利用領域詞彙分類與雙語知識本體詞網輔助主題資訊搜尋

陳永祥　　黃居仁
中央研究院語言學研究所
yxchen@gate.sinica.edu.tw churen@gate.sinica.edu.tw

摘要

數位典藏國家型科技計畫自民國 91 年開始推動，旨在將珍貴的重要文物典藏加以數位化，目前透過聯合目錄系統提供整合的典藏內容查詢介面，但由於典藏品的種類屬性繁多與關鍵詞全文搜尋方法功能上的限制，增加了典藏內容與使用者之間的隔閡及搜尋的困難度。而如何將大量以中文為基礎的典藏資料提供其他外語使用者進行查詢或應用，促進國際合作或研究上接軌亦是相當重要的課題。本研究嘗試提出一領域詞彙對應模式，在使用者對主題資訊的搜尋上提供建議詞彙，可協助全面了解典藏內容並得到更豐富的典藏品資訊。過程中以中央研究院語言學研究所開發之中英雙語知識本體詞網 Sinica BOW 為出發點，結合領域詞彙分類方法 DLT 及建議上層共用知識本體 SUMO，並以數位典藏計畫之聯合目錄展示系統為研究對象進行探討。研究成果以知識本體架構來呈現目前數位典藏計畫中典藏品的分類與分布情形，同時透過語義關係延伸來進行關鍵詞拓展，提供使用者在查詢時更多的建議詞彙。

關鍵詞：查詢拓展，WordNet, SUMO(Suggested Upper Merged Ontology), DLT(Domain Lexico-Taxonomy)

1. 前言

「數位典藏國家型科技計畫」自民國 91 年起開始推動，旨在建立國家數位典藏，以保存文化資產、建構公共資訊系統，促使精緻文化普及、資訊科技與人文融合，並推動產業與經濟發展。一般以數位方式典藏之多媒體數位內容博大精深但主題差異性相當大，加上整理後設資料（Metadata)時用的描述語言與使用者搜尋時使用的白話文字常常有落差，因此除非是領域專家，否則一般使用者不易窺知典藏品之學術上專用術語，甚至不知該如何進行搜尋瀏覽。一般全文檢索的搜尋方式無法滿足此一領域之內容研究、搜尋需求，因此增加了豐富的典藏內容與使用者之間的隔閡及搜尋的困難度。因此有必要針對廣泛範圍的典藏內容設計一套符合使用者需求的搜尋策略。查詢拓展(Query Expansion)是一種以建議詞補充原始查詢短語的方法，可用來提升查詢效率。假如查詢拓展的過程採用與使用者互動的模式，則使用者與系統共同參與查詢拓展的工作，系統通常建議一些建議詞給使用者，而使用者從這些建議詞當中挑選一部分當作真正查詢拓展的拓展詞。設計良好的查詢拓展策略必須能夠表達詞彙之間的語意關係，同時幫助搜尋引擎提升檢索效能。

數位典藏國家型科技計畫目前透過聯合目錄[2]提供一整合的典藏內容查詢介面，目前採用傳統的關鍵詞全文搜尋方式供使用者查詢資訊，礙於典藏品的種類屬性以及關鍵詞全文搜尋方法功能上的限制，在許多情況下並不容易提供給使用者最理想的查詢結果。此外，考量數位典藏國家型科技計畫所投入的人力物力，如何將大量以中文為基礎的典藏資料提供其他外語使用者進行查詢或應用，促進國際合作或研究上接軌亦是相當重要的課題。若能建立一中英雙語交叉查詢系統將可使得系統使用者更容易全面性了解典藏內容並找到更多感興趣的相關資訊。因此本研究以中央研究院語言學研究所開發之中英雙語知識本體詞網(Sinica BOW)[1]為出發點，結合領域詞彙分類方法及建議上層共用知識本體，提出一查詢拓展策略並以數位典藏聯合目錄為研究對象進行探討，嘗試用知識本體架構來了解目前數位典藏計畫中典藏品的分類與分布情形，同時以語義查詢的角度來提供使用者查詢時更多的建議詞彙。

綜上所述，本研究之研究目的主要包含下列三項：

1. 以目前聯合目錄所典藏之項目建構一數位典藏計畫知識內容分布架構。
2. 設計一整合策略提供語義上及知識結構上之關鍵詞查詢拓展建議。
3. 建構數位典藏特殊分類內容之領域詞表，提供相關研究或應用之中英雙語對應資源。

2. 相關研究

2.1 數位典藏聯合目錄

「數位典藏國家型科技計畫」自民國 91 年開始推動，旨在將珍貴的重要文物典藏加以數位化，建立國家數位典藏，以保存文化資產、建構公共資訊系統，促使精緻文化普及、資訊科技與人文融合，並推動產業與經濟發展。迄今開發已有 30 餘個典藏計畫與開放型計劃，共約 50 餘個計畫，已有豐富的成果。因此實有必要開發整合型的成果查詢介面提供各界使用者查詢應用，目前數位典藏計畫中兩個主要的展示系統分別為聯合目錄及公共展示系統。而由中研院語言所主導之語言座標計畫則以自然語言處理技術之應用為出發點，希望藉由語言的中介特性能夠將各領域知識系統化呈現並提供語言詞彙使用上的標準。數位典藏計畫中相關參與者及使用者所接觸之資料內容特性如表 1 所示。

