中央研究院古籍全文資料庫的發展概要

A Survey of full-text Data Bases and Related Techniques for Chinese Ancient Documents in Academia Sinica

謝清俊 Ching-chun Hsieh 中央研究院資訊所研究員
林 晰 Shih Lin 中央研究院計算中心分析師

摘 要

中央研究院利用計算機處理古籍已有十二年，其中以全文資料庫的發展最受矚目，目前上線的全文資料庫文總字數已超過一億一仟萬字，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，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，包括國內的中山、中正、師大各大學，國外的倫敦大學、史丹佛大學、密西根大學、香港中文大學等。本文首先介紹各全文資料庫的發展現況，其次介紹自行開發的相關技術，包括：全文資料庫的結構、文章的標誌系統、資料登錄之管理、缺字造字之管理以及目前各單位相關的研究發展計劃等。

Abstract

A survey of full-text data bases and related text processing techniques for Chinese ancient document in the past 12 years in Academia Sinica is presented in this paper. Five Institutes, (namely the Institute of History and phonology, the Institute of Taiwan History, the Institute of Literature and Philosophy, the Institute of Information Science and the Institute of Modern History) and the Computing Center of Academia Sinica actively participated in this long range project since 1984. Beside, the Archival Library of National History also participated in developing the database of Ching Dynasty. Since 1995, some co-laboration projects with other Universities, such as London University in England, Stanford University, Michigan University in USA, Chinese University in Hong Kong and Chung-Cheng University, Chung-San University and National Taiwan Normal University in Taiwan have been launched to produce more digital texts. Now, the total character count of on-line full-text data bases are over 115 millions, and the data bases of more then 80 million characters are coming. In this report, we also survey some important techniques developed, including the structure of full-text database, the ways of handling missing characters, the management of data entry jobs, the development of markup system, etc. Besides, the status of some on going related research projects are summarized in this paper as a future perspective of the development of digital Chinese ancient documents.

中央研究院資訊科學研究所究所文獻處理實驗室 1997年3月

中央研究院古籍全文資料庫的發展概要

壹、前言

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【註一】，毫不留情的痛陳此弊，引起了西方世界極大的關切。持續至今，如何調和人文和科技這兩個領域，仍然是西方國家內政上的重要課題。反觀國內，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爲大。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。

有鑑於此，爲了中華文化的延續，務必要使古籍能活出現代風貌，不可任其在科技的洪流中式微沒頂，而解決的方法，則是將古籍以電子媒體表達。這就是中央研究院（以下簡稱本院）在1984年7月1日開始推動史籍自動化計劃的初衷。

電子媒體有極優越的性質。古籍的電子版本可無限地複製，且幾無複製的成本。於是，一旦電子古籍開發成功，它就是取之不盡、用之不竭的資源，可供全體人民共享。若有四通八達的電腦網路，則電子古籍可以瞬息千里，也幾乎不須花錢。於是，沒有運輸和分配的問題。再者，資料匯集後會產生新的訊息，經相互鉤稽參照，能發前人所未見，對研究工作非常重要。但是，古籍的匯集異常困難，不僅因量多，而且因原本希少，要匯集大量古籍原本加以整理研讀，光靠人力是幾乎不可能的事，而電子版本卻可輕易做到。此外，它無損耗，不需保養，不怕遺失，儲存方便，體積小，再加上可以用計算機做高速的檢索和種種的應用和處理，古籍的電子化實是使古籍活出現代風貌最佳、也是唯一的選擇【註二】。

本院處理古籍的計劃並不限於只使用全文資料庫技術，有許多資料是用關聯式資料庫處理的。諸如，1985年10月開始試做的「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。這些資料庫雖非本文報告的重點，然而在語文處理技術上和全文資料庫是相輔相成的。

本文以下將先對已完成的及正在建構中的全文資料庫作一綜覽。之後，再介紹製作全文資料庫所發展的一些技術，並討論些製作與管理全文資料庫上相關的問題。

【註一】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，更版爲《The Two Cultures and A Second Look》，目前可得之版本爲後者。

【註二】關於電子文件的性質敘述，請參閱謝清俊〈電子佛典的意義〉人生雜誌第147期，p32-39 1995,11.1及〈談資訊的定義與性質〉，資訊科技與社會轉型學術研討會，1996年12月20日。

貳、古籍的全文資料庫

所謂全文資料庫，是以原文件的所有文字爲素材，以儘量保存文件版面的方式所建構的資料庫。在本節中所列的各個資料庫，都是用同一個軟體工具，即「中文全文檢索系統」【註三】所做成的。所以它們具有相同的資料庫性質與功能，也用相同的檢索語言，對使用者而言相當方便。中文全文檢索系統把中文書籍製成全文資料庫, 並提供檢索、閱讀兩個主要功用。它藉著層級式的(hierarchical)目錄來反應書本的章、節、段落等結構, 使用者得以據其調閱正文, 或訂定檢索的範圍等。此系統也保留原書的頁碼和行次, 一則用來調閱正文, 再則用於報告檢索詞的出處, 方便使用者參照書本。正文具備橫、直兩種顯示方向, 編排的格式儘可能呼應原書, 且能夠區別註文、補文、贅文等與原文。表格的呈現與檢索排版雖較爲繁複,目前已經克服。

檢索條件包含一個或多個字詞, 乃至高達數百詞, 亦無不可。每個詞的前後得附加排除字集,以除去不是要找尋的詞彙。參照檢索結果逐次修正排除字集，可以大幅提高檢索的正確率。各詞之間的關係爲「或」、「且」、「且非」三者之一。檢索的範圍小至段落、大至整個資料庫, 無所不宜。也可限定範圍於特定的文素類別，如選擇正文或注釋。檢索的結果能定爲含檢索詞的句子或段落, 亦得在檢索詞的前後截取定長的正文(collocation), 並按檢索詞及其前後文排序。【註四】

目前本院已上線的全文資料庫，如[表一]所示，其中以史語所擁有的資料庫最多，也最龐大。最早完成的是二十五史資料庫(1990年)，是史語所與計算中心共同開發完成的。資料庫內容，在[表一]之註及附表中有較詳之說明。台灣方志資料庫詳目如[表一之六]所示。

古漢語文獻資料庫的目的與其他者稍異，它是用來發展古漢語語料庫(corpus)及內容分析用的，故並未自由對外開放使用。此資料庫目前正在加標誌(markup)中。所謂標誌是把文章中隱晦的內容用明顯的特殊符號標明，其目的是便於用計算機作後繼的處理。目前語料庫所加的標誌有兩類，其一標示文件，分爲篇章標誌、文章標誌和文體標誌三種；其二標示辭類。篇章和文章標誌分別標明書籍的篇章與文章結構。文體標誌則含三種論說文、三種敘述文及四種混合型文體。辭類的標誌則依古漢語語法逐辭標示【註五】，這些標誌配合著全文資料庫使用，但尚未完全定案，還在檢討改進中。古漢語語料庫計劃和資訊所文獻處理實驗室二者所發展的資料庫有研究和實驗的性質,是故其內涵與其他資料庫略有出入。

【註三】在早期發展的核心檢索軟體（ CTP, Chinese Text Processor）之後，陸續改進並增加了些文獻管理的功能，更名爲中文全文檢索系統（FTMS, Full-Text database Management System）

【註四】關於排除字集、文素等名詞及詳細的操作方式，請參照《中文全文檢索統FTMS第三版使用指引》Jan. 1996, 中央研究院算中心

【註五】關於辭類的標誌，請參閱：張俊盛《Linguistic Markup：ROCLING Text Corpus Exchange Format》SIGMT ROCLING, Dec. 1995

