Ontology based Knowledge Model Research

Fei Wang[1] and Hao Fan
International School of Software
WuHan Univeristy
WuHan City, HuBei Province P.R.China 430079
feiwang@whu.edu.cn
hfan@whu.edu.cn

Abstract

Ontology has been widely applied in biological science (large-scale gene ontology), publishing field (Dublin-core standard and knowledge classification ontology), medicine and health care field (cancer ontology), culture inheritance field (museum and artist ontology). As the fundamental element of semantic web technology architecture, ontology provides expressive framework for knowledge reusability, sharing, representation and reasoning. However, ontology resource is quite rare in social science in China, especially in ancient historical records researching domain, which has turned into a bottleneck of semantic web development and application. On the strength of a semantic system—Chinese Fundamental Historical Records Analysis System, although inherent complexity of Chinese natural language processing (NLP) tools, especially aiming to ancient Chinese characters, this paper tried to make some study on knowledge abstraction, representation and conceptualization based on semantic ontology techniques in specific domain -- Chinese ancient historical records research.

In view of historical records domain features analysis and ontology constructing standard, this paper proposes a knowledge model of Chinese ancient historical records field, which mainly includes domain knowledge features analysis, modeling architecture and applying results. The key part of this paper is the architecture of the knowledge model. There are three levels presented in the model framework, namely resources level, transformation level and representation level. The knowledge resources level is dealing with the scope of raw data, that is what and how much materials are going to be used according to system requirement. The other important task of this fundamental level is to integrate those heterogeneous data to a relatively consistent data structure for further processing. Secondly, the transformation level is the most complex and important part of entire architecture. This level is to turn the initially processed data into ontological conceptual knowledge systems. The difficulty of transformation process is building up rule sets and ontology instances by rule abstraction and knowledge extraction from data objective. Besides, the design and developing of semi-automatic ontology building toolkit would affect the efficiency and precision of ontology knowledge base. The third level is a cluster of various historical records ontologies, which could be called into diversified knowledge systems with different dynasty, people, event and granularity in the light of different application requirement due to scalable and extensible mechanism of OWL ontology. As a result of this knowledge model architecture, a large-scale semantic knowledge system of Chinese ancient historical records was established. There are 122 classes, 32 object properties, 28 datatype properties, 15 reasoning properties and 179,053 instances in the ontology cluster. The amount of processed Chinese Character reached 17,835,000. This model is extensive research of ontology theory and technology in new application domain, which might be beneficial for historical records studying approach, knowledge discovery and in-depth study of domain ontology.

Keywords: ontology, knowledge model, historical records analysis, Chinese ancient character

---

[1] Dr. Fei Wang, Visiting Scholar of Computer Science Department, University of Wisconsin La Crosse, La Crosse, WI 54601,fwang@uwlax.edu

1 Introduction

Ontology has been widely applied in biological science (large-scale gene ontology), publishing field (Dublin-core standard and knowledge classification ontology), medicine and health care field (cancer ontology), culture inheritance field (museum and artist ontology). However, ontology resource is quite rare in social science in China, especially in ancient historical records researching domain.

On the other hand, ontology also plays an important role in semantic web, knowledge engineering and artificial intelligence technology developing, mainly dealing with interoperation of heterogeneous data, reusing and sharing in knowledge engineering, knowledge representation and abstraction. Therefore, ontology is beneficial for information system developing and knowledge modeling in different domains.

However, ontology system and resources are still very rare in historical researching field. Most publicly available ontologies are constricted into existed scientific domains, such as biological, medical, GIS and agricultural research. The ambiguity, subjectivity and controversial features of historical knowledge increase the degree of difficulty of constructing historical ontology. Therefore the study on historical records model based on semantic system explores knowledge organization and representation in Chinese historical scientific resources and improves the application of semantic key technology into social scientific research. This paper proposes historical records knowledge model based on semantic system, whose knowledge scope is 'Historical Records of Twenty-Four Dynasties' (HRTFD) published by Zhong Hua Press, which is the top authorized historical records publishing company of China. And this paper is one of the research findings of the project-'Chinese Basic Historical Records Analysis System' accomplished by the Center for the Studies of Information Resources of Wuhan University and Zhong Hua Press.

