A SEMANTIC-DRIVEN KNOWLEDGE REPRESENTATION MODEL FOR THE MATERIALS ENGINEERING APPLICATION

Xin Cheng[1,2], Changjun Hu[1,2,], and Yang Li[1,2]*

[1]*School of Computer and Communication Engineering, University of Science and Technology Beijing (USTB), Beijing, 100083, China*
[2]*Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing, 100083*
[*]*E-mail:* hu.cj.mail@gmail.com

ABSTRACT

A Materials Engineering Application (MEA) has been presented as a solution for the problems of materials design, solutions simulation, production and processing, and service evaluation. Large amounts of data are generated in the MEA distributed and heterogeneous environment. As the demand for intelligent engineering information applications increases, the challenge is to effectively organize these complex data and provide timely and accurate on-demand services. In this paper, based on the supporting environment of Open Cloud Services Architecture (OCSA) and Virtual DataSpace (VDS), a new semantic-driven knowledge representation model for MEA information is proposed. Faced with the MEA constantly changing user requirements, this model elaborates the semantic representation of data, services and their relationships to support the construction of domain knowledge ontology. Then, based on the ontology modeling in VDS, the semantic representations of association mapping, rule-based reasoning, and evolution tracking are analyzed to support MEA knowledge acquisition. Finally, an application example of knowledge representation in the field of materials engineering is given to illustrate the proposed model, and some experimental comparisons are discussed for evaluating and verifying the effectiveness of this method.

Keywords: Knowledge representation model, Domain ontology, Semantic mapping, Reasoning rule, Evolution tracking, Materials engineering application

1   INTRODUCTION

Materials Engineering Application (MEA) refers to the integration of materials design, solutions simulation, production-manufacturing and processing, and service evaluation. The MEA supports the life expectancy prediction for materials, the service safety of engineering, the design of new materials, etc. In order to achieve optimal application services, the MEA's process in its entirety needs to be studied more deeply. Some of the main challenges currently faced by MEA include capturing needed data accurately and providing timely effective services. To do this analysis, It is necessary to develop a representation of the entire MEA process, which includes relevant data, relationships, requirements, services, and so on. Therefore, the semantic representation of MEA's data services based on domain knowledge has become increasingly important.

In recent years, along with the continuous accumulation of scientific data and constantly changing practical requirements within the scientific data domain, "big data" management issues must now be addressed (Lynch, 2008; Howe, Costanzo, Fey, Gojobori, Hannick, Hide, et al., 2008). The processing of large-scale data sets cannot keep up with the amount of data generated by scientific research and production. Scientists find that today it is difficult to manage, analyze, and share their scientific data accurately and in a timely manner. The effective representation of data services is the foundation and key point for solving these problems. Although much research about the capture and representation of data services has been carried out for many scientific domains, few of the discoveries have been practically implemented in the field of materials engineering. Because different fields have different application features, knowledge representation methods in different areas are not the same. Considering the special characteristics of the materials domain, the issues needing to be resolved about the MEA's knowledge representation are as follows:

- MEA representation lacks causal integration and feedback validation among requirements and services. Almost no patterns of the service application have semantic relationships with user demands patterns. Even the user requirements themselves lack semantic expression.
- MEA heterogeneous information lacks effective abstraction and organization, and its representation lacks semantic mapping and reasoning of the complex associations among the heterogeneous types of information. In particular, material structure data are closely related to material property data, but the semantic representation between them is lacking. As a result, it is difficult for   implied relationships

information to be found and understood in MEA.

- MEA knowledge acquisition lacks semantic representation of evolution tracking. MEA needs to obtain related knowledge in a timely manner, but material data are updated frequently, and the dynamic changes in the MEA information are difficult to track and capture.
- MEA knowledge representation needs an open service supporting environment with which to construct a highly effective data management mode. It should support the treatment of the information resources distributed in different regions and satisfy the MEA's knowledge representation. Both of these issues are complexly associated, dynamically changing, and demand-oriented.

How MEA information is represented is very important for engineering data reuse, service reasoning, and application evaluation. The traditional data representation mode has been unable to deal with the above issues; thus we need to build a new representation model to meet these challenges.
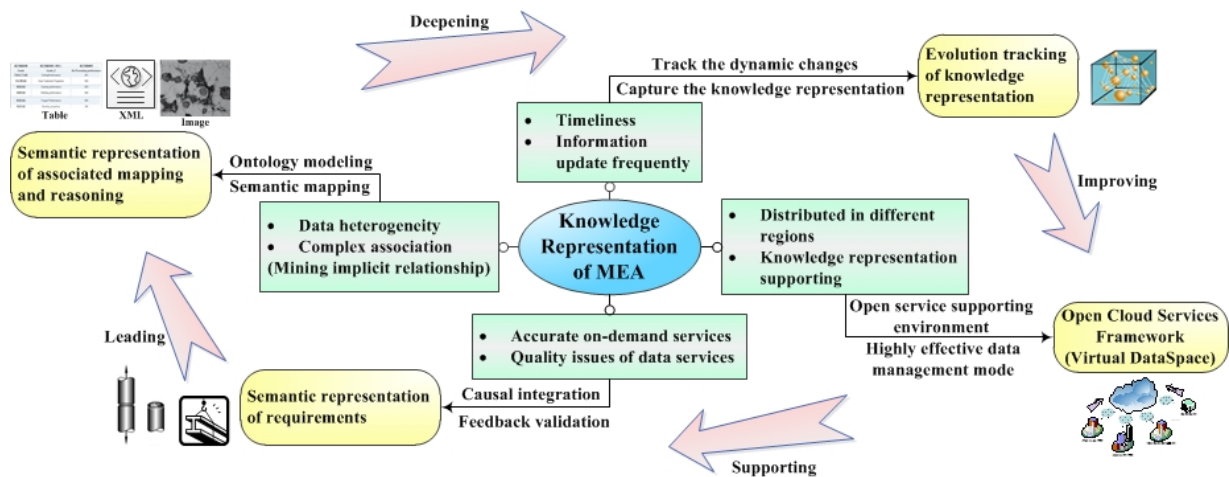

Figure 1. Overview of the issues and contributions

This paper presents a semantic-driven knowledge representation model for the Materials Engineering Application. The main contributions of this paper are shown in Figure 1. First, we propose an Open Cloud Services Architecture (OCSA) with Virtual DataSpace (VDS) to provide an open service supporting environment for MEA knowledge representation. Second, we construct the semantic representation of the requirement mode to support accurate on-demand services. Third, based on the requirement representation, we describe and represent ontology-driven data modeling, semantic mapping, and reasoning. Then, based on more in-depth research, we analyze the dynamic evolution of the association data and represent in a timely fashion the data changes, efficiently capturing all the useful knowledge. Finally, we improve the entire process of MEA knowledge representation based on reasoning and evolution. Doing all this will realize the automatic optimization of knowledge representation in the field of materials engineering.

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 presents the supporting environment of knowledge representation. Section 4 describes, in detail, the MEA semantic-driven knowledge representation model, which includes semantic representation of user requirements, data ontology, association mapping, rule-based reasoning, and knowledge acquisition based on evolution tracking. Section 5 introduces an application example of a knowledge representation model in materials engineering and describes its experimental evaluation. Section 6 gives the conclusions and future prospects for our work.

2 RELATED WORK

2.1 Materials informatics

Different fields have different application features. In particular, materials informatics faces many complex issues (Rajan, 2005; Hunt, 2006). The engineering applications of materials manufacturing systems are very diverse, and they come in many forms and life expectancies. The many different possible designs and materials available also necessitate changes in corresponding manufacturing processes while the changing

materials data need automatic capture, analysis, deployment, and maintenance. The traceability of information needs to be guaranteed (Cebon & Ashby, 2006), and the fast-tracking of new materials needs to be supported (Ferris, Peurrung, & Marder, 2007).

