電子佛學資料庫於行動上網時代的機遇

葉健欣　剎那工坊

【提要】2010 年，蘋果公司推出了 iPad，深刻地改變了數位內容的面貌與生態，以及購買和閱讀書籍的方式。iPad 的成功，並不是偶然，而是奠基於更小更快更便宜的晶片、雲端運算的成熟、HTML5 等等一連串的技術積累。從那一年起，我們邁入了「行動上網時代」，以此切入點，本文嘗試描繪未來三到五年，以電子佛學資料庫為基礎的各種資訊服務可能會具有的面貌。

關鍵詞：行動上網；佛學資料結構；計算文獻學；資訊素養

一、行動上網時代

在行動上網時代，上網的人口將十倍於個人電腦（PC）時代，使用行動設備上網（如智慧型手機）的習性，也和從 PC 上網大有不同，如畫面空間變小，使用時間更短。總的來說，行動上網的用戶，有以下兩個特性：

（一）寬廣的受眾頻譜：下至兩歲小孩，上至百齡老嫗，都可能是潛在的用戶。

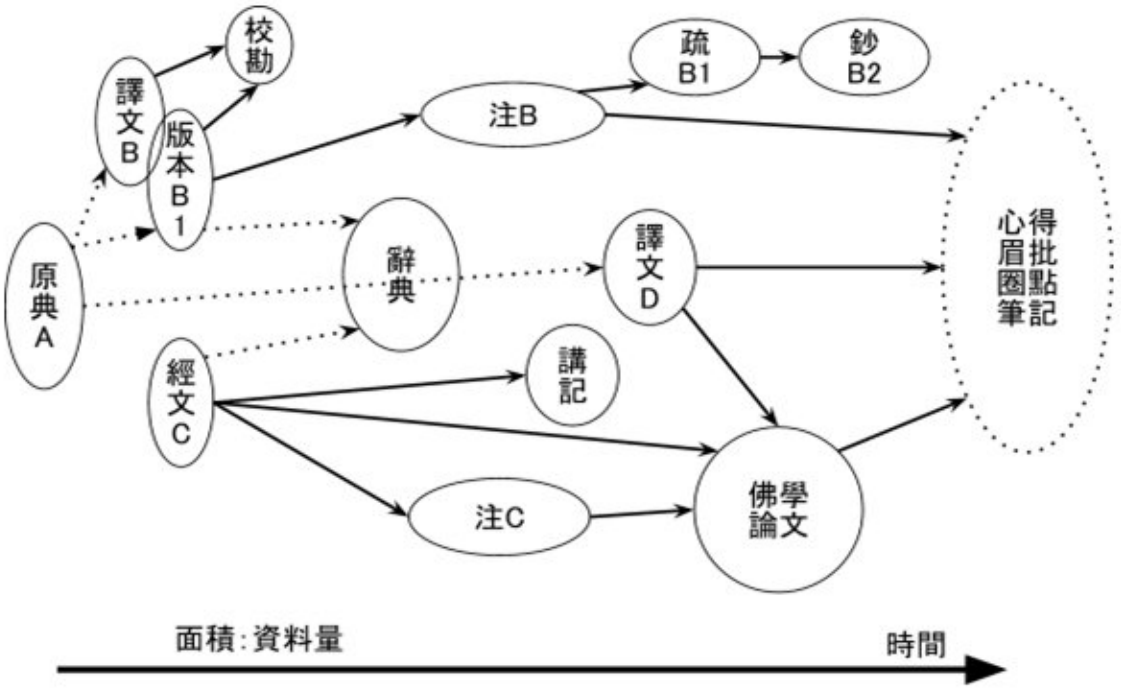
（二）被切碎的注意力：使用者不是端坐在電腦前面，而可能是在會議的中途或等車、等人的空檔，很難擁有使用者完整持續的注意力。

因為第一個特性，未來會勝出的資訊服務，必然是具有親切、直覺的使用介面，才有辦法在短時間內累積出大量的使用者。因為第二個特性，複雜的內容和知識，必須切割成可以在幾分鐘之內消費完畢的片段，然後以簡潔明瞭的方式呈現，最好還帶有一些趣味性。

傳統以線性方式為主，輔以章節目錄結構的內容編排方式，將不適用於行動上網時代。換句話說，如果只是將在電視臺上的一個小時以上的講經節目轉成手機視訊，或是直接將整本大部頭的書轉成 PDF，可能收不到預期的效果。

