# Text-based construction and comparison of domain ontology: A study based on classical poetry

**Chu-Ren Huang**
**Academia Sinica**

1. Knowledge Structure: shared upper ontology and domain ontology

The fact that people from different backgrounds may have knowledge structures unlike ours is a crucial issue to be addressed in knowledge engineering. In order to become sharable and reusable knowledge, all extracted information must first be correctly situated in a knowledge structure. In addition, the situated information must be allowed to transfer from knowledge structure to knowledge structure without losing its meaningful content. This is the vision behind the Suggested Upper Merged Ontology (SUMO, http://www.ontologyportal.org) proposed by an IEEE working group. A shared upper ontology will both anchor the structured transfer of knowledge as well as set a standard for the construction of a middle and lower level ontology for each domain. This vision also has promising applications in the Semantic Web.

The most salient factors dictating variations in knowledge structures are time, space, and domain. These factors are compounded with language, which is both the product and conduit of the conceptual structure of its speakers. In order to demonstrate the felicity of the shared upper ontology approach, we need to show that it can successfully applied to comparative studies of different knowledge structures regardless of their ontological variations. We apply the shared ontology proposal to the interpretation historical texts by adopting the Shakespearean-garden approach towards construction of historical ontology.

2. The Shakespearean-garden approach towards specific ontology

The Shakespearean-garden refers to the common practice in western museums of collecting in a garden all the plants referred to in the Shakespearean texts. This garden then illustrates the flora of the Shakespearean England and will give us the context to interpret his work. In this approach, proposed in Huang et al. (2004b, 2004c), a lexicon of the targeted text, period, or domain is constructed first by segmentation and extraction of lexical items from the collected texts. Once the comprehensive lexicon of that period is collected, a lexical interface based on Sinica BOW

(http://bow.sinica.edu.tw) can be applied. It links each word to a conceptual location on the SUMO ontology, or a synset in WordNet. Since the lexicon from the text represents linguistically instantiated concepts, we use the linked conceptual nodes to construct an ontology for that text. The constructed ontology allows us to both interpret the conceptual structure of that text as well compare its knowledge with our contemporary knowledge.

2.1. SUMO: a reference ontology

The choice of SUMO as the shared reference ontology is worth noting. SUMO represents the shared knowledge structure of our current time, which is in term the sum of human knowledge accumulated through history. It is true that a contemporary ontology necessarily differ from an historical ontology. However, in order to compare the knowledge systems of two historical periods or two domains, it is necessary to have one base reference. The contemporary time seems to be the natural reference not only because this is the knowledge system under which our scientific discourse takes place. The fact that it inherits knowledge from historical ontologies also makes is an effective reference. With this reference ontology, we will be able to observe and generalize systematically which part of the knowledge is different in the specific ontology.

2.2. OntEditor: an online editor for specific ontology

An interface for online editing of specific is under construction. OntEditor will integrate available resources that include: WordNet, Sinica BOW, and segmentation tools for Chinese texts. This interface will allow user to input a domain lexicon or specific lexicl items. It will return all available information from our bilingual versions of WordNet and SUMO ontology. Lastly, it will allow automatically output of the tree representation of the specific ontology after it is constructed with verified lexical information.

3. Construction and Comparative Studies of Specific Ontologies

The two text collections studied are the 300 Tang poems (唐詩三百首) and the collection of poems by Su Shi (蘇軾詩). The ontologies constructed from both text collections allow us to compare and study the knowledge structure of two different historical periods and gain perspective understanding of the different culture and time.

3.1. Tang ontology based on Tang 300

The ontology based on the 300 Tang poems represents our first attempt at a text-based specific ontology. Two sub-lexicons from the Tang 300

Poems were extracted for domain ontology construction: animals, and plants. A total of 123 words were assigned to the three domain lexica: The animals lexicon contains 64 words; and the plants lexicon contains 59 words. These lexical items were manually mapped to SUMO ontology. When there is no direct mapping to SUMO, Sinica BOW is consulted to give the lexical item a wordnet correspondence and relational structure.