Much research has been done on materials informatics engineering applications (Ashino, 2010). Shen et al. (2006) introduced the applications of agent-based systems in intelligent manufacturing. Zhao et al. (2005) presented a novel approach for the design of hard coatings using the elastic properties of transition-metal nitrides calculated from the first-principles density functional theory. Ullah et al. (2008) presented an intelligence-based method to deal with materials selection problems where the design configurations and working conditions as well as the design-relevant information are not precisely known. Zhao et al. (2013) proposed a manufacturing informatics framework for the assessment of manufacturing sustainability. Al Khazraji et al. (2013) presented a Material Information Model (MIM) across the whole product lifecycle for sustainability assessment and developed conceptual ideas with recommendations in distributed cloud-based architecture. Although these proposals have some relevance for our work, almost all of them are at the conceptual level and lack an in-depth analysis of the knowledge representation of the materials engineering application.

2.2   Supporting environment and framework

For engineering informatics applications, researchers have proposed a variety of supporting environments and frameworks. Grossman et al. (2010) proposed an Open Science Data Cloud (OSDC) to support the analysis, processing, and management of large-scale scientific data sets, but this remains only in a support level for high-performance computing. Foster et al. (2005) proposed an Open Grid Services Architecture (OGSA) based on the traditional grid "five layers hourglass structure" and Web Service technology. This supports distributed data processing but lacks a detailed description of data management.

In search of new technology to deal with the new challenges of data management, Franklin et al. (2005) proposed the concept of DataSpace (DS), which is still in its initial stages of development. Only a few prototype systems have been built, for example, personal dataspaces, such as iMeMex (Blunschi, Dittrich, Girard, Karakashian, & Salles, 2007) and Semex (Dong & Halevy, 2005), but these two systems mainly investigated dataspace models, data storage, and query processing. Further, the personal dataspace prototype system OrientSpace (Zhang, Li, & Dou, 2008; Li & Meng, 2008) has been developed, which supports the automatic building of dataspace based on the method of pay-as-you-go evolution and user behavior analysis. However, it abandons full-text indexing; thus its limited query does not satisfy user demands. Elsayed et al. (2006) proposed a dataspace management system architecture that combines the dataspace concept with grid technology (i.e., OGSA). Currently, work combining the concepts of dataspace and cloud computing technology is rare.

2.3   Semantic representation based on domain knowledge

The core of knowledge engineering is the study of the methodologies and technologies for capturing and re-using product and processing engineering knowledge. Its main objective is to reduce the time and cost of product development, which is primarily achieved through design automation enabled by capturing, retaining, and re-using the design knowledge (Verhagen, Bermell-Garcia, Dijk, & Curran, 2012). The transparency and traceability of knowledge is the current research challenge for knowledge representation. To achieve this, researchers have proposed a variety of knowledge representation methods These include semantic representation that introduced ontology technology, an effective method worthy of further exploration (Maedche & Staab, 2001).

Turk (2006) proposed an ontology-based data management method used in the field of construction informatics. Fernandes et al. (2011) proposed a semantic method based on knowledge representation of engineering design to support engineering design innovation. Zhang et al. (2013) proposed a new ontology-based semantic representation model for design rationale (DR) information while presenting the integrated Issue, Solution, Artifact, and Argument (ISAA) model to support product design decisions. Bellazzi et al. (2007) did related research about knowledge-based gene expression data mining, which focused on information evolution and analysis. Tao et al. (2012) proposed terminology representation guidelines for biomedical ontologies in semantic web notations to support the semantic interoperability of biomedical informatics. Alexandridis (2012) proposed the reorganization pattern of social knowledge

representation in evolving semantic networks based on the support of rational propositional inference and formalisms.

Currently, relevant work about knowledge representation in the field of materials engineering is rare. Therefore, although the above application areas are quite different from materials engineering and those methods rarely consider the support of dynamic tracking and timely capturing, they still have a certain utility and significance for the field of materials knowledge representation.

3   SUPPORTING ENVIRONMENT OF KNOWLEDGE REPRESENTATION

As mentioned above, the knowledge representation of the Materials Engineering Application needs an open service support environment to deal with materials data resources that are distributed, heterogeneous, multi-source, associated, variable, and demand-oriented. Based on this, we propose an Open Cloud Services Architecture (OCSA) to emphasize better the characteristics of virtualization and on-demand allocation of data resources. This combines the concept of Virtual DataSpace (VDS) (Liu, Hu, Li, & Hu, 2012) with cloud computing technology (Armbrust, Fox, Griffith, Joseph, Katz, Konwinski, et al., 2010). OCSA characteristics are: (1) data-centricity, i.e., Data as a Service (DaaS); (2) virtualization processing, i.e., "physical dispersion, logical unification"; (3) open associated data evolution tracking; (4) on-demand services; and (5) support for the behavior analysis of user habits, data dissemination, service evolution. The abstract framework for OCSA is shown as in Figure 2. It is defined as follows.

Definition 1. Open Cloud Services Architecture is defined as: OCSA = {DRS, CCSL, VDS, RRM, AIS}, where DRS denotes the data resource set. CCSL denotes the cloud computing support layer that provides the virtualized storage and computing support environment. VDS denotes the Virtual DataSpace that manages the data resources by using semantic mapping and a dynamic evolution mechanism. RRM denotes the requirement representation mode. AIS denotes the application instance set.
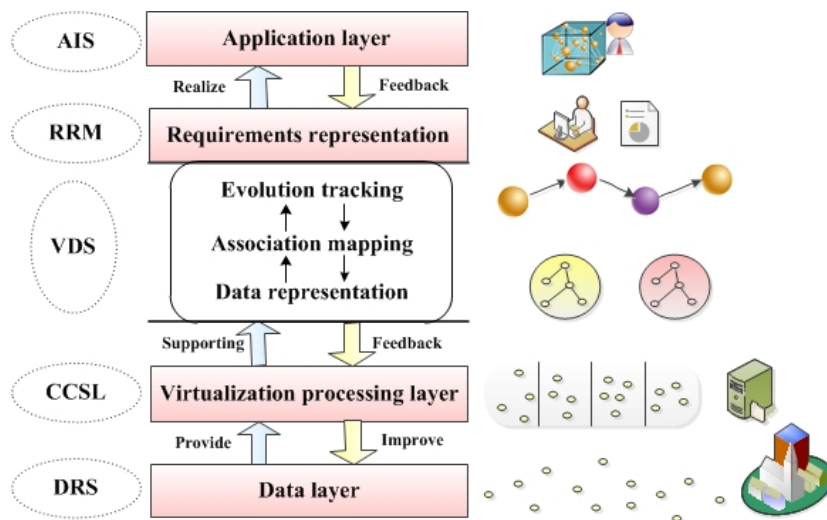

Figure 2. Abstract framework of OCSA

In the OCSA framework, we propose a new data management mode Virtual DataSpace (VDS), which is the set of data, services, and their relationships. Compared with the traditional database (DB) management mode, dataspace (DS) has obvious technological advantages in the aspects of model, operations, objects, relations, and construction costs (Li, Meng, & Zhang, 2008). Further, VDS has additional significant features and advantages, such as the "data first" mode, more emphasis on data association mapping and dynamic evolution, more highlighting of the importance of service, and virtualization processing. The comparison of data management modes among DB, DS, and VDS is shown in Figure 3, and their detailed comparison is described in Table 1. It can be seen that combining VDS with OCSA provides an optimized support environment to solve the issues of knowledge representation, which will then satisfy the MEA knowledge representation that is demand-oriented, complexly associated, and dynamically changing.
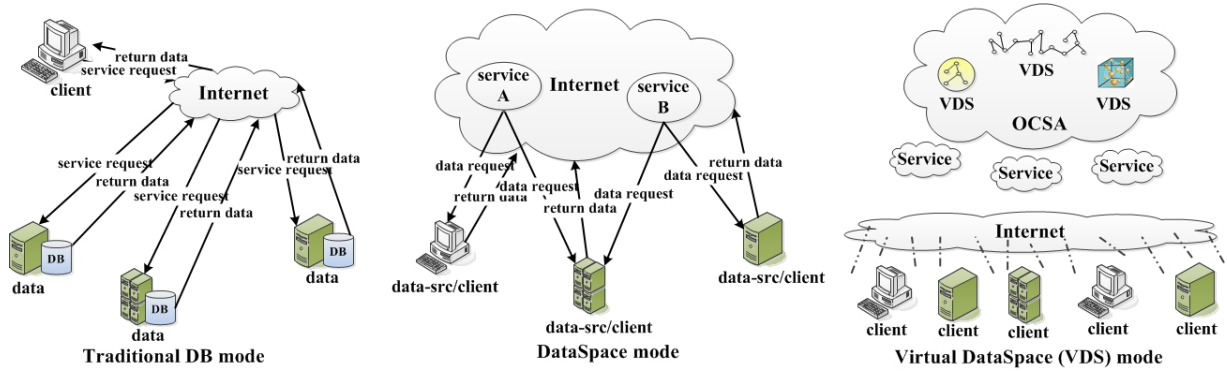
Figure 3. Comparison of DB, DS, and VDS

