2nd International Conference on Automation, Mechanical Control and Computational Engineering (AMCCE 2017)

Research on Domain Ontology Construction Based on Thesaurus of Geographical Science

Duan Linbo[1, a],Qin Ping[2, b],Qian Lingfei[3, c],Xie Ting[4, d]

[1]Library,Nanjing University of Aeronautics and Astronautics,Nanjing,210000,China

[2]Library,Nanjing University of Aeronautics and Astronautics,Nanjing,210000,China

[3]Library,Nanjing University of Aeronautics and Astronautics,Nanjing,210000,China

[4]Library,Nanjing University of Aeronautics and Astronautics,Nanjing,210000,China

[a]email:linboduan@nuaa.edu.cn,[b]email:qplib@nuaa.edu.cn,

[c]email:qianlingfei@nuaa.edu.cn,[d]email:chilli6279@163.com

Keywords: Geography Science; Thesaurus; Domain Ontology; OWL

Abstract. Compared with the traditional domain ontology construction method and the lifecycle thinking of software engineering, the main steps of constructing the domain ontology are to construct purpose, overall design, detailed design, ontology consistency test and ontology evaluation. Using "Thesaurus for Geographic Sciences" as data source and OWL as description language to construct domain ontology, the focus of the process is to determine the concept of the thesaurus, the level of relations between concepts and representations, relationships and so on, realizing conversion of the thesaurus to the domain ontology in the final. By constructing the domain ontology of geographic science, the semantic retrieval based on ontology is realized, solving the problem of traditional information retrieval based on grammatical matching and improving the quality of information retrieval.

1.Introduction

Ontology, which originally belonged to the concept of the metaphysics branch of philosophical domain, was described as a systematic description of objective things, which was later introduced into the field of information science, and there is no unified concept. In the early 1990s, the concept of ontology was widely introduced into computers, especially artificial intelligence, knowledge engineering and other fields, as the basis of knowledge expression. Tom Gruber in Stanford University believes that ontology is a clear specification of the conceptual model, while Studer et al.in the University of Karlsruhe optimize the concept of ontology, concluding that ontology is a clear formal specification of shared conceptual models. At the beginning of this century, China began to study ontology in the field of library and information. At the same time, library and information is also an important place for ontology application, and then ontology got a wide range of applications in semantic web, intelligent information retrieval, natural language processing, knowledge management, heterogeneous information integration and so on. Traditional ontology construction is mainly based on manual construction of domain experts or dictionaries, which is time-consuming and low-automatic. Therefore, it is the focus of this paper to construct ontology with efficient methods and steps.

2.Research Status of Ontology Construction Based on Thesauri

The ontology, a visualization of concept, standardized the concept of abstraction in the real world, which makes the relationship between concept and concept, concept and object, concept and object more clear, and reduces the ambiguity to achieve knowledge reuse and sharing purpose. After the introduction of ontology into the field of computer science, many research institutions at home and abroad have carried out research on the related problems of domain ontology construction based on thesauri. At present, a large number of relatively mature ontology have been developed. Because of

Copyright © 2017, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

ontology construction methods, application fields and so on, these ontology have some differences in scale and complexity. Among them, the main studies abroad include: Agriculture ontology Service Project Group (AOS) established by the FAO, which uses RDFS (RDF Schema) to transform the Agrovoc thesaurus into agricultural ontology; J.Qin and S.Paling of Syracuse University in the United States specifically explored the principles and principles of the conversion of controlled vocabularies in GSM into ontology; SWAD-Europe established a thesaurus research group, through a variety of thesaurus classification research, based on the RDFS language, using thesaurus to describe the ontology organizational system SKOS.

