An Ontology for the Materials Design Domain

Huanyu Li[1,3][0000-0003-1881-3969], Rickard Armiento[2,3][0000-0002-5571-0814], and Patrick Lambrix[1,3] ✉[0000-0002-9084-0470]

[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. In the materials design domain, much of the data from materials calculations are stored in different heterogeneous databases. Materials databases usually have different data models. Therefore, the users have to face the challenges to find the data from adequate sources and integrate data from multiple sources. Ontologies and ontology-based techniques can address such problems as the formal representation of domain knowledge can make data more available and interoperable among different systems. In this paper, we introduce the Materials Design Ontology (MDO), which defines concepts and relations to cover knowledge in the field of materials design. MDO is designed using domain knowledge in materials science (especially in solid-state physics), and is guided by the data from several databases in the materials design field. We show the application of the MDO to materials data retrieved from well-known materials databases.

Resource Type: Ontology
IRI: https://w3id.org/mdo/full/1.0/

Keywords: ontology · materials science · materials design · OPTIMADE · database

1 Introduction

More and more researchers in the field of materials science have realized that data-driven techniques have the potential to accelerate the discovery and design of new materials. Therefore, a large number of research groups and communities have developed data-driven workflows, including data repositories (for an overview see [14]) and task-specific analytical tools. Materials design is a technological process with many applications. The goal is often to achieve a set of desired materials properties for an application under certain limitations in e.g., avoiding or eliminating toxic or critical raw materials. The development of condensed matter theory and materials modeling, has made it possible to achieve quantum mechanics-based simulations that can generate reliable materials data by using computer programs [17]. For instance, in [1] a flow of databases-driven high-throughput materials design in which the database is used to find materials with desirable properties, is shown. A global effort, the Materials Genome Initiative[1], has been proposed to govern databases that contain both experimentally-known and computationally-predicted material properties. The basic idea of this effort is that searching materials databases with desired combinations of properties could help to address some of the challenges of materials design. As these databases are heterogeneous in nature, there are a number of challenges to using them in the materials design workflow. For instance, retrieving data from more than one database means that users have to understand and use different application programming interfaces (APIs) or even different data models to reach an agreement. Nowadays, materials design interoperability is achieved mainly via file-based exchange involving specific formats and, at best, some partial metadata, which is not always adequately documented as it is not guided by an ontology. The second author is closely involved with another ongoing effort, the Open Databases Integration for Materials Design (OPTIMADE[2]) project which aims at making materials databases interoperational by developing a common API. Also this effort would benefit from semantically enabling the system using an ontology, both for search as well as for integrating information from the underlying databases.

These issues relate to the FAIR principles (Findable, Accessible, Interoperable, and Reusable), with the purpose of enabling machines to automatically find and use the data, and individuals to easily reuse the data [23]. Also in the materials science domain, recently, an awareness regarding the importance of such principles for data storage and management is developing and research in this area is starting [6].

To address these challenges and make data FAIR, ontologies and ontology-based techniques have been proposed to play a significant role. For the materials design field there is, therefore, a need for an ontology to represent solid-state physics concepts such as materials' properties, microscopic structure as well as calculations, which are the basis for materials design. Thus, in this paper, we present the Materials Design Ontology (MDO). The development of MDO was guided by the schemas of OPTIMADE as they are based on a consensus reached by several of the materials database providers in the field. Further, we show the use of MDO for data obtained via the OPTIMADE API and via database-specific APIs in the materials design field.

The paper is organized as follows. We introduce some well-known databases and existing ontologies in the materials science domain in Section 2. In Section 3 we present the development of MDO and introduce the concepts, relations and the axiomatization of the ontology. In Section 4 we introduce the envisioned usage of MDO as well as a current implementation. In Section 5 we discuss such

[1] `https://www.mgi.gov/`

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

things as the impact, availability and extendability of MDO as well as future work. Finally, the paper concludes in Section 6 with a small summary.

Availability: MDO is developed and maintained on a GitHub repository[3], and is available from a permanent w3id URL[4].

2 Related Work

In this section we discuss briefly well-known databases as well as ontologies in the materials science field. Further, we briefly introduce OPTIMADE.

2.1 Data and Databases in the Materials Design Domain

In the search for designing new materials, the calculation of electronic structures is an important tool. Calculations take data representing the structure and property of materials as input and generate new such data. A common crystallographic data representation that is widely used by researchers and software vendors for materials design, is CIF[5]. It was developed by the International Union of Crystallography Working Party on Crystallographic Information and was first online in 2006. One of the widely used databases is the Inorganic Crystal Structure Database (ICSD)[6]. ICSD provides data that is used as an important starting point in many calculations in the materials design domain.

As the size of computed data grows, and more and more machine learning and data mining techniques are being used in materials design, frameworks are appearing that not only provide data but also tools. Materials Project, AFLOW and OQMD are well-known examples of such frameworks that are publicly available. **Materials Project** [13] is a central program of the Materials Genome Initiative, focusing on predicting the properties of all known inorganic materials through computations. It provides open web-based data access to computed information on materials, as well as tools to design new materials. To make the data publicly available, the Materials Project provides open Materials API and an open-source python-based programming package (pymatgen). **AFLOW** [4] (Automatic Flow for Materials Discovery) is an automatic framework for high-throughput materials discovery, especially for crystal structure properties of alloys, intermetallics, and inorganic compounds. AFLOW provides a REST API and a python-based programming package (aflow). **OQMD** [19] (The Open Quantum Materials Database) is also a high-throughput database consisting of over 600,000 crystal structures calculated based on density functional theory[7]. OQMD is designed based on a relational data model. OQMD supports a REST API and a python-based programming package (qmpy).