表 1 數位典藏計畫資料內容特性

計畫名稱	單一分項計畫	聯合目錄	公共展示系統	語言座標
參與/使用者	領域專家	領域專家與一般大眾	一般大眾	領域專家與一般大眾
資料內容	專業術語	專業術語加通用詞彙	通用詞彙	通用詞彙
資料數量	少	多	少	多
資料種類	多樣	Meta data	多媒體圖片	語言分析結果
典藏內容	實體物品	實體物品	實體物品	邏輯關係與知識媒介

數位典藏聯合目錄是數位典藏國家型科技計畫所建置的目錄性展示平台，旨在提供全國性數位典藏藏品的檢索與搜尋，以展現數位典藏計畫之成效。透過聯合目錄的單一網站窗口，即可檢索全國近百組跨十餘個學術領域之數位典藏內容，提供民眾資訊的查詢及取用數位典藏資訊。聯合目錄工作小組針對各計畫資料庫欄位與架構，進行後設資料語意、語法、結構之分析比對作業，在後設資料(Metadata)建置上採用 Dublin Core 標準，流程上則包含了訪談與表單回填、內涵分析、系統分析、Metadata 測試、評估等多道嚴謹程序。目前於網路查詢介面提供內容主題、時間分類、地理分類、典藏機構與計畫及 Dublin Core 進階搜尋等五種主要資訊搜尋方式。由於資料量龐大，因此目前依據數位典藏計畫將內容主題劃分為 14 個主題類別，包括：生物、地質、人類學、檔案、地圖與遙測影像、金石拓片、善本古籍、考古、器物、書畫、新聞、漢集全文、影音與建築等，其中以檔案主題包含 699,368 筆內容，典藏資料為最多，地質主題包含 3,309 筆資料量為最少。平均而言，每一主題類別包含有約 115,593 筆典藏資料，而各內容主題類別下各以階層結構包含了若干階層與子類別。對於大多數使用者而言，內容主題查詢為最直觀簡單之查詢方式，因此實有必要針對此一項目作更深入的研究，以期提供更人性化更有效率的檢索結果。

圖 1 聯合目錄系統畫面

2.2 領域詞彙分類法 (Domain Lexico-Taxonomy; DLT)

以領域為基礎的語言處理方法是自然語言處理研究重要的一環，利用領域詞彙分類法可將所有知識區分為幾個主要類別。在 Huang 等人[7]的研究中提出了一個針對多領域的語言處理方法，設計出領域分類法並且半自動地建立領域詞表。每個用來存放領域詞表的分類類別即稱為領域詞彙分類(Domain Lexico-Taxonomy; DLT)， DLT 可以在識別及處理多領域語言資料內容上提供核心詞彙資訊。在領域分類研究中，以人工方式將所有知識區分為 549 個領域，主要的參考依據是中文圖書分類系統、大英百科全書及遠見英漢詞典。

領域分類以階層樹結構呈現主要領域及子領域的上下位關係，領域依抽象程度共分為四個層級，最上層分為 14 個領域，包含人文學科、社會科學、形式科學、自然科學、醫療科學、工程科學、應用產業、藝術、休閒娛樂、專有名詞、語體、各種語言/詞源、各國地名與各國民族，是抽象程度最高的分類。其下第二層具體區分為 147 個子領域，而第三層更細分出 279 個更具體的子領域，最後一個階層則因為並非所有節點均有所延伸而僅細分出 109 個子領域。總計四層樹狀結構中共有 549 個領域分類節點。

每一個領域分類節點中包含了一個小的領域詞表，詞表中的詞彙均是由 WordNet[3]中抽取出的一般常用詞彙，目的是不需要透過特殊的領域詞彙即可以一般性詞彙來定義出一個領域。目前領域分類已可與 WordNet 的同義詞集(synset)進行對應並且可提供中英雙語的配對詞彙。由於 WordNet 中提供了詞彙的下位關係詞彙，因此一個詞彙所屬的領域可以繼承給所有的下位詞彙，研究中將 15,160 個中文詞彙對應至 463 個領域分類中，領域詞表的產生方式可以如圖 2 所示：

