

走向開放化、語意化與關聯化的 《中國分類主題詞表》

范焯

四川大學公共管理學院信息管理技術系副教授

E-mail: fanw@scu.edu.cn

關鍵詞：中國分類主題詞表；關聯數據；詞表資料集

【摘要】

《中國分類主題詞表》是中國大陸地區情報檢索語言研究與實踐的典型代表，也是一種重要的中文知識組織體系。語意網技術與關聯數據實踐為《中國分類主題詞表》研究帶來了新思路與新方法。文章以「開放化—語意化—關聯化」遞進邏輯為寫作脈絡，彙報《中國分類主題詞表》的階段發展。目前《中國分類主題詞表》在開放化與語意化方面已經具備較好的資料基礎與技術支撐，未來將朝向關聯數據集開發與知識服務應用方面重點發展。

前言

《中國分類主題詞表》(Chinese Classified Thesaurus, 以下簡稱《中分表》)是一部集分類法與主題法於一身的綜合性大型詞表，由「分類號—主題詞對應表」與「主題詞—分類號對應表」兩部分構成的分類主題雙向對應表。從情報檢索方法論層面上看，《中分表》既體現了分類法的學科專業性，又保留了主題法的事物直接性，為資訊資源標引與檢索提供了更加靈活、豐富的組織手段。

《中分表》的編制來源於兩部詞表：《中國圖書館分類法》(Chinese Library Classification, CLC, 以下簡稱《中圖法》)與《漢語主題詞表》(Chinese Thesaurus, CT)。《中分表》的分類體系與《中圖法》保持一致，屬於典型的以學科為主要劃分依據的層累制分類體系。《中分表》的主題詞部分以敘詞表結構為主，主題詞數據也是中國國家圖書館主題規範資料的重要組成部分。《中分表》擁有豐富的主題概念，涵蓋哲學、社會科學、自然科學、工程技術等各領域主題概念，目前收錄分類法類目5萬多個，主題詞11萬多條，主題詞串約6萬條，入口詞3千多條(《中國圖書館分類法》編輯委員會，2012)。從結構複雜程度和主題詞數量上看，《中分表》是中國大陸地區現行體量最大的一部綜合詞表。

《中分表》是 1987 年由中國國家圖書館《中圖法》編委會主持牽頭，聯合大陸地區 40 多家圖書情報機構共同編制而成，於 1994 年出版。之後，《中分表》與《中圖法》的日常維護與修訂統一由國家圖書館常設機構《中圖法》編委會負責管理，兩者採用同一套技術系統進行維護，實現了基礎資料的共通共用。在實際管理中《中分表》的修訂維護一般在《中圖法》主要版本推出後進行。《中分表》第一版以《中圖法》第三版與《漢語主題詞表》為基礎，1994 年編制完成。《中分表》第二版以《中圖法》第四版為基礎進行修訂，2005 年同時推出紙質版與光碟版。《中分表》第二版 Web 版於 2010 年 3 月 17 號上線，2014 年 1 月更新到 2.1 版本，與《中圖法》第五版 Web 版同步更新。《中分表》Web 版的誕生，在中文詞表實踐中具有兩個標誌性「打破」意義：一是實現了詞表修訂的動態即時更新，打破了詞表傳統更新週期過長的缺點；二是詞表的網路發佈，打破了詞表的資料孤島，讓詞表可以參與到更廣闊的網路資訊資源組織與服務，從而驅動詞表服務應用創新。下圖是《中分表》發展脈絡一覽圖，以時間軸為主要維度，梳理了《中分表》的主要里程碑、形式特點、數位化研究主要方面以及與《中圖法》的版本對照等，其中的具體內容將在後面各部分論述中詳細討論。

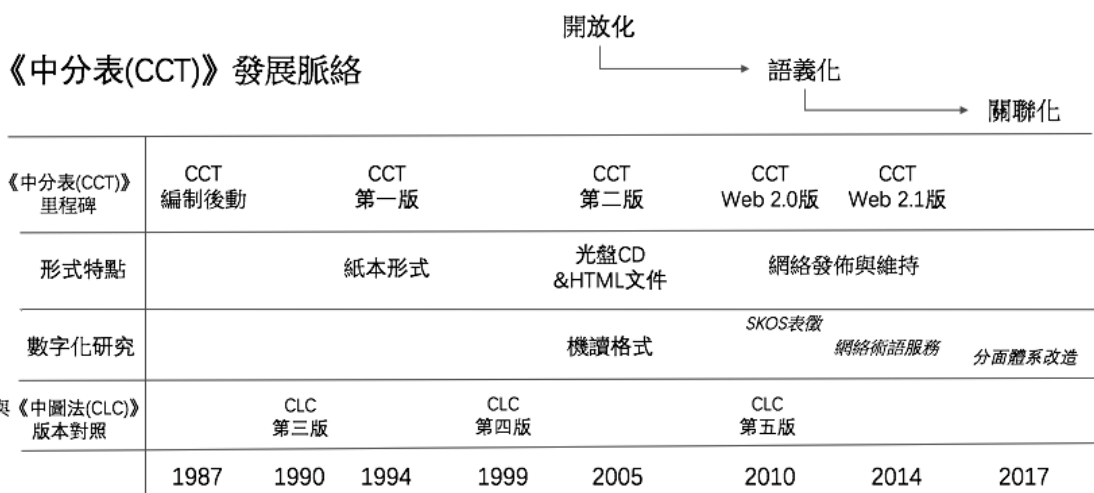


圖 1 《中分表 (CCT)》發展脈絡

從某種意義上講，《中分表》的詞表本身及編制理論方法代表了中國情報檢索語言的一種最佳實踐。從資源角度看，《中分表》包含了豐富的資料資源（主題詞規範資料、分類體系以及概念語意關係等），這是非常寶貴的知識組織財富。如何更好地開發與利用《中分表》？如何讓其參與到更廣闊的網路資訊資源組織與知識服務應用實踐當中，發揮詞彙控制、知識標引與知識組織等作用？這是立足詞表自身發展的主動研究思考。