[3] `https://github.com/huanyu-li/Materials-Design-Ontology`

[4] `https://w3id.org/mdo`

[5] Crystallographic Information Framework, `https://www.iucr.org/resources/cif`

[6] `https://icsd.products.fiz-karlsruhe.de/`

[7] `http://oqmd.org`

2.2 Ontologies and Standards

Within the materials science domain, the use of semantic technologies is in its infancy with the development of ontologies and standards. The ontologies have been developed, focusing on representing general materials domain knowledge and specific sub-domains respectively.

Two ontologies representing general materials domain knowledge and to which our ontology connects are ChEBI and EMMO. **ChEBI** [5] (Chemical Entities of Biological Interest) is a freely available data set of molecular entities focusing on chemical compounds. The representation of such molecular entities as atom, molecule ion, etc. is the basis in both chemistry and physics. The ChEBI ontology is widely used and integrated into other domain ontologies. **EMMO** (European Materials & Modelling Ontology) is an upper ontology that is currently being developed and aims at developing a standard representational ontology framework based on current knowledge of materials modeling and characterization. The EMMO development started from the very bottom level, using the actual picture of the physical world coming from applied sciences, and in particular from physics and material sciences. Although EMMO already covers some sub-domains in materials science, many sub-domains are still lacking, including the domain MDO targets.

Further, a number of ontologies from the materials science domain focus on specific sub-domains (e.g., metals, ceramics, thermal properties, nanotechnology), and have been developed with a specific use in mind (e.g., search, data integration) [14]. For instance, the Materials Ontology [2] was developed for data exchange among thermal property databases, and MatOnto ontology [3] for oxygen ion conducting materials in the fuel cell domain. NanoParticle Ontology [21] represents properties of nanoparticles with the purpose of designing new nanoparticles, while the eNanoMapper ontology [11] focuses on assessing risks related to the use of nanomaterials from the engineering point of view. Extensions to these ontologies in the nanoparticle domain are presented in [18]. An ontology that represents formal knowledge for simulation, modeling, and optimization in computational molecular engineering is presented in [12]. Further, an ontology design pattern to model material transformation in the field of sustainable construction, is proposed in [22]. All the materials science domain ontologies above target different sub-domains from MDO.

There are also efforts on building standards for data export from databases and data integration among tools. To some extent the standards formalize the description of materials knowledge and thereby create ontological knowledge. A recent approach is Novel Materials Discovery (NOMAD[8]) [7] of which the metadata structure is defined to be independent of specific material science theory or methods that could be used as an exchange format [9].

[8] `https://www.nomad-coe.eu/externals/european-centres-of-excellence`

2.3 Open Databases Integration for Materials Design

OPTIMADE is a consortium gathering many database providers. It aims at enabling interoperability between materials databases through a common REST API. During the development OPTIMADE takes widely used materials databases such as those introduced in section 2.1 into account. OPTIMADE has a schema that defines the specification of the OPTIMADE REST API and provides essentially a list of terms for which there is a consensus from different database providers. The OPTIMADE API is taken into account in the development of MDO as shown in section 3.

3 The Materials Design Ontology (MDO)

3.1 The development of MDO

The development of MDO followed the NeOn ontology engineering methodology [20]. It consists of a number of scenarios mapped from a set of common ontology development activities. In particular, we focused on applying scenario 1 (*From Specification to Implementation*), scenario 2 (*Reusing and re-engineering non-ontological resources*), scenario 3 (*Reusing ontological resources*) and scenario 8 (*Restructuring ontological resources*). We used OWL2 DL as the representation language for MDO. During the whole process, two knowledge engineers, and one domain expert from the materials design domain were involved. In the remainder of this section, we introduce the key aspects of the development of MDO.

Requirements Analysis. During this step, we clarified the requirements by proposing Use Cases (UC), Competency Questions (CQ) and additional restrictions.

The use cases, which were identified through literature study and discussion between the domain expert and the knowledge engineers based on experience with the development of OPTIMADE and the use of materials science databases, are listed below.

- UC1: MDO will be used for representing knowledge in basic materials science such as solid-state physics and condensed matter theory.
- UC2: MDO will be used for representing materials calculation and standardizing the publication of the materials calculation data.
- UC3: MDO will be used as a standard to improve the interoperability among heterogeneous databases in the materials design domain.
- UC4: MDO will be mapped to OPTIMADE's schema to improve OPTIMADE's search functionality.

The competency questions are based on discussions with domain experts and contain questions that the databases currently can answer as well as questions that experts would want to ask the databases. For instance, CQ1, CQ2, CQ6, CQ7, CQ8 and CQ9 cannot be asked explicitly through the database APIs, although the original downloadable data contains the answers.

- CQ1: What are the calculated properties and their values produced by a materials calculation?
- CQ2: What are the input and output structures of a materials calculation?
- CQ3: What is the space group type of a structure?
- CQ4: What is the lattice type of a structure?
- CQ5: What is the chemical formula of a structure?
- CQ6: For a series of materials calculations, what are the compositions of materials with a specific range of a calculated property (e.g., band gap)?
- CQ7: For a specific material and a given range of a calculated property (e.g., band gap), what is the lattice type of the structure?
- CQ8: For a specific material and an expected lattice type of output structure, what are the values of calculated properties of the calculations?
- CQ9: What is the computational method used in a materials calculation?
- CQ10: What is the value for a specific parameter (e.g., cutoff energy) of the method used for the calculation?
- CQ11: Which software produced the result of a calculation?
- CQ12: Who are the authors of the calculation?
- CQ13: Which software or code does the calculation run with?
- CQ14: When was the calculation data published to the database?

