聯合目錄——通往百萬國家典藏的窗口

陳克健、溫淳雅*

網際網路的普及，使人們可以便捷的查詢遠端資源，全球蔚為風潮的數位化計畫，更使人們可從網路上取得更多有價值的數位內容。然而，一個個異質且分散的資訊系統，對使用者而言，是福音也是負擔。如果沒有很好的機制整合分散各地的異質系統，則使用者需要個別去連結並使用其不同的檢索功能。就像圖書館裡的書，如果沒有一個良好的目錄，再好的書也會被埋沒在茫茫書海中難被查閱。

民國九十一年起，科技部前身的國科會推動執行了一系列數位典藏與數位學習的國家型科技計畫，其主要目標是將國家重要的文物典藏數位化，促進人文與社會、產業與經濟的發展。參與單位包括國家重要的研究單位及政府機構、博物館、大學等，其典藏縱貫古今、橫跨十餘種不同內容學科領域，例如生物、地質、檔案、書畫、器物、人類學、新聞……等，人文與自然兼具，累積十餘年更達到 500 萬筆以上的龐大資料，可長期作為學術研究、教學與商業加值應用、文創產業素材的資源庫。但是如果沒有共通的目錄，各機構辛苦建立的資訊系統，將不易被人得知及檢索。

這些資料各自採取其學科專門領域的後設資料（Metadata）標準與不同的資料庫系統，呈現方式五花八門，初次接觸的使用者，很容易眼花撩亂難以下手。每一資料庫系統並需一一學習、適應不同的檢索方式。為了減低使用者的負擔，國家型計畫旋即規劃建置一個整合的目錄及其檢索系統：數位典藏聯合目錄 http://catalog.digitalarchives.tw/。聯合目錄擔負著學科統整與跨領域整合的目的，提供跨主題、跨機構整合檢索的便利功能，讓使用者可透過

* 陳克健，中央研究院資訊科學研究所研究員；溫淳雅，中央研究院數位文化中心助理研究員。

聯合目錄檢索瀏覽 500 萬筆以上縱貫古今、橫跨十餘種不同內容學科領域的數位化內容資源的成果。

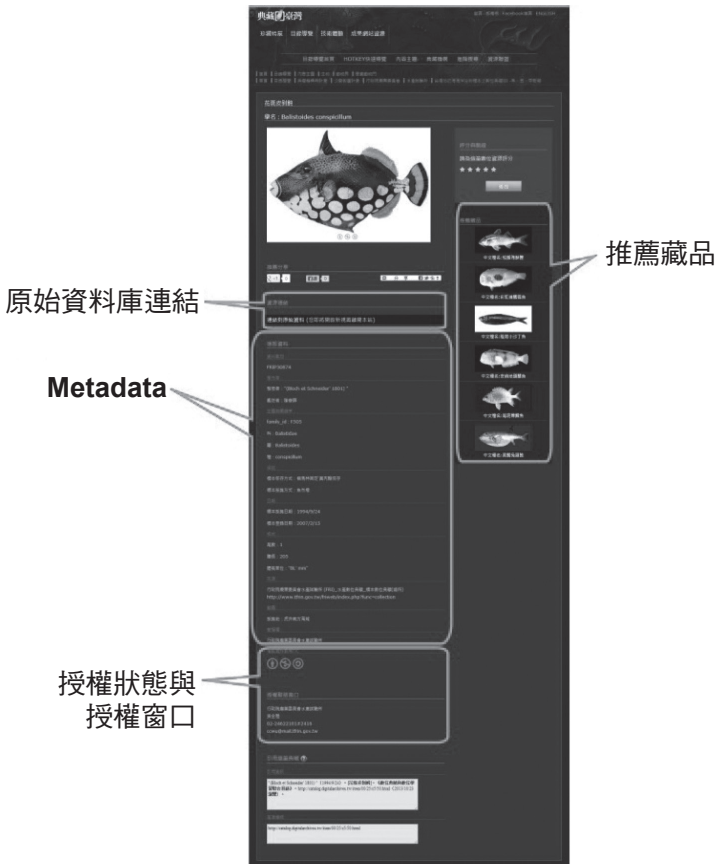
一、聯合目錄目的與功能簡介

什麼是聯合目錄？聯合目錄早先是為了檢索圖書館書目資料而產生的館藏目錄資料庫系統，當使用者有跨館際查詢圖書資料的需求時，聯合目錄系統便應運而生。而數位典藏聯合目錄，則是將相同的概念應用至典藏數位化資源上；因此，不管原件是器物、考古挖掘、生物標本等實體物件；書畫、攝影作品等平面影像；檔案卷宗、善本古籍的文字內容；或是歌曲、戲劇等影音內容，聯合目錄都可透過對藏品共通的描述方式，統合檢索後，將結果提供給使用者。

