Ontology development for unified traditional Chinese medical language system

Xuezhong Zhou\textsuperscript{a,}\textsuperscript{*}, Zhaohui Wu\textsuperscript{a}, Aining Yin\textsuperscript{b}, Lancheng Wu\textsuperscript{b}, Weiyu Fan\textsuperscript{b}, Ruen Zhang\textsuperscript{b}

\textsuperscript{a}College of Computer Science, Zhejiang University, Hangzhou 310027, PR China
\textsuperscript{b}Institute of Information on Traditional Chinese Medicine, China Academy of Traditional Chinese Medicine, Beijing 100700, PR China

Received 31 March 2003; received in revised form 26 October 2003; accepted 17 January 2004

\textbf{1. Introduction}

Traditional Chinese medicine (TCM) is a complete system of medicine encompassing the entire range of human experience. Thousands of scientific studies that support traditional Chinese medical treatments are published yearly in journals around the world. However, even patients who benefit from treatments such as acupuncture or Chinese herbal therapy may not understand all the components of the TCM system. That may be, in part, because TCM is based on a dynamic understanding of energy and flow that has more to do with Western physics than Western medicine. TCM embodies rich dialectical thoughts, such as the holistic connections and the unity of yin and yang. The ideas of integration and \textit{bian\_zhen\_lun\_zhi} are the fundamental infrastructure of TCM \cite{1}.

With the development of information technology and wide use of the Internet, immense amount of disparate isolated medical databases, electronic patient record (EPR), hospital information systems (HIS) and knowledge sources were developed. In 2000, we developed a unified web accessible multi-databases query system of TCM bibliographic databases and specific medical databases to address the distributed, heterogeneous information sources retrieval in TCM. It has been available on the website for registered user online for 5 years \cite{2}. As a complete system with complex disciplines and concepts, TCM has the main obstacle that large amount of ambiguous and polysemous terminologies existed during the information processing procedure. We have initiated the unified traditional Chinese artificial intelligence in medicine (2004) 32, 15—27

\textsuperscript{*} Corresponding author. Tel.: +86-571-8795-1647; fax: +86-571-8795-1647.
\textit{E-mail address:} zxz@cs.edu.cn (X. Zhou).

\textbf{KEYWORDS}
Traditional Chinese medicine; Ontology development; Medical language system

\textbf{Summary} Traditional Chinese medicine (TCM) as a complete knowledge system researches into human health conditions via a different approach compared to orthodox medicine. We are developing a unified traditional Chinese medical language system (UTCMLS) through an ontology approach that will support TCM language knowledge storage, concept-based information retrieval and information integration. UTCMLS is a huge knowledge project, which is a broad collaboration of 16 distributed groups, most of them with no prior experience of formal ontology development. Therefore, the cooperative and comprehensive ontology engineering is crucial. We use Protég\é 2000 for ontology development of concepts and relationships that represent the domain and that will permit storage of TCM knowledge. This paper focuses on the methodology, design and development of ontology for UTCMLS.

© 2004 Published by Elsevier B.V.
medical language system (UTCMLS) project since 2001, which is funded by China Ministry of Science and Technology to study the terminology standardization, knowledge acquisition and integration in TCM. We recognized that there are three main challenges in UTCMLS project.

1. To design a reusable and refinalable ontology that integrates and accommodates all the complex TCM knowledge.
2. To harness a broad collaboration of different domain experts in distributed sites in ontology development.
3. To develop a knowledge infrastructure for semantic web.

This article mainly addresses the former two challenges, which are relevant to the design and development of TCM ontology. Ontology building is a non-trivial and valuable work for domain data, information and knowledge integration, especially for the complicated and comprehensive domain like TCM. Using Protégé 2000, we try to facilitate the development of TCM ontology and alleviate the labor through methodology and structure design of ontology. The rest of this paper is arranged as follows. To illuminate TCM as a complete system and to clarify the methodology of ontology design, Section 2 gives a detailed overview of the knowledge system of TCM from the disciplines perspective. An ontology overview is proposed in Sections 3 and 4 introduces Protégé 2000, the ontology tool we use to build the UTCMLS. Section 5 discusses the methodology, knowledge acquisition and information integration for ontology development. Section 6 gives current main results of ontology development. Finally, the concluding remarks and future work are proposed.

