803

從全文資料庫到數位典藏

——中央研究院的發展經驗談

摘要

1984 年中央研究院開始嘗試開發全文資料庫，至 1990 年，四千萬字的二十五史全文資料庫第一個版本正式啟用。然而，全文資料庫開發工作至此不但沒有停止，反而持續增加。從統計數字來看，90 年代初期每年約增一千萬字，至 1998 年之年增量已近五千萬字。目前，已上線的全文資料庫總字數累積約二億字，尚有二億五千萬字已校對好資料等待上線。

中央研究院全文資料庫的這種高速成長，似乎是有政策、有組織的領導所致，其實不然，這完全是各所自動自發所匯集的結果。為了因應各所開發全文資料庫引發的問題， 1995 年成立了『中央研究院漢籍電子文獻協調小組』，一年後改制為委員會，並負起規劃本院數位典藏之任務。目前所規劃的方向，已從對全文資料庫的製作轉向為對人文社會學科『數位研究環境』的開創，其中包括開發新四庫全書、數位圖書館/博物館、漢學工作站等等構想。

由於計算機科學及信息處理技術突飛猛進，這十多年來計算機環境也變得和以往大不相同了。回顧過去十五餘年所經之歷程，不只是本院漢籍數位化的歷程；也是從圖書館自動化到數位圖書館的過程；也是個人電腦、網際網路等成長和發展，波瀾壯闊的年代。本文將以此背景，報導本院數位化歷程中所學到的經驗，以為同好評議、參考。

謝清俊

中央研究院 資訊科學研究所 研究員

1999.0930

壹、 緣起

自從科技高速發展以來，人文和科技逐漸乖離，在社會中似乎形成了兩個完全陌生且不相往來的團體。在國外，史諾教授一本《兩種文化》的小書，毫不留情的痛陳此弊，引起了西方世界極大的關切【1】。持續至今，如何調和人文和科技這兩個領域，仍然是東西方國家內政上的重要課題。由於我國文化悠久深厚，累積本來豐富，再加上近數十年來對科技生吞活剝地急起直追，使得人文和科技的鴻溝遠比西方國家為大。明顯的徵兆是古代文獻離我們日常生活越來越遠。換言之，我們數千年人文的累積竟越來越無助於時下生活中的問題。

一、史籍自動化計劃

有鑑於此，為了中華文化的延續，而不任其在科技的洪流中式微沒頂，解決的方法則是務必要將古籍以電子媒體表達，使古籍能活出現代風貌。此即中央研究院（以下簡稱本院）推動史籍自動化計劃之初衷。1984 年 7 月首期史籍自動化計劃由史記的食貨志開始，獲得了相當的經驗和成功。 1985 年 7 月開始做前四史，即《史記》《漢書》《後漢書》《三國志》。1986 年前四史順利完成後，便擴充至其餘各史，至 1990 年 6 月大致完成，唯不含表格，且有二千餘缺字待補。1992 年 9 月缺字補齊。1995 年 3 月計算中心開始補充表格部份，至1997 年 1 月，表格部份已全部完成。

二、古籍全文資料庫

本院利用計算機處理古籍以全文資料庫的發展最受矚目【2,3】。所謂全文資料庫，是以原文件的所有文字為素材，以儘量保存文件版面的方式所建構的資料庫。目前上線的全文資料庫，總字數已超過一億八仟萬字，其所用的技術則全由院內同仁自行開發。參與製作資料庫的共有五所：史語所、臺史所、資訊所、近史所、文哲所，以及本院計算中心；院外，則有總統府國史館亦積極參與清史資料庫之開發。1995 年開始，有些大學與本院發展合作關係共享古籍資料，包括國內：中山、中正、師大各大學，國外：香港中文大學、倫敦大學、史丹佛大學、密西根大學等等。

目前要做資料庫，其步驟甚單純，即在選定文獻後，經過繕打、校對、標誌後，即可上機測試，若無毛病，資料庫即已大功告成。從上述的情形看來，本院發展的全文資料庫製作技術已趨成熟，各所均能自行主導開發資料庫，計算中心的協助只是標誌方式的諮詢、中文造字的協調管理、建構資料庫時的機器操作，與必要的技術開發、營運、管理和維護而已。