Table 1. Comparison of DB, DS, and VDS

	DB	DS	VDS
Mode feature	close, focus completion if necessary	open, with "pay-as-you-go" feature	
Data source	only structured data	structured, semi-structured, unstructured data	
Data feature	focus on data stability	emphasis on data association and evolution	
Data storage	remote server	each local	anywhere
Service status	Web service (perpetual)	retrieval services related principal (large concentration)	cloud service (flexible dispersion)
Integration approach	first model, after data	weakening model, highlight data	dilute model, emphasized data, on-demand service
Core	application themes	principal needs	data services
Knowledge representation	does not support	limited support	optimized support
Virtualization	-	-	support virtualization processing
Serviceability	-	-	dynamic, diverse and personalized service
Demand-oriented	-	-	demand-led service coordination, improve service initiative
Behavior analysis	-	-	optimize data services by behavior research, support proactive user-friendly applications

4   A KNOWLEDGE REPRESENTATION MODEL FOR MEA

4.1   Semantic-driven knowledge representation

Based on the open service support environment defined above, the semantic-driven knowledge representation model for MEA is described in Figure 4. First, under the guidance of semantic requirements, this model represents and constructs the semantic representation of data services; second, it describes their semantic association through mapping and reasoning; third, it constantly improves semantic requirements based on evolution representation; and finally, it acquires the required knowledge for the materials engineering application (MEA). Semantic representation is the basis of knowledge acquisition. In view of this, we can obtain

more optimized data services from the knowledge representation of the domain application.
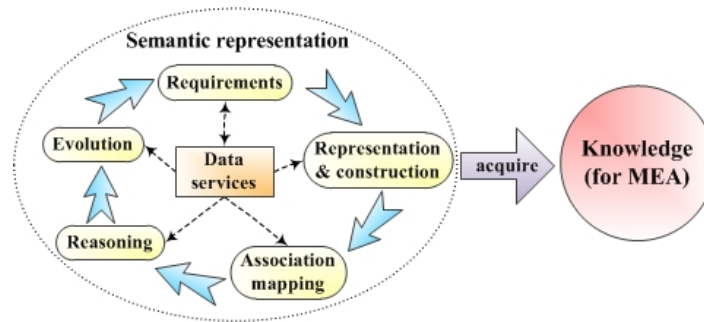

Figure 4. Semantic-driven knowledge representation model for MEA

4.2   Semantic representation of user requirements

A materials engineering application should provide data services according to different user requirements, i.e., it should support on-demand services. To improve the quality of data services, it is necessary to establish a requirement representation mode (RRM) and optimize the semantic relationships between requirements and data services. The RRM is defined as follows.

Definition 2. The requirement representation mode is the semantic representation of user requirements. It consists mainly of two types of contents: one is a description of the various requirements, and the other is the relationship metrics between requirements descriptions and practical applications. RRM can be formally represented as: RRM = {App, Req, W_{R-A}}, where App denotes practical applications, Req denotes user requirements, and W_{R-A} denotes the associated weight between requirements and applications.

The requirement representation mode is illustrated in Figure 5. The common application (App) usually contains the intelligent material selection, safety assessment, life prediction, and the design of new materials. Correspondingly, the requirements include the three aspects listed below Figure 5.
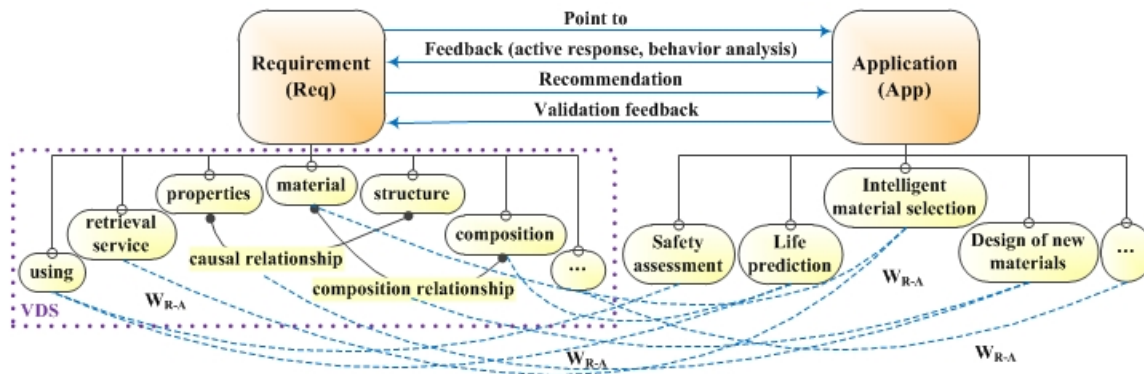
Figure 5. Requirement representation mode (RRM)

1) Data requirements, such as related data about material, property, condition, composition, structure, production, processing, manufacturing, usage, etc. For instance, the safety assessment application needs data concerning how the material is used.
2) Service requirements, such as retrieval, unit conversion, metadata, data upload, quality feedback, traffic monitoring, and so on. For instance, the intelligent material selection application needs the retrieval service.
3) Relationship requirements, such as the causal relationship between property and structure, the composition relationship between material and composition, the inverse relationship between upload service and download service, and the inclusion relationship between retrieval service and property data. For instance, the design of new materials needs to mine the causal relationship between property and structure.

In the RRM, we build a causal feedback mechanism between requirements and applications to improve MEA data service initiative. Normally, the requirement points to the application, and then during the data service, the requirement receives the application feedback in two forms: active responses from users and implicit and automatic acquisition from behavior analysis. Based on the feedback, the requirement recommends appropriate data services to the application and then receives validation feedback from the application, thus constantly improving and optimizing the data service recommendations.

Assuming that the total number of applications in MEA is m, the associated weight (W_{R-A}) between the requirements and the k-th application is described in Eq. (1), where n is the total number of requirements that correspond to the k-th application in MEA. W_{Ri} denotes the satisfaction degree of requirement Ri for the application Ak, and W_{Ai} denotes the demand degree of application Ak for the requirement Ri. W_{Ri} and W_{Ai} are pointing in the opposite directions, and their values are usually different, but close. Their idealized values should be identical. Therefore, in reality, they should be as infinitely close as possible.

$$W_{R-A}(k) = sim_k\left(W_R, W_A\right) = \frac{2\left[\sum_{i=1}^{n}\left(W_{Ri}W_{Ai}\right)\right]}{\sum_{i=1}^{n}W_{Ri} + \sum_{i=1}^{n}W_{Ai}} (k = 1, 2, \cdots, m)$$

(1)

In a more refined manner, RRM also can be expressed as:

RRM = {App<A_name, A_user, A_describe>, Req<R_data, R_services, R_relationships>},

where R_data, R_services, and R_relationships are all weighted semantic representations and have similar forms. For example, R_data can be represented as: R_data = {<data$_1$, Wight$_1$>, <data$_2$, Wight$_2$>… <data$_n$, Wight$_n$>}.

It can be seen that the set of required data, services, and relationships, is in fact the virtual dataspace (VDS). The VDS provides the bottom-up "data leading" to satisfy the service demands. In contrast, RRM provides the top-down "requirement leading" to optimize the data service. We can completely open up the underlying data resources and the upper service applications by combining RRM and VDS.