但其實，每個不同機構針對不同數位化物件所採用的藏品描述後設資料各有不同標準，聯合目錄便為此規劃採用了專為描述網路電子資源的後設資料（Metadata）標準：都柏林核心集（Dublin Core），它具有「簡單易產生或維護、通用易了解的語意、全球通用、彈性高」的特性，便於聯合目錄將各異質資料庫所採用的 Metadata 與 Dublin Core 進行欄位定義上的比對，從中汲取摘要資訊，以統一標準化格式匯入聯合目錄中。目前數位典藏聯合目錄中，共收錄超過 500 萬筆數位化物件的後設資料，並繼續成長增加中。

有了巨量資料，後續便是追求網站流量。透過一般使用者利用搜尋引擎檢索的行為，導引使用者至聯合目錄，並吸引部分使用者在瀏覽過後成為固定使用者。因此，聯合目錄一直以來都執行搜尋引擎最佳化（Search Engine Optimization，簡稱 SEO）策略，進行包括網頁正規化；自數位資源的後設資料中抽取三到六個較具代表性的關鍵詞放入 Meta Tag；設計自動識別網頁 <Title> 標籤；主動提交 Sitemap 至 Google 網站管理員工具，提供搜尋引擎完整圖片與後設資料的索引清單等工作，以提高聯合目錄數位資源頁面於搜尋引擎中的曝光度與排名。實際上聯合目錄的流量組成，一直以搜尋引擎為大宗，經常能保持在 75% 以上的使用者是透過 Google、Yahoo 等搜尋引擎引導進入聯合目錄。

使用者透過檢索結果，連結到聯合目錄數位資源頁面如圖一所示，這是聯合目錄展現數位資源的重要頁面，頁面上會提供該筆資源的 Metadata、數位化檔案預覽、分類架構，還有站內其他資源推薦，也提供連結到原始資料

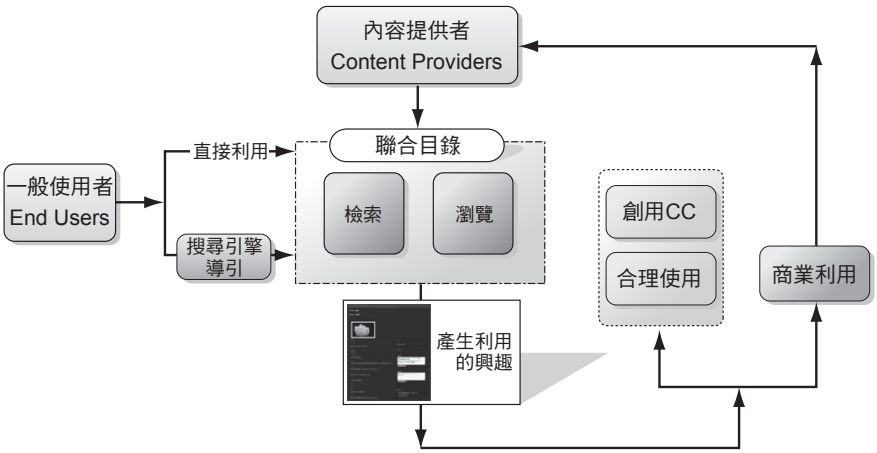


圖一　聯合目錄數位資源頁面設計

庫的功能，以增加資料來源網站（即參與計畫的典藏機構）的流量，並讓使用者能接觸到專業資料庫，形成兩者之前的橋梁。

在瀏覽了數位資源的 Metadata 後，使用者可能會產生興趣，想做進一步的加值利用。頁面下方展示了該筆資源的授權狀態與授權窗口，在著作權規範的合理使用範圍內，就可逕行利用；例如聯合目錄中有部分標註創用 CC 的數位資源，可按照頁面上「姓名標示」、「非商業性」、「相同方式分享」或「禁止改作」等標記，按創用 CC 的規範加以利用。若是有商業利用的需求，則透過授權聯絡窗口聯繫提供內容的典藏機構，依商業利用的形式、對象或範圍，洽談商業授權。（圖二）