文章首先將詞表研究置身於不斷變化的資訊環境中，以《中分表》為典型代表，討論資訊環境變遷中的詞表定位與發展；接下來，以詞表的開放化、語意化與關聯化三個研究遞進展開論述，彙報《中分表》已有研究進展；最後，筆者從自身研究角度對《中分表》的發展與應用實踐進行了展望。

資訊環境變遷中的詞表發展

在圖書情報領域中，詞表的本質屬性是工具性。編制詞表的最初目的一方面是為文獻標引與組織提供規範化與結構化的工作依據，另一方面是為用戶檢索文獻提供更有效的手段與途徑。這兩者的本質是相通的。詞表是資訊資源庫與資訊檢索系統之間的紐帶，詞表本身品質與使用情況在很大程度上決定了資訊資源的組織效果與檢索成效。

由於《中分表》與《中圖法》密切的姊妹關係，中國國家圖書館在中文文獻資料編目系統中集成了《中分表》，文獻標引的書目資源都具備《中分表》賦予的規範化主題詞與《中圖法》分類號。從中國大陸地區圖書情報行業角度看，《中分表》作為文獻標引與組織的工具依然會長期穩固存在下去。

在萬維網（World Wide Web, WWW）出現之前，圖書館、情報機構與檔案館是資訊資源最豐富的擁有機構。然而，隨著萬維網與資訊處理技術的快速發展，海量資訊爆發式增長，通過網路發佈進行發佈與連接。數位化網路已經成為圖書情報領域之外的重要資源陣地。以資訊技術牽引的資訊環境發生著快速變化，詞表及其標引組織的資訊資源需要適時地隨之改變。圍繞網路資源是否需要詞表進行有序化、規範化控制的問題，萬維網應用的發展實踐已經給出了答案。

在萬維網早期，以 Yahoo！目錄、DMOZ 為代表，採用主題與學科混合的分類目錄建立網站索引。在關鍵字搜尋引擎興起後，以 Google、Bing 為代表的搜尋引擎將敘詞表用作檢索系統後端的詞彙控制，實現拼寫建議、查詢擴展與限定等搜索協助工具。在以用戶為中心的 Web 2.0 時代，詞表被網路資訊架構師（Information Architect）收入工具箱，用於網站的詞彙開發、內容主題組織以及導航檢索的有力工具。在大眾參與「打」標籤的 Folksonomy（分眾分類法）中，詞表為 Folksonomy 的集合式鬆散結構與標籤歧義等問題提供了優化思路。在強調語意與結構的語意網（Semantic Web）時代，許多著名的詞表如杜威十進位分類法（Dewey Decimal Classification, DDC）、國際十進位分類法（Universal Decimal Classification, UDC）、藝術與建築索引典（Art & Architecture Thesaurus, AAT）等借助語意網技術很快地成功實現了語意結構的資源描述架構（Resource Description Framework, RDF）表徵、關聯發佈以及擴展為支持自動推理的本體知識庫。在推進智慧化與知識化的泛在智慧時代，詞表用於文本挖掘與知識標引的作用不斷在凸顯。具體而言，數位人文應用中的知識圖譜與機器學習無疑離不

開對領域知識的詞表支撐。

新時代詞表擁有了一個內涵更豐富的專業稱謂——知識組織系統 (Knowledge Organization System, KOS) (Zeng & Chan, 2004)。KOS 涵蓋了分類法、敘詞表、地名詞典與本體 (ontology) 等不同結構功能特點的詞表，引導詞表進入網路各類資訊資源的描述與檢索階段。關聯數據 (Linked Data) 是語意網的一個重要里程碑，為詞表發展帶來新的挑戰與機遇。以 RDF 為基礎的資料開放與關聯應用，成為當前語意網最接地氣的應用實踐。各行業領域借助關聯數據方法開放共用了各類結構化語意資料，以資料集 (dataset) 方式逐漸建立起關聯化的資料網路，即關聯數據雲圖 (Linked Data Cloud, LOD) (Cyganiak & Jentzsch, 2014)。圖書情報機構擁有大量豐富的、高品質的文獻資料，連同文獻資源標引與組織的詞表資源一起參與到關聯數據運動中。美國國會圖書館、OCLC、大英圖書館等都在將書目資料發佈成關聯數據集的同時，提供了詞表服務介面，讓網路使用者能夠更好地訪問與使用書目資料。

詞表除了工具屬性之外，還具有資源屬性。詞表本身也是優質的詞彙資料集資源，其規範化、結構化語意化程度普遍較高。詞表發佈為關聯數據，具體表現為 RDF 詞彙集。在關聯開放詞彙平臺 (Linked Open Vocabulary, LOV) (Schaible, Gottron, Scheglmann & Scherp, 2013) 中可以找到已發佈為關聯數據的各種詞表資源，進一步瞭解有哪些詞表資源可用，以及發現這些詞表之間的隱含關係。由於詞表用於資源標引組織，詞表資料與使用它們進行標引組織的資來源資料產生直接聯繫。通過同一部詞表標引組織文獻或不同詞表標引 (這些詞表之間實現了相容互轉)。在關聯數據網路中，詞表作為連接仲介，最終將搭建起資來源資料之間更廣泛的關聯集成。

