Ontology-Based Technologies for Data Access and Integration in the Materials Design Domain

Huanyu Li

Main Supervisor: Patrick Lambrix
Secondary Supervisors: Nahid Shahmehri, Olaf Hartig, Rickard Armiento

LINKÖPING UNIVERSITY

Agenda

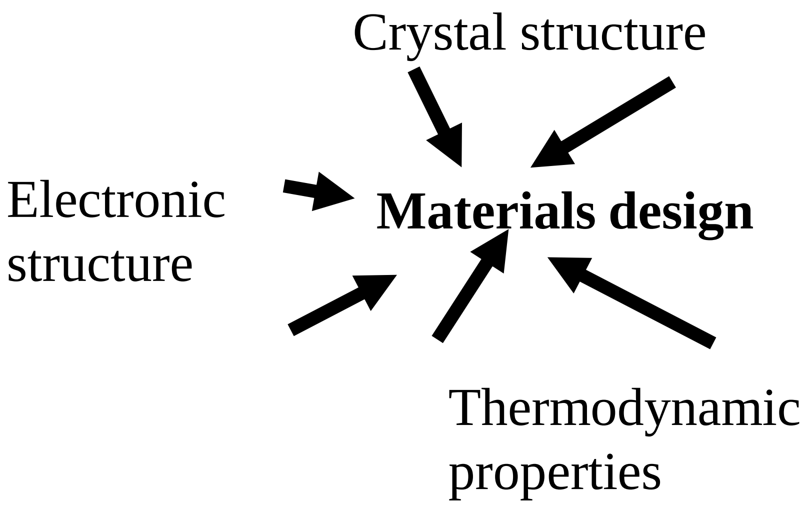
- Research Background
- Motivation
- Problem Formulation
- Finished and Ongoing Work
  - ➢ An Ontology for the Materials Design Domain
  - ➢ A Method for Extending Ontologies with Applications to the Materials Science Domain
  - ➢ Ontology-Driven GraphQL Server Generation for Data Access and Integration
- Conclusion and Future Work

Research Background

LINKÖPING UNIVERSITY

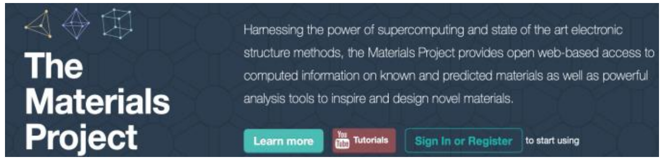
Research Background

- Materials design, design materials with desired properties (e.g., non-toxic, high strength)
- The space of potentially useful materials yet to be discovered is immense (e.g., many billions of possible combinations up to 6 different chemical elements)
  - Exploring by experimental techniques
  - Simulating by computer programs to generate reliable materials data
- Materials calculations run such computer programs with different computational methods



Research Background

- Many software programs can do materials calculations
- A lot of databases provide materials calculation data via Web APIs (e.g., REST APIs)
  - Materials databases are heterogenous in nature
  - APIs follow different data schemas
- A data-driven workflow of materials design will search these databases with desired combinations of properties



Querying Materials Databases

- Searching 'GaN' in Materials Project and OQMD
  - Different number of fields returned, different terminology for the same concept
  - Some semantics could be added
- OPTIMADE (Open Databases Integration for Materials Design)
  - To design a common API

The result from Materials Project

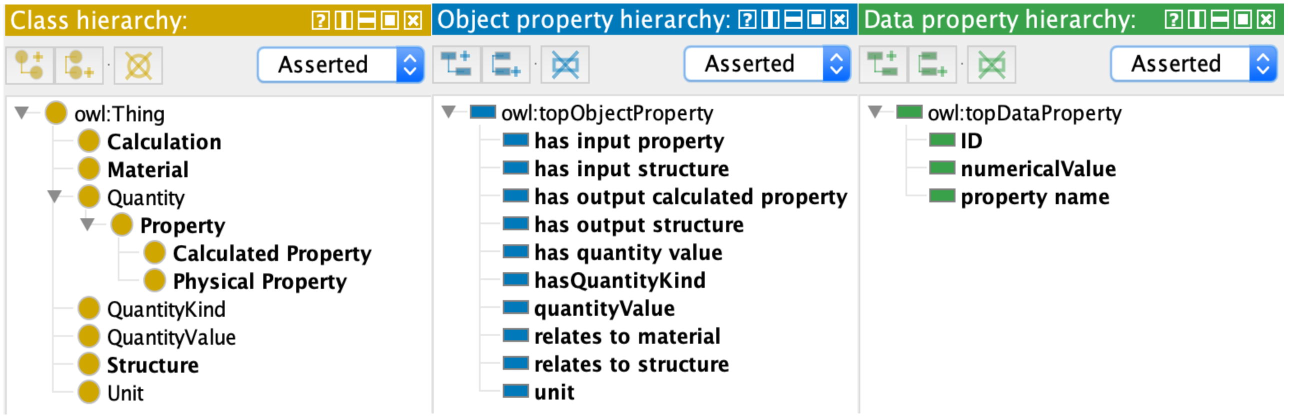
Materials Id	Formula	Spacegroup	Formation Energy (eV)	E Above Hull (eV)	Band Gap (eV)	Volume	Nsites	Density (gm/cc)
mp-830	GaN	$F\bar{4}3m$	-0.657	0.005	1.563	23.48	2	5.921
mp-2853	GaN	$Fm\bar{3}m$	-0.184	0.478	0.395	19.47	2	7.141
mp-1007824	GaN	$P6_3/mmc$	-0.307	0.355	1.105	75.361	4	3.69
mp-804	GaN	$P6_3mc$	-0.663	0	1.743	46.943	4	5.924
mp-1181864	GaN	$P6_3mc$	0.11	0.773	0.671	92.509	8	6.012

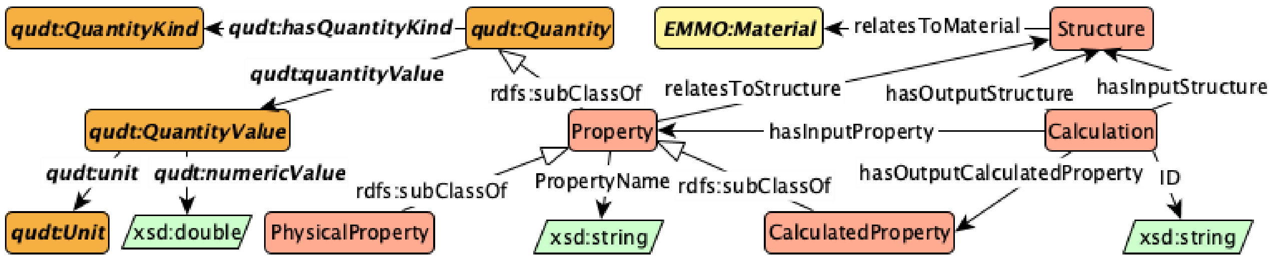
The result from OQMD

ID	Composition	Spacegroup	Formation Energy [eV/atom]	Prototype	# of atoms	Reference
6450	GaN	P63mc	-0.579	ZnS(2H)	4	Schulz, H., Thiemann, K.h.(1977). Crystal structure refinement of Al N and Ga N. Solid State Communications, 23.
97045	GaN	P63mc	-0.579	ZnS(2H)	4	Ren, Y., Gateshki, M., Petkov, V., Choi, J., Gillan, E.g.(2005). Structure of nanocrystalline Ga N from X-ray diffraction, Rietveld and atomic pair distribution function analyses. Journal of Materials Chemistry, 15.
1105884	GaN	Fm-3m	-0.101	B1_NaCl	2	
1234952	GaN	R-3m	-0.100	L1_1_CuPt	2	
1229533	GaN	P-6m2	0.367	B_h_WC	2	
306699	GaN	Pm-3m	1.040	B2_CsCl	2	
101873	GaN	P63mc			8	Gajbhiye, N.s., Bhattacharyya, S., Shivaprasad, S.m.(2008). Synthesis and characterization of epsilon-(Fe3 N)/Ga N, 54/46-composite nanowires. Materials Research Bulletin, 43.
1220368	GaN	Pmma		B19_AuCd	4	

Semantics are needed

- What ontologies can do
  - ✓ Standardized terminology
  - ✓ Relationships between terms
  - ✓ Mapping between terminologies
  - ✓ Making data FAIR
    - Findability
    - Accessibility
    - Interoperability
    - Reusability
- Ontology Construction
- Ontology Extension





Motivation

LINKÖPING
UNIVERSITY

Motivation

- A need for a unified way for querying data among multiple databases
  - Materials design interoperability largely happens via the file-based exchange with certain formats, which is not guided by an ontology
  - Data needs to be FAIR (Findability, Accessibility, Interoperability, Reusability)
- Few methods using existing data as domain knowledge to extend ontologies (semi-)automatically
- The lack of methods accessing and integrating data over multiple heterogeneous data sources where data is shared via Web APIs (e.g., REST APIs)
  - Current ontologies related to materials science do not cover knowledge of materials design (e.g., materials calculations)
  - Traditional ontology-based data access and integration focus on relational data

Problem Formulation

LINKÖPING
UNIVERSITY

Problem Formulation

- RQ1: How can materials domain ontologies be extended by mining unstructured text?
- RQ2: How can ontology-based techniques be used to access and integrate data from heterogeneous sources?
  - RQ2.1: How can ontologies be leveraged to generate GraphQL APIs for data access and data integration?
  - RQ2.2: How to enable ontology-based data access and integration in the materials design domain?

