has its own identity (+O) because it is possible to distinguish a given OU network from other ones and this condition is not inherited. Similarly, *Market, Process, Process Occurrence, Organizational Unit, Organizational Unit Role, Resource Role, Functional Area, Functional SubArea, Utilization* and *Temporal Relationship* classes provide identity (+O). Considering resources, *Facility* and *Human Resource* classes provide identity while *Material Resource*, *Information Resource* and *Financial Resource* do not (-O), because their instances do not share a unique identity criterion. Therefore, these classes should be further specialized.

Table 10. SCOPRO concepts and their classification according the OntoClean philosophical properties

Property	SCOPRO Concepts		
Anti-rigidity (~R)	*Independent OU*	*Dependent OU*	All *Facility* subclasses
Proper Identity (+O)	*Supply Chain*	*Market*	*Process*
	Process Occurrence	*Organizational Unit*	*Organizational Unit Role*
	Functional Area	*Functional Subarea*	*Utilization*
	Temporal Relationship	*Facility*	*Human Resource*
	Resource Role		
No identity -I	*Material Resource*	*Information Resource*	*Financial Resource*
Unity (+U)	*Supply Chain*	*Process*	*Process Occurrence*
	Functional Area	*Functional Subarea*	*Utilization*
	Temporal Relationship	All *Organizational Unit* subclasses	
Without Unity (-U)	*SC Entity*		
Dependence (+D)	*Supply Chain*	*Process*	*Organizational Unit Role*
	Resource Role	*Process Occurrence*	*Dependent OU*
	Functional Area	*Functional Subarea*	*Utilization*

In turn, the *Independent OU*, *Dependent OU* concepts*,* and their subclasses inherit the identity from *Organization Unit* class. Similarly, *Resource Role*, *Process*, *Functional Area*, *Utilization*, *Temporal Relationship* and *Facility* classes, inherit their identity from their superclass (*Organizational Unit Role*).

To analyze SCOPRO under OntoClean unity-related rules it was necessary to identify the classes carrying unity criterion and to verify that their subclasses carry the same unity criterion. Also, it was required to check that the super classes of a concept having unity do not denote anti-unity. In SCOPRO, *the Supply Chain* class carries unity (+U) because it is possible to identify all the organizations belonging to each SC. The following classes also carry a unity criterion (+U): *Process, Process Occurrence*, *Functional Area*, *Functional SubArea*, *Utilization, Temporal Relationship* and all *Organizational Unit* subclasses.

OntoClean defines a class X as a dependent (+D) on another class W if each instance of X needs an instance of W to exist and the later instance is not part of the former. OntoClean states that all the subclasses of each dependent concept are also dependent. Table 11 shows the classes *Organizational Unit Role*, *Resource Role, Process*, *Process Occurrence*, *Dependent OU*, *Functional Area*, *Functional SubArea* and *Utilization,* which are considered as dependent concepts, and the classes on which they depend. All the subclasses of the class shown in the second column of Table 11 are also dependent (+D). Therefore, it can be concluded that the OntoClean rules regarding concept dependency are satisfied.

Since it was verified that the OntoClean rules have been fulfilled in terms of rigidity, identity, unity and dependence of its classes, it was determined that the SCOPRO conceptual models has an adequate structure.

Table 11. SCOPRO constantly dependent concepts

Constantly dependent class (+D)	Depends on
Supply Chain	Organizational Unit
Organizational Unit Role	Organizational Unit and Supply Chain
Resource Role	Resource and Utilization
Process	Organizational Unit Role
Process Occurrence	Process
Dependent OU	Organizational Unit
Functional Area	Organizational Unit
Functional Subarea	Functional Area
Utilization	Process and Resource Role

4.2. SCOPRO content evaluation

The ontology content evaluation has been carried out by comparing SCOPRO against a set of models considered as references in the domain, like IDEF 0, IDEF 3, ARIS, TOVE, among others. The analysis took into account whether a topic is treated or not in a proposal, as well as the depth of such treatment. For this, a score of 1 to 3 was adopted. It was scored 3 when a concept is modelled with precision appropriate to the domain, with a score of 2 if it is only vaguely modeled, and with 1 if the concept is mentioned by the proposal, but is not part of the model. Table 12 presents a summary of the evaluation results. When there is not enough information to demonstrate that a topic is addressed by a certain proposal there is a hyphen. It is important to mention that although 3 points have been assigned to certain proposals in a specific topic, an improvement in the treatment of such topic may be required in some cases.

Table 12. Content included in different proposals in relation to business processes, enterprises and supply chains.

	Process	Resource	Resource Flow	Temporal Information	Organizational Unit	Interorganizational Business Processes	SC Topological Structure	SC Management	SC Organizational Structure
IDEF 0[a]	2	2	2	-	-	-	-	-	-
IDEF 3 [a]	3	2	3	2	-	-	2	-	-
UML Activity Diagram[b]	3	2	2	3	-	2	2	-	-
BPMN[c]	3	2	2	3	-	3	2	-	-
CIMOSA[d]	3	3	3	3	3	2	2	-	-
GRAI-GIM[e]	2	2	2	-	2	-	-	-	-
ARIS	3	3	3	3	3	2	2	-	-
Enterprise Ontology[f]	3	3	3	3	3	-	2	-	-
TOVE[g]	3	3	3	3	3	-	2	-	-
SCOR[h]	3	2	2	2	1	3	2	-	1
GSCF[i]	3	2	2	2	1	3	-	3	3
SCOPRO[j]	3	3	2	3	3	3	2	-	3

- = topic not addressed 1= topic only mentioned 2= topic super partially modeled 3 = topic modeled with enough depth

[a] US-ICAM, 1981; Mayer et al., 1995 - [b] OMG, 2007 - [c] OMG, 2011 - [d] Vernadat, 1992; Kosanke et al., 1999 - [e] Doumeingts et al., 1992 - [f]Scheer & Schneider 2005 - [g] Uschold et al., 1998 - [h] Fox and Grüninger, 1998; Grüninger and Fox, 1994 — [i] Cooper et al., 1997b; Lambert & Cooper, 2000; Croxton et al., 2001; Rogers et al., 2002; Croxton et al., 2002; Lambert, 2008b - [i] Supply-Chain Council, 2012

The analysis has considered the following topics: i) Process; ii) Resource; iii) Resource Flows; iv) Temporal Information; v) Organizational Unit; vi) Interorganizational Business Processes; vii) SC Topological Structure; viii) SC Management; and ix) SC Organizational Structure. Although in some cases the first three topics are vaguely modeled, these topics are included in all the analyzed proposals.

SCOPRO has obtained the highest marks in topics (i) and (ii). In addition, it also models with enough depth topics labeled as (iv), (v), (vi) and (ix), and vaguely addresses topics numbered as (iii) and (vii). Since SCOPRO formalizes and extends the SCOR model, the analysis has shown that both proposals have building blocks for representing the same issues. However, since SCOPRO has incorporated several extensions, this ontology outperforms the SCOR model in the following subjects: Resource, Temporal Information, Organizational Unit, and SC Organizational Structure.

4.3. SCOPRO consistency evaluation

The three sub ontologies of SCONTO have been implemented with the ontology editor Protégé-OWL version 4.3.0. Therefore, the evaluation of the consistency of SCOPRO has been done using two reasoners provided with this editor, which are called HermiT 1.3.8 (Horrocks et al., 2012) and FaCT++ (Tsarkov and Horrocks, 2006). The latter is a description logic-based reasoner that permits satisfiability checking by means of tableaux algorithms (Baader et al., 2010). Hermit 1.3.8 allows satisfiability validation, using new algorithms developed in recent years based on hypertableaux calculations (Motik et al., 2009). Horrocks et al. (2012) state that these new algorithms generate a more complete reasoning in terms of data object properties and instance classifications.

