中央研究院 歷史語言研究所
漢籍全文資料庫系統改版發表會
中華民國九十四三年九月二十二日

漢籍全文資料庫的緣起、沿革與展望（~~930909~~940909 初稿）

謝清俊
玄奘大學 圖書資訊學系 講座教授

光陰似箭，史語所漢籍全文資料庫的發展至今已滿二十年了。若不是為了這個報告，還沒有意識到逝者如斯。為了這個報告，李芳瑩小姐幫忙蒐集了豐富的資料。閱讀這些資料，往事歷歷在目。

二十年來，漢籍全文資料庫的發展從未間斷過。顯然，未來仍將繼續。這是不是中央研究院歷來最長久的研究計畫？即使目前不是，也有一天會是吧！參與此計畫的人甚多，計畫主持人就有毛漢光（73 年 5 月至 75 年 6 月）、丁邦新（75 年 7 月至 78 年 6 月）、謝清俊（計算中心主任身份，74 年 7 月至 75 年 6 月，76 年 6 月至 79 年 6 月）、劉錚雲（78 年 7 月至 79 年 8 月）、黃清連（79 年 9 月至 ~~81~~80 年 ~~1~~ ~~月至~~ 6 月）、柳立言（80 年 1 月至 6 月）、劉增貴（81 年 7 月至 82 年 6 月）、魏培泉（82 年 7 月至 83 年 6 月）、陳若水（83 年 7 月至 86 年 6 月）、劉淑芬（86 年 7 月至 87 年 1 月）、林富士（87 年 1 月至 89 年 12 月）、李建民（90 年 1 月至 92 年 12 月）、邱仲麟（92 年 1 月迄今），等十三位。由此可知，史語所和計算中心對此計畫的重視、熱衷和支持，這也說明了史語所和計算中心二十年來的合作無間。

史語所和計算中心對此計畫持續投入大批人力。於此，恕不能列舉芳名，謹報告一些統計數字作參考。研究人員（包括共同主持人、協同主持人、副主持人、計畫顧問、校對顧問、協同研究人員等；其中大部份是史語所的人，僅幾位計算中心的人）共 ***48*** 人年。漢籍全文資料庫的品質是學界有名的，這不能不歸功於史語所的專業助理、校對及建檔人員的素質和辛勞。這些工作至今總共投入了約 ***260*** 人年。以上，總計史語所投入的人力，在計畫書上列名的，已超過三百人年。

早期（73 年 1 月至 77 年 6 月）計算機中心以 ***21*** 人年的投入，完成了初步的系統及程式設計。之後，約以平均每年三至四人的人力，做漢籍全文資料庫的維護和更新的工作（改版投入的人力，尚未包含在內）。這工作也已持續了 ***16*** 年。所以，計算中心為此計畫投入的人力，在計畫書上列名的，也已約超過一百人年。從計畫規劃和管理的角度來看，人文與資訊科技投入的人力比是三比一。這數字對數位典藏或其他數化的計畫，是值得參考的。

緣起與沿革

關於此計畫最早的記載，是史語所所務會議民國 73 年 1 月 19 日的記錄。~~自~~從此正式展開籌備「史籍自動化」和二十五史全文資料庫計畫。第一個計畫「~~四~~史食貨志」於民國 73 年 5 月 19 日開動。迄今，精確的說，已二十年四個月又四天。私以為，5 月 19 日是個值得慶祝的日子。

史籍自動化的想法萌發甚早，可追朔至民國五十八、九年。當時，毛漢光先生在哈佛大學訪問，筆者在麻省理工學院做博士論文(筆者非麻省理工學院的學生，拿的是交通大學的學位)，而我們正巧分租同一個房子。毛先生和筆者曾談到史籍自動化的想法。不過，當時電腦中還不知如何處理漢字，既無字碼亦無字形，更不必說漢字的鍵盤、檔案系統、顯示螢幕、列印設備、處理程式、輔助記憶設備、以及傳輸和網路了。這些都沒有。所以，當時只是在編織一個美麗的夢、，所談的都不過是一個願望、一個期許罷了。然而，這個夢一直縈繞於心，未曾忘懷。