Finished Work and Ongoing Work

Finished and Ongoing Work

- [1] Lambrix P, Armiento R, Delin A, **Li H**, ***Big Semantic Data Processing in the Materials Design Domain***, Encyclopedia of Big Data Technologies, Springer, 2018.
  - [*] Update version: Lambrix P, Armiento R, Delin A, **Li H**, ***FAIR Big Data in the Materials Design Domain***
- [2] **Li H**, Armiento R, Lambrix P, ***An Ontology for the Materials Design Domain***, The 19th International Semantic Web Conference, Athens Greece (Virtual Conference), 2020.
- [3] **Li H**, Armiento R, Lambrix P, ***A Method for Extending Ontologies with Application to the Materials Science Domain***, Data Science Journal, 2019.
- [4] **Li H**, Hartig O, Armiento R, Lambrix P, ***Ontology-Driven GraphQL Server Generation for Data Access and Integration*** (Ongoing).

Other Publications

- [5] **Li H**, Dragisic Z, Faria D, Ivanova V, Jimenez-Ruiz E, Lambrix P, Pesquita C, ***User validation in ontology alignment: functional assessement and impact***, The Knowledge Engineering Review, 2019.
- [6] Dragisic Z, Ivanova V, **Li H**, Lambrix P, ***Experiences from the anatomy track in the ontology alignment evaluation initiative***, Journal of Biomedical Semantics 8:56, 2017.
- [7] **Li H**, Armiento R, Lambrix P, ***Extending Ontologies in the Nanotechnology Domain using Topic Models and Formal Topical Concept Analysis on Unstructured Text***, ISWC 2019 Posters & Demonstrations, Industry and Outrageous Ideas Tracks, CEUR, 2019.
- [8] Keskisärkkä R, **Li H**, Cheng S, Carlsson N, Lambrix P, ***An Ontology for Ice Hockey***, ISWC 2019 Posters & Demonstrations, Industry and Outrageous Ideas Tracks, CEUR, 2019.
- [9-13] Co-authored in synthesis papers of Results of the Ontology Alignment Evaluation Initiative from 2016 to 2020

Overview

- Materials Design Ontology (MDO) is a domain ontology for the materials design field
  - To provide a conceptual model over multiple databases in the field
  - One of the contributions to address the RQ2
    - How can ontology-based techniques be used to access and integrate data from heterogeneous sources?

The development of MDO

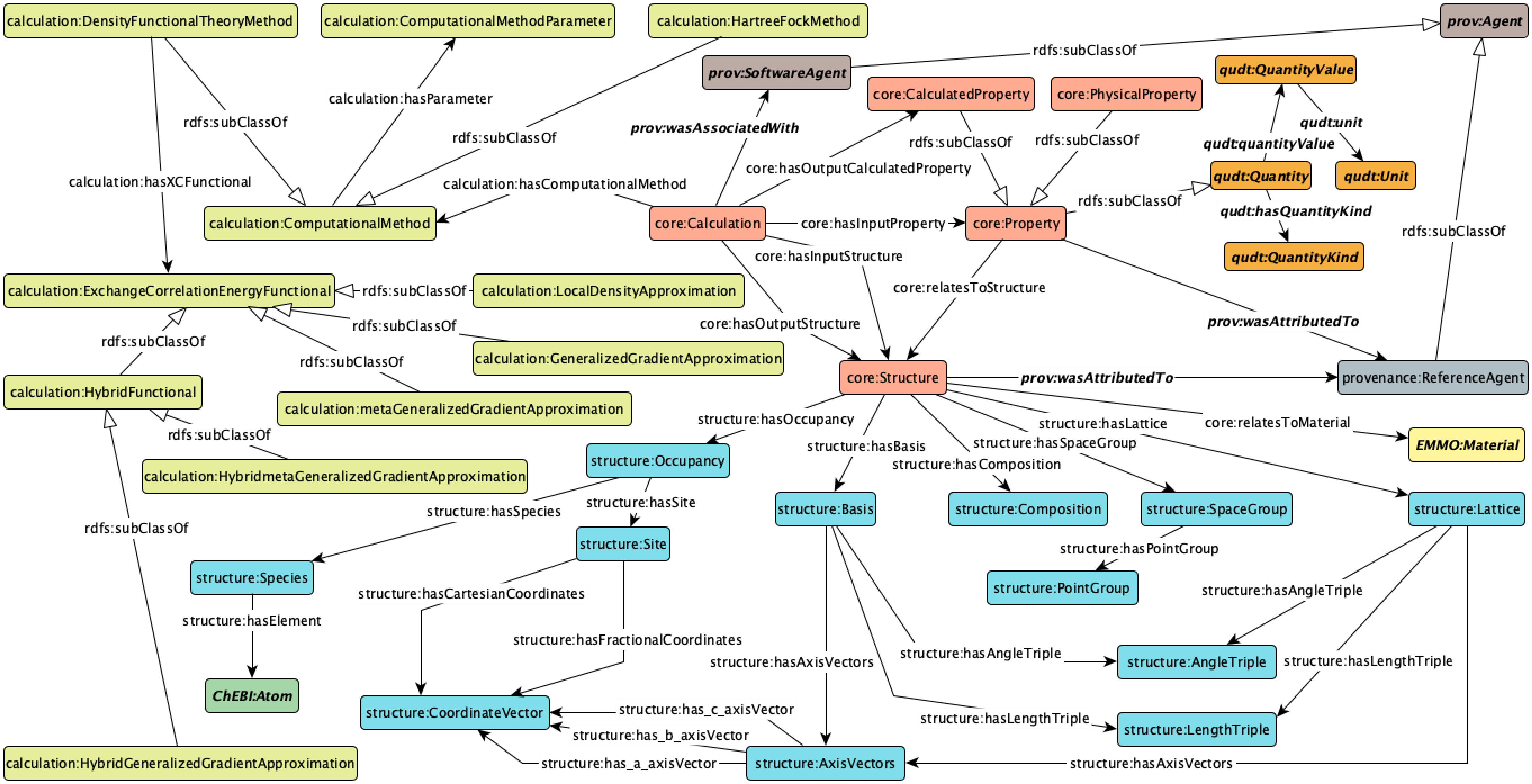
- NeOn methodology in ontology engineering
  - Requirement analysis (Use Cases, Competency Questions, Additional Restrictions)
  - Reusing concepts from existing ontologies
    - PROV-O (PROVenance Ontology), QUDT (Quantities, Units, Dimensions, and Type Ontology)
    - ChEBI (Chemical Entities of Biological Interest Ontology), EMMO (European Materials Modelling Ontology)
- Modular Design
  - Core Module, Structure Module, Calculation Module, Provenance Module
- Discussions with domain expert

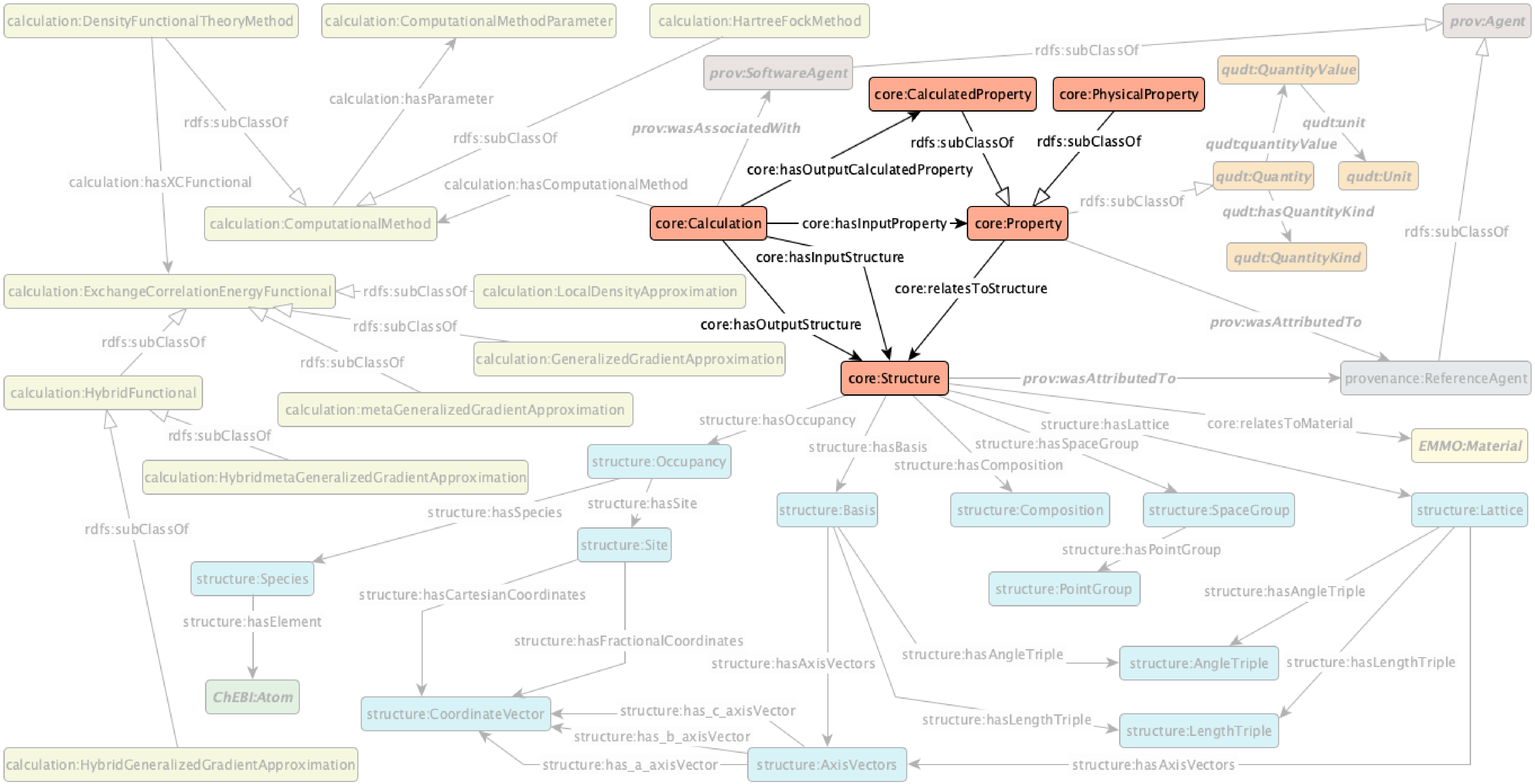
MDO-Use Cases

- MDO can represent knowledge in basic materials science
- MDO can represent materials calculations and standardize the publication of materials calculation data
- MDO can provide a standard to improve the interoperability among heterogeneous databases in the materials design domain
- MDO can map OPTIMADE's schema for enriching its search functionality