The use of both reasoners allowed verifying the consistency of all SCOPRO definitions that have been implemented in OWL 2 (W3C, 2004). Therefore, it is considered that the syntax and semantics of the OWL 2 implementation of the SCONTO ontology are correct. In addition, the OWL implementation has been tested using OOPS! Pitfall Scanner (Poveda-Villalón et al., 2004). The results of this evaluation can be found in the following link: https://industrialonto.github.io/SCOPRO/OnToology/SCOPRO.owl/evaluation/oops.html.

4.4. Quantitative Metrics

Several methodologies, frameworks and metrics have been proposed to quantify the quality of ontologies (Yao et al., 2005; Gangemi et al., 2006; Tartir et al., 2010; Burton-Jones et al., 2005; Zhang et al., 2010; Yu et al., 2009; Manouselis et al., 2010). From the set of metrics proposed in the literature, the following subset has been selected to evaluate SCOPRO structural characteristics:

- NOC (Number of Classes), and NOR (Number of Relations) are simple counts of the number of classes and properties, respectively, defined in the ontology.
- NORC (Number of Root Classes) and NOLC (Number of Leaf Classes) metrics correspond to the number of classes without superclasses and classes without subclasses, respectively.
- RR (Relationship Richness), which is also called relation diversity, reflects the variety of relationships in the ontology. It is defined as the ratio of the number of non-inheritance relationships (P) divided by the total number of relationships, i.e. the sum of the inheritance relationships (H) and the non-inheritance ones (P)
- IR (Inheritance Richness) represents the average number of subclasses per class. It is computed as: FALTA
- DOSH (Depth of subsumption Hierarchy)¸ which is also called depth of inheritance, measures the length of the longest path from a given class C to the root class in an ontology subsumption hierarchy.
- AR (Attribute Richness) is defined as the average number of attributes per class. It is computed as the number attributes for all classes (ATT) divided by the number of classes (C).

The quantitative evaluation of SCOPRO has been done based on the information extracted by using the Protégé tool. The Thing OWL concept (superclass of all concepts in an OWL ontology) has not been considered to compute NOC, NOR, IR and AR. Table 13 shows the values of the adopted metrics that have calculated for SCOPRO.

Table 13. SCOPRO measures

Metric	SCOPRO
Number of Classes	200
Number of Relations	45+195 = 240
Number of root classes	5
Number of Leaf Classes	142
Relationship richness	45/240=0,19
Inheritance richness	195/200 =0,98
Depth of the subsumption hierarchy	6
Attribute richness	0

The result of the formula that computes the Inheritance Richness (IR) is a real number representing the average number of subclasses per class. The computed IR for SCOPRO is 98%. Such measure, together with the DOSH one, suggests that the proposed ontology is of a horizontal nature, which means that it represents a wide range of general knowledge. This was SCOPRO´s original aim, since it was conceived as general purpose ontology, capable to be specialized for different kinds of supply chains. In order to be reusable, the proposed ontology has to represent generic concepts common to different types of industrial organizations. Concept attributes describing particular characteristics of specific industries are out of the scope of SCOPRO. In consequence, the attribute richness metric value is equal to 0 attribute per concept. However, the amount of attributes per classes would increase as SCOPRO will be specialized to represent particular supply chains in the future. The Relationship Richness value is low; it implies that the hierarchy relationships overpass the other kinds of associations.

All these metric values give an idea about the characteristics and complexity of the ontology. However, the complexity measures provide no guarantee about the ontology quality, because there is no consensus or standard to compare with (Marquardt et al., 2010). The analysis of certain ontology characteristics, like usability and reuse, can only be measured and improved by using the ontology in different application contexts.

5. Conclusions

This paper presents the SCONTO ontology that contributes towards the formalization of the SC domain. SCONTO provides the basis for describing the supply chain structure, its associated processes, and evaluation system. This article just focuses on SCOPRO that is the main module of the ontology. This module includes and integrates the structure, process, and resource dimensions, emphasizing the SC processes knowledge. SCOPRO provides the following capabilities

- To describe supply chains that are composed of one or more enterprises.
- To represent the SC structure and the organizations that participate in it.
- To specify organizational processes and their decomposition into interrelated subprocesses.
- To define atomic and composite temporal relationships between processes.
- To describe resources, the effects that activities have on them, and the multiple functions they can fulfill in the SC processes.
- The explicit representation of material movements between SC partners.

This representation made it possible to establish a common vocabulary for all the SC actors, general enough to be valid in supply chains of different industries and having dissimilar sizes. Besides, SCONTO can be extended to allow its specialization to consider the characteristics of particular supply chains.

Acknowledgements

The authors acknowledge the financial support received from CONICET, Project PIP 11220170101131CO, Universidad Tecnológica Nacional, Project UTI5529, and Universidad Nacional del Litoral.

References

Allen J.F. Maintaining knowledge about temporal intervals. Communications of the ACM, Vol. 26 Nº 11, 1983, pp. 832-843

Baader F., Calvanese D., McGuinness D., Nardi D., and Patel-Schneider P.F. (editors). The Description Logic Handbook: Theory, Implementation and Applications, 2nd Edition, United States: Cambridge University Press, 2010.

Böhm, A.C., Leone, H.P., Henning, G.P. A supply chain ontology conceptualization with focus on performance evaluation. ICEIS 2008 - International Conference on Enterprise Information Systems, 2008.

Burton-Jones A., Storey V.C., Sugumaran V. and Ahluwalia P. A semiotic metrics suite for assessing the quality of ontologies. Data and Knowledge Engineering, 2005, Vol. 55 N° 1, pp. 84-102.

Chi, Y.L. Rule-based ontological knowledge base for monitoring partners across supply networks, Expert Systems with Applications, 2010, Vol. 37, pp. 1400–1407.

Doumeingts G., Vallespir B., Zanettin M. and Chen D. GIM-GRAI Integrated Methodology, a Methodology for Designing CIM Systems, Version 1.0, unnumbered Report, LAP/GRAI, Universidad de Bordeaux I, Bordeaux, France, 1992.

Fayez, M., Rabelo, L., Mollaghasemi, M. Ontologies for supply chain simulation modeling. Proceedings of the 2005 Winter Simulation Conference. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.

Fox M.S. and Grüninger M. Enterprise Modeling. AI Magazine, Vol. 19 Nº 3, 1998, pp. 109-121.

Gangemi A., Catenacci C., Ciaramita M. and Lehmann J. Modelling ontology evaluation and validation. In: The Semantic Web: Research and Applications, Germany: Springer, 2006, pp. 140-154.

Gruber T.R. A Translation Approach to Portable Ontologies. Knowledge Acquisition, Vol. 5 N° 2, 1993, pp. 199–220.

Grubic, T., Fan, I-S., Supply chain ontology: Review, analysis and Synthesis. Computers in Industry, Vol 61, 2010, pp. 776-786.

Grubic,T., Veza, I., Bilic, B. Integrating process and ontology to support supply chain modeling, International Journal of Computer Integrated Manufacturing, 2011, Vol. 24, No 9, pp. 847–863.

Grüninger, M. and Fox, M.S. An Activity Ontology for Enterprise Modelling. In: Workshop on Enabling Technologies – Infrastructures for Collaborative Enterprises, West Virginia University, 1994.

Guarino, N. and Welty, C.A. An overview of OntoClean. In: Handbook on Ontologies, Germany: Springer, 2009, pp. 201-220.

Horrocks I., Motik B. and Wang Z., The HermiT OWL reasoner. In: Proceedings of the 1st International Workshop on OWL Reasoner Evaluation, ORE-2012, England.