開發中的資料庫，如[表二]所示。其中仍以史語所最具規模，臺史所次之。史語所開發中的資料庫及其進度詳如[表二之一]。本院建資料庫目前的程序是先找不同的單位將原始資料各打一遍，驗收後用程式協助比對這兩份資料，此即一校。一校經改正後，進行人工二、三校，再一次改正，再四校並改正後，即完成資料登錄工作。這麼多道手續唯一的目的是確保資料的品質。了解這個工作程序，有助於了解[表二]中所示的各種進度。

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。

表一．中央研究院已完成的中文全文資料庫綜覽

1997/2/14日製表

資料庫	字數	製作單位	聯絡人
漢籍全文資料庫		史語所 漢籍全文資料庫計畫	陳弱水先生
二十五史	39,969,533 [說明 1]		
諸子[表一之一]	5,860,450 [說明 2]		
十三經	8,600,316 [說明 3]		
古籍十八種[表一之二]	8,049,602 [說明 4]		
古籍三十四種[表一之三]	12,264,715 [說明 5]		
大正新脩大藏經[表一之四]	5,054,793 [說明 6]		
古漢語文獻語料庫[表一之五]	13,010,490 [說明 7]	史語所 文獻語料庫研究室 資訊所詞庫小組	黃居仁先生 魏培泉先生
臺灣方志[表一之六]	7,537,840 [說明 8]	臺史所 史籍自動化室	詹素娟小姐
臺灣檔案[表一之七]	7,100,885 [說明 9]	臺史所 史籍自動化室	詹素娟小姐
近現代中國史事日誌	2,085,815 [說明 10]	近史所	胡國台先生
文心雕龍[表一之八]	1,700,011 [說明 11]	資訊所 文獻處理實驗室	謝清俊先生
佛經三論[表一之九]	104,257 [說明 12]	資訊所 文獻處理實驗室	謝清俊先生
新清史-本紀	878,629 [說明 13]	國史館 清史組	朱重聖副館長
簡帛金石資料庫[表一之十]	3,391,582 [說明 14]	史語所 簡牘整理小組	邢義田先生
字數總計:	115,608,918		

[說明 1] 廿十五史自 1984 年開始，首期以全史的食貨志爲主,1985 年 7 月開始做前四史,至 1990 年 6 月,廿五史大致完成，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 月計算中心開始補表格的部份，至 1997 年 1 月表之部份已全部完成。

[說明 2] 諸子(又稱古籍十九種)包含抱朴子內篇校釋、莊子集釋、法言義疏、東觀漢記校注、墨子城守各篇簡注、潛夫論箋校正、國語、莊子集解、莊子集解內篇校正、古本竹書紀年輯證、墨子閒詁、列子集釋、晏子春秋集釋、管子輕重篇新詮、四書章句集注、新語校注、戰國策、八家後漢書輯注、老子校釋。

[說明 3] 含阮刻本《十三經注疏》及斷句十三經經文(台北開明書局)。

[說明 4] 古籍十八種包含唐令拾遺、新校搜神記二十卷、齊民要術校釋、世說新語、典論、申鑒、中論、漢官六種、洛陽伽藍記校注、九家舊晉書輯本、顏氏家訓集解、荊楚歲時記、唐律疏議、山海經校注、通典、風俗通義校注、唐會要、後漢紀校注。

[說明 5] 古籍三十四種由史語所漢籍全文資料庫計畫及古漢語語料庫計畫合作開發,內容包括鄧析子、關尹子、太平經合校、鬼谷子(原文)、尹文子、慎子(原文)、孔子家語、鶡冠子、通玄真經(文子)、孔叢子、藝文類聚、論衡校釋、金匱要略(原文)、難經本義新解(原文)、傷寒論(原文)、黃帝內經(原文)、前漢紀、漢魏南北朝墓誌彙編、九章算經點校(原文)、周髀算經(原文)、越絕書、釋名(原文)、方言校箋(原文)、穆天子傳(原文)、西京雜記、吳越春秋(原文)、逸周書(原文)、文獻通考、朱子語類、楚辭補註、敦煌變文集新書、文選、華陽國志校補圖注、古小說鉤沉。前二十八本書於 1996 年 3 月底上線,加上大正新脩大藏經,合稱爲古籍二十九種。1996 年 12 月底新增後六本書,更名爲古籍三十四種。

[說明 6] 自原古籍廿九種抽出大正藏部份，再加新增資料，於 1996 年 12 月底上線。

[說明 7] 古漢語語料庫包含以下五個語料庫: 上古漢語、中古漢語(含大藏經)、近代漢語、其他、出土文獻。部分資料取自史語所漢籍全文資料庫, 故兩者間略有重疊。此計畫另有英國倫敦大學、中山大學、美國史丹佛大學與香港中文大學參與。各資料庫均已在本院內開放使用。本語料庫之出土文獻語料庫, 全部取自史語所漢簡小組所製作的資料庫。

[說明 8] 臺灣方志版本，大部份選自臺灣銀行經濟研究室所出版的「臺灣文獻叢刊」標點本, 合計四十六種、一百一十六冊。其內容大別爲三類：一爲臺灣通志、府志及各縣、廳志，包括重修、續修之各版本；二爲各地採訪冊、相關地區志書及輿圖；三爲補闕的紀略、資料。

[說明 9] 臺灣檔案版本，大部份選自臺灣銀行經濟研究室所出版的「臺灣文獻叢刊」標點本, 合計四十五種、九十三冊。1996 年 12 月中旬上線。

[說明 10]近、現代中國史事日誌版本選自郭廷以先生所編著之「中國史事日誌」、「中華民國史事日誌」等書。

[說明 11]文心雕龍包括詹　《文心雕龍義證》、張立齋《文心雕龍考異》及范文瀾《文心雕龍注》三本書。

[說明 12] 佛經三論包括《中論》《十二門論》《百論》共七卷

[說明 13] 新清史・本記包括：太祖本紀、太宗本紀、世祖本紀、聖祖本紀、世宗本紀、高宗本紀、仁宗本紀、宣宗本紀、文宗本紀、穆宗本紀、德宗本紀、宣統本紀等

[說明 14]簡帛金石資料庫內容包含中山懷王墓文子釋文、中國古代磚文、天水放馬灘日書甲種、包山二號楚墓、石刻題跋索引(漢-隋)、兩漢鏡銘集錄、居延未發表簡、居延新簡、居延漢簡甲乙篇、居延漢簡釋文合校、奏讞書、秦漢金文錄、秦漢南北朝官印徵存、馬王堆(1)52 病方等、馬王堆(1)老子甲本等、馬王堆(3)春秋事語等、馬王堆帛書：二三子問、馬王堆帛書：刑德、馬王堆帛書：周易繫辭、馬王堆帛書：易之要、馬王堆帛書：昭力、馬王堆帛書：要、馬王堆帛書：繆和、疏勒河流域出土漢簡、敦煌漢簡、敦煌漢簡校文、敦煌漢簡釋文、散見簡牘合輯、曾侯乙墓、雲夢龍崗 6 號秦墓釋文、新出石刻關係資料目錄、墓券、漢代石刻集成、漢印文字徵、漢碑集釋、漢簡書目(-1995.12)、睡虎地秦墓竹簡、銀雀山(1)孫子兵法等、醫簡・引書・脈書、帛書《刑德》乙本釋文校讀。

表二．中央研究院開發中的中文全文資料庫綜覽

1997/1/14日製表

資　料　庫	字　數	製　作　單　位	聯 絡 人
漢籍全文資料庫(續)[表二之一]	54,898,000	史語所 漢籍全文資料庫計畫	陳弱水先生
臺灣文獻叢刊[表二之二]	12,700,000 [說明 1]	臺史所 史籍自動化室	詹素娟小姐
道藏及其他文獻[表二之三]		文哲所	
道藏(部分)	1,891,000 [說明 2]		李豐楙先生
姚際恒著作集	950,000 [說明 3]		林慶彰先生
劉宗周全集	1,100,000		鍾彩鈞先生
泉翁大全集	1,000,000		
古漢語文獻續[表二之四]	650,000	史語所 文獻語料庫研究室	黃居仁先生
		資訊所 詞庫小組	陳克健先生
般若經論	1,500,000 [說明 4]	資訊所 文獻處理實驗室	謝清俊先生
樂府詩集	633,000 [說明 5]	師大國文系	季旭昇先生
預估字數總計:	75,322,000		