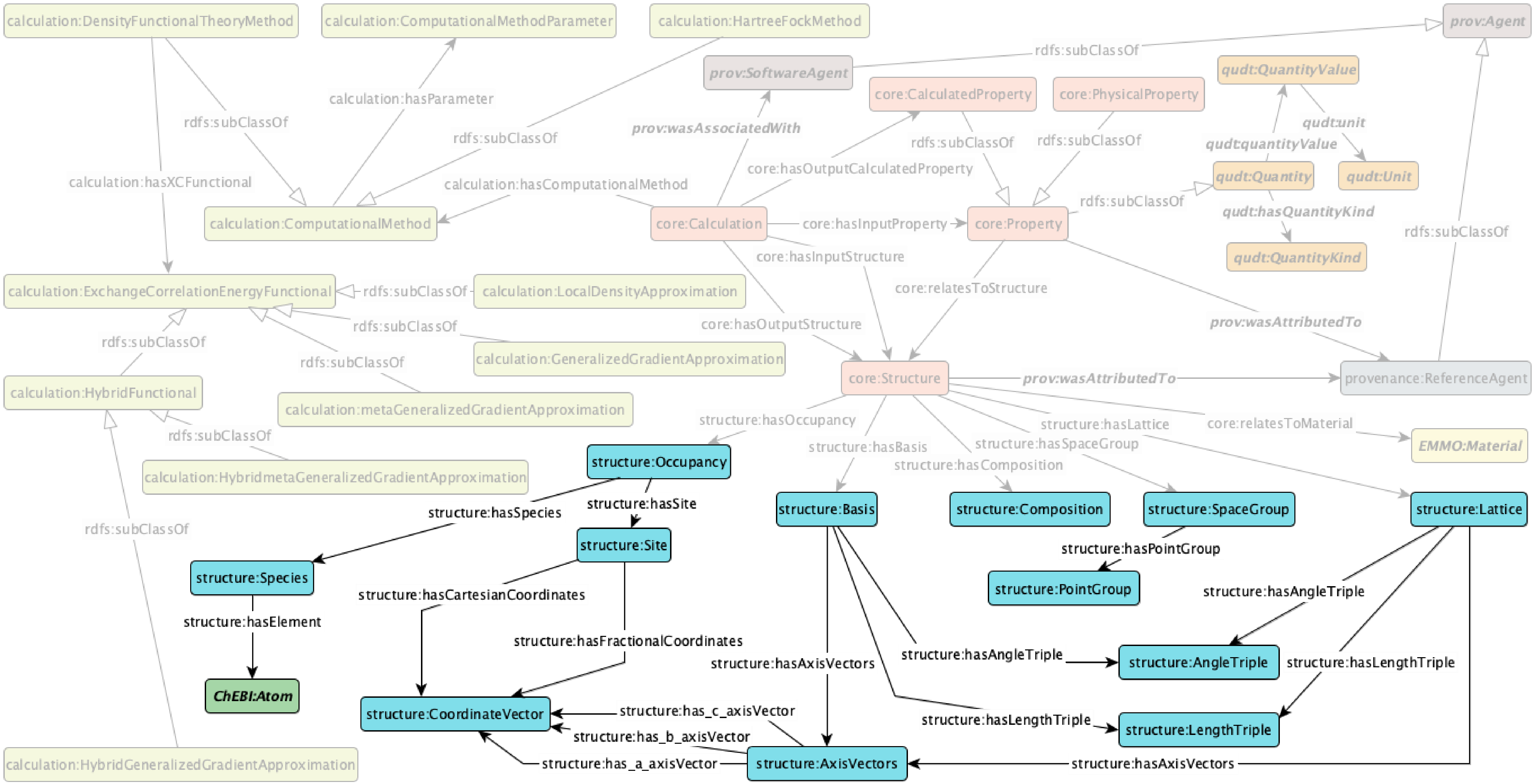
MDO-Competency Questions

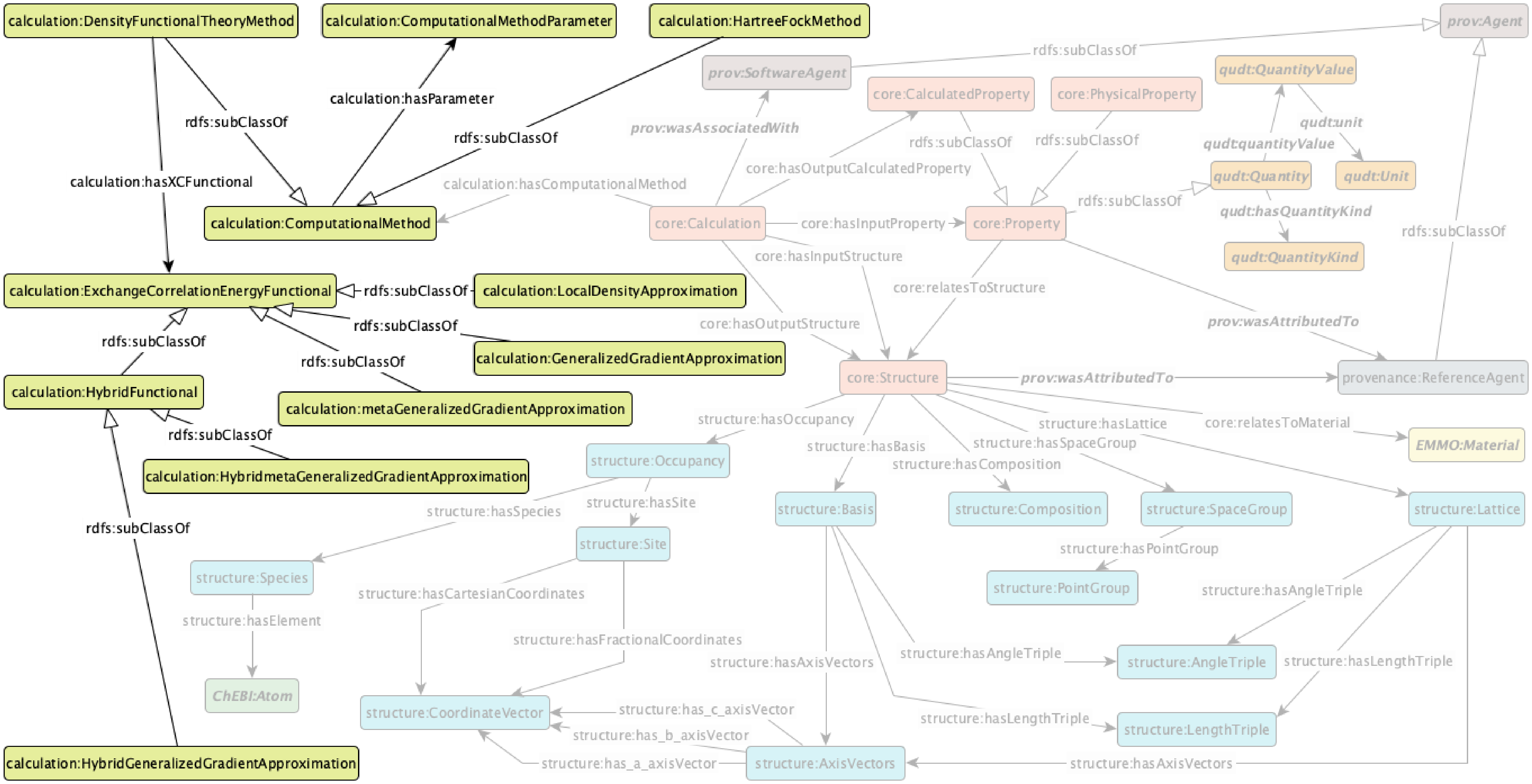
- Materials databases' APIs can /cannot answer, e.g.,:
  - ✓ "What is the chemical formula of a material in the calculation?"
  - ✓ "Which software produced the result of a calculation?"
  - ? "What are the input and output structures of a materials calculation?"
    - e.g., OQMD's API can answer structural information in terms of *'unit_cell'*, *'space group' and 'composition_generic'* but it does not reflect semantic relationships of such terms and *'Structure'* in the API schema
  - ? "For a series of materials calculations, what are the compositions of materials with a specific range of a calculated property (e.g., band gap)?"
    - e.g., Materials Project's API can provide *'anonymous_formula'*, *'full_formula'*, and *'pretty_formula'* for querying but it does not cover the domain knowledge that a *Composition* of a material has several attributes such as *Anonymous Formula, Full Formula*.