Kosanke K., Vernadat F. and Zelm M. CIMOSA: enterprise engineering and integration. Computers in Industry, Vol. 40 Nº 2-3, 1999, pp. 83-97
Lambert D.M. An executive summary of supply chain management: process, partnerships, performance. United States: The Hartley Press, 2008a.
Lambert, D. M. (editor). Supply chain management: processes, partnerships, performance. Supply Chain Management Inst. United States: The Hartley Press, 2008b.
Leukel, J. and Sugumaran, V. Formal Correctness of Supply Chain Design. Decision Support Systems, 2013, Vol. 56, pp. 288-299.
Lu, Y., Panetto, H., Ni, Y. and Gu, X. Ontology alignment for networked enterprises information systems interoperability in supply chain environment, International Journal of Computer Integrated Manufacturing, 2013, Vol. 26, No 1–2, pp. 140–151.
Manouselis N., Sicilia M.A. and Rodríguez D. Exploring ontology metrics in the biomedical domain. Procedia Computer Science, Vol. 1 N° 1, 2010, pp. 2319-2328.
Marquardt W., Morbach J., Wiesner A. and Yang A. OntoCAPE A Re-Usable Ontology for Chemical Process Engineering. Germany: Springer, 2010.
Mayer R.J., Menzel C.P., Painter M.K., de Witte, P.S., Blinn, T. and Perakath, B. Information Integration for Concurrent Engineering (IICE) IDEF3 Process Description Capture Method Report. AL-TR-1995-XXXX. Armstrong Laboratory, Wright-Patterson AFB, Ohio 45433, 1995
Motik B., Shearer R. and Horrocks I. Hypertableau reasoning for description logics. Journal of Artificial Intelligence Research, Vol. 36 N° 1, 2009, pp. 165-228.
Object Management Group (OMG). UML • Unified Modeling Language Specification V2.1.1: Superstructure, OMG Document Number formal/2007-02-05, 2007.
Object Management Group (OMG). Business Process Modeling Notation (BPMN). Version 2.0, OMG Document Number: formal/2011-01-03, 2011.
Object Management Group (OMG). Object Constraint Language Version 2.4, 2014. Available on-line: https://www.omg.org/spec/OCL/2.4/PDF Last access: 2021/04/19
Olivé A. Conceptual Modeling of Information Systems. Germany: Springer, 2007.
Petersen, N., Grangel-González,I., Coskun, G., Auer, S., Frommhold, M., Tramp, S.,Lefrançois, M. and Zimmermann, A. SCORVoc: Vocabulary-based Information Integration and Exchange in Supply Networks. 2016 IEEE Tenth International Conference on Semantic Computing.
Poveda-Villalón, María, Asunción Gómez-Pérez, and Mari Carmen Suárez-Figueroa. "OOPS!(Ontology Pitfall Scanner!): An on-line tool for ontology evaluation." International Journal on Semantic Web and Information Systems (IJSWIS), Vol. 10 No 2, 2014, pp.7-34.
Rodríguez-Enríquez, C.A., Alor-Hernández, G., Sánchez-Ramírez, C. and Córtes-Robles, G. Supply chain knowledge management: A linked data-based approach using SKOS. DYNA, 2015, Vol. 82, No 194, pp. 27-35.
Sakka, O., Millet, P.A. and Botta-Genoulaz V. An ontological approach for strategic alignment: a supply chain operations reference case study, International Journal of Computer Integrated Manufacturing , 2011, Vol. 24, No 11, pp. 1022–1037.
Sheer, A.W., and Schneider, K., 2005. ARIS architecture of integrated information systems. In: Handbook of information systems. Berlin: Springer.
Solanki, M. and Brewster, C. OntoPedigree: Modelling pedigrees for traceability in supply chains. Semantic Web, 2016, Vol. 7, pp. 483–491.
Supply-Chain Council. Supply Chain Operations Reference Model SCOR Version 9.0, 2008.
Supply-Chain Council. Supply Chain Operations Reference Model SCOR Version 10.0, 2010.
Supply-Chain Council. Supply Chain Operations Reference Model SCOR Version 11.0, 2012.
Tartir S., Arpinar I.B. and Sheth A.P. Ontological evaluation and validation. In: Theory and Applications of Ontology: Computer Applications. Paises Bajos:Springer, 2010, pp. 115-130.
Tsarkov D. and Horrocks I. FaCT++ description logic reasoner: System description. In: Automated reasoning, Germany: Springer, 2006, pp. 292-297.
US Air Force Integrated Computer Aided Manufacturing (ICAM). ICAM Architecture, Parte II, Volumen IV-Functional Modeling Manual (IDEF0). FWAL-TR-81-4023. Air Force Materials Laboratory, Wright-Patterson AFB, Ohio 45433, 1981
Uschold M., King M., Moralee S. and Zorgios Y. The Enterprise Ontology. The Knowledge Engineering Review, Vol. 13 Nº 1, 1998, pp. 31-89.
Vernadat F.B. CIMOSA – A European Development for Enterprise Integration. Part 2: Enterprise Modelling. In: Enterprise Integration Modelling, United States: The MIT Press, 1992, pp. 189-204.
W3C. 2004. OWL Web Ontology Language Overview. W3C Recommendation. Ed: McGuinness, D.L., van Harmelen, F., 2004 Available on-line: http://www.w3.org/TR/2004/REC-owl-features-20040210/ - Last access: 2021/04/19
Yao H., Orme A.M. and Etzkorn L. Cohesion metrics for ontology design and application. *Journal of Computer Science*, Vol. 1 N° 1, 2005, pp. 107-113.
Yu J., Thom J.A. and Tam A. Requirements-oriented methodology for evaluating ontologies. Information Systems, Vol. 34 N° 8, 2009, pp. 766-791.
Zhang H., Li Y.F. and Tan, H.B.K. Measuring design complexity of semantic web ontologies. Journal of Systems and Software, Vol. 83 N° 5, 2010, pp. 803-814.
Zdravkovic, M., Panetto, H., and Trajanovic, M. Concept of semantic information pool for manufacturing supply networks. International Journal of Total Quality Management and Excellence, 2010, Vol. 37, Nº 3, pp. 69–74.

An overview of some ontological challenges in engineering maintenance

Stefano Borgo[1], Francesco Compagno[1,2], Nicola Guarino[1], Claudio Masolo[1], and Emilio M. Sanfilippo[1]

[1] ISTC-CNR Laboratory for Applied Ontology (LOA),
via alla cascata 56/C, Povo, Trento, 38123, Italy
{name.surname}@cnr.it
[2] Adige S.P.A, via per Barco, 11, Levico Terme, 38056, Italy
francesco.compagno@istc.cnr.it

Abstract. Maintenance is an important technical aspect that must be considered in engineering practices. In this paper we present a preliminary ontological investigation of questions such as "What is a component of an engineering system?" and "What happens when a component is replaced after a malfunctioning?", which are both fundamental from a maintenance modeling stance. We focus in particular on two inter-related problems, which we call *the missing component* and *the replacement* problems. We describe different approaches dealing with them. First, we start representing *kinds* of components and systems as temporally qualified first-order logic predicates, eventually reified. We then consider a four-dimensionalist (4D) perspective, mainly based on the ISO 15926. Lastly, we briefly mention a novel point of view based on possible worlds. By the end of the paper, we shortly compare the approaches by discussing their advantages and shortcomings.

Keywords: Maintenance · Ontology .