The domestic scholars have also carried on the related research to the ontology construction theory. Liu Wei et al. analyze the research on ontology in the field of library and information science from 2001 to 2009, and divides it into three stages: 2001-2004 is the sprouting stage, 2005-2007 is the rapid growth period, and 2007 is the relatively mature period, used CSSCI as the data source via bibliometric analysis. Chinese Academy of Agricultural Sciences Li Guangda, Chang Chun [6] gives ontology construction methods and the thesaurus ontology transformation of the specific solutions. Professor Ding Shengchun of Nanjing University of Science and Technology[7] pointed out three kinds of treatment schemes in the semiautomatic construction of domain ontology based on space thesaurus to further improve the retrieval efficiency. Deng Zhihong, Tang Shiwei[8] put forward a conceptual model based on book classification system and subject vocabulary in the research of intelligent navigation system of digital library project in Peking University. Shenzhen University Library Zeng Xinhong[9] proposed specific programs based on the OWL representing "China Classification Thesaurus"; Southeast university Zhang Xiang et al.[10] used falcons semantic search engine and hierarchical clustering algorithm to construct a semantic web-based virtual ontology with high semantics.

3.Domain Ontology Construction Based on Geography Science

3.1Thesaurus analysis

Thesauri is also called the subject or dictionary, which is composed of descriptive, non-descriptive, inter-semantic relations and descriptive related words. It can not only reflect the semantic related concepts of a subject area, but also improve as the demand continuing. The thesaurus is mainly used for the automatic control of indexing and indexing, and it is an important way to realize multilingual search and intelligent concept retrieval. The thesaurus serves as the preferred word for a concept rather than a descriptive word as a search entry. The semantic relations between the words can be defined by the six reference characters "Y, D, S, F, Z, C ". Thesaurus is composed of years of efforts of experts and scholars from all disciplines, so it is authoritative, scientific, professional and normative after years of effort.

Through the analysis of the thesauri, it is concluded that there are many similarities between thesauri and ontology, and it is possible to construct semantic domain ontology based on thesauri. The Similarities between thesauri and domain ontology are as follow:

(1)From the perspective of the content of information, both are the concept of a particular discipline to organize a particular subject area of knowledge organization relations, which provides a basis for knowledge sharing.

(2)From the point of view of information organization, both have a hierarchical structure with clear semantics, which can efficiently and accurately map classes or concepts and relationships between words.

(3)From the perspective of information description language, both contain artificial language, can rely on the concept of system rules to express highly complex knowledge.

(4)From the application point of view, with the development needs of various disciplines, can continue to expand and maintain.

Based on the comparison of traditional domain ontology construction methods, this paper draws lessons from software engineering lifecycle thinking and uses Geographic Science Thesaurus as data source, puts forward the construction of ontology based on the thesaurus, the key task is to solve the problem of determining the relationship between concepts and concepts in the thesauri, the representation and refinement of the relationship,using OWL description language[13] to achieve the thesaurus to the domain ontology conversion. This method can not only reduce the workload of domain ontology construction, but also ensure the accuracy of relations between classes in ontology, making the final construction of the ontology more scientific. "Geographic Science Thesaurus" is divided into the main table containing a total of 6721 descriptors, schedule including 6248 formal words and index containing 473 informal words.

3.2The Principle of Ontology Construction of Geography

Based on the thesaurus to construct the ontology of geographic science field, it is necessary to combine the actual situation of Geography Science Thesaurus, according to certain requirements and rules to guide the whole construction process.Based on the structural analysis of the thesaurus and the basic requirements of the design ontology proposed by Tom Gruber [1], this paper proposes five principles for constructing domain ontology based on the Geographic Science Thesaurus:

(1)Integrity. The relationship between the thesauri in the geography science thesaurus is represented by the reference character "with, generation, genus, division, family, and participation", and also contains the English name and category number. As the basis of resource sharing and the importance and particularity of geoscience in practical application, we should, as far as possible, reflect the information in the thesaurus to the ontology so that make the content of the ontology and structure more complete.

(2)Content consistency. Due to the particularity of the domain, the determination of the inter-word relations in the process of ontology construction in the field of geography should have certain logic and rigor, and ensure the effective mapping between concepts and concepts. Through FaCT ++ or HermiT1.3.8 inference engine, it is verified the relationship between concepts inferred from ontology is consistent with the concept relation in geographic science thesaurus.

(3)Scalability. Domain ontology is a shareable specialized vocabulary knowledge base that provides a conceptual basis within the intended task. At present, the field of geography science is developing rapidly, the research is more and more deeply, the number of concepts and the relationship between concepts may be further upgraded and refined in the future. Under the premise of not changing the original definition, the scientific and rational ontology should be able to expand and define the new concept and the relationship between the concepts on the basis of the original ontology.

