Accelon，一個開放的數位古籍平台

萊健欣　剎那搜尋工坊軟體工程師

【摘要】本文以電子佛典發展的三個階段為經，以佛典數位化的種種困難和解決方案為緯，總結筆者這十幾年來從事電子古籍相關工作的心得和感想。探討的範疇包括了缺字、文字編碼、文件格式，以及數位內容所有權等。最後，本文提出了一個開放源碼的數位古籍平臺架構，以降低古籍數位化的製作、維護和發行，以及使用的門檻。

關鍵詞：電子佛典；缺字；全文檢索；數位內容之所有權；開放源碼；維基模式

第一階段：1989-1998

回想筆者和電子佛典的初次結緣，約莫在 1992 年，用 286 1MB RAM 的電腦，加上倚天中文系統和 PE2 純文字編輯軟體，開始輸入一些常用的佛經。動機很單純：一套要價新臺幣十萬元的《大正藏》，如果能夠輸入電腦作廣泛應用，這對佛法的弘揚必定很有助益。

當時抱著這種想法的教界朋友，想必不在少數。從 90 年代初期開始，到 1998 年中華電子佛典協會（CBETA）成立為止，這是佛經數位化的第一個階段，工作內容以佛經的輸入為主，我想有不少朋友也是在這個階段開始投入的。

這個時期，臺灣網際網路還剛剛萌芽，大家各做各的，力量無法整合起來，很多經典被重複輸入，尤其是像《金剛經》、《阿彌陀經》等常用經典，也許在全臺灣被輸入了上千次。從「抄經功德」的角度來說，這並不是什麼壞事；不過就資料庫的觀點來看，這個重複沒有必要，也造成了各經的品質參差不齊。現在回顧起來，當時教界整體投入了很大的力量，但是實際綜效（synergy）並不理想。

數位檔案和紙本檔案有本質上的差別，任何人都可以抄紙本的佛經，字寫得好不好不重要，每個抄本都有其獨一無二的價值。你恭錄了一份《心經》，並不會因為另一個人再抄一次，就失去了價值；而電子佛經不同，數位世界只需要一份錯字最少、校勘最精、後設標誌最全的佛經。

至於各種樣式，無論是做成網頁、電子書，甚至印刷出來。從技術的觀點來看，都可以從這一份主資料（Master Data），由電腦自動轉換而得。換句話說，數位佛經的成熟化，就是從多個互不相容的分散檔案，化成單一權威的版本，能夠根據不同的需要，由電腦即時、自動地轉換成讀者想要閱讀的形式。

第二階段：1998-2006

當時的時空背景，大家體認到需要一個組織，統一進行整部大藏經內文的輸入和校對工作，CBETA 就在這樣的背景之下成立了（註 1）。從 CBETA 醞釀到成立至今，正好十年，這是佛經數位化的第二階段。

這期間，也是電腦技術發展非常迅速的階段，晶片的密度每 18 個月倍增一次（註 2）、儲存媒介的容量、網路速度，整體上朝更大、更快、更便宜的方向發展。

總結來說，這個階段可以歸納出兩條規律：其一，運算愈來愈分散，因為電腦愈做愈小、越便宜，因此運算能力從笨重的桌上型電腦，擴散到筆記型、PDA、手機、數位相機，以及愈來愈多的智慧型設備。換言之，就是每人平均擁有的 CPU 數持續上升。

其二，資料愈來愈集中。這裡所謂的集中，不是指物理上的集結，而是指零碎資料的合併。電腦資料有個性質，整合度越高，價值越大。比方說，個人或單臺收銀機的付款記錄，分散時價值不大，但是如果將整個城市乃至國家，每個人的消費記錄整合起來，這個資料庫就非常有用了，譬如從中分析出消費的習性，可以作為預測消費行為的具體數據。

這兩條規律是數位科技發展的兩條主線，掌握它們，就能掌握數位科技發展的趨勢。

對於第一條，一般人能參與的不多，電腦運算技術的主導權在歐美日等大國。新奇玩意兒可以一直生產出來，但是要不要讓它成為生活的一部分，每個人有充分的選擇權。事實上，在生活當中，新科技很少會完全取代舊方法。舉例來說，便利超商固然可以買到種類繁多的茶飲，但是泡工夫茶的，還是大有人在。人們可以用手機下載慈濟靜思語，也可以洗手焚香，著海青，端身正坐恭誦《大般若經》，兩者並不互斥。甚至在某些時候，正是因為新科技太方便、太普及，才凸顯出傳統方法的專業、稀有和隨之而來的尊榮感。

整合：數位資料庫的威力所在

至於第二條，零散資料的集中化，表面上沒有前者來得明顯可見，但是影響力有過之而無不及。舉個比方，在數位佛典的發展初期，人們會以錯字太多、檔案不全，甚至螢幕傷眼等種種理由而拒用電子版。但是時至今日，只要稍為和佛學甚至是漢學扯上，哪怕只有一點關係的學者，都無法忽視 CBETA 資料庫對研究工作、數量上和質量上的重大影響。事實上，數位世界的種種神通，如全文檢索、資料分析、服務品質及速度，無一不築基於整合式的資料庫。

當然，資料庫的影響力，不限於佛學界。任何學界，當基層資料庫成熟之時，必定對該領域的研究方法產生衝擊，甚至是某種程度的典範轉移。因此，我們瞭解到，促使資料庫的形成，對一個學界來說，是非常重要的基礎建設。今天，道教學界、儒學界乃至國學界，都對佛學界擁有

一個高品質的基本典籍庫，而艷羨不已。他們很清楚地看到，資料庫能對研究工作，以及緊接著會發生的影響，起著非常關鍵的作用。

佛教界先於其他人文領域的原因

資料庫的形成，不是一件輕鬆的事，佛教界會領先其他的領域（有強烈商業動機者除外），有其歷史原因。首先，宗教界向來就勇於嘗試新科技。世界上第一部印刷品是漢文《金剛經》（註3），西方第一部活版印刷品是古騰堡（Gutenberg）的《聖經》（註4）。再者，佛教界很清楚，太平盛世、皇帝英明，編大藏經來彰顯國威是可遇不可求的。因此，由民間發起的倡印佛經的工作，自然成為一個傳統，並且被視為積聚福德的方式。