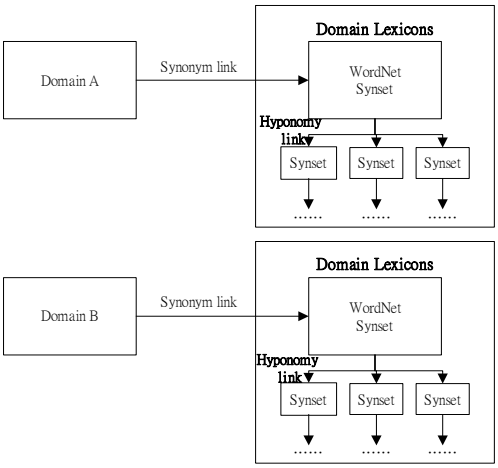

圖 2 DLT 領域詞表建構方法

2.3 建議上層共用知識本體 Suggested Upper Merged Ontology

SUMO（Suggested Upper Merged Ontology，建議上層共用知識本體）[4]是由 IEEE 標準上層知識本體工作小組所提出的知識本體架構，目的是發展成標準的上層知識本體，這將促進資料互通性、資訊搜尋和檢索、自動推理和自然語言處理。知識本體（ontology）類似於一組字典或術語表，但能夠使電腦處理更多內容的細節和其結構。透過知識本體可將人們有興趣的領域正規化為一套概念、關係和定理（axiom）。上層的知識本體被限制在 meta 的概念、一般、抽象或者哲學，因此足夠一般提出（在一定水準上）一個涵蓋廣闊範圍的領域區域[9]。特殊領域具體的概念不被包括在上層知識本體中，但是這樣的知識本體可提供特殊領域（例如：藥、財政、專案...等等）的知識本體結構的建立。SUMO 藉由最高層次的知識本體，鼓勵其他特殊領域知識本體以其為基礎衍生出其他特殊領域的知識本體，並為一般多用途的術語提供定義。目前 SUMO 已經和英語詞彙網路 WordNet1.6 版本作初步的連結。SUMO 中的節點以階層樹方式連結，如圖 3 所示。

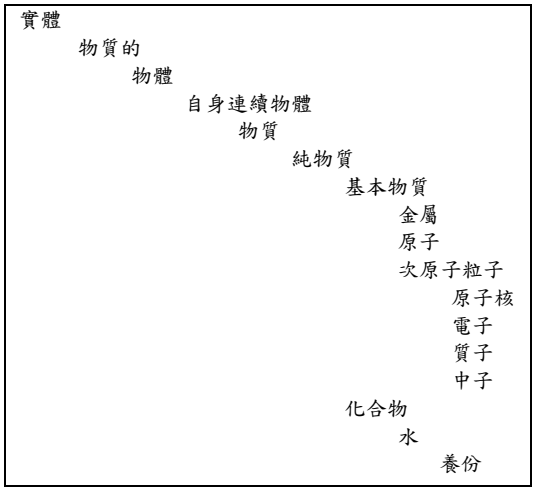

圖 3 SUMO 階層節點示例

2.4 中央研究院中英雙語知識本體詞網(Sinica BOW)

中英雙語知識本體詞網(Sinica BOW)[1]是一結合詞網(WordNet)，知識本體，與領域標記的詞彙知識庫，由中央研究院語言所文獻語料庫小組與資訊所中文詞知識庫小組合作建置，從語言工程的角度，以台灣地區的語言使用為經驗基礎，提供語言和語言、語言和概念以及語言和領域的資訊，甚至是跨語言間的訊息。中英雙語知識本體詞網以建立一完整精確的中英對譯資料庫及檢索介面為目的，作為數位典藏知識國際化的基礎；並持續建立各領域之雙語領域辭典，以作為各領域／典藏之雙語控制詞彙參考標準。中英雙語知識本體詞網同時提供具領域判斷能力之資訊檢索應用。此外，建立附加領域標記之雙語辭典及檢索介面使中英雙語知識本體詞網成為一知識加值雙語電子辭典。

Sinica BOW主要使用的資源包含WordNet、ECTED（English- Chinese Translation Equivalents Database）以及SUMO（Suggested Upper Merged Ontology，建議上層共用知識本體）。其中WordNet[3]是1985年普林斯頓大學認知科學實驗室以現代心理語言學理論所述的人類詞彙記憶為啟發所開發出的語意式電子字典，以每個同義詞集表達一種詞彙概念，將同義詞集區分為四種英文詞類：名詞、動詞、形容詞、副詞，並以二十幾種詞義關係組織同義詞集。由中研院資訊所與語言所合作建構的ECTED以WordNet為基礎，經由現有英中或中英電子辭典的詞形對應，爲每個同義詞集詞義找出可能相對應的中譯詞組，再經由人工檢驗。尋找對譯盡可能的以詞彙而非描述性短語表達，目的在於讓每個同義詞集都有最適當的一至三個左右的中文對譯。[6]

