藏經與佛教工具書的數位化編纂

——以CBETA電子佛典與數位經錄計畫爲例

杜正民　法鼓佛教研修學院副校長
李家名　中華佛學研究所維路資訊室
唐海立、鄭寶蓮、林心雁　國科會佛教藏經目錄數位資料庫計畫

【摘要】本文擬以 CBETA 藏經全文資料爲基礎，輔以資訊技術的運用與配合，就「佛教藏經目錄數位資料庫」及「淨土會哲線上辭典」兩部數位化工具書的進展做報告。

國科會「佛教藏經目錄數位資料庫」，目標爲建構多種語言的經錄資料庫。期待透過數位化技術，完成古人所無法整理的多語言、多版本的經錄比對，進而建立可連結全文的線上閱讀系統。本計畫以 CBETA 現有的數位典藏爲基礎，以歷代佛典經錄建構數位文獻資料，配合當代資訊科技與 XML TEI Markup 等標準規範進行整合開發，將傳統紙本工具書做資訊時代的呈現與應用。

除了數位工具書的成果介紹外，本文擬以資訊技術協助人文研究的角度出發，介紹建立數位工具書相關資源的累積、技術的應用與開展的過程。另外對佛學研究領域外，數位化工作目前所需加強的環節、資訊技術可能的開展等略作描述。

關鍵詞：數位工具書；電子佛典；經錄；佛學數位資料庫；TEI XML

一、前言

因承載資訊的媒介材料的不同，資訊時代數位化的工具書，從製作的方式、使用的資源、服務的功能、到呈現的形式，都與傳統紙本工具書有很大的差別。

在工具書的編纂過程中，數位資料質與量的累積是否足夠、編纂人員（人文領域專家）與技術人員（資訊工作者）溝通合作是否密切等都非常的重要。如果每個環節都能有效配合，則比較起紙本工具書，數位工具書應有製作時間縮短、

內容品質提升、參考功能增加、版本更新便利等優點。

本篇文章將透過經錄專案成果的介紹，來呈現新型態數位工具書的樣貌。

二、理念

1. 人文與資訊的結合——資訊技術協助人文研究

結合人文與資訊兩大領域，利用資訊技術與工具，協助佛學領域的研究與發展，是一個重要的努力方向。若能適當的應用資訊技術與工具，不但可讓專家學者的工作事半功倍，更可以提高其研究品質。

以工具書的編纂為例，若使用資訊工具來整理及準備基礎的文獻，並建立方便清楚的工作平臺，確實可減少編纂者的工作，使專家學者的精力能更專注在最重要的內容處理上。另外，藉由資訊技術大量的儲存分析功能，我們還能更進一步地依不同的主題，結合相關的網路工具，將傳統的內容做更多樣的分類及服務，使得使用者在相同的資料中，能獲得多面向的複合價值的資訊。

2. 新型態的工具書

因資料數位化與網路技術的普及，新型態的線上工具書除型式外，其不同於傳統紙本工具書的方面還有：

2.1 可整理顯示傳統字辭典無法提供的知識概念

因超連結技術的出現，加上線上動態資料顯示技術的成熟，使得工具書除了文字描述以外，還能進一步傳達：

(1) 時間資訊——詞條用法的演變。

(2) 空間資訊——詞條在不同區域的發展現象。

(3) 知識架構——詞條分類及各詞條間的層級關係。

(4) 其他隨時因新技術出現可能的開展。

2.2 與文本的連結及引用整理

傳統工具書在引文的數量、詞彙用法的舉例和其重要性的優先順序等內容上，往往受到篇幅與書本型式的限制。有了電子藏經等數位文本的資源，電子工具書或任何形式的 Metadata 均可以依使用者需要，提供足夠且清晰的文本引用、完整例句及各種排序檢索的功能。

2.3 Web 2.0 的概念——使用者也是建構者

數位化的線上工具書與傳統紙本工具書，最大的不同在於其動態的特性。除了上述的動態資料建構與資訊傳遞功能外，數位化工具書可即時更新、保留與顯示變動歷程的特性，與傳統紙本出版品的版本概念完全不同。