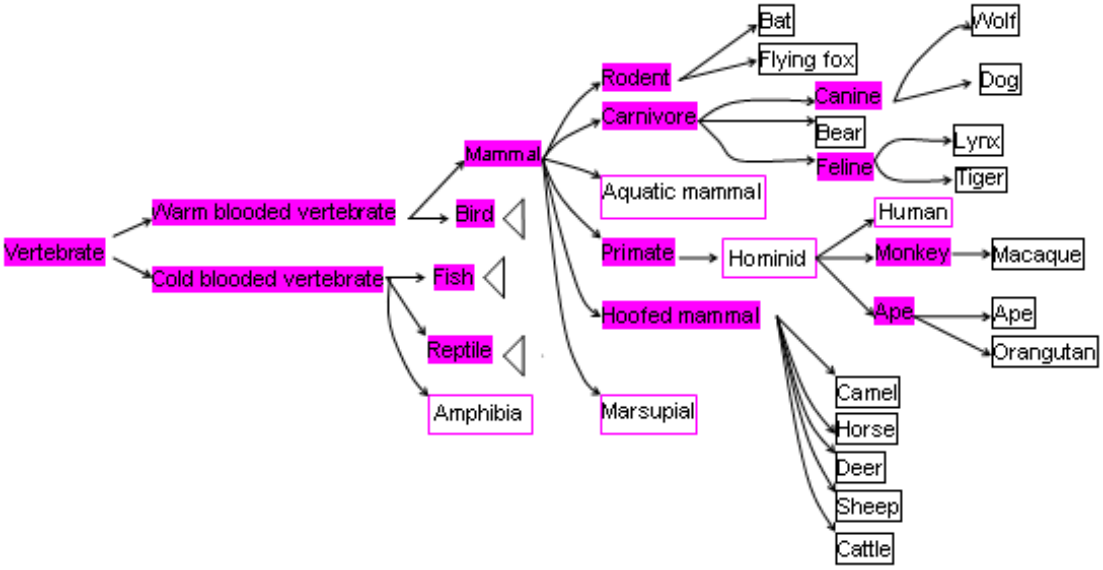


Diagram: Tang Animal Ontology

Diagram 1 gives one of the domain ontology constructed for example. It shows how specific ontology facilitates systematic comparison of knowledge systems. From this ontology, it is easy to observe that the main subclasses of vertebrates not attested in Tang 300 are: Amphibia, Aquatic Mammals, and Marsupial. We were also able to empirically support Tang's fascination with flying by showing that the most often referred to animals in the poems are flying animals: Birds in the vertebrates and insects in the invertebrates.

3.2. Ontology of Poems by Su Shi

A further study built on the foundation of the Tang 300 ontology is the ontology of poems by Su Shi that is being completed. The choice of Su Shi offers more than historical comparison. Su Shi is from the Song dynasty, almost 500 years after the Tang. The time depth allow for comparative study. The collection is also a much larger text than the Tang 300, hence offers a good test case for our new OntEditor. Lastly, Su Shi traveled extensively, and is known to incorporate the local flora and fauna into his poetry. Hence, the collection of his poems offers a much more comprehensive sampling of the

contemporary knowledge.

## 4. Conclusion

We showed that the Shakespearean-garden approach toward construction of domain ontology facilitates representation and insightful studies of different knowledge systems. We also showed that the OntEditor interface help to reduce redundant human efforts in this process. Some insights on how to determine the ontological classifications and additional insights gained from this study will also be discussed.

## Online Resources

CKIP Segmentation and Tagging Program
http://corpus.ling.sinica.edu.tw/project/LanguageArchive/lc_index.html
The Ontolgoy of 300 Tang Poems
http://bow.sinica.edu.tw/ont/ts300_ont.html
Sinica BOW
http://BOW.sinica.edu.tw/
SUMO:
http://www.ontologyportal.org/
Tender Lyrics-The 300 Tang Poems (in Chinese)
http://cls.admin.yzu.edu.tw/300/HOME.HTM
WordNet:
http://www.cogsci.princeton.edu/~wn/

## Bibliography


Fellbaum, C. (ed.), 1998. *WordNet. An Electronic Lexical Database*, Cambridge, The MIT Press.

Huang, Chu-Ren, Chang, Ru-Yng, Lee, Shiang-Bin. 2004a. Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO". Presented at the 4th International Conference on Language Resources and Evaluation (LREC2004). Lisbon. Portugal. 26-28 May, 2004.

Huang, Chu-Ren, Feng-ju Lo, Ru-Yng Chang, and Sueming Chang. 2004b. Reconstructing the Ontology of the Tang Dynasty: A pilot study of the Shakespearean-garden approach. Presented at the OntoLex 2004 Workshop. Lisbon. May 30, 2004.

Huang, Chu-Ren, Feng-ju Lo, Ru-Yng Chang, Sueming Chang. 2004c. Sinica BOW and 300 Tang Poems: An overview of a bilingual ontological wordnet and its application to a small ontology of Tang poetry. Invited talk. Workshop on Possibilities of a Knowledgebase of Tang Civilization: Towards a new comprehensive digital archive of Tang China. Institute for Research in Humanities, Kyoto University. February 20-21.

Niles, I. & Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In Proceedings of the IEEE International Conference on Information and Knowledge Engineering. (IKE 2003), Las Vegas, Nevada.

Niles, I., & Pease, A., (2001). Toward a Standard Upper Ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001). Ogunquit, Maine.