[說明 1] 選自臺灣銀行經濟研究室所出版的「臺灣文獻叢刊」第三批, 計四十二種、九十冊, 約七百六十萬字, 正在二校。另有第四批資料, 亦選自臺灣銀行經濟研究室所出版的「臺灣文獻叢刊」, 計五十五種、七十三冊, 約五百一十萬字, 正在繕打中。

[說明 2] 包括無上祕要、真誥、上清道類事相、三洞珠囊卷、雲笈七籤、道樞、道教義樞、上清經派(六朝前)、正統道藏第五十六冊(上清經派--六朝之後)。

[說明 3] 文哲所提供出版用電子檔案, 正由中心標誌之中。

[說明 4] 般若經論含：江味農《金剛經講義》，周止庵《般若波羅蜜多心經詮注》，藕益大師《金剛經破空論》，黃念祖《金剛經一滴》和《心經略說》，潘重規《敦煌壇經新編》及其附冊，印順法師《般若經講記》等

[說明 5]師大國文系負責資訊輸入及層級結構標誌, 計算中心提供轉碼、基本排版單元(ETU)增補及技術諮詢服務。

表一之一【諸子全文資料庫】目錄表

（總字數：5,860,450 字）

號	書名	字數
1	抱朴子(抱朴子內篇校釋,王明,中華)	207,198 字
2	莊子集釋([清]郭慶藩,中華)	669,447 字
3	法言義疏(汪榮寶,中華)	469,287 字
4	東觀漢記(東觀漢記校注,[東漢]劉珍等)	418,917 字
5	墨子城守各篇簡注(岑仲勉,中華)	76,313 字
6	潛夫論箋(潛夫論箋校正,[清]汪繼培,中華)	305,759 字
7	國語([春秋]左丘明,中華)	238,432 字
8	莊子集解([清]王先謙,中華)	195,879 字
9	莊子集解內篇補正([清]劉武,中華)	152,128 字
10	古本竹書紀年輯證(方詩銘 王修齡)	165,503 字
11	墨子閒詁([清]孫詒讓,中華)	499,758 字
12	列子集釋(楊伯峻,中華)	242,808 字
13	晏子春秋集釋(吳則虞,中華)	378,033 字
14	管子輕重篇新詮(馬非白,中華)	457,385 字
15	點校四書章句集注([宋]朱熹,中華)	260,020 字
16	新語校注(王利器,中華)	149,441 字
17	戰國策([西漢]劉向,上海)	533,143 字
18	八家後漢書(八家後漢書輯注,周天游,上海)	254,773 字
19	老子校釋(朱謙之,中華)	186,221 字

表一之二【古籍十八種全文資料庫】目錄表

（總字數：8,049,602 字）

編	書名	編譯	出版	字數
1	《唐令拾遺》	栗勁等譯	長春出版社	542,418 字
2	《新校搜神記》		世界書局	72,123 字
3	《齊民要術校釋》	繆啓愉校釋	農業出版社	626,950 字
4	《世說新語箋疏》	余嘉錫		476,158 字
5	《典論》		世界書局	8,046 字
6	《申鑒》		世界書局	15,410 字
7	《中論》		世界書局	23,428 字
8	《漢官六種》	孫星衍等輯 周天游點校	中華	100,696 字
9	《洛陽伽藍記校注》	范祥雍校注		291,994 字
10	《九家舊晉書輯本》		鼎文本	223,174 字
11	《顏氏家訓集解》	王利器	上海古籍出版社	416,368 字
12	《荊楚歲時記》	守屋美都雄	《中國古歲時記的研究》	15,342 字
13	《唐律疏議》	劉俊文點校	北京中華書局	319,241 字
14	《山海經校注》	袁珂點校	上海古籍，1980	227,090 字
15	《通典》		中華書局點校本	2,878,643 字
16	《風俗通義校注》	王利器	北京中華，1981	416,648 字
17	《唐會要》		世界書局本影 上海中華書局	952,633 字
18	《後漢紀校注》	周天游	天津古籍出版社	443,236 字

表一之三【古籍三十四種全文資料庫】目錄表（總字數：12,264,715字）

號	書名	出處	作者	出版	字數
1	《鄧析子》	新編諸子集成六	[周]鄧析	世界書局	3,802 字
2	《關尹子》	四部備要	據金壺本校刊	台灣中華書局	13,184 字
3	《太平經合校》		王明	北京中華書局 1960	311,493 字
4	《鬼谷子》		陶弘景注	台灣商務(本文)	9,347 字
5	《尹文子》	新編諸子集成六	[清]錢熙祚校	世界	10,574 字
6	《慎子》	新編諸子集成五	[清]錢熙祚校	世界(原文)	5,330 字
7	《孔子家語》	新編諸子集成二	[魏]王肅注	世界	87,026 字
8	《鶡冠子》	四部備要		台灣中華	19,114 字
9	《通玄真經》	(文子)	四部叢刊		47,192 字
10	《孔叢子》	增訂漢魏叢書(三)	[清]王謨輯	大化書局	30,271 字
11	《藝文類聚》			木鐸，標點本	1,138,846 字
12	《論衡校釋》	諸子集成本	黃暉	北京，中華書局	903,836 字
13	《金匱要略》		李克光	知音出版社(本文)	37,338 字
14	《難經本義新解》		林輝鎖	益群(本文)	15,804 字
15	《傷寒論》		李培生	知音(本文)	49,481 字
16	《黃帝內經》		楊維傑	台聯國風出版社(本文)	179,510 字
17	《前漢紀》	[國學基本叢書四百種]	荀悅撰	台灣商務	212,101 字
18	《漢魏南北朝墓誌彙編》		趙超	天津古籍出版社	322,109 字
19	《九章算經點校》		錢寶琮	九章出版社(本文)	29,641 字
20	《周髀算經》	叢書集成		上海商務(原文)	7,535 字
21	《越絕書》		李步嘉武漢	大學出版社	54,110 字
22	《釋名》	增訂漢魏叢書(一)	[清]王謨輯	大化書局(本文)	22,513 字
23	《方言校箋》	方言校箋附通檢	周祖謨	鼎文(本文)	28,396 字
24	《穆天子傳》	四部叢刊(本文)			8,360 字
25	《西京雜記》	增訂漢魏叢書(二)	[清]王謨輯	大化書局	15,460 字
26	《吳越春秋》	四部叢刊(本文)			48,620 字
27	《逸周書》	皇清經解續編(三)		漢京文化事業有限公司(本文)	40,922 字
28	《文獻通考》		馬端臨	商務書局	3,872,972 字
29	朱子語類		[宋]黎靖德編 王星賢點校	世華出版社	2,009,963 字
30	楚辭補注		[宋]洪興祖	天工	161,375 字
31	敦煌變文集新書		潘重規	中國文化大學	403,432 字
32	文選		[梁]蕭統編 [唐]李善注	文津	1,572,575 字
33	華陽國志校補圖注		任乃強	上海古籍出版社(1987)	387,596 字

34	古小說鉤沈		魯迅	盤庚出版社,文史叢刊	204,879字

表一之四【大正新脩大藏經全文資料庫】目錄表（總字數：5,054,793字）

一、【第一卷 阿含部上】(總字數：98,141字)
1. 一三 長阿含十報法經(二卷)
2. 一四 佛說人本欲生經(一卷)
3. 二三 大樓炭經(六卷)
4. 三一 佛說一切流攝守因經(一卷)
5. 三二 佛說四諦經(一卷)
6. 三六 佛說本相猗致經(一卷)
7. 四六 佛說阿那律八念經(一卷)
8. 四八 佛說是法非法經(一卷)
9. 五七 佛說漏分布經(一卷)
10. 九一 佛說婆羅門子命終愛念不離經(一卷)
11. 九二 佛說十支居士八城人經(一卷)
12. 九八 佛說普法義經(一卷)