除了單向的資料提供服務外，動態模式還可擴大到如 Wikipedia 的使用者互動概念。藉此線上工具書更可進一步發展為雙向互動式的資料收集平臺，建立一個以工具書內容為核心的資料處理及知識匯集中心。

三、步驟與方法

「佛教藏經目錄數位資料庫專案」（http://jinglu.cbeta.org/，以下簡稱「數位經錄」／「經錄」）是國科會數位典藏國家型科技計畫之一，目標為建構多種語言經錄的數位化資料庫。期待透過數位化技術，完成古人所無法整理的多語言、多版本的經錄比對，進而建立可連結全文的線上閱讀系統。本計畫以 CBETA 現有的數位典藏為基礎，以歷代佛經版本經錄、《法寶總目錄》及《法寶義林》等建構數位文獻資料，配合當代資訊科技與 XML、TEI

Markup 等標準規範進行內容開發，將傳統紙本工具書做資訊時代的呈現與應用。

數位經錄計畫運用超文本等資訊時代的重要概念與數位典藏技術，以建置知識架構，建構資訊時代的新資料典範，突破紙型經錄的侷限，拓寬研究的範圍，是一劃時代重要學術資料庫的完成。於此，就該專案進行的步驟與方法做一簡介：

1. 漢文佛藏經錄資料庫建立

第一年計畫完成工作資料庫建檔與連結的有：房山石經、高麗藏、永樂北藏、乾隆藏、卍正藏、大正藏、佛教大藏經、中華藏、新纂卍續藏等 9 部經藏目錄。

第二年計畫完成工作資料庫建檔與連結的有：開寶藏、崇寧藏、毘盧藏、圓覺藏、趙城金藏、資福藏、磧砂藏、宋藏遺珍、普寧藏、洪武南藏、永樂南藏、縮刻藏、頻伽藏、嘉興藏、卍續藏、至元錄等 16 部經錄。

本計畫以現存藏經不同版本的經錄來源資訊，建立各個不同的資料庫，進一步將這些資料庫連結測試，初步建構整合完整的資料庫，將超越現有的傳統的資料庫模式。現有傳統的資料庫設計，皆以表格式設計，底下為實例。（見表一）

此類資料庫有諸多不足，最明顯的是當經目為一經多名，或是朝代譯者有多種稱謂時，即可能無法順利搜尋。在甲藏經是一個名稱，在乙藏經又是另一個名稱，光由文字資料的輸入，無法達到查詢的目的。

表一：傳統資料庫

□高麗藏

類別	部別	經號	經名	經名注文	卷數	起始函號	譯著者朝代	譯著者姓名
大乘經	般若部	1	大般若波羅蜜多經		600	天 1	大唐	大唐三藏法師玄奘奉詔譯
大乘經	般若部	2	放光般若經		20	菜 61	西晉	西晉三藏無羅叉-共竺叔蘭譯
大乘經	般若部	3	摩訶般若經	亦名大品般若經	27	芥 63	姚秦	姚秦三藏鳩摩羅什-共僧叡等譯
大乘經	般若部	4	光讚經		10	鹹 66	西晉	西晉三藏竺法護譯
大乘經	般若部	5	摩訶般若經抄		5	河 67	秦	秦天竺沙門曇摩蜱-共竺佛念譯
大乘經	般若部	6	道行般若經		10	淡 68	後漢	後漢月支國三藏支婁迦讖譯

再者，此類資料庫無法整體性地看出藏經的整體結構，亦無法由朝代看出各經目之間的關係。因此現藏錄資料庫之建立，除了保留傳統的方式，同時整合歷史上各朝代的各種稱謂，建立相對的關係，並建構作譯者完整資料，最後各版經目之間各種一經多名、同經異譯，皆會建立完整資料，讓每一經在各種大藏經的位置，皆能一目瞭然，同時由各朝代或譯者來查詢，亦可完整呈現彼此的關係。（見表二）