同樣地，個人電腦開始普及，等不及政府或研究機構訂計畫、撥經費，三寶弟子們很自然、不約而同地前仆後繼，投入數位佛典的工作。由於數位儲存是個新科技產物，大家一起來摸索，比由政府來「主導」或「輔導」，更能夠快速累積經驗。因此，由民間自發性投入電子佛典的製作，反而是最成功的。

資料無法整合之原因

由於佛教界的示範，大家都見識到資料庫的威力；那麼，為什麼到今天為止，人文學界的免費資料庫的發展，除了中研院的「漢籍電子文獻」（註5）之外，其他免費的資料庫都還未具氣象呢？舉個比方，網路上《孫子兵法》和《紅樓夢》有數十種品質不一的版本，很難從中選出一個最好的版本。因為缺乏一個具公信力的基本資料庫，大家只好花很多力氣來維護自己要用的版本。長遠來說，這是一個很大的浪費，是研究升級的障礙。

在這繁鬧現象的背後，有著相當複雜的成因。筆者觀察多年，至今仍然沒有把握窺得全貌。以下提出幾點心得，請方家不吝指正。

使用不當的工具

佛教界資料數位化進行得早，因此圖形化操作介面還沒有普及之前，就輸入了大量的經文，當時並沒有像今天這樣方便的 Word/PDF/HTML 之類的工具可用，只能單純的輸入經文，很難做視覺格式的變化，但是這反而是一件幸運的事。怎麼說呢？事實上，一旦使用了這些視覺化工具，就掉入了「著相」的歧途。Word/PDF/HTML 是一種末端的「呈現格式」，它讓人很方便地排出漂亮的報告、寫日記，用來編一本書也勉強可以。但是，正因為視覺呈現格式的設定太過方便，反而容易被濫用。

舉個例子來說，假設使用 Word 來輸入佛經，除非有嚴格的規定和監督（這意味著高成本的管理工作），大家傾向於使用各種字體樣式來「美化」經文。有人用「黑體 24pt」來表示經名，有人則喜歡用楷體，最終的結果是各式各樣、花花綠綠的檔案。這對人來說無所謂，但對電腦來說，合併就是一個大麻煩；而要從其中提取出有意義的後設資料（Metadata），更是費時費力。

而在純文字時代，因為不能換字體、換顏色，只好用像「經名：XXXX」的方法來標示，電腦處理起來反而容易許多。因此，數位資料庫最難的第一步：資料合併，佛教界由於開始得早，使用原始的純文字工具，反而進行得比較順利。

其他人文領域數位化開動之時，已是 Word/PDF/HTML 當道的時代，大家理所當然地使用，然後直接將檔案放上網，形成今天這樣的局面。

因為它們實在太誘人，也太方便了，人文工作者很難警覺到它的危險，而主動去尋找其他方案。（註 6）

一位資料庫的設計者，必須有足夠的智慧，穿透問題的表相。而在製作資料庫的初期，選擇適當的格式和工具，攸關計畫的成敗。因此，技術的選用必須非常慎重，一定要吸取過往的經驗。

缺字問題

缺字是個「歷史悠久」的老問題，希望讀者要意識到，缺字是古籍整合最底層的障礙，而這個問題，無法期待電腦科技的發展而自動獲得解決；換句話說，無論 CPU 再快，記憶體、硬碟再大，網路再普及，缺字不會自動被解決。

由於缺字，大家只好都當起了現代倉頡，各自造了一堆字，直到造字區爆滿為止。這個造字檔，表面上解決了寶貝資料的顯示和輸入問題，但是事實上，這正是資料無法彼此互通的元兇。時至今日，少有資料是不打算放上網的，可是隨之而來的問題，就是沒有人願意安裝所提供的造字檔，因為裝了之後，原來的資料就亂掉了（原來的造字檔，被新的造字檔取代）。

對使用者來說，問題還沒有結束，如果有人要從事佛學和醫學的跨領域研究，他很快會發現，佛經和網上下載的醫學典籍無法放在同一個檔案中，無論是用佛經或是醫典的造字檔，總有一些字會牛頭不對馬嘴，於是只好重頭再校對一次，再想辦法補上缺字。

更糟的是，個別的努力，只是沒有必要的重複作業，這些勞動成果非但無法累積起來，反而增加他人的麻煩。因此，這個額外作業將成為永遠的負擔，甚至可以不誇張地說，這個進入數位時代才產生的資料不相容的現象（在紙本時代，可以任意拿任何兩種資料放在一起，反正都只是抄寫），墊高了人文學科之間跨領域溝通的門檻，降低各領域間的交流意願。

有人也許會說，Unicode 不是已經解決了大部分的缺字問題嗎？事實上並沒有這麼單純，Unicode 還是將每個中文字視為個別符號來處理。由於中文字和英文字母不同，是開放集合，在 Unicode 架構下，缺字問題永遠存在。最明顯的例子，就是 Unicode 無法處理新字、錯字和新出土的異體字。而英文卻沒有這個問題，可以用 abc 隨意組合出新字（如 Blog、Wikipedia 等）。這個用更精簡的形式，來表達新觀念的能力，是語言的生機所在。很遺憾地，漢字固有的、精簡表達新事物的能力，被不當的電腦系統架構所扼殺了。

關於文件格式和缺字問題，本文後面有詳細的說明。

缺乏整體規畫

整合的本質，在於增加聯繫、降低溝通的成本。但是，在連線發生的初期，經常是弊大於利，只有在整合的中後期，效益才會凸顯。打個比方，電話和電視剛發明出來時，大家裝設的意願很低，因為成本高而效益不明顯，此時這些都是有錢人和前衛人士的時髦玩意兒，但是等到大部分的人都裝了，此時就非裝不可，否則就會被邊緣化。其他的通訊科技，如 email、skype，也有這個特質。

資料庫也是如此，資料本來就在，整合就是降低存取的成本。整合的初期必須做出很大的投入，但是並沒有明顯的效益。此時，需要有個強而有力的組織，持續進行高密度的投入，統一技術規格，才會大幅縮短累積的時間，提早進入下