三、自發成長

1990 年，四千萬字的二十五史全文資料庫第一個版本，正式啟用。然而，全文資料庫開發的工作至此不但沒有停止，反而持續增加。從統計數字看，90 年代初期每年約增一千萬字，至 1998 年之年增已近五千萬字。目前，已上線的全文資料庫的總字數累積約二億字，尚有二億五千萬字已校對打好資料在等待上線。

本院全文資料庫的這種高速成長，似乎是有政策、有組織的領導所致，其實不然，這完全是各所自動自發所匯集的結果。為了因應各所開發全文資料庫引發起的問題，1995 年成立『中央研究院漢籍電子文獻協調小組』，一年後改制為委員會，並負起規劃本院數位典藏之任務。目前所規劃的方向，已從對全文資料庫的製作轉向為對人文社會學科『數位研究環境』的開創，其中包括了開發新四庫全書、數位圖書館/博物館、漢學工作

站等等構想【3】。

四、多元發展

本院處理古籍的計劃並不限定於全文資料庫技術，有許多資料是使用關聯式資料庫處理。諸如，1985 年 10 月開始嚐試的「漢代墓葬綜合研究資料庫」，1986 年 2 月的「台灣土著語言資料庫」，1986 年 4 月的「台灣日據時代戶籍資料庫」，1987 年 1 月的「清代竹塹地區土地申告書資料庫」，以及 1989 年計算中心所做的「說文解字和玉篇資料庫」等等。也有利用影像處理技術所做的古籍資料庫，如傅斯年圖書館發展的「善本書影像資料庫」，目前已完成該館近半數善本書的典藏，並已開放使用。近四、五年來，本院在近代史料、自然科學、語言學、社會科學和生命科學等資料庫的開發，亦有小成。這些資料庫和古籍的全文資料庫也是相輔相成、相得益彰的。有意一探的讀者，可參考本院的網站：http://www.sinica.edu.tw/ 中的「學術資源」項目。

貳、 從圖書的數位化到文物的數位化

從 1994 以後網際網路(Internet)的普及，逐漸改變了人們溝通、處理信息與知識的方式。對社會來說，在此變遷下，知識的取得、擁有和利用也隨之改變；對學術界而言，新的工具引導了研究環境的改變，也擴展了知識領域範疇，新知獲得的速率是前人所料未及的。這些變化雖已驚人，但僅是資訊科技與網際網路，所引發文化巨大變遷中的幾個例子罷了。這個文化變遷，無疑的，將巨幅改變人類的生活形態、社會結構和文明的內涵；而且，此變遷已經開始，勢之所趨無可規避【4】。處此變局，不可避免地，我們要面對從舊到新的轉化過程，「如何把重要的文化資產數位化」就成了我們必須面對的重要新課題。

從圖書與檔案(archives)的分野來看，檔案數位化應較圖書更具效益。一則是因為目前圖書的流通和使用遠比檔案方便；二來檔案是第一手資料，其價值比圖書更珍貴。然而，由於檔案的稀有，使得它在應用上受到很大限制。如果將檔案數位化，廣予流傳，其效果勢必空前。其實，這情形不僅檔案如此，珍貴文物又何嘗不然？所以，從圖書數位化到檔案、文物的數位化是相當自然的，也是在資訊技術急速發展下勢之所趨。

有鑑於此，本院『漢籍電子文獻協調委員會』，在集思廣益努力下，對於本院珍藏文物的保存和應用，也有全面新體認。本文以下，將以此新構想和體認為藍本，作一簡要的說明。

一、 數位媒介與文化變遷

電子媒體有極優越性質，將它數位化的目的是多重的，要言之：一是利於長久保存；其次是數位化後，幾乎取之不盡、用之不竭，可供全民共享；再次是可以大量匯集知識，經相互鉤稽參照，能發前人所未見，產生相輔相成(synergy)的效果；又，如有四通八達的電腦網路，則數位古籍幾乎不須花錢就可以瞬息千里。其實，文化資產數位化的好處並不止這些，以上不過舉其大者而已【5,6】。這就是為什麼聯合國要推動 Memory of the World 計畫，以及美國推動 American Memory，加拿大做 Canadian Heritage 等等計畫的原因。

