CHINESE MEDICINE

EDITORIAL Open Access

# Semantic Web for data harmonization in Chinese medicine

Kei-Hoi Cheung[1*], Huajun Chen[2]

**Abstract**

Scientific studies to investigate Chinese medicine with Western medicine have been generating a large amount of data to be shared preferably under a global data standard. This article provides an overview of Semantic Web and identifies some representative Semantic Web applications in Chinese medicine. Semantic Web is proposed as a standard for representing Chinese medicine data and facilitating their integration with Western medicine data.

## Background

As the scientific evidence for the preventive and therapeutic efficacy of Chinese medicine (CM) is growing, it is strongly demanded to bridge CM with Western medicine (WM), particularly through the data obtained from biomedical and clinical research. For example, there were acupuncture studies on certain diseases/disorders such as chronic pain [1,2] and cerebral palsy [3], on pharmacological, molecular and therapeutic properties of various Chinese herbs [4,5] using high-throughput technologies such as DNA microarray and mass spectrometry [6,7]. Technical challenges include not only the increasing amount of CM literature but also the wide variety of data among various databases. Some representative databases are as follows:

i) TCMGeneDIT [8] is a database containing disease-gene-herb associations as the results of mining the biomedical literature;
ii) Phytochemical databases of Chinese herbal constituents were constructed [9];
iii) ClinicalTrials http://clinicaltrials.gov/ contains information of a large collection of clinical trials including those that involve CM;
iv) MedlinePlus http://www.nlm.nih.gov/medlineplus/[10] developed by the United States National Library of Medicine provides consumers and health professionals with research information which covers certain herbal supplements;
v) TCM Online http://cowork.cintcm.com/engine/windex1.jsp consists of over 40 categories of CM Databases such as the Traditional Chinese Medical Literature Analysis and Retrieval Database, Clinical Medicine Database, Traditional Chinese Drug Database, Database of Chinese Medical Formula, Traditional Chinese Medicine Enterprises and Productions Database, and State Standards Database.

Data mining and integration of CM and WM databases are of great value but problematic [9,11]. Data mining and integration problems include heterogeneity in data formats and structures as well as a lack of standard terminology. Cultural and linguistic differences further complicate data integration. In the informatics community, methods developed for data integration can be categorized into: (1) data warehousing to translate data and (2) query federation to translate query. Both approaches have their pros and cons. For example, the data warehousing approach has good query performance as data are queried locally, but data are not always up-to-date (data updates are to be made periodically to keep the warehouse in synchrony with the member data sources). The query federation approach guarantees data to be up-to-date, but it may suffer from query performance especially when large volumes of data are queried and joined over the network. Despite their differences, these approaches are based on a common data model. The use of such a model is feasible in either a single enterprise or a small group of enterprises. A common data model which can overcome national, geographical, and cultural boundaries would be different without a global data representation standard. To this end,

* Correspondence: kei.cheung@yale.edu
[1]Yale Center for Medical Informatics and Departments of Anesthesiology and Genetics, School of Medicine, Computer Science Department, Yale University, New Haven, CT 06510, USA

Semantic Web [12] has the potential to help realize data harmonization in CM.

## Semantic Web and its applications in Chinese medicine

Semantic Web is an evolving extension of the World Wide Web in which the semantics of information and services on the Web are defined, making it possible for the Web to "understand" and answer queries in accordance with the Web content. SW's enabling technologies include the Uniform Resource Identifier (URI) http://www.w3.org/Addressing/ and Resource Description Framework (RDF) http://www.w3.org/RDF/, which are the Semantic Web standards for data identification and data representation respectively. The RDF provides a "triple" format for representing a statement that consists of a subject, property and object. Each component of the triple is identified by a URI that serves as a global unique identifier for the Web. For example, the following triple (statement) asserts that an herb-derived drug "Huperzine A" (subject) "inhibit" (property) "NMDA receptor" (object).

