從詞彙庫到知識本體：為專業知識庫許個「語意網」的未來美景

From Lexicon to Ontology:
The Semantic Web is the future of domain knowledgebases

醫藥衛生圖書資源專題講座暨研習會
--網路時代參考服務之應用
二○○三年九月二十五日

黃居仁
中央研究院語言學研究所

一．背景:語意網的遠景

資訊參考服務的發展與時俱進。隨著全球資訊網（WWW）在短短十年內，由默默無聞到無所不在；網路也成了參考服務的主要媒介與訊息來源。其實，網路已到了蛻變的關鍵時機。全球資訊網的發明人伯納-李 (Tim Berner-Lee)於 2001 年五月「科學美國人」的專文（中譯見「科學人」2002 年 8 月號）中宣告「語意網」（Semantic Web）將在未來取代全球資訊網。面對網路的未來發展，專業的網路參考服務是否有相應的對策呢？

語意網的創新關鍵在於以「知識表達」代替檔名，做為網路間訊息交換檢索的依據。而知識內容的瞭解與互通，則要靠每個網路資源上定義知識表達架構的知識本體（Ontology）。也就是說，未來在語意網上，資源與知識檢索的對象是定義完善，可由電腦（或所謂的代理程式 agent）判讀的知識本體。因此，語意網時代的專業參考諮詢服務，需要建立在兩個重要的支柱上：第一是專業知識領域有無定義完善的知識本體，以及與知識本體配合內容完整的知識庫。第二是，專業的參考服務館員是否能有效運用知識本體與代理程式。我相信，第二個問題，在圖書資訊學界堅實的學術與實務基礎上，將來一定會發展出各種完善的理論與方法。本文除了簡介語意網的一些概念外，將集中在討論如何建立個領域互通的知識本體架構，以及語意網上的多語問題。我們也將介紹「中央研究院中英雙語知識本體詞網」(Academia Sinica Bilingual Ontological Wordnet, http://BOW.sinica.edu.tw)建構的初步成果。這個成果不但是將來中文網路轉為語意網必須的基本架構，也提供了建立各別領域詞彙庫與知識體的基礎架構。

二．知識本體與詞網

語意網的構想所面對的重要關鍵問題是：知識與知識架構從何而來、如何架構？語意網之所以規定每個網路資源（間單說來，就是每個網頁）要詳細標出自己的知識本體（ontology），出發點就是語言詞彙與知識體系的變異與多樣性。同樣的事物，在不同的語言/方言/領域中有不同的名稱。同樣的名詞，在不同的語境/用法/領域中有不同的意涵。簡單的以一個「鉛」字為例：關鍵詞中列了「鉛」字的網路資源，有可能是談金屬中毒的療法，溫泉分類，或是水晶的製造（或銷售），甚至是「羅馬帝國興亡錄」的相關討論，等等無法盡舉。也就是說，一個

概念的表達，必須在知道概念背後的知識架構後，才有辦法準確判讀。這也就是「訊息」與「知識」間必須跨越的鴻溝。

可是，知識無涯，知識架構也無法盡數。即使是每個網路資源都附上了自己的知識本體。如何才能保證，從不同知識架構與體系（甚至不同語言）出發的使用者，能「看得懂」這個知識本體，而且可以有效的把知識內容轉換，不失真呢？

可能的答案分兩部分。第一是與語意網中必須有個大家遵循的「上層知識本體」(Upper Ontology)。換句話說，知識架構儘管不同，大家必須同意共用一個上層的基本概念架構。這個架構變成了知識轉換與融合的基準。目前已提出，使用較廣的上層知識本體，最重要的是 SUMO (Suggested Upper Merged Ontology，http://ontology.teknowledge.com/)， 由 IEEE 的工作小組提出。

另一部分的答案是一定要用人類語言中現成的知識架構。知識與內容是由人建立，為人所用的。而人最通用的知識表達方式是語言。我們可以說，每個語言中所有詞彙的集合，就是著個語言所能表達概念的集合。而且詞與詞間，已有相當多的語意關係存在。也就是說，每個語言，都有所有說話者共同遵守，隱含的知識架構。這個知識架構雖然缺乏細部分析，卻是是多人使用的基底。不管領域或主題為何，一旦選定使用的語言，該語言的內含語意關係，也會被採用。因此，語言的詞彙關係架構，可以成為細部知識本體的基礎。更重要的，當我們面對多與知識轉換時，各語言的知識架構，不可或缺。在語言的知識架構上，詞網（wordnet）這個詞彙知識庫架構是目前研究者共用的。

總之，我們為完整的領域知識庫的建立，必須在上層知識本體的綱舉目張下，配合個別語言詞網的共有架構，然後才能有效建構與交換。因此，我們為中文以及中英對譯，建立了一個結合上層知識本體的知識詞網，並亦開始試以「魚類」與「財經」兩個領域來建立領域的詞彙庫與知識本體。