1 Introduction

In this preliminary research work, we address two inter-related problems relative to the ontological conceptualization of experts' knowledge with respect to engineering maintenance. We call these challenges *the missing component problem* and *the replacement problem*. From the perspective taken in the paper, both challenges regard engineering systems, hereby simply understood as products composed of various inter-related components. Both problems are documented in the literature [2,3,8,11] and emerge from real-world engineering practices. The analysis we propose is part of a study that we are carrying out in collaboration with Adige S.P.A (BLM Group), a company specialized in the manufacturing of laser cutting machines (Fig. 1), being developed in the context of the European H2020 project *OntoCommons: Ontology-driven data for industry commons*.[3] The purpose of the collaboration is to support expert decision making about maintenance procedures through the development of an ontology-based maintenance information system. The examples we shall discuss come from this scenario.

[3] Website: `https://ontocommons.eu`, last accessed April 2021.

Some approaches to deal with the aforementioned problems have been already proposed. However, each of them assumes its own ontological principles and formal system, so that it is not clear to which extent they (dis-)agree, nor what their (dis-) advantages are. This paper aims therefore at comparing them to support knowledge engineers in selecting one approach over the others.

Fig. 1. An Adige LT8 laser machine. These machines cut metal tubes using a cutting head moving along three axes, together with the translatory and rotatory motion of the tube.

The paper is structured as follows. Section 2 briefly describes and discusses the aforementioned problems. Section 3 compares four different approaches for modeling engineering systems and their components, discussing for each of them their relation with the missing component problem and the replacement problem. Finally, Section 4 concludes and summarizes our contribution.

2 The two problems

The missing component problem. The problem of the missing component has been documented in the literature in various places (see for instance [2,3,11]). Here we present it as specifically concerning maintenance scenarios. In these contexts, indeed, components may be physically removed from their hosting systems, e.g., to be controlled, cleaned or repaired (and sometimes even replaced). However, even when a component is not installed in the system, technicians may refer to it as if it were already in its right place; e.g., they can say *"This cable leads to the laser head"*, "*The laser head is placed in this position*", "*The laser head of this machine has not been installed yet*". This talk seems to presuppose a sort of augmented reality scenario, which in certain situations would indeed be useful (so that it is often artificially created nowadays): for instance, in the context of an assembly task, it may be useful to visualize a component in its expected position, even though it is not physically present. Of course, the technical problem is how to clarify the semantics of the statements above, and, in particular, how to make sense of the fact that engineers talk and reason about an entity that is not physically present (see Fig. 2 for an example of laser cutting head).

Fig. 2. The LT8 laser cutting head in action.

The replacement problem. Let us consider a particular laser cutting machine, like the one in Fig. 1, that has at time t a protective window[4] installed in the appropriate position within the cutting head, we call it *protective-window$_1$*. Assume that, due to malfunctioning, *protective-window$_1$* is replaced at time $t+1$ with a new one, *protective-window$_2$*. Technicians who replace the protective window are therefore primarily concerned with two specific physical objects: the defective one and the new one. In some cases, however, the same technicians seem to refer to something else. For instance, when claiming *"The protective window of this cutting head has been replaced 3 times in the last year"*, they clearly don't refer with the expression 'the protective window of this cutting head' to an ordinary physical object (since no physical object was replaced three times), but to some other entity whose ontological nature is unclear.

The reader should notice the strong connection between the two problems just mentioned. In both cases, engineers talk about entities whose nature is different from that of ordinary physical objects; first, by referring to them even when they are not physically present, second, by thinking of them as items that keep their identity even when replaced. One has therefore to bite the bullet and make sense of these entities from an ontological stance. In particular, the two problems above are intimately related to the possibility for a system to lack some component or have a component replaced. This involves subtle notions concerning the existence conditions of engineering systems and their persistence through time.

In engineering practice, the design process commonly ends up with the production of a technical specification (a.k.a. design model). Hence, to understand when a material

[4] The protective window is a particular glass that protects the laser beam optics from, e.g., the metal sparks flying around during a cutting process.

realization of a certain specification physically exists, it is crucial to clarify the conditions by which a physical system *complies* with the specification. For instance, suppose to have an object a that has only some of the components required by the specification. There are at least two possibilities here: (i) no system (of the given kind) exists because the object a does not fully comply with the specification; (ii) a is a system (of a certain kind) because it *sufficiently* complies with the specification. To define what 'sufficiently' means, one may refer to the *relevance* of the components, e.g., it may be necessary for a to have only the *fundamental* components. This approach would however require to characterize fundamental vs. non-fundamental, or even optional, components while being sure that this manner of framing compliance matches well with engineering practices. The possibility of replacing the components of a system makes the overall situation more complex, because (at least in principle) fundamental components can be replaced and maintained, too. During maintenance operations, therefore, a system may lack some of its fundamental components and, thus, stop existing, at least in principle.[5]

3 Overview of approaches

We provide in this section an overview of some ontological approaches that address the problems previously mentioned. We will assume that an engineering system of a certain *kind* (e.g., a mechanical assembly) can be described by means of technical specifications in terms of the *features* of its components and the *relations* holding among them. To depict a specification, we use a graph like the one in Fig. 3, where nodes stand for components (with their characterizing features), and arcs for relations between components. For instance, the graph in Fig. 3 describes a generic system with four components, denoted $\mathbb{C}i$, each characterized by some (complex) feature[6], F_i, and five relationships R_{ij} holding between them.

Generally speaking, the $\mathbb{C}i$s could be recursively specified and further decomposed into components. However, we assume that the $\mathbb{C}i$s do not have proper parts. The main designing activity regards therefore the choice of both the components and the way in which they are structurally related. For the sake of the discussion, we will often refer to the example in Fig. 3, although the analysis is generalizable to arbitrary specifications given in terms of features and relations between components.

[5] A practical choice to avoid this problem (not discussed here) might be to admit that, in a service and maintenance context, the identity of the system to be repaired or maintained is assumed by convention, independently of the status of its components. For instance, we may say that, as long as a certain machine maintains its serial number, and a service contract concerning a machine with that serial number is still operational, the machine to be serviced exists (although possibly in a nonfunctional state), independently of the presence and functional state of its components

[6] We assume that a complex feature is a boolean combination of a property that characterizes the component's kind (say, *pump* or just *physical object*) plus one or more qualitative properties describing a particular shape, size, color, etc. For instance, F_1 could mean that $\mathbb{C}_1$ is a physical object that weighs 2kg $\pm$ 0.1kg, is made of metal or plastic but not wood etc.

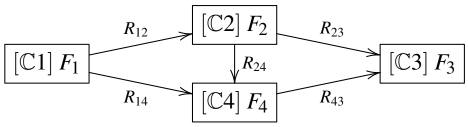


Fig. 3. An example of graph-based technical specification

3.1 Approach 1: system kinds as predicates

In this first approach, the specification of a certain system is conceived as a specification of its most specific kind (i.e., the most specific property it shares with all its duplicates). In turn, the latter consists in stating the necessary and sufficient conditions for being, at a certain time, an instance of that kind. Formally, a temporally indexed monadic predicate $\mathtt{K}_t$ will be used to represent the property of being an instance of kind K at time t. Let us assume that Fig. 3 specifies a system of kind K. The corresponding logical specification would be expressed by (f1), where $o_1 \oplus \ldots \oplus o_n$ denotes the mereological sum of n mutually disjoint objects, and $o_1 \equiv_t o_2$ means that, at time t, the objects o_1 and o_2 have the same parts. Formula (f1) ensures therefore the *full compliance* of all K-systems with respect to the specification in Fig. 3.

f1 $\mathtt{K}_t x \leftrightarrow \exists abcd(x \equiv_t a \oplus b \oplus c \oplus d \wedge \mathtt{F1}_t a \wedge \mathtt{F2}_t b \wedge \mathtt{F3}_t c \wedge \mathtt{F4}_t d \wedge \mathtt{R12}_t ab \wedge \mathtt{R14}_t ad \wedge \mathtt{R23}_t bc \wedge \mathtt{R43}_t dc \wedge \mathtt{R24}_t bd)$