表二：權威控制後的資料舉例

欄位	內容
經碼	0000556
冊數	3
經號	0158
經名	大乘莊嚴論經
朝代	後秦
作譯者	馬鳴菩薩 造　鳩摩羅什 譯
漢名漢語拼音	Ta chuang yen lun ching
翻譯概述	Translation by Kumārajīva: between the 4th and 14th years of Hung Shih (弘始), Later Ch'in dynasty (後秦) (A.D. 402-412).T. 2151-359a:16; T. 2151-359b:26; T. 2154514c:26.
朝代碼	[D034]
作譯者碼	[B1757:1][B1113:2]
朝代校勘	
作譯者校勘	
譯經起年_中	弘始四年
譯經迄年_中	弘始十年
譯經起年_西	402
譯經迄年_西	412
譯經地	-

除了單經的資料要建構完整之外，彼此縱橫的關係亦會加入資料庫中。縱者，同一經之經、論、注、疏及相關各家說明，皆會建立關係，當任何一經被查詢時，相關的同類經皆能同時列出參照。橫者，諸版大藏經中，大經往往包含小經，小經又常常是各大經的一部分，諸如此類的關係，亦會建立相關資料，讓同一經散布各處的身分，皆能一覽無疑。

有了這樣的資料架構，不論由時代、人物或藏經分類來查詢，皆能清楚而完整，在查詢或統計時，才能鉅細靡遺地得到正確的資訊，經由正確的資訊，才能發掘出更多有價值的成果，這正是本計畫設置佛經資料庫的目的之一。

2. 經錄 Markup 標記作業

除了目前 CBETA 數位檔內經錄可用的 XML TEI 基本標記外，經錄計畫更進一步加注內容標記（content markup），以達各項規劃功能的預計目標，茲列舉譯者與經名標記如下：

2.1 **標記例** XML TEI / Authors and Translators（**譯者列舉**）

<lb n="0477c18"/>騰<note type="inline">一部一卷經</note>

<lb n="0477c19"/><item>沙門竺法蘭<note type="inline">四部一十五卷經</note></item>

2.2 **標記例** XML TEI / Sutra Title（**經名列舉**）

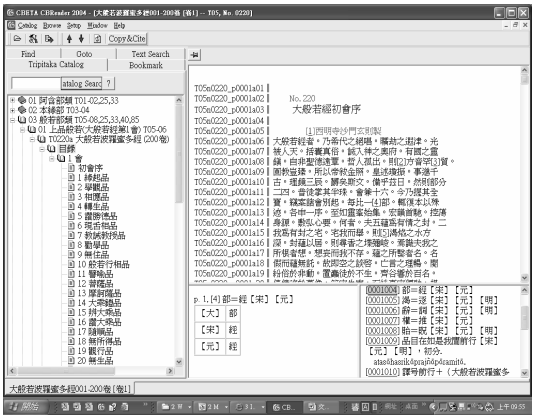
<lb n="0478a05"/>卷新附</note></item></list>

<lb n="0478a04"/><item>新舊諸失譯經<note type="inline">一百四十一部一百五十<anchor id="fnT55047801"/>一八卷五十九部七十六卷舊集八十二部八十二

3. 經文知識架構

由於良好的內容標記與經錄的搭配，可以經由目錄迅速得知該經的知識架構，如圖一所示：

圖一：CBETA 經目架構



4. 其他語言資料庫建立

4.1 巴利藏經錄資料庫建立

有關巴利語藏經的流傳，學術界研究界定為早期佛教聖典，「四部尼科耶」與北傳漢譯佛典的「四阿含」雖屬不同傳本，但經學者研究證明其中有相當多經典相互對應。這些相對應經典在佛教研究上，具有版本學、文獻學校勘、語言學等學術研究價值。因此，本計畫除建立巴利藏經錄資料庫外，為方便學者研究，並建立南傳四部尼科耶與漢譯四部阿含經之對應經群組的資料庫，以及提供巴利藏經文線上全文閱讀。

4.2 西藏藏經經錄資料庫建立

西藏大藏經傳統裝訂方式是屬傳統梵夾裝，所以與現代書本裝訂不同。欄位的建立比較詳細，並提供巴利藏與大正藏對應經連結。其中經名分成藏文、中文、梵文、羅馬轉寫四種欄位，方便學者閱讀與學習。