雖然資訊環境不斷變化，但詞表的本質作用沒有發生改變，變化的只不過是詞表的外在形式、參與資源組織的方式而已。詞表主要經歷了三種形態：面向文獻組織的標引工具、面向網路資訊資源的知識組織系統，以及關聯數據語境中的 RDF 詞彙集。在資訊環境變化中定位詞表研究，這對深入剖析與研究《中分表》本身有著非常重要的方向性作用。

具體就《中分表》而言，本文提出三個問題：1. 為什麼普遍觀點認為《中分表》只是圖書情報領域文獻標引工具？2. 為什麼《中分表》跟不上、滿足不了多元化網路資訊資源的描述與組織要求？3. 為什麼《中分表》的詞彙、概念與語意關係無法被網路協力廠商應用與服務開發者所利用？圍繞這三個問題，文章將從詞表的開放化、語意化與關聯化等三個方面對《中分表》發展進行梳理與探討。

《中分表》的開放化

開放是萬維網的精神所在。萬維網打破了行業界限，開放的資料格式、網路通訊協定與技術規範讓原本一個個獨立存在的資訊孤島能夠連接起來。通過數位化與網路化手段，圖書

館、情報機構與檔案館擁有的文獻資源讓普通使用者可以通過網路進行訪問、檢索與利用。這就是開放的直接表現。

文獻資源通過網路實現了開放，用於文獻標引與組織的詞表隨著資源的網路開放也在逐漸走向開放。若詞表不開放，僅存在於文獻編目業務流程的標引工作中，其使用價值與應用面就會被侷限起來。詞表的開放意味著文獻資源的共用與集成這個前提條件得以成立，以詞表作為仲介的資源關聯層面實現也就成為可能。

開放為資源的交換與共用帶來可能性與便利性。開放的反義詞是封閉，資訊孤島一直以來都是最大的問題。作為中國大陸地區最大規模的分類主題一體化標引工具，《中分表》從封閉走向開放，主要經歷了以下五個階段：

一、《中分表》的紙質形式

紙質形式是《中分表》早期開放的最大障礙。詞表要開放，首先必須數位化。《中分表》自編制完成以來，按照慣例出版發行紙質本。《中分表》的紙質本無法進行資料開放與共用，其作用可視為圖書情報領域編目人員手頭的一部工具書，對文獻進行主題標引時的參考手冊。從資料共用與開放層面，紙質本的《中分表》是封閉的，不開放的。《中分表》第一版（1994）出版紙質本，《中分表》第二版（2005）同時出版紙質本與光碟版。從短期發展看，作為圖書情報從業人員的文獻標引工具，《中分表》仍然會出版紙質本，但不會成為主要發佈形式。

二、《中分表》的機讀數據化

隨著電腦在圖書情報領域的普及與應用，MARC 機讀目錄格式系列為《中分表》的數位化提供了實現支援。《中分表》的機讀目錄格式包括《中圖法》機讀數據格式（CLCMARC）與主題規範資料格式兩部分構成。《中圖法》編委會自 1999 年開始研製《中圖法》的機讀目錄格式，2000 年根據國際圖聯分類法研究小組頒佈的 UNIMARC 分類資料格式，研製出 CLCMARC（國家圖書館《中國圖書館分類法》編輯委員會，2006）。主題詞規範資料是基於《中國機讀規範格式》（China MARC Format of Authorities），補充了主題詞對應的《中圖法》的類號欄位。在萬維網之前的單機時代，《中分表》的機讀數據化做到了與國際接軌，《中分表》從紙本形式全面轉換到以資料庫驅動的數位化存儲、管理與維護。然而，機讀目錄格式一直以來是圖書情報領域的專有資料格式，對外開放與共用具有一定難度。雖然《中分表》在文獻資源管理集成系統中做到了與書目資料的融合，但從資料開放程度上看，《中分表》的機讀數據仍然是局限在圖書館專有資訊系統裡的封閉資料。

三、《中分表》的電子版發行

《中分表》的機讀數據化帶來了電腦可處理的資料基礎，為《中分表》的電子版發行做好準備。2005 年《中分表》第二版發佈紙質本的同時發佈了電子版，以光碟為介質的《中分

表》電子版系統。《中分表》的電子版是可獨立運行的 Windows 應用和在局域網內部署的服務系統，後臺的《中分表》資料與應用系統捆綁在一起對外發行。《中分表》的電子版實現了無紙化的使用環境，對使用者而言，比起翻閱紙質本，多個聯動顯示視窗能夠快速有效地進行主題詞檢索和瀏覽，《中分表》的使用效率大大得到提升。雖然《中分表》的資料隨電子版對外開放，但光碟介質分發時與資料捆綁，對《中分表》資料更新與維護造成了斷裂。但是，值得一提的是，《中分表》的電子版光碟中附帶一組 HTML 編碼的網頁檔，借助 HTML 超連結功能，能夠模擬出《中分表》紙本瀏覽效果。原本只是作為用戶的參考檔，但這組靜態 HTML 檔包含了《中分表》分類體系與主題詞兩部分的全部資料，這讓詞表研究者與技術開發人員能夠獲得到最直接的《中分表》資料檔案。通過對 HTML 頁面元素解析，可以讓《中分表》提取出研究所用的部分結構與主題詞成為可能。從某種意義上講，這可視為《中分表》詞表資源整體對外開放的最早方式。

