RESEARCH PAPER

Research on Intelligent Organization and Application of Multi-source Heterogeneous Knowledge Resources for Energy Internet[①]

Wang Yuxuan, Luo Liqun, Li Guangjian[†]

Department of Information Management, Peking University, Haidian District, Beijing 100871, China

Keywords: Energy internet; Knowledge resources; Body of knowledge; Intelligent knowledge organization

Citation: Wang, Y.X., Luo, L.Q., Li, G.J.: Research on Intelligent Organization and Application of Multi-source Heterogeneous Knowledge Resources for Energy Internet. Data Intelligence 5(1), 75-99 (2023). doi: dint_a_00158
Received: February 28, 2022; Revised: April 9, 2022; Accepted: May 2, 2022

ABSTRACT

To improve the informationization and intelligence of the energy Internet industry and enhance the capability of knowledge services, it is necessary to organize the energy Internet body of knowledge from existing knowledge resources of the State Grid, which have the characteristics of large scale, multiple sources, and heterogeneity. At the same time, the business fields of State Grid cover a wide range. There are many sub-fields under each business field, and the relationship between fields is diverse and complex. The key to establishing the energy Internet body of knowledge is how to fuse the heterogeneous knowledge resources from multiple sources, extract the knowledge contents from them, and organize the different relationships. This paper considers transforming the original knowledge resources of State Grid into a unified and well-organized knowledge system described in OWL language to meet the requirements of heterogeneous resource integration, multi-source resource organization, and knowledge service provision. For the State Grid knowledge resources mainly in XML format, this paper proposes a Knowledge Automatic Fusion and Organization idea and method based on XSD Directed Graph. According to the method, the XML corresponding XSD documents are transformed into a directed graph in the first stage during which the graph neural network detects hidden knowledge inside the structure to add semantic information to the graph.

In the second stage, for other structured knowledge resources (e.g., databases, spreadsheets), the knowledge contents and the relationships are analyzed manually to establish the mappings from structured resources to graph structures, using which the original knowledge resources are transformed into graph structures, and

[†] Corresponding author: Guangjian Li (E-mail: ligj@pku.edu.cn; ORCID: 0000-0002-2897-6246).

[①] This paper is the research result of "Research and Application of Dynamic Knowledge Map Technology for Energy Internet (1200/2021-66002A), a science and technology project of the State Grid Corporation of China.

merged with the directed graphs obtained in the first stage to achieve the fusion of heterogeneous knowledge resources. And expert knowledge is introduced for heterogeneous knowledge fusion to further extend the directed graph. And in the third stage, the expanded directed graph is converted to the body of knowledge in the form of OWL. This paper takes the knowledge resources in the field of human resources of the State Grid as an example, to establish the ontology of the human resources training field in a unified manner, initially demonstrating the effectiveness of the proposed method.

1. INTRODUCTION

State Grid Corporation of China is currently building a super-large-scale, unified, business-oriented energy Internet knowledge system and knowledge services, which can effectively break the knowledge barriers among State Grid departments, units, professions, and systems, and better complete the convergence and flow of knowledge elements within the State Grid, so that the users(people or machines) in the energy Internet can conveniently access and use authoritative, reliable and valuable power grid knowledge, and improve the overall intelligence of State Grid. However, in the process of developing the knowledge system and knowledge services of the energy Internet, the primary challenge faced by State Grid is how to efficiently and accurately transform the large-scale multi-source heterogeneous knowledge resources into a unified body of knowledge, to help form a unified energy Internet standardized knowledge and knowledge organization system among the provincial and municipal companies and grid business segments of the State Grid. This will break knowledge silos one by one and contribute to the free flow and exchange of advanced knowledge and experience in the State Grid.

This paper investigates the existing knowledge resource transformation and organization methods, and finds that there are few researches on the multi-source heterogeneous intelligent transformation and organization of the Energy Internet. Large-scale practice and application are still lacking. In the State Grid, some original knowledge resources to describe and represent the business are mainly in the Extensible Markup Language (XML) format. At the same time, other original resources have many other structured formats (such as databases, spreadsheets, etc.). To build a body of knowledge, the fusion of these heterogeneous knowledge resources and multi-source business knowledge is needed to convert the original resources into ontology descriptions.

The existing ontology description languages mainly include Resource Description Framework (RDF), RDF Schema (RDFS), and Web Ontology Language (OWL). RDF format has no definition of classes and can only describe concrete entities and the relationships between them, and it is difficult to reveal the structure and correlation of various types of knowledge contents inside and outside business domains. The degree of organization of knowledge contents is low. RDFS adds a schema layer to the data layer, defines classes, attributes and relationships to describe resources, but the relationships between classes are limited to hierarchical relationships, which cannot reveal the rich semantic information in the knowledge resources of the State Grid. The description ability of the knowledge content is weak.

The OWL format further enriches the semantic expression capability of RDF and RDFS by dividing properties into data property and object property, distinguishes the properties of the entity itself, and expands the ability to define relationships between classes. It can meet the needs of abstract fusion of knowledge content between different fields and within a specific area. Therefore, it is considered to convert heterogeneous resources into OWL format uniformly for the fusion of heterogeneous knowledge resources, and to fuse heterogeneous knowledge through the rich semantic expression of OWL.

In order to carry out the fusion of heterogeneous knowledge resources, it is necessary to use a type of structure as the intermediary for the transformation from knowledge resources to OWL. Since the original resources of the State Grid are mainly in XML format, considering the tree-like structure feature of XML language, this paper proposes to use the XSD Directed Graph as the base for knowledge resource fusion, convert other data structures into graph structures and expand the XSD Directed Graph to implement the fusion of heterogeneous knowledge resources. In order to carry out the fusion of heterogeneous knowledge, this paper uses graph neural network (GNN) to process the XSD Directed Graph to obtain richer knowledge contents embedded in XML, and introduces expert knowledge to establish multi-source knowledge association to obtain a unified and ordered body of knowledge. Through the large-scale resource transformation and knowledge organization of multiple source heterogeneity in the field of human resources of the State Grid, the effectiveness of the proposed method is initially proved, and the subsequent large-scale application will be carried out in other fields of the State Grid.

2. EXISTING CONVERSION METHODS AND THE STATE GRID CONVERSION REQUIREMENTS

2.1 Existing Methods

In the process of fusing heterogeneous knowledge resources and establishing a body of knowledge, the original structured knowledge resources need to be transformed into ontologies described using the OWL language. The existing methods mainly provide two ideas: (1) using the RML standard developed by W3C to transform structured data into RDF and further into OWL afterward; (2) transforming the original knowledge resources into OWL directly. Since the State Grid knowledge resources are mainly represented in XML, the conversions by the RML method and the method from XML to OWL are introduced separately next.

2.1.1 RML Method

The RML, a custom mapping language from structured data formats to the RDF data model, is being developed by W3C as a superset of the previous R2RML standard designed to express customized mappings from relational databases to RDF [1]. RML currently provides a general way to define a mapping to RDF for structured data formats such as CSV, XML, JSON, etc., following the same syntax as the previous R2RML, with mapping definitions represented as RDF graphs.

RML accepts a structured data format as an input data source and then maps the input to an RDF format through RML mapping, a structure containing one or more triples maps. Each triples map must include or reference a logical source that specifies the data source to be transformed, a subject map that defines the mapping rules from the XML element to the subject, and zero or more predicate maps or object maps corresponding to subject. In the end, for each XML element, one or more triple elements with the same subject are obtained.

Wang et al. [2] used the R2RML Parser tool to establish R2RML mapping documents from relational database fields to RDF, such as terms, concepts, and category classes, according to the database storage characteristics of STKOS technical vocabulary, to realize the conversion of STKOS to RDF. Wu et al. [3] adopt the R2RML language to establish different mapping schemas to store Medical Subject Headings (MeSH) to realize the transformation from MeSH to RDF. Kyriakos et al. [4] used GepTriples to generate and process extended R2RML and RML mappings and realize the transformation of structural geospatial data stored in files such as XML, GML, GeoJSON, and in spatially-enabled RDBMS into RDF maps using vocabularies like GeoSPARQL and stSPARQL.