To address the replacement problem, we need now to suitably characterize the tags $\mathbb{C}i$, which denote the unique role played by each component in the system. In general, this role depends on all the relations holding among components, so, for example, $\mathbb{C}1$ would be defined as follows:[7]

f2 $\mathtt{C1}_t x \leftrightarrow \mathtt{F1}_t x \wedge \exists sbcd(s \equiv_t x \oplus b \oplus c \oplus d \wedge \mathtt{F2}_t b \wedge \mathtt{F3}_t c \wedge \mathtt{F4}_t d \wedge \mathtt{R12}_t xb \wedge \mathtt{R14}_t xd \wedge \mathtt{R23}_t bc \wedge \mathtt{R43}_t dc \wedge \mathtt{R24}_t bd)$

However, in some cases it is possible to uniquely characterize a certain $\mathbb{C}i$ just in terms of the relations holding between the i-th component and its neighbor components, as in formula (f3), where $o_1 \leq_t o_2$ stands for 'at time t, the object o_1 is part of the object o_2'.

f3 $\mathtt{C1}_t x \leftrightarrow \mathtt{F1}_t x \wedge \exists sbd(\mathtt{K}_t s \wedge x{+}b{+}d \leq_t s \wedge \mathtt{F2}_t b \wedge \mathtt{F4}_t d \wedge \mathtt{R12}_t xb \wedge \mathtt{R14}_t xd)$

Note that, in this particular case, the necessary and sufficient conditions in (f3) may be further weakened out, since $\mathbb{C}1$ is the only component with feature $F1$ in a K-system, but, in general, relational constraints are needed when systems have several components with the same features (which is typically the case). For instance, should we only consider constraints on features, $\mathbb{C}2$ and $\mathbb{C}4$ would turn out to be indistinguishable if $F2 = F4$, the relations they have with the other components are the same (i.e., $R_{12} = R_{14}$ and $R_{23} = R_{43}$), and R_{24} is symmetric.

7 If a K system is defined by (f1), then (f2) implies that x can be a $\mathbb{C}1$-instance only if there exists a K-system s which has x as a part, i.e., $\mathbb{C}1$-components are existentially dependent on K-systems.

Note also that the lack of a single $\mathbb{C}i$-component could induce the lack of all other kinds of components. In the example in Fig. 3 this happens when a $\mathbb{C}2$- or $\mathbb{C}4$-component is missing. The mutual existential dependencies between components need therefore to be carefully taken into account.

Let us show now how we can address the replacement problem with this approach. Observe first that component definitions can be *relativized* to a specific system by adding a system-argument in the antecedent of the formula and discarding the existential quantification on systems in the consequent; see (f4) representing x as $\mathbb{C}1$-component of the specific system s. The replacement of the $\mathbb{C}1$-component of s can be then represented as in (f5).

f4 $\mathtt{C1}_t xs \leftrightarrow \mathtt{F1}_t x \wedge \exists bd(\mathtt{K}_t s \wedge x{+}b{+}d \leq_t s \wedge \mathtt{F2}_t b \wedge \mathtt{F2}_t d \wedge \mathtt{R12}_t xb \wedge \mathtt{R14}_t xd)$
f5 $\mathtt{C1}_{t_0} xs \wedge \mathtt{C1}_{t_1} ys \wedge \mathtt{C1}_{t_2} zs \wedge x \neq y \neq z$

What (f5) says is that the role of 'being the $\mathbb{C}1$-component of s' is played by three different objects at different times. Note that no single individual corresponding to the expression 'the $\mathbb{C}1$-component of s' is present in the domain of quantification.

However, this approach does not allow dealing with the missing component problem in the strict sense, since only the actual physical components are included in the domain of discourse. On the other hand, however, having isolated the properties of system-kinds and their components, one can talk of systems and their components in general without pointing to their physical realizations. According to the defenders of this approach, this view seems close to design practices where experts commonly talk about and reason over design specifications and their background knowledge without necessarily pointing to physical items.

3.2 Approach 2: system-kinds as individual constants

The approach considered in this section is conceptually similar to the previous one, the main difference being technical. System-kinds are indeed not represented via predicates but by means of individual constants: $\mathtt{K}_t x$ is now replaced by $x ::_t \mathtt{k}$ where $\mathtt{k}$ represents the kind K and :: stands for the *instantiation* relation so that $x ::_t \mathtt{k}$ reads 'x is an instance of kind K.' Similarly for relations, e.g., $\mathtt{R}_t xy$ is replaced by $xy ::_t \mathtt{r}$ where the *new* instantiation relation has four arguments (i.e., x, y, t and $\mathtt{r}$).[8] We can follow the discussion in Sect.3.1 by observing that the previous formulas can be rewritten in this new framework; e.g., (f1) can be rewritten as in (f6).

f6 $x ::_t \mathtt{k} \leftrightarrow \exists abcd(x \equiv_t a \oplus b \oplus c \oplus d \wedge a ::_t \mathtt{f1} \wedge b ::_t \mathtt{f2} \wedge c ::_t \mathtt{f3} \wedge d ::_t \mathtt{f2} \wedge$
$ab ::_t \mathtt{r12} \wedge ad ::_t \mathtt{r12} \wedge bc ::_t \mathtt{r23} \wedge dc ::_t \mathtt{r23} \wedge bd ::_t \mathtt{r24})$

The introduction of system-kinds in the domain of quantification allows taking into account their *intensional* and *intentional* dimensions. Different kinds could have the same instances and their difference can be grounded on some meta-information, e.g., that a design feature (or an entire specification) has been designed by an engineer working in the company, therefore copyrights apply to it. Furthermore, rather than relying on formulas like (f6) or on the counterpart of (f5), system-kinds can be considered as

[8] For simplicity we write in the same way all the instantiation primitives.

composed by their component-kinds (e.g., by writing $\mathtt{k} = \mathtt{c1} + \mathtt{c2} + \mathtt{c3} + \mathtt{c4}$), which is an approach similar to what done in formal ontology with respect to the debate on structural universals [9]. In this view the mereological structure of a system-kind is aligned with the one of its instances; see (f7)-(f8).

f7 $x ::_t k \wedge c \leq k \rightarrow \exists y(y ::_t c \wedge y \leq_t x)$
f8 $x ::_t k \wedge y ::_t c \wedge y \leq_t x \rightarrow c \leq k$

This idea can be pushed further towards the introduction of component-kinds *relativized* to a specific system. These component-kinds—that indeed, at a given time, can have only one instance —can be intended as the subject of replacements, i.e., they can have different realizations (instances) through time. Furthermore, they can be deployed also to address the missing component problem because their existence is independent from the one of their material instances, i.e., they can be 'empty' (i.e., not instantiated) at some times. However, in order to exist they require the whole system to exist, therefore one needs to accept partial compliance with the drawbacks discussed before.