**Subject** – http://en.wikipedia.org/wiki/Huperzine_A

**Property** – http://en.wikipedia.org/wiki/inhibit

**Object** – http://en.wikipedia.org/wiki/NMDA_Receptor

The above example demonstrates that the Wikipedia URIs are used to identify and define the subject, property and object (this is only for demonstration purposes). The statement indicates an "inhibitory" effect of the drug "Huperzine A" on "NMDA Receptor" (drug target). A collection of linked RDF statements forms a directed acyclic graph (DAG). Such collections of statements represent the knowledge of a domain. To query and manipulate RDF statements, we may use "SPARQL" http://www.w3.org/TR/rdf-sparql-query/, which is the RDF query language standard. SPARQL is analogous to SQL http://en.wikipedia.org/wiki/SQL for querying relational databases.

To capture richer data semantics to support computational inference and reasoning, the RDF Schema (RDFS) http://www.w3.org/TR/rdf-schema/ and the Web Ontology Language (OWL) http://www.w3.org/TR/owl-features/ have been used to encode ontologies in the biomedical domains [13,14]. RDFS provides the *rdfs:Class* construct to declare a resource as a class, e.g. Herb. A hierarchy of classes can be defined using the *rdfs:subClassOf* construct. For example, "Huperzia serrata" is a subclass of "Herb". Most of the RDFS components are included OWL, which is more expressive than RDFS. OWL has the built-in property *owl:sameA*s that allows a synonymous relationship between two classes (e.g. "Huperzine A" and "Huperzia serrata"). Cardinality constraints can be applied to properties (e.g. the "inhibit" property can have a minimum cardinality of one and cardinality with a maximum of a positive integer). While OWL is semantically richer than RDF or RDFS, it can be expressed using the RDF syntax. OWL reasoners such as Pellet [15] and Racer [16] can be used to make inferences out of OWL ontologies.

Adoption of the Semantic Web has been significantly important to health care and life sciences. In part, the adoption has been driven by the World Wide Web Consortium (W3C), which launched the Semantic Web for Health Care and Life Sciences Interest Group (HCLS IG) http://www.w3.org/2001/sw/hcls/. The group has been chartered to develop, adopt, and support the use of Semantic Web technologies and practices to improve collaboration, research and development in health care and the life sciences.

As RDF/OWL-formatted datasets are growing in terms of the number and size , efficient data storage and manipulation become big issues. To this end, a variety of triplestore technologies have emerged, including Virtuoso http://virtuoso.openlinksw.com/, Oracle http://www.oracle.com/technology/tech/semantic_technologies, AllegroGraph http://agraph.franz.com/allegrograph/, and Sesame http://www.openrdf.org/. While some of these technologies (e.g. Oracle and Virtuoso) are proprietary, others (e.g. Sesame) are open source. Some of them (e.g. Virtuoso, AllegroGraph and Sesame) support SPARQL, but some others (e.g. Oracle) have their own RDF query languages. To provide a uniform query access, many triplestores provide a so-called "SPARQL endpoint" so that queries can be issued by client programs against the triplestores via the SPARQL language. For example, even though Oracle does not support SPARQL internally, it can be configured to provide an external SPARQL endpoint through the Jena adaptor http://www.oracle.com/technology/tech/semantic_technologies/htdocs/documentation.html. Triplestores such as Oracle provide their own native OWL reasoners, while some others (e.g., Sesame) can be integrated with external reasoners.

Linked Data [17] is a new method of exposing, sharing, and connecting data via dereferenceable HTTP URI's on the Semantic Web. A dereferenceable HTTP URI serves as both an identifier and a locator. The key idea is that useful information should be provided to data consumers when its URI is dereferenced. Using the Linked Data approach, not only do data providers make their data available in the form of RDF graphs, but data linkers can also create new RDF graphs that consist of links between independently developed RDF graphs provided by different sources. Examples of Linked Data, e.g. DBpedia http://wiki.dbpedia.org/OnlineAccess, are listed on Linking Open Data (LOD) http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData. A similar effort has been launched by the Linking

Open Drug Data task force of the HCLS IG to use the linked data approach to link drug-related data.