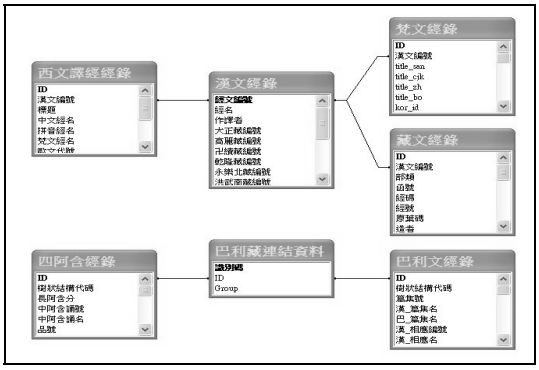
4.3 西文經錄資料庫建立

本多語言經錄資料庫除漢、巴、藏等語言外，更納入西文藏經的經錄，加上國際間的重要漢文經錄，取得授權東西方合作，進行數位化作業，貫通中西經錄文獻。提供外國學者方便查詢使用。本資料庫主要版本來源：Mahāyāna Texts translated into Western Languages. Bonn: Brill, 1986。

5. 資料庫結構

目前已完成收集梵文、藏文、巴利文與西文譯經的藏經目錄，圖二是各語系經錄的資料庫結構圖。

圖二：各語系經錄的資料庫結構圖



在上圖結構中，巴利文經錄是比較特殊的，因為巴利藏尼科耶與北傳四阿含是相對於大正藏的四部經（若包含《別譯雜阿含》則為五部經），只由四部經無法建立詳細對照關係，所以另外建立四阿含各小經的資料庫，用以和巴利藏尼科耶各小經對應。在對應資料中，除了巴利藏尼科耶與四部阿含各小經相互對應之外，也會有對應至大正藏阿含部其他相關的單經。

在資料記錄上，梵文、藏文、西文譯經除了各藏應有的經號、經名、作譯造者等基本資料外，也要有相應於漢文藏經的欄位，我們即是用此欄位與漢文大藏經連結，然而此連結並不是一對一的結構，所以在設計上採用了一對多的格式。

四、成果介紹

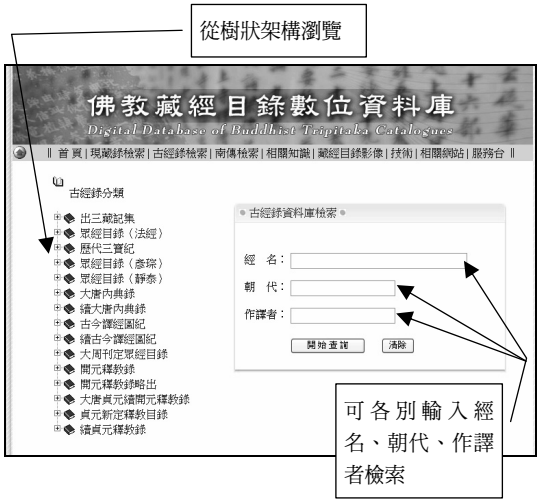
佛教經錄網頁主要功能是多語言、多版本佛典強大效能的【經錄資料庫檢索】。提供對應相關經群線上閱讀連結等相關資料的查詢，豐富並加強使用者的需求與方便性。

學者已不需出門上圖書館、翻閱目錄，輕易能從本資料庫檢索到某部經，找到佛典收錄版本冊頁出處及漢、藏、巴等相關經群對應，並同時連結全文線上閱讀。「佛教藏經目錄數位資料庫」的完成，堪稱是佛教學界的一大貢獻。