2 Related Work

There are two basic categories of ontology related to social science, namely universal category and particular category. The typical universal ontology includes WordNet [1], Cyc [2], IEEE SUMO [3] etc., which covers almost all aspects of scientific researching. Getty Thesaurus of Geographic Names [4], Union List of Artist Names [5], the CULTOS ontology [6] are quite specific to the particular domain.

There are not many ontology projects relating to historical research. VICODI [7] built the ontology of European history by Karlsruhe University in Germany, which could be used by heuristics during automatic contextualization and context-based search as a knowledge base via semantic annotation. And E-Culture project [8] funded by The Dutch Cultural Heritage has semantic image retrieval function. As to Chinese history ontology projects, 'ZiZhiTongJian' ontology built by Beijing University is relatively much smaller in scale. The ontology of 'Chinese Nationalist Party and Communist Party Cooperation History' is based on modern Chinese NLP systems which is much easier than ancient Chinese natural language process.

Therefore Chinese Basic Historical Records Analysis System based on Semantic System built up the ontology with massive data by ancient Chinese NLP technique. And HRTFD is unprecedented with its history-specific background. The historical records knowledge model proposed by this paper is the theoretical exploration according to the HRTFD ontology.

3 Architecture of Historical Records Knowledge Model

According to ontology standard and knowledge modeling approach, this paper proposed historical records knowledge model based on ontology being combined with the characteristics of HRTFD's knowledge system.

3.1 Characteristics of Historical Records Knowledge

In this paper, HRTFD are the research objectives of historical records analyzing. And characteristic analysis is the necessary base of the historical records knowledge modeling.

3.1.1 Widespread scope of knowledge resources

As we all know, Chinese HRTFD is the biggest official historical records in the world.

1) The content of this series of historical records covers almost every aspect of social life and natural world. For example, 'ShiHuoZhi' records economical system of different dynasties. 'DiLiZhi' mainly presents local population and administrative districts. And 'TianWenZhi' recounts ancient astronomy findings of China. Besides, there are a lot of materials recording historical situation of other area and countries, such as Central Asia area including Kazakhstan, Tadzhikistan, Turkmenistan, Kyrgyzstan and Uzbekistan. As a result, the HRTFD is also regarded as the most important evidence to research economy, politics, culture, geography, history, ethnicity and local customs of Central Asia area. The scope of recorded explicit knowledge and implicit knowledge on these historical records is extremely extensive.

2) As to the time span, HRTFD covers more than four thousand years from 2550 B.C. to A.C. 1644. From the first record 'ShiJi' to the last record 'MingShi', it took more than 1,800 years to form this huge series of historical records. During such long period, dozens of authors devoted most their lifetime to finish the historical records of different dynasties.

3) From the angle of quantity of knowledge processing, the amount of text is rather massive. There are totally 3213 series and about forty million Chinese ancient characters to process. What's more, different from modern Chinese language representation, ancient Chinese described event and people in a concise pattern. One-character word was rather popular in ancient Chinese language. Especially the recording style of the first four histories, namely, 'ShiJi', 'HanShu', 'HouHanShu' and 'SanGuoZhi', was even more concise. The concise describing style signifies that the quantity of knowledge resources contained in the simple texts is even larger. And undoubtedly the massive amount of

knowledge resources is a challenge to construct historical records knowledge system based on ontology.

3.1.2 Widespread scope of knowledge resources

Complicated relationship among concepts: The main conceptual categories include people, time, location and event. The relationships among those concepts comprised different changing multi-level networks. Especially the relationships between people are not always linear and one-direction. In the historical records of such long time, people could be friends at some time and be enemies at some other time.

3.1.3 Widespread scope of knowledge resources

Free representation of natural language: All the content of HRTFD was recorded by ancient Chinese characters. There is no NLP tool for ancient Chinese language currently. The well-known ICTCLAS and LTP toolkit s only deal with modern Chinese language participle and parsing. Therefore the extraction of concept and semantic annotation are big challenges for ontology construction of Twenty-Four Histories.

3.2 Historical Records Knowledge Model Contents

Combing with the characteristic analysis of historical records and developing of ontology application toolkit, this paper proposed Historical Records Knowledge Model based on ontology. The structure of the model is shown as Figure 1. There are four levels, namely the knowledge resources level, the knowledge domain analyzer level, the knowledge transforming level and the knowledge representation Level. The relationships between different levels do not simply follow the bottom-up sequence, which are incremental iterative procedure with repairing and providing feedback mechanism.