As the relational database technology is prevalent in the health care and life science domains, many of the CM databases are currently in the relational format. While these relational databases serve the specific needs of individual labs or institutions, their accessibility by other labs or institutions is limited. An object or data record is identified by a unique identifier (primary key) that is local to the database. In other words, the same identifier does not identify the same object (data records) in different relational databases. Another issue with the relational databases is that relationships are defined based on links between primary and foreign keys. These links are not to convey some meaning semantically. Semantic Web can be used to address this problem by allowing a semantic layer to be created on top of existing relational databases. Semantically rich queries (based on meaningful relationship names) can be formulated at the semantic layer (built using the Semantic Web technology) and then be mapped to the local queries against the underlying relational databases. DartGrid [18] is a system demonstrating the use of this semantic web approach to integrate CM databases. The advantage of this approach is that existing relational databases and applications accessing these databases need not be abandoned, while new powerful applications can be developed to make use of the Semantic Web features.

As data are increasingly available in RDF/OWL format, new warehouses and federated query systems have been built from scratch using Semantic Web technologies to allow direct access by programs. As part of the HCLS IG effort, a subset of TCMGeneDIT was converted into RDF format and loaded into an RDF triplestore [19]. In addition, the BioRDF task force of the HCLS IG has undertaken the effort of implementing query federation using the Semantic Web [20].

Ontologies encoded by Semantic Web enable expressive knowledge representation, integration, and discovery. Ontology research is active in the biomedical informatics community. Examples include the OBO Foundry [21] and BioPortal [22] that provide access to a large collection of biomedical ontologies. These ontologies are relevant to CM research especially when relating CM to WM. In addition, efforts have begun to create new ontologies specifically for CM. For example, China Academy of Traditional Chinese Medicine has created a CM ontology that defines more than 8,000 classes and over 50,000 instances and may help integrate heterogeneous and disparate databases [23].

Some information technologies such as text mining, Grid computing, and Web services have been using the Semantic Web. These technologies combined with the Semantic Web can further empower CM researchers to carry out *in silico* research.

## Discussion

Given the long history of CM, most of the CM documents were written in Chinese. While the Web is multilingual, a simple literal translation, however, is not sufficient in terms of making the CM knowledge accessible by Western researchers. An example is the translation of signs and symptoms between CM and WM. For example, the term *Re* (which literally means "Heat") in CM may be referred to as high fever and irritability in WM. The theories behind WM and various CM can be fundamentally different, leading to the difficulty to make alignments among their domain ontologies. For example, CM practitioners interpret human body and organs based on Chinese philosophical ideas of "yin-yang" and "five-elements". They are aware of the efficacy of the herb, Huperzia serrata (HS), in aging disorders, and interpret the action mechanism of this herb as strengthening the *Shen* (kidney). Biomedical scientists analyze some experimental evidence, and deduce that a compound of the herb HS acting on the brain can serve as a potential therapy for the Alzheimer's disease. In this case, HS targets the brain (WM) instead of the *Shen* (kidney).

These language gaps limit the communication and interaction between WM and CM in both directions. On the one hand, scientific communities have not reached the full potential of utilizing CM knowledge. On the other hand, best practices of WM are not widely adopted in the regions where CM is predominant form of healthcare service. To bridge these gaps, we need to establish an infrastructure that can support communication and collaboration in integrative medicine studies. The infrastructure should also be able to capture and publish the results of these integrative medicine studies to extend the actionable knowledge shared among communities.

Data sharing is a key to advancing science in the digital age [24]. For example, the Human Genome Project [25] made public release of data to the scientific community. This open access culture should be widely encouraged and supported by the CM community. At the same time, we need to address the concerns of sharing data. Among these concerns is the intellectual property including data ownership, attribution, and licensing. The legal complication should never be underestimated, as the laws affecting data sharing vary from one country to another. The Consortium for Globalization of Chinese Medicine http://www.tcmedicine.org/ was formed to promote data sharing as well as collaboration among academia, industry and regulatory agencies in various countries.