2. The principle and knowledge system of TCM

TCM is a medical science that embodies Chinese culture and philosophical principle, which is the basis and essence of the simple and naïve materialism in China. Because the understanding of TCM concept, theory and philosophy is vital to this ontology’s development, we give a brief overview of TCM knowledge system.

Traditional Chinese medicine embodies rich dialectical thought, such as that of the holistic connections and the unity of yin and yang. It deals with many facets of human anatomy and physiology: zang-fu (organs), meridians (main and collateral channels), qi (vital energy), blood, jing (essence of life), body fluid, the inside and outside of the body, as well as the connections between the whole and the parts. It also examines the effect of the social and natural environment—the universe, the sun and moon, the weather, the seasons and geography on the interrelations and conditioning of yin and yang. The result has been the formation of a system of thought about the interrelations behind spirit and organism, zang and fu, and the inside and outside of the body. Traditional Chinese medicine puts the human body into a large system for observation and explores the interrelationship among formations, factors and variables, both within and outside the body; it regards and deals with these interrelations with reference to data that are correspondingly interrelated; it uses the principle of stabilization to ‘harmonize yin and yang to reach a state of equilibrium’, adjust their relationship so that they remain in a healthy state. The detailed discussions and remarks about the methodology of TCM can be found in [1,3].

Given the difficulties to TCM concept (i.e. its complexity, vast, variable and non-standard), and the importance given to that enormous amount of ancient literatures, which have been a general and core knowledge sources in TCM, it is a great challenge and central role to develop an ontology of formally specified concepts and relationships for UTCMLS. The design, implement and use will be described in detail in the next several sections.

3. What is an ontology

Ontology is a branch of philosophy concerned with the study of what exists. ‘Ontology’ is often used by philosophers as a synonym for ‘metaphysics’ [4]. Philosophical ontology is a descriptive enterprise. It is distinguished from the special sciences not only in its radical generality but also in its primary goal or focus: it seeks, not predication or explanation, but rather taxonomy. Formal ontologies have been proposed since the 18th century, including recent ones such as those by Carnap [5] and Bunge [6].

It was McCarthy [7] who first recognized the overlap between work done in philosophical ontology and the activity of building the logical theories of AI systems. McCarthy affirmed already in 1980 that builders of logic-based intelligent systems must first ‘list everything that exists, building an ontology of our world’. According to Gruber [8] an ontology is a ‘specification of a conceptualization’. While Guarino [9] argued that ‘an ontology is a logical theory accounting for the intended meaning of a formal vocabulary’. Ontologies are essential for developing and using knowledge-based systems. Every knowledge model has an ontological...
commitment [10], that is, a partial semantic account of the intended conceptualization of a logical theory. Ontologies form the foundation for major projects in knowledge representation such as CYC [11], TOVE [12], KACTUS [13] and SENSUS [14]. Medicine has been the active ontology research and construction area for large knowledge bases. There are several distinguished efforts in medical terminology systems like SNOMED-RT [15] and ‘Canon group’ [16]. The semantic network in unified medical language system (UMLS) [17] is also considered a distinguished terminological ontology [18]. GALEN [19] is a project developing medical terminology servers and data entry systems based on ontology, the common reference model, which is formulated in a specialized description logic, GRAIL [20]. Also reusable medical ontologies are strongly recommended by Schreiber and Musen [21,22] in intelligent systems, furthermore, Van Heijst et al. [23,24] have a case study in construction of library of reusable ontologies and they proposed several important and useful principles to address the corresponding hugeness problem and the interaction problem. TCM is a specific domain with large amount of knowledge. The goal of TCM ontology development is to facilitate the development of TCM terminological KBS by providing a reusable core generic ontology and relevant skeleton sub-ontologies.

4. Protégé 2000: the tool we use

We use Protégé 2000 [25] as an ontology editor (also as knowledge acquisition tool) with RDFS as the underlying representation language. Protégé 2000 is a frame-based knowledge base development and management system that offers classes, slots, facets and instances as the building blocks for representing knowledge. The Classes Tab is an ontology editor, which designs classes in flexible style and organizes classes as hierarchy. Classes have slots whose value may or may not be inherited. Facets specify the cardinality and data type of the slot value. The Instances Tab can help the user acquire the knowledge of domain. The ontology plus instances can be viewed as a domain knowledge base. We choose Protégé 2000 because:

1. it integrates ontology editors and knowledge acquisition tools in a single application to facilitate the knowledge engineering process;
2. it has the extensible component architecture; and
3. it defines a flexible metaclass architecture and supports many formats such as OKBC, RDF/RDFS and database storage [26], and the DAML + OIL is also supported by the additional plug-ins in the current version.