民國六十年，馬志欽教授在國科會呼籲從事電腦處理漢字的研究。這是以國家政策推動電腦處理漢字文獻的肇始。民國六十三、四年，初步解決了電腦處理漢字字形的問題 [1.2.]。之後於七十四年左右，完成電腦裡中文字碼的設計與工業標準的建立 [3.]。當這些基本的工作稍具規模時，筆者曾於七十一年協助潘重規教授以電腦處理《紅樓夢》，這經驗可說是創作二十五史全文資料庫的濫觴。當時隨筆者做這些研究的學生，丁之侃和曾士熊，也於七十三年來中央研究院計算中心。至此，中央研究院有了可以發展全文資料庫的環境 [4]。

有一件特殊的因緣必需一提。民國七十二年春，筆者偶爾遇到了毛漢光先生。當時，筆者向毛先生請教轉任中央研究院的可能。毛先生立即表示熱情的歡迎之意。他提起題起我們在劍橋談到的夢想，並且指出：中央研究院是最適合從事漢籍資料庫開發和成長的地方。毛先生鼓勵筆者到中央研究院來和他共同完圓

~~此~~夢~~想~~。這段談話，促成了筆者於七十二年八月轉任中央研究院資訊科學研究所。此因緣，感念至今~~仍感念於心~~。

當第一個計畫「史籍自動化：食貨志輸入電腦」[5]啟動的時候，美國的各種全文資料庫計畫也正剛開始。所以，就全文資料庫的研發而言，臺灣幾乎與世界領先的學術研究同步。以下，謹談談兩件研發上的花絮，以為沿革資料，供讀者茶餘之資。

其一，當時國外的全文資料庫，幾乎都是以書本的「頁」作為檔案分割的基本單位。筆者以為不妥，要求系統設計人員依文章的篇章節段為基本單位。此事現在看來是理所當然，但是當時卻遇到明顯的反對。反對的主要原因是，若依篇章節段劃分，~~則~~技術上複雜很多，而且沒有任何文獻可以參考（沒人做過）。幾經溝通，系統設計人員都不以為然。最後，只好命令他們依篇章節段劃分收場。與此類似的，還有螢幕直式中文~~的~~顯示的研發。這功能也是在強制下才能推出的，可惜目前已~~刪除~~看不到了。~~，但是~~當時~~確~~它是吸引人文學者的有力賣點。

其二，當時尚無文獻標誌的標準。國際標準組織（***ISO***）的標準通用標誌語言（Standard General Markup Language, ***SGML***）在 ~~1986~~ 民國 75 年才發表。所以，我們一邊設計自己的文獻標誌標準，一邊各方打聽、設法，希望盡早取得 ***SGML*** 標準的初稿（draft standard）和內部設計資料。75~~1985~~ 年夏，終於如願以償。我們自己做的文獻標誌標準，是以數學裡樹狀結構為骨幹，配合著篇章節段而構成

的。與 ***SGML*** 標準初稿比對，發現二者完全相容（***SGML*** 以制式語言為發展的基礎，其功能較樹狀結構為大）。據此，工作團隊信心大增，即以自己的文獻標誌標準，於 ~~1988~~77 年完成了系統設計，當時系統內的資料只有前四史。順便一提，由於這段因緣，***SGML*** 是本計畫將引進臺灣的。這是鮮為人知的往事[6.]。

從整體看來，民國 73 至 77 年是全文資料庫的初期研發階段。此時，全文資料庫的雛型業已完成，而史料則只輸入了前四史。77 至 79 年主要的工作是全文資料庫的改善和全部二十五史資料庫的完成。80 年後的工作則約分兩路進行，史語所持續擴充數化史料的量，計算中心則做系統維護、更新、以及改版的事，其中為網際網路、***HTML*** 所改的版，最為人知。

展望

筆者曾半開玩笑地對朋友說過，做個資料庫比生個孩子還累；生孩子也許只要照顧二十年，可是照顧資料庫卻是永無休止的負擔。此言照顧，主要包括資料庫的維護、更新和改版。其中，改版一事又好比是資料庫在資訊科技的無常（進步）裡輪迴；若不「投胎轉世」，即配合電腦硬體、軟體和網路的環境改版而再生，則資料庫遲早會「形神具滅」。歷年來，「形神具滅」的資料庫，可說是死傷不計其數。漢籍全文資料庫能存活二十年，是大家努力的成就，值得慶幸、珍惜。雖然如此，依筆者一己之見，漢籍全文資料庫的發展仍有不足之處，願提出來與讀者討論。