Due to the multi-source heterogeneity characteristics of the State Grid knowledge resources, RML's processing method will encounter the following problems when dealing with these resources

(1) Low processing efficiency for large-scale, multi-source and heterogeneous resources. RML is a custom mapping. For XML documents with simpler structure, corresponding transformation rules can be defined according to XML structure and semantic information. However, the knowledge resources of the State Grid are large in scale and involve many domains, and the nested relationships in large-scale XML documents are complex, so manually defining mapping rules has a great workload and may fall into some trivial details that improve the final ontology effect less, thus decreasing the overall effectiveness. Meanwhile, resources in different domains usually require corresponding domain experts for analysis, which further increases the labor cost.
(2) Insufficient revealing of semantic information inside XML. As RML maintains backward compatibility with R2RML, the core ideas used by RML when dealing with data formats such as XML have not changed much from R2RML, which deals with relational databases, and still achieves RDF generation by establishing rules for mapping the original XML elements to subjects, predicates, and objects. This is still an abstract type-to-abstract type mapping method, where the names and values of elements are only used as instantiations of types, and the semantics embedded in them have no impact on the relationships between types, which in fact limits the extent to which the semantics of XML can be revealed by mapping rules defined.

2.1.2 XML to OWL Conversion

Existing methods mainly focus on the direct transformation from XML resources to OWL. These methods can be divided into two types according to the transformation goals, namely, targeting OWL architecture and targeting complete OWL ontology. The former has only one stage of generating the structural framework

of an OWL ontology. In contrast, the latter has two steps to generate a complete OWL ontology by transforming instances based on the former. The content and characteristics of these two transformation methods are categorized and introduced here.

1. Conversions targeting OWL architecture

This transformation type mainly focuses on extracting corresponding semantic information from XML documents and transforming them into related OWL classes and attributes. Since this type of transformation mainly requires structural information in the XML document, it is primarily based on the validation document of the XML document, that is, the corresponding XML Schema (XSD) document or Document Type Definition (DTD) document. The result of this transformation is only a structured explicit description of the original XML document and the underlying semantic information in it, forming an OWL architecture, but not a complete OWL ontology.

Specifically, the X2OWL [5] method uses XSD documents for transformation, extracts elements and attributes from XML documents by creating XML Schema Graph (XSG), generates XPath expressions for XSG tree nodes, and produces OWL documents after element deduplication. The XS2OWL [6] method uses XSD documents and allows the transformation of elements in them. In addition to supporting SPARQL to XQuery transformation, the entire transformation model can be implemented as XML Style Sheet (XSLT) format. Methods such as Janus [7], and S-Trans [8] also require the verification documents. However, The X2R-R2O method [9] does not require an existing verification document, and the entire conversion process is divided into two parts. The first part, the X2R method, directly utilizes DOM4J, a XML parsing tool, to traverse the XML document node tree, and transform the XML document into a relational database model (ER-Model) based on the rules. The second part, the R2O method, extracts the association between concepts based on the concept description rules in a relational database to form OWL ontology. The characteristics of the above methods are summarized in Table 1.

Table 1. Characteristics of transformation methods targeting OWL architecture.

Method name	Features
X2OWL	Generate instances according to user needs, solve the problem of repeated elements, and provide graphical interface tools
XS2OWL	Support SPARQL to XQuery conversion. The entire model is treated in XSLT format for other methods to use
Janus	Employ Turtle syntax
S-Trans	Define the calculation method of element similarity to realize deduplication and merging
X2R-R2O	Directly use DOM4J and a relational database as a conversion intermediary

2. Conversions targeting the complete OWL ontology

This type of transformation usually consists of two phases. The first phase targets the OWL ontology similarly to the architecture transformation. The second phase targets the expansion of the OWL ontology instance. This type of transformation is based on architecture transformation, extracts specific elements and attributes, and expands the OWL ontology, but the extraction workload is also more significant.

Specifically, the XML2OWL [10] method is based on predefined rules, using XML documents from relational databases in the first phase and transforming them to OWL according to XSD documents; in the second phase, it expands instances using XSLT extracted from XSD documents at the same time as the OWL ontology is generated in the first phase. In the XSD2OWL [11] method, in the first stage, the user manually establishes the mapping rules from XSD to OWL and describes them by RDF language; in the second stage, according to the mapping rules, the OWL instance is directly extracted and expanded from the XML document. The DTD2OWL [12] method automatically generates OWL documents in accordance with DTD in the first stage, but only the commonly used DTD elements and attributes are given in the rules, which still need to be expanded; in the second stage, XSL is exploited to convert XML documents into OWL instance. EXCO (An Efficient XML To OWL COnverter) [13] using the schema merging method, the OWLMAP [14] method using RDF language, and a set of C++ software tools [15] for transforming XML validation documents and instance documents into OWL classes and instances all fall into this category Types of. However, JXML 2 OWL [16], B2BISS [17], etc. can also convert XML to OWL without a validation document. The characteristics of the above methods are shown in Table 2.

Table 2. Characteristics of transformation methods targeting OWL ontology.

Method name	Features
XML2OWL	The method is more mature and widely used
XSD2OWL	Apply RDF language as a description of matching rules, and apply Protégé as graphical interface
DTD2OWL	Use the DTD document as the verification document
EXCO	The first stage uses pattern merging to identify internal references of different XSDs, and has an accuracy test after conversion
OWLMAP	Take RDF as the instance description and augment the instance with the OWL inference engine
C++software toolset	A set of tools work together to complete the conversion process
JXML2OWL	No need to use validation documents, use XPath to locate tags directly, and use IDs, for instance deduplication
B2BISS	There is no need to use verification documents, directly match source documents and rules, and realize heterogeneous data fusion of XML-based standards such as cXML and ebXML

2.2 The State Grid Resource Conversion Needs

The original resources to be transformed in the State Grid are characterized by large size, multiple sources, and heterogeneity. Here, this paper takes the human resource domain as an example to show the resource structure of the State Grid. The overall structure of resources is shown in Figure 1.

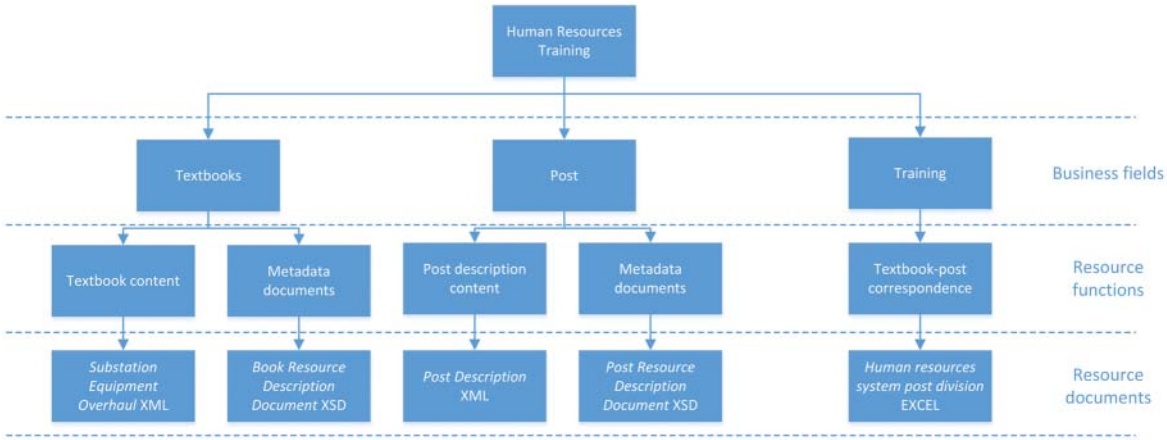

Figure 1. Diagram of the overall structure of knowledge resources in the field of human resources training.

The original resource structure of the human resources training field is mainly divided into three layers, among which the business field is the segmented business objects involved in the human resources field, consisting of three parts: the teaching materials as training resources, the positions of the training subjects and the training that describes the relationship between the training resources and the training subjects. The functional layer of resources is the functional description of specific resources in each subdivision of business objects, which consists of XML documents describing particular resources in the field and the XSD documents for all fields. The correspondence between positions and teaching materials in the training field is given by EXCEL tables. The particular knowledge resources involved in each business object are provided in the resource document layer. It can be found that in the face of such a resource structure, the above transformation methods have the following problems.

(1) Weak applicability to large-scale resources. Since the resources of the State Grid mainly exist in the form of books, which involve complex and diverse tags, the above method cannot classify different tags well, making the final ontology less structured as the scale of resources increases. Especially for books, some title information may be scattered into multiple tags, which need to be summarized and integrated to represent the organized resources better.