Further, we proposed a list of additional restrictions that help in defining concepts. Some examples are shown below. The full list of additional restrictions can be found at the GitHub repository[9].

- A materials property can relate to a structure.
- A materials calculation has exactly one corresponding computational method.
- A structure corresponds to one specific space group.
- A materials calculation is performed by some software programs or codes.

Reusing and re-engineering non-ontological resources. To obtain the knowledge for building the ontology, we followed two steps: (1) the collection and analysis of non-ontological resources that are relevant to the materials design domain, and (2) discussions with the domain expert regarding the concepts and relationships to be modeled in the ontology. The collection of non-ontological resources comes from: (1) the dictionaries of CIF and International Tables for Crystallography; (2) the APIs from different databases (e.g., Materials Project, AFLOW, OQMD) and OPTIMADE.

Modular development aiming at building design patterns. We identified a pattern related to provenance information in the repository of Ontology Design Patterns (ODPs) that could be reused or re-engineered for MDO. This has led to the reuse of entities in PROV-O [15]. Further, we built MDO in modules considering the possibility for each module to be an ontology design pattern, e.g., the calculation module.

[9] `https://github.com/huanyu-li/Materials-Design-Ontology/blob/master/requirements.md`

Connection and Integration of Existing Ontologies. MDO is connected to EMMO by reusing the concept 'Material', and to ChEBI by reusing the concept 'atom'. Further, we reuse the concepts 'Agent' and 'SoftwareAgent' from PROV-O. In terms of representation of units we reuse the 'Quantity', 'QuantityValue', 'QuantityKind' and 'Unit' concepts from QUDT (Quantities, Units, Dimensions and Data Types Ontologies) [10]. We use the metadata terms from the Dublin Core Metadata Initiative (DCMI)[10] to represent the metadata of MDO.

3.2 Description of MDO

MDO consists of one basic module, *Core*, and two domain-specific modules, *Structure* and *Calculation*, importing the *Core* module. In addition, the *Provenance* module, which also imports *Core*, models provenance information. In total, the OWL2 DL representation of the ontology contains 37 classes, 32 object properties, and 32 data properties. Figure 9 shows an overview of the ontology. The ontology specification is also publicly accessible at w3id.org[11]. The competency questions can be answered using the concepts and relations in the different modules (CQ1 and CQ2 by *Core*, CQ3 to CQ8 by *Structure*, CQ9 and CQ10 by *Calculation*, and CQ11 to CQ14 by *Provenance*).

The **Core** module as shown in Figure 1, consists of the top-level concepts and relations of MDO, which are also reused in other modules. Figure 2 shows the description logic axioms for the *Core* module. The module represents general information of materials calculations. The concepts *Calculation* and *Structure* represent materials calculations and materials' structures, respectively, while *Property* represents materials properties. *Property* is specialized into the disjoint concepts *CalculatedProperty* and *PhysicalProperty* (Core1, Core2, Core3). *Property*, which can be viewed as a quantifiable aspect of one material or materials system, is defined as a sub concept of *Quantity* from QUDT (Core4). *Properties* are also related to *structures* (Core5). When a calculation is applied on materials structures, each *calculation* takes some *structures* and *properties* as input, and may output *structures* and *calculated properties* (Core6, Core7). Further, we use EMMO's concept *Material* and state that each *structure* is related to some *material* (Core8).

The **Structure** module as shown in Figure 3, represents the structural information of materials. Figure 4 shows the description logic axioms for the *Structure* module. Each *structure* has exact one *composition* which represents what chemical elements compose the structure and the ratio of elements in the *structure* (Struc1). The *composition* has different representations of chemical formulas. The *occupancy* of a structure relates the *sites* with the *species*, i.e. the specific chemical elements, that occupy the *site* (Struc2 - Struc5). Each *site* has at most one representation of coordinates in Cartesian format and at most one in fractional format (Struc6, Struc7). The spatial information regarding structures is essential to reflect physical characteristics such as melting point and strength of

[10] `http://purl.org/dc/terms/`

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

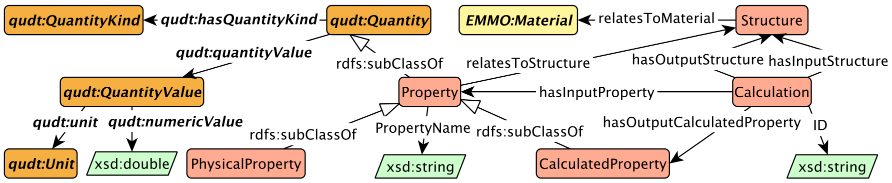

Fig. 1: Concepts and relations in the Core module.