而整體聯合目錄網站的建置，最初於 2004 年 8 月推出，其後歷經兩次重要的改版，皆是為了改善系統、提供更佳的使用者體驗。目前，聯合目錄



圖二　聯合目錄運作機制圖

圖三　聯合目錄首頁

屬於整體計畫成果入口網「典藏臺灣 digitalarchives.tw」的「目錄導覽」，以「內容主題」、「典藏機構」兩種分類方式為主要分類瀏覽路徑，首頁如圖三所示是以「聯合目錄精選藏品」為主，精選聯合目錄資源隨機輪播，吸引使用者

點選；側欄則提供上述兩種分類瀏覽功能的快速選單，讓使用者可以直接挑選感興趣的主題或者機構瀏覽。

聯合目錄首頁上方中央的醒目位置，提供了全文檢索功能，能根據使用者鍵入的關鍵字，提示字首相同的建議關鍵詞，例如圖四輸入「關」字，便會自動推薦「關公」、「關務」等其他相關詞彙，若以「關公」為檢索詞得到如圖五的結果。如果搜尋出的結果數量太大還可以輸入其他關鍵詞作進一步檢索找到更確切需要的標的物件。系統同時也會提供搜尋結果的分類項目，讓使

圖四　搜尋功能的建議關鍵詞

圖五　全文檢索結果頁面

用者根據自己的需求選取感興趣的內容主題分類，例如想了解「關公蟹」，就點選「生物」主題；或是想看看有什麼相關的藝術創作，便點選「器物」或「書畫」類別，如此能更快篩選出所需的結果。

為了讓使用者更方便搜尋到標的物，聯合目錄除了提供上述關鍵詞搜尋功能外，我們同時對龐大資料內容進行相關知識連結的知識化工程，將文件內容中的關鍵詞與相關藏品作超連結。聯合目錄也將物件與物件之間作虛擬的相關性連結，除了能自動根據使用者的檢索詞，推薦高相關性的其他關鍵詞，提供「相關搜尋」推薦功能，可達到知識延伸的功效。例如搜尋「委陵菜」，會推薦諸如「生瓜菜」、「臺北堇菜」等內容（圖六）。也能在物件頁面推薦其他藏品，如圖一左側所顯示的各種相關魚類。

圖六　根據檢索詞提供的相關搜尋

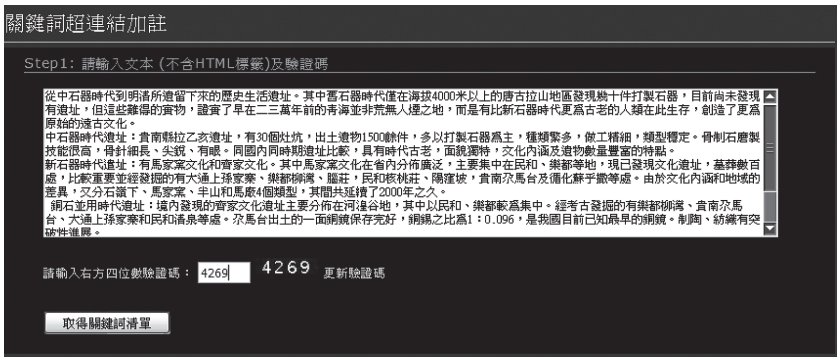
二、加強知識連結

聯合目錄迄今已收錄 500 萬筆以上數位資源，隨著數位資源的數量日漸成長，除了現階段已有的目錄瀏覽、全文檢索等功能外，針對巨量資料（Big Data）進行 Medatada 分析與關鍵詞抽取（Keyword Extraction），進而形成易於瀏覽與檢索的結構化資料（Structured Data），是進一步將數位資源更加活化利用的基礎。因此，聯合目錄設立了知識化工程相關工作，目標是能提供完整快速且精準的檢索及聯想機制，進而引起更多的發想及回應形成社群，同