(2) The knowledge content contained in the XML document is less revealed. The above approach is based mainly on the tag structure of XML, which only reflects the original nesting relationships. The information about the different nesting levels and their nesting positions is not utilized, i.e., the deeper knowledge content in the XML structure tree is not well analyzed. These contents often include the common knowledge of multiple sources in a business domain by the way they represent the content organization.

(3) Weak support for the fusion of heterogeneous resources. The relationship between heterogeneous resources usually needs to be determined by domain experts based on different domains and the Grid's overall knowledge structure. One of the difficulties is that the various forms of knowledge organization increase the difficulty of establishing the relationship between heterogeneous resources.

The above methods mainly focus on the original structure. They do not bridge the knowledge representation forms of heterogeneous resources well, thus facilitating the role of expert knowledge in the fusion of heterogeneous resources.

For the sake of the above problems, this paper proposes a method of XML to OWL transformation based on XSD Directed Graph. By analyzing the characteristics of XML structure, using the directed graph as the common situation of heterogeneous knowledge representation, and deeply analyzing the graph structure and node characteristics through graph neural network, the core knowledge content contained in XML structure can be derived. More semantic information contained in the structure can be discovered. Meanwhile, the flexibility of graph structure can be used to reduce the difficulty of joining expert knowledge, and facilitate the fusion of heterogeneous knowledge resources. The expandability of the graph structure also reduces the problem of adding expert knowledge, facilitates the fusion of heterogeneous resources while obtaining the commonality of multiple sources, and enhances the capability of knowledge services.

3. MODEL DESIGN

The model is designed in the following aspects to meet the needs of the State Grid for knowledge resource transformation: (1) Extraction of core contents of large-scale resources. The scale of the State Grid resources is large, so it is difficult to extract the core information and organize and integrate them manually. (2) Multi-source resource integration. As the State Grid involves many business contents, the resources under each business area often come from different sources, but these resources have a certain homogeneity, so the key structure and components need to be extracted to form a unified domain resource ontology. Due to the large variability of different XML documents, and the State Grid mainly uses the unified DocBook format to describe resources and only provides a common XSD for DocBook format in all fields, it is needed to extract field-specific XSD documents from original XML resources for transformation. (3) Fusion of heterogeneous resources. The resource descriptions between different business domains of the State Grid are heterogeneous and eventually need to form a unified domain ontology, which requires the intervention of domain experts. In order to facilitate the incorporation of expert knowledge, it is necessary to open up the original heterogeneous knowledge representation and reduce the difficulty of establishing association relations while presenting it in a unified way so that the heterogeneous knowledge can be integrated more deeply.

Synthesizing the above aspects, it can be found that the key issue lies in the need of a new form of knowledge representation that represents large-scale multi-source heterogeneous resources using a unified form and represents the knowledge structure embedded in the resources, and carries out the transformation on this basis, to improve the efficiency and accuracy of knowledge organization and enrich semantic knowledge expression. Considering that the XML document itself has a definite tree structure, the XML-to-OWL transformation is carried out here using directed graphs to represent knowledge in XML documents. The overall conversion model is shown in Figure 2.

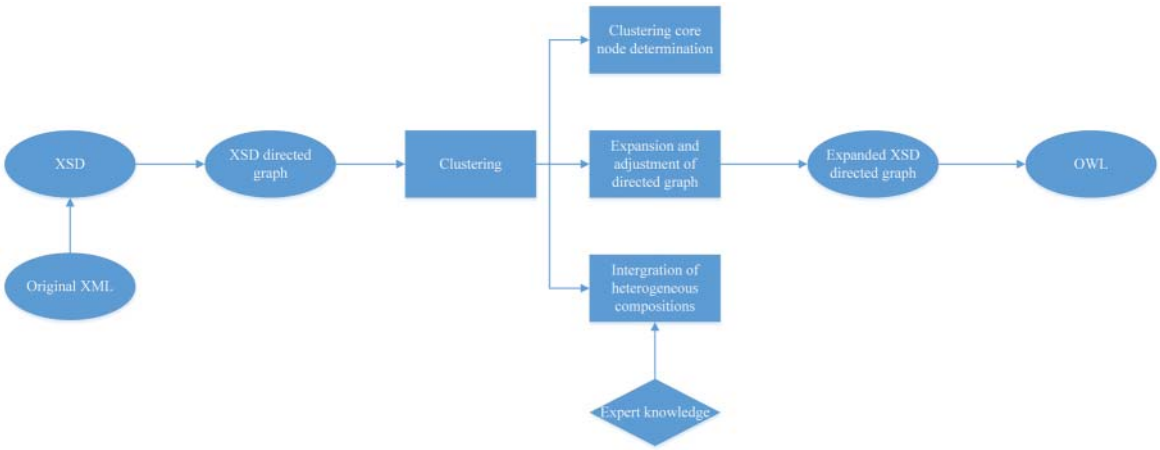

Figure 2. The overall model of XML to OWL conversion.

3.1 XSD to Directed Graph Transformation

The State Grid only provides XSD documents that are common to XML documents in various fields, and does not distinguish between fields. However, there are differences in the content of elements contained in each field's XML documents, and corresponding XSD documents need to be generated to improve the accuracy of the description of field knowledge resources. Here, the Trang API for Java [18] is applied to generate XSD documents from the original XML documents. In the specific transformation, both the complex elements and the defined element types in the XSD document are transformed into nodes. The complex elements are used as class nodes, representing elements that will be transformed into OWL classes (owl:Class) later. At the same time, all nested relations are used as edges, and the nested relations between class nodes and named descendant nodes with constraints in class nodes are used as class edges, which represent the properties that will be transformed into OWL object properties (owl: ObjectProperty) afterward. The resulting nodes and edges form a directed graph and a class directed graph representation of the original XSD structure.

3.2 Clustering of Directed Graphs

Traditional XML to OWL transformation methods defines the rules directly after obtaining the directed graph to generate an OWL document. This transformation is actually only a conversion of the XML structure. In order to reveal more information in the XML document, a clustering process is considered for the obtained directed graph, as follows.

(1) The vectorized representation of nodes is obtained using the node2vec method. The general idea of the node2vec algorithm is to obtain the vectorized representation of nodes by optimizing the neighborhood retention objective through learning to adapt to different definitions of neighborhoods by simulating a biased random walk. [19] Sampling is performed by second-order random wandering

to obtain the neighborhood set of each node, after which the vectorized representation of the node is obtained from the node, and the neighboring nodes jointly learned.

(2) Learning the embedding representation of vertices using the GraphSAGE algorithm. [20] The traditional GCN approach combines the network structure and node information to learn the embedding representation of vertices but cannot generalize to nodes that have not appeared during training. Considering the multi-source heterogeneity and scalability of the State Grid resources, there are limitations in using the GCN approach here. Therefore, the Graph SAmple and aggreGatE (GraphSAGE) method is chosen. The core idea of this method is to obtain the embedding vector of the target node by learning the aggregation function of the node's neighboring nodes, where unsupervised learning is mainly used, and the graph-based loss function is as follows:

$$J_G(\boldsymbol{z}_u) = -log\left(\sigma\left(\boldsymbol{z}_u^T \boldsymbol{z}_v\right)\right) - Q \cdot E_{v_n \sim P_n(v)} log\left(\sigma\left(-\boldsymbol{z}_u^T \boldsymbol{z}_v\right)\right)$$

Where u is the target node, v is a fixed-length random wandering node appearing near u, Pn is the probability distribution of negative sampling, and Q is the number of negative samples. The node embedding representation obtained from learning according to this objective function can be directly used in downstream clustering tasks.

(3) Clustering using the HDBSCAN algorithm. Since the XML tag structure of the knowledge resources of the State Grid may have different densities after being transformed into a directed graph, the HDBSCAN algorithm transforms the DBSCAN algorithm into hierarchical clustering by constructing a minimum spanning tree according to the transformed space of densities, after which the cluster hierarchy is established. And the hierarchy is compressed according to the size of the minimum clusters so that stable clusters can be extracted from it. HDBSCAN algorithm has strong robustness to extract the noise in the graph, and this feature also helps to extract the core content in large-scale multi-source heterogeneous resources.