(Core1) $\mathit{CalculatedProperty} \sqsubseteq \mathit{Property}$

(Core2) $\mathit{PhysicalProperty} \sqsubseteq \mathit{Property}$

(Core3) $\mathit{CalculatedProperty} \sqcap \mathit{PhysicalProperty} \sqsubseteq \bot$

(Core4) $\mathit{Property} \sqsubseteq \mathit{Quantity}$

(Core5) $\mathit{Property} \sqsubseteq \forall\ \mathit{relatesToStructure.Structure}$

(Core6) $\mathit{Calculation} \sqsubseteq \exists\ \mathit{hasInputStructure.Structure} \sqcap \forall\ \mathit{hasInputStructure.Structure}$
$\sqcap\ \forall\ \mathit{hasOutputStructure.Structure}$

(Core7) $\mathit{Calculation} \sqsubseteq \exists\ \mathit{hasInputProperty.Property} \sqcap \forall\ \mathit{hasInputProperty.Property}$
$\sqcap\ \forall\ \mathit{hasOutputCalculatedProperty.CalculatedProperty}$

(Core8) $\mathit{Structure} \sqsubseteq \exists\ \mathit{relatesToMaterial.Material} \sqcap \forall\ \mathit{relatesToMaterial.Material}$

Fig. 2: Description logic axioms for the Core module.

materials. To represent this spatial information, we state that each *structure* is represented by some *bases* and a (periodic) *structure* can also be represented by one or more *lattices* (Struc8). Each *basis* and each *lattice* can be identified by one *axis-vectors* set or one *length triple* together with one *angle triple* (Struc9, Struc10). An *axis-vectors* set has three connections to *coordinate vector* representing the coordinates of three translation vectors respectively, which are used to represent a (minimal) repeating unit (Struc11). These three translation vectors are often called a, b, and c. Point groups and space groups are used to represent information of the symmetry of a structure. The *space group* represents a symmetry group of patterns in three dimensions of a *structure* and the *point group* represents a group of linear mappings which correspond to the group of motions in space to determine the symmetry of a *structure*. Each *structure* has one corresponding *space group* (Struc12). Based on the definition from International Tables for Crystallography, each *space group* also has some corresponding *point groups* (Struc13).

The **Calculation** module as shown in Figure 5, represents the classification of different computational methods. Figure 6 shows the description logic axioms for the *Calculation* module. Each *calculation* is achieved by a specific *computational method* (Cal1). Each *computational method* has some *parameters* (Cal2). In the current version of this module, we represent two different methods, the *density functional theory method* and the *HartreeFock method* (Cal3, Cal4). In particular, the density functional theory method is frequently used in materials design to investigate the electronic structure. Such method has at least one

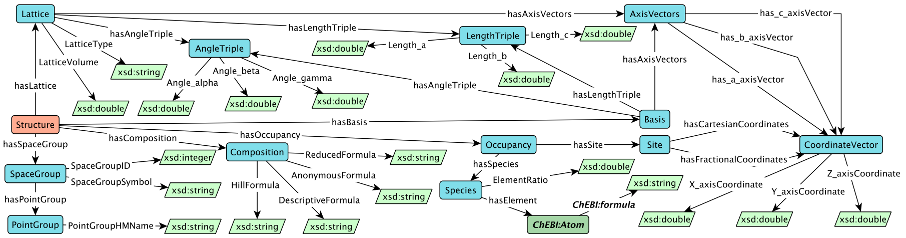

Fig. 3: Concepts and relations in the Structure module.

$$
\begin{array}{ll}
\textit{(Struc1)} & Structure \sqsubseteq\ =1\ hasComposition.Composition \\
 & \sqcap\ \forall\ hasComposition.Composition \\
\textit{(Struc2)} & Structure \sqsubseteq \exists\ hasOccupancy.Occupancy \sqcap \forall\ hasOccupancy.Occupancy \\
\textit{(Struc3)} & Occupancy \sqsubseteq \exists\ hasSpecies.Species \sqcap \forall\ hasSpecies.Species \\
\textit{(Struc4)} & Occupancy \sqsubseteq \exists\ hasSite.Site \sqcap \forall\ hasSite.Site \\
\textit{(Struc5)} & Species \sqsubseteq\ =1\ hasElement.Atom \\
\textit{(Struc6)} & Site \sqsubseteq\ \leq 1\ hasCartesianCoordinates.CoordinateVector \\
 & \sqcap\ \forall\ hasCartesianCoordinates.CoordinateVector \\
\textit{(Struc7)} & Site \sqsubseteq\ \leq 1\ hasFractionalCoordinates.CoordinateVector \\
 & \sqcap\ \forall\ hasFractionalCoordinates.CoordinateVector \\
\textit{(Struc8)} & Structure \sqsubseteq \exists\ hasBasis.Basis \sqcap \forall\ hasBasis.Basis \sqcap \forall\ hasLattice.Lattice \\
\textit{(Struc9)} & Basis \sqsubseteq\ =1\ hasAxisVectors.AxisVectors \sqcup \\
 & (=1\ hasLengthTriple.LengthTriple \sqcap\ =1\ hasAngleTriple.AngleTriple) \\
\textit{(Struc10)} & Lattice \sqsubseteq\ =1\ hasAxisVectors.AxisVectors \sqcup \\
 & (=1\ hasLengthTriple.LengthTriple \sqcap\ =1\ hasAngleTriple.AngleTriple) \\
\textit{(Struc11)} & AxisVectors \sqsubseteq\ =1\ has\_a\_axisVector.CoordinateVector \\
 & \sqcap\ =1\ has\_b\_axisVector.CoordinateVector \\
 & \sqcap\ =1\ has\_c\_axisVector.CoordinateVector \\
\textit{(Struc12)} & Structure \sqsubseteq\ =1\ hasSpaceGroup.SpaceGroup \sqcap \forall\ hasSpaceGroup.SpaceGroup \\
\textit{(Struc13)} & SpaceGroup \sqsubseteq \exists\ hasPointGroup.PointGroup \sqcap \forall\ hasPointGroup.PointGroup
\end{array}
$$

Fig. 4: Description logic axioms for the Structure module.

corresponding *exchange correlation energy functional* (Cal5) which is used to calculate the exchange-correlation energy of a system. There are different kinds of functionals to calculate exchange–correlation energy (Cal6 - Cal11).