漢籍全文資料庫初期的發展，無論在理論和技術方面，在世界上都居領先的地位（國外的全文資料庫沒有中文資訊處理的技術）。可是，自從數化的《四庫全書》[7]和數化的《大正藏》[8]全文系統出現之後，顯然已形成各領風騷的場面，漢籍全文資料庫已失去了唯我獨尊的地位。《四庫全書》的主要優點是有原書的圖檔，並能與打字的文本互相對映。《大正藏》則有符合 **TEI** 標準的文獻標誌系統和處理多版本的功能。根據這些事實，我們不能不認為，漢籍全文資料庫近十年來改版的速率的確慢了些，已被後起之秀~~超越~~追上了。

為了方便討論漢籍全文資料庫的展望，~~筆者~~讓我們以~~特別提出個~~「文物數位化通用模式」[9]作為~~討論的~~引子。此模式分為兩部份：其一是單一文物的數化模式，其次是數化文物之間的關係。茲將此模式簡述如表一，並請參考圖一。

表一：文物數位化通用模式

壹、單一文物的數化分三個層次：

其一是外觀的數位化：通常是把文物打字、掃描，如文獻中的文字、版面、照片...等。一件文物的數位化，外觀是必要的基礎成份。

其次是背景資料的數位化：舉凡一份文件的書目、作者、譯者、屬性、圖說，或類似的支援資料等，都屬此類。背景資料是一些客觀的資料，包含有時下流行的後設資料。背景資料的對錯是只可以考證而不容詮釋的。

再次是內容的數位化：內容的數位化包含對內容的注解、詮釋、分析、考據等。內容的詮注可以依某一個文獻、主義、理論、研究的方法等作為立論所在。內容的數位化即是文物相關知識的整理和再現。沒能做到這一部份，數化就失去了應有的深度和廣度。

貳、數化文物之間亦分為三種關係

其一是互為文本(Inter-textuality)：***Julia Kristeva*** 說，任何文學史學之間沒有獨立存在的文字，這些文字一定跟其他的資料之間有某種關聯。如果把這些關聯找出來就是一個文學的知識庫（knowledge base）。文史資料跟其他各學科之間還有關係，這類不限領域文件之間關係都屬互為文本的參照。

其次是情境(context)的參照：從背景說，情境的參照可分為文化背景、社會背景、以及個人背景的情境；從傳播的角度來講，有作者情境及讀者情境。早期的口語傳播沒有這個問題，因為口語傳播是面對面，讀者跟作者處在同一個情境。當傳播或溝通的機制越來越進步，便幾乎把作者情境及讀者情境斬斷，成為越來越獨立而交集越來越少的狀態，越分越遠。情境的參照是目前數化尚未接觸的領域，這些不同情境的資料怎麼表達，是文物數位化面臨到較大的挑戰。

再次是文與物彼此的參照：如紅樓夢一書中對食譜、建築、花卉...等的參照就是很好的例子。這是人文跟自然的結合。

依表一的模式，對漢籍全文資料庫的展望，可作如下的討論。

一、從單一文物數化的角度檢討：對於外觀的數位化、背景資料的數位化兩方面，都有值得改善之處。如增加原書之影像（像《四庫全書》），或增加各古籍之後設資料以資利用等。對於內容之數化方面，可考慮發展內容標誌標準（包括注疏）、~~增進注疏，或~~增加版本之處理功能（如《大正藏》）等。

二、從數化文物之間的關係檢討：目前漢籍全文資料庫大致上是一個封閉的系統。也就是說，它孤芳自賞，除了與歷史地圖的地理資訊系統曾嘗試作聯繫之外，和其他數化系統間所作的聯繫是很值得改進的。

在互為文本方面，系統本身中各古籍之間彼此間的相互參照，即有加強的需要。如，二十五史與十三經之間的聯繫。在文與物之參照方面，則與數位典藏系統之間的聯繫，應是可規劃的方向。對情境參照方面，似可利用