3.3 Expansion and Refinement of Class Directed Graphs

3.3.1 Clustering Core Node Determination

After getting the clustering results of the directed graph, as the goal is to construct an ontology, the core nodes in each category need to be extracted as the representative of the class and set up the association relationships with the other members of the category to explore more the knowledge content embedded in the original XML document. Since the core nodes need to be inclusive of other nodes, the nesting levels of different nodes in the original XML document are mainly considered here as the basis for core node selection.

3.3.2 Expansion and Adjustment of Directed Graph

The relationship between the core nodes and other nodes in the obtained class group is not available in the original directed graph and needs to be added to the graph, where the edges connecting the class nodes are then used as a complement to the class edges. At this point, the obtained directed graph actually

includes class nodes and other nodes, and the graph needs to be adjusted for direct conversion to OWL documents. Here, the non-class nodes are first removed from the graph, after which a number of isolated nodes are created that are not connected to other nodes. These nodes have less representation of the overall XML core content, so they are also removed directly, and the resulting graph is an expanded class directed graph that can be directly transformed into an OWL document. The removed non-class nodes become the data properties of the connected class nodes (owl: DataProperty).

3.3.3 Fusion of Heterogeneous Compositions

Since the heterogeneous resources of the State Grid usually originate from different fields, the business problems within each field are highly specialized, and the business associations between fields usually exist in the form of expert experience, and there are few formed knowledge resources to illustrate this empirical knowledge. It is difficult to identify these field associations by automated processing alone, and the participation of experts' knowledge is required. To this end, this paper uses the human-in-the-loop (HITL) approach in the fusion of heterogeneous field knowledge representations, as shown in Figure 3.

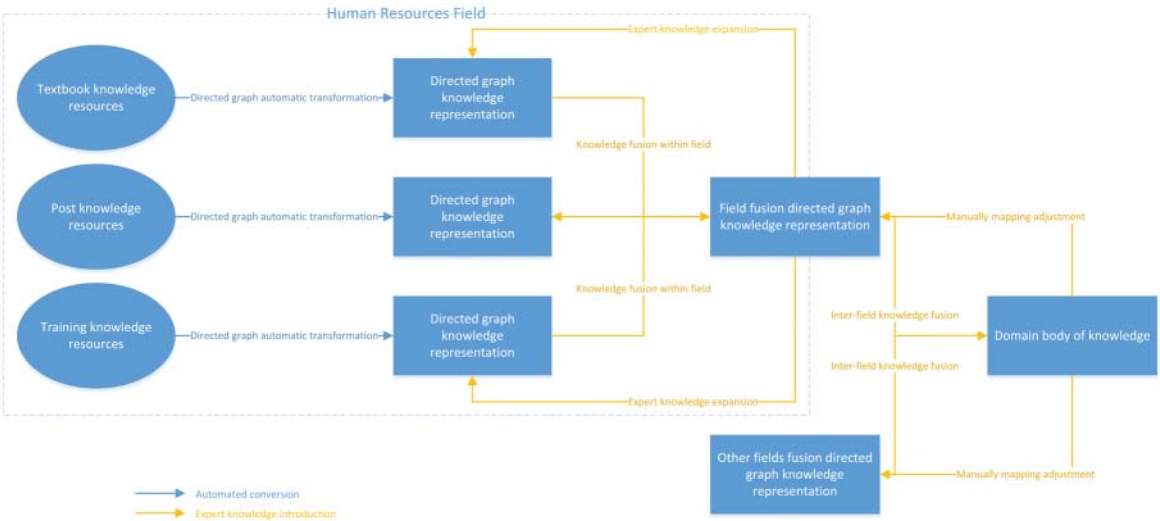

Figure 3. The process of human-in-the-loop heterogeneous knowledge fusion.

After transforming XML and other format resources in different fields into the corresponding directed graphs, expert knowledge is introduced to determine the associations between different nodes in the graph, specify the nodes that need to be associated, and the specific types and directions of the association between nodes. At the same time, in this step, experts can also adjust and expand the directed graphs in different fields so that the results are more in line with actual business needs. For the knowledge fusion between areas, a similar method is used to establish and adjust the mapping to improve the organization of knowledge content further when transforming from directed graphs to OWL and establishing a domain body of knowledge.

3.4 Directed Graph to OWL Transformation

After getting the fused and expanded class directed graph, the transformation of the class directed graph to OWL can be performed directly. It should be noted that the class directed graph here is only to distinguish data attributes from ontology classes, and all nodes and edge relationships in the whole directed graph need to be transformed to get a complete OWL representation of the XSD document. The transformation rules are as Table 3 shown here.

Table 3. Transformation rules for the directed graph to OWL.

Directed graph elements	OWL elements
Class nodes	owl: Class
Class edges (edges between class nodes)	owl: ObjectProperty domain: source class node range: target class node
Non-class nodes	owl: DataProperty domain: corresponding class node range: corresponds to the type of the non-class node

4. MODEL EXPERIMENT

4.1 Experimental Data

This paper uses knowledge resources from the State Grid Human Resource Training area for the transformation experiments. The specific resources used in each section are presented separately, according to the structure in 2.3.

(1) Textbook resources. This paper uses the training material "Substation Equipment Overhaul. 330kV and above" (hereinafter referred to as the "training material") in the field of human resources of the State Grid. The training material is described in DocBook format, which is a text-centered XML standard with XSD documents. Part of the training materials is shown in Figure 4, and part of the XSD document is shown in Figure 5.

```
<section>
        <title>模块1  SN10-12Ⅱ（Ⅲ）型少油断路器 大小修（Z14E2001Ⅰ）</title>
        <section>
                <title>【模块描述】</title>
                <para>本模块包含SN10-12Ⅱ（Ⅲ）型少油断路器大小修的主要作业内容及质量标准。通过结构分析、图例展示、要点归纳、操作技能训练，掌握SN10-12Ⅱ（Ⅲ）型
        </section>
        <section>
                <title>【模块内容】</title>
                <section>
                        <title>一、SN10-12Ⅱ（Ⅲ）型少油断路器的结构</title>
                        <para>SN10-12Ⅱ（Ⅲ）型少油断器采用纵、横吹灭弧原理，利用绝缘油作为灭弧介质，因此用油量较少。该类断路器主要配用CD10系列直流电磁操动机构
                        <section>
                                <title>1.SN10-12Ⅱ（Ⅲ）型少油断路器的结构</title>
                                <para>SN10-12Ⅱ（Ⅲ）型少油断路器由本体、框架、传动系统、操动机构等部分组成。SN10-12Ⅱ型断路器结构剖面如图Z14E2001Ⅰ-1所示
                        </section>
                        <section>
                                <title>2.CD10电磁操动机构</title>
                                <para>CD10电磁操动机构主要由分、合闸电磁机构，四连杆机构，脱扣器，辅助开关等部分组成，其结构如图Z14E2001Ⅰ-3所示。</para>
                        </section>
                </section>
```

Figure 4. Example of part of the training materials.

```
<xs:element name="original" minOccurs="0">
        <xs:annotation>
                <xs:documentation>原版原版书名、原版作者（姓名、简介、著作方式）原版书号、原版版权声明
                </xs:documentation>
        </xs:annotation>
        <xs:complexType>
                <xs:choice minOccurs="0" maxOccurs="1">
                        <xs:element ref="docbook:title">
                                <xs:annotation>
                                        <xs:documentation>原版书名</xs:documentation>
                                </xs:annotation>
                        </xs:element>
                        <xs:element ref="docbook:authorgroup">
                                <xs:annotation>
                                        <xs:documentation>原版作者组</xs:documentation>
                                </xs:annotation>
                        </xs:element>
                        <xs:element ref="docbook:biblioid">
                                <xs:annotation>
                                        <xs:documentation>原版书号</xs:documentation>
                                </xs:annotation>
                        </xs:element>
                        <xs:element ref="docbook:copyright">
                                <xs:annotation>
                                        <xs:documentation>原版版权声明</xs:documentation>
                                </xs:annotation>
                        </xs:element>
                </xs:choice>
        </xs:complexType>
</xs:element>
```

Figure 5. Example of part of XSD document.