The **Provenance** module as shown in Figure 7, represents the provenance information of materials data and calculation. Figure 8 shows the description logic axioms for the *Provenance* module. We reuse part of PROV-O and define a new concept *ReferenceAgent* as a sub-concept of PROV-O's agent (Prov1). We state that each *structure* and *property* can be published by *reference agents* which could be databases or publications (Prov2, Prov3). Each *calculation* is produced by a specific *software* (Prov4).

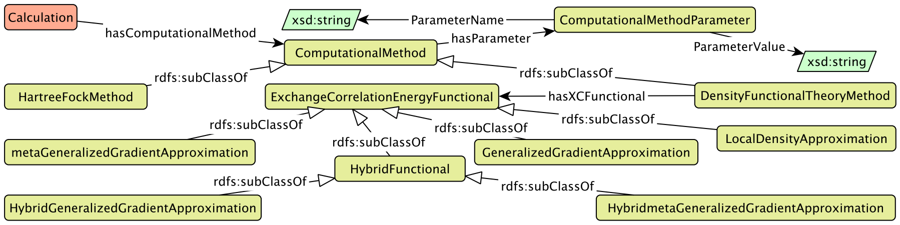

Fig. 5: Concepts and relations in the Calculation module.

(*Cal1*) $\mathit{Calculation} \sqsubseteq\ = 1\ \mathit{hasComputationalMethod.ComputationalMethod}$
(*Cal2*) $\mathit{ComputationalMethod} \sqsubseteq \exists\ \mathit{hasParameter.ComputationalMethodParameter}$
$\sqcap\ \forall\ \mathit{hasParameter.ComputationalMethodParameter}$
(*Cal3*) $\mathit{DensityFunctionalTheoryMethod} \sqsubseteq \mathit{ComputationalMethod}$
(*Cal4*) $\mathit{HartreeFockMethod} \sqsubseteq \mathit{ComputationalMethod}$
(*Cal5*) $\mathit{DensityFunctionalTheoryMethod} \sqsubseteq$
$\exists\ \mathit{hasXCFunctional.ExchangeCorrelationEnergyFunctional}$
$\sqcap\ \forall\ \mathit{hasXCFunctional.ExchangeCorrelationEnergyFunctional}$
(*Cal6*) $\mathit{GeneralizedGradientApproximation} \sqsubseteq \mathit{ExchangeCorrelationEnergyFunctional}$
(*Cal7*) $\mathit{LocalDensityApproximation} \sqsubseteq \mathit{ExchangeCorrelationEnergyFunctional}$
(*Cal8*) $\mathit{metaGeneralizedGradientApproximation} \sqsubseteq$
$\mathit{ExchangeCorrelationEnergyFunctional}$
(*Cal9*) $\mathit{HybridFunctional} \sqsubseteq \mathit{ExchangeCorrelationEnergyFunctional}$
(*Cal10*) $\mathit{HybridGeneralizedGradientApproximation} \sqsubseteq \mathit{HybridFunctional}$
(*Cal11*) $\mathit{HybridmetaGeneralizedGradientApproximation} \sqsubseteq \mathit{HybridFunctional}$

Fig. 6: Description logic axioms for the Calculation module.

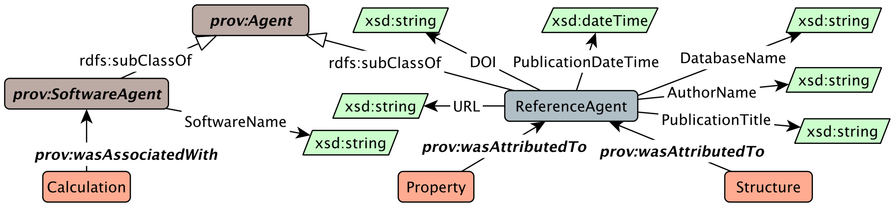

Fig. 7: Concepts and relations in the Provenance module.

(*Prov1*) $\mathit{ReferenceAgent} \sqsubseteq \mathit{Agent}$
(*Prov2*) $\mathit{Structure} \sqsubseteq \forall\ \mathit{wasAttributedTo.ReferenceAgent}$
(*Prov3*) $\mathit{Property} \sqsubseteq \forall\ \mathit{wasAttributedTo.ReferenceAgent}$
(*Prov4*) $\mathit{Calculation} \sqsubseteq \exists\ \mathit{wasAssociatedwith.SoftwareAgent}$

Fig. 8: Description logic axioms for the Provenance module.