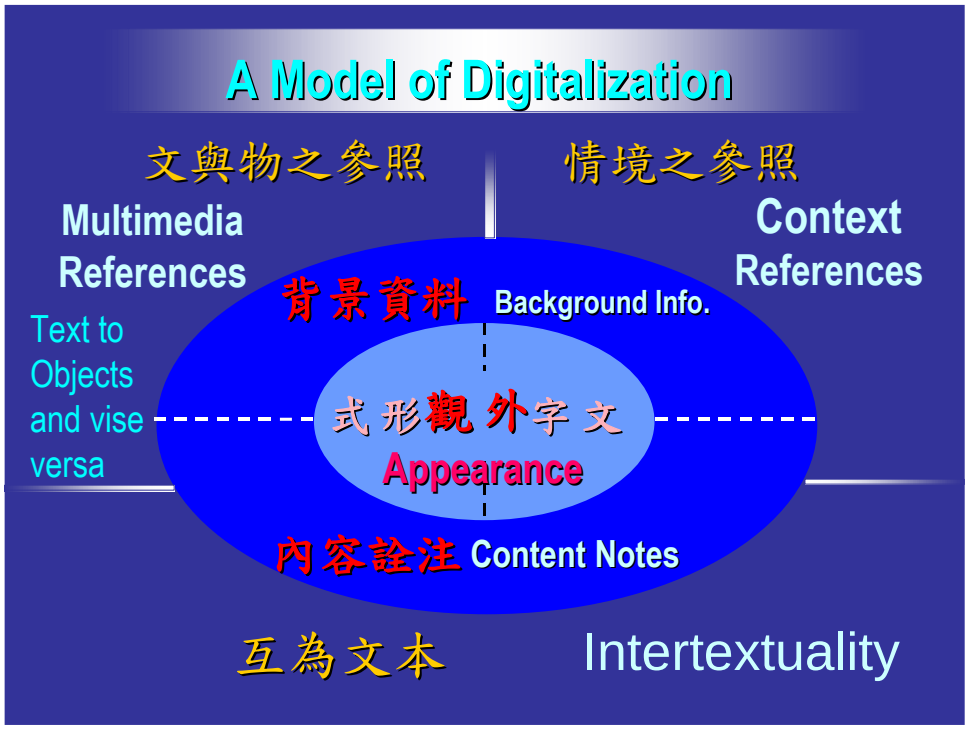
數位典藏所發展的各式後設資料，來增進系統的功能。

結語

上述的討論，並不是很仔細、深入的，僅僅是舉些例子，期以拋磚引玉。除此之外，關於國際上數化文獻相關標準的引用，文獻處理資訊工具的開發，漢學研究工具的數化，使用者的聯絡、組織和需求的調查，以及改善整體系統結構和功能的努力，則是眾所周知的經常性~~的~~工作和責任。~~，~~

今天欣逢漢籍全文資料庫改版的盛事，除了賀喜之外，也向歷史語言研究所和計算中心表示敬意，是他們~~也是~~使漢籍全文資料庫能夠誕生、成長、茁壯和永續經營~~必要的承擔~~。

圖一：**文物數位化通用模式示意圖**



注：

1. 請參考：謝清俊等著，中文字根的分析，交大學刊，第六卷・第一期，1973 年 2 月。

2. 請參考：謝清俊等著，中文字根的貯存和文字的合成，交大學刊，第六卷・第一期，1973 年 2 月。

3. 請參考：謝清俊、黃克東著《國字整理小組十年》資訊應用國字整理小組，文建會發行，1989 年 12 月。

4 曾士熊、楊鍵樵、謝清俊，***An Experimental Model of Chinese Textual Database***：（中文全文資料庫之實驗模型），中國工程學刊，13 卷 6 期(1990 年)：頁 607-622。

5. 毛漢光等《史籍自動化：食貨志輸入電腦 第一年總報告》臺北市，中央研究院計算中心，1985 年 7 月。

6. 本計畫也介紹了兩位教授，蘇克毅和朱四明，與筆者三人共同參與 ***ISO*** 文獻標準技術工作小組（包括 ***SGML***），為起草東方語文相關的標準服務，達七年之久。

7. 香港迪志公司出版。

8. 法鼓山中華佛學研究所研發的系統。

9. 謝清俊,〈文物數位化雜議〉**,** ***Workshop on Possibilities of a Knowledgebase of Tang Civilization,*** 京都大學人文科學研究所, 論文集排印中, 2004 年 2 月 21 至 22 日。