(4)Maintainability. The field of geographical science will be applied to different areas of the geographical direction, experts in the field participate in development process of thesaurus, with a certain degree of scientific and authoritative. With the increasing demand, the original domain ontology can not meet the needs of scientific research and application. The completeness of terminology and the relationships between terms need to be further improved and merged with other existing ontologies. Therefore, the scientific field ontology should have a good maintainability.

Domain ontology building steps

The traditional domain ontology construction methods mainly include KACTUS engineering method, SENSUS method, skeleton method, TOVE method, IDEF5 method, METHONTOLOGY method and seven-step method. The most mature ontology construction method is domain ontology construction method developed by Stanford University, that is seven-step method. These ontology construction methods are mostly the experience summed up in the process of constructing the domain ontology, and have not been authorized by the authority standard institutions, and have certain limitations. This paper draws on the cycle acquisition method proposed by Alexander Maedche[14] and the concept of ontology modularization proposed by Fu Ling[15], National Science Library of the Chinese Academy of Sciences, and combines the thought of software engineering life cycle and the actual situation of geographic science thesaurus. This paper proposes a domain ontology construction flow chart based on thesauri. As is shown in Figure 1.

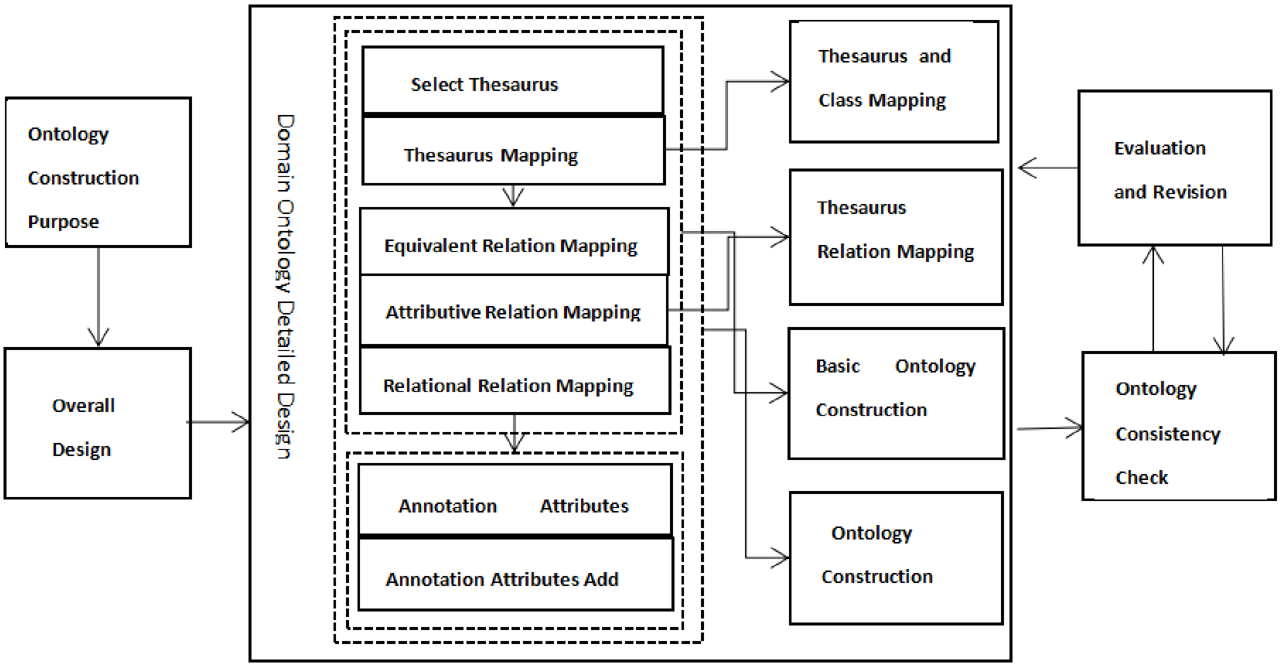
Ontology Construction Purpose
Overall Design
Domain Ontology Detailed Design
Select Thesaurus
Thesaurus Mapping
Equivalent Relation Mapping
Attributive Relation Mapping
Relational Relation Mapping
Annotation Attributes
Annotation Attributes Add
Thesaurus and Class Mapping
Thesaurus Relation Mapping
Basic Ontology Construction
Ontology Construction
Evaluation and Revision
Ontology Consistency Check

Fig.1 Domain Ontology Construction Process

In the construction of the whole domain ontology, using the top-down approach to construct the process, the construction process is divided into five steps: domain ontology construction purpose; overall design; detailed design; ontology consistency test and ontology evaluation and revision.