二、【第二卷 阿含部下】(總字數：23,810字)
1. 一0五 五陰譬喻經(一卷)
2. 一0九 佛說轉法輪經(一卷)
3. 一一二 佛說八正道經(一卷)
4. 一一四 佛說馬有三相經(一卷)
5. 一一五 佛說馬有八態譬人經(一卷)
6. 一三一 佛說婆羅門避死經(一卷)
7. 一三七 舍利弗摩訶目連遊四衢經(一卷)
8. 一四0 阿那邠邸化七子經(一卷)
9. 一四九 佛說阿難同學經(一卷)
10. 一五0A 佛說七處三觀經(一卷)
11. 一五0B 佛說九橫經(一卷)
12. 一五一 佛說阿含正行經(一卷)

三、【第三卷 本緣部上】(總字數：784,560字)
1. 一五二 六度集經(八卷)
2. 一五四 生經(五卷)
3. 一五七 悲華經(十卷)
4. 一六七 佛說太子慕魄經(一卷)
5. 一八四 修行本起經(二卷)
6. 一八六 佛說普曜經(八卷)
7. 一九0 佛本行集經(六十卷)

四、【第四卷 本緣部下】(總字數：603,705字)
1. 一九六 中本起經(二卷)
2. 一九七 佛說興起行經(二卷)
3. 一九八 佛說義足經(二卷)
4. 二0二 賢愚經(十三卷)
5. 二0四 雜譬喻經(一卷)
6. 二0八 經撰雜譬喻(二卷)
7. 二0九 百喻經(四卷)
8. 二一0 法句經(二卷)
9. 二一一 法句譬喻經(四卷)
10. 二一二 出曜經(三十卷)

五、【第八卷 般若部四】(總字數：474,724字)
1. 二二三 摩訶般若波羅蜜經(二十七卷)
2. 二二四 道行般若經(十卷)
3. 二二五 大明度經(六卷)

六、【第十卷 華嚴部下】(總字數：2,633字)
1. 二八0 佛說兜沙經(一卷)

七、【第十一卷 寶積部上】(總字數：20,863字)
1. 三一三 阿●佛國經(二卷)

八、【第十二卷 寶積部下槃部全】(總字數：59,092字)
1. 三二二 法鏡經(一卷)
2. 三四八 佛說大乘方等要慧經(一卷)
3. 三五0 佛說遺日摩尼寶經(一卷)
4. 三五六 佛說寶積三昧文殊師利菩薩問法身經(一卷)
5. 三六一 佛說無量清淨平等覺經(四卷)

九、【第十三卷 大集部全】(總字數：33,470字)
1. 四一七 佛說般舟三昧經(一卷)
2. 四一八 般舟三昧經(三卷)

一0、【第十四卷 經集部一】(總字數：37,733字)
1. 四五八 文殊師利問菩薩署經(一卷)
2. 四九二 佛說阿難問事佛吉凶經(一卷)
阿難問事佛吉凶經(一卷・別本)
3. 五0六 犍陀國王經(一卷)
4. 五二五 佛說長者子懊惱三處經(一卷)
5. 五二六 佛說長者子制經(一卷)
6. 五五一 佛說摩鄧女經(一卷)
7. 五五三 佛說 女祇域因緣經(一卷)
8. 五五四 佛說柰女耆婆經(一卷)

一一、【第十五卷 經集部二】(總字數：123,434字)
1. 六0二 佛說大安般守意經(二卷)
2. 六0三 陰持入經(二卷)
六0四 佛說禪行三十七品經(一卷)
3. 六0五 禪行法想經(一卷)
4. 六0七 道地經(一卷)
5. 六0八 小道地經(一卷)
6. 六二一 佛說佛印三昧經(一卷)
7. 六二二 佛說自誓三昧經(一卷)
8. 六二四 佛說佗真陀羅所問如來三昧經(三卷)
9. 六二六 佛說阿闍世王經(二卷)
10. 六三0 佛說成具光明定意經(一卷)

一二、【第十六卷 經集部三】(總字數：2,054字
1. 六八四 佛說父母恩難報經(一卷)
2. 七0一 佛說溫室洗浴 僧經(一卷)

一三、【第十七卷 經集部四】(總字數：42,983字)
1. 七二四 佛說罪業應報教化地獄經(一卷)
2. 七二九 佛說分別善惡所起經(一卷)
3. 七三0 佛說處處經(一卷)
4. 七三一 佛說十八泥犁經(一卷)
5. 七三二 佛說罵意經(一卷)
6. 七三三 佛說堅意經(一卷)
7. 七三五 佛說四願經(一卷)
8. 七七八 佛說菩薩內習六波羅蜜經(一卷)
9. 七七九 佛說八大人覺經(一卷)
10. 七八四 四十二章經(一卷)
11. 七九一 佛說出家緣經(一卷)
12. 七九二 佛說法受塵經(一卷)
13. 八0七 佛說內藏百寶經(一卷)

一四、【第二十二卷 律部一】(總字數：714,955字)
1. 一四二八 四分律(六十卷)

一五、【第二十三卷 律部二】(總字數：789,109字)
1. 一四三五 十誦律(六十一卷)

一六、【第二十四卷 律部三】(總字數：25,040字)
1. 一四六七 佛說犯戒罪報輕重經(一卷)
佛說犯戒罪報輕重經(一卷・別本)
2. 一四七0 大比丘三千威儀(二卷)
3. 一四九二 佛說舍利弗悔過經(一卷)

一七、【第二十五卷 釋經論部上】(總字數：3,918字)
1. 一五0八 阿含口解十二因緣經(一卷)

一八、【第二十八卷 毘曇部三】(總字數：5,160字)
1. 一五五七 阿毘曇五法行經(一卷)

一九、【第四十九卷 史傳部一】(總字數：4,113字)
1. 二0二七 迦葉結經(一卷)

二0、【第五十卷 史傳部二】(總字數：617,966字)
1. 二0五九 高僧傳(十四卷)
2. 二0六0 續高僧傳(三十卷)

二一、【第五十二卷 史傳部四】(總字數：587,327字)
1. 二一0二 弘明集(十四卷)
2. 二一0三 廣弘明集(三十卷)

表一之五【古漢語語料庫/】目錄表（總字數：13,010,490字）

（說明1：加∨者爲與史語所《漢籍全文資料庫》計畫合製部份；
加＊者爲本計畫自行建製之文獻；加○者爲與中山大學合製部份；
加☆者爲與中正大學交換部份；加◇者爲與史丹佛大學合製部份；
未加上述任何標記者爲史語所提供文件原文之數位檔案，再由本計畫製成語料庫。）
（說明2：加※者爲已對院內公開的資料。）
（說明3：部份與表一之一及一之三重覆。）