There is huge amount of knowledge and data that is required to put into knowledge base. Protégé 2000s user interfaces are intuitive for domain experts to work with for knowledge acquisition. Fig. 1 is an interface screen shot for ontology development in Protégé 2000. As it can be seen on the left frame of Protégé interface, we defined six core top-level classes, namely semantic type, semantic relationship, concept name, concept definition, concept interpretation and concept relationship in TCM ontology. The right frame contains the specific definitions of a class (i.e. name, documentation, constraints, role, template slots). Section 5 gives the detailed description of the 6 core top-level classes definitions.

5. Ontology design and development for UTCMLS

The development of ontologies is a modeling activity that needs the ontological engineers (also called ontologists) that have sufficient understanding of the domain of concern and are familiar with knowledge representation languages. Problems like ontology and conceptual modeling need to be studied under a highly interdisciplinary perspective [27]. Ontology construction is a complex and labor-intensive activity. There have existed several controlled vocabularies and many special terminology lexicons in TCM, but combining and synchronizing individual versions of existing medical terminology vocabularies into a unified terminological system is still a problem, because of the heterogeneity and indistinctness in the terminology used to describe the terminological systems [28]. It is accelerated by the various and non-standard using of words in the clinical practice. The ontology development for UTCMLS is still in the preliminary stage. Only small part of sub-ontologies (e.g. the basic theory of TCM, formula of herbal medicine, Chinese materia medica and acupuncture) has been developed. About 8000 class concepts and 50,000 instance concepts are defined in the current TCM ontology. Whereas we estimate the number of concepts of TCM will up to several hundreds of thousands, maybe even reaches several millions. Furthermore, because the terminology knowledge is mastered and used in practice by different groups of experts, there should be a broad co-operation in knowledge acquisition procedure. A methodology of loosely coupled development and quality assurance is needed to build a formal final ontology. Although
agreement on a high level schema is a prerequisite for effective co-operation between different groups of modelers, it is not practically possible to build a philosophically perfect model. We would rather aim to assure the ontology to fit closely enough with most usage requirements, to be refinable and to be reusable. We give a summary discussion of the methodology, knowledge acquisition, design and development of ontology in TCM in the successive sections.

5.1. Methodology of ontology development

The use and importance of ontologies is widespread, however, building ontologies is largely a black art. All the methodologies, such as TOVE [12], SENSUS [15], etc. are task-specific. Uschold [18] emphasized that no unified methodology is suitable to all the jobs, but different approaches are required for different circumstances. Jones et al. [30] had an overview of ontology methodologies and proposed guidelines for ontology development. It shows that ontological engineering is still a craft rather than a science at present. There are two main tasks involved in content development for the TCM ontology: (1) knowledge acquisition and conceptualization from disparate TCM information sources; and (2) formulization and implementation of ontology schema. The building of TCM domain ontology conforms to several principles.

1. Refinement is needed. The methodology, which based on the prototype ontology, is preferred in TCM ontology building for the vast, complex and dynamic knowledge system in TCM.
2. Development should be based on informal ontologies. There are many terminology lexicons and a number of controlled vocabularies. These knowledge sources can be viewed as the informal ontologies at the start point.
3. Evaluation is essential and important. Ambiguity and polysemy is the characteristic phenomenon of TCM concepts. No standard has been agreed on the concept structure and relationship in the TCM discipline. The ontology development of TCM is a continuous procedure of evaluation and development.
4. The methodology of distributed loosely coupled development is required. The task of building TCM ontology is distributed across 16 sites in China, and each site is responsible for corresponding sub-domain terminology work of TCM.