4.3   Data representation based on ontology modeling

Extremely small changes in materials property data can cause huge differences in conclusions; therefore, MEA data representation needs a high degree of accuracy. Modeling the ontology of semantic information means constructing conceptual models that represent the concepts and the relationships between them in engineering applications (Zhang, Luo, Li, & Buis, 2013). The data representation method based on ontology modeling can shield the heterogeneity of the information and then effectively abstract and organize the complex data and relationships.

4.3.1   Semantic representation in VDS

As mentioned above, requirement representation guides the expression of data, services, and relationships, i.e., RRM guides the semantic expression of the VDS. For abstracting and capturing the MEA domain knowledge, we must build a conceptual model of the semantic representation in VDS.

Definition 3. Virtual DataSpace (VDS) is the set of data and services and their relationships. It supports the management of data resources by using the support mechanisms of semantic representation, association mapping, and dynamic evolution. Broadly speaking, VDS means the entire public virtual dataspace (P-VDS), which points to all the application instance sets (AIS). P-VDS is defined as: P-VDS = {ADS, ASS, ARS}, where ADS denotes all the data sets, ASS denotes all the services sets, and ARS denotes all the relationships sets.

Specifically VDS means the subject related data, services, and relationships, i.e., the sub virtual dataspace (S-VDS). S-VDS is the subset of P-VDS, i.e., P-VDS = $\sum$S-VDS$_i$, i=1, 2,…,V^i. S-VDS$_i$ points to the specific application instance set (AIS$_i$), where V^i is the number of sub VDSs in the P-VDS, i.e., V^i is also the number of application instance sets that are related to the corresponding subject i. For the subject i, S-VDS$_i$ is formally represented as: S-VDS$_i$ = {DS$_i$, SS$_i$, RS$_i$, AIS$_i$}, where DS$_i$ denotes the data set relevant to the subject i, SS$_i$

denotes the services set relevant to the subject i, RS_i denotes the relationships set relevant to the subject i, and AIS_i denotes the application instance set relevant to the subject i.

The data set of S-VDS$_i$ is expressed as: $DS_i = \sum DE_{id}$, d=1,2,...,D^i, where D^i is the number of data entities in the data set of this sub VDS that are related to the subject i, and DE_{id} is the data entity. DE_{id} is defined as: DE_{id} = {D_{type}, D_{level}, D_{sour}, D_{Mat}, D_{onto}, D_{desc}, W_{i-d}}, where D_{type} denotes the data type, D_{level} denotes the data level, D_{sour} denotes the source of data, i.e., the storage location of the original data, D_{Mat} denotes the material features of the data entity, D_{onto} denotes the semantic representation of the data ontology as the unique identifier of the data entity, D_{desc} denotes the data description, and W_{i-d} denotes the weight of data d for subject i. D_{type} includes the numeric type, text, image, etc. The data type guides the form of expression of D_{desc}. For example, for numeric data, D_{desc} is described as <value, range, accuracy, max, min, unit>; for text data, D_{desc} is described as <text content, text length, text url>; for image data, D_{desc} is described as <image name, image type, image description, image size, image url>. D_{level} includes mainly class, property, individual, etc. By combining the ontology description (D_{onto}) and the different data types (D_{type}), we build the semantic ontology of the data representation for different data levels (D_{level}) in a corresponding manner. For a detailed description, see Section 4.3.2. D_{Mat} is performance data, structural data, experimental data, simulated data, and so on. Because there are associated logical organizations among these data, when they are combined with the data weight (W_{i-d}), we are able to discover the foundation of mapping and reasoning. For a detailed description, see Section 4.4.

In P-VDS, the total number of data is expressed as $N_{data} = \sum D^i$, i=1,2,..., V^i.

Similarly, the services set S-VDS$_i$ is expressed as: $SS_i = \sum SE_{is}$, s=1,2,...,S^i, where S^i is the number of service entities in the services set of this sub VDS that are related to the subject i, and SE_{is} is the service entity defined as: SE_{is} = {S_{type}, S_{Mat}, S_{onto}, S_{desc}, W_{i-s}}.

This also holds for the descriptions of SE_{is} and DE_{id}, except that S_{type} usually includes browse, inquiry, interact, simulation design, etc.

Correspondingly, the relationships set of S-VDS$_i$ is expressed as: $RS_i = \sum RE_{ir}$, r=1,2,...,R^i, where R^i is the number of relationship entities in the relationships set of this sub VDS that are related to the subject i, and RE_{ir} is the relationship entity defined as: RE_{ir} = {R_{type}, R_{Mat}, R_{onto}, R_{desc}, W_{i-r}}.

RE_{ir} and SE_{is} are described similarly, except that R_{type} usually includes similar, opposition, neighbor, causation, and so on relationships. R_{desc} is described as: R_{desc} = {<Item1, Item2, ..., ItemN>, CD_{items}}, where the 'Item' is data, service, or the relationship itself. N is usually 2 or more but can be 1 when it represents a reflexive relationship. CD_{items} denotes the correlation degree among these items.

The application instance set of S-VDS$_i$ is expressed as: $AIS_i = \sum AI_{ia}$, a=1,2,...,A^i, where A^i is the number of application instances in the application instance set of this sub VDS that are related to the subject I, and AI_{ia} is the application instance defined as: AI_{ia} = {AI_{onto}, AI_{desc}, AI_{user}, W_{i-a}}, where AI_{user} denotes the relevant user of the application instance.

It can be seen that based on the semantic representation and analysis of the relevant parameters of the VDS and RRM, we have communication between the upper demands and the underlying data. Thus we achieve top-down demand-driven data services and bottom-up data-affected intelligence applications. See Figure 6.
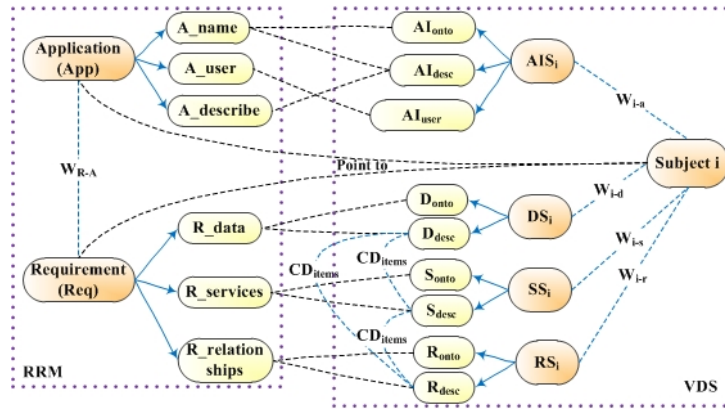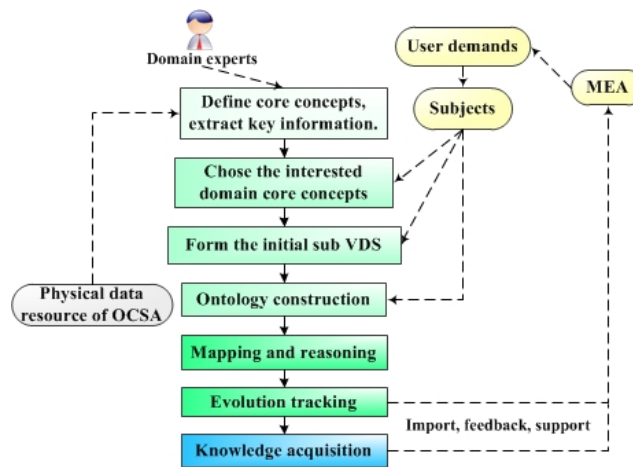
Figure 6. Get through the RRM and VDS

Figure 7. Building process of the VDS

We build the VDS as follows.
Step 1. Domain experts define the core concepts in the materials field, and then key information from the physical data resources of OCSA is extracted;
Step 2. The interested domain core concepts around specific subjects is chosen and form an initial sub VDS on-demand;
Step 3. The semantic ontology of the data representation is constructed;
Step 4. The association mapping and reasoning is established; and
Step 5. Based on the evolution tracking, knowledge acquisition is achieved.

The process of building the VDS is illustrated in Figure 7. Through a dynamic evolution cycle, the VDS supports the continuous improvement and optimization of MEA's intelligent services.