1. 漢文經錄檢索

圖三：現藏錄檢索，現藏錄可綜合總資料庫檢索或單一版本資料庫檢索「經名」、「作譯者」及「朝代」。

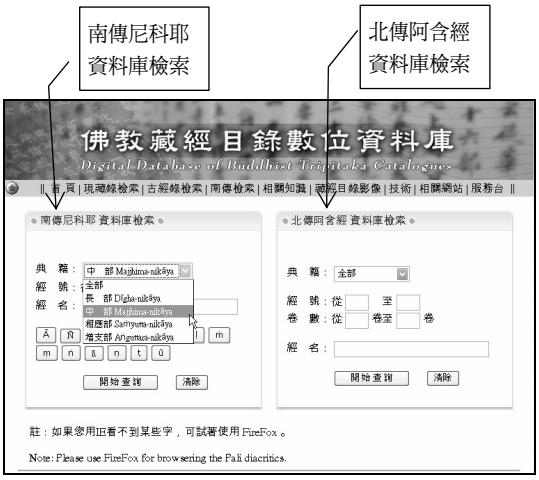
圖四：古經錄檢索，可從下圖樹狀架構或直接輸入「經名」、「作譯者」及「朝代」進行檢索。



2. 巴利藏經錄檢索

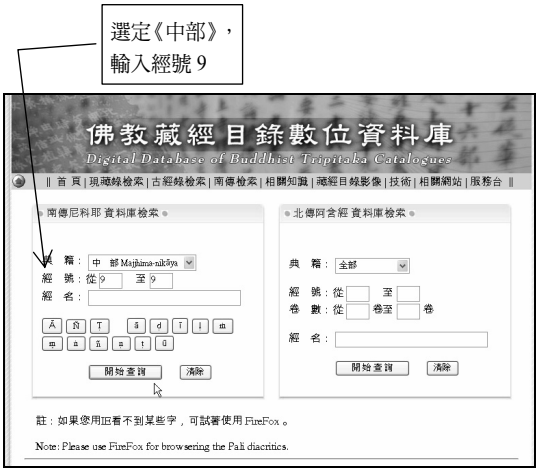
如圖五巴利藏檢索系統分成「南傳尼科耶」與「北傳阿含經」資料庫檢索。「南傳尼科耶」資料庫檢索可分成「經號」及「經名」二種檢索方式；「北傳阿含經」資料庫檢索則可分成「經號」、「經名」及「卷數」檢索。二者皆可選定一經或多經檢索。以下以「南傳尼科耶」資料庫檢索為例，「北傳阿含經」資料庫檢索大致相同，不再另做舉例說明。

圖五：巴利藏資料庫檢索



2.1 經號檢索

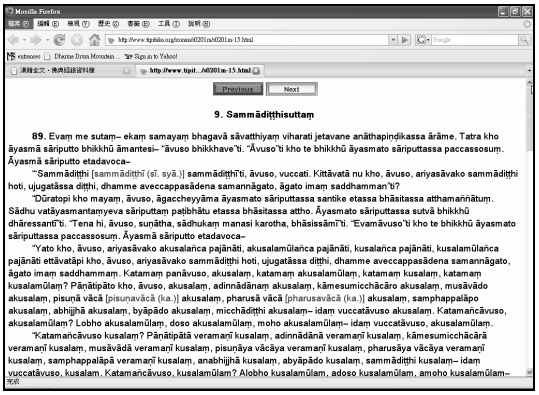
圖六：經號檢索頁，選擇部別，並輸入經號，則可進行「開始查詢」。



圖七：經號檢索結果頁，詳細經目內容資料可點選「更多」。

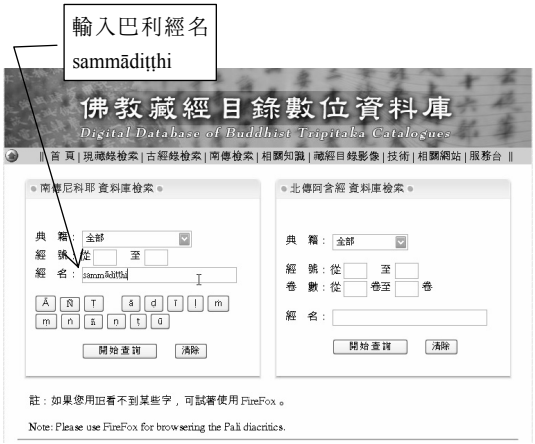
圖八：經目詳細資料頁，除有檢索巴利經的詳細經名、篇名、品名、經號、中譯本冊頁出處及連結全文網頁，有檢索經的南北傳相關經文，並可進行相關經文之詳細資料瀏覽。