四、《中分表》的網路版上線

為了順應網路環境的資訊資源組織、檢索與利用新需求，《中分表》的網路版是《中分表》與時俱進，走向開放的一個里程碑。《中分表》的網路版實現了通過瀏覽器直接進行訪問與檢索。《中分表》的網路版延續了電子版的多視窗聯動理念，將類目與主題詞的互動在檢索瀏覽過程中充分調用起來，也體現了分類—主題雙向對照的豐富語意關係。《中分表》的網路版需要註冊之後才可以瀏覽檢索，免費用戶可查看到 3 級類目，付費用戶（按年付費）則可以查看到全部類目資料。考慮到《中分表》版權與資料濫用等問題，《中分表》的網路版不提供直接下載資料功能，出於編表、研究實驗以及應用系統開發等目的，可通過申請流程，選擇性下載特定類目與主題詞的 MARC 或 XML 資料。《中分表》的網路版提供的 XML 資料是從 MARC 欄位直接轉換格式而來的，進一步研究使用的話，需要進行相應的解析處理。除此之外，《中分表》的網路版還提供與 OPAC 的對接檢索，支援用戶對分類款目與主題詞進行評注，體現了一定的系統開放性與使用者參與性。《中分表》的網路版在研製過程中保留了電子版既有的優點之外，跳出與書目資料系統的固有約束，以網路化新姿態面向公眾提供知識體系，這是開放道路上的又一次進步。《中分表》的網路版一方面鞏固了其在文獻主題標引與組織的專業地位，又以詞表資源實體的形式在網路上提供基礎瀏覽與檢索，這也是開放共用的體現。

五、《中分表》的術語服務試驗

雖然《中分表》的網路版是目前對外開放的最新形態，但圍繞《中分表》的細粒度開放服務的研究實驗一直在進行中。術語服務 (Terminology Service, TS) 是近些年詞表開展網路化服務應用的一種主流形式 (Tudhope, Koch & Heery, 2006)。與詞表資料整體發佈不同的是，術語服務深入到詞表中每個主題概念，為使用者提供基於知識單元的語意檢索與主題資源發現等深度服務。術語服務與關聯數據是一拍即合的技術應用組合，關聯數據為詞表的術語服

務提供了資料規範、發佈規則與技術路徑。《中分表》目前已經實現了主題詞規範資料部分的術語服務試驗，該試驗系統在中國國家圖書館局域網裡演示運行，未來會考慮對外全面發佈。

從《中分表》的紙本形式、機讀數據化、電子版發行、網路版上線以及術語服務試驗等主要階段可以看出《中分表》發展的開放化脈絡。從詞表自身發展與應用服務角度看，《中分表》的開放化是一個技術問題，其底層的資料需要不斷適應新技術環境的資源描述與組織要求，跟上技術的變化才能保證開放的同步性。但是，《中分表》的開放化又不僅僅是單純的技術問題，詞表資料使用與許可策略的不明朗背後一直以來是圖書情報領域對詞表爭論所有權的態度問題。在開放資料的網路大背景下，詞表愈開放，才會引導出更多的應用實踐。開放是連同《中分表》在內的中文詞表共同面臨的問題。

《中分表》走向開放化進程中，對其改造以適應新環境的自身發展問題得到了充分的重視。《中分表》不是在真空中自我鍛造的，而是與資訊資源環境發展緊密相關。在開放化的基礎上，對《中分表》進行語意化升級改造的觀點是下一節論述的核心思想。

《中分表》的語意化

《中分表》的開放化目標實現，需要與時俱進地升級改造自身的詞表結構與功能。《中分表》的結構是「分類—主題」雙向對照表，其實質是同一主題概念的分類號與主題詞兩種檢索標識的轉換系統（張琪玉，1997）。類號標識與語詞標識都是主題概念的表示形式，那麼以概念為中心的語意化建模與語意網理念是一致的。圍繞《中分表》的語意化改造措施是採用語意網相關技術與方法對詞表進行語意表徵。在語意網發展早期，電腦領域推出的本體（ontology）以及 W3C 的網路本體語言（Ontology Web Language, OWL）（W3C OWL Working Group, 2012）曾經一度是主題詞表升級改造的目標指向。一些學者對《中分表》的 OWL 表徵研究進行過深入的理論建模討論（Panzer & Zeng, 2009）。《中分表》分類體系的層級使用 OWL 來表徵不存在太大問題，但對類號標識、類號組配以及分類—主題映射關係等諸多詞表細節的表達難度較大。

《中分表》主題詞規範資料使用 OWL 來表示又過於嚴格。雖然 ontology 與主題詞表有相通之處，但兩者的出發點並不相同。本體構建的直接目標指向是人工智慧要求的基於嚴謹知識結構的自動推理。OWL 語言的強語意表達與嚴格的推理約束等特點，將《中分表》整體改造為 OWL 本體在目前來看是不太現實的發展思路。換個角度看，基於《中分表》的主題詞規範資料與部分學科分類機構進行領域本體的構建倒是一條可行的詞表衍生思路。關聯數據的適時出現，讓詞表擁有者與研究者意識到：詞表向本體轉換的願景是長期存在的。現階段將詞表轉化成 RDF 詞彙集，促進基礎語意關聯的資源聚合與發現，這才是更為現實可行的現實思路。