4.3.2 Ontology construction of data representation

Considering the parameter D_{type} in definition DE_{id}, we classify the different data types into three main categories: structured data, semi-structured data, and unstructured data. Making table, XML, and image the respective examples, we construct an ontology of the data representation using different methods. The ontology construction method is illustrated in Figure 8.
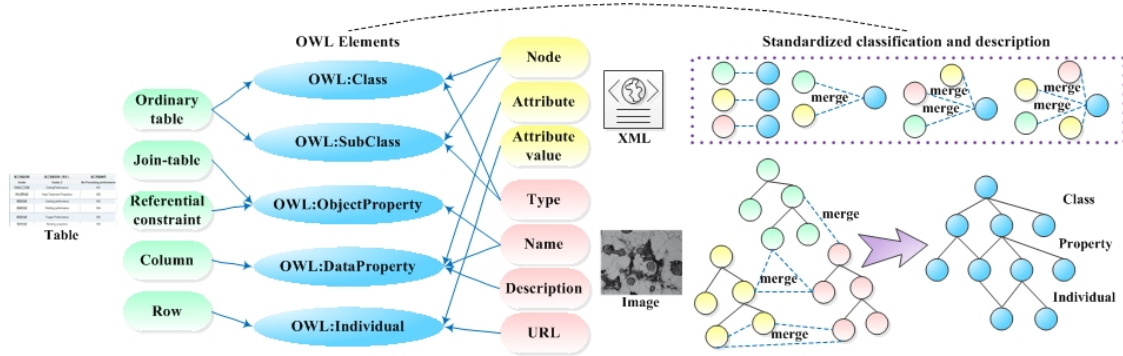
Figure 8. Ontology construction method for data representation

(a) For structured tables in a relational database, we adopt the following conversion rules to build the OWL ontologies.
Rule a1: Ordinary tables convert into classes or subclasses (OWL:Class or OWL:SubClass).
Rule a2: Join-tables and referential constraints of tables convert into object properties (OWL:ObjectProperty).
Rule a3: Columns of tables convert into data properties (OWL:DataProperty).
Rule a4: Rows of tables convert into individuals (OWL:Individual).

(b) For a semi-structured XML whose elements include node and attribute, the conversion from XML to OWL ontology is divided into content convert and relationship convert. We adopt the following conversion rules to build the OWL ontologies.
Rule b1: XML nodes convert into OWL classes.
Rule b2: XML attributes convert into OWL data properties.
Rule b3: XML attribute values convert into OWL individuals.
Rule b4: Parent-child relationships in XML are converted into class-subclass relationships of OWL ontology.
Rule b5: Element-attribute relationships in XML are converted into class-data property relationships of OWL ontology.

(c) For an unstructured image, we adopt the following conversion rules to build the OWL ontologies.
Rule c1: Image types convert into classes or subclasses.
Rule c2: Image names convert into object properties or data properties.
Rule c3: Image descriptions convert into data properties.
Rule c4: Image URLs convert into individuals.

The converted data contents are classified according to the parameter D_{level}, i.e., the data levels are divided into class, property, individual, etc. Meanwhile, the semantics representation of the data is uniformly described using parameter D_{onto}. Based on the semantic level, we merge the same data entities and directly convert the other data entities, thus achieving the standardized ontology construction of the data representation.

4.4   Semantic representation of association mapping and reasoning

The constructed ontology model still lacks semantic representation of the association mapping and reasoning. This is difficult to find and understand for the implied relationships within the MEA information. In particular, the material structure data are closely related to the material property data and discovering the structure-property relationships is a crucial and meaningful issue (Rajan, 2005). However, the ontology lacks a semantic representation between these two types of data, and these relationships are often not linear. Therefore, we must build a multi-scale semantic representation pattern based on the association mapping and reasoning to accurately capture such related complex information in the MEA.

4.4.1   Associated semantic mapping

From the perspective of its material characteristics, the parameter D_{Mat} in definition DE_{id}, mainly includes material grade data, material classification data, structural data, performance data, condition data, composition data, appearance data, production process data, auxiliary data, and so on. On the other hand, the parameter D_{Mat}

from the perspective of the application process mainly includes issue data, target data, design data, solution data, simulated data, experimental data, usage data, evaluation data, and so on. By considering both these aspects, we build the associated logical organization between them, illustrated in Figure 9.
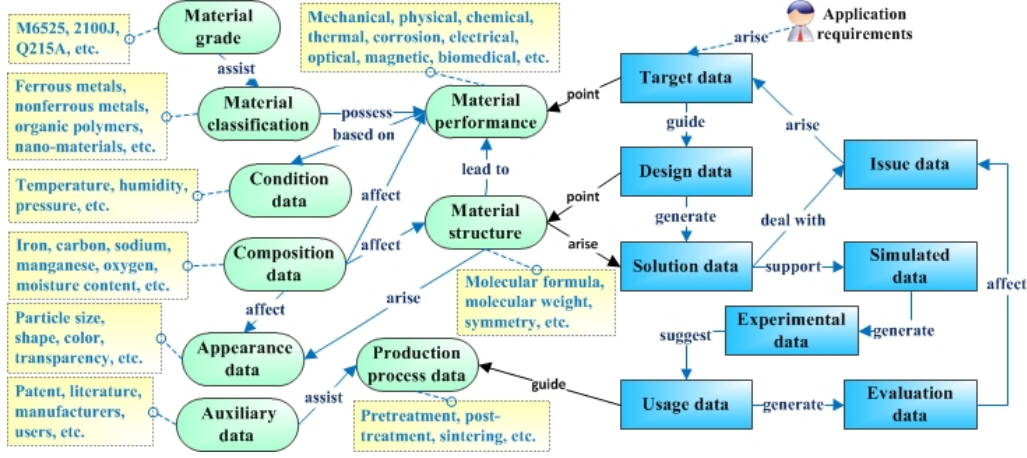

Figure 9. Associated logical organization of the material data

The most important relationship in this organization is that between the material structure and performance. Therefore, we design a materials structure from the required material performance. This process borrows the idea of reverse engineering in gene expression (Bellazzi & Zupan, 2007) and uses it to mine deeply hidden information in the MEA. Thus it supports in-depth exploration using the principle of material composition, becoming the theoretical foundation of the "Materials Genome Project" (Kalil & Wadia, 2011). Analogously, based on the associated logical organization of the material data, we construct the associated semantic mapping by using the parameter D_{Mat} in definition DE_{id}, parameter S_{Mat} in definition SE_{is}, and parameter R_{Mat} in definition RE_{ir}. Of course, other parameters also affect the mapping establishment.

For example, around subject i, there are two data entities:
(1) DE_{i1} = {text, individual, node_1, structure: molecular formula, Lithium niobate, <"LiNbO$_3$", 20kb, local link>, 0.67}
(2) DE_{i2} = {numeric, individual, node_2, performance: thermal: melting point, Lithium niobate, <1528.0, 20, 1, null, null, ℃>, 0.72}

The D_{onto} parameters are the same, i.e., $DE_{i1}.D_{onto}$ = $DE_{i2}.D_{onto}$ = "Lithium niobate", and they have the same data level, i.e., $DE_{i1}.D_{level}$ = $DE_{i2}.D_{level}$ = "individual". In the same way, the D_{Mat} parameters have a logical association "structure leads to performance". Therefore, we can establish a new association mapping between DE_{i1} and DE_{i2}. This mapping is described as "DE_{i1} leads to DE_{i2}". Meanwhile, combining this with the data weight (W_{i-d}), we calculate the similarity between DE_{i1} and DE_{i2}. See Eq. (2).

$$W_{i1-i2} = \frac{2W_{i-1}W_{i-2}}{W_{i-1}+W_{i-2}} = \frac{2 \times 0.67 \times 0.72}{0.67 + 0.72} \approx 0.694$$

(2)

Accordingly, we construct the associated semantic mapping shown in Figure 10. The semantic representation of the association mapping is described as follows.