圖九：連結 VRI 網頁版經文，線上全文閱讀。



2.2 經名檢索

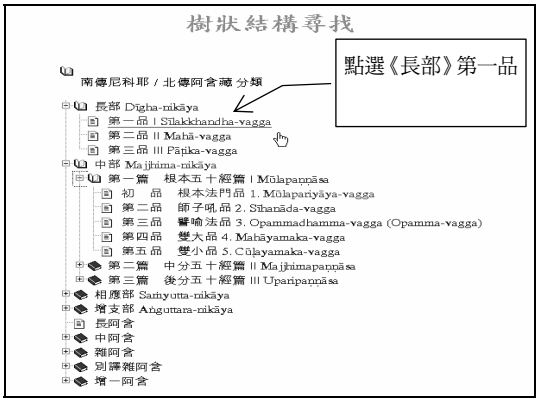
圖十：輸入經名檢索



圖十一：經名檢索結果頁，以下程序同上經號檢索，可點選「更多」繼續瀏覽。

2.3 樹狀架構檢索

圖十二：樹狀架構圖，四部「尼科耶」皆可層層點選開啟尋找品名及其經目。

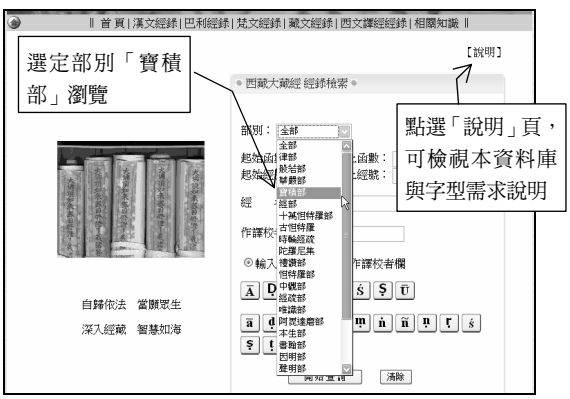


圖十三：《長部》第一品檢索結果頁，以下程序同上經號檢索，每一經皆可點選「更多」繼續瀏覽。

3. 藏文藏經經錄檢索

3.1 部別檢索

圖十四：可選定部別，如下例可進行所有「寶積部」經群的瀏覽。

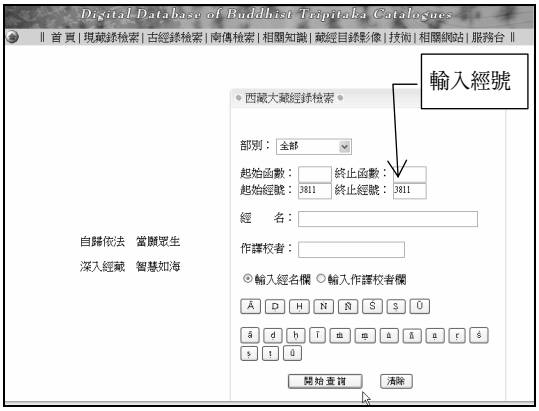


圖十五：寶積部檢索結果頁，從 45 經～74 經共有 30 部經。可點選第 74 經「更多」並查看詳目。

圖十六：寶積部第 74 經經目詳細資料

3.2 經號檢索

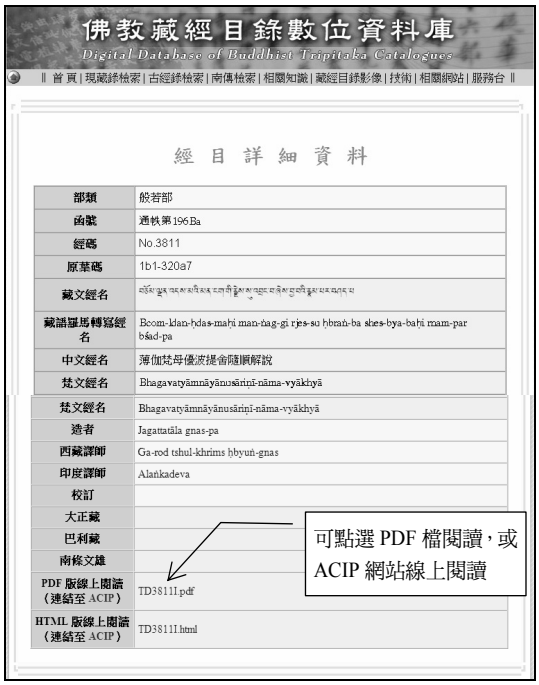
圖十七：輸入經號檢索。



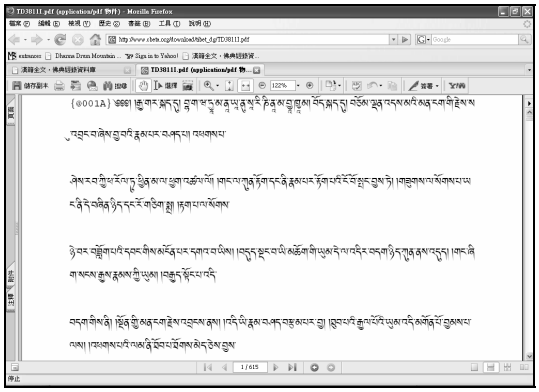
圖十八：經號檢索結果頁，接下來瀏覽程序同上部別檢索瀏覽。

圖十九：第 3811 經經目詳細資料

※PDF 線上閱讀沒有字型需求，若是連結 ACIP 網站線上閱讀，則需加裝 Tibetan Machine Web 字型。安裝方法與連結下載請參考「說明」頁。



圖二十：第 3811 經 PDF 檔全文



3.3 經名檢索

圖二十一：經名檢索，可從鍵盤表輸入羅馬轉寫經名

圖二十二：經名檢索結果頁，可點選「更多」，以下瀏覽程序同上。

五、延伸與開展

1. 以藏經全文及目錄為基礎

如上所述，我們知道數位經錄資料庫已超越傳統紙本資料庫的內容。透過這個完整的Metadata，讓我們向下可以透過經號與原典（CBETA）對應，向上可以藉由GIS時空地理資訊系統的使用，呈現從經文到人物等資料的時間及空間變化。往後如能加強技術與文獻整合、善用時空資訊處理技術來連結本計畫資料庫，便能開展出新的知識架構，完成時空屬性之佛教電子文化地圖。