The purpose of this paper is to build the ontology of geographical sciences based on the geographical science thesaurus, to use the vocabulary and the relationships between words of the Geographical Science Thesaurus as the data source, to transform the descriptors in the thesauri as the ontology class and concepts, and transforms the relationship between the words in the thesaurus as the relation between classes in the ontology. The ontology of geographic science is constructed and displayed by OWL ontology description language, which provides a common understanding of the knowledge of the field of geography, so that make the ontology in the field as the mutual understanding semantic basis between people and computers, computer and computer.

In the domain ontology construction process, the most important step is the domain ontology detailed design, this process is divided into two steps:

(1)Based on Domain Ontology Design. After completing the construction of domain ontology, domain scope, terminology selection and overall design, we define class and class hierarchy, class attributes and relationships. This stage mainly completes the mapping of the thesauri and the ontology class and the relation between the lexeme and the ontology class in the thesaurus, and uses OWL Lite[16] to describe the language.

(2)Domain Ontology Design. The main task of this stage is to enrich and normalize the content of the previous stage. The specific work is to further optimize the ontology class and semantic relation through the OWL DL[16] ontology description language, combined with the thesaurus vocabulary characteristics and semantic relation accuracy and avoid semantic inconsistency. At the same time we add class annotation attributes, examples, descriptions and other content for the class, making the whole ontology system more perfect.

Ontology description language solves the problem of ontology representation, that is, how to express the concept in the thesaurus in a formalized way. This paper adopts OWL[17,18]as the ontology description language and Protege[19,20] as ontology editing tool. OWL is an ontology description language standard recommended by W3C on the basis of RDF (S) and DAML + OIL. As an extension of RDFS, OWL constructs a complex conceptual relation network by describing concepts, concept attributes and relationships among them. Protege is a multi-language ontology compilation tool developed by Stanford University Medical Department, which can provide convenient operation platform for ontology building concept class, relation, attribute and instance addition and deletion.

3.4Domain Ontology Detailed Design

The Geographic Science Thesaurus lists all the professional vocabularies that cover a wide range of geography and related disciplines, and use the six reference symbols of "use, generation, genus, division, family and reference" to indicate the various relationships between words. The reference symbol "Y, D" indicates the synonymy relationship between formal and informal words, such as the relationship between "geographical environment" and "natural environment". In the transformation of relations between classes, we regard the two as the equivalent relationship, the purpose is to ensure the consistency of the index and the literature recall rate. The attributive relation corresponds to the reference symbol "S, F, Z". "S, F" representing the upper and lower relations is a supplementary approach to indexing and access to the literature. "Z" refers to the headword of the subordinate subordinate, that is, the highest epistatic subject, which is converted into the superclass of descriptor when mapping. There are "F" items without "S" items of the family head, the right with "*", such as the relationship between "geological map" , "Seismogram" , "Geographical Map" and "Mapping"; "C" corresponding to the relationship refers to keywords related to the subject,, referring to the close relationship between them but do not have a description of the sub-relations, such as the relationship between "geographical circle" and "geographical shell". When mapping, We should base on the actual situation using attribute declaration tag and attribute description label for customization.

3.4.1The Descriptors of the Thesaurus and the Mapping of Ontology

In constructing the domain ontology, the descriptors and non-descriptors in the thesaurus can be co-mapped into classes in the domain ontology. The OWL description language uses the <owl: Class> tag to represent classes, taking "Geography", "Geography Circle" and "Geographical System" as Examples in Geographic Science Thesaurus.

```
<owl:Class rdf:ID=“geography”></owl:Class>
<owl:Class rdf:ID=“Geographical circle”></owl:Class>
<owl:Class rdf:ID=“Geographical system”></owl:Class>
```

When using the "<owl: Class>" tag, the built-in attribute "ID" is used to represent the concrete class.