calculation:DensityFunctionalTheoryMethod
calculation:ComputationalMethodParameter
calculation:HartreeFockMethod
prov:Agent
prov:SoftwareAgent
rdfs:subClassOf
core:CalculatedProperty
core:PhysicalProperty
qudt:QuantityValue
qudt:unit
qudt:quantityValue
qudt:Quantity
qudt:Unit
qudt:hasQuantityKind
qudt:QuantityKind
calculation:hasParameter
prov:wasAssociatedWith
core:hasOutputCalculatedProperty
calculation:hasXCFunctional
calculation:hasComputationalMethod
calculation:ComputationalMethod
core:Calculation
core:hasInputProperty
core:Property
core:hasInputStructure
core:relatesToStructure
calculation:ExchangeCorrelationEnergyFunctional
calculation:LocalDensityApproximation
calculation:GeneralizedGradientApproximation
calculation:HybridFunctional
calculation:metaGeneralizedGradientApproximation
calculation:HybridmetaGeneralizedGradientApproximation
calculation:HybridGeneralizedGradientApproximation
core:hasOutputStructure
prov:wasAttributedTo
core:Structure
provenance:ReferenceAgent
structure:hasOccupancy
structure:hasBasis
structure:hasLattice
structure:hasSpaceGroup
structure:hasComposition
core:relatesToMaterial
EMMO:Material
structure:Occupancy
structure:hasSite
structure:hasSpecies
structure:Basis
structure:Composition
structure:SpaceGroup
structure:Lattice
structure:Site
structure:hasPointGroup
structure:PointGroup
structure:Species
structure:hasCartesianCoordinates
structure:hasAngleTriple
structure:hasElement
structure:hasFractionalCoordinates
structure:hasAxisVectors
structure:AngleTriple
structure:hasLengthTriple
ChEBI:Atom
structure:CoordinateVector
structure:has_c_axisVector
structure:has_b_axisVector
structure:has_a_axisVector
structure:LengthTriple
structure:AxisVectors