Definition 4. Associated Semantic Mapping (ASM) means the semantic representation of mapping between entities that can be any data, services, or relations. ASM is defined as: ASM = Mapping ($E_1 \rightarrow E_2$) = {E_1, E_2, M_{type}, M_{onto}, W_{map}}, where E_1 denotes the mapping starting point, E_2 denotes the mapping ending point, M_{type} denotes the mapping type, M_{onto} denotes the ontology description of the mapping, and W_{map} denotes the associated mapping weights, i.e., the similarity between E_1 and E_2.
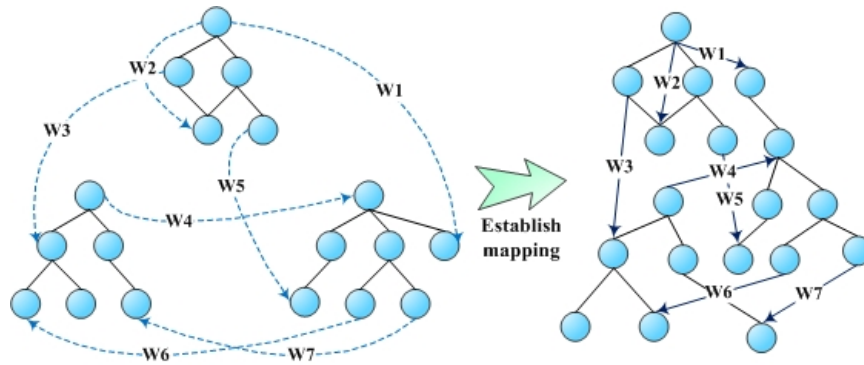
Figure 10. Construction of the associated semantic mapping

When establishing the semantic mapping, we add the relevant mapping parameters into the corresponding relationship entity (RE_{ir}), i.e., new relationships are created according to the mapping. Using the above example, we build the mapping between DE_{i1} and DE_{i2}: Mapping ($DE_{i1} \rightarrow DE_{i2}$) = {$DE_{i1}$, DE_{i2}, causality, lead to, 0.694}. Then we create the corresponding relationship entity: RE_{ir} = {causality, structure-performance, lead to, {<DE_{i1}, DE_{i2}>, 0.694}, W_{i-r}}. It can be seen that corresponding relationships exist between the parameters M_{type} and R_{type}, M_{onto} and R_{onto}, E_1 and Item1, E_2 and Item2, W_{map} and CD_{items}, where N=2 for ItemN in parameter $R_{desc.}$, and R_{Mat} is the <u>joint</u> of parameters D_{Mat}, S_{Mat}, or R_{Mat} in E_1 and E_2. Initially, we set the parameter W_{i-r} as CD_{items}, i.e., W_{i-r} = CD_{items} = W_{map}.

4.4.2   Rule-based reasoning

In order to discover and define the implicit semantic relations and then further optimize MEA knowledge representation, we chose the Semantic Web Rule Language (SWRL) (Horrocks, Patel-Schneider, Boley, Tabet, Grosof, & Dean, 2004), based on OWL, the Web Ontology Language, to define new rules for assisting the VDS ontology to realize MEA's semantic reasoning. The typical reasoning rules in the materials field are described as follows.

Rule 1: Classified integration rule

If several entities have the same mapping type of associated relationship with the same entity, then the entities with the same entity type, data level, or material feature might also have weight values similar to this specific entity; meanwhile, a "similar relation" is also likely to exist among these entities. The reasoning process of classified integration rule is shown in Figure 11. This rule supports the discovery of similar relations and also partly optimizes the weight values of the existing mappings.
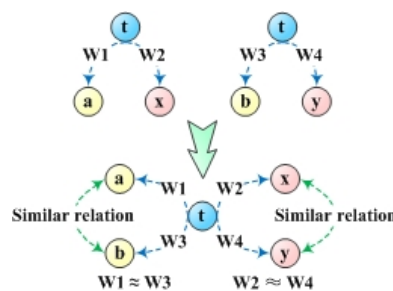
Figure 11. Reasoning process of the classified integration rule

Rule 2: Propagation rule of transitive relation

Many associated relationships, to a certain extent, have transitivity, such as the mapping relations of "affect", "based on", "guide", "generate", "support", and so on. The influence of these association mappings is

propagated through corresponding entities and relations. The propagation rule's reasoning process is shown in Figure 12. This rule supports the discovery of new relations and also partly optimizes the weight values of the new mappings.

Figure 12. Reasoning process of the propagation rule

The new weights depend on all the existing related weight values. When the respective weights are equal, the more occurrences there are of the transitive entities and the larger their associated weight values, i.e., $W_D > W_E \approx W_G > W_F$; otherwise, all of these weights need to be calculated according to their corresponding equations. Normally, we assume the initial associated entities are i1, i2, …, ik for the new entity "X" and its frequency of occurrence "n". The process for calculating W_X is shown in Eq. (3).

$$W_X = \frac{2 \sum_{I=i1}^{ik} W_I \times W_{IX}}{\sum_{I=i1}^{ik} (W_I^{n-1} + W_{IX}^{n-1})}$$

(3)

Rule 3: Reverse rule of inverse relations

Some associated relationships have reversibility, for example, "based on" has the inverse relation "arise", "lead to" has the inverse relation "source tracing", etc. According to this, the reverse rule supports the discovery of inverse relations. Then, based on reverse engineering, it assists scientific innovation such as the design of new materials. The reverse rule's reasoning process is shown in Figure 13.
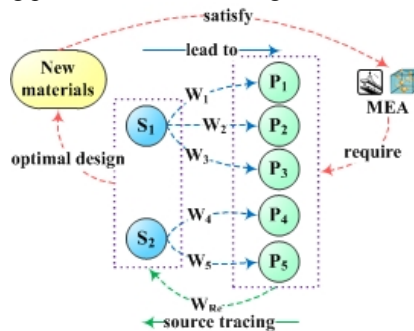
Figure 13. Reasoning process of reverse rule

For example, coating materials for an aircraft engine need to have the performance characteristics of low specific gravity, thermostability, antioxidative, and high tenacity. Considering that aluminum (Al) has the performance advantages of low density and good ductility and that titanium (Ti) has the performance advantages of corrosion and high temperature resistance, we can integrate their advantages and use knowledge about phase diagrams, crystal structures, materials processing, etc. to design an appropriate coating material. The characteristics of low specific gravity and high tenacity source trace to aluminum (Al) and the characteristics of antioxidative and thermostability source trace to titanium (Ti). Therefore titanium aluminum alloy materials are considered to be a key target.

Meanwhile, different kinds of phase diagrams, crystal structures, and materials processing "lead to" different material properties so that specific material properties "source trace" to specific phase diagrams, crystal

structures, and materials processing. We can, therefore, design optimal materials according to the required material properties by considering the characteristics of the elemental materials (Al and Ti), phase diagrams, crystal structures, materials processing, and so on. This process is precisely the usage of the inverse relation between "lead to" and "source tracing". It shows that this reverse rule has a good supporting role in the optimized designing and innovative discovery of new materials.

4.5   Knowledge acquisition based on evolution tracking

Considering that material data update frequently and that their dynamic changes are difficult to capture, the timely acquisition of MEA related knowledge needs the semantic representation of evolution tracking. As mentioned above, changing data service user demands power the VDS evolution. Although requirements are constantly changing, they do provide corresponding feedback. Combined with the requirement representation mode, the three types of feedback are as follows.

(1) Active response: active acceptance or negation from users, including validation feedback from an application after getting a corresponding recommendation.
(2) Behavior analysis: discovering hidden information through an analysis of user behavior in data manipulation, service usage, page browsing, and so on.
(3) Changing data source: through an active response or implied mining, a data source change can also have corresponding feedback.

Based on the above feedback, the dynamic evolution of the knowledge representation (i.e., data, services, and semantic relations) can be traced respectively through time and space dimensions. For a positive response or analysis (acceptance or frequent behavior), the weight value of related knowledge representation will increase. For a negative response or analysis (negation or rare behavior), the weight value of related knowledge representation will decrease. For a change of data source, timely detection of changes is mainly supported by the association mapping and reasoning. Accordingly, we can continuously optimize knowledge representation and acquire required knowledge in a timely manner.