依據SUMO2002年版資料，黃居仁等人[8]將系統介面以及概念節點進行中文化，其涵蓋11大類的概念，每大類又區分為二至五個類別，總共囊括3,912個概念。SUMO已經與WordNet1.6版本結合，且以同義（synonymy）、上位（hypernym）、體例（instantiation）這三種類別顯示同義詞集和SUMO概念間的對應關係，例如：同義詞集cell（細胞）與細胞概念（cell）是同義。hockey（曲棍球）屬於運動概念（sport），兩者間的關係為上位，也就是說運動涵蓋hockey（曲棍球）。China（中國大陸）屬於國家（nation）這概念的體例。除此，更以「中國圖書分類法」為基準，並參考各知識分類與實際研究經驗，提出：包含九大類的知識分類（Knowledge Content），涵蓋427個領域。另外，並因應語言資源特性加入下列語言使用（Language Usage）的各類訊息：專名（說明文字符號的指涉）（Proper Name）、語體（說明文字符號的使用）（Genre/Strata）、各種語言／詞源（Language/Etymology）、各國地名（Country Name）。領域階層的建立在於替不同詞義中的詞彙項目區別其使用的領域，例如：stock作「股票」和

「家畜」兩個不同解釋時，分屬於財政學裡的資本以及動物學的脊椎動物學。加註領域信息可降低詞彙歧異性，增加資料交換時的互通性，輔助領域詞彙庫之建構。Sinica BOW透過WordNet1.6 offset延伸所產生的識別碼作為媒介，進行串連，將每個資源以及各類訊息連結。因WordNet1.6 offset延伸的識別碼可獲得原本WordNet存在的詞類、解釋、英文例句、同義詞集、各同義詞集間的詞義關係及其所屬詞彙。而SUMO概念與WordNet的連結，使得可透過該識別碼獲取詞義與概念搭配的訊息。以WordNet為基礎所建置的ECTED與針對WordNet同義詞集的各詞彙項目所給予的領域值，也是透過該識別碼獲取。而特殊領域詞彙庫，加上相對應的Sinica BOW識別碼，也可保留原始資源的資料庫格式和WordNet連結。又，領域知識本體則是在SUMO某些概念下進行延伸發展。每個特殊領域詞彙庫中的詞彙一樣具有所屬的概念，其所屬概念可能是SUMO或特殊領域知識本體的某一概念，特殊領域詞彙庫和領域知識本體的結合，使得透過該識別碼又串起所有的訊息。Sinica BOW的資源和架構如圖4所示。由於透過WordNet可以和同是以WordNet為基礎架構所建置的其他語系WordNet資源加以連結，例如：EuroWordNet[9]，因此以此基礎架構可編製成多語的詞彙網路，成為多語環境中所需之語言知識結構的基礎資料。

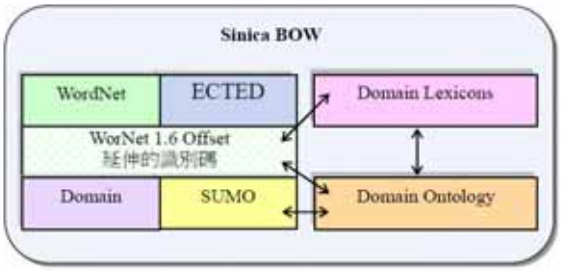

圖 4 SINICA BOW 架構圖

3. 領域詞彙知識對應架構

根據研究目的與相關研究探討，本研究結合DLT、WordNet 與 SUMO 等架構所提供之橫向與縱向語言資源，將聯合目錄中各主題分類對應至一建議架構中如圖 6 所示，由此架構連結各項語言資源提供領域分類之細項資訊，除可將數位典藏計畫之典藏品類別對應至知識本體中觀察目前典藏品在整體知識架構中的分布情形，同時亦可擴充聯合目錄使用者在進行主題分類檢索時的查詢拓展詞彙。