一個階段。

比較棘手的是，如果這樣的組織一直沒有出現，而因為實際的需求，大家各做各的，熱熱鬧鬧地鋪路拉線，等到要整合的時候，才發現由於沒有統一的規畫，彼此的成果無法合併。整合者將面臨兩難的局面，要麼推倒重來，犧牲某些不相容的成果，但是這樣做對被放棄的一方是非常傷感情的事；要麼就要開發大量的「轉換器」來連結不同的子系統。如果選擇後者的話，表面上不傷和氣，但整合的效益大打折扣，就像一個國家如果有十幾種不同寬度的鐵軌，無論如何轉換，都不會順暢運行。

但是這樣的局面，在數位世界卻是常態，大家使用不同工具，產生不同的檔案格式、不同的造字檔、不同的字體設定，每一種不同，都是交流的障礙；每一次的轉換，都增加了溝通的成本。由於數位檔案的複製和傳布幾乎不必成本，因此最大的負擔，就在於處理和轉換這些不同的格式。

既然如此，是不是要規定大家都用同一套工具集，同一個造字檔，並嚴禁設定字體顏色樣式呢？當然不能這麼霸道，何況這是完全不切實際的規定，沒有實施的可能。

人文與數位科技的橋樑

數位的危險之處，在於它是一個急速進展的科技，不但是身在其中的科技界人士，對層出不窮的新技術感到無所適從，一般的文史工作者，對琳瑯滿目的各種技術，也不知如何選擇——作業系統有 Windows/Linux/Mac 三大門派，輸入法有好幾十種，程式語言常見的也有十幾種，資料庫軟體、編輯軟體莫不如此。因為新事物不斷湧現，即使是專家，也無法確定哪一種是最佳的組合選擇。

於是，人文工作者就面臨了一個困難的局面：一是要花很多工夫去嘗試各種所謂的「解決方案」，時時追趕技術新知；二是以不變應萬變，堅持自己習慣的做法，對環境的改變置若罔聞。選擇前者，人文工作者會逐漸變成了科技專家，離自己的本業漸行漸遠，想要役物反而為物所役，這就有違使用數位科技的初衷；後者則是閉門造車，最後做出來的車子能不能上路，還是未知之數。

在這種情況下，我們亟需在人文學界和數位科技之間，搭起更多的橋樑。基於這個理由，1998 年，筆者離開第一線的經文輸入工作之後，展開了為期八年尋師訪道和磨練技術的旅程。

沒有事前的規畫，也不知道應該學些什麼，只是抱持著一個簡單的信念：想解決佛經最關鍵的幾個技術問題。這個題目，決定我必須同時理解古籍和數位科技的特質。這段個人經歷，和本文接著要提出的架構密切相關。

從 1998 年開始，我有幸為《印順法師佛學著作集》光碟製作搜尋軟體。當時我意識到，不能直接把既有造字檔放進去而要求使用者安裝，出版社印書可以要求每個編輯和排版這樣做，但是發行光碟行不通。於是，我將《漢語大字典》五萬四千個字形，做成不占用造字區的 TTF 向量字形檔。這個字庫，後來被中研院採用，成為「漢字構形資料庫」的一部分；直到數年後，Unicode 3.0 七萬字 TTF 漸漸普及之後，才慢慢退休。

另外，值得一提的是，印順導師光碟的全文檢索功能，除了七百萬字的導師著作之外，還收錄了三十二冊的《大正藏》，有數千萬的字數。在當時，要在光碟上直接檢索這個數量的資料，是一項技術

上的挑戰。中研院謝清俊教授在 1985 年就開始做古籍資料庫，90 年代初就有實用成果。我因為謝教授的因緣，得以結識「漢籍電子文獻」全文檢索功能的設計者林晰先生，他很慷慨地和我分享全文檢索技術的一些訣竅，雖然我沒有機會看到該系統的原始程式，但是由於他的啟發，我有了自己動手打造一個全文檢索系統的信心。

Accelon1

2000 年，網路泡沫化，我回到老家馬來西亞，寫了大半年程式，成果即是 Accelon1，導師光碟的第二版及《中華佛教百科全書》光碟版使用了這個系統。Accelon1 是我第一個具有真正全文檢索引擎的作品，這個引擎的成熟度和規模當然無法和 Google 相提並論，但是就技術而言，它們的本質是一樣的。所謂的檢索引擎，就是搜尋的速度不會因資料的增加而等比例變慢；換句話說，當資料增加 10 倍，系統並不會變慢 10 倍。全文檢索的效能，取決於索引結構的設計，理論上和資料量關係不大。

之後，我開始將注意力轉向缺字問題。越深入研究，越發現背後的學問不簡單。當時，謝教授的方法，確實從理論的高度上，解決缺字的關鍵部分，即字形的「制式表達」，但是在實用化部分，還停留在實驗室階段，對一般的使用者來說，造字還是無法避免；換句話說，這個研究成果應得的利益並沒有完全在民間實現。

經過反覆的思索，我發現中研院的方案缺乏一個關鍵的模組：依據字形表達式，轉換成人們習慣的方塊形式。簡單來說，這個模組的任務，就是將一個線性的式子，如「方方土」，轉換為真正的字形「堃」。整個缺字的關鍵問題點，就是因為目前所有電腦的中文字形，都是根據定好的內碼，而預先製作出來的。因此，只要某個中文字沒有被編成內碼，任何字形檔都不會有。從這個角度來看，造字法其實不在「造字」，因為字本來就在，我們只是隨意指派一個內碼給它而已，這是缺字無法交換的癥結所在。

因為這個觀念非常重要，我這裡再舉一個例子，想必大家對「閔」、「鉢」、「睺」這幾個字，必定印象深刻吧？佛教團體和出版社，大概找不到沒有造過這幾個字的吧。正因為內碼隨意指定，成果不能分享，所以每個單位都要個別重造一次。造字並不便宜，但是大家已經習慣把它當作成本的一部分，除了承接造字服務的公司之外，沒有人對這個情況感到滿意。

動態字形產生器

所以解決缺字的訣竅，其實非常簡單：使用「門众」、「釒本」、「目侯」這樣的式子來代替任意造字。這些式子由系統已知的漢字「部件」所構成，約莫一千個「部件」，就可以組合出無限個字形來；這和 26 個字母就可以組成無限個英文單字，道理上是相通的。