3.4.2Relationship Between Words and the Mapping of Ontology Class Relations

(1) OWL representation of equivalence relation. The thesaurus uses the reference "Y, D" to represent the equivalence relations among the thesauri, and uses the "<owl: equivalentClass>" tag in the OWL language to express the equivalence relation. In the thesaurus, "geographical environment" and "natural environment", "geomorphic environment" are the same relationship, specifically as follows:

```
<Owl: Class rdf: ID = "Geographic Environment">
   <Owl: equivalentClass rdf: resource = "#Natural"> </ owl: equivalentClass>
   <Owl: equivalentClass rdf: resource = "# Geomorphic Environment"> </ owl: equivalentClass>
</ Owl: Class>
```

(2) OWL representation of the Attributive Relation. In the thesaurus, the reference character "S, F, Z" represents the attributive relations among the thesauri, "S, F" represents the direct relationship between the thesaurus, "Z" represents the relationship between the thesaurus and the head of the family. The "<rdfs: subClassOf>" and "<rdfs: supperClassOf>" tags are used in the OWL language, and are represented as follows:

```
<Owl: Class rdf: ID = "Geographic Environment">
   <Rdfs: subClassOf rdf: resource = "# Geography" />
   <Rdfs: supperClassOf rdf: resource = "# marsh ecological environment" />
</ Owl: Class>
```

(3) OWL representation of the correlation. In the OWL language, there is currently no label be used to express the correlation between the thesaurus, so we need to use the OWL language to define a label to express the relationship between semantics to achieve the relationship between the descriptor and the mapping between classes. Combining the symmetry of the correlation, define and use the "relevant" attribute, as follows:

```
<Owl: ObjectProperty rdf: ID = "Relevant">
   <Rdf: type rdf: resource = "& owl; SymmetricProperty" />
   <Rdfs: label xml: lang = "en"> RelateTerm </ rdfs: label>
   <Rdfs: label xml: lang = "ch"> Related </ rdfs: label>
</ Owl: ObjectProperty>
<Owl: Class rdf: ID = "Geosphere">
   <Owl: RelateTerm rdf: resource = "# Geographic Shell">
</ Owl: Class>
```

Through the mapping of the thesauri and the class, the thesaurus and the relations between classes, we complete the basic domain ontology construction task at the detailed design stage, and then use Protege software to complete the OWL grammar test.

3.4.3The label attribute is added in the ontology

Building domain ontologies, it is not enough to describe the relationships between classes and classes. The thesauri provide descriptive information such as the English name and classification number of the descriptive words by describing the relationship between them. Adding ontology attributes to the ontology is a rich and normalized treatment of the contents of the basic domain ontology. According to the domain ontology construction principle, the ontology construction of geoscience domain should map the information in the Geographic Science Thesaurus to the ontology, so it is necessary to map the attribute information about the descriptor to the corresponding class. Take the "geographical environment" in the thesaurus as an example:

```
<Owl: AnnotationProperty rdf: ID = "English name" />
<Owl: AnnotationProperty rdf: ID = "category number" />
```

3.4.4The annotation attribute description of the class in the ontology is added

After declaring the attributes, you need to describe and restrict the attributes. Use the <rdf: type>, <rdf: commment>, <rdf: range> tags, and take the "English name" and "classification codes" as examples, details are as follows :

```
<Owl: AnnotationProperty rdf: about = "English name">
   <Rdf: comment rdf: datatype = "& asd; string"> This attribute is a string type, unique
   </ Rdf: commment>
   <Rdf: type rdf: resource = "& owl; FunctionalProperty" />
   <Rdf: type rdf: resource = "& owl; DatatypeProperty" />
</ Owl: AnnotationProperty>
<Owl: Class rdf: ID = "Geographic Environment">
   <Rdfs: label xml: lang = "en"> geographical environment </ rdfs: label>
   <Rdfs: label xml: lang = "ch"> Geographic Environment </ rdfs: label>
   <Rdf: comment> The geographical location of the society and the sum of the various natural
conditions associated with it
   </ Rdf: commnet>
   <English name rdf: datatype = "& xsd; string"> geographical environment </ English name>
   <Category number rdf: datatype = "& xsd; string"> 20A </ category number>
</ Owl: Class>
```