那什麼才是理想的內容格式呢？在回答這個問題之前，我們先來探討一下，佛學資料的幾種類型。

二、佛學資料的類型



圖一：佛學資料 DAG 結構圖

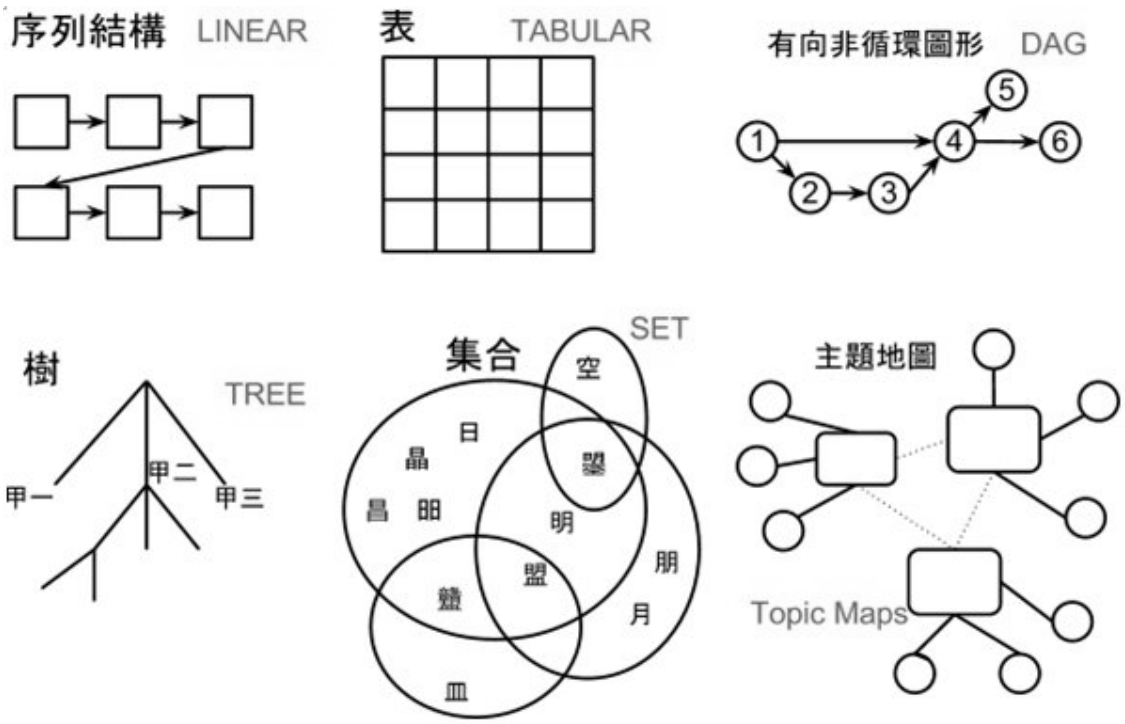
圖一以 DAG（非循環有向圖形，Directed Acyclic Graph）的結構，呈現各種佛學資料。

以往在教界，我們談的數位化，多半是既有資產的數位化，比方說三藏經文、某部佛學辭典、某位法師的演講記錄，但我們很少關心廣大使用者所參與創造的數位內容。事實上，近十年來最重要的網站，都是遵循「將使用者創造的資訊整合起來，變成有用的服務」的模式而崛起。比方說 Google 將全球的網頁，製成全文檢索資料庫；Facebook 靠使用者打造出的人際關係網絡；Wikipedia 整合全球的游兵散勇，完成了人類有史以來規模最龐大的百科全書。

過去二十年，我們在經文的數位化方面取得非凡的成就；未來發展的重心，將會是「重現創作者情境」和「集結使用者加值」兩大主軸，例如：漢文到巴利文的溯源、各譯本之間的對照、歷代注疏和當代法師對經文的引用。更進一步，使用者群的心得、筆記，彙集成新的注疏，為後人留下我們這一代對經文的解讀和實踐記錄。

接下來，我們從電腦的觀點，來看看佛學資訊在電腦內部的儲存方式，即所謂的「資料結構」。

三、電子佛典的資料結構



圖二：電子資料結構圖

圖二列出六種資料的結構方式。

（一）序列結構：比方說大家很熟悉的純文字檔，是電子檔案最初期的階段，一般以這樣的形式來建構。

（二）樹狀結構：最常見的例子就是 HTML 和 XML，大約可以對應到紙本的目錄和科判。

（三）表格式：應用非常廣泛，適合結構（欄位）相對固定的數據，試算表、關聯性資料庫都屬於第三種。

（四）集合：很適合用來描述字與部件之間的關係；另外，論文開頭的「關鍵字群」，也是一種集合——交集、聯集的運算，提供了非常靈活的分類和檢索途徑。

（五）有向非循環圖形，是分散式版本管理系統的核心資料結構，每個節點可以開枝散葉，數個節點也可以合併成一個新的版本，這個結構準確地表達了文件的版本演化進程。分散式版本管理系統目前最成功的應用，是在開放軟體原始碼的管理。這個技術突破了傳統必須同時同地的編輯模式，而以極低的成本，匯聚全球腦力的涓涓細流，完成不可思議的龐大工程（如