三．「中央研究院中英雙語知識本體詞網檢索介面」功能簡介：

「中央研究院中英雙語知識本體詞網」（簡稱「研究院知識詞網」，Sinica BOW）目前開放的雛型是由中研院與國家數位典藏計畫中的「語言座標」計畫所建構完成。主要計畫成員在中央研究院。所引用的資料除了中央研究院詞庫小組（資訊所），文獻語料庫（語言所）及計算中心開發的資料外。國外的有 IEEE 批准執行的 SUMO(Suggested Upper Merged Ontology)知識本體(teknowledge.com 管理)及普林斯頓（Princeton University）的 WordNet。國內主要有來自遠見科技股份有限公司（包括該公司自有資料及與中研院共同開發資料）以及教育部國語會的辭典。

1. 跨語言資訊轉換：

目前以中英雙語查詢為基礎。中文與其他語言轉換為長程目標。但因為「語言座標」採用了「詞彙網路」(WordNet)的架構，為國際間詞彙知識庫通用的架構。找到了英文對譯詞後，可藉 EuroWordNet 等網路上開放的資料庫，對應到 20 幾種語言。

2. 語言資訊與概念架構（知識本體）的連結

目前連結到 SUMO 這個上層共用知識本體。也就是說可以由每個詞查到該詞在概念架構上的歸屬。利用知識本體架構作知識內容分類，與簡單推理。

3. 詞義的區分與詞義關係的連結

語言的特色之一是同一個詞可能有好幾個意義（中文的「機關」可以指機構組織，也可以指害人的陷阱等），因此正確解讀需辨明詞義。另一個特色是詞與詞之間有複雜的語意關係。因而產生了許多的替代說法與推論判斷（如某人喜歡籃球，是喜歡這個運動，而非喜歡這個球體。這些都是建立在「籃球」這個詞的詞義與詞義關係上的）。詞彙網路的架構，提供了多重詞義與詞義關係的檢索。

4. 使用領域

詞語因領域不同而有不同的解釋與用法。我們就現有資料提供了部分的領域標記。此外，領域的使用，也可藉詞彙在不同領域（包括時代，區域，學門等）辭典中的分佈判定。因此我們特別提供了每個詞彙在參考資料中出現分佈的訊息。比如說，如果某個詞在國小課本中出現，應該是簡單基本詞彙。

四。結語

回溯網路的歷史，如果「語意網」的美夢成真，他可能會在短的幾年內衝擊我們。中文的知識內容供給者準備好了嗎？國內個個領域的參考資訊提供者有心理與操作上的準備嗎？「中央研究院中英雙語知識本體詞網」提供了一個關鍵的環節。希望我們可以與國際同步，站在者一波知識革命的先端。

參考文獻

黃居仁，2003，語意網、詞網與知識本體：淺談未來網路上的知識運籌. 佛教圖書館館訊，33 期. 6-21 頁.。

黃居仁 張如瑩 蔡柏生，2003，語意網時代的網路華語教學-兼介中英雙語知識本體與領域檢索介面。Chinese Language Education and the Developing Semantic Web: An Introduction to Chinese-English Bilingual Ontology Interface. To be presented at the Third International Conference of Internet Chinese Education. Taipei, Oct. 24-26. 2003.

Berners-Lee, Tim, James Hendler and Ora Lassila. The Semantic Web. Scientific American. May 2001. [中譯：2002.高虹譯, 黃居仁審, 語意網（Semantic Web）科學人，2002 年 5 月號]。

John DeFrancis,Alphabetically-Based Computerized Chinese-English Dictionary (Honolulu: University of Hawai'i Press ,1996)

Huang, Chu-Ren, Zhao-ming Gao, Claude C.C. Shen, and Keh-jian Chen. 1998. Quantitative Criteria for Computational Chinese Lexicography: A Study based on a Standard Reference Lexicon for Chinese NLP. Proceedings of ROCLING XI. 87-108.

Huang, **Chu-Ren.** Elanna I. J. Tseng, Dylan B. S. Tsai, Brian Murphy. 2003. Cross-lingual Portability of Semantic relations: Bootstrapping Chinese WordNet with English WordNet Relations. To Appear in Languages and Linguistics. 4.2.

Niles, I."Mapping WordNet to the SUMO Ontology."(Teknowledge, 2003.) Technical Report.

網路資源

Sinica BOW「中央研究院中英雙語知識本體詞網」

Academia Sinica Bilingual Ontological Wordnet
http://BOW.sinica.edu.tw
SUMO（Suggested Upper Merged Ontology）.
http://ontology.teknowledge.com/
WordNet.
http://www.cogsci.princeton.edu/~wn/
HowNet Knowledge Database. http://www.keenage.com/
EuroWordNet. http://www.illc.uva.nl/EuroWordNet/
語言座標
http://LingAnchor.sinica.edu.tw
語言典藏
http://LanguageArchives.sinica.edu.tw
中文分詞標準計劃相關資料下載區－四、成果效益檢討內容
http://ckip.iis.sinica.edu.tw/ROCLING/4.htm
「文國尋寶記」
http://www.sinica.edu.tw/wen/
「中央研究院現代漢語平衡語料庫」。
http://www.sinica.edu.tw/SinicaCorpus/
British National Corpus,BNC.
http://info.ox.ac.uk/bnc/
CIDE: Cambridge Internal Dictionary of English.
http://dictionary.cambridge.org/