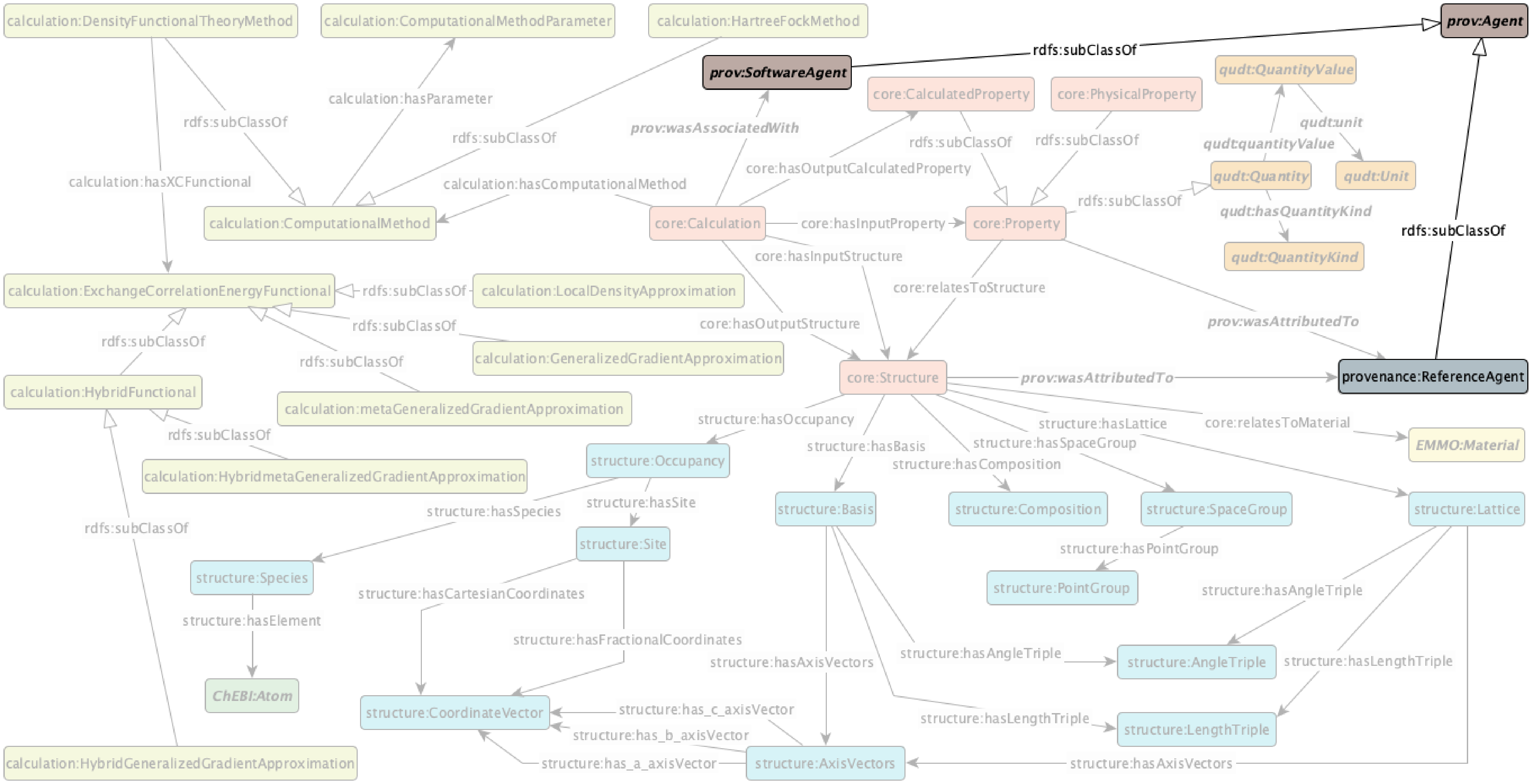


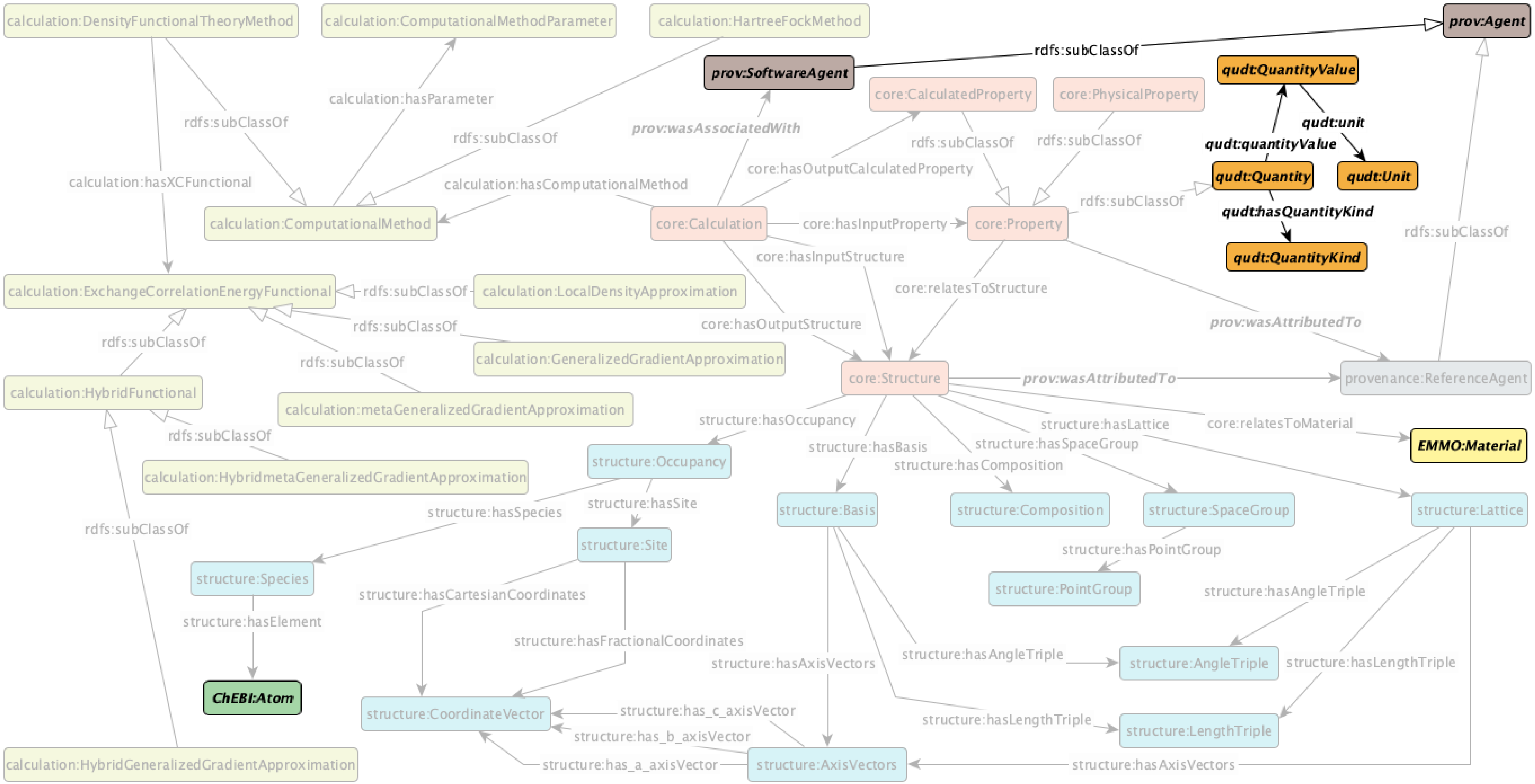
LINKÖPING UNIVERSITY


calculation:DensityFunctionalTheoryMethod
calculation:ComputationalMethodParameter
calculation:HartreeFockMethod
prov:Agent
prov:SoftwareAgent
core:CalculatedProperty
core:PhysicalProperty
qudt:QuantityValue
qudt:Quantity
qudt:Unit
qudt:QuantityKind
calculation:ComputationalMethod
core:Calculation
core:Property
calculation:ExchangeCorrelationEnergyFunctional
calculation:LocalDensityApproximation
calculation:GeneralizedGradientApproximation
calculation:HybridFunctional
calculation:metaGeneralizedGradientApproximation
calculation:HybridmetaGeneralizedGradientApproximation
calculation:HybridGeneralizedGradientApproximation
core:Structure
provenance:ReferenceAgent
EMMO:Material
structure:Occupancy
structure:Site
structure:Species
ChEBI:Atom
structure:CoordinateVector
structure:Basis
structure:Composition
structure:SpaceGroup
structure:PointGroup
structure:Lattice
structure:AngleTriple
structure:LengthTriple
structure:AxisVectors
rdfs:subClassOf
calculation:hasParameter
calculation:hasXCFunctional
calculation:hasComputationalMethod
prov:wasAssociatedWith
core:hasOutputCalculatedProperty
core:hasInputProperty
core:hasInputStructure
core:hasOutputStructure
core:relatesToStructure
prov:wasAttributedTo
qudt:unit
qudt:quantityValue
qudt:hasQuantityKind
core:relatesToMaterial
structure:hasOccupancy
structure:hasBasis
structure:hasComposition
structure:hasSpaceGroup
structure:hasLattice
structure:hasSpecies
structure:hasSite
structure:hasElement
structure:hasCartesianCoordinates
structure:hasFractionalCoordinates
structure:hasPointGroup
structure:hasAngleTriple
structure:hasLengthTriple
structure:hasAxisVectors
structure:has_c_axisVector
structure:has_b_axisVector
structure:has_a_axisVector

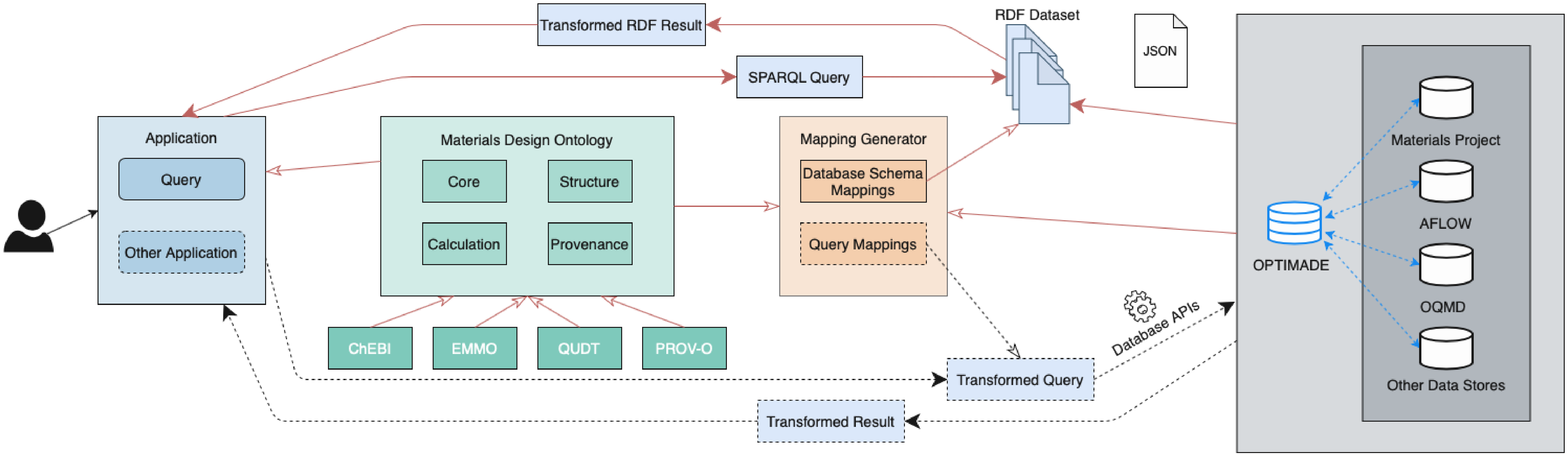


LINKÖPING UNIVERSITY

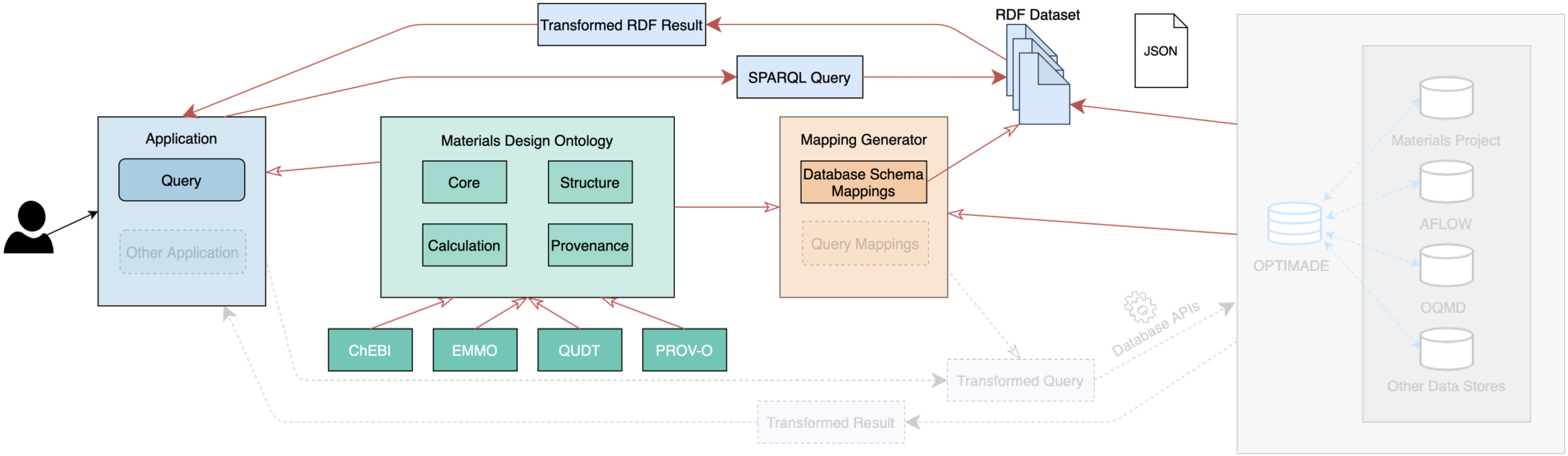

calculation:DensityFunctionalTheoryMethod
calculation:ComputationalMethodParameter
calculation:HartreeFockMethod
prov:Agent
rdfs:subClassOf
prov:SoftwareAgent
qudt:QuantityValue
core:CalculatedProperty
core:PhysicalProperty
calculation:hasParameter
prov:wasAssociatedWith
qudt:unit
qudt:quantityValue
qudt:Quantity
qudt:Unit
calculation:hasXCFunctional
core:hasOutputCalculatedProperty
calculation:hasComputationalMethod
qudt:hasQuantityKind
calculation:ComputationalMethod
core:Calculation
core:hasInputProperty
core:Property
qudt:QuantityKind
core:hasInputStructure
core:relatesToStructure
calculation:ExchangeCorrelationEnergyFunctional
calculation:LocalDensityApproximation
core:hasOutputStructure
prov:wasAttributedTo
calculation:GeneralizedGradientApproximation
calculation:HybridFunctional
core:Structure
provenance:ReferenceAgent
calculation:metaGeneralizedGradientApproximation
structure:hasOccupancy
structure:hasLattice
core:relatesToMaterial
EMMO:Material
structure:hasBasis
structure:hasSpaceGroup
structure:hasComposition
structure:Occupancy
calculation:HybridmetaGeneralizedGradientApproximation
structure:hasSpecies
structure:hasSite
structure:Basis
structure:Composition
structure:SpaceGroup
structure:Lattice
structure:Site
structure:hasPointGroup
structure:Species
structure:PointGroup
structure:hasCartesianCoordinates
structure:hasAngleTriple
structure:hasElement
structure:hasFractionalCoordinates
structure:hasAxisVectors
structure:AngleTriple
structure:hasLengthTriple
ChEBI:Atom
structure:CoordinateVector
structure:has_c_axisVector
structure:has_b_axisVector
structure:LengthTriple
calculation:HybridGeneralizedGradientApproximation
structure:has_a_axisVector
structure:AxisVectors


calculation:DensityFunctionalTheoryMethod
calculation:ComputationalMethodParameter
calculation:HartreeFockMethod
prov:Agent
rdfs:subClassOf
prov:SoftwareAgent
qudt:QuantityValue
core:CalculatedProperty
core:PhysicalProperty
calculation:hasParameter
qudt:unit
prov:wasAssociatedWith
qudt:quantityValue
calculation:hasXCFunctional
core:hasOutputCalculatedProperty
qudt:Quantity
qudt:Unit
calculation:hasComputationalMethod
calculation:ComputationalMethod
core:Calculation
core:hasInputProperty
core:Property
qudt:hasQuantityKind
qudt:QuantityKind
core:hasInputStructure
calculation:ExchangeCorrelationEnergyFunctional
calculation:LocalDensityApproximation
core:relatesToStructure
core:hasOutputStructure
prov:wasAttributedTo
calculation:GeneralizedGradientApproximation
core:Structure
provenance:ReferenceAgent
calculation:HybridFunctional
calculation:metaGeneralizedGradientApproximation
structure:hasOccupancy
structure:hasLattice
core:relatesToMaterial
EMMO:Material
structure:hasBasis
structure:hasSpaceGroup
structure:Occupancy
structure:hasComposition
calculation:HybridmetaGeneralizedGradientApproximation
structure:hasSite
structure:hasSpecies
structure:Basis
structure:Composition
structure:SpaceGroup
structure:Lattice
structure:Site
structure:hasPointGroup
structure:Species
structure:PointGroup
structure:hasCartesianCoordinates
structure:hasAngleTriple
structure:hasElement
structure:hasFractionalCoordinates
structure:hasAxisVectors
structure:AngleTriple
structure:hasLengthTriple
ChEBI:Atom
structure:has_c_axisVector
structure:CoordinateVector
structure:has_b_axisVector
structure:LengthTriple
structure:has_a_axisVector
structure:AxisVectors
calculation:HybridGeneralizedGradientApproximation