While the Semantic Web is a candidate for standardizing the format of CM data sharing, it needs to be used in conjunction with other standardization efforts that are underway in the CM community, e.g. the regulatory standards for quality control of Chinese medicinal materials [26]. This also brings up the question of how much information needs to be provided for describing different types of CM data for reproducibility, quality, and safety purposes. In the fields of genomics and proteomics, standards such as MIAME [27] and MIAPE [28] are available for specifying the minimum amount of information to be provided for microarray experiments and proteomics experiments, respectively. Similar standards are needed for sharing scientific data in CM.

There is a broad spectrum of international Semantic Web research related to the health care and life sciences. Semantic Web research effects in CM are mainly in Asia. It would be beneficial to integrate CM into these international activities. More use cases are needed to demonstrate how the Semantic Web can be used to harmonize CM and WM through data linking and integration as well as community collaboration.

## Concluding remarks

As the interest of using Semantic Web in the health care and life sciences is growing, it has the potential to facilitate cross-disciplinary data integration between Chinese Medicine and Western Medicine. The Semantic Web could potentially play an important role in Chinese medicine informatics involving a new breed of informaticians who are able to bridge multiple scientific and cultural disciplines.

#### Acknowledgements

The work of KC is supported in part by NIH grants P01 DC04732 and R01 DA021253. We would also like to thank the editorial team of Chinese Medicine for their input and advice.

#### Author details

[1]Yale Center for Medical Informatics and Departments of Anesthesiology and Genetics, School of Medicine, Computer Science Department, Yale University, New Haven, CT 06510, USA. [2]College of Computer Science, Zhejiang University, Hangzhou, Zhejiang, 310027, PR China.

#### Authors' contributions

Both authors took part in the discussion and writing of this article. They also read and approved the final version of the manuscript.

#### Competing interests

The authors declare that they have no competing interests.

