Abstract—The World Wide Web connects a wide variety of communities in medical domain, and provides a platform for knowledge exchange and integration between Traditional Chinese Medicine (TCM) and biomedicine. However, the cultural gaps between TCM and Western Medicine hinder cross-cultural communication. We utilize Semantic Web technologies to build a knowledge base that integrates distributed and heterogeneous knowledge elements from both biomedicine and TCM, and provides various information retrieval and knowledge discovery services. We explain how this integrated knowledge base helps to bridge the linguistic, semantic, and ontological gaps between different communities, and fosters cross-cultural scientific collaboration.

Keywords—traditional Chinese medicine; ontology; semantic web; knowledge base; knowledge integration; knowledge services

I. INTRODUCTION

Ancient civilizations have independently developed their own medical systems, such as Traditional Chinese Medicine (TCM), and Ayurveda (Traditional Indian Medicine) [1]. Scientists in biomedicine have been increasingly interested in the knowledge elements of Traditional Medicine, and progressively integrated them into their knowledge base [2]. This trend is commonly known as the Integrated Medicine (IM) [2][3]. Integrated studies include herb-inspired drug discovery [4], scientific interpretation of TCM theory and practices (such as TCM diagnoses [5] and herbal formulae [6]), and evidence-based Chinese medicine [7]. It is a multidisciplinary endeavor involving diverse participants, and requires different pieces of knowledge and best practices to be published and accessed on a global basis.

In the past ten years, TCM practitioners have made efforts to computerize TCM knowledge assets for clinical decision-making, drug discovery, and education. A series of seminal works have made progress in computer-mediated knowledge sharing and integration for TCM, such as ontology engineering [8], heterogeneous data integration [9], knowledge representation [6], knowledge management [7], and knowledge discovery [5]. For example, China Academy of Chinese Medical Sciences (CACMS) has constructed over 70 databases and a variety of Web applications that provide knowledge and information services for TCM practitioners [9]. In our experience, collaboration and knowledge sharing between TCM and Western Medicine is beneficial and can be supported by information technologies such as databases and the Internet.

To illustrate the benefits of integrated medicine, we present a case study in which herbal therapies are analyzed to inspire modern drug discovery for the Alzheimer’s disease (AD) [4]. Here, an analyst discovers that TCM practitioners frequently use the herb Huperzia Serrata (HS) to treat the TCM Syndromes that are associated with the symptom Memory Loss, which leads to the hypothesis that “HS seems to affect Memory Loss”. However, scientific validation for the hypothesis goes beyond the scope of TCM, so that the hypothesis is published to solicit evidences. Scientists from such disciplines as pharmacetics, biomedicine, or chemistry can then publish evidences to prove/falsify this hypothesis. For example, a thread of evidences that support the original hypothesis can be: (1) HS contains the bioactive chemical Huperzine A; (2) Huperzine A is an inhibitor of the enzyme cholinesterase, which is a drug target for Alzheimer’s disease; and (3) Alzheimer’s disease is often associated with Memory Loss. To support this genre of analysis, complementary knowledge elements should be integrated from distributed sources so that their relations can be discovered and reasoned. This case study illustrates that practitioners from both TCM and biomedicine can collaborate to explore the knowledge base of TCM, yet also reveals three levels of cultural gaps among these participants.

1) Linguistic Gap: While biomedical communities achieve a consensus to use English as the common language of discourse, TCM practitioners largely express their knowledge and evidences in native languages. A unified mechanism is needed to represent multilingual knowledge bases that are accessible to persons from different cultural backgrounds.

2) Semantic Gap: Practitioners of TCM and modern biomedicine may interpret concepts with different meanings. The same concept may have different meanings in different contexts. For example, the “Kidney” in TCM refers to a different organ than the same concept in Biomedicine refers to. On the other hand, different concepts from different domains may have roughly the same meaning. For example, TCM refers to diabetes as Wasting-Thirst syndrome. A unified mechanism is needed to differentiate and align the concepts that are used in different domains.

3) Ontological Gap: Practitioners of TCM and modern biomedicine have different world-views in terms of how the
human body functions (or malfunctions) in an environment and what interventions should be made. We refer the difference in world-views as the ontological gap. For example, the TCM practitioners are aware of the efficacy of the herb, Huperzia Serrata (HS), in aging disorders, and interpret the Mechanism of Action (MOA) of this herb as "strengthening the kidney". Biomedical scientists analyze this evidence, and deduce that a bioactive compound of the herb HS can serve as a potential therapy for the Alzheimer’s disease (which is targeted at brain instead of kidney). Therefore, a unified mechanism is needed to specify the domain in which a knowledge element is originated from and applied to.