2009 年 W3C 的簡單知識組織系統（Simple Knowledge Organization Systems, SKOS）推薦標準正式出臺（Miles & Bechhofer, 2009），這意味著圖書情報領域的詞表全面適應語意網

的開始。SKOS 以主題概念為中心，定義了主題詞表結構中最基本的語意屬性，它在現實詞表與本體知識庫之間搭建了漸進的橋樑，以付出較小的轉換成本代價來實現詞表通用部分的語意化表徵，於此同時為未來預留了本體改造空間。SKOS 將概念空間與詞彙空間做出了明確的劃分處理，概念相關的詞彙關係梳理則留給了 SKOS-XL 擴展。SKOS 應用於主題詞表的表徵是成功的，主流的詞表均提供部分或全部的 SKOS 表徵資料。

《中分表》語意化改造的一種現實思路是從主題詞規範資料入手，使用 SKOS 對主題詞規範資料進行描述。在此基礎上，借助 SKOS 的上下位類關係實現了分類體系資料的部分表徵。以此為行動綱要，《中分表》的主題詞規範資料部分被定義為 `skos:ConceptScheme`，其 URI 為 `http://cct.nlc.cn/Subject#conceptScheme`。主題詞概念被定義為 `skos:Concept`。主題詞概念的首選標籤有中文簡體、中文拼音以及英文三種表達形式；非首選標籤表示代項 (D) 均為中文相關詞彙。主題詞概念的參照 (C) 關係使用 `skos:related` 表示，屬 (S) 與分 (F) 關係使用 `skos:broader/skos:narrower` 表示。需要說明一點，由於 SKOS 定義的 `skos:broader/skos:narrower` 是直接上下位關係，因此主題詞的直接屬分關係使用 `skos:broader/skos:narrower`。與之類似的 `skos:broaderTransitive/skos:narrowerTransitive` 具備可傳遞推理，主要用於族首詞體系的構建。族首詞是主題詞規範資料中一類重要主題概念，使用 `skos:isTopConcept` 表示，與所屬的概念體系之間的關係使用 `skos:hasTopConcept` 表示。主題詞的類號目前使用 `skos:notation` 表示分類—主題的映射結果。通過主題詞的 `skos:notation` 屬性可以構造出相應的類號標識體系，形成《中分表》的分類主幹。由於 `skos:notation` 僅是普通文本，未來考慮到分類-主題的動態映射與類號組配構造，還需進一步擴展《中分表》分類體系的系統化語意建模。

以下是「敦煌石窟」主題詞數據及 SKOS 描述片段。石窟是一個族首詞，族性體系只有一級，其下羅列各類石窟。因此，敦煌石窟的族 (Z) 項與屬 (S) 項相同，這只是一種特例情況。

表 1 「敦煌石窟」主題詞數據及 SKOS 描述片段

<pre><skos:Concept rdf:about="http://cct.nlc.cn/Subject/S017466#concept"> <skos:inScheme rdf:resource="http://cct.nlc.cn/Subject#conceptScheme"/> <skos:prefLabel xml:lang="zh">敦煌石窟</skos:prefLabel> <skos:prefLabel xml:lang="zh-pinyin">dun huang shi ku</skos:prefLabel> <skos:prefLabel xml:lang="en">Dunhuang Grottoes</skos:prefLabel> <skos:altLabel xml:lang="zh">敦煌莫高窟</skos:altLabel> <skos:altLabel xml:lang="zh">莫高窟</skos:altLabel> <skos:broader rdf:resource="http://cct.nlc.cn/Subject/S067358#concept"/> <skos:broaderTransitive rdf:resource="http://cct.nlc.cn/Subject/S067358#concept"/> <skos:notation>K879.21⑨</skos:notation> </skos:Concept></pre>	<pre>dun huang shi ku 敦煌石窟 Dunhuang Grottoes K879.21⑨ D 敦煌莫高窟 D 莫高窟 Z 石窟 S 石窟 記錄控制號：S017466</pre>
---	---

族首詞的表徵以「美術」為例進行說明。美術作為族首詞，整個詞族包含 11 個二級主題概念，每個二級主題概念又包含若干個層級的主題概念，例如，版畫又包含了兩個層級的分類。通過 skos:broaderTransitive/skos:narrowTransitive 的關係傳遞，最終由電腦自動推理出整個族首詞體系。下表中間列是美術主題概念，相應的 SKOS 表徵片段位於左側列，右側列是美術的二級類——版畫。