從歷史觀之，任何新媒介的引用，都將導致溝通行為及其效果的改變，也會使得人們在知識處理和發現上產生革命性變化。這兩個變因都對文化影響至為巨大：即新媒介的引用將引發新文明產生。從歷史上觀察，這個因果關係是絕對真實的，因為從來沒有例外。

有鑑於此，數位媒體將帶動史無前例的文化變遷，這已是學術界不爭的共識。古籍

數位化實在是使古籍活出現代風貌最佳、也是唯一的選擇。我國累積有豐富的文化資產，是世界的瑰寶，然而，我們有能力這麼做嗎？我們的環境適合這麼做嗎？時機對嗎？這麼做有什麼社會效益？有什麼前景？想探一探這些問題的企圖心，正是促成本文思考的背景。

二、 全觀

我們明白，如果要將優良的文化資產數位化，是一個極鉅大艱難的過程，不是少數人能做到的，也不是某一個專業可以做成的。這需要全民參與、跨領域合作、並作長期奮鬥，才能成功。因此，鼓勵跨領域，尤其是跨人文社會和科技的合作，至為重要。於此，重視相關計畫之間的溝通協調和合作，發展公用的工具、規格和標準的設立與共享、資訊的透通無阻，以及重視社會整體發展機制之建立和社會效益的評估等，便使得做事的觀念、態度和方法上起了根本上的變化。

三、 數位典藏的文化工程

為建立上述文化主題的數位典藏，必須作該特殊領域中專門知識在電腦中如何表達、如何應用、推廣的研究。於此，應強調對既有資料和知識的整理，以及它們在計算機內部的表達，而不在對文化主題資料的收集，也不在對文化新知的追求，雖然執行的結果一定會獲得相當程度的新知識、新技術和在文化上的創新。數位典藏的建立，實是一項碩大無比的文化工程，含有對固有文化全盤的整理、系統化、以及詮釋和表達等等工作。此工程在資訊技術上，要選擇適當的技術配合，作創意性的應用。建構數位典藏重點之一在於，資訊技術應用於文化上的意義和創新(是 know how 而不在 know why)。是故，技術內涵應具有創新性、通用性、相容性和規範性。

四、 文化的共同座標

由於文化主題有共同的時、空和語言文字基礎，建構共同的時空和語文「坐標」用以安置文化內容，就成為製作文化主題的數位典藏必要的基礎建設。如，共用或相容的地理資訊系統（含時空兩相度）、中外曆法系統、各式語文系統（含時間、語文兩相度）等，它們是可建立為一個共同的環境，讓所有文化主題和所有使用者共用共享的。

五、 公共資訊系統

文化無國界，文化資產是全世界人民共有的。於此可推知，應優先鼓勵建構「公共資訊系統【7】」以鼓勵文化交流和共享。此所謂「公共」，對於一個國家而言，是指其使用權是公開屬於大眾、全體納稅人所共有的。換言之，在「正當使用」範圍內，大眾可自由使用。公共資訊著作權和所有權必需清楚，不可有爭議。任何使用者可依清晰合理且公平的條件作二次加值開發。對於公共訊的內容或功能之加值機會，應是人人平等的。然而，目前真正全然屬於公共的資料庫或檔案極少；即使資料是屬於公共的，各式各樣的工具和軟硬體卻未必然。於此，對使用資料庫和網路相關的軟體，如搜尋、瀏覽之工具等應避免壟斷，以使資料庫和資料能共享、可攜（portable）和永續使用。