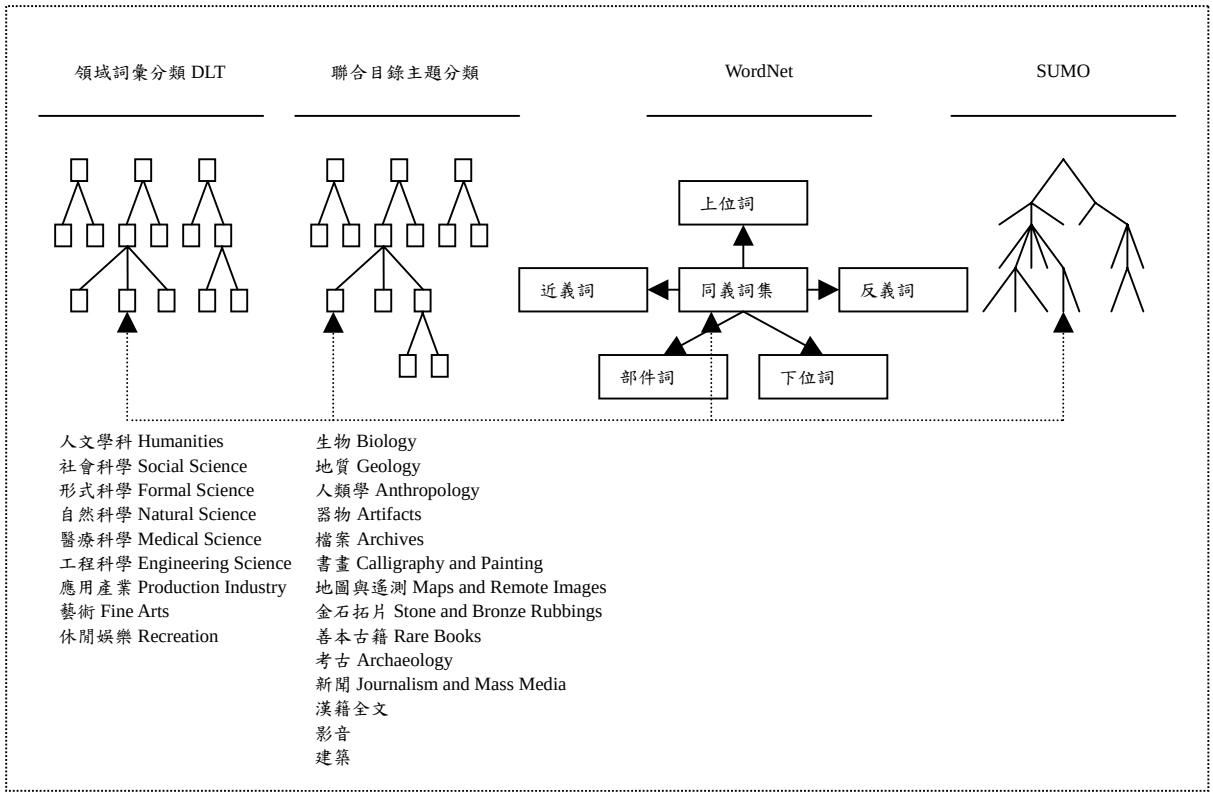

圖 6 領域詞彙對應模式

表 2 聯合目錄主題分類資料對應表

聯合目錄主題分類	領域分類	WordNet 同義詞集 (同義詞集 offset)	SUMO 對應關係 ； SUMO 概念
生物 Biology	自然科學 -> 生物學	biota, biology (05987709)	Hypernym ; Organism
地質 Geology	自然科學 -> 地球科學 -> 地質	geological formation, geology, formation (06691504)	Hypernym ; SelfConnectedObject
人類學 Anthropology	社會科學 -> 人類學	social anthropology, cultural anthropology (04673338)	Instantiation ; FieldOfStudy
器物 Artifacts	藝術 -> 造形藝術	artificiality (03757381)	Hypernym ; SubjectiveAssessmentAttribute
	藝術 -> 裝飾藝術 -> 手工藝		
檔案 Archives	人文學科 -> 圖書館學	archive, archives (02206789)	Hypernym ; EducationalOrganization
	人文學科 -> 史學		
書畫 Calligraphy and Painting	藝術 -> 圖畫藝術	calligraphy, penmanship (04826894)	Hypernym ; Text
	藝術 -> 裝飾藝術 -> 書法	painting, picture (03079051)	Hypernym ; ArtWork
地圖與遙測 Maps and Remote Images	藝術 -> 攝影 -> 遠距攝影	Map(02965788)	Hypernym ; Icon
	藝術 -> 圖畫藝術 -> 製圖法		
	應用產業 -> 交通		
	自然科學 -> 地球科學		
金石拓片 Stone and Bronze Rubbings	人文學科 -> 古文字學	Rubbing (03259790)	Hypernym ; copy
		Bronze (02342693)	Hypernym ; ArtWork
善本古籍 Rare Books	人文學科 -> 史學	book (04831824)	Synonymy ; Book
	人文學科 -> 哲學	papyrus (04868693)	Hypernym ; Text
	人文學科 -> 文學		
考古 Archaeology	社會科學 -> 考古學	archeology, archaeology (04670536)	Instantiation ; FieldOfStudy
新聞 Journalism and Mass Media	工程科學 -> 電信通訊	journalism (00403517)	Hypernym ; OccupationalRole
		information bulletin (05010288)	Hypernym ; Text
		report, news report, story, account, write up (05009327)	Hypernym ; Text
漢籍全文	人文學科 -> 宗教	Literature (04798932)	Hypernym ; Text
影音	藝術 -> 音樂	dramaturgy, dramatic art, dramatics, theater, theatre (05256340)	Instantiation ; FieldOfStudy
	藝術 -> 表演藝術	play, drama (05257517)	Hypernym ; Text
建築	工程科學 -> 建築	construction, building (00715519)	Hypernym ; Constructing

