Taylor & Francis
Taylor & Francis Group

Science and Technology of Advanced Materials: Methods

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstm20

Structuring superconductor data with ontology: reproducing historical datasets as knowledge bases

Masashi Ishii & Koichi Sakamoto

To cite this article: Masashi Ishii & Koichi Sakamoto (2023) Structuring superconductor data with ontology: reproducing historical datasets as knowledge bases, Science and Technology of Advanced Materials: Methods, 3:1, 2223051, DOI: 10.1080/27660400.2023.2223051

To link to this article: https://doi.org/10.1080/27660400.2023.2223051

SCIENCE AND TECHNOLOGY OF ADVANCED MATERIALS: METHODS
2023, VOL. 3, NO. 1, 2223051
https://doi.org/10.1080/27660400.2023.2223051

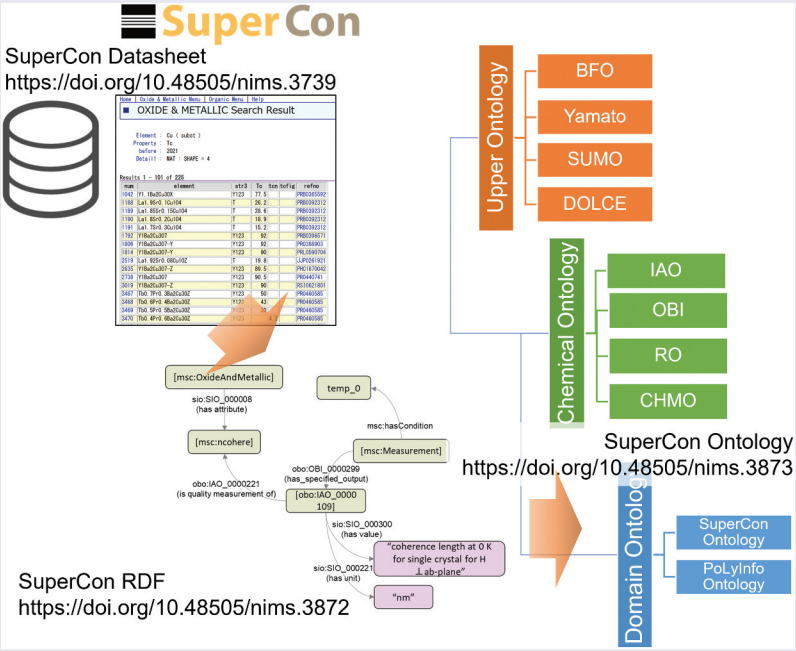
Structuring superconductor data with ontology: reproducing historical datasets as knowledge bases

Masashi Ishii and Koichi Sakamoto

Center for Basic Research on Materials, National Institute for Materials Science (NIMS), Tsukuba, Japan

ABSTRACT

Applying historical human-readable databases to recent data-driven science is a natural concept. However, this cannot be realized by simply converting a database into a tabular format because the meaning of each table column and the relationships between columns need to be rewritten in a machine-readable format. In particular, eliminating implicit notations that can only be understood by experts in the fields covered by each database and making them machine-readable under a unified academic system is necessary when integrating data across fields with a view to solving specific social issues and social implementation. In this study, we constructed a superconducting materials ontology for SuperCon, a legacy superconductor database that was recently republished as a datasheet, based on the well-known Basic Formal Ontology (BFO) top ontology, and designed a schema for material composition, structure, properties, and processes, among others. Using this schema, we constructed and published the Resource Description Framework (RDF) for each instance in the SuperCon datasheet. We also discuss the machine-readable format of data common to materials science discovered in this process.



ARTICLE HISTORY
Received 12 April 2023
Revised 15 May 2023
Accepted 1 June 2023

KEYWORDS
Ontology; RDF; SuperCon; superconducting material; knowledge base

1. Introduction

Systematic data curation is a common challenge in materials science for data-driven material research and development. In addition to high-throughput experiments [1] and calculations [2], public information facilitates the collection of physical-property data under a variety of experimental conditions [3], and examining vast amounts of data is expected to provide general guidelines for developing materials [4].

The release of open data, a major source of public information, has evolved from paper-based sharing and electronic-media distribution to web-based search

CONTACT Masashi Ishii ISHII.Masashi@nims.go.jp Center for Basic Research on Materials, National Institute for Materials Science (NIMS), Tsukuba 305-0044, Japan

Supplemental data for this article can be accessed online at https://doi.org/10.1080/27660400.2023.2223051.

systems. This evolution has resulted in large amounts of human-only readable information that is difficult to transform into high-quality data. In addition, data were manually curated in the past, while recent technological advances in artificial intelligence have automated these processes [5]. While such data are compatible with machine-readable formats, complex links that are comparable to human knowledge are not easily achieved. However, recent significant advances in the automatic linking of other entities and related data, such as processes and measurement conditions, may lead to a paradigm shift [6]. With this background in mind, data curation not only aims to collect data but also to structure and link them machine-readably.

To reconsider this problem as a bottom-up one from a data-driven research site, the dataset needed here is organized into a single table, with individual data uniquely identified by triads; that is, table column information, row information, and values at row/column intersections. For example, rows are organized by sample name or ID, and columns are organized by the names and conditions of various properties [7]. However, the columns, physicochemical meanings of the columns themselves, and the taxonomy behind the rows in tables need to be bridged to ensure tool versatility or to rationalize the model created by each tool [8].

With the aim of linking data and constructing a knowledge base that enables machine reasoning, in this study, we discuss our attempt to create an ontology-based data structure for the SuperCon superconductivity database, which was recently released as a datasheet at the Materials Data Repository (MDR) [9,10] in the National Institute for Materials Science (NIMS), after its web-based graphical user interface (GUI) service was decommissioned in 2021. We first briefly summarize the history of SuperCon, the subject of the discussion, and summarize its database policy because data structuring is not uniquely determined and depends greatly on the purpose of the database and the handled contents.

2. SuperCon background

The 'Multi-Core' project was launched in 1987 at the then Science and Technology Agency (current Ministry of Education, Culture, Sports, Science and Technology, MEXT) following the discovery of high-superconducting critical-temperature (Tc) oxide superconductors in 1986 [11]. A 'Database Core' was established as part of this project, and the construction of databases of standard data and numerical data from academic papers began. Three fields were established: high-Tc oxide superconductors, metal-alloy superconductors, and organic superconductors. The initial metal and alloy data were collected from the original papers listed in the so-called 'Roberts' Table' [12] and from various data books, while data from 1990 were collected from selected papers. The SuperCon database has been republished on the internet as two databases: 'OXIDE & METALLIC' for inorganic superconducting materials and 'ORGANIC' for organic superconducting materials.

Typical properties in the database include: Tc, extrapolation of the pressure P dependence of Tc at $P = 0$ (dTc/dP), lower critical field (Hc1), upper critical field (Hc2), temperature dependence of Hc2 at the critical temperature (dHc2/dT), coherence length (ξ), penetration depth (λ), isotope effects, energy gaps, specific heat (C), and related properties (thermal conductivity (k), thermoelectric capacity (S), Hall coefficient (R), etc.). The temperature dependences of these properties were determined from graphs using digitizers, edited, and redisplayed graphically. The graphical data include the Tc(x), Tc(P), C(T), Hc2(T), S(T), k(T), and R(T) values, where the variables in parentheses (x, P, and T) correspond to composition, pressure, and temperature, respectively.

Interestingly, the construction policy at the start of the project (almost 40 years ago) included creating a data-oriented database rather than a literature-oriented type, which had been the dominant policy to that point in time. Data were restructured irrespective of the literature framework and envisaged for use in material development research and research theme exploration by making the sample the smallest unit of record. Compositionally different samples from the same original literature were treated as independent data and managed using serial numbers as sample IDs. Furthermore, when physical properties were unable to be expressed by a single value (e.g. temperature dependence, composition dependence, etc.), the ID was associated with representative values and their graphical data. SuperCon adopted a data collection policy aimed at creating data structures suitable for machine readability and data-driven research as early as the 1980s. Therefore, SuperCon is an appropriate target for reproducing historical data as a knowledge base in this study.