雖然目前公共的資料庫或檔案極少，然而，屬於公共的原始資訊卻並不少，無數沒有智慧產權的書本、文獻、檔案散居於各機構。因此，將這批資料數位化，便可能建立

相當可觀的公共資訊系統。從另一方面看，任何目前私有的資訊，其智慧產權都是有固定期限的；因此，當時過境遷，這些私有資訊終將化為屬於公共的原始資訊。換言之，公共資訊系統的建立，是勢之所趨，無可規避。且公共資訊的認定是建立有價資訊的基礎。換言之，沒有良好的公共資訊系統，則所有的資訊加值將無所依據。此事悠關民生至鉅，建立公用的國家數位典藏作為公共資訊系統的主軸，應該是目前政府不可輕忽的責任。

六、社會效益

此所指之社會效益，大致上可以分為三個方向考慮：一是對教育和學術上的；其次是對相關產業的；再次是對一般國民的貢獻。為求社會效益之成效，應注重其使用時的易用性和親和性。

在考慮社會效益方面，研究、開發、產業和應用四者必須要有密切的聯繫與配合。對產業而言，如軟體產業、加值產業(value-added industry)、內容產業(content industry)等，這些產業雖然都是以資訊業為主，然而其影響與應用絕不限於資訊業，在文化、藝術、社會等等各行各業的應用是可預見的。例如：[表一] 第四點產出清單中，可得知各層次產品均有不同的產業效益。

參、 實踐的個案—數位博物館專案計畫

1998 年六月，在國科會推動下，遴選臺灣大學和本院為主力，從事數位博物館計畫的嘗試。在此計畫中，共建構八項文化主題：在自然科學方面有：『臺灣的植物』、『臺灣的魚類』和『臺灣的蝴蝶』，在傳統人文方面有：『不朽的殿堂—漢代的墓葬文化』、『古代的文學與思想—四書、老莊、和唐詩』、『火器與明清戰爭』，綜合性的有：『淡水河溯源』、『臺灣的原住民—平埔族』等。這些主題的內涵，都以豐富的學術資料為基礎，其正確性、精確性、週延性、擴充性、適用性等等，市面上商品均無可比擬。換言之，我們試著建立的是一個共同的知識結構，此知識結構的內容可以透過界面作各種呈現，以適應不同的應用之需，如：研究、教學、商品開發和個人使用等。

此外，在共同數位環境開發上，則有五項支援技術，它們是：『人文與自然資源地圖』、『語文知識網路—搜文解字』、『資源組織與檢索之規範』、『系統評估』和『數位典藏系統先導計畫』的開發等。人文與自然資源地圖和語文知識網路兩者，是試圖建立共同時空和語文環境。資源組織與檢索之規範計畫則著眼於文本內容的標誌(markup)、文物的後設說明(matadata)、檢索機制（如索引典）等程序知識的開發。系統評估是試圖將評估作業納入系統開發的程序之中。數位典藏系統先導計畫則是以開發共用資訊技術為主，如加密技術、影音技術和共用的工具程式等等。

數位博物館專案第一期結構圖如 [圖一] 所示，到今年九月底，第一期將告一段落。在過去的一年多裡，數位博物館專案計畫已取得相當的進展【8】。[表一] 所列者，是此計畫建構數位主題的稽核條目。相信不久，此計畫將陸續公開一些類似 [表一] 的資訊供各界採用。由於數位博物館計畫初步成功，1999 年七月行政院主辦「電子、通訊、與資訊策略會議」中，通過了將重要國家文化典藏數位化的提案。此計畫將於 2001 年正式

展開。

肆、 結語

當網路時代來臨時，處理知識的角色全都面臨巨大變遷考驗，研究者、教師、圖書館、博物館、檔案典藏都不能例外。在思考這些問題時，不可避免的都要回溯至問題最基本的源頭—包括其眼光、旨趣、環境、目標等等。而這些思考又不能獨善其身，資訊科技早已撤除了傳統藩籬、模糊了存在的疆界，迫使我們要顛覆和重組既有的架構。這局面正是目前面臨文化變遷中的一環。處此情境，放開心胸勇於嘗試是應該鼓勵的，但是，甚麼才是正確的方向？甚麼才是正確的手段？甚麼才是正確的步驟？甚麼才是正確的時機？恐怕只能盡你我之力，並於事後印證了。