再者，若能以已完成數位化佛典經錄文獻為基礎，進行佛經目錄詞彙與跨語詞彙抽取應用；以統計分析、資訊檢索及抽取、文件分類與分群、資料探勘等各項語言學應用工作，也將是佛學研究領域中另一個新的開展。

2. 資訊協助人文工作的實例

人文領域內的專家學者在整個人文資訊結合中應扮演核心的角色。資訊技術是輔助，技術與工具要能有效發揮，須賴專家學者的參與。畢竟若不是符合人文研究需求的資源或工具，即使使用的資訊技術再先進、功能再強大，做出來也不過是徒勞無功。

以下將以本校與西蓮淨苑淨土辭典編纂小組合作進行的「淨土薈萃線上辭典」編纂工作為例，介紹資訊技術在CBETA與經錄的基礎上，如何建立索引與應用資料庫，以符合淨土辭典編纂小組對「文本多樣檢索」與「詞條網狀知識架構建立」上的需求。

2.1 索引的建立與應用

對全文資源的索引能力是處理數位文本的基礎，能夠迅速準確的在 CBETA 一億四千多萬字的全文中隨意找到所需的文字資料，是面對各種服務與需求所需累積的重要技術。在本篇介紹的數位工具書範圍中，我們對 CBETA 全文建立了兩種索引：

表三：CBETA 全文的兩種索引

A. Suffix Array

索引步驟	實例節錄
取得純文字內容	有關法顯大師的歷史資料，……
依文章順序記錄每個字的 offset * offset 是檔案裡每個字元的 bit 數	0 有（關法顯大師的歷史資料） 2 關（法顯大師的歷史資料） 4 法（顯大師的歷史資料） 6 顯（大師的歷史資料） 8 大（師的歷史資料） 10 師（的歷史資料） 12 的（歷史資料） 14 歷（史資料） 16 史（資料） 18 資（料） 20 料
依單字的 unicode 字碼順序排列	16 史 8 大 10 師 20 料 0 有 14 歷 4 法 12 的 18 資 2 關 6 顯
取得 offset 順序成為索引資料	16, 8, 10, 20, 0, 14, 4, 12, 18, 2, 6