一、【上古漢語語料庫】（總字數：4,966,693 字）

1.※ 尚書<古漢語壹>
2.※ 毛詩<其他>
3.※ 周易<古漢語壹>
4.※ 儀禮<古漢語壹>
5.※ 周禮<其他>
6.※ 禮記<古漢語參>
7.※ 春秋公羊傳<古漢語參>
8.※ 春秋穀梁傳<古漢語參>
9.※ 春秋左傳<古漢語壹>
10.※ 國語<古漢語壹>
11.※ 戰國策<古漢語壹>
12.※ 論語<古漢語壹>(又見【論孟老莊全文資料庫】)
13.※ 孟子<古漢語壹>(又見【論孟老莊全文資料庫】)
14.※ 墨子<古漢語壹>
15.※ 莊子<古漢語壹>(又見【論孟老莊全文資料庫】)
＊16.※ 荀子<古漢語壹>
＊17.※ 韓非子<古漢語壹>
＊18.※ 呂氏春秋<古漢語壹>
19.※ 老子（又見【論孟老莊全文資料庫】）
＊20.※ 商君書<古漢語壹>
＊21.※ 管子<古漢語壹>
22.※ 晏子春秋<古漢語參>
＊23.※ 孫子<古漢語參>
＊24.※ 大戴禮記<古漢語參>
＊25.※ 韓詩外傳<古漢語參>
＊26.※ 吳子<古漢語參>
＊27.※ 尉繚子<古漢語參>
＊28.※ 六韜<古漢語參>
＊29.※ 司馬法<古漢語參>
∨30.※ 愼子
∨31.※ 文子
∨32.※ 關尹子
∨33.※ 鶡冠子
＊34.※ 公孫龍子<古漢語參>
∨35.※ 鄧析子
∨36.※ 尹文子
∨37.※ 鬼谷子
＊38.※ 燕丹子
39.※ 列子集釋
40.※ 孝經<古漢語參>
41.※ 爾雅<其他>
∨42.※ 周髀算經
∨43.※ 九章算經
∨44.※ 黃帝內經素問
∨45.※ 黃帝內經靈樞
∨46.※ 難經
47.※ 古本竹書紀年輯證
∨48.※ 逸周書
∨49.※ 穆天子傳
∨50.※ 孔子家語
∨51.※ 孔叢子
∨52.※ 吳越春秋
∨53.※ 越絕書
54.※ 史記<其他>
55.※ 漢書<古漢語貳>
＊56.※ 新書<西漢基準文獻>
57.※ 新語<古漢語貳>
＊59.※ 淮南子<古漢語貳>
＊60.※ 新序<古漢語貳>
＊61.※ 說苑<古漢語貳>
＊62.※ 列女傳<古漢語貳>
＊63.※ 鹽鐵論<古漢語貳>
64.※ 法言<古漢語貳>
∨65.※ 西京雜記
66.※ 前漢紀

二之一、【中古漢語語料庫之一】（總字數：231,479 字）

∨1.※ 風俗通義
∨2.※ 釋名
∨3.※ 傷寒論
∨4.※ 金匱要略（又見【先秦西漢語料庫】）
∨5.※ 世說新語

二之二、【中古漢語語料庫之二】（總字數：3,709,731 字）

∨1. [784](東漢)(迦葉摩騰共法蘭)四十二章經
∨2.※ [13](東漢)(安世高)長阿含十報法經
∨3.※ [14](東漢)(安世高)佛說人本欲生經
∨4.※ [31](東漢)(安世高)佛說一切流攝守因經
∨5.※ [32](東漢)(安世高)佛說四諦經
∨6.※ [36](東漢)(安世高)佛說本相猗致經
∨7.※ [48](東漢)(安世高)佛說是法非法經
∨8.※ [57](東漢)(安世高)佛說漏分布經
∨9.※ [91](東漢)(安世高)佛說婆羅門子命終愛念不離經
∨10.※ [92](東漢)(安世高)佛說十支居士八城人經
∨11.※ [98](東漢)(安世高)佛說普法義經
∨12.※ [105](東漢)(安世高)五陰譬喻經
∨13.※ [109](東漢)(安世高)佛說轉法輪經
∨14.※ [112](東漢)(安世高)佛說八正道經
∨15.※ [131](東漢)(安世高)佛說婆羅門避死經
∨16.※ [140](東漢)(安世高)阿那邠邸化七子經
∨17.※ [149](東漢)(安世高)佛說阿難同學經
∨18.※ [150A](東漢)(安世高)佛說七處三觀經
∨19.※ [151](東漢)(安世高)佛說阿含正行經
∨20.※ [167](東漢)(安世高)佛說太子慕魄經
∨21. [348](東漢)(安世高)佛說大乘方等要慧經
∨22. [356](東漢)(安世高)佛說寶積三昧文殊師利菩薩問法身經
∨23. [492](東漢)(安世高)佛說阿難問事佛吉凶經
∨24. [506](東漢)(安世高)犍陀國王經
∨25. [525](東漢)(安世高)佛說長者子懊惱三處經
∨26. [526](東漢)(安世高)佛說長者子制經
∨27. [551](東漢)(安世高)佛說摩鄧女經
∨28. [553](東漢)(安世高)佛說　女祇域因緣經
∨29. [554](東漢)(安世高)佛說柰女耆婆經
∨30. [602](東漢)(安世高)佛說大安般守意經
∨31. [603](東漢)(安世高)除持入經
∨32. [604](東漢)(安世高)佛說禪行三十七品經
∨33. [605](東漢)(安世高)禪行法想經
∨34. [607](東漢)(安世高)道地經
∨35. [621](東漢)(安世高)佛說佛印三昧經
∨36. [622](東漢)(安世高)佛說自誓三昧經
∨37. [684](東漢)(安世高)佛說父母恩難報經
∨38. [701](東漢)(安世高)佛說溫室洗浴　僧經
∨39. [724](東漢)(安世高)佛說罪業應報教化地獄經
∨40. [729](東漢)(安世高)佛說分別善惡所起
∨41. [730](東漢)(安世高)佛說處處經
∨42. [731](東漢)(安世高)佛說十八泥犁經
∨43. [732](東漢)(安世高)佛說罵意經
∨44. [733](東漢)(安世高)佛說堅意經