表 2 「美術」足首詞及二級類「版畫」舉例

<pre> <skos:Concept rdf:about="http://cct.nlc.cn/Subject/S051901#concept"> <skos:inScheme rdf:resource="http://cct.nlc.cn/Subject#conceptScheme"/> <skos:prefLabel xml:lang="zh">美術</skos:prefLabel> <skos:prefLabel xml:lang="zh-pinyin">mei shu</skos:prefLabel> <skos:prefLabel xml:lang="en">Fine arts</skos:prefLabel> <skos:narrower rdf:resource="http://cct.nlc.cn/Subject/S001840#concept"/> <skos:narrowerTransitive rdf:resource="http://cct.nlc.cn/Subject/S001840#concept"/> <skos:narrower rdf:resource="http://cct.nlc.cn/Subject/S015764#concept"/> <skos:narrowerTransitive rdf:resource="http://cct.nlc.cn/Subject/S015764#concept"/> <skos:related rdf:resource="http://cct.nlc.cn/Subject/S051902#concept"/> <skos:topConceptOf rdf:resource="http://cct.nlc.cn/Subject#conceptScheme"/> <skos:notation>J06</skos:notation> </skos:Concept> </pre>	<pre> mei shu 美術 Fine arts J06 · 版畫 · 雕塑 · 工藝美術 · 繪畫 · 建築藝術 · 攝影藝術 · 書法 · 業餘美術 · 原始美術 · 民間美術 · 書畫藝術 C 美術創作 記錄控制號： S051901 </pre>	<pre> ban hua 版畫 Graphic art J217 Z 美術 S 美術 · 版畫 · · 玻璃版畫 · · 麻膠版畫 · · 木刻 · · · 浮水印木刻 · · · 套色木刻 · · · 油印木刻 · · 石版畫 · · 絲漏版畫 · · 桃花塢木版年畫 · · 銅版畫 · · 楊柳青木版年畫 記錄控制號：S001840 </pre>
--	--	---

《中分表》的主題詞規範資料以集合方式使用 skos:Collection 表示，主要包含 7 個主題概念集合：普通主題詞集合、人名主題詞集合、地名主題詞集合、團體會議主題詞集合、題名主題詞集合、家族主題詞集合以及漢語複雜主題詞集合。每個主題詞集合內部的主題詞是無序的。

目前《中分表》所有的主題詞規範資料已經實現了 SKOS 語意表徵，通過主題概念體系與概念集合組織在一起。《中分表》分類體系部分的語意化表徵還有待進一步實驗，一種可能的研究思路是借助 OWL 的形式化邏輯實現複雜的類號組配與複分推理等。

《中分表》的語意化改造經驗告訴我們，《中分表》的分類主題一體化結構並不適合整體改造。從主題概念出發，對分類—主題映射關係應做邏輯性拆解。在分類與主題兩部分語意

化改造完成準備後之後，再進行兩部分對應表的耦合式集成。關聯數據及相關技術方法為《中分表》能夠儘快參與到資料網路中提供了落地方案，《中分表》的主體資料已經轉換為 RDF 詞彙集，下一步進入關聯發佈與服務應用階段。

《中分表》的關聯化

《中分表》的編制初衷就是基於同一概念的類號標識與主題標識之間的等同映射關係。從這個意義上講，《中分表》自身就是《中圖法》與《漢語主題詞表》兩部詞表的關聯集成產物。當前探討《中分表》的關聯化問題，實則包含兩方面：一是知識組織體系的語意交互操作研究，即詞表的相容轉換。這方面一直以來是詞表研究的重要議題，通過對《中分表》與 DDC、UDC 以及其他綜合詞表的對照映射，借由詞表之間的關聯對應，最終服務於文獻資源的共用與集成。二是符合關聯數據發佈的關聯要求，能夠遵守語意映射規則，實現事物主題概念的關聯。

在 Tim Berners Lee 提出的關聯數據 4 條原則中，最後一條原則是：開放的 RDF 資料集要成為關聯數據必須與其他 RDF 資料集通過 URI 的連結指向，進而發現更多的相關資料集。關聯數據 5 星評價的第 5 顆星是指關聯起來的 RDF 資料集 (Berners-Lee, 2009)。

以關聯數據的相關原則與評價為參考依據，《中分表》的關聯化維度可以從以下幾方面考慮：一是基於詞表相容轉換的研究結果，使用關聯數據方法建立 RDF 詞彙集之間的語意關聯；二是隨著書目資料的關聯發佈，在書目資料與《中分表》之間原有的主題標引之上建立關聯；三是《中分表》的主題詞規範資料部分可以考慮與虛擬國際化規範文檔 (Virtual International Authority File, VIAF) 等國內外權威的名稱規範資料集建立關聯。

《中分表》的關聯化除了理論方法層面的討論之外，還需要技術系統的支撐。在《中分表》的關聯化中，語意存儲、Web Service 與 SPARQL Endpoint 檢索介面、JSON-LD/RDF/XML/NT/N3 等序列資料格式的轉換輸出也是必備的服務元件。

目前《中分表》對主題詞部分的 SKOS 表徵資料，採用 Fuseki+TDB 技術架構，已經實現了術語服務風格的網路發佈，能夠檢索《中分表》的每個主題概念並進行資料視覺化 (Fan, Bu & Zou, 2013)，但距離真正意義上的詞表關聯化還有一定距離。

《中分表》的未來

《中分表》的開放化、語意化與關聯化研究是層層遞進的：開放化是前提，在《中分表》開放化進程中研究語意化；語意化側重以概念為中心的語意建模，遵守 W3C 頒佈的相關語意網標準與協定；《中分表》的開放化與語意化共同促進了關聯化。目前《中圖法》編委會正在著手論證《中分表》註冊與關聯數據服務平臺功能需求，不久的將來會推出基於關聯數據的《中分表》詞表服務。

未來《中分表》將以更開放的姿態，發揮其作為中文詞表的帶頭作用，採用語意化深入與開放關聯化的並行發展思路，以建設成為中文詞彙集中樞（Chinese Vocabulary Hub）為目標而繼續前進。