3.2.1 Knowledge resources level

The main task of this level is different knowledge resources obtainment and collection. This level of model could define the resource assemblages that satisfy the system functional requirement. Knowledge resources are collected. The obtainment of heterogeneous data from different domains lays the groundwork for further domain analyzing and complex semantic content mining. The objects of knowledge resources level almost include all kinds of structured, semi-structured and unstructured data. The relational database, XML files, TXT files DOC files and PDF files are covered. The images, audio files and video files are also regarded as knowledge resources objects with URI. All kinds of resources objects could be transformed and integrated into the historical records knowledge system through the Historical Records Knowledge Model based on ontology.

3.2.2   Knowledge transformation level

This level is the key of entire model, which is also the most difficult. The aim of transformation process is building up rule sets and ontology instances by means of knowledge extraction and rule extraction of data objectives. The implementation

approach is the developing of semi-automatic ontology building toolkit combined with domain expert's involvement. The ontology building toolkit cannot accomplish the entire ontology building process because of the ontology conceptual model's complexity and precision. Consequently the involvement of domain expert is indispensable to the Historical Records Knowledge Model. The most important results in this level are the domain rule set and ontology. And the design and developing of semi-automatic building toolkit would affect the efficiency and precision of ontology knowledge base building.

1) Domain rule set formulation: The reasoning rules are the characteristic principles of different domain. In this paper, the selected domain is historical records analysis. Therefore the recorded people's blood relationships and social connections are the basic principles. The rule set could be either the existed research findings, or the results concluded by domain experts. After the domain reasoning principles are confirmed, the series of specific rules with causality and conditional relation should be proposed to form the rules set of historical records analysis. As a result, domain reasoning mechanism means that rules set is the outcome of domain rules extracting procedure which domain experts have to be involved in the checking and revising. Certainly domain experts would present custom rules directly that could be transformed by logic representing tools, such as regular expression, and be added to the rules set. Ontology could be constructed based on the knowledge extraction from heterogeneous data resources according to rules set. In the light of ancient Chinese grammar, considering organizing and modeling of historical people knowledge, the representing regular pattern how people take a post is as follows.

<People's name>行<People's post>

<People's name>领<People's post>

以<People's name>为<People's post>

拜<People's name><People's post>

……

From those rules, computers could extract the knowledge about people's posts and offices. Ontology knowledge content and values of properties could be concluded as much as possible via such inferring rules. And the more accurate the rules are summarized, the more authentic semantic content extracted from knowledge resources is. The semantic content of knowledge is represented by triple when formalizing knowledge extracting procedure. For example,