3. Ontologies for materials science

Ontology research has been developed under several top ontologies, including the Basic Formal Ontology (BFO) [13], Cyc [14], and the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [15]. Attempts to integrate general concepts and improve the knowability of local topics in the field of materials science include the Materials Ontology by Ashino [16], European Materials and Modelling Ontology (EMMO) [17], and MatOnto [18] based on DOLCE. While providing a global view of materials science is undoubtedly important, local ontologies based on specialized knowledge is important

when considering solutions to specific issues close to practical use. However, it is essential to ensure logical consistency of the local ontology with the global one. In fact, there is no clear boundary between global and local, and a number of intermediates overlap to form concepts, such as ChEBI [19] (which deals with biologically interesting molecules in accordance with BFO), CHEMINF [20] (which provides an overview of chemical information), CHMO [21] (which formalizes experimental chemistry methods), and RXNO [22] (which deals broadly with chemistry). These intermediate ontologies deal with more general concepts than the ontologies addressed in this study. The ceramics ontology [23] and the crystal defect ontology [24] are examples of local ontologies that are as specialized as that in this study.

NIMS has been promoting the systematization of chemical-data-centered databases, such as PoLyInfo [25], into a knowledge base. Therefore, we consider that other material databases should be systematized using the same policy as much as possible. We decided to use BFO, which is highly compatible with large-scale chemical databases, such as PubChem [26], to systematize the NIMS databases and datasheets.

4. SuperCon ontology (MSC ontology)

4.1 Ontology creation policy

The data items in the SuperCon datasheet discussed in this paper are too voluminous to include; please refer to the list published in the MDR (https://doi.org/10.48505/nims.3740). The following policy was created for the SuperCon ontology developed in this study (hereafter referred to as the 'MSC ontology', where 'MSC' is the abbreviation for 'MDR SuperCon'). The created ontology defines the column names in the MDR SuperCon datasheet and the object properties (predicates) that relate the columns to each other; these relationships are depicted in the schemata presented in the next section. The Resource Description Framework (RDF) [27] of schema-based SuperCon data (instances) ultimately forms a superconductor knowledge base.

MSC ontology creation policy:

(1) The OXIDE & METALLIC datasheet, which accounts for most of the SuperCon data, is the MSC-ontology-creation target.
(2) SuperCon datasheets should be completely structured so that the column names are URIs (Uniform Resource Identifiers) and the contents listed on each line are instances.
(3) Separate concepts are required if physical conditions are incorporated into the column name (i.e. class name), with the aim of working toward machine-readable definitions of classes.
(4) Upper classes and predicates are defined as prominent external ontologies as much as possible.

Background and purpose are provided for each of the policies listed here:

(1) The OXIDE & METALLIC datasheet, which account for most of the SuperCon data, is the ontology target. As mentioned in Section 2, SuperCon consists of two datasheets: 'OXIDE & METALLIC' for inorganic superconducting materials, and 'ORGANIC' for organic superconducting materials. These datasheets contained 33,407, and 568 samples, respectively (as of April 2023), with the former constituting more than 98% of the total. While creating an integrated ontology that covers both datasheets is possible by establishing a universal electronic model of superconductivity, the two concepts are hard to merge since inorganic and organic chemistry have electronically different research paradigms. Therefore, in this paper, we first create an MSC ontology specifically for OXIDE & METALLIC (hereafter referred to as 'SuperCon O&M'). SuperCon ontology is denoted by adding 'msc': as a URI prefix in the following discussion. Therefore, the msc: entries in this study are represented as URIs if 'msc': is replaced by http://dice.nims.go.jp/ontology/SuperCon-ont/Schema#, where the definition of the msc: entry is provided.
(2) With the exception of literal comment columns, those with duplicate content, and some that are not suitable for structuring due to notation fluctuations, there are approximately 200 target columns in total number that require structuring. These columns need to be hierarchized by defining upper classes for coupling with the external upper ontology and for linking with fields other than superconductivity. In fact, each column in the datasheet is classified into one of the following eight upper classes, with their URIs shown in parentheses:

- Magnetic property (msc:MagneticProperty)
- Material property (msc:MaterialProperty)
- Mechanical property (msc:MechanicalProperty)
- Normal state property (msc:NormalState Property)
- Relational property (msc:RelationalProperty)
- Structural property (msc:StructuralProperty)
- Superconducting property (msc:Super ConductingProperty)
- Thermal property (msc:ThermalProperty)

It should be noted that while these upper classes have commonalities with other fields, they also contain 'relational properties' (msc:RelationalProperty) that

are not physically intuitive due to ontological requirements; 'msc:RelationalProperty' refers to physical properties that assume the presence of other materials, such as isotopes in SuperCon O&M. The isotope effect is a phenomenon in which Tc changes when some of the metal ions in the superconducting material are replaced with isotopes [28]. Physical properties that require the presence of other substances, such as isotopes in this case, belong to this upper class. Although details of the isotope effect are discussed later, as in this example, the structuring in this work is not only associated with building classes and their hierarchical structures, but also about structuring data to include the physical meanings of the physical properties, which is also related to (3) below.

(3) Column names with implied physical conditions are split and defined in machine-readable format. In a similar manner to other domains, the superconductivity domain discusses important physical concepts among experts without defining them. For example, only researchers in this field are implicitly aware of the differences and physical importance associated with whether the magnetic field is applied parallel or perpendicular to the ab plane when studying the superconducting properties of single crystals. Unfortunately, many SuperCon O&M properties incorporate these basic conditions into column names (class names in the MSC ontology), and their resolution is essential for data sharing. For example, the column name 'abreshe' (Column number #165 in the datasheet) is composed of ab+res+he, which describes the resistivity of a single crystal at liquid helium temperature when a current is applied parallel to the ab-plane. Clearly, in order to share data widely with SuperCon Organic and other researchers outside of the domain in the future, these 'implicit' understandings of a particular domain need to be defined in the forms of general concepts. The creation of the MSC ontology not only benefits the superconductivity domain, but also the greater purpose of contributing to knowledge sharing with exterior domains.

(4) The upper concepts of each class and predicate are defined, as far as possible, in a well-known external ontology. Many general phenomena, not just in physics, are defined within well-established external ontologies, and the upper-level concepts of SuperCon O&M must ultimately connect to them in a consistent manner both in and out of the ontology. The following external ontologies are used in the MSC ontology created in this study:

- Basic Formal Ontology, BFO (http://basic-formal-ontology.org/). BFO is a top ontology designed to support information search, analysis, and integration in science and other domains, and is used in many initiatives comparable to this activity.
- Information Artifact Ontology, IAO (https://obofoundry.org/ontology/iao.html). The IAO is a collection of entities related to information developed by the OBO foundry [29] as well as the BFO.
- Ontology for Biomedical Investigations, OBI (https://obi-ontology.org/). OBI is an ontology for scientific investigations that involve a biomedical scope and includes technical terms such as 'allergen', which has nothing to do with superconductivity; here we only use the general super term 'Planned Process' (http://purl.obolibrary. org/obo/OBI_0000011). The logical definition here is consistent with other terms in the OBO community, and can be used in conjunction with the abovementioned ontologies.
- CHemical Methods Ontology, CHMO (https://obofoundry.org/ontology/chmo.html). CHMO is an ontology published by the OBO to complement the OBI. It defines the equipment used in physical chemistry, for data collection, and sample synthesis.

In addition, the Relation Ontology (RO) of OBO is marginally included in the constraints of the predicates defined in the MSC ontology; however, this is not discussed in this paper. Table 1 summarizes the prefixes used in the MSC ontology, including the external ontologies used in the predicates.