Received: 21 December 2009
Accepted: 12 January 2010 Published: 12 January 2010

#### References

1. Manheimer E, White A, Berman B, Forys K, Ernst E: **Meta-analysis: acupuncture for low back pain.** *Ann Intern Med* 2005, **142(8)**:651-663.
2. Trinh K, Graham N, Gross A, Goldsmith C, Wang E, Cameron I, Kay T: **Acupuncture for neck disorders.** *Spine* 2007, **32(2)**:236-243.
3. Sun JG, Ko CH, Wong V, Sun XR: **Randomised control trial of tongue acupuncture versus sham acupuncture in improving functional outcome in cerebral palsy.** *J Neurol Neurosurg Psychiatry* 2004, **75(7)**:1054-1057.
4. Wang R, Tang XC: **Neuroprotective effects of huperzine A. A natural cholinesterase inhibitor for the treatment of Alzheimer's disease.** *Neurosignals* 2005, **14(1-2)**:71-82.
5. Ruan CJ, Si JY, Zhang L, Chen DH, Du GH, Sun L: **Protective effect of stilbenes containing extract-fraction from Cajanus cajan L. on Aβ25-35-induced cognitive deficits in mice.** *Neurosci Lett* 2009, **467(2)**:159-163.
6. Liu S, Yi LZ, Liang YZ: **Traditional Chinese medicine and separation science.** *J Sep Sci* 2008, **31(11)**:2113-2137.
7. Zhang YB, Wang J, Wang ZT, But PPH, Shaw PC: **DNA microarray for identification of the herb of dendrobium species from Chinese medicinal formulations.** *Planta Med* 2003, **69(12)**:1172-1174.
8. Fang YC, Huang HC, Chen HH, Juan HF: **TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining.** *BMC Complement Altern Med* 2008, **8**:58.
9. Ehrman TM, Barlow DJ, Hylands PJ: **Phytochemical databases of Chinese herbal constituents and bioactive plant compounds with known target specificities.** *J Chem Inf Model* 2007, **47(2)**:254-263.
10. Schloman BF: **MedlinePlus: key resource for both health consumers and health professionals.** *Online J Issues Nurs* 2006, **11(2)**:9.
11. Ehrman TM, Barlow DJ, Hylands PJ: **Virtual screening of Chinese herbs with random forest.** *J Chem Inf Model* 2007, **47(2)**:264-278.
12. Berners-Lee T, Hendler J, Lassila O: **The semantic web.** *Scientific Am* 2001, **284(5)**:34-43.
13. Cheung KH, Qi P, Tuck D, Krauthammer M: **A Semantic web approach to biological pathway data reasoning and integration.** *Web Semant* 2006, **4(3)**:207-215.
14. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, Doherty D, Forsberg K, Gao Y, Kashyap V, Kinoshita J, Luciano J, Marshall MS, Ogbuji C, Rees J, Stephens S, Wong GT, Wu E, Zaccagnini D, Hongsermeier T, Neumann E, Herman I, Cheung KH: **Advancing translational research with the Semantic Web.** *BMC Bioinformatics* 2007, **8(Suppl 3)**:S2.
15. Sirin E, Parsia B, Grau BC, Kalyanpur A, Katz Y: **Pellet: a practical OWL-DL reasoner.** *Web Semant* 2007, **5(2)**:51-53.
16. Haarslev V, Möller R, Straeten RVD, Wessel M: **Extended query facilities for racer and an application to software-engineering problems.** *Proceedings of the 2004 International Workshop on Description Logics (DL-2004): 6-8 June 2004; Whistler, BC, Canada* 2004, 148-157.
17. Zhao J, Miles A, Klyne G, Shotton D: **Linked data and provenance in biological data webs.** *Brief Bioinform* 2008, **10(2)**:139-152.
18. Chen H, Wu Z, Mao Y, Zheng G: **DartGrid: a semantic infrastructure for building database grid applications.** *Concurrency and Computation: Practice and Experience* 2006, **18(14)**:1811-1828.
19. Zhao J, Jentzsch A, Samwald M, Cheung KH: **Linked data for connecting traditional Chinese medicine and Western medicine.** *The Sixth International Workshop of Data Integration in the Life Sciences (Poster&Demo). Manchester, UK* 2009, 13.
20. Cheung KH, Frost HR, Marshall MS, Prud'hommeaux E, Samwald M, Zhao J, Paschke A: **A journey to Semantic Web query federation in the life sciences.** *BMC Bioinformatics* 2009, **10(Suppl 10)**:S10.
21. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Consortium O, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S: **The OBO foundry: coordinated evolution of ontologies to support biomedical data integration.** *Nat Biotechnol* 2007, **25(11)**:1251-1255.
22. Musen MA, Shah NH, Noy NF, Dai BY, Dorf M, Griffith N, Buntrock J, Jonquet C, Montegut MJ, Rubin DL: **BioPortal: ontologies and data resources with the click of a mouse.** *AMIA Annu Symp Proc* 2008, 1223-1224.
23. Zhou X, Wu Z, Yin A, Wu L, Fan W, Zhang R: **Ontology development for unified traditional Chinese medical language system.** *Artif Intell Med* 2004, **32(1)**:15-27.
24. Anonymous author: **Data's shameful neglect.** *Nature* 2009, **461(7261)**:145.
25. Cantor CR: **Orchestrating the human genome project.** *Science* 1990, **248**:49-51.
26. Chan K, Leung KSY, Zhao SS: **Harmonization of monographic standards is needed to ensure the quality of Chinese medicinal materials.** *Chin Med* 2009, **4**:18.

27. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FCP, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M: **Minimum information about a microarray experiment (MIAME) - toward standards for microarray data.** *Nat Genet* 2001, **29**:365-371.
28. Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK, Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW, Dunn MJ, Heck AJR, Leitner A, Macht M, Mann M, Martens L, Neubert TA, Patterson SD, Ping P, Seymour SL, Souda P, Tsugita A, Vandekerckhove J, Vondriska TM, Whitelegge JP, Wilkins MR, Xenarios I, Yates JR, Hermjakob H: **The minimum information about a proteomics experiment (MIAPE).** *Nat Biotechnol* 2007, **25(8)**:887-893.

doi:10.1186/1749-8546-5-2
**Cite this article as:** Cheung and Chen: **Semantic Web for data harmonization in Chinese medicine.** *Chinese Medicine* 2010 **5**:2.