Rector et al. [31] had the practice of distributed cooperative ontology development in medical domain, in which an intermediate representation at the knowledge level has been used to control the quality of ontology development. We also adopt the intermediate representation mechanism based on tabular and graph notations. The knowledge acquisition, conceptualization, integration, implementation, evaluation and documentation activities are involved in the ontology development. The evaluation mainly focuses on the quality control of the intermediate representation from the distributed domain experts. We also make guidelines and the

![Figure 1](image)

**Figure 1** Protégé 2000 user interface and the overview of TCM ontology in editing.
criteria of Gruber [8] for domain experts to control process of knowledge acquisition and conceptualization. According to the above principles, we applied development criterion to both whole TCM ontology and individual sub-ontologies as shown in Fig. 2. The current practice showed that this criterion assured the quality control of ontology, interaction, communication between domain experts and central knowledge engineering team.

A core generic framework of TCM ontology is designed at the start point of ontology development. About 14 sub-ontologies and six core top-level classes (Fig. 4 depicts the definitions of each class) are defined as the initial skeleton ontology. The sub-ontologies are defined according to the disciplines of TCM based on the most domain experts’ viewpoints as shown in Table 1.

5.2. Knowledge acquisition

In 2001, China Academy of Traditional Chinese Medicine and College of Computer Science, Zhejiang University initiate and organize the project of UTCMLS as the basis of all the TCM information related work. We aim to implement a unified TCM language system, which stimulates the knowledge and concept integration in TCM information processing. The National Library of Medicine in the United States has assembled a large multidisciplinary, multi-site team to work on the unified medical language system [17], aimed at reducing fundamental barriers to the application of computers to medicine [29].

UMLS is a successful task in medical terminology research, which inspired our research of TCM knowledge and terminology problem. Much good experience has been learned from UMLS. The structure of TCM ontology is heavily influenced by the semantic network of UMLS. However, the work of TCM as a complete discipline system and complex giant system is much more complicated. Some principles should be adhered to during the knowledge acquisition procedure.

1. Deep analysis of specialized lexicons as the knowledge source. The scientific control of the conceptual glossary of a discipline is the most import issue in natural and artificial language processing. We combine the pre-controlled vocabulary and post-controlled vocabulary as the whole and have a multi-level description of conceptual glossary such as morphologic, lexics, semantics and pragmatics, etc.

2. A good construction of the TCM oriented concept framework. TCM involves the complete discipline system including medicine, nature, agriculture and humanities, etc. The knowledge system of TCM has complex semantic concept structures, types and relationships, which refers to multi-disciplines content. Therefore, comprehensive analysis and research of the terminology of TCM is needed before the ontology design. A TCM oriented concept framework should be constructed to address all the knowledge engineering problems of related disciplines.

3. Efficient combination of controlled vocabularies and specialized lexicons. As TCM has various concept taxonomical frameworks, a unified concept infrastructure should be established on the basis of controlled vocabularies and specialized lexicons. The controlled vocabulary can be viewed as an ontology, which has no instances. The specialist lexicon can be viewed as the instances of ontology. Both of them constitute a knowledge base.

4. On basis of the TCM science and referring to the other relevant disciplines. We develop the UTCMLS not only for the information integration and processing in TCM field, but also for those in the agriculture, pharmaceutical technology and western scientific medicine, etc. This is coordinated along with the characteristics of TCM discipline system.

From the perspective of TCM discipline, considering the medical concept and its relationship, we define the TCM knowledge system by two components: (1) concept system; and (2) semantic system.
The concept system initially contains about 14 sub-ontologies according to the division of TCM disciplines and four basic top-level classes to define each concept. The semantic system concerns the semantic type and semantic relationship of concept (Fig. 3).

According to the time, function, space, entity and concept attributes of TCM knowledge, we have

![Semantic System Diagram](image)

**Figure 3** The semantic system framework. The detailed class definitions and instances of semantic types and semantic relationships are provided in Section 6.
defined 59 kinds of semantic relationships between concepts and about one hundred and four kinds of TCM semantic types plus all the semantic types of UMLS. In Section 6, we will introduce the current version of TCM ontology.