Linux）。可惜的是，目前它的使用介面，大概只有每天必須和程式碼打交道的人才會覺得方便；但隨著時間推移，應該會漸漸變得親切易用，它的發展，值得文史工作者深入關注。

（六）主題地圖和本體論（Ontology），是近年來資訊探勘領域相當熱門的話題。在主題地圖中，知識以連結的形式存在，更接近我們大腦中的運作方式；結合類神經網路（如Hopfield Network）、潛在語意分析（Latent Semantic Analysis）等的文本探勘技術，以應付自然語言中一形多義（Polysemy）和多形一義（Synonymy）現象。

接下來，我用「異體字及同形字」、「視覺化多版本比對」以及「模糊搜尋」三個實例，來說明上述資料結構的運用情形。

四、異體字及同形字

漢字處理，是漢文電子佛典的核心難題之一。過去二十年來，擴充字集的辦法，表面上舒緩了罕用「字形」的顯示問題，實際上卻造成輸入（檢字）和搜尋的困難（如戶户戶）。字與字形，是多對多的關係。舉例來說，「倉」是「答」的異體字、也是「會」的異體字。「叶」，是古文「協」字，也是「葉」的簡化字。像這樣複雜的關係，很難用一字一碼的方式描述，必須用集合和圖（Graph）的結構，才可以精準表達。

異體字可能是一個古文字的不同隸定，或是轉注字、錯別字。而同形字，可能是不同古文的相同隸定，例如「  （zhong）」「（zhong4）」都隸定為「中」字；古文「上」，二畫上短下長，看起來和數字「二」一樣（數字「二」原為上下均齊的兩橫，字作「𠄎」）。

字義、字形、字音之間的關係，遠比想像中的複雜。Unicode 組織當年以為漢字只要 2 萬字（0x4E00~0x9FFF）就夠了，於是擴充 A 擺在 2 萬字的前面（0x3400~0x4DFF），造成排序的混亂；後來又陸續增加的擴充 B（0x20000）、擴充 C 和擴充 D，不久前又加了 IVS（Ideographic Variation Sequence）以表達微小的字形差異，於是一個 Extension B 的字加上 IVS 在 UTF16 的編碼下，竟要 8 bytes。

如此不斷地疊床架屋，不要說文史人員，大部分的資訊專業人員都弄不明白。這說明了 Unicode 最初設計缺乏周全考慮，暴露了對漢字內在結構認識的不足。一個由文字學家主導，輔以技術專家的團隊，才有機會克服這個問題。

五、視覺化多版本比對

多版本比對，是在分散式程式碼版本管理工具（Distributed Source Code Version Control System）中，必備的功能之一。在程式碼中，一個英文字母之差，也可能帶來完全不同的執行結果。當很多人在不同時間、不同地點編寫同一份程式時，如何直覺地檢查別人的更動記錄，就變得很重要。

古籍在漫漫歷史長河，被不斷地傳抄、翻印，文句必然會發生失真變異的現象；此外，佛經還有不同譯本的情況，在資料變異的性質上，和多人編寫程式碼非常接近。因此，我們利用程式碼版本比對的技術，很容易將同一經文的各個版本之間的差異，以視覺方式呈現出來。

圖三為玄奘所譯的《心經》和其他四種譯本一一比較後的視覺化呈現（共 5P2 = 20 種）。綠色（■）表示各版獨有的字，紅色（■）表示各版所無的字，黑色（■）表示兩種版本相同。

有了自動化的版本比對機制，現代統計學方法就派得上用場了。舉例來說，我們用這個工具分析歷年法律條規修訂的情況，發現某一個時期，「山胞」大量被替換成「原住民」。相信同樣的方法，也可以應用在解讀文獻的詞句，從中解讀出肉眼難以發現的規律。

六、模糊搜尋

我們以多年全文檢索的經驗為基礎，針對古籍的需求，發展了一個模糊比對的技術，它能以每秒約二十萬次的速度，計算文言文短句與短句之間的相似度。目前我們發現兩項應用：（一）近似句查詢（查詢重複），（二）節引出處溯源。舉例如下：