<Sanguo:#YuanShao>

<Sanguo:hasTitle>

<Sanguo:TaiWei>

The value of subject 'YuanShao' is a name of historical person. And the value of object 'TaiWei' is an ancient Chinese post name.
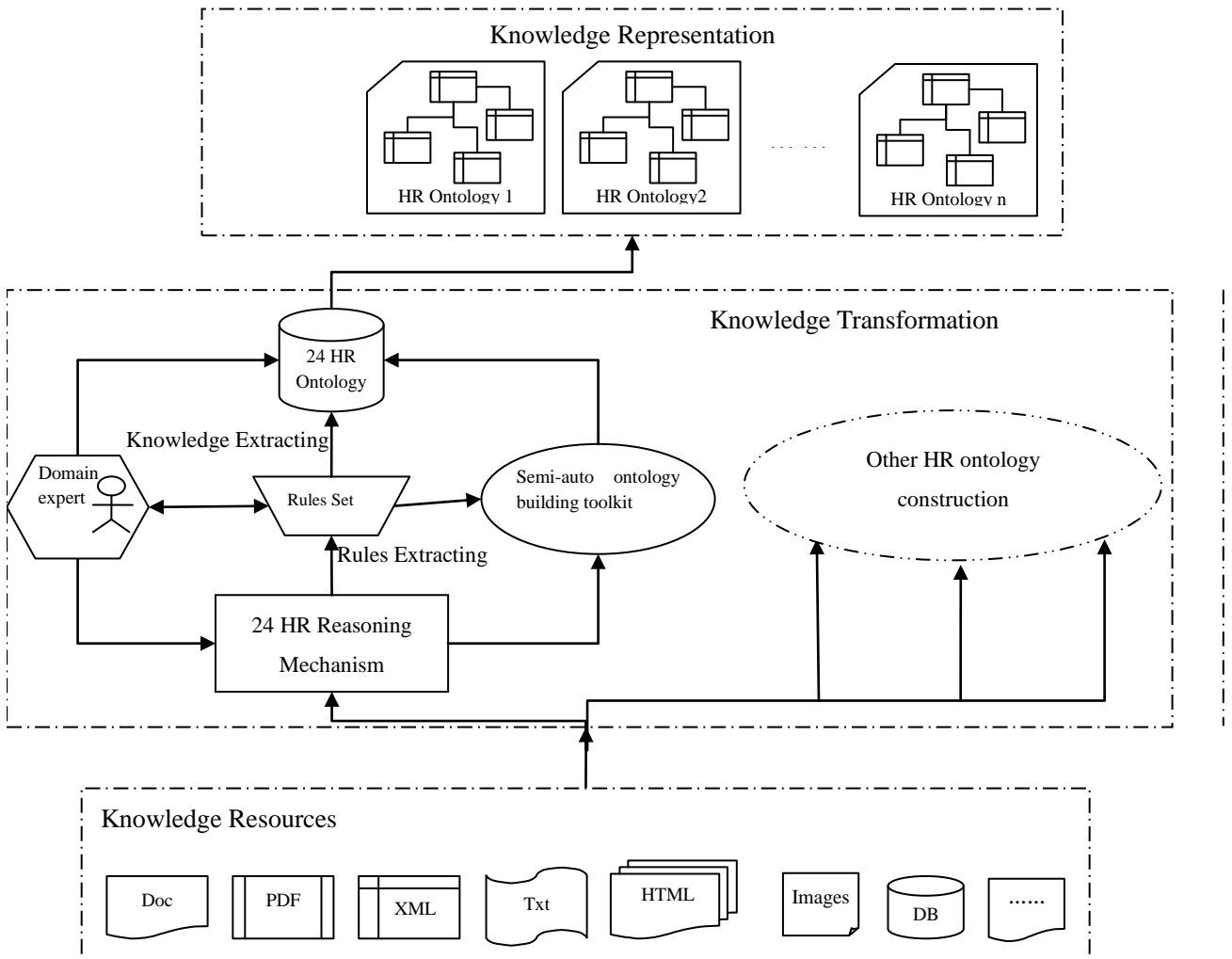

Figure 1 Historical Records Knowledge Model

In a word, the knowledge transforming task could be divided into two steps, those are 'domain rules extraction' and 'domain knowledge extraction'. The transforming object is heterogeneous data from knowledge resources level. The basis of transforming is the reasoning mechanism which represents the main domain principles. And the transforming results include domain rules set and domain ontology. All transforming procedures and results are related to domain knowledge. The involvement of domain expert is to make sure that historical records knowledge system based on ontology is correct and fitting the facts. The combination of domain expert's artificial involvement and semi-automatic ontology building toolkit is the implementing approach of knowledge transforming level.

2) Semi-automatic ontology building toolkit. At present, as we all know, ontology could not be possibly constructed automatically [9]. Because of Chinese NLP basis, universality of different domains and strict ontology modeling requirement, semi-automatic ontology building toolkit developing has been a research difficulty. Therefore the semi-automatic ontology building toolkit plays an indispensable role in the architecture of Historical Records Knowledge Model. Especially for dealing with massive data historical records objectives, such as HRTFD, it is unimaginable that manual operation is the only method to build up ontologies. However this paper will not discuss semi-automatic ontology building approach and implementation in detail which will be studied in other papers.