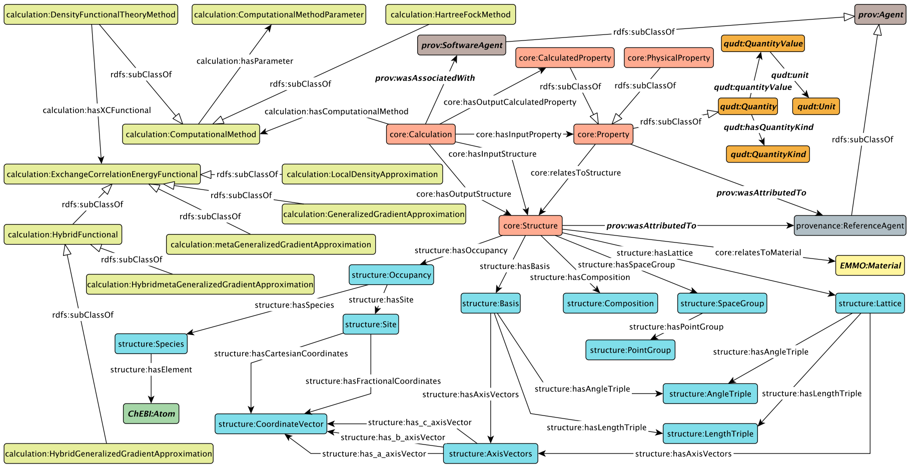

Fig. 9: An overview of MDO.

4 MDO Usage

In Figure 10, we show the vision for the use of MDO for semantic search over OPTIMADE and materials science databases. By generating mappings between MDO and the schemas of materials databases, we can create MDO-enabled query interfaces. The querying can occur, for instance, via MDO-based query expansion, MDO-based mediation or through MDO-enabled data warehouses.

As a proof of concept (full lines in the figure), we created mappings between MDO and the schemas of OPTIMADE and part of Materials Project. Using the mappings we created an RDF data set with data from Materials project. Further, we built a SPARQL query application that can be used to query the RDF data set using MDO terminology. Examples are given below.

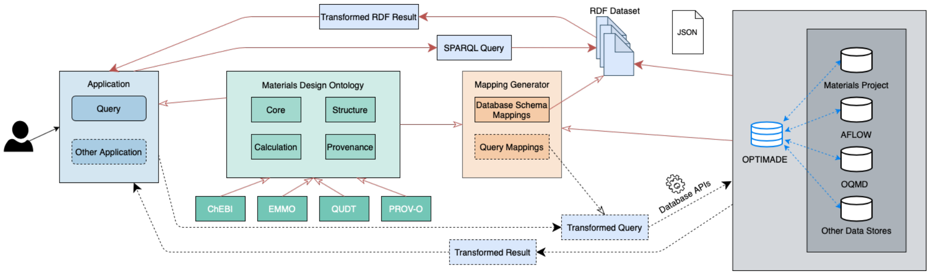

Fig. 10: The vision of the use of MDO. The full-lined components in the figure are currently implemented in a prototype.

Instantiating a materials calculation using MDO. In Figure 11 we exemplify the use of MDO to represent a specific materials calculation and related data in an instantiation. The example is from one of the 85 stable materials published in Materials Project in [8]. The calculation is about one kind of elpasolites, with the composition $Rb_2Li_1Ti_1Cl_6$. To not overcrowd the figure, we only show the instances corresponding to the calculation's output structure, and for multiple calculated properties, species and sites, we only show one instance respectively. Connected to the instances of the Core module's concepts, are instances representing the structural information of the output structure, the provenance information of the output structure and calculated property, and the information about the computational method used for the calculation.

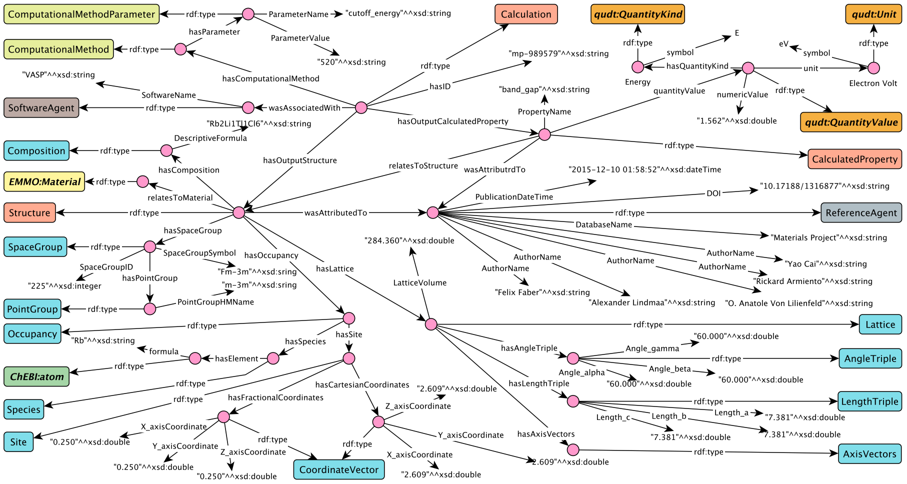

Fig. 11: An instantiated materials calculation.

Mapping the data from a materials database to RDF using MDO. As presented in section 2.1, data from many materials databases are provided through the providers' APIs. A commonly used format is JSON. Our current implementation mapped all JSON data related to the 85 stable materials from [8] to RDF. We constructed the mappings by using SPARQL-Generate [16]. Listing 1.1 shows a simple example on how to write the mappings on 'band gap' which is a *CalculatedProperty*. The result is shown in Listing 1.2. The final RDF dataset contains 42,956 triples. The SPARQL-generate script and the RDF dataset are available from the GitHub repository[12]. This RDF dataset is used for executing SPARQL queries such as the one presented below.