5.3. Integrating and merging of TCM ontology

Information integration is a major application area for ontologies. Ontology integration is possible only if the intended models of the original conceptualizations that the two ontologies associated with overlap [9]. In the procedure of TCM ontology development, we must let ontology be built by different experts in distributed environment for the very complex knowledge acquisition work of TCM. We use the top-down approach to develop the 14 sub-ontologies and other six core top-level classes and distribute the 14 sub-ontologies to the domain experts of about 16 TCM research organizations in China. The bottom-up approach is used during the development of each sub-ontology. Therefore, ontology merging (information integration) is a must. We use IMPORT to merge the sub-ontologies from different sources to a unified TCM ontology. IMPORT is a plug-in of the Protégé 2000, which is the latest version of SMART [32]. It is showed from the TCM ontology practice that IMPORT is an effective tool to merge ontologies.

6. Results

The development of UTCMLS is a process of building a systematized general knowledge oriented TCM terminological system through an ontology approach. We have a nomenclature committee consisting of seven TCM linguistic and terminological experts to evaluate the nomenclature standard and fix the final definitions with the other participant linguistic and domain experts. More than 30 experts in the fields of traditional Chinese medicine, medical informatics, knowledge engineering, and medical administration were consulted about the development of TCM ontology. The categories of the structure of the TCM ontology are formed according to the structures of the controlled vocabularies and the standard textbooks. More than one hundred controlled vocabularies and terminologies have been chosen as the sources for TCM ontology, which are stored in a simple ontology named reference book (18) in Fig. 4. Some of the reference books as main knowledge sources are Chinese Traditional Medicine and Materia Medica Subject Headings [33], Traditional Chinese Medical Subject Headings [34], Chinese Library Classification (4th ed.) [35], National Standard [36–38] and Pharmacopoeia of the People’s Republic of China [39], etc.

Based on these existing controlled vocabularies and terminologies, manual knowledge distilling and knowledge extraction are the approaches taken to concept definition and organization. A basic principle we conform to is that the three controlled vocabularies, namely Chinese Traditional Medicine and Materia Medica Subject Headings, Traditional Chinese Medical Subject Headings and Chinese Library Classification (4th ed.) are considered as the main knowledge sources of ontology, but we prefer to use National Standard and Clinical Diagnosis and Treatment Terms (e.g. Classification and codes of diseases and ZHENG of traditional Chinese medicine/GB/T 15657-1995, Clinic terminology of traditional Chinese medical diagnosis and treatment—diseases, syndromes, therapeutic methods/GB/T 16751.1-1997, GB/T 16751.2-1997, GB/T 16751.3-1997) when the terminology definitions of the above three controlled vocabularies are conflicts with those of the two terminological systems, and then the final definitions are defined by the participant domain experts and nomenclature committee. The translation is done from those sources to the TCM ontology by building the relations between terms. Sixteen medical information institutes or medical libraries joined the research group to establish the principles and rules for TCM ontology development, as well as to build the ontology. The translation from the sources to the TCM ontology was done according to the following principles. The relationships between terms were built based on the concepts. Different terms from various sources with the same concepts were connected by this way. The synonyms in different forms were translated by the system into the corresponding subject headings. All the terms with same concepts were selected from the sources first, and then the relationships between terms were built. The nomenclature committee and experts defined the subclasses of each category in TCM ontology. There were some intensive debates. For example, there were 12 subclasses under the category of the basic theory of traditional Chinese medicine on the first draft of the structure, but some of experts did not agree to this classification. After the discussion, a new structure with six subclasses was developed. The UTCMLS project is still in progress. The core top-level categories such as the concept relevant categories, semantic type and semantic relationship, and the first level sub-class definitions of 14 essential sub-ontologies are currently finished. Furthermore, the complete ontology definitions and knowledge acquisition of some sub-ontologies (e.g. the basic theory of TCM, acupuncture and formula of herbal medicine) also
Figure 4  The core top-level categories of TCM ontology. In which: (1) is the highest-level ontology has six basic slots and 14 sub-ontologies; (2), (3), (4), (5) constitute the concept system; (6), (7) form the semantic system; and (8) is a simple ontology of reference book.
have been completed. This section provides the current main results and experience of the ontology development by introducing the whole framework of the skeleton top-level categories of TCM ontology and the semantic types. Fig. 4 shows the whole framework of the skeleton top-level class definitions of TCM ontology in Protégé 2000 (RDFS is the underlying knowledge representation language and storage format).