由於漢字不是線性文字，因此我們需要一個模組，根據式子「動態」地畫出二維的字形。這就是這個模組命名為「動態字形產生器」的理由。換言之，我們看到「方方土」這個式子，為什麼會知道是哪個字形？因為懂漢字的人，腦海中都有這樣的產生器。這個模組的功能，就是讓電腦接手這個任務，替我們做字形組合動作。（註 7）

倉頡系統、文化傳信

事實上，動態字形並不是新觀念，早在 70 年

代，就有幾位學者專家提出這樣的想法。其中以朱邦復先生最為知名，他是少數有實作經驗並公開這個技術的先行者(註 8)。當時朱先生在澳門，經過幾次網上交談，我很快下定決心向他學藝。從 2000 年到 2001 年，前後約一年時間，我一方面擔任澳門文傳科技的軟體工程師（朱先生為該集團副總裁），從事將朱先生的系統移植到 Linux 圖形核心的工作，工作之餘，主要的精力放在理解朱先生的系統。

過了幾個月，我大致掌握了這個系統的設計理念，頗為其精巧高速而折服。可是，我也發現了兩個架構上的缺陷：其一，該系統完全以 8086 組合語言（註 9）撰寫，幾乎無法跨平臺。為求精簡，使用大量的 8086 特殊技巧，資料和演算法的偶合（coupling）程度太高，除了口傳心授的嫡系弟子，別人很難一窺堂奧。其二，這個系統是根據倉頡輸入法的編碼來產生字形。倉頡輸入法本質上是一個遷就鍵盤，將漢字進行主觀拆解，然後設法分布到 26 個按鍵上。字形拆分主要的考慮是降低重碼率，而不是依據字源。舉例來說，倉頡輸入法中並沒有設計「門」這個部件（註 10），而是把「門」拆解為「日弓」，這顯然是違反文字學語源。

因為這個設計，倉頡字形產生系統中必須處理大量的例外。比方說「日弓人」，可以同時指「昅」、「閃」，這樣字形產生器就必須設定例外規則予以化解，無形中增加了很多複雜度。因為與文字學割裂，也註定這個系統無法表達新字、異體字、錯字，以及如「招財進寶」這樣的合文。

當發現了這個事實之後，我面臨兩個選擇，一是留在文傳，等到資歷夠深，也許會被允許改動倉頡輸入法字根的設計；要不然只能掛冠求去，因為無法說服自己，繼續投入在一個並不認同的架構。正巧此時，認識了原本想和文傳合作中文 CPU 的臺灣易符科技（註 11），我向易符的股東說明了有必要從頭打造一個更理想的中文環境，這樣的理念得到易符的認同。於是，我在 2001 年底離開文傳，加入易符科技，從事動態組字的研發工作。我們約定，易符資助我的研究，成果將應用於易符開發的電腦晶片，以取得商業上的回報；而我則可以將開發出來的成果，以軟體形式，自由散布給所有需要的人。

易符科技

在易符的期間，是非常值得紀念的日子，在這三年半的時光裡，一方面做中文字形產生器的工作，另一方面也學習一個稱為 Forth 的精巧語言，參與嵌入式系統的設計和製作。學會 Forth 對提升個人編寫程式的能力，以及對整個電腦系統的深入理解，起著很大的作用。

2003 年中，我們做出了第一代的中文字形產生器和編輯器，這個產生器採用了中研院的構形資料，加上自行開發的部件筆畫比例資料庫，以及畫字的演算法。年底，這個成果發表於中研院（註 12）和日本京都大學（註 13），取得一些回響。不過，我知道這只是一個小小的里程碑，還有一大段路要走。

中文字形是一個非常基礎的工程，凡是越基礎的工程，投資期越長，只有在大功告成之後，才會產生長遠的利益。但是一般的個人和公司，無法承受長時間沒有回報的投入。所以，這樣的工作應該是由國家來進行的。現在回想起來，當

初以一家規模不大的公司，投入這項工作，確實是有點天真而不自量力。有夢最美，這三年多，股東偶而會為回收遙遙無期而發愁，但是大家基本上覺得可以參與這樣有意義的工作，都感到非常快樂。

Accelon3

由於我們重研發而不善經營，2004 年中之後，易符的資金告急，整個開發工作也不得不中斷，我也得另謀出路。於是我又回到老本行，開發全文檢索軟體，畢竟比起中文系統，這是唯一能兼顧生計和志趣的工作。我有一個好友王尚智，每月都要往返兩岸三地，出國就像我出門那麼頻繁。我在他七十坪大的工作室掛單了一年多，在一頭波斯貓的陪伴之下，開發了 Accelon3。

細心閱讀本文的朋友會問，那麼 Accelon2 呢？事實上是有的，Accelon2 是在易符期間，專門為 PDA 開發的搜尋引擎，前後花了近一年的時間，總投資額一百萬。承蒙中研院計算中心和 CBETA 捧場，各買了一套，回收了五萬塊。就商業投資來說，算是慘賠。

Accelon3 就個人而言，算是一個頗堪告慰的作品，除了老客戶印順文教基金會，將之應用於太虛大師全集和第四版印順導師光碟外，還蒙慈濟基金會的採用，作為慈濟四十週年紀念光碟的搜尋系統（註 14）。另外，我也從靜思精舍的師父處得知，證嚴上人也是本系統的愛用者，這個消息讓我樂不可支，高興了好幾天。從此以後，每當我開發程式碰到瓶頸，陷於苦思之際，一想到自己的作品竟然有這麼顯赫的使用者，都能令我激起無比的勇氣，獲得繼續奮戰的力量。

通用平臺

Accelon 的設計理念，其一，要替文史工作者處理瑣碎的技術細節（主要就是缺字處理、搜尋和瀏覽功能，以及軟體的發布），讓他們可以不必為技術問題傷腦筋，將精力專注於內容的製作。

其二，要有良好的彈性，必須讓不同性質的資料，都能在同一個系統上使用。一來是為了避免每做一種資料庫，就被某個程式綁住的窘境；一個理想的系統，程式和資料內容必須分離，不能互相糾纏，否則對使用者來說，每種資料庫要分別安裝不同的程式，本身就是一個大麻煩。二來，軟體環境改變，原來的程式不能跑，資料庫就無法使用了。