Since the original XSD document of the DocBook document of the State Grid does not describe the various nested relationships involved in the original resource of on e specific field well and has some problems, a new XSD document is generated here by using the Trang toolkit based on the original XSD document. The generated XSD document is partially shown in Figure 6.

```
<xs:element name="book">
 <xs:complexType>
  <xs:sequence>
   <xs:element ref="docbook:title"/>
   <xs:element ref="docbook:etitle"/>
   <xs:element ref="docbook:info"/>
   <xs:element maxOccurs="unbounded" ref="docbook:attachment"/>
   <xs:element maxOccurs="unbounded" ref="docbook:formerAidText"/>
   <xs:element ref="docbook:toc"/>
   <xs:element maxOccurs="unbounded" ref="docbook:part"/>
  </xs:sequence>
  <xs:attribute name="version" use="required" type="xs:decimal"/>
 </xs:complexType>
</xs:element>
```

Figure 6. XSD document generated from the original resource

(2) Posts and training resources. State Grid's original post resources are described in XML format and have corresponding XSD documents. However, since the hierarchical structure of posts has been given in the training resources, the ontology can be constructed directly according to it so as to meet the business needs. Therefore, the posts and training resources use the corresponding EXCEL table of the State Grid human resources training posts-teaching materials, as shown in Table 4.

Table 4. Example of content of Part of the EXCEL form for the division of labor in the human resources system.

Post division (11 categories)	Subpost	State Grid 54 training materials	ISBN
Transmission Line Operation and Inspection	Transmission line operation and inspection (330kV and above)	State Grid Co., Ltd. professional training materials for skilled personnel Transmission line operation and Inspection (330kV and above) (upper and lower volumes)	978-7-5198-4466-0
	Transmission line operation and inspection (220kV and below)	State Grid Co., Ltd. professional training materials for skilled personnel Transmission line operation and inspection (220kV and below) (upper and lower volumes)	978-7-5198-4451-6
	Transmission cable operation and inspection	State Grid Limited professional training materials for skilled personnel transmission cable operation and inspection	978-7-5198-4490-5

4.2 XSD to Directed Graph Transformation

In the generated XSD document, all elements with actual names (attribute "name") are first transformed into nodes, where complex elements or data types are used as class nodes and simple elements are used as data attribute nodes, and directed edges are created between all elements with nested relationships, pointing from higher-level elements to lower-level elements.

After that, since the complex elements in nested elements are identified by two kinds of attributes, maxOccurs and minOccurs, the elements with these two attributes and the elements in which these elements are nested can be located. Thereafter, the edges between the corresponding nodes of these elements are used as class edges, pointing from the higher-level elements to the lower-level elements, named "has" + the name of the subordinate node. It should be noted that, according to the book structure, the nested relationship where the <choice> element is located is not used as a class edge because this element is only used to declare that the higher-level element contains specific textual content, which is less meaningful to establish as an OWL class. The obtained directed graph is shown in Figure 7.

4.3 Clustering of Directed Graph

(1) Using nodevec2 to get the vectorized representation of nodes. Here the node2vec toolkit from Stanford University [21] is used directly to transform each node into a vector of 64 dimensions as a feature representation of the node. The result of vectorization is shown in Table 5.

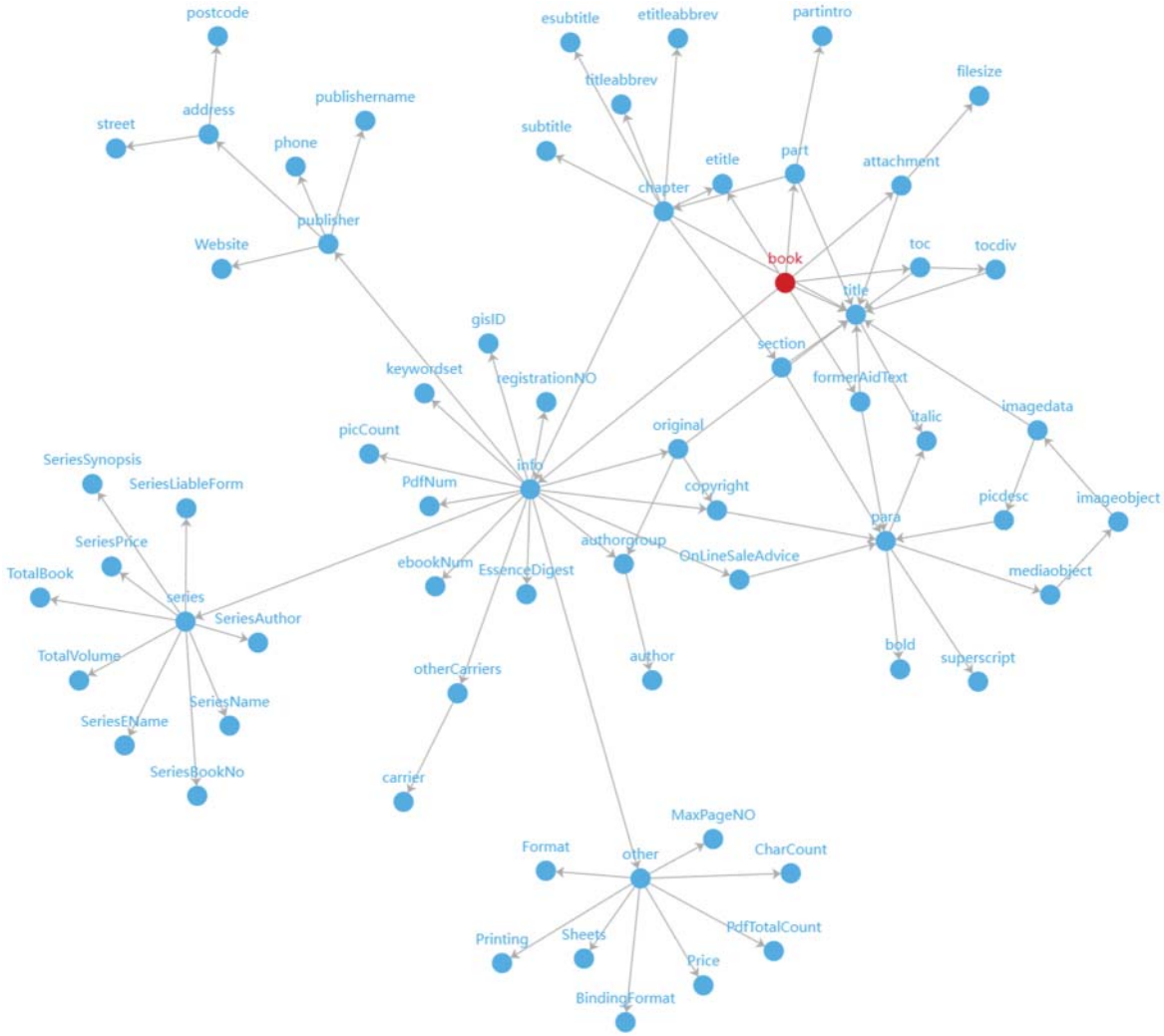

Figure 7. The whole directed graph obtained by transforming the XSD file (where the number of edges connected by the node book is 7).

Table 5. Partial results of node vectorization.

name	id	0	1	2	...	62	63
book	0	0.164828	-0.286466	0.266627	...	-0.240514	0.080081
title	1	0.234903	-0.285622	0.412794	...	-0.227231	0.094118
etitle	2	0.154292	-0.293889	0.461764	...	-0.143793	0.239639

(2) The embedding representation of the nodes is obtained using the GraphSAGE algorithm. Here, the directed graph is first represented using the StellarGraph framework for the convenience of subsequent transformations. [22] This framework is mainly used to solve problems related to graph structure data, such as graph node clustering. It is built on top of TensorFlow, which can directly implement many graph neural network algorithms. After the transformation, 6 nodes and 43 edges are obtained, nodes are characterized as 65-dimensional vectors (64-dimensional vectorization results with one dimension for id), and edges are pointed from source to target. After this, the model is trained to fit the data by using the unsupervised GraphSAGE method, setting batch_size = 20 and epochs = 60, and the last round of model binary accuracy is 0.7810. This model is then used to learn node embeddings, with each node represented as a 15-dimensional vector. The embedding results are shown in Table 6.