以此架構進行查詢拓展可得到縱向及橫向延伸之相關詞彙資料，為優於傳統查詢拓展方式之設計。根據聯合目錄系統對典藏品所區分之14類主題，可以一對多方式對應至DLT領域詞彙分類中，而各DLT領域詞表中的詞彙即可稱為是該聯合目錄主題之基本詞彙。此外，透過SUMO之邏輯關係可將主題分類對應至SUMO架構上特定節點，並由該節點延伸納入相關子節點。以WordNet synset之offset number 找出包含上下位詞及近義反義詞作為領域主題查詢延展時之參考延伸詞彙。在建議詞策略上，一般搜尋引擎採用的是被動式使用者行為觀點: 以回饋資料為導向，計算使用者行為而得到相關資料頁面。主動式知識架構觀點則以嚴謹的學術分類為基礎，架構出明確體系，可透過語言提供知識交換基礎平台。

4. 研究結果

根據前文所提出之領域詞彙知識對應架構，可以清楚觀察聯合目錄中採用的主題分類在第一層的 14 種主題分類中如何對應到 DLT 的領域分類，其中某些主題由於缺少直接對應的分類項目，因此列出兩個以上的類似分類。由於階層樹狀結構的特性，第二層以下較具體的聯合目錄細項主題分類亦可以相同方式連結至相對應的 DLT 領域分類當中，亦可得到對應的 WordNet synset 資料以及 SUMO 節點資訊。

4.1 數位典藏計畫知識內容分布架構

以第一層的聯合目錄主題分類對應至 SUMO 結構中可以觀察目前數位典藏計畫所建置的典藏品內容在整體人類知識中所佔的位置，如表 3 所示。其中 SUMO 節點以節點代號數字串代表，例如“生物體 1.1.1.1.2.4.8.,organism,生物體” 代表著如同位於圖 7 的階層結構，以此架構可以了解目前在數位典藏計畫中的典藏品以有形的物質居多，其中又以文本為最大宗；在知識本體上屬於抽象項目的則以專業的研究領域如人類學及考古學最為明顯。

表 3 SUMO 節點對應表

聯合目錄主題分類	SUMO 節點位置
生物 Biology	1.1.1.1.2.4.8.,organism,生物體
地質 Geology	1.1.1.1.,self connected object,自身連續物體
人類學 Anthropology	1.2.7.32.,field of study,研究領域
器物 Artifacts	1.2.4.13.61.211.,subjective assessment attribute,主觀評價屬性
檔案 Archives	1.1.1.4.12.31.51.,educational organization,教育組織
書畫 Calligraphy and Painting	1.1.1.1.3.8.22.,text,文本 1.1.1.1.2.5.13.,art work,藝術品
地圖與遙測 Maps and Remote Images	1.1.1.1.3.7.,icon,圖示
金石拓片 Stone and Bronze Rubbings	1.2.5.14.62.214.202.142.215.,copy,複製 1.1.1.1.2.5.13.,art work,藝術品
善本古籍 Rare Books	1.1.1.1.3.8.22.45.,book,書籍 1.1.1.1.3.8.22.,text,文本
考古 Archaeology	1.2.7.32.,field of study,研究領域
新聞 Journalism and Mass Media	1.2.4.13.60.210.,occupational role,職業角色 1.1.1.1.3.8.22.,text,文本
漢籍全文	1.1.1.1.3.8.22.,text,文本
影音	1.2.7.32.,field of study,研究領域 1.1.1.1.3.8.22.,text,文本
建築	1.1.2.8.43.93.94.,constructing,建構