The Vision of MDO's Usage



MDO's Usage-Querying over a mapped RDF dataset



MDO's Usage-Querying over a mapped RDF dataset

- SPARQL-Generate is used to define mappings
- An example query is "What are the materials of which the value of band gap is higher than 5eV (electron volt)?"

```
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")
}
```

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

The availability of MDO

- Open Source and Open Access
  - On Github project: https://github.com/huanyu-li/Materials-Design-Ontology
  - On W3ID: https://w3id.org/mdo/full/1.0/
  - On LOV (Linked Open Vocabularies): https://lov.linkeddata.es/dataset/lov/vocabs/mdo

Summary

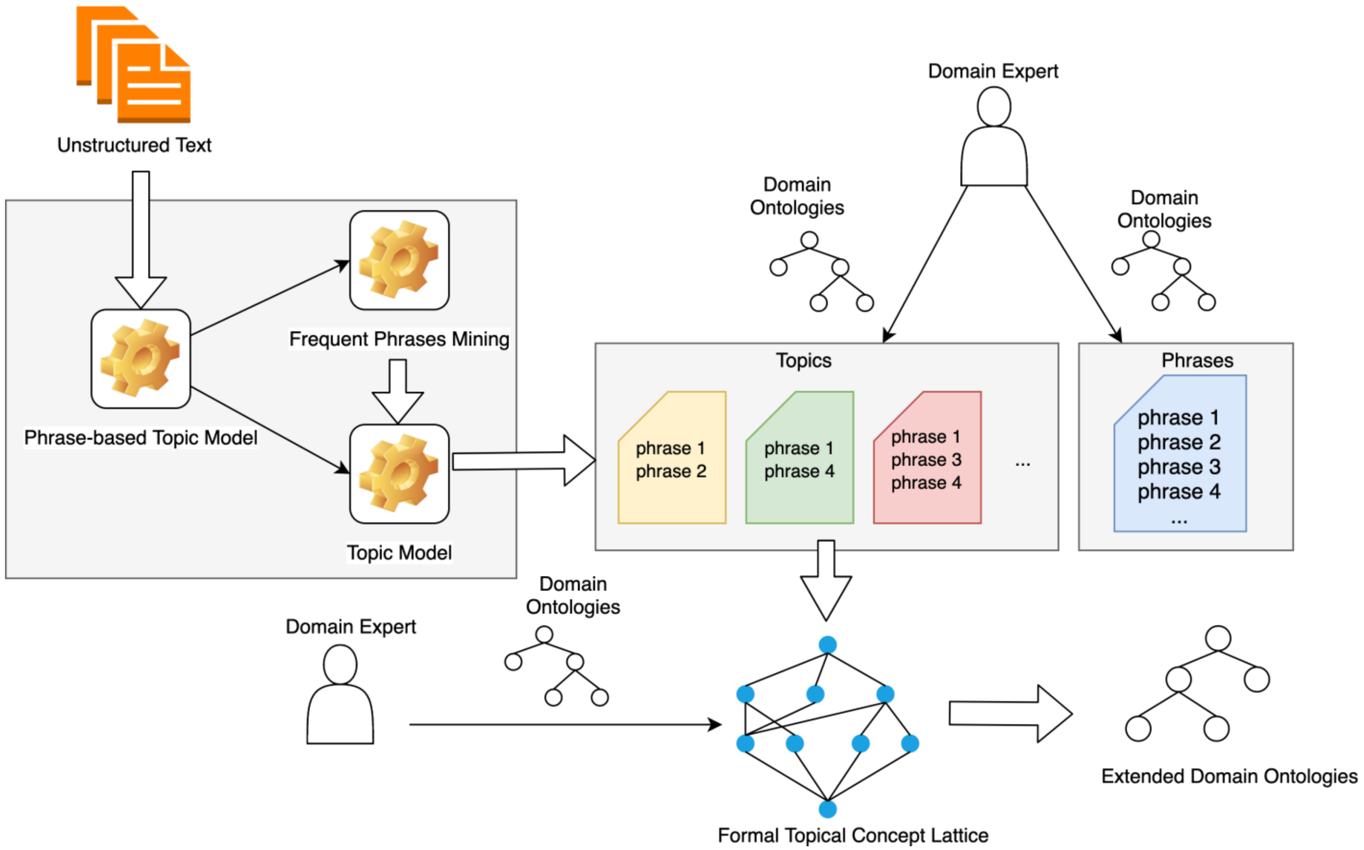
- ✓ MDO is capable to represent basic domain knowledge
- ✓ MDO can be used for mapping different materials databases' schemas
- ✓ MDO can be used for semantically enabling materials database search
- ✓ MDO is publicly accessible via Github, W3ID URL, LOV

Overview

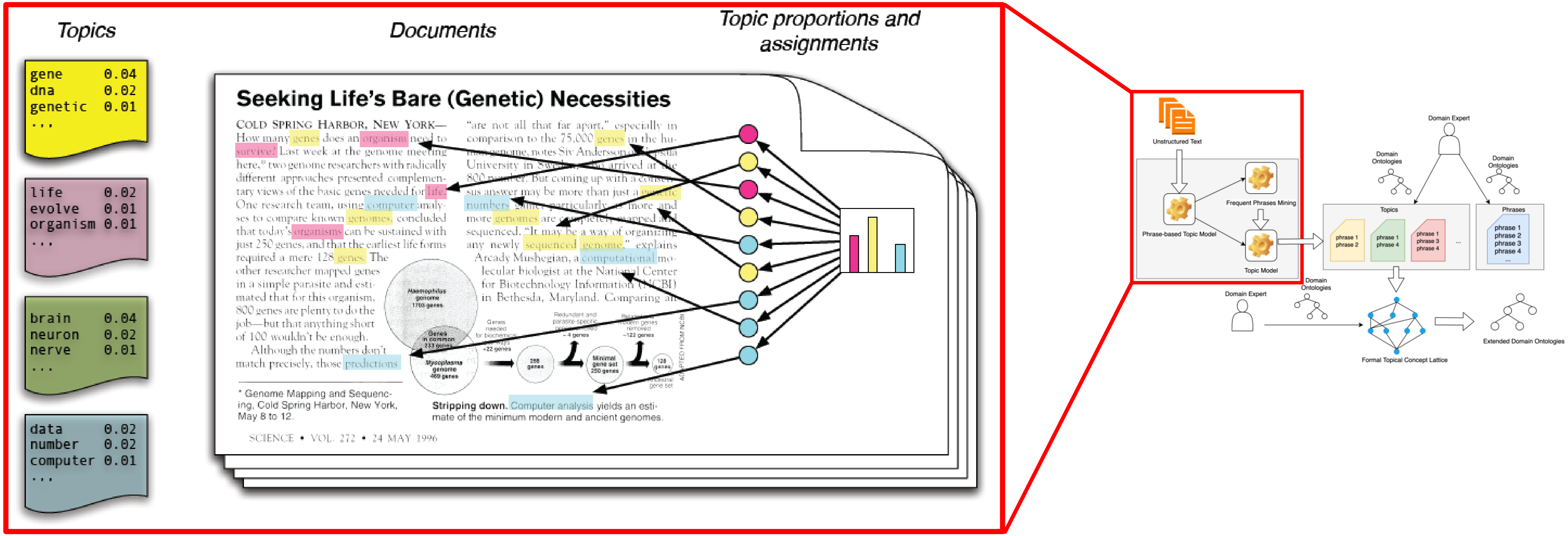
- A topic model and formal concept analysis-based method for extending ontologies
  - A contribution to address the RQ1
    - How can materials domain ontologies be extended by mining unstructured text?