Table 6. Partial results of node embedding representation.

name	0	1	2	...	13	14
book	0.0269398	-0.3902382	0.22185883	...	-0.15399809	0.09135856
title	0.03119678	-0.45190296	0.05691655	...	0.2662341	-0.11103718
etitle	0.13692315	-0.11827449	-0.0860365	...	-0.09366022	0.47177184

(3) Clustering using HDBSCAN. After obtaining the node embedding representation, the clustering is performed directly using the HDBSCAN method to get eight total categories. One of them being the noise category. The results of each category are listed in Table 7.

Table 7. Partial clustering results.

Category tags	Class Nodes
-1 (Noise category)	attachment, filesize, otherCarries, …
0	publisher, publishername, Website, …
1	address, stress, postcode
2	mediaobject, imageobject, imagedata, …
3	info, copyright, original, biblioid
4	authorgroup, author, editor, …
5	book, title, para, section, …
6	part, partintro, chapter, …

It can be found that the noise category, which is an attachment to some resources or an overall external property description, basically does not reflect the knowledge content and cannot be reflected in the ontology. The remaining categories have some practical significance, such as category 0 for publication information, 1 for address information, 2 for multimedia element information, etc. It is worth noting that the classes in this category are actually an expression of multi-modal knowledge content in the original knowledge resources. For example, image object can be used to present the image information contained in the original knowledge resource, while image data is a specific description for different pictures, such as by providing picture metadata, etc. These multi-modal knowledge resources can be combined with textual knowledge resources, which is more conducive to establishing a more efficient body of knowledge.

Category 5 is actually a rich text content class, i.e., the elements which will contain richer text content internally and the semantic information which can be further mined; the two aspects of part and chapter in category 6 only play the role of title in the original book resources and do not have a large amount of actual content, so this class can be regarded as reflecting the structure of the knowledge resources. The above clustering results are verified by manually using the original XSD documents, and the results are found to be practically meaningful, reflecting the knowledge content that cannot be directly reflected by the XSD tag structure.

4.4 Expansion and Refinement of Class Directed Graph

The above clustering results are complementary to the edge relationships in the original directed graph, and in order to add them to the original graph, it is necessary to determine the basis for establishing edges, i.e., to derive class-cluster core nodes and establish relationships with other nodes in the same class-cluster. Considering that the element nesting level reflects the element's hierarchical position in the XML document, the average nesting level of the element is used here as the basis for selecting the core nodes of each class cluster. The number of ancestor elements before the position of each element is the nesting level of that element, and the average nesting level is calculated as follows:

$$\text{Average nesting level} = \frac{\sum nesting\ level}{num\ of\ occurrences}$$

The resulting nesting levels can reflect the more common nesting depth of elements and, to some extent, attenuate the influence of element occurrence order and element declarations on the determination of element level positions. The final core nodes of each category are shown in Table 8.

Table 8. Core Nodes by Category.

Category tags	Core Nodes
-1	attachment
0	publisher
1	address
2	mediaobject
3	info
4	authorgroup
5	book
6	part

All of the above core nodes satisfy the requirement of using the outermost element as the core element (node) for elements that appear in the same nested structure and also match the actual meaning of each category, such as core node publisher of category 0, core node address of category 1, etc.

After this, the core node of each class is the source of the directed edge, pointing to other target nodes in the class, and the edge is named "hasMember" + node name as the edge to be transformed for the object attributes. At this point, since the clustering uses the data of the whole directed graph, which will involve

the nodes that should actually be transformed into data attributes, i.e., simple elements, the existing directed graph is filtered according to the previously derived class nodes, and finally, the edges between the class nodes are the actually transformed edges of the object attributes (class edges). It should be noted that the purpose of this step is to facilitate the transformation, and all nodes in the final whole directed graph should be transformed into the corresponding OWL elements. The expanded whole directed graph is shown in Figure 8, and the directed graph containing only class nodes is shown in Figure 9.

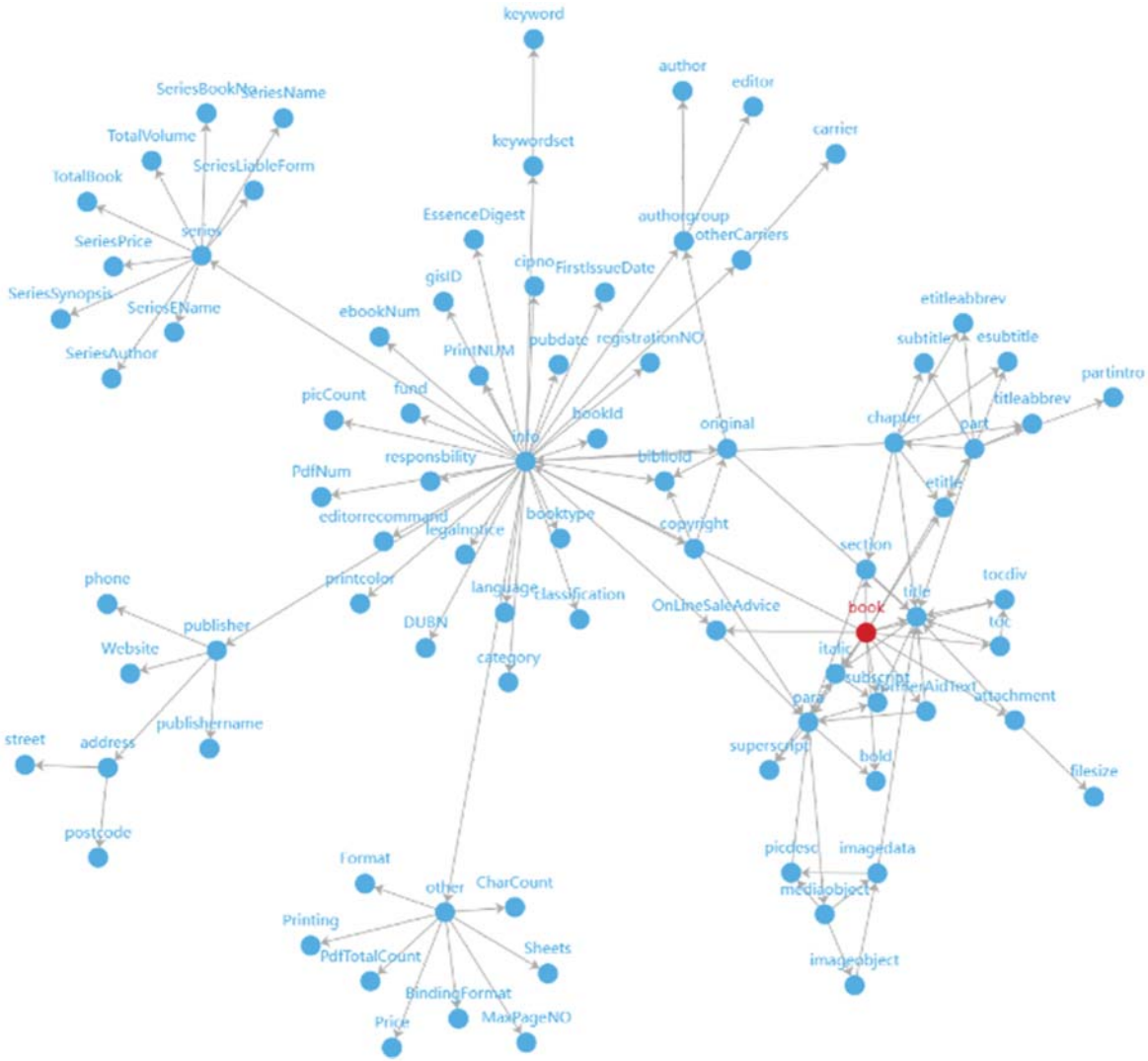

Figure 8. The expanded whole directed graph (where the number of edges connected by the node book is 15).