6.1. The core top-level categories

To unify, and initiate the whole ontology development and knowledge acquisition, we have defined the core top-level categories of TCM ontology with essential instances (e.g. 104 TCM semantic types and 59 semantic relationships). The semantic types and relationships are depicted in detail in Section 6.2. We provide six core top-level categories, which are treated as the instances of the metaclass: STANDARD-CLASS in Protégé 2000, and on the basis of them define all the intentional and extensional content of concepts in UTCMLS. Meanwhile, 14 sub-ontologies with first level sub-classes are defined, but the second or deeper level sub-class definitions are mainly determined by the corresponding groups of domain experts who take change of the knowledge acquisition work, because the bottom-up method is used to facilitate the knowledge acquisition and to decrease the labor efforts. As is shown in Fig. 4, three basic components constitute the core top-level TCM ontology.

6.1.1. Sub-ontologies and the hierarchical structure

The sub-ontologies and their hierarchical structure reflect the organization and content of TCM knowledge. The initial 14 sub-ontologies (Table 1 illustrates the definitions of sub-ontologies and their contents) and the six basic slots have been defined ((1) of Fig. 4). The six basic slots are concept name, concept definition, concept interpretation, relevant term, interrelated concept and semantic type. Concept name slot gives the standard name of a concept, and the concept definition slot and interpretation slot give the descriptive content about the meaning of a concept. The relevant term slot defines the different terms of a concept used in the other relevant vocabulary sources, by which we can construct the relations between UTCMLS and other terminological sources. The interrelated concept slot defines the concepts, which has some kinds of semantic relationships with a concept. The semantic type slot gives the semantic type definition of a concept. More slots can be defined for an individual sub-ontology if necessary. The complete sub-ontologies with concept instances will become a medical language knowledge base system. Now there are 8000 concepts (e.g. herbal medicine, chemistry of herbal medicine and disease, etc.) and 50,000 concept instances (e.g. Rhubarb, Rhein and diabetes, etc.) in UTCMLS.

6.1.2. Concept structure

Using Protégé 2000, we provide a knowledge representation method to define the concept in a unified mode. We consider that every TCM concept consists of three basic intentional attributes, namely definition, interpretation and name, hence three classes, namely concept definition, concept name and concept interpretation (2), (3), (4) in Fig. 4) are defined to construct a terminological concept. The class of concept definition involves the definitions of essential meanings of a concept. The class of concept interpretation gives the explanation of a concept. The class of concept name defines the synonyms, abbreviations and lexical variants of a concept, that is to say, the concept name gives the relevant terminological names of a concept from different controlled vocabularies. Together with the concept name slot of each sub-ontology, these three classes give the lexicon level knowledge of a concept. However, the semantic structure aims at the semantic level knowledge of a concept.

6.1.3. Semantic structure

The semantic type and relationship classes (6), (7) in Fig. 4) form the foundation of semantic level knowledge of a concept. The semantic types provide a consistent categorization of all concepts represented in UTCMLS, and the semantic relationships define the relations may hold between the semantic types. Classification and inference are supported by these definitions of semantic type and relationship of a concept. As Fig. 4 shows, we define semantic relationship as a slot of concept relationship class and semantic type as an essential slot of TCM ontology to assign the semantic content to concept. Each concept of UTCMLS is assigned at least one semantic type. In all cases, the most specific semantic type available in the hierarchy is assigned to the concept. The semantic structure enables to construct an abstract semantic network of all the concepts in UTCMLS, so is vital to TCM ontology. This paper will list all the TCM oriented semantic types in Section 6.2.

6.2. Semantic types and semantic relationships

The semantic network of UMLS [17] is a high-level representation of the biomedical domain based on
semantic types under which all the Metathesaurus concepts are categorized. The semantic network makes UMLS a unified medical language system, which is different from other classification and terminological system. We define the semantic structure in UTCMLS to construct the semantic level knowledge framework of TCM concept from the idea of semantic network of UMLS. We define the semantic type and semantic relationship as one of the two top-level categories of TCM ontology. The structures of semantic type and semantic relationship are defined as Fig. 4 shows. Most of semantic relationships in TCM ontology are same as UMLS, but there are five more semantic relationships in TCM ontology than UMLS, namely being exterior-interiorly related, produces, transform to, restricts, and cooperates with. Those are special ones used in the unique system of traditional Chinese medicine. For example, being exterior-interiorly related is the relationship between the concepts with zang-fu semantic type, which is a special semantic category in TCM. We have defined 104 semantic types (including 40 entity definitions and 64 event definitions) to describe the high-level
The concept categories in TCM. The 40 TCM oriented entity semantic types and the hierarchical structure are depicted in Fig. 5, and the 64 TCM event semantic types are listed in Fig. 6. All the definitions of semantic system are finally fixed by intensive debates and rigorous evaluation. However, incompleteness and imperfection is practically allowed for. The detailed definitions of semantic types and relationships, and the rules of their applications to concepts are given in the technical report [40] and also are contained in the core top-level ontology of TCM. The whole ontology will be published and shared on the Internet when finished, so this article does not give further descriptions of them.
7. Concluding remarks and future work