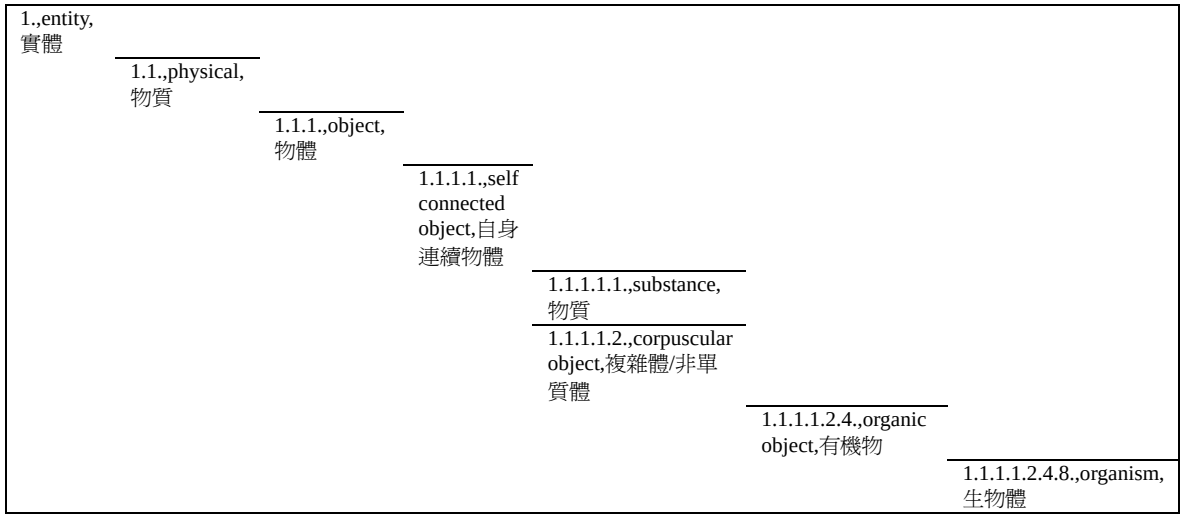

圖 7“生物”節點在 SUMO 中的位置

4.2 數位典藏分類內容之領域常用詞表

透過與 DLT 及 WordNet 的對應連結可以爲每一個數位典藏聯合目錄中的主題分類提供領域常用詞表。正如同 DLT 的設計精神，以領域常用詞表作為定義領域的方式可以在不熟悉艱澀的專業術語情況下爲特定領域定義出範圍，做為知識分類的工具。

分類內容之領域常用詞表同時可應用於資訊檢索上，透過領域常用詞表進行查詢拓展時可以協助使用者通盤了解該領域中的相關詞彙及這些常用詞彙之間的關係，有助於全面了解領域中相關內容及知識結構。以”地質”領域為例，可以透過 DLT 及 WordNet 的連結找出如表 4 的領域常用詞表，詞表中包含了研究地質主題的各學門名稱，以及相關的氣象學及礦物學等等。透過這些常用詞表即可概略了解地質學所研究的內容與項目，對於陌生的使用者而言可提供相當大的幫助。單純以地質為關鍵詞進行查詢時，在聯合目錄系統上可以查詢到 4,760 筆資料，但若採用領域常用詞表中的建議詞彙進行查詢，則可獲取 9,750 筆資料，可得到兩倍以上的資料量。

表 4 地質領域詞表

”地質”領域詞表
大地測量學(geodesy),水文學(hydrology),火山學(volcanology),古地質學(paleogeology),石油地質學(petroleum_geology),地球物理學(geophysics),地質學(geology),地震學(seismology),地學(geology),形態學(morphology),岩石學(petrology),洞穴學(speleology),洞窟學(speleology),氣候學(climatology),氣象學(meteorology),測地學(geodesy),結構地質學(tectonics),經濟地質學(economic_geology),構造地質學(tectonics),構造運動學(tectonics),磁學(magnetism),礦物地質學(mining_geology),礦物學(mineralogy),

4.3 詞彙查詢拓展

聯合目錄查詢系統提供了使用者方便的查詢介面，累積至 2006 年 7 月止，使用者查詢頻率較高的詞彙如表 5 所示。其中可以發現與蝴蝶相關的詞彙相當多，但部分詞彙屬於較為艱深的專業詞彙。此一現象說明部份艱深的專業詞彙在實際查詢使用上有其實用性，然而未受過專業訓練的使用者可能未具備足夠的專業知識可直接以這些詞彙進行查詢。本研究所提出之對應模式即可在此一情況下針對使用者感興趣之詞彙進行詞彙拓展，增加使用者查詢時的參考依據。

以”蝴蝶”作為使用者搜尋詞為例，可以透過 DLT 及 WordNet 的連結找出蝴蝶所屬的領域常用詞表如表 6，詞表中包含了具代表性的蝴蝶名稱，以及分類上相近的蛾類等。透過這些常用詞表可概略了解蝴蝶與”昆蟲”、”蛹”等詞彙歸屬同一領域，這對於陌生的使用者而言可提供有效的查詢提示。

而 WordNet 亦提供了同義詞集、上位詞及下位詞等相關詞彙，表 7 列出”蝴蝶”在 WordNet 中的關係詞彙。另外，由圖 8 則可了解”蝴蝶”概念在 SUMO 架構中的位置以及所屬的知識體系。

在本研究所提出的模式中，使用者可分別由領域詞、關係詞、知識架構等三個面向得到關鍵詞彙的拓展。觀察以”蝴蝶”進行詞彙拓展的結果可以發現，許多拓展所得的詞彙與聯合目錄統計之使用者高頻查詢詞相符合，因此說明詞彙拓展可以在詞型比對之外，以詞義分析方法提供使用者有效的查詢詞彙拓展協助。

表 5 聯合目錄高頻查詢詞

關鍵詞
胡台麗；陶；王建民；酒器；虱目魚；陶 AND 陶器；西拉雅；**蝴蝶**；聯勤；蓮；台中；大乘；董氏針灸；大稻埕；**蝶**；文徵明；坐骨 AND 神 AND 痛；棒球；翠玉白菜；鬼；…；**紫斑蝶**；**昆蟲**；**挵蝶科**；**紫斑蝴蝶**；**蝶科**；**鳳蝶總科**；**鳳蝶**；**鱗翅目**；……