4.2 Overview of the created MSC ontology

The ontology designed according to the policies in Section 4.1 has been published at https://doi.org/10.48505/nims.3873 and can be reviewed in detail using an appropriate ontology editor. The number of object properties, classes, and named individuals defined in OWL (c.f., owl:ObjectProperty, owl:Class, and owl:NamedIndividual) used in the MSC ontology are listed below.

Table 1. External ontologies and prefixes used in the MSC ontology.

Prefix	Actual prefix in URI
dc	http://purl.org/dc/elements/1.1/
dcterms	http://purl.org/dc/terms/
msc	http://dice.nims.go.jp/ontology/SuperCon-ont/Schema#
obo	http://purl.obolibrary.org/obo/
owl	http://www.w3.org/2002/07/owl#
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
skos	http://www.w3.org/2004/02/skos/core#
xml	http://www.w3.org/XML/1998/namespace
xsd	http://www.w3.org/2001/XMLSchema#

Table 2. Summarizing external ontology classes and their labels, corresponding top-level classes in the MSC ontology, and their representative subclasses.

#	External ontology class	Label	msc ontology top-level class	Representative subclass
1	obo:BFO_0000040	material entity	msc:Specimen	msc:OxideAndMetallic
2	obo:IAO_0000030	information content entity	msc:Structure	msc:InternalStructure
3	obo:IAO_0000027	data item	msc:Descriptor	msc:MaterialStructureDescriptor
4	obo:BFO_0000020	specifically dependent continuant	msc:Property	msc:SuperConductingProperty
5	obo:IAO_0000003	measurement unit label	msc:Unit	msc:SuperConductingPropertyUnit
6	obo:OBI_0000011	planned process	msc:Measurement msc:SampleProcess	msc:CrystallographicMeasurement msc:SamplePreparation
7	obo:IAO_0000104	plan specification	msc:MeasurementMethod msc:PreparationMethod	msc:mcohere*msc:method* msc:method*
8	obo:BFO_0000023	role	msc:Role	msc:Component
9	obo:IAO_0000300	textual entity	msc:FigureNumber	msc:tcfig*

*These are the same as SuperCon column names, where msc:mcohere is the coherence length measurement method and msc:method is an instance of the fabrication method, such as CVD.

- Object property 13
- Class 248
- Named Individual 80

Most classes corresponded to approximately 200 SuperCon O&M columns. The remaining classes are newly defined upper classes as described in 4.1 (2) that are eventually connected to external ontology concepts. Table 2 lists the external ontology classes and their labels, the top MSC ontology classes, and their representative subclasses. This table shows that the MSC ontology consists of nine series: materials (#1), structures and their characteristics (#2 and #3), properties and their values (#4 and #5), measurement and fabrication processes and methods (#6 and #7), roles (#8), and textual information (#9), in alignment with the external ontology. All concepts in the MSC ontology are classified into one of these nine series. While these series are understood by definitions in the external ontology, #8 (role), which reflects the original MSC ontological idea, will be discussed in the next section where actual applications are described. Object properties and named individuals will be discussed in 5 using schema that describe actual datasheets.

5. Constructing an RDF for SuperCon O&M

5.1 Overview of the RDF

The data contained in SuperCon O&M can be structured in terms of the MSC ontology as followings:

- Composition of the superconducting sample (msc-cmp).
- Structure of the superconducting sample (msc-str).
- Physical property/property values of the superconducting sample (msc-prop).
- Process for preparing the superconducting sample (msc-prc).

The concept space (namespace) handled in SuperCon O&M can be comprehensively described by adding the ID and chemical name of the superconductor sample (msc-smp and msc-elm) and the RDF of the bibliographic information associated with the data source (msc-ref). The prefix for the namespace created by each piece of information is indicated in parentheses. The correspondence between the prefix and the actual URI can be found on the RDF public site of this project (https://doi.org/10.48505/nims.3872). For example, 'msc-cmp': which indicates the namespace of the sample composition in the RDF, can be replaced with http://dice.nims.go.jp/ontology/SuperCon-ont/supercon-rdf/cmp# to obtain the actual URI.

These four namespaces are discussed in the following sections using typical schema diagrams. In particular, the physical property schema for isotope effects is intensively discussed in Section 5.4 as a representative academic issue for data structuring in materials science. The following schema and the actual RDF are seen to share sample IDs (URIs), even though they are in different namespaces. In other words, queries for the graphical database can be generated by tracing triples around the sample URI.

5.2 Structuring the composition of a superconducting sample (msc-cmp)

As with most inorganic materials, SuperCon O&M focuses on the composition of the sample; therefore, each composition is managed using a column in the datasheet. Determining the optimal composition or exploring alternative materials with the same composition (in the case of a compound) is a natural developmental approach once a promising material is found. To illustrate the significance of composition, the MSC ontology defines the sample as a subclass of the material entity (obo:BFO_0000040), while component is defined as a 'bearer of the mixture role' (msc:MixtureRole) that forms a superconductive material. In other words, the ontology explicitly states that

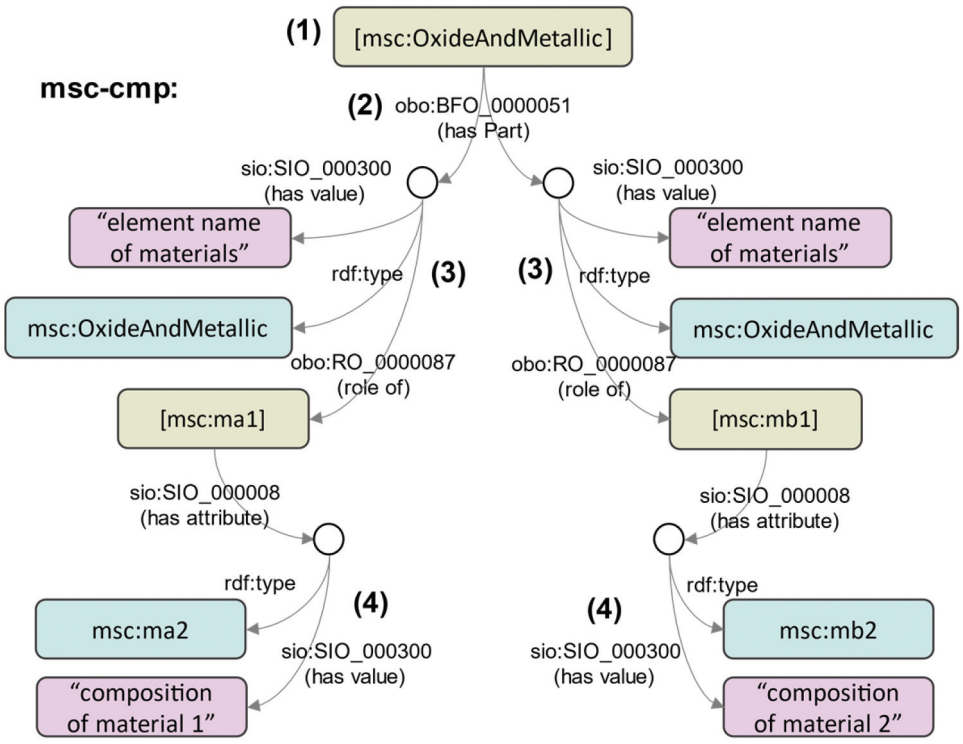


Figure 1. MSC-CMP schema for describing superconductor composition.

components are materials that do not exist on their own using the mixture role. The composition is a descriptor defined in msc:Descriptor, and is an attribute of msc:MixtureRole rather than the material entity. Figure 1 shows an msc-cmp schema based on this concept. In the following, '(1)' and '(2)' correspond to the numbers shown in the figure.