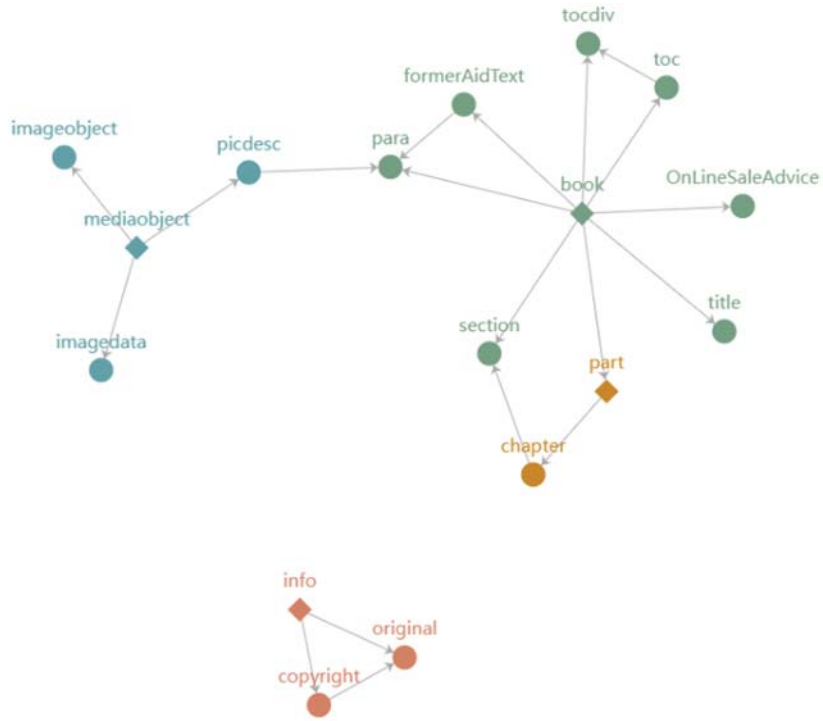

Figure 9. Directed graph with only class nodes preserved (where the number of edges connected by the node book is 8, colors are used to identify classes and the diamond symbols are used to represent the core nodes of each class).

At this point, the organization of the original knowledge resources from the internal semantic level is formed by combining the division of categories and the determination of core nodes in each category. The core nodes in each category actually play the role of identifying the semantic topic of each category, and the category system thus obtained is a kind of extraction and organization of the knowledge content in the original resources. The overall body of knowledge of the business domain can be obtained by fusing the knowledge contents of heterogeneous knowledge resources.

4.5 Fusion of Heterogeneous Graphs and Conversion of Directed Graphs to OWL

The heterogeneous diagram here mainly refers to the fusion between the directed graph obtained from the post-textbook EXCEL table transformation and the directed graph obtained from the above training materials. It is relatively simple to transform the post-textbook EXCEL table into a directed graph, with each column as a node and all columns except the ISBN number column as class nodes. The hierarchical relationship between the post columns (post and sub-post) is transformed into a "has" type of edge, the relationship between the book and the ISBN number is transformed into a non-class edge, and a class edge, "useBook", is established between the sub-post and the book from the sub-post to the book. It should be noted that in this step, more refined correlations can be established between nodes based on actual business

requirements in combination with expert knowledge, and the experts' empirical knowledge about the correlations between business domains can be transformed into the correlations of knowledge nodes to improve the degree of fusion of knowledge contents between different domains.

At this point, a complete human resources domain directed graph can be obtained, and the human resources body of knowledge in the form of domain ontology can be obtained by directly converting the directed graph to OWL according to the conversion rules defined in 3.5. The ontology of the human resources field body of knowledge system obtained is shown in Figure 10, part of the converted OWL document is shown in Figure 11.

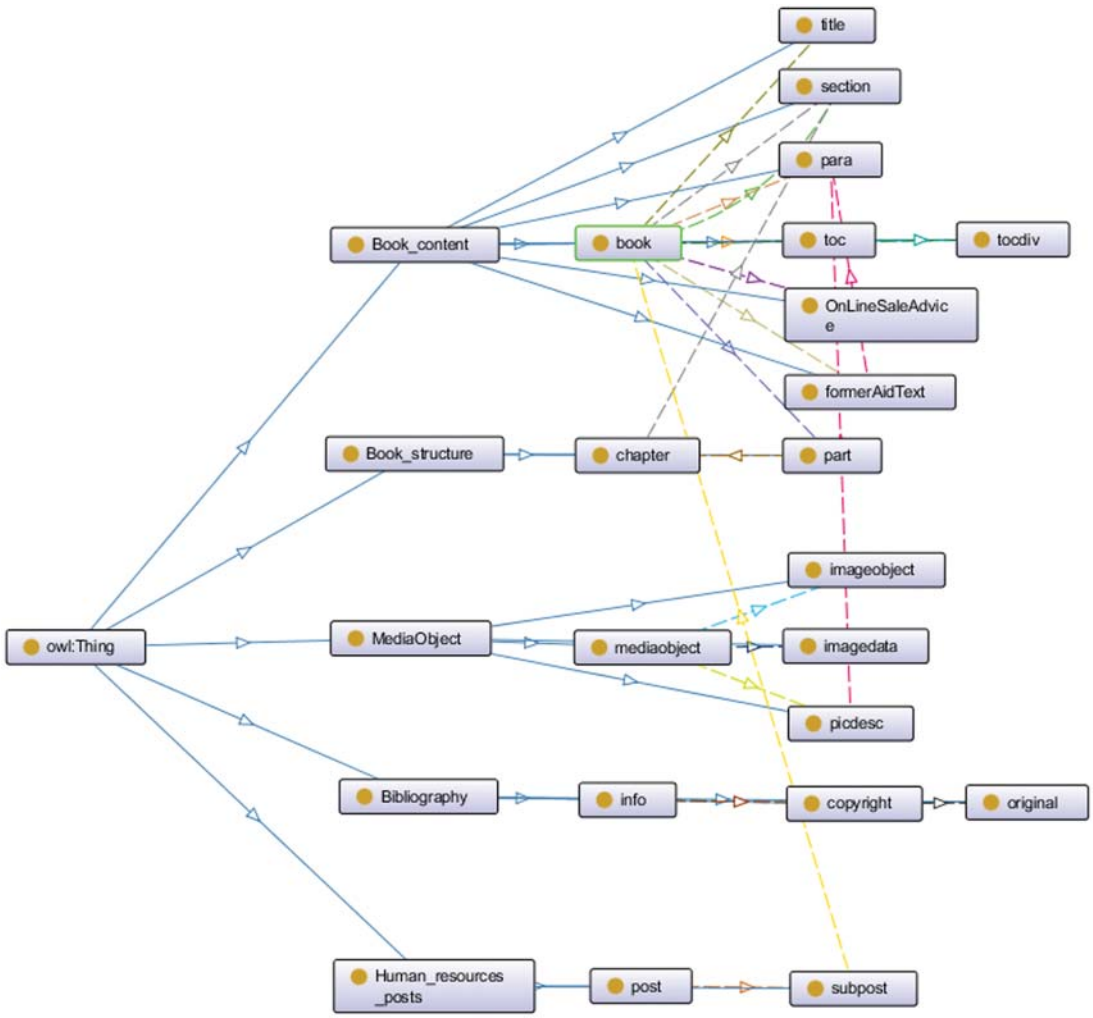

Figure 10. The body of knowledge in the field of human resources obtained from the fusion Where the second column on the left is the name of each group, the third column is the core class of each group, and the fourth column and after are other classes. (The class book has ten object properties).

```
### http://www.semanticweb.org/dell/ontologies/2021/7/untitled-ontology-6#hasChapter
hr:hasChapter rdf:type owl:ObjectProperty ;
              rdfs:domain hr:Part ;
              rdfs:range hr:Chapter .


### http://www.semanticweb.org/dell/ontologies/2021/7/untitled-ontology-6#hasPart
hr:hasPart rdf:type owl:ObjectProperty ;
           rdfs:domain hr:Book ;
           rdfs:range hr:Part .


### http://www.semanticweb.org/dell/ontologies/2021/7/untitled-ontology-6#hasSection
hr:hasSection rdf:type owl:ObjectProperty ;
              rdfs:domain hr:Chapter ;
              rdfs:range hr:Section .
```

Figure 11. Example of a part of converted OWL document.

5. SUMMARY AND DISCUSSION

Aiming at the needs of the State Grid for energy Internet knowledge organization and the basic form of existing resources, this paper proposes an XML-to-OWL transformation method based on the directed graphs. This method has the following characteristics.

(1) The XML document has a tree structure, i.e., a directed acyclic graph. The use of graph structures facilitates the mining of richer semantic information contained within the original structure of XML. This content is difficult to extract manually from large-scale documents, but the content organization of the document itself is actually a representation of this semantic information. At the same time, the graph structure makes the original resource content more clearly presented, and different nodes can be directly connected, which makes it easier to show the relationship between the knowledge content within the resource. In addition, the graph structure only consists of nodes and edges, and there is no nested structure, which is conducive to the transformation between different formats.