In the current literature on knowledge management, it is often observed that the main challenges are in the realm of human organizational culture and practices. The key to providing useful support for knowledge management lies in how meaning is embedded in information models as defined in ontologies. In China, various kinds of TCM terminology systems with differing purposes have been developed during the past two decades. Unfortunately, most of them are paper-based that cannot satisfy anymore the desiderata of healthcare information systems, such as the demand for re-use and sharing of patient data. The unambiguous communication of complex and detailed medical concepts is now a crucial feature of medical information systems. An ontological approach to the description of terminology systems will allow a better integration and reuse of these systems. Ontology development for the UTCMLS is a systematic and comprehensive procedure of knowledge acquisition and knowledge integration in TCM. TCM embodies knowledge of systematics, cybernetics and informatics, and involves a wide range of multi-discipline terminologies. It is tempting to build a unified knowledge oriented terminological system with consistent, formal and extensible structure to integrate the existing terminology systems, but it will be a huge, labor and intelligence intensive work, like the UMLS project. Sixteen TCM institutes and colleges with several hundred researchers have been involved in this project to build the ontology and to acquire the knowledge. To satisfy the requirement of distributed development, scalable and extensible methodology is focused on in TCM ontology development. The process of ontology development for UTCMLS is also a course of medical concept standardization and unification. Due to the vast knowledge storage and very much complex knowledge system involved in TCM research, the UTCMLS and ontology development is a continuous refinement procedure.

More precisely, this article presents a preliminary ontology development experience of TCM medical language system. The main efforts of this article are to provide an ontology approach to standardize TCM medical terminology. Furthermore, the TCM oriented methodology of ontology development and an ontology structure are defined to facilitate the development of UTCMLS KBS. According to the knowledge system of TCM, two sub-systems (e.g. concept system and semantic system) are defined to describe and construct the terminological system. The structure of the semantic system is inherited from the semantic network of UMLS, but many more semantic types (about 104 additional semantic types) are defined to reflect the essential of TCM science. Four core top-level categories, namely concept definition, concept name, concept interpretation and concept relationship constitute the concept system, which build a unified approach to define the TCM concept. We aim to build a unified ontology framework to effectively organize and integrate the terminological knowledge of TCM. Although the core top-level categories and essential ontology structures are defined, much more work (e.g. terminology standardization, problem-solving methods and knowledge acquisition) should be done to construct a final version of TCM ontology and medical terminology system, which will support various medical applications.

Now the applications of UTCMLS such as concept-based medical information retrieval and TCM specific semantic browser (exploring the information and structure of UTCMLS online) are in progress. Future work on UTCMLS includes: (1) completion of TCM ontology and expansion of the sub-ontologies; (2) completion of knowledge acquisition to build a final medical language system; (3) interrelating to the existing medical database sources; (4) development of a TCM semantic web terminology server; and (5) development of healthcare information systems based on UTCMLS terminology server.

Acknowledgements

Our special thanks go to Dr. Yongsheng Gao (Salford Health Informatics Research Environment, School of Health Care Professions, The University of Salford, Great Manchester, UK) who has read the whole draft and provided valuable comments. We are also appreciative of comments by the three anonymous reviewers, which prompt a considerable revision and improvement to the final article. We gratefully acknowledge all the domain experts for their collaboration and hard work. We are also grateful to all the members of the UTCMLS project for the valuable suggestions and innovative discussions. This research is partly supported by National Basic Research Priorities Programme of China Ministry of Science and Technology under grant number 2001DEA30039.

References

Ontology development for unified traditional Chinese medical language system