(1) A sample is an instance of the entire material (msc:OxideAndMetallic) considered in SuperCon O&M. In this study, instances of a class are denoted in brackets; thus, samples are denoted by [msc:OxideAndMetallic].
(2) The sample has 'parts' represented by blank nodes (open circles in the figure) that indicate parts whose boundaries, such as chemical bonds, are unclear; however, each element can be identified.
(3) The 'parts' are also instances of msc:OxideAndMetallic as well as the sample, and have component roles defined by msc:ma1 or msc:mb1, as shown in the figure as a typical example. The component-role URIs correspond to column names in SuperCon O&M: first component (msc:ma1), second component (msc:mb1), and third component (msc:mc1). The tenth constituent (msc:mj1) and the specially conceptualized role of oxygen (more precisely, oxygen in oxides), msc:mo1, define 11 species, which indicates that up to 11 constituents, including oxygen, are described in the datasheet. In fact, several 10-element superconductors, such as MgCNaSnCaAlSiNiInPb and $[(Py\text{-}CnH_{2n+1})\text{-}_2HgI_4]Bi_2(Sr,Ca)_2CaCu_2O_z$, are included in SuperCon O&M, which suggests that the existence of superconductors with more than 11 elements cannot be ruled out; however, the flexibility to accommodate superconducting materials with more than 11 elements is not provided by this ontology that structures SuperCon O&M.
(4) The superconducting material composition concept is given by the SuperCon O&M column names, as in (3), and the MSC ontology defines the first composition (msc:ma2), the second composition (msc:mb2), and the tenth composition (msc:mj2); furthermore, adding the oxygen role composition (msc:mo2) leads to 11 defined compositions. Composition is considered to be an attribute of the component role, not the component element itself, and is defined as a descriptor in the MSC ontology; the values that specifically appear in the molecular formula are provided in the literature.

As discussed in part (4) of 4.1, ontology creation policy, components, and composition classes lead to an external ontology by following the upper classes, as illustrated in Figure 2. The lower part of the figure represents the msc-cmp schema, where msc:ma1 and msc:ma2 (shown in Figure 1) are used as examples, while the upper part is the concept hierarchy defined in the MSC ontology. As indicated by the boundary between the upper and lower parts, the MSC ontology

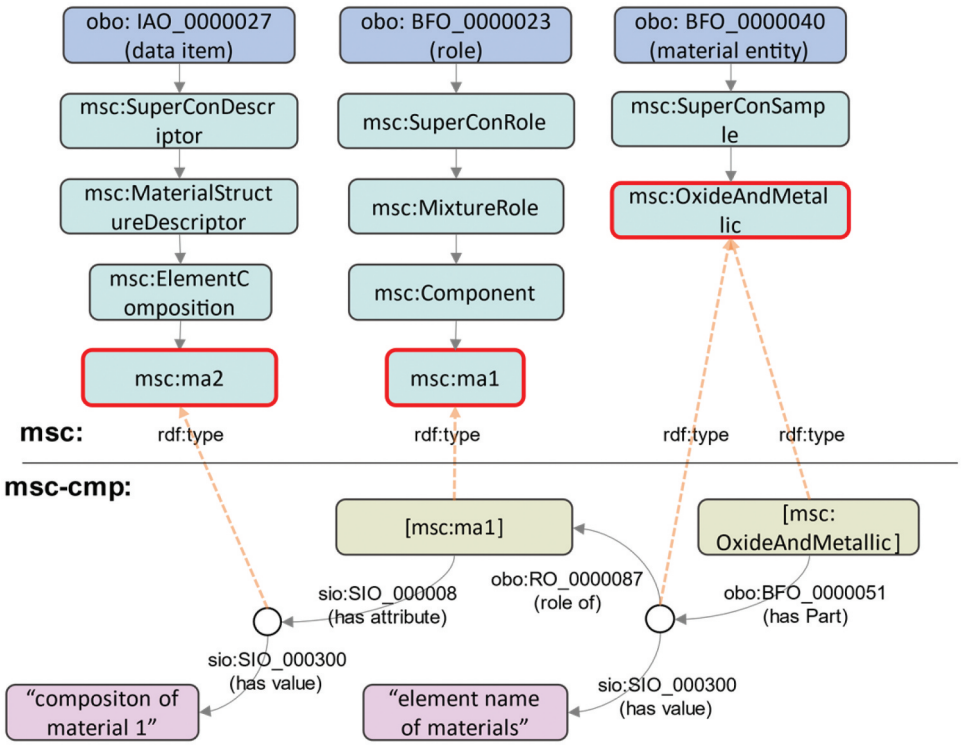


Figure 2. Relationship between schema, superclasses, and external ontology, with msc-cmp as an example.

and msc-cmp are connected by rdf:type. In addition, the IAO and BFO classes of external ontologies are shown at the top of the MSC ontology to ensure harmonization with external concepts.

5.3 Structuring crystal information for superconducting samples (msc-str)

The 'msc-str' namespace provides information on the internal structure (msc:InternalStruture) of the sample and physically corresponds to crystal-structure information, specifically the crystal system (msc:str1), crystal space group (msc:spaceg), common structure name (msc:str3), and lattice constant (msc:latc). With the exception of lattice constant, which is a physical property of the sample, all attributes of msc:InternalStruture are defined as descriptors, that is, subclasses of msc:CrystalState in msc:MaterialStructureDescriptor. The remaining lattice constants are structured under msc:StructuralProperty, which indicates that measured values are treated as properties rather than descriptors.

Figure 3 shows an msc-str schema, the details of which are discussed below, with '(1)' and '(2)' corresponding to the numbers in Figure 3.

(1) As in Section 5.2, all schemas created in this study start from the instance of the sample [msc:OxideAndMetallic]. This common starting point for sample IDs has the practical advantage of simplifying the description in SPARQL (SPARQL Protocol and RDF Query Language), which is an RDF query language.

(2) The sample has [msc:InternalStruture] as the entirety of the information treated in msc-str.

(3) This information is consists of descriptors related to the crystal structure, as shown in the obo:BFO_0000051 (has Part) predicate defined in OBO. Space groups (msc:spaceg) and general structure names (msc:str3) are defined as classes in the MSC ontology. In the schema, blank nodes, which are instances of these classes, have actual data in the literals that are connected by rdfs:comments.

(4) A possible seven crystal systems (msc:str1), such as the 'cubic' structure, also use the obo:BFO_0000051 (has Part) predicate to specify the internal structure. The following URIs are used for each crystal system:

msc:str1_01	cubic
msc:str1_02	tetragonal
msc:str1_03	orthorhombic
msc:str1_04	monoclinic
msc:str1_05	triclinic
msc:str1_06	trigonal (rhombohedral)
msc:str1_07	hexagonal

Essentially, such crystal information should be linked to external knowledge [30].

In this schema, the 'msc:str1' and 'msc:spaceg' parallel notations clearly duplicate information because msc:spaceg is a subclass of (rdfs:subClassOf) msc:str1 in the MSC ontology. This peculiar notation is the