After the detailed design phase is completed, the generated OWL files are imported into the Protege software, the FaCT ++ or HermiT 1.3.8 inference engine is started, the reasoning rules are combined with the OntoGraf function modules to visualize the "Geography" class and its subclasses. Part of the relationship is shown in Figure 2 below:

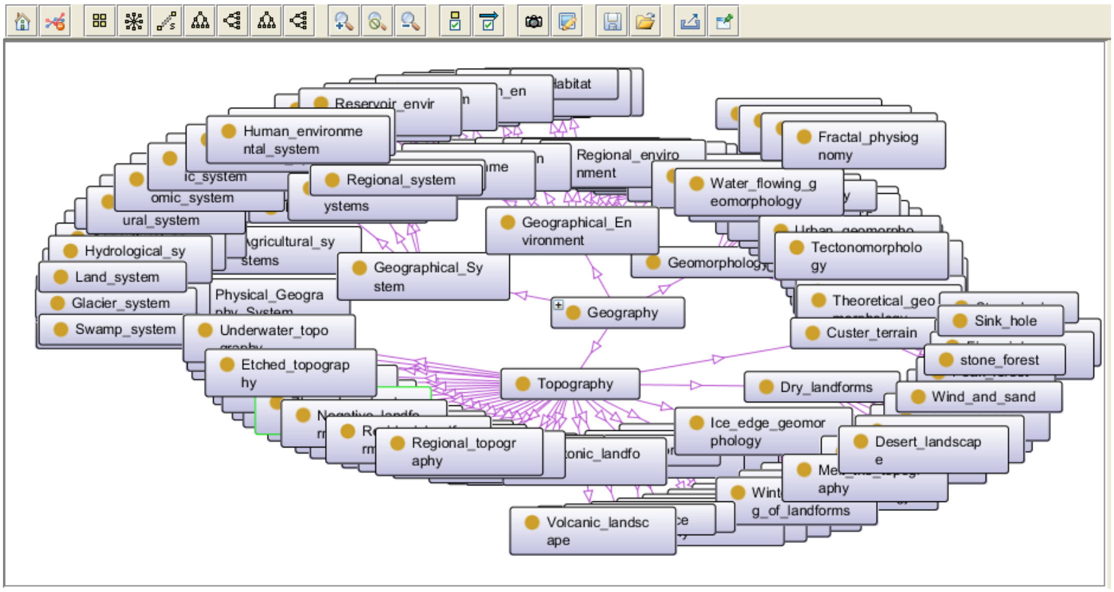

Fig.2 Partial Ontology Visualization Display

3.5Consistency check

Ontology construction is not a one-step process, we need to continue to cycle, in order to make domain ontology to meet the design requirements in the whole construction process. After the detailed design phase of domain ontology is completed, the consistency of ontology is tested from the following three aspects:

(1)Grammar check. After the detailed design phase of domain ontology is finished, the generated OWL text is imported into Protege software, and the domain ontology logical consistency test is carried out using its own FaCT ++ or HermiT 1.3.8 inference engine. If the domain ontology is constructed correctly, Query, reasoning, and presentation of relationships between classes can be performed.

(2)Content Consistency Check. Protege comes with the DL Query function module, you can query class Super Class, Equivalent Class, SubClasses, Individuals, etc., to verify the consistency of visual class and the relationship between classes. If there is inconsistency, we need to adjust the ontology class relations on the basis of the original ontology, and carry on grammar test again..

(3)Quantity consistency test. The number of classes in the domain ontology is compared with the number of descriptors included in the main table of Geographic Science Thesaurus, and the relations between thesauri and thesauri are mapped to ontology as far as possible.

4. Summary and Prospect

Based on the idea of software engineering and combining the authoritative and scientific characteristics of the Geographic Science Thesaurus, this paper realizes the transformation from thesaurus to the domain ontology by OWL ontology description language. But it is necessary to study the ontology evaluation method in the domain ontology construction process. It is necessary to further study the ontology evaluation process from the perspective of theory, technology and application, and how to use AHP and science to set up domain ontology evaluation index system. This process also requires the participation and help of domain experts to explore.