∨45. [779](東漢)(安世高)佛說八大人覺經
∨46. [791](東漢)(安世高)佛說出家緣經
∨47. [792](東漢)(安世高)佛說法受塵經
∨48. [1467](東漢)(安世高)佛說犯戒罪報輕重經
∨49. [1470](東漢)(安世高)大比丘三千威儀
∨50. [1492](東漢)(安世高)佛說舍利弗悔過經
∨51. [1557](東漢)(安世高)阿毘曇五法行經
∨52. [2027](東漢)(安世高)迦葉結經
∨53. [204](東漢)(支婁迦讖)雜譬喻經
∨54. [224](東漢)(支婁迦讖)道行般若經
∨55. [280](東漢)(支婁迦讖)佛說兜沙經
∨56. [313](東漢)(支婁迦讖)阿●佛國經
∨57. [350](東漢)(支婁迦讖)佛說遺日摩尼寶經
∨58. [361](東漢)(支婁迦讖)佛說無量清淨平等覺經
∨59. [417](東漢)(支婁迦讖)佛說般舟三昧經
∨60. [418](東漢)(支婁迦讖)般舟三昧經
∨61. [458](東漢)(支婁迦讖)文殊師利問菩薩署經
∨62. [624](東漢)(支婁迦讖)佛說佗真陀羅所問如來三昧經
∨63. [626](東漢)(支婁迦讖)佛說阿闍世王經"
∨64. [807](東漢)(支婁迦讖)佛說內藏百寶經
∨65. [778](東漢)(嚴佛調)佛說菩薩內習六波羅蜜經
∨66. [322](東漢)(安玄共嚴佛調)法鏡經
∨67. [1508](東漢)(安玄共嚴佛調)阿含口解十二因緣經
∨68.※ [46](東漢)(支曜)佛說阿那律八念經
∨69.※ [114](東漢)(支曜)佛說馬有三相經
∨70.※ [115](東漢)(支曜)佛說馬有八態譬人經
∨71. [608](東漢)(支曜)小道地經
∨72. [630](東漢)(支曜)佛說成具光明定意經
∨73.※ [137](東漢)(康孟詳)舍利弗摩訶目連遊四衢經
∨74.※ [197](東漢)(康孟詳)佛說興起行經
∨75.※ [196](東漢)(曇果共康孟詳)中本起經
∨76.※ [184](東漢)(康孟詳共竺大力)修行本起經
○77.※ [153](吳)(支謙)菩薩本緣經
○78.※ [185](吳)(支謙)佛說太子瑞應本起經
∨79. [198](吳)(支謙)佛說義足經
○80.※ [200](吳)(支謙)撰集百緣經
∨81. [225](吳)(支謙)大明度經
○82.※ [362](吳)(支謙)佛說阿彌陀三耶三佛薩樓佛檀過度人道經
○83.※ [632](吳)(支謙)佛說慧印三昧經
∨84. [735](吳)(支謙)佛說四願經
○85.※ [790](吳)(支謙)佛說孛經抄
∨86. [210](吳)(維祇難)法句經
∨87.※ [152](吳)(康僧會)六度集經
∨88.※ [154](西晉)(竺法護)生經
○89.※ [170](西晉)(竺法護)佛說德光太子經
∨90.※ [186](西晉)(竺法護)佛說普曜經
○91.※ [222](西晉)(竺法護)光讚經
○92.※ [263](西晉)(竺法護)正法華經
○93.※ [266](西晉)(竺法護)佛說阿惟越致遮經
○94.※ [285](西晉)(竺法護)漸備一切智德經
○95.※ [398](西晉)(竺法護)大哀經
○96.※ [221](西晉)(無羅叉)放光般若經
∨97. [211](西晉)(法矩共法立)法句譬喻經
○98.※ [638](西晉)(聶承遠)佛說超日明三昧經
○99.※ [2042](西晉)(安法欽)阿育王傳
○100. [26](東晉)(僧伽提婆)中阿含經
∨101.※ [212](姚秦)(竺佛念)出曜經
○102.※ [201](後秦)(鳩摩羅什)大莊嚴論經
∨103. [208](後秦)(鳩摩羅什)眾經撰雜譬喻
∨104.※ [157](北涼)(曇無讖)悲華經
∨105.※ [202](元魏)(慧覺)賢愚經
∨106.※ [209](蕭齊)(求那毘地)百喻經
∨107.※ [190](隋)(闍那崛多)佛本行集經
☆108. [47](西晉)(竺法護)佛說離睡經
☆109. [103](西晉)(竺法護)聖法印經
☆110. [118](西晉)(竺法護)佛說鴦掘摩經
☆111. [135](西晉)(竺法護)佛說力士移山經
☆112. [168](西晉)(竺法護)佛說太子墓魄經
☆113. [180](西晉)(竺法護)佛說過去世佛分衛經
☆114. [182](西晉)(竺法護)佛說鹿母經
☆115. [199](西晉)(竺法護)佛五百弟子自說本起經
☆116. [274](西晉)(竺法護)佛說濟諸方等學經
☆117. [283](西晉)(竺法護)菩薩十住行道品一
☆118. [288](西晉)(竺法護)等目菩薩所問三昧經
☆119. [315](西晉)(竺法護)佛說普門品經
☆120. [317](西晉)(竺法護)佛說胞胎經
☆121. [318](西晉)(竺法護)文殊師利佛土嚴淨經
☆122. [334](西晉)(竺法護)佛說須摩提菩薩經
☆123. [345](西晉)(竺法護)慧上菩薩問大善權經
☆124. [378](西晉)(竺法護)佛說方等般泥洹經
☆125. [428](西晉)(竺法護)佛說八陽神咒經
☆126. [481](西晉)(竺法護)持人菩薩經
☆127. [496](西晉)(竺法護)佛說大迦葉本經
☆128. [589](西晉)(竺法護)佛說魔逆經
☆129. [627](西晉)(竺法護)普超三昧經
☆130. [636](西晉)(竺法護)無極寶三昧經
☆131. [685](西晉)(竺法護)佛說盂蘭盆經
☆132. [501](西晉)(法炬)佛說沙曷比丘功德經
☆133. [739](西晉)(法炬)佛說慢法經
☆134. [502](西晉)(法炬)佛爲少年比丘說正事經
☆135. [503](西晉)(法炬)比丘避女惡名欲自殺經
☆136. [508](西晉)(法炬)阿闍世王問五逆經
☆137. [537](西晉)(聶承遠)佛說越難經

三、【近代漢語語料庫】（總字數：1,992,985 字）
◇1.※ 老乞大諺解
◇2.※ 朴通事諺解
◇3.※ 訓世評話
○4.※ 醒世姻緣
○5.※ 六祖壇經
○6.※ 神會語錄
○7.※ 入唐求法巡禮行記
○8.※ 遊仙窟
＊9.※ 王梵志詩
○10. 鏡花緣
○11. 元刊雜劇三十種
○12. 新刊大宋宣和遺事
○13. 大唐三藏取經詩話
○14. 桃花扇
○15. 關漢卿戲曲集
○16. 五代史平話
○17. 永樂大典戲文

四、【其他語料庫】（總字數：827,203 字）
○1.※ 日知錄
○2.※ 亭林詩集
○3.※ 亭林文集
○4.※ 熹廟諒陰記事
○5.※ 四存編

五、【出土文獻語料庫】（總字數：1,282,399 字）
此部份資料全部取自史語所「簡牘整理小組」所整理之
【簡帛金石資料庫】
1. 馬王堆漢墓帛書(壹)
2. 馬王堆漢墓帛書(三)
3. 馬王堆漢墓帛書(肆)
4. 睡虎地秦墓竹簡
5. 銀雀山漢墓竹簡(壹)
6. 居延漢簡甲乙編上
7. 居延漢簡甲乙編下
8. 居延新簡上
9. 居延新簡下
10. 敦煌漢簡釋文
11. 敦煌漢簡

12. 敦煌漢簡校文
13. 武威漢代醫簡
14. 張家山漢簡引書
15. 脈書
16. 疏勒河流域出土漢簡
17. 甲種《日書》
18. 散見簡牘合集
19. 漢碑集釋
20. 秦漢金文錄
21. 墓券
22. 秦漢南北朝官印徵存

表一之六【臺灣方志全文資料庫】目錄表（總字數：7,537,840字）

臺灣文獻叢刊第一批　凡四十六種共一百一十六冊

一、 通志、府志、縣志、廳志

代碼	叢刊號	書名	作/編者	冊數
A	68	清一統志台灣府		1
B	84	福建通志台灣府		6
C	30	台灣通志		4
D	65	台灣府志	高拱乾	3
E	66	重修台灣府志	周元文	3
F	74	重修福建通志台灣府	劉良璧	4
G	105	重修台灣府志	范咸	5
H	121	續修台灣府志	余文儀	6
I	75	恆春縣志	屠繼善	2
J	103	台灣縣志	陳文達	2
K	113	重修台灣縣志	王必昌	4
L	140	續修台灣縣志	謝金鑾	4
M	124	鳳山縣志	陳文達	2
N	146	重修鳳山縣志	王瑛曾	3
O	141	諸羅縣志	周鍾瑄	2
P	156	彰化縣志	周璽	3
Q	159	苗栗縣志	沈茂蔭	2
R	160	噶瑪蘭廳志	陳淑均	4
S	164	澎湖廳志	林豪	3
T	172	淡水廳志	陳培桂	3

二、采訪錄、一般志書與輿圖

代碼	叢刊號	書名	作/編者	冊數
U	37	雲林縣采訪冊	倪贊元	2
V	55	台灣縣采訪冊	諸家	2
W	58	嘉義管內采訪冊		1
X	73	鳳山縣采訪冊	盧德嘉	3
Y	81	台東州采訪冊	胡傳	1
Z	145	新竹縣采訪冊	陳朝龍	2
AA	48	苑裏志	蔡振豐	1
AB	63	樹杞林志	諸家	1
AC	80	金門志	林焜熿	3
AD	95	廈門志	周凱	5
AE	61	新竹縣志初稿	諸家	2
AF	101	新竹縣制度考		1
AG	92	噶瑪蘭志略	柯培元	1
AH	181	台灣府輿圖纂要		3
AI	185	台灣地輿全圖		1
AJ	195	福建通志列傳選	陳衍	3
AK	233	泉州府志選錄		1
AL	232	漳州府志選錄		1