致謝

感謝《中圖法》執行主編卜書慶老師以及詞表組成員的研究合作與支持。

參考文獻

- Berners-Lee, T. (2009, June 18). Linked Data. *Design Issues – W3C*. Retrieved from <https://www.w3.org/DesignIssues/LinkedData.html>
- Cyganiak, R., & Jentzsch, A. (2014). *Linking Open Data Cloud Diagram*. Retrieved from <http://lod-cloud.net/>
- Fan, W., Bu, S., & Zou Q. (2013). Semantic visualization for subject authority data of Chinese Classified Thesaurus. In Sylvie Davies (Chair), *Classification and visualization: interfaces to knowledge: proceedings of the International UDC Seminar*. Symposium conducted at the meeting of UDC Consortium, Hague, Holland. doi: 10.13140/2.1.5007.0087 Retrieved from https://www.researchgate.net/profile/Jason_Zou/publication/258484571_Seantic_visualization_for_subject_authority_data_of_Chinese_Classified_Thesaurus/links/54672d160cf20dedafcdf77e.pdf
- Miles, A., & Bechhofer, S. (Eds.) (2009). SKOS Simple Knowledge Organization System Reference. Retrieved from <https://www.w3.org/TR/2009/REC-skos-reference-20090818/>
- Panzer, M., & Zeng, M. L. (2009, October). *Modeling classification systems in SKOS: some challenges and best-practice recommendations* (pp. 3-14). Paper presented at the DCMI International Conference on Dublin Core and Metadata Applications, Seoul, Korea.
- Schaible, J., Gottron, T., Scheglmann, S., & Scherp, A. (2013). *Lover: support for modeling data using linked open vocabularies*. Proceedings of the Joint EDBT/ICDT 2013 Workshops, 89-92. doi: 10.1145/2457317.2457332.
- Tudhope, D., Koch, T., & Heery, R. (2006). *Terminology services and technology: JISC state of the art review*. United Kingdom: Joint Information Systems Committee. Retrieved from <http://www.ukoln.ac.uk/terminology/JISC-review2006.html>
- W3C OWL Working Group (Eds.) (2012). *OWL 2 Web Ontology Language Document Overview*. Retrieved from <https://www.w3.org/TR/owl2-overview/>
- Zeng, M.L., & Chan, L.M. (2004). Trends and issues in establishing interoperability among knowledge organization systems. *Journal of the American Society for information science and technology*, 55(5), 377-395. doi: 10.1002/asi.10387
- 【《中國圖書館分類法》編輯委員會（2012）。《中國分類主題詞表》概況。檢自：<http://clc.nlc.cn/ztfzfbgk.jsp>】
- 【《Chungkuo tushukuan fenleifa》 pienchi weiyuanhui (2012). 《Chungkuo fenlei chutituzupiao》 kaikuang. Retrieved from <http://clc.nlc.cn/ztfzfbgk.jsp>】
- 國家圖書館《中國圖書館分類法》編輯委員會（編著）（2006）。《中國分類主題詞表》（第二版）及其電子版手冊（80頁）。中國北京市：北京圖書館出版社。

【Kuochia tushukuan 「Chungkuo tushukuan fenleifa」 pienchi weiyuanhui. (editors) (2006). 「Chungkuo fenlei chutitzupiao」 (second edition) chi chi tientzupan shoutse, (80). Beijing, China: National Library of China Publishing House.】

張琪玉 (1997)。情報語言學基礎 (增訂二版) (246 頁)。中國武漢市：武漢大學出版社。

【Chang, Chi-Yu (1997). Ch'ingpao yuyensueh chichu (Second edition), (pp. 246). Wuhan, China: Wuhan University Press.】

Towards Open, Semantic, and Linked Chinese Classified Thesaurus

Wei Fan

Associate Professor, Department of Information Management,
Sichuan University, China (P.R.C.)
E-mail: fanw@scu.edu.cn

Keywords: Chinese Classified Thesaurus; Linked Data; Vocabulary Dataset

【Abstract】

Chinese Classified Thesaurus (CCT) is one type of best practices of Chinese information retrieval language. As the information environment changes, vocabularies need to be advance with the time. CCT has an upgrading process to become open, semantic and linked. The article takes open-semantic-linked as three steps and reports the current development of CCT. In the Linked Data context, CCT has already transformed subject authority data into RDF datasets and experimented the technology solution with Fuseki and TDB. In the future, CCT will focus on linked vocabulary dataset and knowledge service.

【Long Abstract】

Background

Chinese Classified Thesaurus is the typical representative of research and practice of information retrieval language in Mainland China, as well as an important Chinese knowledge organization system. The subject of *Chinese Classified Thesaurus* is mainly a thesaurus framework. Subject data is also an important composition covered by subject authority data of National Library of China. *Chinese Classified Thesaurus* possesses rich subject concepts and covers the subject concepts of various fields, such as philosophy, social sciences, natural sciences, and engineering technology. It collects more than 50,000 classification categories, 110,000 subject concepts, 60,000 pre-coordinated subject strings, and 3,000 entry terms (*Chinese Library Classification* Editorial Board, 2012). This study started with three questions: (1) Why is *Chinese Classified Thesaurus* generally perceived as a literature indexing tool in the field of library and information? (2) Why cannot *Chinese Classified Thesaurus* keep up with or meet

the descriptive and organizational requirements of diversified network information resources? (3) Why aren't the vocabularies, concepts, and semantic relationships in *Chinese Classified Thesaurus* applied by network third party and utilized by service developers?

Discussions

In order to answer these three questions, this study investigated the development of *Chinese Classified Thesaurus* from three aspects: openness, semantization, and linkage of thesaurus.