Acknowledgement

This paper is one of the achievements of the National Social Science Fund Project "Research on the Automatic Measurement of the Academic Innovation Power of the Humanities and Social Sciences under the Large Data Environment" (15BTQ058).

References

[1] Thomas R, Gruber. Howard Principles for the Design of Ontologies Used for Knowledge Sharing [J]. International Journal of Human-Computer Studies, 1995, 43 (5-6): 907-928.

[2] Studer R, Benjamins VR, Fensel D. Knowledge engineering, principles and methods. Data and Knowledge Engineering, 1998, 25 (112).

[3] Zhang Hongyan.Visualization Analysis on the Ontology of Library Information Field [J] .Research of Library Science, 2012,06: 7-12.

[4] Sudath R. Heiyanthuduwage, Rolf Schwitter, Mehmet A. Orgun. OWL 2 learn profile: an ontology sublanguage for the learning domain [J]. SpringerPlus, 2016, 51: 13-16.

[5] Tang Jing.Translation of thesauri into Ontology [J] .Information Theory and Practice, 2004,06: 642-645.

[6] Liu Wei, Wang Ying, Li Peijin. Bibliometrics Analysis of Ontology Research in Library and Information Field from 2001 to 2009 [J]. Information Science, 2010,11: 1673-1678.

[7] Li Guangda, Chang Chun, Zhang Jun-feng, Zheng Huai-guo, TanCui-ping. Study on Domain Ontology Visualization Construction [J] .info, 2013,09: 171-174 +127.

[8] Ding Sheng-chun, FU Zhu.Study on semi-automatic construction of domain ontology based on the space thesaurus [J] .Journal of Intelligence Theory and Practice, 2011,11: 113-116.

[7] DENG Zhi-hong, TANG Shi-wei, YANG Dong-qing, ZHANG Ming.Study on Ontology Representation and Retrieval Technology Based on XML [J]. Computer Engineering and Applications, 2002,03: 14-15 + 67.

[10] Xian Guo Jian. Agricultural Science Thesaurus to the agricultural ontology system of research and implementation [D]. Chinese Academy of Agricultural Sciences, 2008.

[11] PSin-Jae Kang, PJong-Hyeok Lee.Semi-automatic practical ontology construction by using a thesaurus, computational dictionaries, and large corpora [C] .Proceeding of the working on Human Language Technology and Knowledge Management, 2001: 362- 365.

[12] Daniel Kless, Simon Milton, Edmund Kazmierczak, Jutta Lindenthal. Thesaurus and ontology structure: Formal and pragmatic differences and similarities [J]. J Assn Inf Sci Tec, 2015, 667: 11-14.

[13] M.Uschold and M.King. Towards a methodology for building ontologier [C]. In Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95, Montreal, Canada, 1995.

[15] ZHANG Xiang, LI Xing, WEN Yun-Qing, SHEN Kai, HAO Jing-Kun.Virtual Ontology Construction of Semantic Web [J]. Journal of Southeast University (Natural Science Edition), 2015, 04: 652-656.

[15] SHANG Xin-li.Comparative Analysis of Foreign Ontology Construction Methods [J] .Library and Information Work, 2012,04: 116-119.

[16] Fu Ling, Liu Zheng. Ontology Modularization of Research[J]. Book and Information, 2013,01: 17-22.

[17] Djungu, Saint-Jean A O, Ekionea, Jean-Pierre Booto. OnLearn: The ontology for managing e-learning resources[J]. International Journal of Computer Science Issues (IJCSI), 2015, 123: 22-25.

[18] G. Antoniou, F V Harmelon. Web Ontology Language: OWl[J] .Handbook on Ontologies, 2004:

67-92.

[19] Aditya Kalyanpur, Bijan Parsia, Matthew Horridge, et al. Finding All Justifications of OWL DL Entailments [C]. Proceeding of the Sixth International Semantic Web Conference, Korea, 2007: 261-263.

[20] Al-Saiyd, Nedhal A, Al-Samarae, Muna F, Al-Sayed, Intisar A. A MULTI-AGENT SYSTEMS ENGINEERING FOR SEMANTIC SEARCH OF REUSE SOFTWARE COMPONENTS [J]. International Journal of Computer Science Issues (IJCSI) , 2014, 116: 11-12.