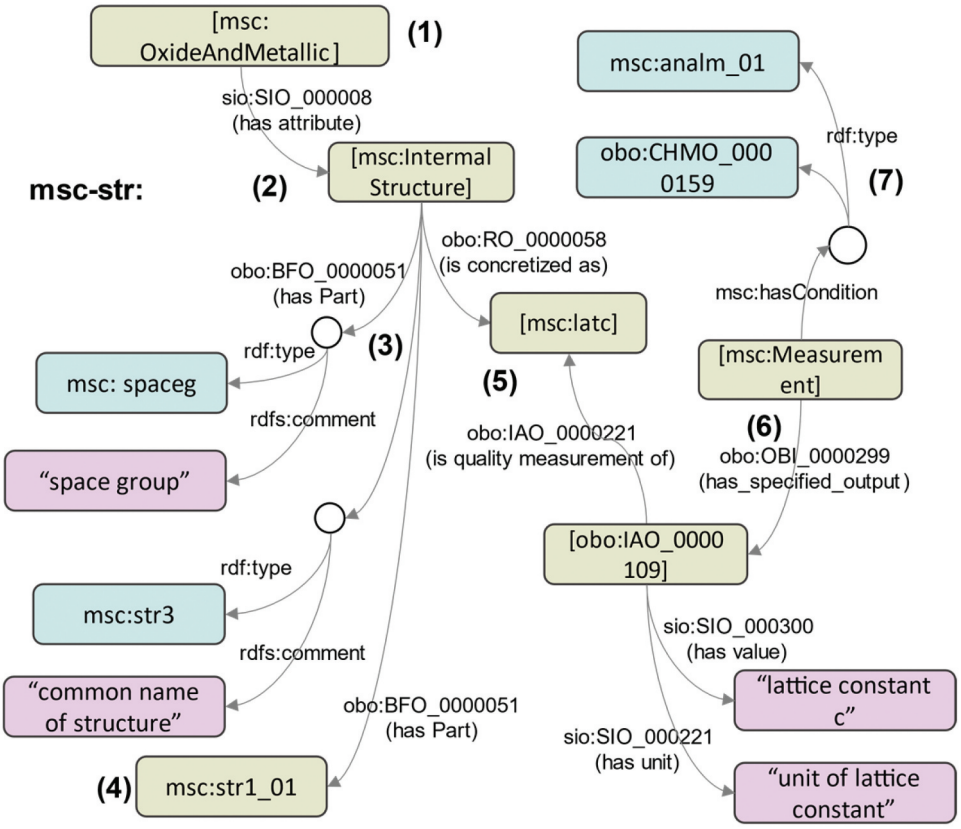


Figure 3. An msc-str schema for describing superconductor structure.

result of actual experimental limitations; only the crystal system of the upper concept provides useful information for samples with unidentifiable space groups. Consequently, both column names are treated equally in this schema.

(5) SuperCon records the measured values of experimentally estimated lattice constants, as shown in this schema, in which the c-axis lattice constant (msc:latc) is used as an example, although the a-axis and b-axis lattice constants (msc:lata and msc:latb, respectively) can be expressed in the same manner. The measured values and units were formulated using the measurement datum (obo:IAO_0000109); hence, this notation follows the IAO restrictions.

(6) The datum is the output of a measurement process (msc:Measurement).

(7) The actual measurement method (msc:analm_01 in this example) is given as a condition for msc: Measurement, and typical crystallographic measurement methods are defined as classes in the MSC ontology. Well-known methods can be linked to classes in the CHMO external ontology, as shown in Table 3. This fact reveals that the URIs of both the MSC ontology and CHMO can be queried through SPARQL, as illustrated below (section 6.2). However, only msc:analm_02 (neutron crystallography) is not defined in CHMO; therefore, we refer to the 'neutron diffraction' upper class.

These crystallographic methods are nominalized and classified according to SuperCon policy. Simple sequential numbers were used as IDs for nominalization and are displayed on the web database GUI (as they were) as well as in the current MDR SuperCon datasheet, which indicates that the importance of vocabulary management through ID was potentially recognized as the database began to be constructed. The MSC ontology also retains this ID by adding the prefix 'analm_', thereby ensuring compatibility

Table 3. Crystallographic method URIs and labels used in the MSC ontology and corresponding CHMO URIs and labels.

URI	Lable of msc ontology URIs	Corresponding URI in CHMO	Label of CHMO URIs
msc:analm_01	X-ray crystallography	obo:CHMO_0000159	single crystal X-ray diffraction
msc:analm_02	Neutron crystallography	obo:CHMO_0000698	neutron diffraction
msc:analm_03	Powder x-ray diffraction	obo:CHMO_0000158	powder X-ray diffraction
msc:analm_04	Powder neutron diffraction	obo:CHMO_0000699	neutron powder diffraction
msc:analm_05	Electron diffraction	obo:CHMO_0000142	electron diffraction

between the MSC ontology and MDR SuperCon datasheets. The advantage of the MSC ontology is that a human-readable representation of the measurement method can also be obtained with SPARQL using 'rdfs: label'.

5.4 Physical property/property value of the superconducting sample (msc-prop)

Figure 3 includes a schema for the 'measurement of lattice constant' property; the msc-prop that structures the measurement of the various properties discussed here is essentially the same as this schema. This compatibility suggests that the measurement schema is independent of the namespace and is a universal representation of the measurements themselves. We will discuss two probable cases where additions to this universal schema are necessary:

(a) The structure of the physical property column names that incorporate the measurement conditions.
(b) The structure of the physical properties presupposing the presence of other substances.

We first consider case (a) using coherence length ξ as an example, where ξ corresponds to the spatial size of the Cooper pair and is known to be infinite at Tc [31]. In particular, ξ at 0 K is an important superconducting property according to BCS theory [32]; its value depends on the applied magnetic field and is considered to be key to understanding the superconductivity mechanism. Meanwhile, the high-Tc cuprate superconductor crystal has a planar CuO_2 structure in the ab plane and is further arranged in a layered structure in the c-axis direction. Hence, the direction of the magnetic field applied to the crystal structure is an important experimental condition for ξ. All of the information detailed here is condensed under the following two column names in the SuperCon dataset:

- ncohere: coherence length at 0 K for a single-crystal sample in the H⊥ab plane
- pcohere: coherence length at 0 K for a single-crystal sample in the H//ab plane

The combined 'ncohere' abbreviation for 'n+cohere' is found in a cryptogram that is only understandable to an expert, and does not machine-readably describe the physical meaning of the coherence length at 0 K when a magnetic field is applied normal to the ab crystal plane. By analogy, the machine cannot understand 'pcohere' to mean 'coherence length when a magnetic field is applied parallel to the ab crystal plane'. For example, a schema for ncohere created using only msc:ncohere without the aforementioned physical background is shown in Figure 4(a). Figures 3 (5–7) and 4(a) (1–3) are identical, despite the ξ-measurement method being expressed in literals; consequently, experimental conditions, such as magnetic field direction and temperature, are not machine-readably expressed. Therefore, we redesigned Figure 4(a) and constructed Figure 4(b). The experimental conditions not described in Figure 4(a) are expressed as shown in Figure 4(b) using one class (owl:Class), one object property (owl:ObjectProperty), and two named individuals (owl:NamedIndividual) (Table 4). Figure 4(b) shows that ncohere expresses under conditions in which (1) the magnetic field is applied perpendicular to the ab plane of the sample crystal, and (2) the temperature is absolute zero. Similarly, msc:nchohere and msc:pchohere are machine-readably described by introducing the 'msc:ParallelTo' predicate, which means 'applied in parallel'.

There are many crystal-orientation-dependent superconducting properties. For example, the critical

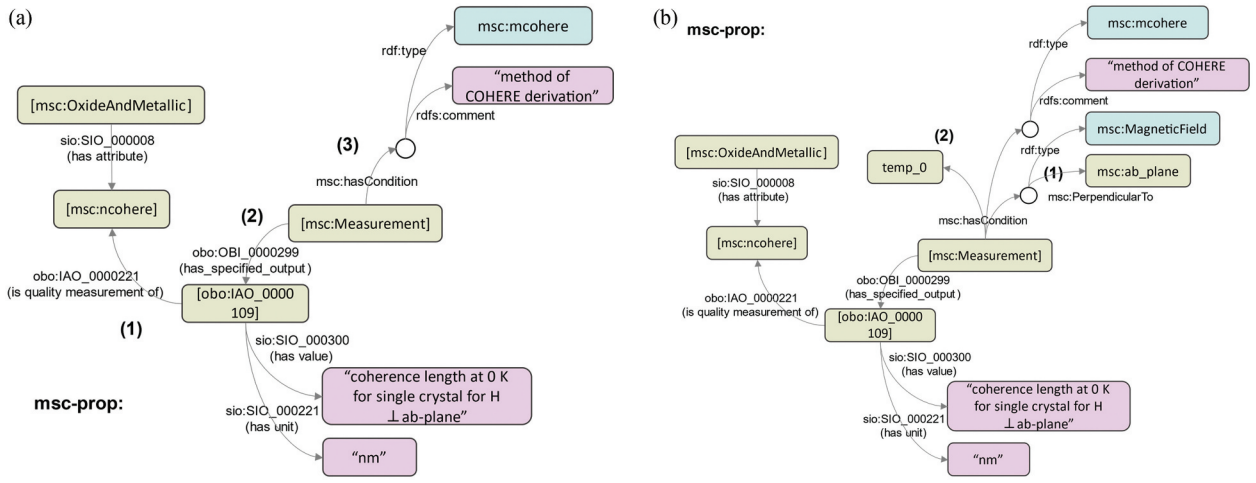


Figure 4. (a) An MSCC-prop schema for 'ncohere' (coherence length at 0 K for a single crystal sample under the H⊥ab plane) created using only msc:ncohere in the absence of physical background. (b) 'Ncohere' expressed with conditions: (1) magnetic field is applied perpendicular to the ab-plane of the sample crystal, and (2) the temperature is absolute zero.