3.3 Approach 3: systems as four-dimensionalism objects

Four-dimensionalism (4D) is the philosophical perspective according to which the objects of everyday experience have both spatial and temporal parts. To understand this, let us shortly comment on the opposite position, called three-dimensionalism (3D). According to 3D, objects have only spatial parts; e.g., a laser cutting machine has a cutting head, a laser source, a water chiller, etc. The machine can lose and acquire parts, but whenever it is present in time, it consists of all parts that it has at that time. Differently, for four-dimensionalists an object x exists at a time t if and only if there exists its temporal slice $x_{@t}$ (which is present only at t). That is, if my laser cutting machine c exists at both t and t', it has two different temporal slices at those times, i.e., $c_{@t}$ and $c_{@t'}$, respectively. A whole object consists therefore in the mereological sum of all its temporal parts.[9]

In the landscape of applied ontology, West [11] has developed a 4D-approach for engineering which has eventually led to the standard ISO 15926 [5]. Similarly to Approach 1 (see Section 3.1), according to this theory, system-kinds are intended in an extensional way but both physical systems and their components are now conceived as 4D-objects in the sense just introduced. Following the 4D approach, a proposition $\mathtt{P}_t x$, which reads 'x satisfies P at t', can be reduced to $\mathtt{P}x_{@t}$, i.e., 'the temporal slice of x existing at and only at t satisfies P'. In the same line, (f1) can be rewritten as in (f9).

f9 $\mathtt{K}_t x \leftrightarrow \exists abcd(x_{@t} = a_{@t} \oplus b_{@t} \oplus c_{@t} \oplus d_{@t} \wedge \mathtt{F1}a_{@t} \wedge \mathtt{F2}b_{@t} \wedge \mathtt{F3}c_{@t} \wedge \mathtt{F2}d_{@t} \wedge \mathtt{R12}a_{@t}b_{@t} \wedge \mathtt{R12}a_{@t}d_{@t} \wedge \mathtt{R23}b_{@t}c_{@t} \wedge \mathtt{R23}d_{@t}c_{@t} \wedge \mathtt{R24}b_{@t}d_{@t})$

Looking at West's position more deeply, he argues that systems' components (1) are existentially dependent on the systems they are part of, and (2) are non-ordinary physical objects which can undergo moments of non-existence. Recall the example of a particular LT8 cutting machine and consider its cutting head. According to West's

[9] Admittedly, this is a simple way to look at the distinction between 3D and 4D. The reader can refer to [4] for more information.

first claim, the cutting head exists as a system component of the whole machine only when the latter exists. In this sense, *being a system component* is a sort of *role property* which an object might satisfy at some time. For the second claim, the idea is that a component of a certain system s is the mereological sum of all the temporal slices of the ordinary objects which have played the corresponding component-role in relation to s. For instance, the $\mathbb{C}1$-component of a system s can be defined as in (f10) (note that the formula makes explicit the existential dependence of components on the system s).

f10 $\mathtt{C1}xs \leftrightarrow x = \sigma y(\exists tz(y = z_{@t} \wedge y \leq s \wedge \mathtt{C1}y))$

By (f10), x is a single individual that represents the succession of all the (temporal slices of the) ordinary objects playing the $\mathbb{C}1$-role. This is an interesting move to deal with the replacement problem. According to West, indeed, the 4D-component of a system (the object x in (f10)) is the sum of the temporal slices of the different 'ordinary' objects (the slices $z_{@t}$ in (f10)) that are placed in the system during its existence. It follows, as noted by West, that components' existence can be intermittent, i.e., when no ordinary object plays the $\mathbb{C}1$-role in the system s, the $\mathbb{C}1$-component of s does not exist. Hence, in principle, also the whole system s can intermittently exist, e.g., when one of its (fundamental) components is missing (a fact which is reflected in (f9)). However, it should be clear that in order to talk about missing components of a system s, one needs to include a notion of partial compliance into the framework in such a way to guarantee the existence of the system s even when it lacks components. For, a component can be said to be missing at some time t only if the system of which it is a component exists at t notwithstanding it lacks parts.

3.4 Approach 4: Embracing possibilism

Let us now briefly introduce a fourth approach based on some former work [3], which is however still a research hypothesis that deserves a proper formalization. Differently from Approach 1, we assume here that the C_i do not denote properties, but specific parts of a certain system. The tags are therefore *part names*, as usual when describing assemblies, and not property names. We also assume that a system s is an individual that exists even at design time, when the system has not been physically realized yet. Of course, we need to characterize the ontological nature of these entities, which turns out to be definitely non-standard.

We shall consider *design objects* as *possible* (although non-actual) physical objects. Under this view, designing an object means *choosing* a possible physical object among many others, and *describing* it. For example, the design object satisfying the specification of Fig. 3 would be simply described as follows(where s, c_1, c_2, c_3, c_4 are constants denoting distinct elements of the domain of discourse);

f11 $s = c_1 \oplus c_2 \oplus c_3 \oplus c_4 \wedge \mathtt{F1}c_1 \wedge \mathtt{F2}c_2 \wedge \mathtt{F3}c_3 \wedge \mathtt{F4}c_4 \wedge$
$\mathtt{R}_{12}c_1c_2 \wedge \mathtt{R}_{14}c_1c_4 \wedge \mathtt{R}_{23}c_2c_3 \wedge \mathtt{R}_{43}c_4c_3 \wedge \mathtt{R}_{24}c_2c_4$

Differently from the previous cases, where the formulas fully specify the *conditions* for an object to be of a particular kind K, corresponding to a certain product model, this formula simply presupposes that a certain object s exists in the mind of the designer, and provides a description of its nature in terms of the properties (P_i) of its components and the relations (R_{ij}) among them. So, under this view, the content of a design activity (i.e., *what* is designed) is an individual and not a kind. The corresponding kind K would be defined as the most specific property shared by all *duplicates* of s, i.e. (forgetting tolerances), by all individuals which share s's intrinsic properties.

To account for the peculiar way design objects exist, we shall adopt Lewis' possibilist realism [7,12], according to which a possible (physical) object is simply a part of a maximally spatiotemporally related whole, i.e., a whole among whose parts spatiotemporally relations hold. More exactly, forgetting temporal relations for the sake of simplicity, we shall assume that a possible physical object is an object such that a certain *spatial distance* relation holds among its parts (this means that all the relations occurring in Fig. 3 imply some spatial relation between their arguments). Under this view, objects are *physical* not because they are located in a region of space, but simply because they are *extended in space*, in the sense that spatial relations hold between their parts. Of course, they may also have a location, but only relative to some other physical objects.

Following this line of thought, we shall think of a design object as a maximally spatially related whole. Forgetting temporality, this means that a design object is just a possible world, in Lewis' terms. So, we can think that formula (f11) holds in a world (different from the actual one) where only an instance of K exists. For example, a laser cutting machine that has been designed but has not been built yet is something like a possible object floating in the void, i.e., a world of which it is the only inhabitant.

We shall assume that the *actual physical world* is the largest spatially related whole that includes us. For a physical object to exist means to be part of some maximal spatially related whole, and to actually exist means to be part of the actual world. So, we can have a physical object that exists, but it is not actual. If, in addition, this object is the content of an intentional design act (whose author is an agent who inhabits the actual world), then it is a design object: design objects exist but they are not actual.

We have therefore a solution to a variant of the missing component problem: when engineers talk of a certain component within a system (possibly even before the system is realized), they may talk of it in a generic way, as in the statement 'the protective window is a part of the cutting head of the LT8 machine.' In this case we shall say that they are talking of a design object, i.e., of a physical object that exists in a non-actual world and is the content of an intentional design act.

In most cases, however, the missing component problem concerns a specific, actually existing system, one of whose components has been removed. To address this case we must discuss the relationship between design objects and actual objects. We shall say that an object x is a *realization* of a design object y iff x is a duplicate of y (forgetting tolerance) and x exists in the actual world, that is, the spatial distance of its parts from the other objects belonging to the actual world is defined. We shall assume that, when engineers talk of a missing component as if it was there, they talk of a *virtual* entity, which is a sort of *projection* of a design system's component on the actual world: it is a duplicate of that component which is assumed to have a location (and therefore exist) in the actual world, although it is not *real*, since it is not capable of causal interactions with the actual world. The actual world can host therefore both *real* and *virtual* entities. The latter are intentional entities (like design objects), which are existentially dependent on one or more agents. In conclusion, when technicians talk of a missing component as if it were there (e.g., using the expression *'This frame supports the cutting head's protective window'* they refer to a virtual component, which is a projection of a design object on the actual world. Such a virtual component may be visualized by using augmented reality techniques.