(2) Applicable to the transformation of large-scale multi-source XML documents. The core problem of large-scale multi-source XML document transformation is the high labor cost. This method transforms the original XSD document into the form of a directed graph, and by analyzing the graph structure and node features, the core content of the large-scale XML document is derived semi-automatically, and the corresponding XML document can be generated directly, which reduces the manual workload by using the XSD structure itself.

(3) Revealing richer knowledge about XML documents. The traditional XML-to-OWL method mostly uses XML hierarchical structure as OWL semantic information, but this method uses a neural network to analyze the XML structure deeply and combine the semantic information of the XML tags themselves to cluster the directed graphs and derive additional knowledge content embedded in the XML structure, which enriches the semantic representation of the ontology and reveals the knowledge in the original resource to a relatively higher.

(4) Strong flexibility and extensibility. The various methods of generating nodes and edges in the graph structure and the richness of the knowledge content are expressed to make it possible to flexibly expand the graph after conversion into a directed graph according to the actual business needs, which facilitates the integration of heterogeneous resources. In this paper, heterogeneous format of the EXCEL table is taken as an example to demonstrate the convenience of converting other structures into graph structures as a way to unify heterogeneous knowledge resources. At the same time, the presentation form of graph structure is also more convenient for experts' knowledge to be added, which simplifies the process of establishing relationships between different business areas of the State Grid to form the body of knowledge, and only requires experts to establish edge relationships between corresponding nodes according to business knowledge, providing a way for heterogeneous knowledge fusion.

Accordingly, this directed graph-based transformation approach has certain shortcomings. The existing transformation is mainly based on XSD documents, and the rich knowledge content in the original XML files is less involved. By using nodes as text content carriers, the text content in the corresponding elements of nodes can be added to the graph clustering. The clustering results contain more semantic information and reveal deeper knowledge content. At the same time, the fusion of existing heterogeneous knowledge mainly relies on expert knowledge, so selecting representative text contents according to the descriptions of text contents of different domains and extracting association relations between domains automatically by means of natural language processing as a supplement to expert knowledge to improve the intelligence of heterogeneous knowledge fusion are considered. Furthermore, the RML method can be integrated and extended to enhance the capability of fusing different structured data formats (such as JSON, etc.) by converting them to OWL.

AUTHOR CONTRIBUTION STATEMENT

Y. Wang (imwyx@pku.edu.cn) is responsible for the specific paper writing, model design, and experiments. L. Luo (liqun@pku.edu.cn) is responsible for the overall framework conception and model design of the thesis. G. Li (ligj@pku.edu.cn) is responsible for the overall framework of the thesis.

SUPPORTING INFORMATION

S1 Human Resources Field Training Materials of the State Grid (Textbook).xml
S1 Title: Substation Equipment Overhaul. 330kV and above
S2 Post-Textbook correspondence table.xlsx
S2 Title: Explanation of the correspondence between posts and training materials

REFERENCES

[1] RDF Mapping Language (RML) (2020). Available at: https://rml.io/specs/rml/#conformance. Accessed 1 May 2022

[2] Wang, Y., Wu, S.: Converting STKOS metathesaurus to RDF triples with R2RML. Data Analysis and Knowledge Discovery 2(12), 89–97 (2018)

[3] Wu, S., et al.: Conversion of Medical Subject Headings(MeSH) to RDF based on R2RML. Journal of Medical Informatics 40(05), 65–71 (2019)

[4] Kyzirakos, K., et al.: GeoTriples: Transforming geospatial data into RDF graphs using R2RML and RML mappings. Journal of Web Semantics 52, 16–32 (2018)

[5] Ghawi, R., Cullot, N.: Building ontologies from XML data sources. In: International Workshop on Database and Expert Systems Application, pp. 480–484 (2009)

[6] Tsinaraki, C., Christodoulakis, S.: Interoperability of XML schema applications with OWL domain. In: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", pp. 850–869 (2007)

[7] Bedini, I., et al.: Transforming XML schema to OWL using patterns. In: IEEE International Conference on Semantic Computing, pp. 102–109 (2011)

[8] Pham, T., Lee, S.: S-Trans: Semantic transformation of XML healthcare data into an OWL ontology. Knowledge-Based Systems. 35, 349–356 (2012)

[9] Li, W.: Research on XML to OWL document generation method. Master thesis, China University of Petroleum (2008). Available at: https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CMFD&dbname=CMFD2009&filename=2008199985.nh&uniplatform=NZKPT&v=twZzzfXXjaG7A7OLxq84ADLov0Z0z4yztOfUJd7IPWSiKmBtze1n1Z0WdjgrPmpM. Accessed 1 May 2022

[10] Bohring, H., Auer, S.: Mapping XML to OWL ontologies. Marktplatz Internet: Von e-Learning bis e-Payment, 13. Leipziger Informatik-Tage (LIT 2005), pp. 147–156 (2015)

[11] Cruz, C., Nicolle, N.: Ontology enrichment and automatic population from XML data. In: International VLDB Workshop on Ontology-Based Techniques for Databases in Information Systems and Knowledge Systems, ODBIS, pp. 17–20 (2008)

[12] Pham, T., Lee, Y., Lee, S.: DTD2OWL: Automatic transforming XML documents into OWL Ontology. In: International Conference on Interaction Sciences: Information Technology, Culture, and Human, pp. 125–131 (2009)

[13] Lacoste, D., Sawant, K., Roy, S.: An efficient XML to OWL converter. In: India Software Engineering Conference, pp. 145–154 (2011)

[14] Ferdinand, M., Zirpins, C., Trastour, D.: Lifting XML schema to OWL. In: International Conference on Web Engineering (ICWE), pp. 354–358 (2004)

[15] Kramer, T., et al.: Software Tools for XML to OWL Translation (2015). Available at: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=915506. Accessed 29 Oct 2021

[16] Rodrigues, T., Rosa, P., Cardoso, J.: Mapping XML to existing OWL ontologies. In: International Conference on WWW/ Internet, pp. 72–77 (2006)

[17] Cardoso, J., Bussler, C.: Mapping between heterogeneous XML and OWL transaction representations in B2B integration. Data & Knowledge Engineering 70, 1046–1069 (2011)

[18] Trang Multi-format schema converter based on RELAXNG (2008). Available at: https://relaxng.org/jclark/trang.html. Accessed 25 Feb 2022.

[19] Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855-864 (2016)

[20] Hamilton, W., et al.: Inductive representation learning on large graphs. Advances in Neural Information Processing Systems 30 (2017)
[21] node2vec: Scalable Feature Learning for Networks (2016). Available at: https://snap.stanford.edu/node2vec/. Accessed 26 Feb 2022
[22] StellarGraph—Machine Learning on Graphs (2020). Available at: https://www.stellargraph.io/. Accesses 26 Feb 2022

AUTHOR BIOGRAPHY

Wang Yuxuan is now studying in the Department of Information Management of Peking University, majoring in big data management and application. At present, his main research interests are ontology engineering and knowledge management. He has participated in provincial projects and won several awards in scientific research competitions.
ORCID: 0000-0003-2441-8850

Luo Liqun is a researcher at the Department of Information Management of Peking University, Jiangsu Industry Professor. Graduated from Nanjing University and Imperial College London, he has been engaged in research on intelligent decision-making and knowledge computing. He presided over 8 national and provincial-level projects, owned 4 invention patents, published 22 SCI/CSSCI papers, and obtained 18 software copyrights.
ORCID: 0000-0002-4860-7050

Li Guangjian, Ph.D., Professor and PhD Supervisor, Department of Information Management, Peking University, expert committee member of the National Public Cultural Service System Construction, member of the National Technical Committee 4 on Information and Document Standardization Administration of China (SAC/TC4),member of the computer network service system expert committee of the National Science and Technology Library (NSTL), and member of the expert advisory committee of the National Engineering Center of Science and Technology Information; Executive member of China Society for Scientific and Technical Information (CSSTI), council member of Chinese Information Society of social sciences (CISSS) and council member of Library Society of China; Editorial board members of many academic journals such as Journal of the China Society for Scientific and Technical Information, Data Analysis and Knowledge Discovery, and Digital Library Forum.
ORCID: 0000-0002-2897-6246