Table 4. Items introduced to make 'ncohere' machine-readable.

URI	Description (rdfs:comment)	Upper class	Type
msc:MagneticField	Magnetic field, H	msc:PhysicalDescriptor	owl:Class
msc:PerpendicularTo	Perpendicular To	msc:hasDescriptor	owl:ObjectProperty
msc:ab_plane	ab-plane of a crystal	-	msc:CrystalOrientation, owl:NamedIndividual
msc:temp_0	a virtual temperature of 0 K, known as absolute zero.	-	msc:Temperature, owl:NamedIndividual

magnetic field (msc:nhc1zero and msc:phc1zero), penetration depth (msc:npenet and msc:ppenet), and Hall coefficient (msc:rh300n and msc:rh300p) are sensitive to the direction of the applied magnetic field, while resistivity (msc:abreshe and msc:creshe) must consider the direction of an applied electric field. For the latter, we defined the electric field concept (e.g. msc:ElectricField), which is not defined in the database, in the MSC ontology so that the schema in Figure 4(b) can be applied universally in an external-field-type independent manner.

To extend the universal msc-prop schema, we next discuss (b) 'structure of the physical properties presupposing the presence of other substances' using isotope effect as an example. An outline of the isotope effect in superconductors is provided above. Because isotopes do not exhibit different chemical properties when forming superconducting materials, the isotopic component for quantifying phonon effects must be introduced separately from the chemical component of the sample. In Section 5.2, we assigned msc:MixtureRole to the stoichiometric component of the superconducting material, while here we assign the isotope role (msc:IsotopeRole) instead. Here, the isotope element (msc:isoel) must be the isotope of a component element, which is considered a physical constraint of the isotope effect rather than an ontological constraint of msc:isoel, which is defined as an isotope with the msc:hasIsotope predicate in MSC ontology. Therefore, this constraint will be introduced in reasoning.

The isotope effect is represented by the schema shown in Figure 5, which extends the abovementioned ontological design. This figure reveals that: (1) the sample has an attribute of change in Tc (msc:dtc) due to the isotope effect, and (2) its actual value of msc:dtc is given via the measurement datum (obo:IAO_0000109), i.e. the same protocol as that used for the lattice-constant measurement (Section 5.2) is used. Here, we formulated (3) such that msc:dtc presupposes the existence of msc:isoel by the predicate msc:dependsOn. (4) This isotope plays the role of msc:IsotopeRole as discussed above, and (5) the isotope conversion rate (msc:isorat) is given as an attribute of that role, which is consistent with (4) in Figure 1. Thus, the physical property is expressed as a class that 'depends on' other substances, and attributes (in this case conversion rates) are formulated by assigning them to the role of other substances rather than the other substances themselves.

5.5 Superconducting sample preparation process (msc-prc)

The schema designed for the sample preparation process is described in Figure 6. Here, comments by the data curator are provided for the entire process (msc:SampleProcess), including pretreatment and post-treatment, and for sample preparation (msc:SamplePreparation), excluding pretreatment and post-treatment. Although these comments are free-text (human-readable), they provide important information that enables reproducibility and reliability to be better understood by datasheet users. The relationship between (1) process (msc:SampleProcess) and preparation (msc:SamplePreparation) is clearly stated to maintain this important information in the schema, with a literal comment provided for each. (2) In the case of sputtering, a typical sample preparation method, the sputter targets, substrates, and products are defined as having their own roles, and the process is considered to realize (obo: BFO_0000055) these roles; it should be emphasized that this follows the OBO process format. (3) These roles are assigned to each material using the 'has role' predicate (obo:RO_0000087). (4) The sample ID (URI) is assigned to the material with the product role (msc:ProductRole). (5) The specific process method is given as a condition of msc:SamplePreparation. Here, the process method msc:method_01 show in this figure means 'powder sintering method' and a total of 21 registered process methods can be identified in the MSC ontology published on the web. For example, https://dice.nims.go.jp/ontology/SuperCon-ont/Schema#method_20 shows that method_20 is a vapor growth method.

6. Establishing a graphical database

Using the schema discussed above, we converted all entries in the MDR SuperCon O&M datasheet into RDF format using an original script. The results are published at https://doi.org/10.48505/nims.3872. The data contained 2,850,362 triples. In addition, there were 1,412 triples in the MSC ontology (https://doi.org/10.48505/nims.3873). By deploying them in a suitable RDF store, SPARQL-based searches are easily performed in a local environment. The data are no longer represented in the SuperCon datasheet, but rather as a revived database and renewed

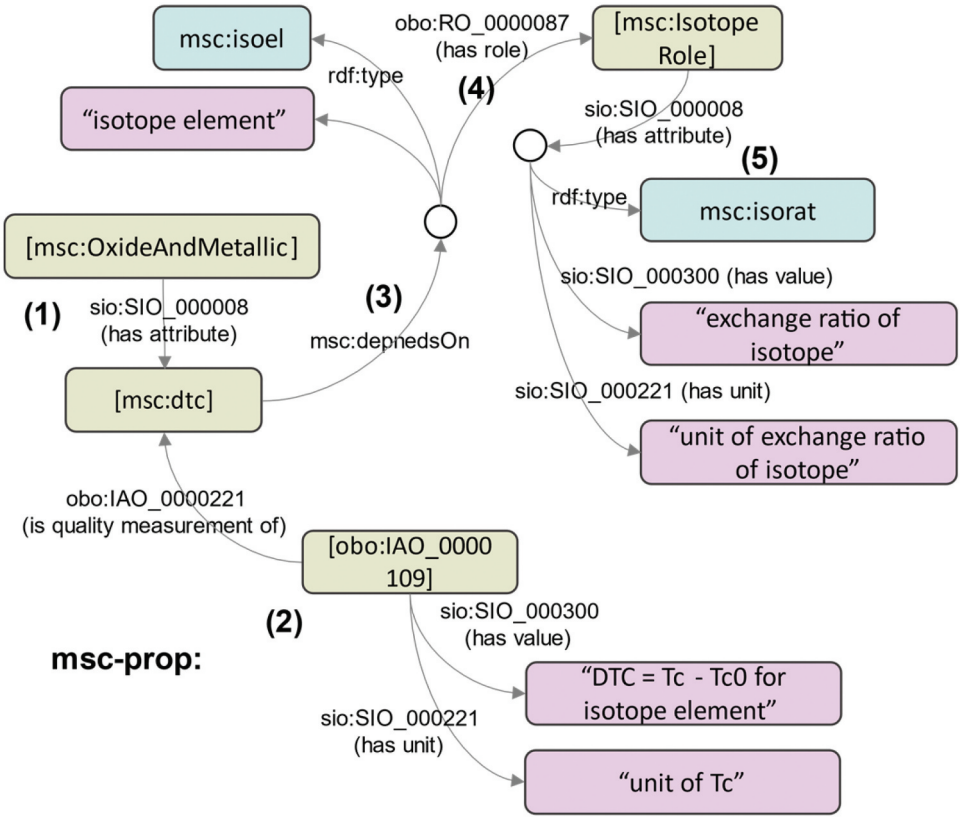


Figure 5. A schema for structuring physical property that presupposes the presence of other substances, with isotope effect as an example.