Finally, let us briefly mention how the replacement problem may be addressed in this approach. The problem is the reference of a certain identifier *i*, which in our example is *'the protective window of this cutting head'*, used to talk of something related to a real object existing in the actual world. Assuming that *i* denotes a virtual object does not work, since when the protective window is installed we expect *i* to refer to the real window. On the other hand, one cannot assume that the *i* denotes a real physical object, since the reference link would break as soon as the object is removed.

A solution is to assume that the identifier refers to a *variable object*, i.e., following Fine [1], an object that is a *variable embodiment* of other objects. Suppose to have a function f that, at each time when an object exists, tells us what its parts are. Such object is called the *variable embodiment of f*, and the sum of its parts at t is called its *manifestation* at t. In our case, we can assume that *'the protective window of the cutting head'* denotes a variable embodiment such that, when a physical protective window is installed, its manifestation is just that physical window. When the physical window is removed, its manifestation is the virtual window.

4 Discussion and conclusions

The paper presented an initial analysis of ontological problems in the maintenance domain. In particular, it addressed what happens in the world when system components are missing or replaced. These problems are inextricably intertwined with compliance conditions, as one must consider how a system can be compliant with its specification when some of its components are not in place, and how it survives the replacement of some components.

We have presented four different modeling approaches with the purpose of analyzing their ontological commitments and modeling (dis-)advantages. The first two (Sect. 3.1 and Sect. 3.2) are very similar from a conceptual perspective in that system kinds and their characterizing features (all represented through design specifications) stand for *properties* that particular physical systems satisfy (if compliant at some suitable degree). In both approaches, dealing with the missing component and the replacement problems means looking at the physical items that, at different times, do physically exist. As we have seen, a sentence like *'the protective window of this cutting head has been replaced twice'* needs to be rephrased in something like *'this cutting head had three different protective windows at different times.'* As said, the only difference between the two approaches is that in the second one properties are reified. Among other advantages, we have noticed that this choice allows one to deal with the missing component problem, because one can (at least) point to the reified properties that the missing component is meant to satisfy, although, again, there is no way to account for the superficial semantics of sentences like *This cable leads to the laser head.*

The third approach (see Sect. 3.3), based on West [11], relies on a 4D ontological framework. Similarly to Approach 1, system-kinds, their components, and features are (first-order) properties that particular entities satisfy. Hence, although West does not explicitly address the representation of compliance conditions, the considerations done in Sect. 3.1 and Sect. 3.2 could be tuned to this approach by adopting a 4D representational system. For the replacement problem, this approach offers an interesting solution. Indeed, a whole 4D-component of a certain system can well correspond to the mereological sum of the temporal slices of different physical objects, which (at different times) are linked to that system. Similarly to the previous cases, the missing component problem remains problematic: if at a certain time there is no temporal slice in the world to which one can point to, there is no 4D-component altogether at that time.

Finally, the paper (preliminarily) introduced a fourth approach (Sect. 3.4), which relies on the introduction of a new type of objects, namely design objects, which depend on the intention of an agent, have a spatial extension, and yet are not actual (i.e., they do not physically exist in the common-sense reality). Differently from the previous approaches where formulas represent system kinds (as represented in specifications), this one is mainly focused on design objects, logically treated as possible individuals that may be *realized* in the actual world, originating real individuals, or just ‘projected’ onto the actual world, originating virtual individuals. From this perspective, when a real component is missing from a larger system, one can still refer to its virtual counterpart in the physical world. The replacement problem remains more challenging and we discussed a way to deal with it based on Fine’s [1] theory on variable embodiments.

To conclude, the results of the paper are still preliminary and further work is necessary at both the conceptual and representational levels. All approaches have their (dis-) advantages. For instance, the 4D perspective, adopted in the third approach, is not very popular in applied ontology. The fourth approach is appealing but it commits to modal realism, whose assumptions are hotly debated in analytic ontology [12]. In addition, since design objects are individual entities, this approach may face some problems when dealing with design tolerances and variants, which are however crucial in engineering design, since a design specification is typically underdetermined. From this perspective, the first two approaches are more flexible; in addition, they both rely on a 3D ontological framework that is standardly used in engineering (see, e.g., [6,10] and the references thereby quoted). Detailing the engineering use case may help to set a suitable benchmark for comparing the approaches and their consequences on related notions.

Acknowledgments: the research presented in the paper is partially supported by the CSA European project H2020 *OntoCommons: Ontology-driven data for industry commons* (GA 958371).

References

1. Fine, K.: Things and their parts. Midwest Studies in Philosophy **23**(1), 61–74 (1999)
2. Galle, P.: Candidate worldviews for design theory. Design Studies **29**(3), 267–303 (2008)
3. Guarino, N.: Artefactual systems, missing components and replaceability. In: Artefact Kinds, pp. 191–206. Springer (2014)
4. Hawley, K.: Temporal Parts. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University (2020)
5. ISO, I.: 15926: Industrial automation systems and integration–integration of life-cycle data for process plants including oil and gas production facilities. Tech. rep., ISO (2004)
6. Karray, M.H., Ameri, F., Hodkiewicz, M., Louge, T.: Romain: Towards a BFO compliant reference ontology for industrial maintenance. Applied Ontology **14**(2), 155–177 (2019)
7. Lewis, D.: On the plurality of worlds, vol. 322. Oxford Blackwell (1986)
8. Masolo, C., Sanfilippo, E.M.: Technical artefact theories: A comparative study and a new empirical approach. Review of Philosophy and Psychology **11**(4), 831–858 (2020)
9. Masolo, C., Vieu, L.: Graph-based approaches to structural universals and complex states of affairs. In: Formal ontology in information systems. Proceedings of the 10th International conference (FOIS). pp. 69–82. IOS press (2018)
10. Sanfilippo, E., Kitamura, Y., Young, R.I.: Formal ontologies in manufacturing. Applied Ontology **14**(2), 119–125 (2019)
11. West, M.: Developing high quality data models. Elsevier (2011)
12. Yagisawa, T.: Possible Objects. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University (2018)

Materials' Tribological Characterisation: an OntoCommons Use Case

Iker Esnaola-Gonzalez[0000−0001−6542−2878], Izaskun Fernandez

TEKNIKER, Basque Research and Technology Alliance (BRTA)
Iñaki Goenaga 5, 20600, Eibar, Spain
iker.esnaola@tekniker.es

Abstract

Tribology is the science that deals with the design, friction, wear, and lubrication of interacting surfaces in relative motion (as in bearings or gears), being one of the key enabling technologies in the development of novel products and the driving of new materials into sustainable solutions within any machine or mechanical system. The goal of one of the OntoCommons project's use case is to help tribologists shortening the time, number and size of experiments required to identify the behaviour of a material or combination of them with respect to specific operation conditions. To do so, semantic technologies and more specifically, ontologies are proposed to abstract tribologists from the underlying vaguely documented experiments' data structures and ease their finding. Furthermore, the mapping of the needed ontologies with relevant information sources will enable developing a holistic Semantic Federated Search service to look for relevant information in additional sources such as patents and repositories containing scientific articles.

1 Introduction

Although the literal translation of the word tribology, which is derived from the Greek, would be the science of rubbing, it is actually understood as the science and technology of interacting surfaces in relative motion and of related subjects and practices [1]. Generally speaking, the tribology is the science that deals with the design, friction, wear, and lubrication of interacting surfaces in relative motion (as in bearings or gears), being one of the key enabling technologies in the development of novel products and the driving of new materials into sustainable solutions within any machine or mechanical system. Wherever moving bodies are in contact with each other, the constitutive materials contributing to friction and wear are also determining the tribological performance of engineering components and machines.