這個開放性，對人文界尤其重要，文史工作者對電腦科技的掌握本來就不足，更因為看不出明顯的商業利益，一般軟體廠商當然也不會予以重視。有人曾不無怨懟地對我說：「真像是數位世界的孤兒。」

因此，如果可以整合文史工作者的一般需求，設計出一個通用的工作平臺，將可以大幅降低資料庫的建置、散布和學習的成本。

這聽起來似乎是非常棒的主意，但是實行起來有相當的難度，理由如下：

1. 一般的文史機構和工作者，縱然有能力自行開發應用系統，對本身的需求已自顧無暇，不太可能將別人的需求納入系統的設計，增加無謂的複雜度。再者，文史研究，可以關起門來搞個十年八年，相較於變動劇烈的科技界，開發新工具的動機較小。因此，為這個通用平臺開立準確的規格，本身就是一個相當大的挑戰。
2. 文史界對自己規格不清楚，科技界更不用說。

規格沒有確定，找來電腦專家也枉然。

3. 相較於針對某個特殊應用而量身打造的系統，一個通用的系統，彈性更大、變數更多，因此實作的難度很大。

回顧筆者十餘年來的經歷，似乎冥冥之中就已經由老天做出了安排。漢籍資料方面，處理過古文、今文、辭書，累積高達十億字的資料量。外文方面，處理過目前世界最大的百科 Wikipedia，約莫 20 國語言、數十億字的檔案。此外，這兩年還替印度 VRI 撰寫巴利藏的網路版搜尋引擎（註 15），以及一些藏文大藏經的轉換工作。這些經驗的累積，讓我對大型古籍資料的結構、標誌和特徵，有比較全面的認識。

至於技術方面，這十餘年來，全靠自己摸索，技藝得到不斷的磨練，突然有一天，發現自己可以直接和電腦溝通，對技術的發展和未來的可能性，有了某種程度的預感，因此能夠安於能做和該做的事，不再為層出不窮的新技術所惑。

第三階段：開放式數位古籍平臺

歷史給了我這樣的機遇，同時掌握資料和技術的性質。沒有任何來自國家或創投的資助，唯一的優勢就是由匱乏而來的創意。今天，一個藍圖在我腦海中逐漸成形，這就是本文的主旨：開放式數位古籍平臺的架構。

首先，挑選佛典作為電子古籍的代表，有幾個原因：其一，電子佛典在古籍之中難度極高，比如缺字、目錄結構等。別的古籍不存在的技術問題，在佛經典籍中會發生；別的古籍有的問題，佛經典籍則更為嚴重。其二，資料量極大，包含佛學相關的研究則更為可觀。其三，道家、儒家的古籍，以漢文為主；對佛學資料來說，巴利文、梵文、日文、英文、德文、泰文、緬甸文等，都必須納入考慮。

因此，佛典在芸芸古籍中，有指標性的地位。可以推論，一旦解決了電子佛典，其他古籍數位化的問題也大致克服。

我從三個面向來說明這個架構：其一，所有權；其二，內容架構；其三，技術規格。

資訊的所有權

所有權是資訊服務的核心議題。所有權的歸屬，決定了這個資料庫的形態、服務對象和影響力。這裡依資訊的所有權，大致分為三個類別：

1. **內容及程式私有，付費使用**

內容私有，表示製作者保有內容的大部分權力，其他人不得未經授權使用。修改、販售更是嚴禁，違犯者是要吃官司的。程式亦復如是，必須付費，以取得內容和程式的「使用權」。大部分的文史資料都屬於此類，教界比較有代表性的是《佛光大辭典》、《佛光大藏經》及《中華佛教百科全書》。

2. **內容及程式私有，免費使用，禁止商業行為**

製作者很慷慨地將成果和大眾分享，在道德方面能夠得到比較高的評價，可以將之視為數位形態的「結緣善書」。但是，內容和程式的所有權，依然掌握在製作者的手中；換句話說，使用者可以自由取閱「善書」，但是不能拿去賣錢，也無法直接參與「善書」品質的提升。這個類別，教界的代表作品有《印順法師佛學著作集》光碟，以及眾所周知的《CBETA 電子佛典集成》光碟。

免費結緣固然陳義很高，但是不可能取代有

價商品，變成獲取佛經的唯一管道。因為大部分群眾只會通過正規的市場管道（如書局）來取得資訊，而不會特地跑去寺廟請經。而且，有價的商品會受到市場嚴酷的檢驗和汰選，一般來說都會比較快速地反映客戶的需要，以及擁有較高的品質。

我們說佛經是無價之寶，無價（priceless）是超越性的意思。因此，無論賣錢與否，只是手段不同，目的都是為了佛法的弘揚。現實的情況是，商品受版權法保護，無法任意地複製，這對弱勢者不便；而免費結緣品則排斥商業行為，偏離了市場，意味著更接近廣大群眾，化解這兩者的對立，是能兼容自由免費又不排斥商業行為的機制，其代表是在電腦界發起的自由軟體運動。

3. **內容公開（GFDL），程式公開（GPL）**

GPL（註 16）的全稱是 GNU General Public License，是由美籍電腦工程師 Richard Stallman 在 1989 年起草的。這是他對版權主義（Copyright），即資料和程式私有化所造成的商業壟斷的一種反動。使用 GPL 最有代表性的，是一個名為 LINUX 的作業系統，最初只是一個大學生的習作，十餘年來經由全球有志之士共同努力，加上 IBM 等大企業的全力挹注，功能和穩定性已非同凡響，被視為微軟視窗系統的頭號對手。此外，還有無數採用 GPL 條款的自由軟體，今天掛在全球網路上超過七成以上的電腦，背後所執行的都是這些自由軟體。

GFDL（註 17）（GNU Free Document License），是大約在 GPL 實施十年取得重大成功之後，在 1999 年比照 GPL 的精神，將自由開放的理念，擴大到所有文字資料。目前 GFDL 最有名的案例是維基百科。這個誰都可以編輯的百科全書，以驚人的成長率（註 18），累積資料和人氣，短短數年間，就把《大英百科全書》這個百年老店拋在後頭。維基的創作模式，其影響力只是初試劍鋒，往後的十年間，我們將會目睹這個全球化運動，逐步滲透到各行各業，對傳統的創作模式，產生重大衝擊，甚至塑造出全新的面貌。（註 19）