[12] https://github.com/huanyu-li/Materials-Design-Ontology/tree/master/mapping_generator

Listing 1.1: A simple example of mapping

```
BASE <https://w3id.org/mdo/data/1.0/>
PREFIX fun: <http://w3id.org/sparql-generate/fn/>
PREFIX core: <https://w3id.org/mdo/core/>
PREFIX qudt: <http://qudt.org/schema/qudt/
PREFIX qudt_unit: <http://qudt.org/vocab/unit/>

GENERATE {
        ?band_gap_node a core:CalculatedProperty;
        qudt:quantityValue ?band_gap_quantity_value;
        core:hasPropertyName "band_gap"
        GENERATE {
            ?band_gap_quantity a qudt:QuantityValue;
            qudt:unit qudt_unit:EV;
            qudt:numericValue "band_gap"
}.
}
SOURCE <http://example.com/mp-989579_Rb2LiTlCl6.json>
      AS ?source
WHERE {
  BIND(fun:JSONPath(?source,"$.band_gap") AS ?band_gap)
  BIND(BNODE() AS ?band_gap_node)
  BIND(BNODE() AS ?band_gap_quantity_value)
}
```

Listing 1.2: RDF data

```
@prefix  core: <https://w3id.org/mdo/core/> .
@prefix  qudt: <http://qudt.org/schema/qudt/ .
@prefix  qudt_unit: <http://qudt.org/vocab/unit/> .

<https://w3id.org/mdo/data/1.0/mp-989579_band_gap>
        a core:CalculatedProperty ;
        core:hasPropertyName "band_gap" ;
        qudt:quantityValue[ a qudt:QuantityValue ;
        qudt:numericValue 1.5623e0 ;
        qudt:unit  qudt_unit:EV
        ];
```

A SPARQL Query Example. As an example, we show a SPARQL query related to CQ6 in Listing 1.3. The result contains 7 records, which are shown in Table 1. The query is:

– "What are the materials of which the value of band gap is higher than 5eV?" (The result should contain the formula, and the value of band gap.)

Listing 1.3: A SPARQL query example on Materials Project's dataset

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX core: <https://w3id.org/mdo/core/>
PREFIX structure: <https://w3id.org/mdo/structure/>
PREFIX qudt: <http://qudt.org/schema/qudt/>