In this context, experiments are carried out towards the understanding of the behaviour of different materials with regards to mechanisms of friction, mechanical wear, chemical wear (corrosion) and the wear-corrosion synergy at different scales. Many of these experiments are similar in terms of the type of materials used. However, this type of experiments are not adequately documented nor publicly available.

The goal of one of the OntoCommons project's[1] use case is to help tribologists shortening the time, number and size of experiments required to identify the behaviour of a material or combination of them (e.g., metal, coating, lubricant) with respect to specific operation conditions. To do so, semantic technologies and more specifically, ontologies are proposed to abstract tribologists from the underlying vaguely documented experiments' data structures and ease their finding.

The rest of the article is structured as follows. Section 2 presents the approach proposed and in Section 3 the next steps are discussed.

2 The proposed approach

In order to support tribologists and shorten the time, number, and size of experiments that they need to perform to identify the behaviour of a given material under certain operation conditions, this section describes the approach proposed by the OntoCommons Use Case 4.

First of all, let us explain that the scenario presented in this article takes place in the context of the i-TRIBOMAT H2020 project[2]. i-TRIBOMAT is aimed at providing an open innovation test bed dedicated to validating and upscaling new materials, thereby enabling intelligent tribological materials characterisation and fostering industrial innovation in the European manufacturing industry, through a completely new, cross-institutional collaborative approaches in sharing infrastructure, competence and data approach.

The modelling of the information is one of the key and basic aspects for a success full sharing approach. Without a common data model, the integration of data coming from different sources becomes almost an unfeasible work: difference in used units, result variable descriptors and configuration parameter names need to be avoided in order to combine data. A well-organised data model can be exploited to find relationships and similarities between different characteristics such as experiment conditions and sample materials. This can be further enhanced with the use of semantic reasoning, complementing the data with information coming from external databases and material research approaches [5].

Since most of the times, the results from tribological experiments are not made open to external stakeholders, in this scenario, other sources of information will also be considered towards helping tribologists. Namely, databases where

[1] `https://ontocommons.eu/`
[2] `https://www.i-tribomat.eu/`

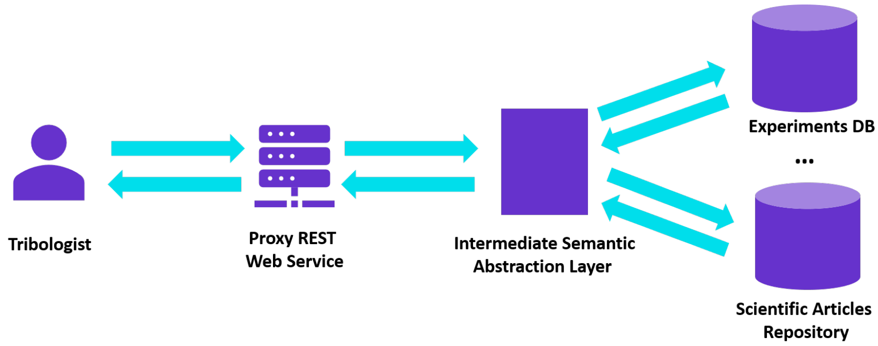


Figure 1: Semantic Federated Search approach

materials' non-tribological information (i.e., mechanical and chemical properties) are stored, as well as patents and repositories containing scientific articles will be targeted. In order to avoid performing manual searches in each of these repositories, the proposed approach advocates for developing an intermediate abstraction layer that enables accessing all this information in a homogeneous way, even dealing with normalisation issues and providing filters for advanced searching options such as alternative materials able to behave in a similar manner under certain conditions. Furthermore, in order to enhance security and abstract the underlying configuration of the whole system, RESTful APIs are considered. Figure 1 summarises the proposed approach.

Semantic Technologies in general, and ontologies in particular, will play a key role in the proposed approach, not only providing a formal and shared representation of the data, but also providing an homogeneous access to heterogeneous data stored with different structures and in different systems via the intermediate abstraction layer. As a matter of fact, once data is annotated with ontological resources, there is no need for the user to be aware of raw data's underlying structure. A common data model for tribological experiments has been proposed in the context of the i-TRIBOMAT project, although this model is not formalised and it could benefit from being aligned to ontologies because of the aforementioned reasons.

In this regard, ontologies for describing the information contained in different storage systems will be necessary. Namely, ontologies for describing tribological experiments, materials information relevant from a tribological point of view, scientific contributions and patents. Following the Semantic Web best practices, the reuse of existing ontologies will be aimed. However, likewise to other domains (e.g. the building domain), not every existing ontology is reusable [3]. This is a direct consequence of neglecting factors that influence the quality of an ontology such as the lack of an explicit license, a proper documentation page or careful metadata with explanatory description of the intended meaning of the ontology terms [4].

Although a thorough analysis of existing ontologies is necessary in order to decide which of them could be reused (if there is such a possibility), there are some ontologies that have already been identified to be considered. The European Materials and Modelling Ontology (EMMO) is an upper ontology to establish semantic standards that apply at the highest possible level of abstraction, under which all conceivable domain ontologies can be subsumed and interoperated [7]. The TribAIn [6] ontology aims to provide a formal and explicit specification of knowledge in the domain of tribology to enable semantic annotation and the search of experimental setups and results. Likewise, other ontologies may be necessary for representing the rest of areas of knowledge such as the BIBO ontology [2] for representing the information gathered from the scientific publications or patents repositories according to best practices of the Semantic Web.

3 Challenges and next steps

After performing a methodical and thorough analysis of the existing ontologies covering the targeted areas of knowledge, this use case will focus on the formalisation of the final ontology that will be the basis for the *Intermediate Semantic Abstraction Layer* (see Figure 1). This task may involve dealing with low-quality ontologies (e.g., ontologies with a scarce documentation or insufficient metadata), as well as the need of extending them to address the use case requirements.

Once the ontology is defined, the next step will consist in the mapping of the ontology with relevant information sources to be able to provide a holistic Semantic Federated Search service.

Acknowledgements

This work is supported by the i-TRIBOMAT and OntoCommons projects which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements no. 814494 and 958371, respectively.

References

[1] Bhushan, B., Ko, P.L.: Introduction to tribology. Appl. Mech. Rev. **56**(1), B6–B7 (2003)

[2] Dimić Surla, B., Segedinac, M., Ivanović, D.: A BIBO ontology extension for evaluation of scientific research results. In: Proceedings of the Fifth Balkan Conference in Informatics. pp. 275–278 (2012)

[3] Esnaola-Gonzalez, I.: On the reusability of ontologies for observations and actuations in buildings. In: Proceedings of the 9th Linked Data in Architecture and Construction Workshop (LDAC 2021) (To be published)

[4] Esnaola-Gonzalez, I., Bermúdez, J., Fernandez, I., Arnaiz, A.: Ontologies for Observations and Actuations in Buildings: A Survey. Semantic Web **11**(4), 593–621 (2020). https://doi.org/10.3233/SW-200378

[5] Hofmann-Amtenbrink, M., Hofmann, H., Koopmans, R., Schuring, E., Meneve, J., Gauss, R., Igartua, A., Fernández, P.: Towards a system approach for materials research, development and innovation for Europe (2019)

[6] Kügler, P., Marian, M., Schleich, B., Tremmel, S., Wartzack, S.: tribAIn—Towards an Explicit Specification of Shared Tribological Understanding. Applied Sciences **10**(13), 4421 (2020). https://doi.org/10.3390/app10134421

[7] Morgado, J.F., Ghedini, E., Goldbeck, G., Hashibon, A., Schmitz, G.J., Friis, J., de Baas, A.: Mechanical testing ontology for digital-twins: A roadmap based on emmo. SeDiT 2020: Semantic Digital Twins 2020 p. 3 (2020)