GPL 和 GFDL 的高明之處，在於強調「自由」而不是「免費」，它並不禁止以商業形式來推廣這個大家創作的成果；也就是說，每個人都可以拿 Linux 和維基百科來賣錢。但有趣的是，正因為生產需要的資料和工具完全攤在陽光底下，反而有效地杜絕了商業壟斷。舉例來說，GPL 並不反對將每套 Linux 以一萬元的高價賣出。但網路上既然可以免費下載，又有誰要買？假使有人願意出這個價錢，那想要取得的不只是 Linux 本身，而是所提供的服務（安裝、維修、教育訓練、諮詢等）。如果提供的服務並不值得這個價錢，那麼收費更便宜的競爭者隨時會出現；換句話說，在自由軟體的生態之下，沒有人可以壟斷商品生產的機密，暴利無法存在，只能收取合理的工本費。

佛經本來就 GFDL

佛典經本，自古以來雖無 GFDL 之名，但是已有其實。在過去，沒有人可以壟斷佛經的所有權。我們可以寫信向佛陀教育基金會索取，或者到文物流通處購買燙金大字精裝本，有錢的大老闆甚至可以從拍賣會標得名書法家手書或是從某石窟挖出的孤本。這個選擇的自由，正是我們要捍衛的。

但是從《大正藏》開始，這樣的情形有了微

妙的變化。日本的原版非常貴（註 20），不要說是一般佛教徒，圖書館和寺廟也不一定消費得起。於是有出版社應大眾要求，推出了價格約為五分之一的影印本（註 21）。從法律的觀點來看是盜版；但是從佛法流通的角度來看，難道不是大功德一件嗎？

感謝 CBETA 爭取到正式的授權，我們得以享用免費的藏經。《大正藏》從很貴到不用錢，是很大的進展，不過所有權還是握在日本人手上，他們可以稍作修改，或者乾脆修改法律以延長著作權的期限（註 22）。

因此，問題就來了，如果有個出版社要印一本佛經導讀，要麼事前取得商業授權，要麼就要自己重新輸入、校對，法律方面才完全無虞。教界花了那麼多力氣，做成了這麼好的資料庫，依然不是公共財富。學者和個人固然可以非營利使用，但是商業使用被排拒在門外。

因此，佛經典籍勢必走向 GFDL，所有權會再度回歸到所有大眾。任何一套電子藏經，只要率先宣布為 GFDL，會以沛莫能禦之勢，聚集了網民和商業的力量，迅速完成標點、內容標記等加值作業，成為新一代的標準。

據此，下一代的電子佛典已經呼之欲出，就是「經文以 GFDL 釋出；製作和運用經文的相關工具，以 GPL 釋出」。對於前者，我心有餘而力不足。但是，製作和運用佛經的工具與平臺，正是筆者所長。

文字編碼層次

即使是將內容限制在佛學研究相關的資料，就有藏經、辭典、百科全書、叢刊、論文、提要、年鑑、年表、名錄等不同的資料。如何將那麼多性質不同的資料共冶一爐，並應付層出不窮的需求？我們必須先回到資料的本質，先從計算機的角度來看待它們。在這裡，我將資料分為文字編碼、文件和資料庫三個層次來探討。

文字的編碼是非常基礎的，在沒有 Unicode 的年代，除了英文之外，其他的語言都要使用特殊的字符編碼，比方說我們用 Big5、大陸簡化字用 GB、日文用 JIS、藏文、梵文等都有自己獨特的編碼方式。

這種不同的編碼，造成資料無法整合。即使勉強放在一起，也會造成處理和查詢上很大的麻煩。而佛學的資料，先天都有多語言的特質，一篇用中文寫的佛學論文，同時用到梵文、日文的機會很高。解決方案就是使用 Unicode，以及搭配相應的字形檔，並且一開始就使用 Unicode。如果已經用了其他的編碼，那麼轉換到 Unicode 的工作越早進行越好。如果到了資訊化的後期才進行轉換作業，就不是「另存新檔」那麼簡單，要面對的是應用程式和整套工具的汰換，以及根深柢固的使用習慣，轉換成本將會非常高昂。

時至今日，大部分的軟體工具已經完成 Unicode 化，除了早期開發的應用程式外，目前最常見的「非 Unicode」應用，就是像 Dia, Foreign1, kh2 等的字形檔。所以，為了您資料的「前途」著想，請儘早改用 Unicode。

文件層次

首先，將文件拆成兩個部分來看，一為內容本身，二為數位儲存的格式。舉例來說，《心經》是內容，而.doc 檔、html 網頁、列印用的 PDF 檔，或資料庫中的一筆資料，就是不同的格式。

格式非常重要，文件的編輯工具、維護成本

和應用範圍皆取決於檔案格式。根據不同的需求，挑選適合的格式，是數位化非常重要的一環。

其中至關重要的，是區分出「來源資料（Master data）」和「延伸應用（Derived works）」，用列表的方式來說明。（見表一）

表一：數位儲存格式

	來源資料	延伸應用
儲存資訊	完整資訊	部分資訊+應用所需的特定資訊
編輯	人工	唯讀、不做人工編輯；或來源資料的變動，必須能夠立即反映
產生方式	人工收集	自動產生
考量點	適合電腦處理	方便人類使用
範例	純文字，XML	Pdf，Doc，Html，排版檔案

來源格式必須使用開放的格式，不能使用特定軟體的專屬格式。就目前來說，個人認為最理想的是採用 XML 作為來源格式，其他的延伸應用，可以使用 XSLT 的技術，自動轉換而得。

來源檔案是人工編輯的主要對象，內容的任何改動，必須能夠自動擴散到其他的格式，如果不這樣做的話，延伸應用越多，維護成本會呈等比級數上升，並且大幅增加資料不一致的機率。這個維護成本和不一致性，是制約資料規模增大的主因。