These cultural gaps limit the effectiveness of communication between practitioners of TCM and biomedicine. To bridge these cultural gaps, we need to establish a Web-based infrastructure that can support the exchange, publishing, and integration of knowledge in integrated studies, in order to extend the actionable knowledge shared among communities. We aim to utilize Semantic Web technologies to address these challenges. In this paper, we will introduce the Semantic Web technology, and explain how the Semantic Web is applied in ontology engineering, knowledge integration, knowledge discovery for integrated medicine.

II. OVERVIEW OF THE SEMANTIC WEB TECHNOLOGY

The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries [10]. The Semantic Web builds on a language specification called Resources Description Framework (RDF) [11]. In this framework, a resource is any object (such as a concrete thing, an abstract entity, and a generic concept) that has a Uniform Resource Identifier (URI). A Knowledge Base (KB) is composed of a set of RDF Graphs, and each RDF graph is a set of statements in the form of Subject-Property-Object triple. Subjects are in practice (though not restricted to) resources, and Objects can be resources or literals. Properties define binary relations between two resources or between a resource and a literal. An RDF graph represents a triple with (1) a node for the subject, (2) a node for the object, and (3) an arc for the predicate, directed from the subject node to the object node. The merging of two RDF graphs is essentially the union of the two underlining sets of triples. Within a KB containing multiple RDF graphs, a member RDF graph has a URI so that its provenance information can be traced. This model gives an elegant solution to connect data from different sources, and to express complex relationships between concepts across domains.

Semantic Web languages and technologies can be used for the annotation, classification, and organization of knowledge assets and digital artifacts based on biomedical ontologies [12]. The health care and life sciences communities have taken efforts in the adoption of Semantic Web technologies [13]. Notably, the World Wide Web Consortium (W3C) established a Semantic Web interest group to focus on Health Care and Life Sciences (HCLSIG) in 2005. HCLSIG has conducted a series of successful projects including ontology engineering, data integration, and application development, which demonstrate that the Semantic Web is a feasible technical framework for knowledge representation, integration, and discovery in these fields. These projects, however, focus mainly on orthodoxy medicine, and not on Traditional Medicine. Therefore, we intend to fill this gap by using the Semantic Web to integrate the knowledge resources from both TCM and biomedicine.

III. OUR APPROACH

We utilize Semantic Web technologies to build a knowledge base that integrates the TCM knowledge elements that are distributed and heterogeneous. The Semantic Web plays an important role in unifying and linking the underlying relational databases, to enhance the interoperability, reusability, and utilization of biomedical knowledge resources. Particularly, we emphasize on making connections between TCM and biomedicine in the purpose of facilitating cross-cultural knowledge management, integration, and discovery.

Domain analysis reveals that the two independent medical systems have connecting points, such as: (1) Electronic Medical Records (EMR) that records the methods and results of integrative clinical practices, and (2) bioactive compounds from herbal medicine, which serve for drug discovery and safety analysis. We first engineer the IM domain ontology, and then use this ontology to integrate heterogeneous relational databases from both TCM and Western Medicine, and to support various Web-based information retrieval and knowledge discovery applications.

A. Ontology Engineering

An ontology is a formal, explicit specification of a shared conceptualization for a domain [14]. An ontology can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms. Shared domain ontologies play a critical role in the sharing and integration of knowledge across interrelated communities.

As a major component of the Semantic Web, the Web Ontology Language (OWL) is a standardized language for defining and instantiating Web ontologies [15]. An OWL ontology may include descriptions of classes (such as “Drug”, “Herb”, and “Disease”), properties (such as “treats”) and instances (such as “Huperzine A”). We designed and engineered a high-level ontology for integrated medicine, including concepts from both TCM and biomedicine.

We will describe this ontology by explaining some basic concepts within it. This ontology start with the most basic concept that is the same for both TCM and biomedicine: the patient. The class “Patient” is defined as the subclass of the class “Person”, and the class “Person” is reused from the FOAF ontology. A patient typically has clinical conditions that are recorded in clinical cases. Accordingly, the ontology defines classes such as “Clinical Condition” and “Clinical Case”: the class “Clinical Condition” is defined as the subclass of the
class “Concept” in the SKOS ontology; and the class “Clinical Case” is defined as a subclass of the FOAF class “Document”.

This ontology also defines properties as binary relations between things, and specifies their domain and range. For example, we can represent the assertion that “to treat is for something is to counteract a clinical condition” by defining the property “treats”, and declaring that the property “treats” has the domain of “Thing”, and the range of “Condition”. In addition, we define the property “has” as a binary relation from things to things, and the property “isSubjectOf” as a binary relation from things to documents.