三、補闕

代碼	叢刊號	書名	作/編者	冊數
AM	104	澎湖台灣紀略	諸家	1
AN	109	澎湖紀略	胡建偉	2
AO	115	澎湖續編	蔣鏞	1
AP	52	安平縣雜紀		1
AQ	120	台灣通紀	陳衍	2
AR	243	清史稿 台灣資料集輯		6
AS	148	台灣志略	李元春	1
AT		台灣府志	蔣毓英	2

表一之七【臺灣檔案全文資料庫】目錄表（總字數：7,100,885 字）

號碼	叢刊號	書名	冊數/頁數
1	027	劉壯肅公奏議	3 冊/ 449 頁
2	029	福建臺灣奏摺	1 冊/ 93 頁
3	031	臺案彙錄甲集	3 冊/ 251 頁
4	038	同治甲戌日兵侵臺始末	2 冊/ 297 頁
5	049	東溟奏稿	1 冊/ 180 頁
6	062	楊勇　公奏議	1 冊/ 69 頁
7	088	左文襄公奏牘	1 冊/ 142 頁
8	110	臺灣海防檔	2 冊/ 203 頁
9	158	清世祖實錄選輯	1 冊/ 188 頁
10	165	清聖祖實錄選輯	1 冊/ 180 頁
11	167	清世宗實錄選輯	1 冊/ 52 頁
12	173	臺案彙錄乙集	4 冊/ 574 頁
13	176	臺案彙錄丙集	2 冊/ 344 頁
14	178	臺案彙錄丁集	2 冊/ 320 頁
15	179	臺案彙錄戊集	3 冊/ 392 頁
16	186	清高宗實錄選輯	4 冊/ 736 頁
17	187	清仁宗實錄選輯	1 冊/ 194 頁
18	188	清宣宗實錄選輯	3 冊/ 520 頁
19	189	清文宗實錄選輯	1 冊/ 68 頁
20	190	清穆宗實錄選輯	1 冊/ 172 頁
21	191	臺案彙錄己集	3 冊/ 410 頁
22	192	法軍侵臺檔	4 冊/ 568 頁
23	193	清德宗實錄選輯	2 冊/ 306 頁
24	200	臺案彙錄庚集	5 冊/ 842 頁
25	203	籌辦夷務始末選輯	3 冊/ 422 頁
26	204	法軍侵臺檔補編	1 冊/ 126 頁
27	205	臺案彙錄辛集	2 冊/ 312 頁
28	210	清光緒朝中日交涉史料選輯	3 冊/ 440 頁
29	226	清會典臺灣事例	2 冊/ 218 頁
30	227	臺案彙錄壬集	1 冊/ 114 頁
31	228	臺案彙錄癸集	1 冊/ 142 頁
32	231	吳光祿使閩奏稿選錄	1 冊/ 70 頁
33	236	籌辦夷務始末選輯補編	1 冊/ 60 頁
34	247	清季申報臺灣紀事輯錄	8 冊/1126 頁
35	253	述報法兵侵臺紀事殘輯	2 冊/ 468 頁
36	256	清奏疏選彙	1 冊/ 94 頁
37	262	東華錄選輯	2 冊/ 322 頁
38	273	東華續錄選輯	2 冊/ 340 頁
39	276	劉銘傳撫臺前後檔案	2 冊/ 272 頁
40	277	光緒朝東華續錄選輯	2 冊/ 238 頁
41	278	清季臺灣洋務史料	1 冊/ 98 頁
42	285	李文襄公奏疏與文移	3 冊/ 524 頁
43	288	道咸同光四朝奏議選輯	3 冊/ 408 頁
44	290	臺灣對外關係史料	1 冊/ 104 頁
45	300	雍正硃批奏摺選輯	2 冊/ 262 頁

[說明]選自臺灣銀行經濟研究室所出版之「臺灣文獻叢刊」第二批，凡四十五種、共九十三冊。

表一之八【文心雕龍全文資料庫】(字數： 1,700,011 字)

號	內　　容	作　者	字　　數
1	文心雕龍義證	詹	1,141,622 字
2	文心雕龍考異	張立齋	71,333 字
3	文心雕龍注	范文瀾	487,055 字

表一之九【佛經三論全文資料庫】(字數： 104,257 字)

號	內　　容	字　　數
1	中論	64,143 字
2	十二門論	14,768 字
3	百論	25,345 字

表一之十【簡帛金石資料庫】字數:3,391,582字

編號	內　　容	字　　數
1	中山懷王墓文子釋文	6,205 字
2	中國古代磚文	19,929 字
3	天水放馬灘日書甲種	3,207 字
4	包山二號楚墓	19,966 字
5	石刻題跋索引(漢-隋)	362,984 字
6	兩漢鏡銘集錄	115,197 字
7	居延未發表簡	108,675 字
8	居延新簡	246,421 字
9	居延漢簡甲乙篇	306,376 字
10	居延漢簡釋文合校	246,078 字
11	秦讞書	9,758 字
12	秦漢金文錄	57,946 字
13	秦漢南北朝官印徵存	111,961 字
14	馬王堆(1)52 病方等	7,441 字
15	馬王堆(1)老子甲本等	67,872 字
16	馬王堆(3)春秋事語等	23,530 字
17	馬王堆帛書：二三子問	3,857 字
18	馬王堆帛書：刑德	2,319 字
19	馬王堆帛書：周易繫辭	5,620 字
20	馬王堆帛書：易之要	4,474 字

編號	內　　容	字　　數
21	馬王堆帛書：昭力	1,478 字
22	馬王堆帛書：要	1,677 字
23	馬王堆帛書：繆和	7,215 字
24	疏勒河流域出土漢簡	56,699 字
25	敦煌漢簡	154,030 字
26	敦煌漢簡校文	386,745 字
27	敦煌漢簡釋文	104,524 字
28	散見簡牘合輯	60,006 字
29	曾侯乙墓	11,046 字
30	雲夢龍崗 6 號秦墓釋文	4,562 字
31	新出石刻關係資料目錄	218,003 字
32	墓券	34,673 字
33	漢代石刻集成	51,769 字
34	漢印文字徵	120,204 字
35	漢碑集釋	51,340 字
36	漢簡書目(-1995.12)	242,251 字
37	睡虎地秦墓竹簡	87,069 字
38	銀雀山(1)孫子兵法等	47,024 字
39	醫簡・引書・脈書	16,507 字
40	帛書《刑德》乙本釋文校讀	4,944 字

表二之一 史語所進行中的資料庫 (總字數：約54,898,000字)

一、大正藏: 約 4,349,000 字

1985-2006,2012.禪宗語錄等二十三種

[宋]贊寧等: 2061.宋高僧傳

[梁]寶唱: 2063.比丘尼傳

[宋]道原: 2076.景德傳燈錄

[唐]道宣: 2104.集古今佛道論衡

[唐]智昇: 2105.續集古今佛道論衡

[梁]寶唱等: 2121.經律異相

[唐]道世: 2122.法苑珠林

[梁]僧祐: 2145.出三藏記集

二、其他典籍：約 42,389,000 字

1.水經注	陳橋驛校熊本	2,275,000 字
2.太平御覽	商務，四部叢刊三編本	4,365,000 字
3.全上古三代秦漢三國六朝文	[清]嚴可均校輯，中華書局	5,390,000 字
4.續資治通鑑長編(2-34 冊)	[宋]李燾，北京中華書局	7,110,000 字
5.明實錄(1-133 冊)	中央研究院歷史語言研究所校勘	12,320,000 字
6.大正藏經		1,500,000 字
歷代三寶紀	No.2034 [隋]費長房	
佛祖統紀	No.2034 [宋]志磐	
佛祖歷代通載	No.2036 [元]念常	
大唐西域求法高僧傳	No.2066 [唐]義淨	
大唐西域記	No.2087 [唐]玄奘、辯機	