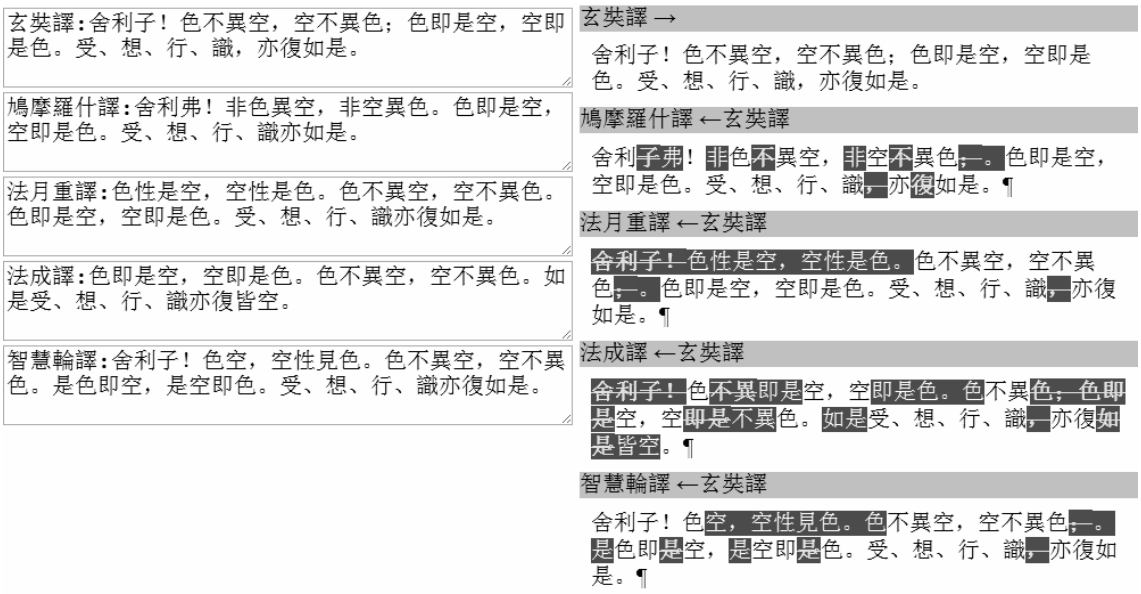
（一）近似句查詢

在標點過的康熙字典全文中（註 1），挑出引用的句型，兩兩相比（1,000 句的樣本，要做五十萬次的運算，大約需時 2.5 秒），找出一大批相似的句子，例如：

儩=《唐書·王叔文傳》：儩然以為天下無人。（《康熙字典》） ← 《舊唐書》用字。

憪=《唐書·王叔文傳》：憪然以為天下無人。（《康熙字典》） ← 《新唐書》用字。

開頭的一字之差，經過文史人員核對原文，查出前一句應節引自《舊唐書》，後一句則節引自《新唐書》。



圖三：《心經》五種譯本

（二）節引出處溯源

《康熙字典．利》：「《莊子．駢拇篇》：小人以身殉利。」電腦檢得出處如下：《莊子．駢拇篇》：小人則以身殉利，士則以身殉名，大夫則以身殉家，聖人則以身殉天下。

《康熙字典．母》：「《詩．小雅》：豈弟君子，民之父母。」電腦查出「豈弟君子，民之父母」事實上出自《詩．大雅．泂酌》。而《詩經．小雅．南山有臺》的類似句是：「樂只君子，民之父母」。

藉由自動出處查詢的機制，讀者很容易從節引文字跳到原書出處，透過前後文的關聯，更好地把握原作者及引文者的意圖，避免斷章取義。

葉聖陶在《十三經索引》序言提到：「目光馳驟於紙面，如牧人之偵亡畜，久乃得之，甚矣其憊。……每有所遇，似曾相識，而隸屬何篇，上下何文，往往弗省。」葉公為了節省士人查核原書的時間精力，把母親、妻子、姑母都拉下水，投入於《十三經索引》的編製，以致「寒夜一燈，指僵若失，夏炎罷扇，汗濕衣衫，顧皆為之弗倦。友人戲謂此家庭手工業也」。

讀罷這段文字，一方面對前人的努力，致以崇高的敬意，同時也感到深重的使命感；今天有了電腦技術，很多苦差得以避免。換句話說，資訊人員的任務，是讓文史工作者掌握適當的資訊工具，讓他們從低價值的重複勞務中解放出來，將時間精力留給高價值的判讀、解說和創造活動。另一方面，文史工作者也要有清楚的認識，哪些工作可以交給電腦代勞，否則忙了幾個月，才發現相同的工作用電腦來做，只需費一點電力而已。