3.2.3 Knowledge representation level

This level is the ultimate result of Historical Records Knowledge Model based on ontology. There should be a cluster of different historical records ontologies. According to the OWL standard of W3C, there is a flexible 'Import' mechanism for OWL ontology. The ontology system could 'Import' different ontology knowledge base in the light of different application requirement. These different HR ontologies together constitute a general historical records knowledge system based on ontology, which is not simply superposition of different domainial ontologies, but an integrated Historical Records Knowledge System, which is also could be regarded as the macroscopic collection of loosely-coupled system with subsets of different historical records ontologies. Because of the same ontology modeling standard, the general historical records knowledge system could be applied as a whole. And one part or more could be picked up and applied into specific requirement, forming historical records ontologies that satisfy the particular knowledge managing requirement of different granularity. There could be multi-dynasty historical records ontologies, such as 'Sui Dang Period' ontology and 'Pre-Qin Period' ontology. And there could be other kinds of ontology besides people-based ones, such as 'War Event' ontology, 'Earthquake Event' ontology and 'SanGuo Location' ontology.

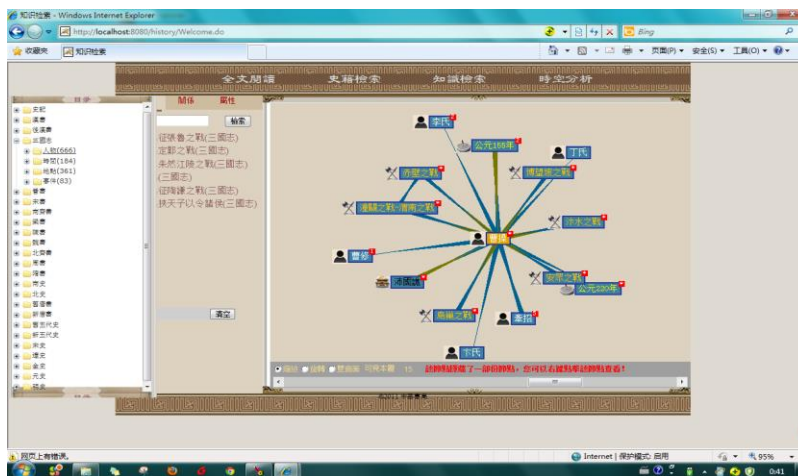# 4 Historical Records Knowledge Model Application

Figure 2 Knowledge Base of HRTFD

6

The Historical Records Knowledge Model proposed in this paper was actually applied into several practical projects successfully. There is the ontology knowledge system of 'Chinese Nationalist Party and Communist Party Cooperation History', which includes six ontologies, namely, event, people, location, institute, resource and time. The related ontologies could be invoked altogether or just one or more of them according to user requirement and system application. Such flexibility is one of advantages to construct knowledge system on the basis of ontological standard.

For example, in semantic analysis system constructing project of HRTFD, a large-scale historical records knowledge system based on ontology was established according to above-mentioned knowledge model architecture. Figure 2 implies that the basic structure of historical records ontology knowledge base of HRTFD and application interface of the semantic knowledge base system. As a result of knowledge model proposed in this paper, there are 122 ontology classes, 32 object properties, 28 data type properties and 15 reasoning properties in the HRTFD ontology knowledge base. The amount of ontology instance reaches 179,053. And the amount of processed Chinese character is 107,835,000 in total. And the size of ontology knowledge base is incomparable of Chinese historical records domain.

References

 [1] WordNet 3.0 Reference Manual [EB/OL].[2013-04-24].
https://wordnet.princeton.edu/wordnet/documentation/
[2] Cyc's knowledge-based solution address. [EB/OL].[2013-04-24].
http://www.cyc.com/
[3] Suggested Upper Merged Ontology. [EB/OL].[2013-04-24].
http://www.ontologyportal.org/
[4] Getty Thesaurus of Geographic Names.[EB/OL].[2013-04-24].
http://www.getty.edu/research/conducting_research/vocabularies/aat/
[5] Union List of Artist Names. [EB/OL].[2013-04-24].
http://www.getty.edu/research/conducting_research/vocabularies/aat
[6] Multimedia knowledge management tools for culture and arts. [EB/OL].[2013-04-24].
http://www.cultos.org/
[7] VICODI objectives [EB/OL].[2013-04-24] http://www.vicodi.org/
[8] MultimediaN N9C Eculture project document [EB/OL].[2013-04-24]. http://e-culture.multimedian.nl
[9] LiWenRui, BaiChengXi.The design and implementation of ontology learning model.[J].Journal of HeNan University (Natural Science), 2006, no. 12, pp. 100-102.