In addition to declare classes and properties, the ontology also includes their formal, logical definitions. For example, we can define the class “Patient” as a person who has some conditions (as is observed in a clinical context) as is recorded in some clinical cases. We translate the above definition of patient into OWL class “Patient” as follows:

Person and has some Condition and isSubjectOf some Case.

Also, TCM Syndromes are a set of concepts describing morbid conditions of human body, which are different from diseases in biomedicine [5]. To model this fact, we assert that the class “TCM Syndrome” and “Disease” are both a subclass of the class “Condition”, and they are disjoint from each other. In addition, a “Drug” can be defined as a thing that treats some clinical conditions, and “TCM Herb” can be defined as subclass of “Drug”.

In summary, we have briefly described the content of the high-level ontology, and this ontology contains more classes and properties that cannot be fully mentioned in this paper. This high-level ontology enables domain experts to collaborate in ontology engineering, resulting into a large-scale Semantic Web ontology that includes more than 5,000 resources (including classes, properties, and individuals). This ontology serves as the foundation for Semantic Web applications such as semantic integration of disparate databases, ontology learning for information extraction, knowledge reference, and clinical decision-support.

B. Semantic Encyclopedia

We developed a semantic wiki tool named DartWiki, to build ontology-based digital encyclopedia for the biomedicine domain [12]. Next, we use the DartWiki tool to build a “semantic encyclopedia” for integrated medicine, which enables users to look up and edit bilingual descriptions for medical concepts. The major role of the encyclopedia is to store and manage Human-Friendly Annotations (HFAs), which are RDF literals that are used to describe concepts for humans to comprehend. The knowledge base contains the labels and the literal comments of the things, and also the documents in which the information or the definitions of the things can be found. A HFA is represented as a plain RDF literal, which has a lexical form and optionally a language, normalized to lowercase.

The vocabulary of RDF Schema [16] is used to annotate entities with different types of HFAs. For example, “rdfs:label” is an instance of “rdf:Property” that may be used to provide a human-readable version of a resource’s name. We can use “rdfs:label” to annotate the concept “Huperzine A” with two labels (Chinese and English).

The semantic encyclopedia provides a scalable platform for domain experts to edit the description about any domain concepts. Whereas a shared domain ontology and a semantic encyclopedia is the crucial first step towards multilingual knowledge integration, we also need to integrate the massive factual data (clinical cases, clinical trials, and experimental data) into the integrated knowledge base.

C. Semantic Integration of Relational Databases

Based on the ontology, we build a knowledge base through semantic integration of knowledge resources, such as clinical cases, clinical trials, and various databases around the topics of “formulas”, “drugs”, “syndromes”, and “diseases”, etc. This knowledge base can be conceptualized as a set of interlinked RDF graphs. This “multi-graph” connects the TCM data with biomedicine data, and enables the discovery of semantic relations across database boundaries.

The biomedical knowledge resources, viewed as graphs in the conceptual level, are physically stored in relational databases. In order to publish and interconnect these knowledge resources, we need to map relational databases to the Semantic Web layer. In our approach, a semantic mediator is used to map relational schemas to a shared domain ontology. Each relational table is defined as a semantic view over the ontology [9]. A semantic view defines the mapping between a relational table and an ontology, in the form of “head := body”, with the “head” as a relational predicate, and the “body” as a set of RDF triples. A semantic view defines the semantics of a relational predicate from the perspective of the ontology. In general, we have the following mapping rules:

1) Map a table to a class;
2) Map a column to a property;
3) Map a row to an individual (or an instance).

A visualized tool was developed for defining semantic views. This tool utilizes visualization features to simplify the operations and to deliver vivid user experience. The database administrators can specify table-to-class mappings in an interactive manner aided by the intelligent schema reasoner. Given table-to-class mappings, the reasoner infers potential table joins based on semantic associations inferred in the ontology. If two tables have foreign key relationships in relational schema, the reasoner is able to define the semantics of this join automatically. Class-level matching can also be inferred based on instance-level similarities. This tool can improve efficiency and reduce errors compared with editing configuration files. The generated views can be deployed in a query-rewriting engine via a semantic registration service and take effect in the query-rewriting process.
The SPARQL (Simple Protocol and RDF Query Language) language is the query language on the Semantic Web [17]. It is informally a graph pattern involving a set of variables, the class and property constraints for these variables, and the relationships between these variables. The query-rewriting engine translates a SPARQL query into a series of SQL queries against underlying relational schemas, based on mapping rules derived from semantic views.

In a practical application, an ontology can be mapped into multiple distributed and heterogeneous databases, and therefore the rewriting engine needs to translate a SPARQL query into a set of SQL queries against related databases. Next, the engine dispatches the SQL queries into specific databases, merges the results of SQL queries and transforms the results back to semantically-enriched format. This ontology-based database integration capability supports intelligent information retrieval services, such as query, search, and navigation, based on one virtual distributed database.