Openness of *Chinese Classified Thesaurus*

Openness creates possibility and convenience for the exchange and sharing of resources. The antonym of openness is closeness. The isolated and information island has always been a major issue. As the classification subject integration indexing tool of the largest scale in Mainland China, *Chinese Classified Thesaurus* mainly experienced the following five stages to step from closeness to openness:

(1) Paper-based form of *Chinese Classified Thesaurus*

The Hardcopy of *Chinese Classified Thesaurus* First Edition (1994) was published, and then both the hardcopy and CD of *Chinese Classified Thesaurus* Second Edition (2005) were published. From the perspective of short-term development, the hardcopy of *Chinese Classified Thesaurus*, which is the literature indexing tool of personnel engaging in the field of library and information, will still be published. However, it will not become the main release form.

(2) Machine-readable digitization of *Chinese Classified Thesaurus*

The *Chinese Library Classification* Editorial Board initiated the R&D of machine-readable cataloging format of *Chinese Library Classification* in 1999. In 2000, the CLCMARC (National Library of *Chinese Library Classification* Editorial Board, 2006) was developed according to the UNIMARC classification data format promulgated by the Research Group of International Federation of Library Associations and Institutions (IFLA). Subject authority data supplemented the corresponding *Chinese Library Classification* class number and field of subjects based on *China MARC Format of Authorities*. Although *Chinese Classified Thesaurus* achieved the integration with bibliographic data in the literature resource management integration system, from the perspective of level of data openness, the machine-readable data of *Chinese Classified Thesaurus* was still closed data limited to specific library information systems.

(3) Electronic distribution of *Chinese Classified Thesaurus*

In 2005, both the hardcopy and electronic version of *Chinese Classified Thesaurus* Second Edition were published. From user's perspective, compared with reading hardcopy, using an

additional linkage display window enables them to rapidly and effectively search and browse subjects. A HTML-encoded web page file is enclosed with the electronic version CD of the Chinese Classified Thesaurus. The hyperlink function of HTML could be used to simulate hardcopy browsing effect of Chinese Classified Thesaurus. In a sense, this can be regarded as the first model of external openness of overall thesaurus resources for the Chinese Classified Thesaurus.

(4) Web version of Chinese Classified Thesaurus

The web version of the *Chinese Classified Thesaurus* extends the multi-windows linkage concept of electronic version to fully reflecting the interaction between categories and subjects during the searching and browsing processes, as well as embodies the abundant semantic relationships of two-way comparison of categories—subjects. Providing fundamental browsing and searching functions online in the form of thesaurus resource entity is also an exemplification of openness and sharing.

(5) Terminology service testing of *Chinese Classified Thesaurus*

The terminology service testing of subject authority data was completed in *Chinese Classified Thesaurus*. This testing system is demonstrated and operated in the area network of National Library of China, and will be considered to be fully released externally in the future.

Semantization of *Chinese Classified Thesaurus*

A realistic way of thinking of semantic transformation of *Chinese Classified Thesaurus* is to start from subject authority data and to use SKOS to describe subject authority data. The semantic transformation measures revolving around *Chinese Classified Thesaurus* are to use Semantic Web-related techniques and methods for semantic representation of thesaurus. In the early development of Semantic Webs, ontologies and Web Ontology Language (OWL) of W3C promoted by the computer field (W3C OWL Working Group, 2012) were once the target/direction of upgrade. At the current stage, the realistic and practical idea is to convert thesaurus into RDF vocabulary sets to promote the resource aggregation and discovery of basic semantic associations.

Subject authority data of *Chinese Classified Thesaurus* is defined as skos:ConceptScheme, and subject concept is defined as skos:Concept. Because the skos:broader/skos:narrower defined by SKOS is a direct upper and lower relationship, skos:broader/skos:narrower is used as the direct subordinate relationship of subjects. Top term is an important subject concept in subject authority data, and skos:isTopConcept is used to represent it. skos:hasTopConcept is used to represent the relationship between top term and the concept system to which it belongs. For the class notation of subjects, skos:notation is currently used to

represent the mapping result of classification-subject. The subject data of *Chinese Classified Thesaurus* has been converted into RDF vocabulary sets, and the next stage is linkage release and service application.

Linkage of *Chinese Classified Thesaurus*

The original intention of editing *Chinese Classified Thesaurus* is to reflect the equivalent mapping relationship between class number identification and subject identification based on the same concept. In this sense, *Chinese Classified Thesaurus* itself is the product of associative integration of two thesauri, *Chinese Library Classification* and *Chinese Thesaurus*. The comparison mapping of *Chinese Classified Thesaurus*, DDC, UDC and other comprehensive thesauri and the associative correspondences among thesauri eventually serve for the sharing and integration of literature resources.

Conclusion

The openness and semantization of *Chinese Classified Thesaurus* are supported by better data foundation and technologies. The focuses of future development are the development of associative data sets and knowledge service application. The openness, semantization, and linkage of *Chinese Classified Thesaurus* are developed step by step: openness is the premise, and semantization is studied during the openness process of *Chinese Classified Thesaurus*; semantization focuses on concept-centered semantic modeling and follows the standards and agreements on relevant Semantic Web promulgated by W3C; the openness and semantization of *Chinese Classified Thesaurus* jointly promote linkages. *Chinese Library Classification* Editorial Board is currently investigating the functional needs of registration and associative data service platforms of *Chinese Classified Thesaurus*. In the near future, the associative data-based thesaurus services of *Chinese Classified Thesaurus* will be promoted.

[Romanization of Chinese references is offered in the paper.]