Framework

- Stage 1: Phrase-based Topic Modelling
- Stage 2: Formal Concept Analysis
- Domain Expert Validation



Topic Model


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
SCIENCE • VOL. 272 • 24 MAY 1996
Unstructured Text
Phrase-based Topic Model
Frequent Phrases Mining
Topic Model
Domain Expert
Domain Ontologies
Topics
Phrases
phrase 1
phrase 2
phrase 3
phrase 4
Formal Topical Concept Lattice
Extended Domain Ontologies

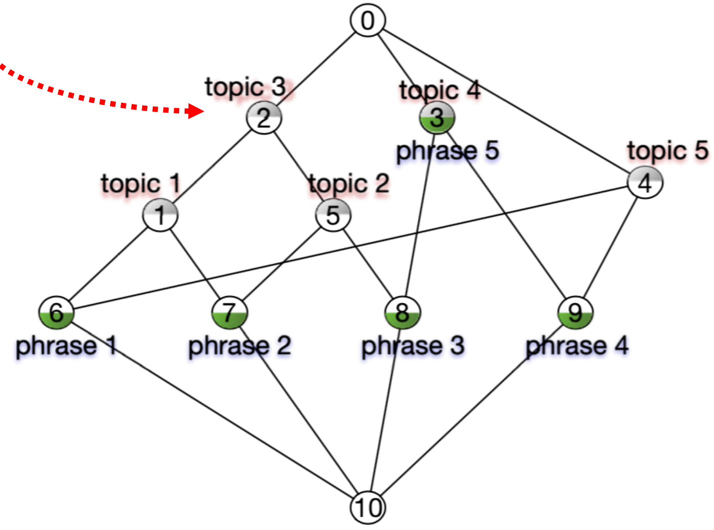
Formal Concept Analysis over the result of a Phrase-based Topic Model

- phrase-topic table
- A formal concept is (P, T) where every phrase in set P is a representative of all topics in set T
- Node 2: ((phrases 1, phrase 2, phrase 3), topic 3)

	topic 1	topic 2	topic 3	topic 4	topic 5
phrase 1	✓		✓		✓
phrase 2	✓	✓	✓		
phrase 3		✓	✓	✓	
phrase 4				✓	✓
phrase 5				✓	

a: Example of phrases occurrence in topics

b: Example of Formal Topical Concept Lattice

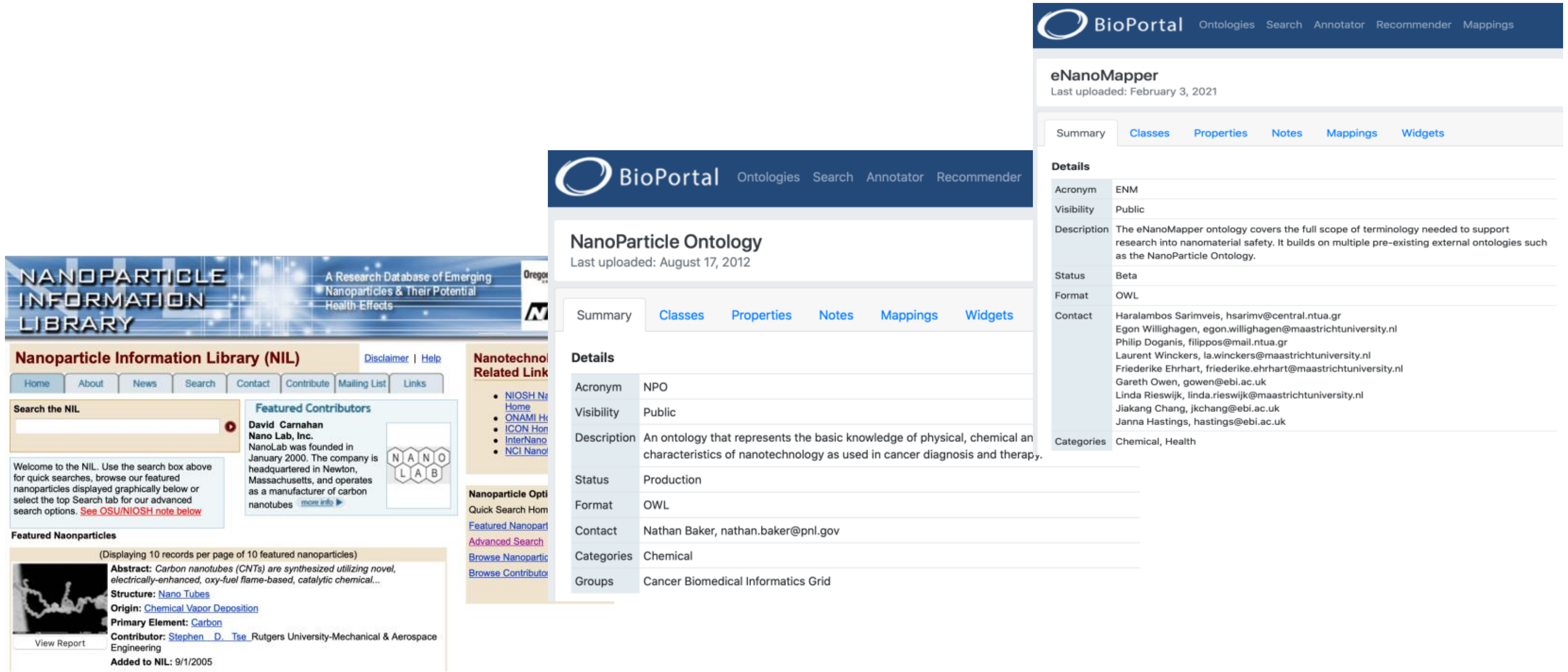


Domain Expert Validation

- Interpret all phrases appearing in topics
  - *Is a phrase meaningful in the domain? (No(-g))*
  - *Does a phrase exist in the ontology already? (EXIST(-m))*
  - *Should a phrase be added into the ontology? (ADD(-m))*
- Interpret topics
  - Choose representative phrases for a topic or give a label
    - If the label of the topic is too specific *(Q)*
  - *Does a topic exist in the ontology already? (EXIST(-m))*
  - *Should a topic be added into the ontology? (ADD(-m))*

Experiments

- Dataset
  - 627 abstracts of publications from Nanoparticle Information Library (http://nanoparticlelibrary.net)
- Existing Ontologies
  - Nanoparticle ontology
  - eNanoMapper ontology



Result-1

- To interpret phrases
- Examples:
  - ‘amorphous silicon’ (ADD)
  - ‘gold nanoparticle’ (EXIST)

ADD(-m): New knowledge for the ontologies
EXIST(-m): EXISTING knowledge for the ontologies

	ADD	ADD-m	EXIST	EXIST-m	No-g	No
20, low, Nanoparticle	32	4	26	19	16	9
20, low, eNanoMapper	29	3	24	25	14	12
30, low, Nanoparticle	30	4	26	18	16	9
30, low, eNanoMapper	28	3	24	26	12	11
40, low, Nanoparticle	32	4	26	15	16	10
40, low, eNanoMapper	29	3	24	22	14	12
20, high, Nanoparticle	9	1	14	7	4	0
20, high, eNanoMapper	8	2	12	10	3	0
30, high, Nanoparticle	8	2	14	8	0	1
30, high, eNanoMapper	7	1	12	10	0	1
40, high, Nanoparticle	9	2	14	12	4	4
40, high, eNanoMapper	9	2	12	14	2	4

Result-2

- To interpret topics
- Examples:
  - 'mesoporous silica nanosphere' (ADD)
  - 'chemical vapor deposition' (EXIST)

ADD(-m): New knowledge for the ontologies
EXIST(-m): EXISTING knowledge for the ontologies

	ADD	ADD-m	EXIST	EXIST-m	No-g	No	Q
20, low, both	3(1)	0	2	0	1	0	13
30, low, both	8(2)	0	4	0	1	0	13
40, low, both	16(1)	0	11	1	2	5	10
20, high, both	8(1)	0	3	2	0	0	7
30, high, both	3(2)	0	10	2	0	0	7
40, high, Nanoparticle	10(2)	0	10	3	2	2	3
40, high, eNanoMapper	10(2)	0	9	4	2	2	3

Comparison to Text2Onto

- Combine text mining algorithms
- Experiments with different number of candidates

# of elements	Algorithm	ADD	ADD-m	EXIST	EXIST-m	No-g	No	precision
100	Entropy	5	0	39	19	4	33	0.67
	C-value/NC-value	5	0	39	19	4	33	0.67
	Relative term frequency	5	0	39	20	4	32	0.68
	TF-IDF	17	0	22	12	6	43	0.57
200	Entropy	7	1	63	43	8	79	0.60
	C-value/NC-value	7	1	63	43	7	79	0.60
	Relative term frequency	7	1	63	42	8	79	0.60
	TF-IDF	24	1	38	19	19	99	0.50
300	Entropy	12	1	80	52	16	139	0.53
	C-value/NC-value	12	1	80	52	16	139	0.53
	Relative term frequency	13	1	78	52	16	140	0.53
	TF-IDF	28	1	58	36	29	148	0.50
400	Entropy	18	1	98	62	20	199	0.50
	C-value/NC-value	18	1	98	62	20	199	0.50
	Relative term frequency	19	1	100	61	20	199	0.50
	TF-IDF	36	1	70	44	38	211	0.47

	ADD	ADD-m	EXIST	EXIST-m	No-g	No	precision
Text2Onto-100	20	0	51	27	11	71	0.60
Text2Onto-200	29	1	84	55	26	164	0.54
Text2Onto-300	39	1	118	78	44	266	0.51
Text2Onto-400	41	1	120	73	47	313	0.47
Our Method	32	3	25	18	14	22	0.80

Comparison to Text2Onto

- New Concepts
  - Our Method vs Text2Onto

Concepts	Our method	Text2Onto
growth mechanism	✓	✓
intensity		✓
resolution	✓	
layer by layer growth	✓	
liquid solid	✓	
pressure		✓
MCM 41	✓	
mechanical property	✓	
melting		✓
melt spin	✓	
mesoporous silica nanoparticle	✓	
mesoporous silica nanosphere	✓	
microcrystalline silicon	✓	✓
nano colloid		✓
nano composite		✓
nanocrystal	✓	✓
nano crystalline silicon particle		✓
nanogrid		✓
nano ribbon		✓
nanotube array	✓	✓
nanowire array		✓
oxidation		✓

Summary

- Phrase-based topic model + Formal concept analysis
- Domain expert interprets in terms of phrases, topics
- Confirmation of ontological concepts
- Addition of new ontological concepts

Overview

- A GraphQL-based framework for data access and integration where an ontology drives the generation of the GraphQL server
- The lack of methods accessing and integrating data over multiple heterogeneous data sources where data is shared via Web APIs (e.g., REST APIs)
- One of the contributions to address the RQ2
  - How can ontology-based techniques be used to access and integrate data from heterogeneous sources?
    - RQ2.1: How can ontologies be leveraged to generate GraphQL APIs for data access and data integration?