D. Intelligent Search portal

We have developed an intelligent search portal based on the integrated knowledge base. This portal supports interactive discovery of biomedical knowledge across database boundaries. Users can fluently switch between the paradigms of query, search, and navigation based on shared domain ontology. In response to a keyword search, the portal displays content entries in the middle, and a recommended list of synonyms and associated concepts on the right. Using the classes and concepts in the searching results, the user can start to specify an accurate semantic query. Disparate pieces of information about the same concept or topic are synthesized. Based on the semantic associations defined at the ontological level, the user can keep searching and navigating information and knowledge unaware of database boundaries. In summary, the portal provides three major paradigms: Semantic Search, Semantic Query, and Semantic Navigation, which are explained as follows:

1) Semantic Search: the system accepts one or more keywords, performs a thorough content search on the Semantic Web, and presents searching results as a logical chain of evidences. The TCM knowledge is not organized as explicitly-linked Web resources, which makes link-based analysis not applicable for searching the TCM knowledge, so we utilize the TCM domain ontology to improve the quality of search. The multimedia content, such as textual documents, images, and reporting tables, are annotated and indexed in terms of the ontology. The semantic associations of concepts are analyzed to determine the relevance and importance scores of concepts and content entries.

2) Semantic Query: the user interface facilitates interactive and dynamic generation of the SPARQL query. The system guides users through specifying and refining a query in an intuitive way. First, users can specify interesting classes by navigating on a graphical view of domain ontology. Second, a query form corresponding to property definitions of each selected class will be automatically generated and displayed. Third, users can check and select interesting properties and input query constraints into the text boxes. Finally, the constructed semantic query is translated into SQL queries based on the mapping rules derived from semantic views.

3) Semantic Navigation: the user interface enables users to navigate knowledge on the semantic level. For example, the association between a patient’s disease and a therapy that cure the disease can be critical for a clinical decision-making. However, the patient’s EMR and the document for the therapy may not have explicit links or shared keywords. In our approach, semantic links are inferred and ranked based on the ontology together with other knowledge resources. Semantic links are used to enhance the connectivity of the knowledge base. A user can navigate the integrated knowledge by following semantic links. For example, a thread of navigation can be from AD to one of its symptoms (Memory Loss), then to a related TCM Syndrome (Kidney Yang Deficiency, KYD), then to a TCM herb (Huperzia serrata, HS) used to treat the syndrome, then to a Plant-Derived Chemical Huperzine A. The user can then analyze the mechanism of action of Huperzine A, and find that Huperzine A is an inhibitor of the enzyme cholinesterase, which is a drug target for AD. The user can then find out related evidences from Clinical Trials and Electronic Medical Records to make the right decision.

IV. Conclusions

Most works in the Semantic Web HCLS community focus on the standardization and integration of knowledge elements in biomedicine, for the major purpose of accelerating scientific progress. In this work, we aim to broaden the perspective to cover various forms of traditional medicines, in order to retain the balance between accelerating scientific progress and preserving cultural heritage. We put a special emphasis on bridging various gaps between TCM and biomedicine in order to facilitate cross-cultural communication and interdisciplinary knowledge discovery. For example, we provide a semantic encyclopedia system for users to edit multilingual labels and annotations, in order to bridge the linguistic gap. Next, we provide a unified ontology platform for users to define concepts and their semantic relations, in order to bridge the semantic gap. Finally, we provide an intelligent search portal that provides knowledge discovery services to users with different backgrounds, in order to bridge the ontological gap.

This study reveals that the challenges for knowledge integration across sub-domains of biomedicine go beyond technological realm and lie in the creation of shared culture among different experts and communities. The first step is for scientific communities to recognize TCM as an independent medical system, not as complementary and alternative methods. Then we can build up ontologies and knowledge bases for this medical system and publish them as Web resources. Finally, we can make semantic links between these Web resources based on the results from integrative studies. These efforts are crucial to realize the vision of integrated medicine, in which knowledge from different medical systems can be integrated for the benefits of human health. The Semantic Web, by enabling the publication of domain ontologies and data for sharing and reuse on a global
basis, provides us with the necessary toolkit to realize this vision. In the future, the methodology and technologies that are used in this application might be applied in the preservation and utilization of other medicinal or cultural systems.

ACKNOWLEDGMENT

This work is supported by “the Fundamental Research Funds for the Central public welfare research institutes (NO. ZZ080324, NO. ZZ090305, NO. ZZ070804, NO. ZZ090301).”

REFERENCES