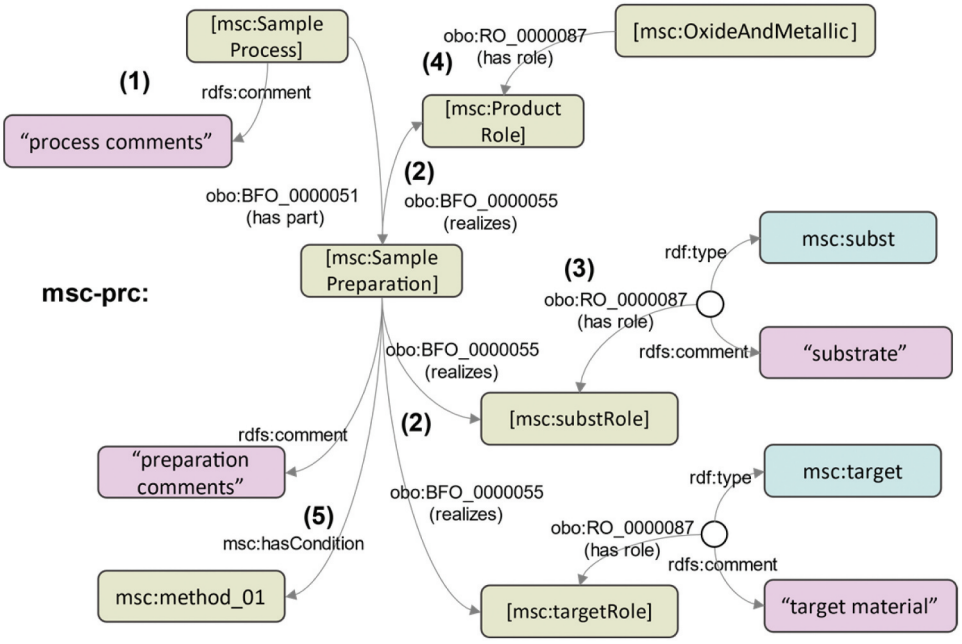


Figure 6. msc-prc schema for describing the superconductor fabrication process.

knowledge base referred to as 'SuperCon RDF' in the following discussion.

SuperCon RDF introduced semantic technology to the historical database, and as a result, the database became machine-readable knowledge. This enables general scientists to recognize databases that previously required expert background knowledge and use them for data sharing and application development. Examples of SPARQL queries that perform semantic SuperCon RDF searches can be obtained at the same public website. The following discussion can be confirmed by executing them.

6.1 RDF data statistics by physical property category

The MDR SuperCon datasheet does not include any physical property categories. The total sample numbers for each category introduced in the 4.1 Ontology creation policy are:

Magnetic property (msc:MagneticProperty):	4,047
Material Property (msc:MaterialProperty):	4,790
Mechanical property (msc:MechanicalProperty):	697
Normal state property (msc:NormalState Property):	5,160
Relational property (msc:RelationalProperty):	571
Structural property (msc:StructuralProperty):	81
Superconducting property (msc:Super ConductingProperty):	76176
Thermal property (msc:ThermalProperty):	5,700

There are a total of 97,222 physical property values in the SuperCon RDF.

SuperCon O&M has 33,407 samples; hence, there are approximately 2.91 properties per sample. However, this metric does not quantify the results associated with the introduction of ontologies and schemas or by transforming historical datasheet knowledge; 85.3 triples per sample (2,850,362 triples/33,407 samples) for knowledge-based metrics is considered to be reasonable.

6.2 External ontology cooperation

As discussed above, the MSC ontology ultimately leads to a higher-level external ontology, which is powerful when linking cross-disciplinary data. A CHMO example of practical data linkage is shown below. This external knowledge can be exploited if publicly available CHMO (https://github.com/rsc-ontologies/rsc-cmo) is introduced into the same triple store as the SuperCon RDF. Figure 3 shows that the definition of CHMO can be used to select the means of obtaining crystal information. In fact, SuperCon RDF shows that there are 91 and 957 of X-ray diffraction data for single crystals (obo: CHMO_0000159) and powders (obo:CHMO_0000158), respectively. Investigating the effects of crystal orientation from 1048 diffraction patterns obtained using the higher-level measurement concept of 'X-ray diffraction (obo: CHMO_0000156)' is also possible. More importantly, it facilitates data integration with external data that reference CHMO. Extending knowledge by sharing physical concepts common to many fields through ontologies is an important aspect of this approach.

6.3 Utilizing structured physical property column names with RDF

The physical meanings within the column names were made machine-readable, as discussed in Section 5.4(a). Herein, we demonstrate a semantic search for the direction of an applied external field, which is an important parameter for copper oxide superconductivity. The SuperCon RDF led to 2,392 samples with physical property data when an external field is applied parallel to the ab plane. In contrast, there were 417 for perpendicularly applied fields. Each target physical property and external field applied during its measurement can be summarized using a SPARQL query, as shown in Tables S1–S3 of Supplementary Material (The figures and tables discussed in this paper will be available on the MSC ontology and SuperCon RDF sites, including the tables in this Supplementary Material). Selecting the direction of the applied external field and providing an overview of physical properties is possible using SPARQL, which highlights the effectiveness of the semantic search.

In another demonstration, samples with measurements conducted both parallel to the ab plane and parallel to the c axis were selected after specifying the measured temperature, external field, and physical properties. Here, 29 samples with msc:abres77 and msc:cres77 measurements described in Table S3 were obtained. Thus, the SureCon RDF enabled semantic search by creating a dataset with a clear physical meaning.

The SuperCon RDF site also includes an isotope effect query example, as discussed in 5.4(b). In that query, the name of the superconductor, the isotope and its ratio, and the corresponding physical properties can be obtained by linking relationships among the many columns. In principle, even longer links are possible, and the query is flexible. This initiative enables the necessary information (including external resources) to be freely retrieved.

6.4 Sustainable development of ontologies and RDF

The constructed ontology is still being validated using the queries shown on the SuperCon RDF public site as well as queries that assume various integrated databases. Nonetheless, there may well be revisions to the concept itself in the future, in addition to inevitable errors. For example, the number of components currently limited to 11 could be expanded immediately, however, considering the constraints of reproducing SuperCon datasheets, it may be necessary to add a canonicalized ontology consistent with the external materials ontologies, in conjunction with the MSC ontology. The project has assigned DOIs to the SuperCon datasheet, MSC ontology, and RDF to provide version control and to encourage widespread use under the FAIR Principles [33]. We recommend that the latest version be used in the future, irrespective of the DOIs presented in this paper. The license for their

 By making the ontology and RDF reusable in this way, we hope that it will become one ontology that covers the entire materials.

In the recent prominence of materials informatics, the idea of integrating databases and using ontologies as a useful source of data has also been advocated [34,35]. In addition, extending the current attempt to reproduce the database accurately and joining knowledge by introducing certain assertions to develop new materials is probably one of the most promising aspects of ontology use. Ontology-based reasoning is widely known, and machine-based reasoning about materials with desired properties is expected to be realized in the near future. In such an era, the relevance and feasibility of the challenge will be determined by how many knowledge bases are integrated. Alignment with top ontologies and coupling between ontologies have been extended to the field of superconductivity in this study.

7. Conclusion

Relational information is inevitably missing when a database published as a web service is represented as a simple large table (datasheet). Given that the long-published SuperCon superconductor database has recently been converted into a datasheet, in this study we built an ontology and structured the properties and processes described in the datasheet using a schema. The ontology behind the schema was aligned with a well-known external upper-level concept that facilitates data linkage and reuse.