電腦儲存和運算的能力，這幾十年來都是每 18 個月倍增；相對而言，人力成本不降反升。從這個意義來看，如果以電腦作為主要工具，但是工作效率卻沒有隨著電腦的發展而提升，就是對人力的浪費。或許有人會認為，電腦要花錢買，而義工不必支薪又源源不絕，何來成本云云？對於抱持這樣想法的人，個人覺得非常遺憾！人活著就是巨大的成本，不支薪表示義工自願負擔時間和通勤的支出，並不代表人力不值錢。我們不能只關心切身的利害，而無視於轉嫁到社會的成本。因此，與其用「發心」、「修福」和「耐煩」來開示可憐的義工，不如試著去瞭解如何替電腦和人力分工。當有一天發現，義工忙了幾個星期的工作，用電腦來做只要花幾分鐘以後，將會為揮霍掉的人力，感到無比心痛。

文件發展的三個階段

來源格式的觀點建立之後，就可以探討文件發展的三個階段：1. 純文字；2. 後設資料；3. 內容標誌。

1. 純文字

純文字是記錄內容的基石，能夠穿梭於其他格式之間；換言之，在格式轉換的過程中，其他資訊很容易被丟棄，只有純文字記錄的內容被保存下來。瞭解這個特質，就會明白什麼內容需要用純文字的形式來保存。

2. 後設資料

後設資料（Metadata）是記錄非內容本身的資訊，它有幾個性質：

(1) 會隨著資料規模的擴大而成長。比方說，一個人記筆記的時候，並不需要署名；當很多人的筆記被集結起來時，就必須加上「作者」欄位；而要印成書前，就要分章節和編頁碼。這些本來內容所無，而為支撐整體結構所必須的，就

是後設資料。

(2) 客觀性：後設資料不涉及對內容的詮釋，是客觀的資料。

(3) 每建立一項後設資料，就是增加一種索引。後設資料越豐富，文字內容的應用越廣。

實務上，後設資料有以下幾種：(1) 內容的結構，如篇、章、節、條、項、目、段落。(2) 數位化前的物理結構，如冊、卷、頁、欄、行。(3) 數位版面的資訊：字體、大小、顏色、粗細。(4) 與其他系統整合的介面，如編目格式等。

變動少的後設資料，應該和內容一起，如(1)和(2)。會因不同需求而變化的後設資料，則應該採取和內容分離的作法。例如版面的資訊，應該使用像 CSS、XSLT 的方式處理（註 23）。

以上這幾個結構會有重疊的情況，而由於 XML 只允許巢狀結構，不容許重疊標記，因此這裡提出一種解決辦法，大家可以參考看看。如：將<頁 n="5">內容</頁>，換成<頁 n="5"/>。通常下頁的開始，就是本頁的結束，如果有必要加以指定，就用<頁開始 n="5"/>內容<頁結束 n="5"/>。

3. 內容標誌

目前的數位化，都停留在後設資料的階段。謝教授提出了「內容標誌」的概念，為電子文件開啟了一個嶄新的方向。簡言之，內容標誌提供了一個人文和科技合作的框架，人文學者利用內容標誌，將對於內容的理解和情境，表達給電腦知道。舉個例子，用 5W1H（註 24），來標記新聞內容，因為人事時地物的判定，人遠比電腦勝任。此外，當看到一篇文章中對於情感的描寫，和蘇東坡的〈江城子〉（註 25）很類似，就可以用內容標誌來表達這兩段文字的關係。

對於內容標誌，謝教授已經有精闢的論述（註 26），這裡僅探討如何做技術準備。首先，需要一個夠大的整合式資料庫。對佛學來說，至少得包含藏經、辭典、重要論文等；對文史研究來說，《四庫》、《康熙字典》等是必備的。其中，字辭典占有一個特殊的地位，辭典是原典內容的總索引，從這裡會連結到所有重要原典，並且從連結的頻率和分布情況，可以計算出原典對某個研究的重要程度。

再者，需要一個工具系統作為這個總資料庫的統一存取介面，而這個系統必須非常容易使用，以便一般沒有受過專業電腦訓練的人文專家也可以輕鬆上手。人文專家的參與程度，決定內容標誌的品質。

接下來所要探討的，是這個整合式資料庫的設計方針。

資料庫層次

從技術的觀點來看，常用的資料庫有兩大類別：表格式（註 27）和全文式。

就佛學來說，辭典、百科全書、年表、名錄，可以做成表格式資料；藏經、著作、開示錄，屬於全文式資料。

表格式適用於資料量大、結構明確、修改頻繁的場合，優點是可以對每個欄位做排序、加總之類的運算，並且可以對資料的著錄施以嚴格的檢查，比方說「身分證字號」一欄必須符合一定的法則。表格式資料庫適合電腦處理資料的安排方式，通常人無法直接下指令取用資料，而需要透過程式介面。

雖然表格式資料庫軟體很容易設計出不同的表格結構，不過為了使用上的便利，需要撰寫相

應的程式碼來存取資料，當結構改變時，程式碼要做相應的調整，造成維護的負擔和資料交換的困難。

而全文式的資料庫，優缺點剛好和表格式資料庫相反。全文式資料庫對資料的定義比較寬鬆，比較偏向人類習慣的資料組織方式，無須撰寫特定的存取介面，一般人都可以輕鬆使用。但是也正因為太過自由，除非規定嚴格的標誌，否則電腦很難提取某個文章元素所代表的意義。

一直以來，這兩者有相當不同的工具集。以表格式資料庫來說，有各種不同的 SQL 資料庫引擎（註 28），並且可以搭配各種程式語言來撰寫操作介面。一般而言，需要有比較多的技術知識，才能設計出一個以表格式為基礎的資料庫系統。而全文式資料庫通常會使用現成的工具，如編輯器、Grep 搜尋程式（註 29）等，對技術知識的要求較低。

維基百科使用的 MediaWiki（註 30）軟體，結合了兩者的優點，允許多人同時以熟悉的文字編輯方式來編修資料。再者，MediaWiki 的表格結構是固定的，因此很容易轉換為 XML 格式。就古籍資料來說，MediaWiki 是非常合適的。

由於 MediaWiki 是一個伺服器軟體，需要在網路主機上執行，存放在裡頭的內容，很難整份搬到一般的個人電腦上使用。因此，我規畫了「剎那古籍平臺」，著重於全文檢索、標記介面、缺字處理、單機執行等功能，以補 MediaWiki 之不足。這兩者結合，可以提供從資料生成、多人編輯、內容加值，一直到成果發行的「一站式、一條龍」服務。