5   APPLICATION EVALUATION IN MATERIALS ENGINEERING

To satisfy the materials engineering application based on the semantic-driven knowledge representation model, we have developed a "Materials Scientific Data Sharing Service Platform" as shown in Figure 14.
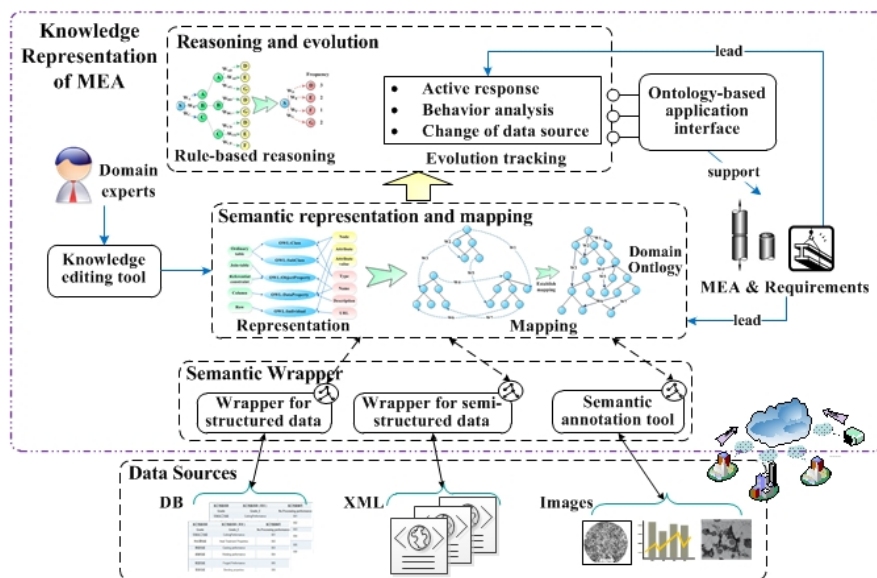
Figure 14. Framework of the Materials Scientific Data Sharing Service Platform

In the materials field, this platform integrates data sources that are massive, heterogeneous, and distributed in different regions. These data sources are aggregated in a cloud platform and organized according to their different data types. They can be structured data such as two-dimensional data tables from databases, semi-structured data such as XML or HTML files, or unstructured data such as images or PDF files. The

different data sources are displayed on the platform page in various forms, such as metadata browsing, data navigation, keyword search, and data visualization. They are updated through the data submit module. On this basis, we use different formalization methods to reconstruct these data sources and form the virtualized data processing layer to reorganize and analyze the data. For different types of data, we have developed different rules to convert the data sources into unified structured data expression patterns. For structured data, such as two-dimensional data tables from databases, we use the rows and columns conversion method to extract the data mode. For semi-structured data, such as XML or HTML files, we use the data nodes conversion method to extract the data mode. For unstructured data, such as images, we use semantic annotation technology to extract the data mode. For unstructured data, such as PDF files, we use natural language processing technology to extract the data mode.

The whole process needs to consider the metadata dictionary while the converted data mode forms the domain ontology using the semantic mechanism. Based on an open cloud service environment, to meet various needs of MEA, the semantic representation of different types of data is built by extracting semantic information from complex data sources as shown in Figure 15. We realize the materials field's knowledge acquisition through semantic mapping, reasoning analysis, and evolution tracking. We use Protégé, an open-source ontology editor, to generate the materials ontology in an OWL file and then use interactive graphics software, TouchGraph, to generate the MEA's visual semantic model. See Figure 16.
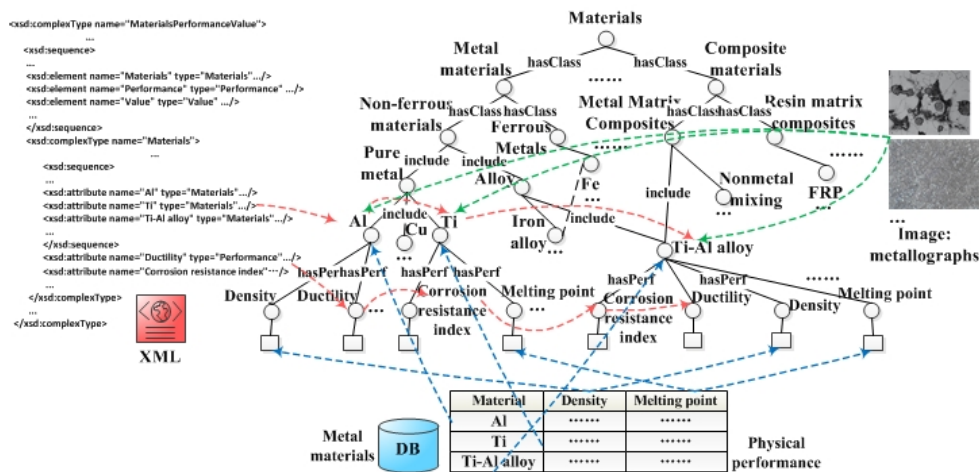
Figure 15. Partial semantic extraction and representation of the MEA
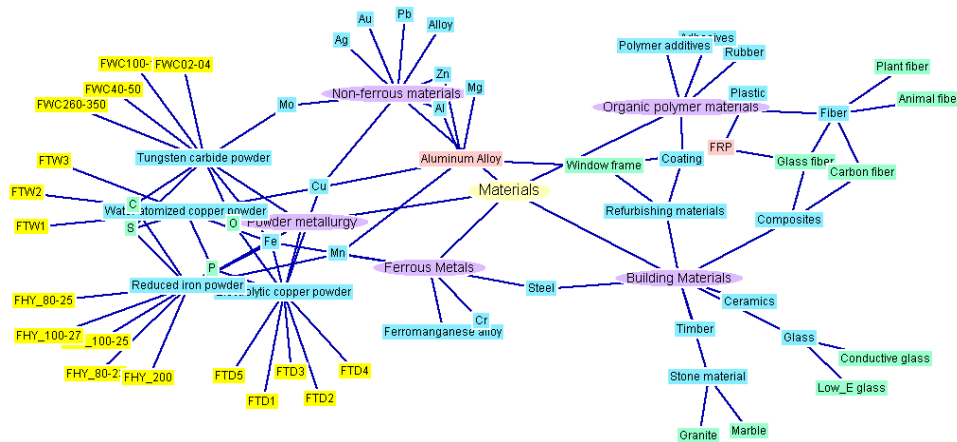
Figure 16. Partial visual knowledge representation model of the MEA

Figure 17. An application instance of a materials selection recommendation

Based on the semantic-driven knowledge representation model, we obtain the MEA domain knowledge to provide timely and accurate service applications. A typical application of the materials field, "materials selection recommendation", is illustrated in Figure 17. First, we select the required material category and performance parameters; then we retrieve the relevant material information according to a free combination of conditions. For specific materials, we further view detailed information and recommend related materials, literature, etc. Finally, we sort the recommended contents according to their degree of similarity. This application case demonstrates the effectiveness of the semantic-driven knowledge representation model and achieves optimal and intelligent domain data services.

At present, the materials scientific data sharing service platform (http://matsec.ustb.edu.cn/matsharing) has been on-line for 2 years, and the number of visits has reached 170,000. This platform has collected nearly 600,000 items of data resources, the data volume has reached 1 terabyte, and it still growing rapidly. We have accumulated 1791 data tables, 672 xml files, and 4216 unstructured data files, which include images, videos, etc. For these complex and changeable MEA data resources, the converted semantic information is classified and counted, as shown in Table 2.