Furthermore, we proposed transforming column names, which could not previously be understood without expert prerequisite knowledge, into knowledge schema by decomposing them into a machine-readable format. By converting datasheet instances into an RDF based on this schema, we created a graph database and published it with ontology. Considering the success of bioinformatics, which increasingly uses semantic data, we created part of a knowledge base that is expected to advance material informatics.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the MEXT Program: Data Creation and Utilization-Type Material Research and Development Project (Digital Transformation Initiative Center for Magnetic Materials), Grant Number [JPMXP1122715503].

ORCID

Masashi Ishii http://orcid.org/0000-0003-0357-2832

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request. The data and queries necessary to reproduce the results are available on the MDR SuperCon public website (https://doi.org/10.48505/nims.3872 and https://doi.org/10.48505/nims.3873).

References

[1] Qin M, Lin Z, Wei Z, et al. High-throughput research on superconductivity. Chin Phys B. 2018;27(12):127402. doi: 10.1088/1674-1056/27/12/127402
[2] Shipley A, Hutcheon M, Needs R, et al. High-throughput discovery of high-temperature conventional superconductors. Phys Rev B. 2021;104(5):054501. doi: 10.1103/PhysRevB.104.054501
[3] Foppiano L, Dieb S, Suzuki A, et al. SuperMat: construction of a linked annotated dataset from superconductors-related publications. STAM. 2021;1:34–44. doi: 10.1080/27660400.2021.1918396
[4] Cheng B, Griffiths R, Wengert S, et al. Mapping materials and molecules. Acc Chem Res. 2020;53(9):1981–1991. doi: 10.1021/acs.accounts.0c00403
[5] Foppiano L, Castro PB, Suarez PO, et al. Automatic extraction of materials and properties from superconductors scientific literature. STAM. 2023;3:2153633. doi: 10.1080/27660400.2022.2153633
[6] Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. Proceedings of Advances in Neural Information Processing Systems 33 (NeurIPS 2020); 2020 Dec 6–12; Neural Information Processing Systems Foundation, Inc. (NeurIPS); 2020.
[7] Oka H, Yoshizawa A, Shindo H, et al. Machine extraction of polymer data from tables using XML versions of scientific articles. STAM. 2021;1:12–23.
[8] Feunang YD, Eisner R, Knox C, et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminform. 2016;8(1):61. doi: 10.1186/s13321-016-0174-y
[9] MDR SuperCon Datasheet Ver.220808 [Internet]. Tsukuba, Japan: National Institute for Materials Science; [cited 2023 Mar 31]. doi: 10.48505/nims.3837
[10] MDR SuperCon Datasheet Readme [Internet]. Tsukuba, Japan: National Institute for Materials Science; [cited 2023 Mar 31]. doi: 10.48505/nims.3740
[11] Bednorz JG, Müller KA. Possible High Tc Superconductivity in the Ba - La- Cu- 0 System. Z Phys B. 1986;64:189–193. doi: 10.1007/BF01303701
[12] Roberts BW. Survey of superconductive materials and critical evaluation of selected properties. J Phys Chem Ref Data. 1976 [cited 2023 Mar 31];5(3):581. doi: 10.1063/1.555540
[13] Temal L, Rosier A, Dameron O, et al., editors. Mapping BFO and DOLCE. Proceedings of the 13th World Congress on Medical Informatics (MEDINFO 2010); 2010 Sep 12–15; Cape Town, Republic of South Africa: IOS Press; 2010.

[14] Lenat DB, Guha RV, Pittman K, et al. Cyc: toward programs with common sense. Commun ACM. 1990;33(8):30–49. doi: 10.1145/79173.79176
[15] Masolo C, Borgo S, Gangemi A, et al. WonderWeb Deliverable D18 - Ontology Library (final). Italy: laboratory for Applied Ontology - ISTC-CNR; 2001 (IST Project 2001-33052 WonderWeb: Ontology Infrastructure for the Semantic Web).
[16] Ashino T. Materials ontology: an infrastructure for exchanging materials information and knowledge. Data Sci J. 2010;9:54–61. doi: 10.2481/dsj.008-041
[17] European Materials Modelling Ontology [Internet]. Belgium: European Materials Modelling Counsil (EMMC); 2020 [cited 2023 Mar 31]. Available from: https://emmo-repo.github.io/versions/1.0.0-beta/emmo.html
[18] Kwok C, Drennan J, Hunter J Towards an Ontology for Data-driven Discovery of New Materials. AAAI Spring Symposium: Semantic Scientific Knowledge Integration; 2008 Mar 26-28; Stanford (CA) USA.
[19] Degtyarenko K, Matos PD, Ennis M. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 2008;36:D344–D350. doi: 10.1093/nar/gkm791
[20] Ontology Search [Internet]. Cambridgeshire (UK): EMBL-EBI; [cited 2023 Mar 31]. Available from: https://www.ebi.ac.uk/ols/ontologies/cheminf
[21] The Chemical Methods Ontology [Internet]. San Francisco (CA) USA: GitHub, Inc.; [cited 2023 Mar 31]. Available from: https://github.com/rsc-ontologies/rsc-cmo
[22] Garay-Ruiz D, Bo C. Chemical reaction network knowledge graphs: the OntoRXN ontology. J Cheminformatics. 2022;14(1):29. doi: 10.1186/s13321-022-00610-x
[23] Vet PVE, Speel PH, Mars NJI The Plinius ontology of ceramic materials. Computer Science. 1994 [cited 2023 Mar 31]. https://www.semanticscholar.org/paper/The-Plinius-ontology-of-ceramic-materialsPaul-Vet-Speel/5f2806ddecc02786a84a051a2ed0e8f8f20af0c6
[24] Ihsan AZ, Dessi D, Alam M, et al., editors. Steps towards a dislocation ontology for crystalline materials. Proceedings of the Second International Workshop on Semantic Digital Twins co-located with the 18th Extended Semantic Web Conference (ESWC 2021); 2021 Jun 6; Hersonissos, Greece: CEUR-WS.org. ISSN 16130073. ISSN 16130073
[25] PoLyInfo [Internet]. Tsukuba, Japan: National Institute for Materials Science; [cited 2023 Mar 31]. Available from: https://polymer.nims.go.jp/
[26] Fu G, Batchelor C, Dumontier CM, et al. PubChemRDF: towards the semantic annotation of PubChem compound and substance databases. J Cheminform. 2015;7(34):1–15. doi: 10.1186/s13321-015-0084-4
[27] RDF [Internet]. Cambridge (MA): world wide web consortium; [cited 2023 Mar 31]. Available from: https://www.w3.org/RDF/
[28] Chen XJ, Struzhkin VV, Wu Z, et al. Unified picture of the oxygen isotope effect in cuprate superconductors. Proc Natl Acad Sci, USA. 2007;104:3732–3735.
[29] Information-artifact-ontology/IAO [Internet]. San Francisco (CA): GitHub, Inc. Available from: https://github.com/information-artifact-ontology/IAO/
[30] Hall SR, McMahon B. The Implementation and evolution of STAR/CIF ontologies: interoperability and preservation of structured data. Data Sci J. 2016;15:1–15.
[31] Cooper JR, Forro L, Keszeit B. Direct evidence for a very large penetration depth in superconducting $Bi_2Sr_2CaCu_2O_8$ single crystals. Nature. 1990;343(6257):444–446. doi: 10.1038/343444a0
[32] Bardeen J, Cooper L, Schrieffer JR. Microscopic theory of superconductivity. Phys Rev. 1957;108:162–164. doi: 10.1103/PhysRev.106.162
[33] FAIR Principles - GO FAIR. [cited 2022 Oct 14]. Available from: https://www.go-fair.org/fair-principles/
[34] Takahashi L, Takahashi K. Visualizing scientists' cognitive representation of materials data through the application of ontology. J Phys Chem Lett. 2019;10(23):7482–7491. doi: 10.1021/acs.jpclett.9b02976
[35] Zhao S, Quan Q. Ontology based heterogeneous materials database integration and semantic query. AIP Adv. 2017;7:105325. doi: 10.1063/1.4999209