時也能提供更多數位媒材合作契機。

關鍵詞抽取與編輯是知識化工程核心且基礎的工作項目，當專屬於聯合目錄的關鍵詞詞典收錄的數量越多，系統就能更有效且正確的標識數位資源的重要詞彙，提高資源的獨特重要性。而當關鍵詞累積到一定數量後，聯合目錄也開發應用工具，以將聯合目錄數位資源串聯更多文本內容。目前所開發的關鍵詞超連結標記工具（Hyperlink Tagging Tool，http://knowledge.digitalarchives.tw/），能將使用者的文章自動斷詞，並連結聯合目錄數位資源，如此既能為文章增加關鍵字重點提示，也能具象文章的描述內容。它以步驟式方法引導使用者一步一步完成文章標記，是一個簡單、容易上手的延伸閱讀工具，大家都可以使用。

使用的方法，第一步在文本輸入區塊中，貼上想要標記的文章，按照系統提示輸入驗證碼後，點擊「取得關鍵詞清單」便可得到對應聯合目錄數位資源的關鍵詞（如圖七）：



圖七　輸入文章及驗證碼（步驟一）

這個清單代表了聯合目錄關鍵詞詞典中，能對應這篇文章的詞彙，每個詞彙各有若干數位資源，有些數量多，有些數量少。不過這些關鍵詞並不一定是作者想要在文章中強調的，而列出來的關鍵詞也可以自由選擇要展示多少數位資源，因此第二步，就是要從關鍵詞清單中再行勾選（如圖八）。

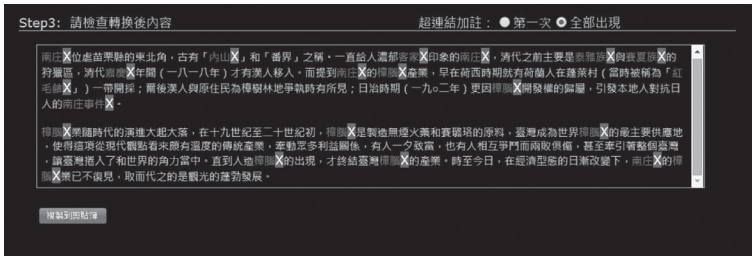
從清單中任選一個關鍵詞，會開啟挑選數位資源的視窗，可依主題類別挑選，也可連結至資料頁面觀看詳細內容，以確定與文章是否相關。但有時，有些關鍵詞對應的資料太多，使用者若一一針對每個關鍵詞挑選會耗費不少時間，因此也開發了各種篩選與自動推薦機制。可自動勾選單一主題前

圖八　選擇標記關鍵詞及其對應資源（步驟二）

20 筆，或是單頁全選……等，而聯合目錄知識化工程也透過分析文本關鍵詞對應主題的比例，自動排序推薦與輸入的文章最相關的資源（如圖九）。

圖九　依主題類別挑選關鍵詞所連結的資源（步驟二）

挑選對應的關鍵詞與數位資源之後，第三步進到檢查預覽標記內容。在預覽區塊中，勾選的關鍵詞會自動帶入連結，使用者也可以選擇這個關鍵詞在文章中每次出現都要附帶連結、或是只在第一次出現時附帶連結（如圖十）。



圖十　確認標記後的文本（步驟三）

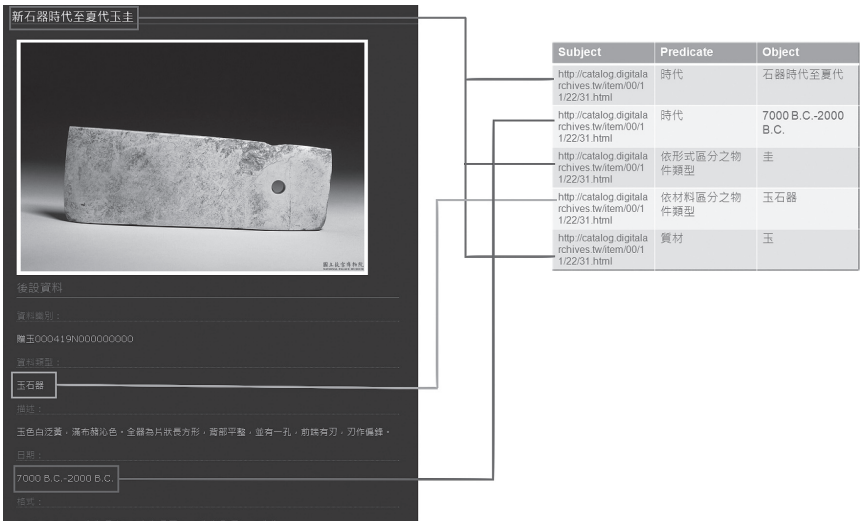
最後，按下「複製到剪貼簿」按鈕，將帶有 html 語法的文章貼到使用者想要發表的平台上發布，便可得到帶有關鍵詞延伸連結的豐富內容（如圖十一）。