Table 2. The classification and statistics of MEA data resources

Data type	Class	Property	Instances
Structured data files (tables)	1612	18,719	421,268
Semi-structured data files (xml)	1239	6728	78,902
Unstructured data files (images, etc.)	476	82,134	82,134

Fifty thousand different types of data items have been selected as sample sets with which to compare accuracy and timeliness among the SQL query for tables, the XPath query for XML files, a simple search for unstructured data files, and a semantic query based on the knowledge representation model for all data types. As shown in Figure 18, as the number of data items increases, this model has the most significant increase in accuracy and the smallest increase in time. Through an experimental evaluation, it can be seen that the semantic-driven knowledge representation model has a significant advantage over any other single method. This knowledge representation model supports timely, accurate, and personalized MEA data services.
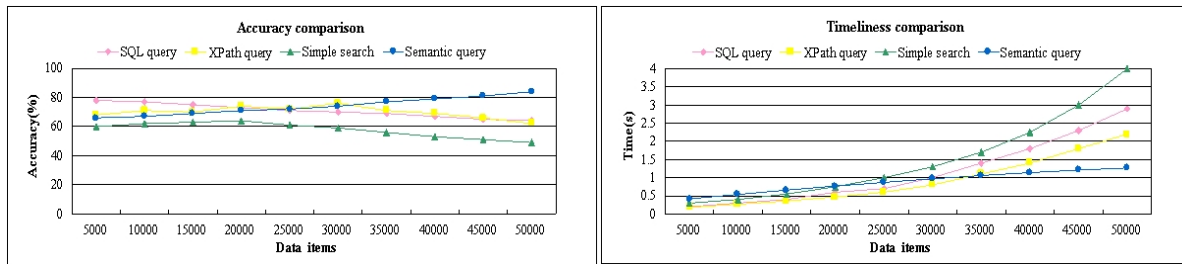
Figure 18. Experimental evaluation of the MEA knowledge representation model

6   CONCLUSIONS AND FUTURE WORK

This paper proposes a semantic-driven knowledge representation model for the materials engineering application (MEA). Using a supporting environment of open cloud services architecture (OCSA), the semantic representation of requirements, data, services, and their relationships has been constructed. Based on ontology modeling in the VDS, the semantic representation of association mapping, rule-based reasoning, and evolution tracking has been discussed in support of MEA knowledge acquisition. Through evaluating and analyzing the application in the field of materials engineering, the accuracy and timeliness of this model have been validated. Future research work should have three aspects: (1) further refining MEA knowledge representation model; (2) improving the mechanism of semantic mapping and reasoning; and (3) an in-depth study of the evolution tracking issue of MEA domain knowledge.

7   ACKNOWLEDGEMENTS

8   REFERENCES

Alexandridis, K. & Maru, Y. (2012) Collapse and reorganization patterns of social knowledge representation in evolving semantic networks. *Information Sciences 200*(1), 1-21.

AlKhazraji, Q.Y., Saldana, C., & Kumara, S. (2013) Material Information Model across Product Lifecycle for Sustainability Assessment. *Re-engineering Manufacturing for Sustainability*, ,Springer: Singapore, 511-514.

Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., et al. (2010) A view of cloud computing. *Communications of the ACM 53*(4), 50-58.

Ashino, T. (2010) Materials Ontology: An Infrastructure for Exchanging Materials Information and Knowledge. *Data Science Journal 9*, 54-61.

Bellazzi, R. & Zupan, B. (2007) Towards knowledge-based gene expression data mining. *Journal of Biomedical Informatics 40*(6), 787-802.

Blunschi, L., Dittrich, J.P., Girard, O.R., Karakashian, S.K., & Salles, M.A.V. (2007) A dataspace odyssey: The iMeMex personal dataspace management system. *Proc. 3rd Biennial Conference on Innovative Data Systems Research* (CIDR), Asilomar, California, USA, 114-119.

Cebon, D. & Ashby, M.F. (2006) Engineering Materials Informatics. *MRS Bulletin 31*(12), 1004-1012.

Dong, X. & Halevy, A. (2005) A platform for personal information management and integration. *Proc. 2005 VLDB PhD Workshop* (CIDR), 26-30.

Elsayed, I., Brezany, P., & Tjoa, A.M. (2006) Towards Realization of Dataspaces. *Proc. 17th International Conference on Database and Expert Systems Applications* (DEXA'06), Krakow, Poland, 266-272.

Fernandes, R.P., Grosse, I.R., Krishnamurty, S., Witherell, P., & Wileden, J.C. (2011) Semantic methods supporting engineering design innovation. *Advanced Engineering Informatics 25*(2), 185-192.

Ferris, K.F., Peurrung, L.M., & Marder, J.M. (2007) Materials informatics: Fast track to new materials. *Advanced Materials & Processes 165*(1), 50-51.

Foster, I., Berry, D., Djaoui, A., Grimshaw, A., Horn, B., Kishimoto, H., et al. (2005) The Open Grid Services Architecture. *Global Grid Forum* (GFD-I) 30.

Franklin, M., Halevy, A., & Maier, D. (2005) From databases to dataspaces: a new abstraction for information management. *ACM Sigmod Record 34*, 27-33.

Grossman, R.L., Gu, Y.H., Mambretti, J., Sabala, M., Szalay, A., & White, K. (2010) An overview of the Open Science Data Cloud. *Proc. 19th ACM International Symposium on High Performance Distributed Computing* (HPDC'10), Chicago, Illinois, USA, 377-384..

Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B., & Dean, M. (2004) SWRL: A semantic web rule language combining OWL and RuleML. W3C Member submission. Retrieved November 27, 2013 from the World Wide Web: http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., et al. (2008) Big data: The future of biocuration. *Nature 455*(7209), 47-50.

Hunt, W.H. (2006) Materials informatics: Growing from the bio world. *JOM Journal of the Minerals, Metals and Materials Society 58*(7), 88.

Kalil, T. & Wadia, C. (2011) *Materials Genome Initiative: A Renaissance of American Manufacturing*. Retrieved from the World Wide Web April 21, 2014: http://www.whitehouse.gov/blog/2011/06/24/materials-genome-initiative-renaissance-american-manufacturing

Li, Y.K. & Meng, X.F. (2008) Research on personal dataspace management. *Proc. 2008 SIGMOD PhD Workshop,* Vancouver, BC, Canada, 7-12.

Li, Y.K., Meng, X.F., & Zhang, X.Y. (2008) Dataspace technology research. *Journal of Software 19*(8), 2018-2031.

Liu, Z., Hu, C., Li, Y., & Hu, J. (2012) DSDC: A Domain Scientific Data Cloud Based on Virtual Dataspaces. *Proc. Parallel and Distributed Processing Symposium Workshops & PhD Forum* (IPDPSW), Shanghai, China, 2176-2182

Lynch, C. (2008) Big data: How do your data grow? *Nature 455*(7209), 28-29.

Maedche, A. & Staab, S. (2001) Ontology learning for the semantic web. *Intelligent Systems 16*(2), 72-79.

Rajan, K. (2005) Materials informatics. *Materials Today 8*(10), 38-45.

Shen, W., Hao, Q., Yoon, H.J., & Norrie, D.H. (2006) Applications of agent-based systems in intelligent manufacturing: An updated review. *Advanced Engineering Informatics 20*(4), 415-431.

Tao, C., Pathak, J., Solbrig, H.R., Wei, W.Q., & Chute, C.G. (2012) Terminology representation guidelines for biomedical ontologies in the semantic web notations. *Journal of Biomedical Informatics 46*(1), 128-138.

Turk, Ž. (2006) Construction informatics: Definition and ontology. *Advanced Engineering Informatics 20*(2), 187-199.

Ullah, A.M.M. & Harib, K.H. (2008) An intelligent method for selecting optimal materials and its application. *Advanced Engineering Informatics 22*(4), 473-483.

Verhagen, W.J.C., Bermell-Garcia, P., Dijk, R.E.C.V., & Curran, R. (2012) A critical review of Knowledge-Based Engineering: An identification of research challenges. *Advanced Engineering Informatics 26*(1), 5-15.

Zhang, X.Z., Li, Y.K., & Dou, Y.B. (2008) OrientSpace: Personal dataspace management prototype system. WAMDM Technical Report.

Zhang, Y., Luo, X., Li, J., & Buis, J.J. (2013) A semantic representation model for design rationale of products. *Advanced Engineering Informatics 27*(1), 13-26.

Zhao, L.R., Chen, K., Yang, Q., Rodgers, J.R., & Chiou, S.H. (2005) Materials informatics for the design of novel coatings. *Surface and Coatings Technology 200*(5), 1595-1599.

Zhao, Y.F., Perry, N., & Andriankaja, H. (2013) A Manufacturing Informatics Framework for Manufacturing Sustainability Assessment. *Re-engineering Manufacturing for Sustainability,* Springer: Singapore, 475-480.