SELECT ?descriptive_formula ?value WHERE {
  ?calculation rdf:type core:Calculation;
               core:hasOutputCalculatedProperty ?property;
               core:hasOutputStructure ?output_structure.
  ?property qudt:quantityValue ?quantity_value;
            core:hasPropertyName ?name.
  ?quantity_value rdf:type qudt:QuantityValue;
                  qudt:numericValue ?value.
  ?output_structure structure:hasComposition ?composition.
  ?composition structure:hasDescriptiveFormula ?descriptive_formula.
  FILTER (?value>5 && ?name="band_gap")
}
```

Table 1: The result of the query

descriptive formula	value
$Cs_2Rb_1In_1F_6$	5.3759
$Cs_2Rb_1Ga_1F_6$	5.9392
$Cs_2K_1In_1F_6$	5.4629
$Rb_2Na_1In_1F_6$	5.2687
$Cs_2Rb_1Ga_1F_6$	5.5428
$Rb_2Na_1Ga_1F_6$	5.9026
$Cs_2K_1Ga_1F_6$	6.0426

We show more SPARQL query examples and the corresponding result in the GitHub repository[13].

5 Discussion and Future Work

To our knowledge, MDO is the first OWL ontology representing solid-state physics concepts, which are the basis for materials design.

The ontology fills a need for semantically enabling access to and integration of materials databases, and for realizing FAIR data in the materials design field.

[13] https://github.com/huanyu-li/Materials-Design-Ontology/tree/master/sparql_query

This will have a large impact on the effectiveness and efficiency of finding relevant materials data and calculations, thereby augmenting the speed and the quality of the materials design process. Through our connection with OPTIMADE and because of the fact that we have created mappings between MDO and some major materials databases, the potential for impact is large.

The development of MDO followed well-known practices from the ontology engineering point of view (NeOn methodology and modular design). Further, we reused concepts from PROV-O, ChEBI, QUDT and EMMO. A permanent URL is reserved from w3id.org for MDO. MDO is maintained on a GitHub repository from where the ontology in OWL2 DL, visualizations of the ontology and modules, UCs, CQs and restrictions are available. It is licensed via an MIT license[14].

Due to our modular approach MDO can be extended with other modules, for instance, regarding different types of calculations and their specific properties. We identified, for instance, the need for an *X Ray Diffraction* module to model the experimental data of the diffraction used to explore the structural information of materials, and an *Elastic Tensor* module to model data in a calculation that represents a structure's elasticity. We may also refine the current ontology. For instance, it may be interesting to model workflows containing *multiple calculations*.

6 Conclusion

In this paper, we presented MDO, an ontology which defines concepts and relations to cover the knowledge in the field of materials design and which reuses concepts from other ontologies. We discussed the ontology development process showing use cases and competency questions. Further, we showed the use of MDO for semantically enabling materials database search. As a proof of concept, we mapped MDO to OPTIMADE and part of Materials Project and showed querying functionality using SPARQL on a dataset from Materials Project.

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. Armiento, R.: Database-driven high-throughput calculations and machine learning models for materials design. In: Schütt, K.T., Chmiela, S., von Lilienfeld, O.A., Tkatchenko, A., Tsuda, K., Müller, K.R. (eds.) Machine Learning Meets Quantum Physics. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-40245-7_17

[14] `https://github.com/huanyu-li/Materials-Design-Ontology/blob/master/LICENSE`

2. 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
3. Cheung, K., Drennan, J., Hunter, J.: Towards an ontology for data-driven discovery of new materials. In: AAAI Spring Symposium: Semantic Scientific Knowledge Integration. pp. 9–14 (2008)
4. Curtarolo, S., Setyawan, W., Hart, G.L., Jahnatek, M., Chepulskii, R.V., Taylor, R.H., Wang, S., Xue, J., Yang, K., Levy, O., Mehl, M.J., Stokes, H.T., Denis Demchenko, D., Morgan, D.: AFLOW: an automatic framework for high-throughput materials discovery. Computational Materials Science **58**, 218–226 (2012). https://doi.org/10.1016/j.commatsci.2012.02.005
5. Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., Ashburner, M.: ChEBI: a database and ontology for chemical entities of biological interest. Nucleic acids research **36**(suppl_1), D344–D350 (2008). https://doi.org/10.1093/nar/gkm791
6. 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
7. Draxl, C., Scheffler, M.: The NOMAD laboratory: from data sharing to artificial intelligence. Journal of Physics: Materials **2**(3), 036001 (2019). https://doi.org/10.1088/2515-7639/ab13bb
8. Faber, F.A., Lindmaa, A., Von Lilienfeld, O.A., Armiento, R.: Machine learning energies of 2 million elpasolite (a b c 2 d 6) crystals. Physical review letters **117**(13), 135502 (2016). https://doi.org/10.1103/PhysRevLett.117.135502
9. Ghiringhelli, L.M., Carbogno, C., Levchenko, S., Mohamed, F., Huhs, G., Lueders, M., Oliveira, M., Scheffler, M.: Towards a Common Format for Computational Materials Science Data. PSI-K Scientific Highlights **July** (2016)
10. Haas, R., Keller, P.J., Hodges, J., Spivak, J.: Quantities, units, dimensions and data types ontologies (qudt). `http://qudt.org`, accessed: 2020-08-03
11. 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**(1), 10 (2015). https://doi.org/10.1186/s13326-015-0005-5
12. 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
13. Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., Persson, K.: The Materials Project: A materials genome approach to accelerating materials innovation. APL Materials **1**(1), 011002 (2013). https://doi.org/10.1063/1.4812323
14. Lambrix, P., Armiento, R., Delin, A., Li, H.: Big semantic data processing in the materials design domain. In: Sakr, S., Zomaya, A.Y. (eds.) Encyclopedia of Big Data Technologies. Springer (2019). https://doi.org/10.1007/978-3-319-63962-8_293-1
15. Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: Prov-o: The prov ontology. `https://www.w3.org/TR/prov-o/` (2013), accessed: 2020-04

16. Lefrançois, M., Zimmermann, A., Bakerally, N.: A SPARQL extension for generating RDF from heterogeneous formats. In: European Semantic Web Conference. pp. 35–50. Springer (2017). https://doi.org/10.1007/978-3-319-58068-5_3
17. Lejaeghere, K., Bihlmayer, G., Björkman, T., Blaha, P., Blügel, S., Blum, V., Caliste, D., Castelli, I.E., Clark, S.J., Dal Corso, A., de Gironcoli, S., Deutsch, T., Dewhurst, J.K., Di Marco, I., Draxl, C., Dulak, M., Eriksson, O., Flores-Livas, J.A., Garrity, K.F., Genovese, L., Giannozzi, P., Giantomassi, M., Goedecker, S., Gonze, X., Grånäs, O., Gross, E.K.U., Gulans, A., Gygi, F., Hamann, D.R., Hasnip, P.J., Holzwarth, N.A.W., Iusan, D., Jochym, D.B., Jollet, F., Jones, D., Kresse, G., Koepernik, K., Kücükbenli, E., Kvashnin, Y.O., Locht, I.L.M., Lubeck, S., Marsman, M., Marzari, N., Nitzsche, U., Nordström, L., Ozaki, T., Paulatto, L., Pickard, C.J., Poelmans, W., Probert, M.I.J., Refson, K., Richter, M., Rignanese, G.M., Saha, S., Scheffler, M., Schlipf, M., Schwarz, K., Sharma, S., Tavazza, F., Thunström, P., Tkatchenko, A., Torrent, M., Vanderbilt, D., van Setten, M.J., Speybroeck, V.V., Wills, J.M., Yates, J.R., Zhang, G.X., Cottenier, S.: Reproducibility in density functional theory calculations of solids. Science **351**(6280), aad3000 (2016). https://doi.org/10.1126/science.aad3000
18. 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
19. Saal, J.E., Kirklin, S., Aykol, M., Meredig, B., Wolverton, C.: Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). Jom **65**(11), 1501–1509 (2013). https://doi.org/10.1007/s11837-013-0755-4
20. Suárez-Figueroa, M.C., Gómez-Pérez, A., Fernández-López, M.: The neon methodology for ontology engineering. In: Ontology engineering in a networked world, pp. 9–34. Springer (2012). https://doi.org/10.1007/978-3-642-24794-1_2
21. 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
22. Vardeman II, C.F., Krisnadhi, A.A., Cheatham, M., Janowicz, K., Ferguson, H., Hitzler, P., Buccellato, A.P.C.: An ontology design pattern and its use case for modeling material transformation. Semantic Web **8**(5), 719–731 (2017). https://doi.org/10.3233/SW-160231
23. 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