圖十一　關鍵詞結果顯示列表頁

三、未來發展方向

在建立了專屬聯合目錄的關鍵詞詞典後，知識化工程要進一步推動語意分析的工程，提升聯合目錄檢索的召回率與精確率。由於聯合目錄是一個涵蓋多種內容主題的資料集合，同樣的詞彙在不同內容主題領域中可能存在歧異，例如「金」這個詞彙可以是一種質材，也可以是一種顏色。若是資料庫中，只記錄了某筆數位資源含有「金」這個詞彙，這樣只能單純記錄數位資源與關鍵詞互有關連，無法清楚表現「金」這個詞彙在藏品中的正確語意。

而如果不僅僅是單一的關鍵詞，而是用「物件－屬性－屬性值（subject-predicate-object）」三個欄位一組的物件屬性值（Triple），就能更明確表述一個句子的語意關係，更能進一步鞏固知識結構。聯合目錄針對眾多數位資源，依其內容主題特性進行分析，建置知識架構樣板（Schema Template Extraction）來進行詞彙與藏品資源對應，就能進而形成物件屬性資料庫（Triple Store），更加精確表達知識。目前，聯合目錄先以「中國器物」作為試做標的建置物件屬性資料庫，開發針對主題的物件屬性值自動化抽取對應程式，建置初級物件屬性值資料庫。再以人工檢視驗證屬性值的正確性（如圖十二）。



圖十二　抽取數位物件屬性值的示意圖

這樣的工作可進一步做許多應用，包括在聯合目錄現有數位資源資料頁面「推薦藏品」功能上，可更精確推薦相同屬性的資料。例如，原先「推薦藏品」的範圍僅限於具有相同內容主題目錄的藏品，以圖例而言，「北宋 政和鼎」因屬於「銅器與金屬器」目錄，因此「推薦藏品」便會從此目錄中挑選六筆藏品作為推薦清單（如圖十三）：

圖十三　導入物件屬性資料庫前的推薦藏品清單

而當「推薦藏品」功能導入物件屬性值之後，推薦規則改為：前三筆來自物件屬性資料庫中相同「物件類型」的藏品，後三筆才是相同內容主題目錄的藏品。圖十四顯示修改後的推薦結果，前三筆分別為「獸面紋鼎」、「西周 蟠夔紋鼎」與「明 嬌黃錐拱獸面紋鼎」，明顯與此筆藏品「北宋 政和鼎」相關性高；而後三筆藏品則是來自同為「銅器與金屬器」內容主題目錄的藏品。物件屬性值語意推薦規則的準確度優於過去的相同內容主題目錄。

圖十四　導入物件屬性資料庫（Triple Store）後的推薦藏品清單

物件屬性值（Triples）未來將以國際標準化的 RDF 格式發布，便能與全世界所有其他國家所產生的物件屬性值接軌，預期未來標準化的物件屬性值便能串聯起來形成一個龐大的語義知識網絡，可以將聯合目錄的內容經查詢國際標準物件屬性資料庫自動連結到維基百科，如此一來如果在聯合目錄看到蘇東坡的字畫就可以自動連結到維基百科看到蘇東坡的生平，從蘇東坡的生平可以進一步連結到相關的人事時地物，無窮無盡的知識就在彈指之間獲得。知識化工程還可以利用知識架構、藏品間的知識關連，繼續發展多語檢索與索引典、網頁語意標註、搜尋引擎最佳化等應用方向，提供使用者良好的瀏覽體驗，建立更友善的瀏覽環境。