這個平臺能大幅降低古籍資料庫的製作、維護和發行成本。如果這個平臺可以促進古籍資料庫的發展，相關的延伸應用與產品，也將會有質和量的飛躍。

為了達成這個任務，我們在 2006 年初成立了「剎那搜尋工坊」，是一個獨立運作、不帶任何宗派色彩的工作室。我們專注於解決技術問題，希望人文學者多多提供寶貴意見，共同努力掃平數位古籍的所有障礙。祈願祖先留下來的文化至寶，在數位時代繼續發光。

剎那古籍平臺技術規格

1. 正式名稱：剎那古籍平臺
2. 開發團隊：剎那搜尋工坊
   ksanaservice@gmail.com
3. 下載網址：http://www.ksana.tw（請加入群組以取得最新動態）
4. 授權方式：以 GPL 授權公眾使用。
5. 搜尋速度：1GB（五億字）以下，平均反應時間小於 0.1 秒；10GB 以下，平均反應小於 1 秒。
6. 索引速度：每秒約 2~4MB，由 XML 產成《大正藏》和維基中文百科資料庫，各需約五分鐘。（運行條件為 2007 年，單價兩萬元左右之個人電腦）
7. 支援格式：純文字，XML，Wiki markup，TEI（可自由擴充）。
8. 缺字處理：動態組字系統，解決所有缺字、新字、錯字的顯示和搜尋。
9. 本系統使用中華民國發明專利 I254863 號。本團隊已於 2007 年 3 月 17 日釋放為公共財，任何人皆可自由運用，免權利金。http://www.zhongwen.tw。
10. 支援語系：目前支援中（繁簡自動轉換）、英、

日、巴利、梵、泰、緬等二十餘種語言，持續增加中。

11. 支援平臺：Windows 2000/XP/Vista、Linux、OLPC、Mac OS X、WinCE PDA/智慧型手機。
12. 其他特色：免安裝，可於光碟、隨身碟、記憶卡直接執行。
13. 發行方式：單機版、光碟版、網路版、PDA 版，使用同一套工具和資料庫，不必分別製作。

（感謝王志攀先生審訂本文）

【附註】

註 1：CBETA 於 1998 年 2 月 15 日成立，http://www.cbeta.org/intro/origin.htm。

註 2：參考摩爾定律 http://en.wikipedia.org/wiki/Moore%27s_Law。

註 3：目前世界上現存最早的有明確時間記載的印刷品是唐咸通九年（868 年）出版的《金剛經》，目前藏於大英博物館，http://en.wikipedia.org/wiki/Diamond_Sutra。

註 4：古騰堡，活版印刷的發明人，http://en.wikipedia.org/wiki/Johannes_Gutenberg。

註 5：漢籍電子文獻 http://www.sinica.edu.tw/ftms-bin/ftmsw3。

註 6：洪朝貴，〈我不用.doc 檔〉，http://people.ofset.org/~ckhung/a/c041.php。

註 7：動態組字的展示，http://www.ksana.tw/accelon/ccg/。

註 8：朱邦復，〈倉頡輸入法與中文字形產生器〉，http://www.cbflabs.com/book/gif_cg/gif_cg/。

註 9：8086 是 Intel 1978 年推出的 CPU，http://en.wikipedia.org/wiki/Intel_8086。

註 10：因為以「門」為部件的字不夠多，如果將「門」配在一個按鍵，會增加重碼率。

註 11：易符智慧科技 http://www.eforth.com.tw。

註 12：「漢字智慧編碼與應用研討會」2003 年 3 月 17 至 19 日，http://www2.ndap.org.tw/newsletter06/news/read_news.php?nid=627。

註 13：書体・組版ワークショップ 2003 年 11 月 28 日至 29 日，http://coe21.zinbun.kyoto-u.ac.jp/ws-type-2003.html.ja。

註 14：慈濟法髓 http://www.ksana.tw/tzuchi_40years_help/。

註 15：VRI 內觀研究中心的巴利藏全文檢索，http://www.tipitaka.org/search。

註 16：GPL http://en.wikipedia.org/wiki/GNU_General_Public_License。

註 17：GFDL http://en.wikipedia.org/wiki/GFDL。

註 18：在 2002-2006 年間，英文維基百科條目數每年倍增，http://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth。

註 19：詳見商智《維基經濟學》，2007 年 8 月 2 日初版。

註 20：http://daizoshuppan.bunkensystem.co.jp/，日本大正藏價目表，全套 88 冊日幣 1,564,500 円，折新臺幣 447,154 元，平均每本 5,081 元。不包含運費。

註 21：根據 http://www.swfc.com.tw/，大正藏 100 冊新臺幣 105,000 元，每本 1,050 元。

註 22：這幾十年，美國已數次修法，將著作權從二十幾年延長到七十五年，http://en.wikipedia.org/wiki/Sonny_Bono_Copyright_Term_Extension_Act。

註 23：關於 XML 相關技術，http://www.zvon.org/ 有非常詳盡的教材。

註 24：5W1H = who, what, where, when, why, how，一篇新聞報導的六個要素。

註 25：蘇軾〈江城子〉抒發了妻子逝世之痛，以及對她的懷念之情。「十年生死兩茫茫，不思量，自難忘。千里孤墳，無處話淒涼。縱使相逢應不識，塵滿面，鬢如霜。夜來幽夢忽還鄉，小軒窗，正梳妝。相顧無言，惟有淚千行。料得年年腸斷處，明月夜，短松岡。」

註 26：詳見謝清俊教授，〈後設資料與內容標誌〉，http://www2.ndap.org.tw/newsletter06/news/read_news.php?nid=1498。

註 27：正式的名稱叫關聯式資料庫 RDBMS，http://en.wikipedia.org/wiki/Relational_database_management_system。

註 28：SQL 資料庫引擎 http://en.wikipedia.org/wiki/SQL。

註 29：廣為使用的文字檔搜尋工具，但中文詞被斷行隔開則無法搜尋。http://en.wikipedia.org/wiki/Grep。

註 30：MediaWiki 只是 Wiki Engine 其中一種，是最為知名，http://en.wikipedia.org/wiki/Mediawiki。