GraphQL

- What is GraphQL?
  - GraphQL is a conceptual framework for building Web APIs
  - GraphQL can work with existing APIs of a system
  - A GraphQL server contains GraphQL schema and GraphQL resolver
  - Clients use the GraphQL query language to make requests to a GraphQL server

```
type Book {
  title: String
}

type Author {
  first_name: String
  books: [Book]
}
type Query {
    List_Books: [Book]
    List_Authors: [Author]
}
```

GraphQL Schema

```
List_Books {
    title
}

List_Authors {
    first_name
    books {
        title
    }
}
```

GraphQL Query

```
const resolvers = {
  Query: {
    List_Books() {
      # query underlying data source to get books with titles
      # return books with titles in json object
    },
    List_Authors() {
      # query underlying data source to get Authors with first names
      # and referencing books
      # return authors with first names and referencing books in json object
    }
  }
};
```

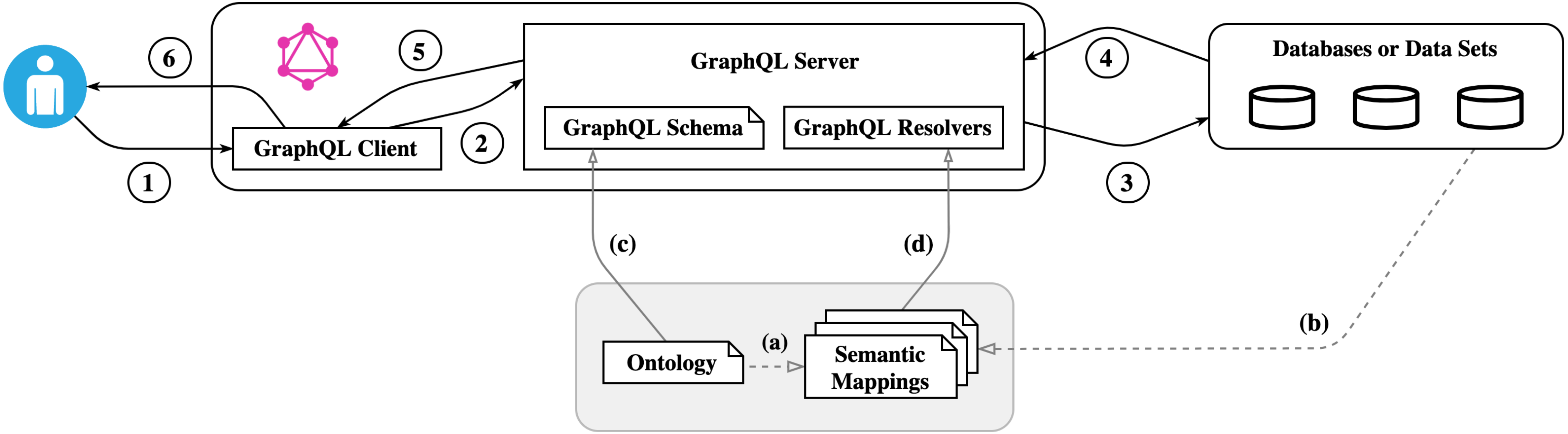
GraphQL Resolver

GraphQL

- Why GraphQL? (In theory / In practice)
  - GraphQL schema can help to provide a unified view of data from multiple sources
  - GraphQL resolver can mediate the accesses to different data sources via APIs (e.g., web data source, SQL database data source)
  - GraphQL has self-documenting specification
  - Users can discover the semantics reflected in API specification when writing GraphQL queries
- Why ontology-driven GraphQL server generation?
  - Ontology can provide the conceptual model over multiple data sources
  - To avoid manually create GraphQL schema and GraphQL resolver from scratch

Framework

- Stage 1: Ontology and semantic mappings constructions, arrows (a) and (b)
- Stage 2: GraphQL server generation, arrows (c) and (d)
- Stage 3: Query answering process, arrows (1) to (6)



Stage 2: Ontology-based GraphQL Schema Generation

- Input: General concept inclusions from an ontology
- Output: a GraphQL schema

$CalculatedProperty \sqsubseteq Property$
$PhysicalProperty \sqsubseteq Property$
$Property \sqsubseteq =1\ PropertyName.xsd{:}string$
$Property \sqsubseteq \forall\ relatesToStructure.Structure$
$Calculation \sqsubseteq \forall\ hasInputStructure.Structure \sqcap \forall\ hasOutputStructure.Structure$
$Calculation \sqsubseteq \forall\ hasInputProperty.Property \sqcap \forall\ hasOutputCalculatedProperty.CalculatedProperty$
$Calculation \sqsubseteq =1\ ID.xsd{:}string$
$Property \sqsubseteq \forall\ quantityValue.QuantityValue$
$QuantityValue \sqsubseteq \forall\ numericValue.xsd{:}float$

MDO logic axioms

```
interface Structure{
  iri: String!
}
type Structure_obj implements Structure{
  iri: String!
}
interface Property{
  iri: String!
  PropertyName: String!
  relatesToStructure: [Structure!]
}
type Property_obj implements Property{
  iri: String!
  PropertyName: String!
  relatesToStructure: [Structure!]
}
interface Calculation{
  iri: String!
  ID: String!
  hasInputStructure: [Structure!]
  hasOutputStructure: [Structure!]
  hasInputProperty: [Property!]
  hasOutputProperty: [CalculatedProperty!]
}
type Calculation_obj implements Calculation{
  iri: String!
  ID: String!
  hasInputStructure: [Structure!]
  hasOutputStructure: [Structure!]
  hasInputProperty: [Property!]
  hasOutputProperty: [CalculatedProperty!]
}
```

```
interface CalculatedProperty implements Property{
  iri: String!
  PropertyName: String!
  relatesToStructure: [Structure!]
}
type CalculatedProperty_obj implements CalculatedProperty & Property{
  iri: String!
  PropertyName: String!
  relatesToStructure: [Structure!]
}
interface PhysicalProperty implements Property{
  iri: String!
  PropertyName: String!
  relatesToStructure: [Structure!]
}
type PhysicalProperty_obj implements PhysicalProperty & Property{
  iri: String!
  PropertyName: String!
  relatesToStructure: [Structure!]
}
type Query{
  CalculationList(IDs: [String!]): [Calculation!]!
  StructureList(): [Structure!]!
  PropertyList(PropertyNames: [String!]): [Property!]!
  PhysicalPropertyList(PropertyNames: [String!]):
    [PhysicalProperty!]!
  CalculatedPropertyList(PropertyNames: [String!]):
    [CalculatedProperty!]!
}
```

GraphQL schema

Stage 2: GraphQL Resolver Generation

- Semantic mappings Processing
  - Data from databases can be mapped as instances of ontologies
- Duplicates Detection and Data Fusion
  - Detect records from data which represent the same real-world object
  - Handle data conflicts in terms of contradiction and uncertainty

Conclusion and Future Work

LINKÖPING UNIVERSITY

Conclusion

- We conducted an overview study of state of the art in terms of the usage of materials data and current efforts related to semantic technologies
- We proposed a method for extending ontologies based on topic modelling, formal concept analysis and domain expert validation
- We proposed a domain ontology (MDO) aiming to alleviate problems arisen when accessing and integrating data from heterogeneous materials databases
- We currently focus on introducing GraphQL for data access and integration for materials design domain;
  - We focus on generating GraphQL server based on ontologies, and merging data in the context of GraphQL

Future Work

- [1] Lambrix P, Armiento R, Delin A, **Li H**, ***Big Semantic Data Processing in the Materials Design Domain***, Encyclopedia of Big Data Technologies, Springer, 2018.
  - [*] Update version: Lambrix P, Armiento R, Delin A, **Li H**, ***FAIR Big Data in the Materials Design Domain***
- [2] **Li H**, Armiento R, Lambrix P, ***An Ontology for the Materials Design Domain***, The 19th International Semantic Web Conference, Athens Greece (Virtual Conference), 2020.
- [3] **Li H**, Armiento R, Lambrix P, ***A Method for Extending Ontologies with Application to the Materials Science Domain***, Data Science Journal, 2019.
- [4] **Li H**, Hartig O, Armiento R, Lambrix P, ***Ontology-Driven GraphQL Server Generation for Data Access and Integration*** (Ongoing).

➢ Extending the Materials Design Ontology [2] with the method presented in [3]

➢ To develop a system with the implementation of the framework in [4] for the materials design domain

Conclusion

- We conducted an overview study of state of the art in terms of the usage of materials data and current efforts related to semantic technologies
- We proposed a method for extending ontologies based on topic modelling, formal concept analysis and domain expert validation
- We proposed a domain ontology (MDO) aiming to alleviate problems arisen when accessing and integrating data from heterogeneous materials databases
- We currently focus on introducing GraphQL for data access and integration for materials design domain;
  - We focus on generating GraphQL server based on ontologies, and merging data in the context of GraphQL