B. Signature Files

句 1：……《俱舍論》在中、印佛教思想史上，被譽為「聰明論」……

句 2：……正好可以想我剛剛講授的印度佛教美術上相銜接。……

將每個字與文句的對應關係記錄如下，該文句有此單字為 1，無此單字為 0：

表四

	俱中	想印	剛美	上佛相	明	教	思論	聰在	……
句 1	1	1	0	1	1	1	1	1	……
句 2	0	1	1	1	0	1	0	0	……
……									

將每雙字與文句的對應關係記錄如下，該文句有此雙字為 1，無此雙字為 0：

表五

	俱舍 在中	想我 印佛	剛剛 教美	史上 佛教 相銜	明論	教美	思想 被譽	聰明 正好	……
句 1	1	1	0	1	1	0	1	1	……
句 2	0	1	1	1	0	1	0	1	……
……									

上表的第一列可以一字（或一雙字）一欄，也可以多字（或多雙字）一欄，主要的考量在索引的檔案大小及做交集計算的效能。在表三、表四的範例中，顯示的是多字（或多雙字）一欄的狀況。

以上述的索引技術為基礎，我們利用 Suffix Array 做大量的全文比對及自動抽詞工作，再建立詞彙與段落間關係的記錄，進而可以整理出詞彙與詞彙間關連性的數據，利用這樣的數據，可以提供編纂工具書時建立相關詞彙的基礎。更多有關自動抽詞語統計比對的技術說明，可參考國科會計畫「中文詞彙與跨語詞彙抽取技術在數位佛典上的研發與應用」（http://211.23.77.92/BuddhistTermExtract）網站資料。

Signature Files 則大量應用在模糊比對的工作上，由於許多詞彙未必會完全符合的出現在文本中，因此利用模糊比對找出可能相關的文本段落，是索引技術可以協助遍搜全文的重要功能。例如詞條「一心不亂」與《阿彌陀經通贊疏》卷 3：「一日乃至七日一心者，更無間隔故名曰一心；不亂者，專注無散也。」兩者間詞彙與文本的關聯，就必須用雙字詞的模糊比對來找到。

2.2 資料庫應用

在資料庫應用方面，淨土薈萃編纂使用 SQL 關聯式資料庫來處理各項 Metadata 及對照表。另外，對於樹狀及網狀的知識架構的處理，我們也應用 SQL 關聯式資料庫來記錄與應用。相關的資料表有：

a. 藏經目錄資料表（參見表一）

d. 中國古今地名地理資訊對照表

e. 朝代年表

b. 主題分類樹狀結構資料表

c. 詞語網路網狀結構資料表

在樹狀結構與網狀結構的呈現方面，我們使用關聯式資料庫，給予每一個主題詞條（節點）一個唯一代碼，並記錄每一個主題詞條（節點）上一層（parent）的代碼。有了這樹狀結構的基本紀錄後，便可配合瀏覽器呈現工具在網頁上檢索及顯示樹狀或網狀知識結構的模樣。

六、結論

由於傳統上，人文學者對文本的研究多是以紙本出版品為對象，要立即改變習慣面對電子版本的工具，確實會有對電子資源信任度等等的問題。這種狀況需要資訊方面相關工作的持續累積，並依照人文研究習慣的需求逐漸提高品質，藉此希望能讓人文研究的領域漸漸看到資訊工作協助下的方便，以期能吸引更多人的興趣、提高人文研究與計畫參與的意願，使得人文與資訊整合的成績能持續的往前推進。

人文研究為主，資訊為用。數位工具書是最好的人文研究輔助工具。如果人文學者能多了解並測試在人文的研究領域裡重要的資訊工具，對往後人文研究的廣度及人文資訊領域的開展都會有相當的助益。

〔專科工具書編輯研討會　第四場研討〕

▲主持人自行法師（右二），發表人杜正民副校長（左二）、李家名先生（左一）、永本法師（右一）（編輯組提供）