接下來我們談談文史工作者面對排山倒海而來的資訊科技，應具備哪些素養，以做出正確的選擇。

七、文史工作者的資訊素養

五四運動以來，重科技、輕人文，成為普遍的現象，流毒至今，在文史工作者和資訊科技人員必須協作的場合，常會聽到這樣的對話：「拜託一下修改這裡！」「不行，這功能技術上辦不到。」文史工作者的需求得不到滿足，只好無奈地回到人工的作業方式。

敦厚的文史工作者也許不知道，這句話還會指涉兩種語意，一是超過能力所及做不出來，二是覺得麻煩不想做。有鑑於此，剎那工坊規劃了一系列的課程，協助青年文史工作者升級為「機械化文史尖兵」。簡言之，這是一個能夠用精確的資訊行話來描述文史需求、並能跳過代理人，直接指揮電腦工作的新職種。成為尖兵的步驟是：

（一）心理建設：不要覺得程式技術是資訊人員的事，也不要覺得電腦很難。電腦只是飛快但完全不懂揣摩上意的笨奴隸。我們要做的只是弄懂它的習性，用它聽得懂的語言，指揮它幹活。文史工作者是主人，主人只要會下命令即可，沒有必要成為駭客。

（二）計算機概論：花幾個週末的時間，翻一翻《圖解電腦作用與原理》之類的書。讀完杜林機（Turing Machine）一節，明白什麼是「可計算性」和「時間空間複雜度」之後，一般的資訊人員就很難用「這做不出來」的理由來欺矇你了。

（三）少用 Word，改用純文字編輯器，掌握巨集、Regular expression、XSLT、SQL 等批次文本轉換和資訊擷取工具，可大幅提高工作效率。如果實在沒空學那麼多，掌握 Text Pipe、Google Refine 等應用，很多時候也就夠了。經由批次作業，慢慢體會到什麼工作可以交給電腦代勞，因此在建立資料時，會優先從電腦方便處理的角度來思考，而不再拘泥於人類習慣的方式。

（四）完成以上三項，你將能從瑣碎的資料編輯和整理工作中釋放出很多時間，而且在嚐到甜頭之後，學習進階程式技術的熱情會空前高漲。此時，還在潛龍勿用的階段，先不要讓老闆知道你已今非昔比，日常工作還是按以往的節奏完成。空下來的精力，投入至少半年到一年的時間，熟悉 Javascript 和 HTML5，它能替你的心血結晶插上翅膀，翱翔於個人電腦、智慧型手機、電子書以及未來的行動上網載具上。

八、結語

電腦程式語言經過幾十年的發展，目前看來，Javascript 成了強勢物種。在其他程式語言中，累積多年的成熟觀念和設計架構，正大規模、高速地往 Javascript「移民」。換言之，它很可能會像英文一樣，成為調遣雲端資源的通用語。

曾榮汾在〈漢籍工具書編輯經驗談〉提到：「如果說漢籍整理是個資管工程，則一件資管工程的從開動到完成需要的幾個條件，正是：眼光、人才、技術、恆心。眼光短淺則無視輕重緩急；人才缺乏則徒勞無功；技術未精則事倍功半；恆心難繼則事難究竟。」（註 2）

古往今來，能同時具備四個條件的個人或團體，應該是寥若晨星吧！從歷史的角度來看，資訊人員就是工人，和古時的製紙、雕版工人並沒有本質上的不同，只有飽讀聖賢詩書的士人，才知道大義之所在。長期刻字的工人，也許某一天也能吟幾首詩，但是真正能令文化界產生質變的，是文化人願意親身來理解科技、運用科技來解決問題。相信四海之內，一定還有不少默默的文史鬥士，汗流浹背地拿著圓鍬，奮力替我們民族的未來開路。剎那工坊（註 3）以開發新型「怪手」為職志，歡迎有志青年，來學習如何駕馭它。如果您是長者，我們恭迎入座、聽候差遣。

【附註】

註 1：http://kangxi.adcs.org.tw/。

註 2：曾榮汾，〈漢籍工具書編輯經驗談〉，2004 年古籍學術研討會，輔仁大學文學院．圖書資訊學系暨中國古籍整理學程、輔仁大學圖書館主辦，民國 93 年 6 月 11 日，http://diction.sg1002.myweb.hinet.net/dict/01lunwen/25doc.pdf。

註 3：剎那工坊 E-mail: yapcheahshen@gmail.com。