表 6 蝴蝶領域詞表

”蝴蝶”領域詞表
衣蛾(tineid),衣蛾(clothes_moth),夜盜蛾(armyworm),夜盜蛾(armyworm),昆蟲(insect),昆蟲(bug),昆蟲(coreid_bug),枯葉蛾(lasiocampid),枯葉蛾(eggar),麥蛾(gelechiid),麥蛾(Gelechia_gossypiella),麥蛾(angoumois_moth),菜粉蝶(small_white),菜粉蝶(southern_cabbage_butterfly),葉蟲(leaf_bug),葉蟲(mirid_bug),蛹(pupa),蛹(chrysalis),蛾(moth),蛾(gypsy_moth),鳳蝶(emperor_butterfly),鳳蝶(emperor),穀蛾(tineid),穀蛾(grain_moth),燈蛾(arctiid),燈蛾(tiger_moth),蠶蛾(bombycid),蠶蛾(giant_silkworm_moth)

表 7 蝴蝶在 WordNet 中的關係詞

同義詞集
蝴蝶、蝶
上位詞
鱗翅目昆蟲
下位詞
粉蝶、熱帶臭蝶、小灰蝶、蛺蝶科的蝴蝶

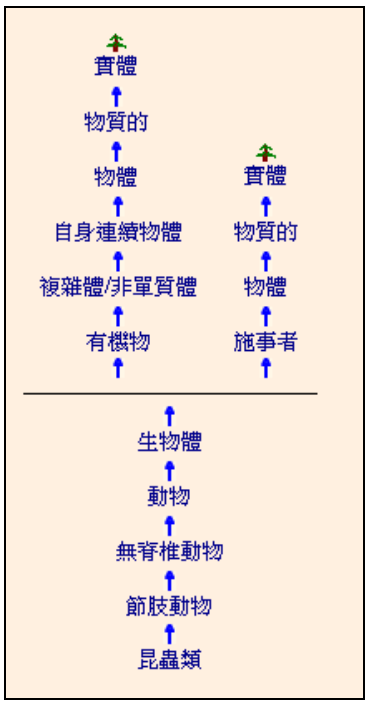

圖 8 “蝴蝶”所屬的知識體系

5. 結果與討論

數位典藏國家型計畫為一規模龐大之典藏計畫，涵蓋的資料量與知識層次均相當龐大，各主題彼此之間的差異亦大，在建構一整合性之展示系統時如何有效呈現資料內容是一相當重要之挑戰。本研究透過連結 DLT, Sinica BOW 與 WordNet，針對網路使用者在瀏覽檢索上提供一查詢拓展模式，DLT 中的領域分類系統主要基於中文圖書分類系統，同時參考大英百科全書及遠見英漢詞典等語言資源。可以與數位典藏之主題分類進行對應並提供相關的領域詞彙。而 WordNet 所提供的透過各種關係所聯結的相關詞彙可以有效幫助了解詞彙的使用及詞彙間關係，亦對於使用者全盤了解某領域知識並進行查詢有相當大的助益。另外，由 Sinica BOW 所提供之中英雙語對應詞彙，可供作為聯合目錄展示系統後續發展中英雙語查詢介面之使用。最後，本研究透過主題分類與 SUMO 節點的對應，可供了解數位典藏計畫中典藏品在整體人類知識架構中所在之環境位置，亦對於延伸了解相關知識提供了方向。因此，由本研究可瞭解自然語言處理技術與語言資源在數位典藏計畫中的定位及應用方向。後續研究將以整合連結其他語言資源，發展語義查詢拓展為首要目標，在應用上則計劃建置自動化模組輔助數位典藏計畫之相關展示系統處理查詢拓展應用。

參考文獻

[1] 中央研究院中英雙語知識本體詞網 The Academia Sinica Bilingual Ontological Wordnet (Sinica BOW)，http://BOW.sinica.edu.tw

[2] 數位典藏聯合目錄，http://catalog.ndap.org.tw/

[3] WordNet, http://www.cogsci.princeton.edu/~wn/

[4] Suggested Upper Merged Ontology, http://www.ontologyportal.org/

[5] Bourret, XML and Databases, 2004, http://www.rpbourret.com/xml/XMLAndDatabases.htm

[6] Huang, Chu-Ren, Elanna I. J. Tseng, Dylan B. S. Tsai, and Brian Murphy, 2003, Cross-lingual Portability ofSemantic relations: Bootstrapping Chinese WordNet with English WordNet Relations, *Language and Linguistics*, 4.3, pp.509-532.

[7] Huang, Chu-Ren, Xiang-Bing Li, Jia-Fei Hong, 2004, Domain Lexico-Taxonomy:An Approach Towards Multi-domain Language Processing, *Proceedings of the Asian Symposium on Natural Language Processing to Overcome Language Barriers*, pp. 54-60, March 25-26, Hainan Island.

[8] Huang, Chu-Ren, Ru-Yng Chang, and Shiang-Bin Lee, 2004, Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. *4th International Conference on Language Resources and Evaluation (LREC2004)*, Lisbon. Portugal.

[9] Niles, I., and Pease, A., 2001, Toward a Standard Upper Ontology, *In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, Chris Welty and Barry Smith, eds, Ogunquit, Maine.