從全文資料庫到國家數位典藏這一段歷程中，有太多刺激、太多學習、太多顛簸和太多感慨。凡此種種不可能在這一篇短文中忠實全然地呈現。本文所報告的，是我們自以為是的一些重要觀點和想法。讓我們謹以此文恭賀新亞書院五十週年大慶，並謝謝主辦單位讓我們有機會在此就教於各位，各位的指正和批評正是我們精益求精的良機。

[表一]：建構數位主題的稽核條目

一 原始典藏概述

1 主題名稱
2 典藏主題之文化特色與價值
3 目前典藏內容概述
說明文物之品質、數量、來源、結構和組織、歷史背景和相關文件、文獻等。
4 典藏文物之所在地、管理維護單位。
5 典藏機構之同意書，和對智慧產權的要求。

二 建構之數位典藏說明

1 欲數位化之文物清單。
2 各種文物數位化之方式和規格（請註明理由）。
3 數位典藏主題之內容與其結構、功能。
4 說明數位典藏之呈現方式、劇本構想、導覽結構（圖）、參照、注釋解說、檢索功能和使用者界面等。
5 逐項說明對此數位典藏智慧產權的要求。

三 資訊技術之運用

1 說明整理文物所援用的主要技術，以及整理工作的內容。
2 說明典藏文物之資訊，如何在電腦中表達 (representation)。
3 說明典藏文物之資訊，在網際網路上傳遞與在終端機上呈現之構想。
4 逐項說明所引用之標準和規範，並說明理由。

四 產出之清單

1 各個原始的數位檔案（原數位檔）。
2 各個已加標誌的原數位檔（標誌檔）。
3 各個屬性或背景資料的檔案。
4 各個已有主題且已結構化的屬性或背景資料檔案，如主題詞表、索引詞表、索引典、破音字表、近（同）義詞表、詞網等等。
5 各個資料庫。
6 各個可供參考或共用之工作流程，包括行政上和技術上。
7 各個可供參考或共用之程式和系統。
8 各個其他申請人認為有價值的產出。

[圖一]：1999 年度數位博物館專案主題計畫及技術支援計畫關係圖

【主題計畫】 \ 【技術支援計畫】	人文與自然資源地圖	搜文解字－語文知識網路	資源組織與檢索之規範	系統評估	數位典藏系統先導計畫
鄉土風情					
人文 淡水河溯源			•	•	
臺灣原住民 —— 平埔族群		•	•		•
自然 蝴蝶生態面面觀			•		
台灣的魚類 台灣的植物	•	•	•		•
傳統文化					
傳統思想與文學(四書、老莊、唐詩)		•	•		
不朽的殿堂--漢代的墓葬與文化	•	•	•		•
火器與明清戰爭			•		

參考文獻

1. 史諾 (C.P.SNOW) 在 1959 年發表了《The Two Cultures》一書引起世人的注意。之後，在 1964 年增加了一些對外界意見的回響，更版為《The Two Cultures and A Second Look》
2. 謝清俊 林晰 〈中央研究院古籍全文資料庫的發展概要〉1997，見網頁：http ://www.sinica.edu.tw/~tdbproj/handy/thesis.html
3. 黃寬重 劉增貴〈中央研究院人文計算的回顧與前瞻〉中央研究院 歷史語言研究所，1998
4. 謝清俊 〈Digital Media and Culture Change〉Digital Museum Seminar 1999, keynote speech, 中央研究院 1999.7
5. 謝清俊〈談資訊的定義與性質〉資訊科技與社會轉型學術研討會，中央研究院，臺北，1996.12
6. 謝清俊等《資訊科技對人文、社會的衝擊》經計建會委託研究報告，1997
7. 謝清俊 〈公共資訊系統概說〉行政院科技顧問組，1996
8. 請參考其網頁：http://www.dmpo.sinica.edu.tw