集神州三寶感通錄	No.2106 [唐]道宣	
7.本草綱目	[明]李時珍，北京人民衛生出版社，1975	2,180,000字
8.水滸全傳	[元]施耐庵、羅貫中，王利器校訂，貫雅文化	1,010,000字
9.繡像金瓶梅詞話	雪山圖書有限公司(故宮藏萬曆丁巳本)	900,000字
10.西遊記	[明]吳承恩，桂冠圖書公司(世德堂本)	865,000字
11.紅樓夢校注	[清]曹雪芹、高鶚,里仁書局(庚辰本、程甲本合配本)	950,000字
12.兒女英雄傳	桂冠圖書公司(北京聚珍活字本)	505,000字
13.儒林外史	[清]吳敬梓，桂冠圖書公司(臥賢草堂本)	320,000字
14.唐語林校證	[宋]王讜撰，周勛初校證，北京中華書局，1987	560,000字
15.四朝聞見錄	[宋]葉紹翁撰，沈錫麟、馮惠民點校,北京中華書局，1989	138,000字
16.大唐新語	[唐]劉肅撰，許德楠、李鼎霞點校,北京中華書局，1984	120,000字
17.歸田錄	[宋]歐陽修撰，李偉國點校，北京中華書局，1981	39,000字
18.泊宅編	[宋]方勺撰，許沛藻、楊立揚點校 北京中華書局，1983	74,000字
19.清波雜志校注	[宋]周輝撰，劉永翔校注，北京中華書局，1994	370,000字
20.湘山野錄，續湘山野錄	[宋]文瑩撰，鄭世剛、楊立揚點校 北京中華書局，1984	63,000字
21.玉壺清話	[宋]文瑩撰，鄭世剛、楊立揚點校 北京中華書局，1984	74,000字
22.舊聞證誤	[宋]李心傳撰，崔文印點校，北京中華書局，1981	46,000字
23.遊宦紀聞	[宋]張世南撰，張茂鵬點校，北京中華書局，1981	58,000字
24.鐵圍山叢談	[宋]蔡絛撰，馮惠民、沈錫麟點校 北京中華書局，1983	76,000字
25.青箱雜記	[宋]吳處厚撰，李裕民點校，北京中華書局,1985	90,000字
26.鶴林玉露	[宋]羅大經撰，王瑞來點校，北京中華書局,1983	235,000字
27.齊東野語	[宋]周密撰，張茂鵬點校，北京中華書局，1983	232,000字
28.龍川別志	[宋]蘇轍撰，俞宗憲點校，北京中華書局，1982	27,000字
29.龍川略志	[宋]蘇轍撰，俞宗憲點校，北京中華書局，1982	46,000字
30.涑水記聞	[宋]司馬光撰，鄧廣銘、張希清點校 北京中華書局，1989	277,000字
31.澠水燕談錄	[宋]王闢之撰，呂友仁點校，北京中華書局，1981	80,000字
32. 肋編	[宋]莊綽撰，蕭魯陽點校，北京中華書局，1983	94,000字
三．準備輸入：8,160,000 字		
明實錄附錄(3-21冊)		2,220,000字
明實錄校勘記(29冊)		5,940,000字

表二之二　臺史進行中的資料庫（總字數：約12,700,000字）

一、臺灣文獻叢刊第三批(台灣銀行出版，共計四十二種，九十冊)：(二校：7,600,000字)

叢刊號	書　　名	作　者	冊數,頁數	叢刊號	書　　名	作　者	冊數,頁數
2	東瀛識略	丁紹儀	1冊,104頁	83	中復堂選集	姚瑩	2冊,262頁
3	小琉球漫誌	朱仕玠	1冊,102頁	87	斯未信齋文編	徐宗幹	1冊,182頁
4	臺海使槎錄	黃叔璥	1冊,177頁	89	臺灣遊記	諸家	1冊,96頁
6	臺游日記	蔣師轍	1冊,141頁	90	番社采風圖考	六十七	1冊,104頁
7	東槎紀略	姚瑩	1冊,126頁	91	臺灣私法商事編		2冊,332頁
8	東瀛紀事	林豪	1冊,69頁	116	陳清端公文選	陳璸	1冊,54頁
9	蠡測彙鈔	鄧傳安	1冊,64頁	117	臺灣私法人事編		5冊,852頁
12	東征集	藍鼎元	1冊,107頁	125	欽定福建省外海戰船則例		3冊,364頁
14	平臺紀略	藍鼎元	1冊,72頁	128	臺灣通史	連橫	6冊,1064頁
17	治臺必告錄	丁曰健	4冊,598頁	139	臺灣府賦役冊		1冊,84頁
19	海東札記	朱景英	1冊,63頁	150	臺灣私法物權編		9冊,1712頁
20	臺陽筆記	翟灝	1冊,39頁	151	臺灣中部碑文集成		1冊,176頁
21	巡臺退思錄	劉璈	3冊,286頁	152	清代臺灣大租調查書		6冊,1116頁
36	臺灣紀事	吳子光	1冊,117頁	180	清職貢圖選		1冊,60頁
44	裨海紀遊	郁永河	1冊,72頁	184	臺灣土地制度考查報告書		1冊,92頁
45	臺灣與圖	夏獻綸	1冊,82頁	196	琉求與雞籠山	諸家	1冊,108頁
46	臺灣番事產與商務		1冊,122頁	197	淡新鳳三縣簡明總括圖冊		1冊,152頁
51	臺灣生熟番紀事	黃逢昶	1冊,55頁	199	福建省例		8冊,1222頁
71	臺灣日記與稟啓	胡傳	2冊,282頁	216	臺灣輿地彙鈔諸家		1冊,142頁
78	清代臺灣職官印錄		1冊,163頁	218	臺灣南部碑文集成		6冊,784頁
79	臺灣司法債權編		2冊,250頁	295	淡新檔案選錄行政編初集		4冊,594頁

二、臺灣文獻叢刊第四批(台灣銀行出版，共計五十五種，七十三冊)：

(輸入中：5,100,000字)

代碼	叢刊號	書名	冊數	頁數	作者	出版年
4A	001	臺灣割據志	1冊	87	川口長孺	46
4B	005	臺灣鄭氏紀事	1冊	78	川口長孺	47
4C	010	赤嵌集	1冊	83	孫元衡	47
4D	011	閩海紀要	1冊	78	夏琳	47
4E	013	靖海紀事	1冊	101	施琅	47
4F	015	臺灣鄭氏始末	1冊	87	沈雲	47
4G	016	平臺紀事本末	1冊	74		47
4H	022	海紀輯要	1冊	78	夏琳	47
4I	023	閩海紀略	1冊	66		47
4J	024	海上見聞錄	1冊	63	阮旻錫	47
4K	025	賜姓始末	1冊	98	黃宗羲	47
4L	026	海國聞見錄	1冊	81	陳倫炯	47
4M	028	臺灣雜詠合刻	1冊	78	諸家	47
4N	030	臺陽見聞錄	2冊	200	唐贊袞	47
4O	032	從征實錄	1冊	194	楊英	47
4P	034	臺陽詩話	1冊	92	王松	48
4Q	039	甲戌公牘鈔存	1冊	161	王元	48
4R	040	臺海思慟錄	1冊	65	思痛子	48
4S	041	北郭園詩鈔	1冊	92	鄭用錫	48
4T	042	海南雜著	1冊	62	蔡廷蘭	48
4U	043	馬關議和中之伊李問答	1冊	87		48
4V	047	戴施兩案紀略	1冊	116	吳德功	48
4W	050	滄海遺民賸稿	1冊	70	王松	48
4X	053	臺戰演義	1冊	52		48
4Y	054	臺灣教育碑記	1冊	92		48
4Z	060	臺灣外記	3冊	448	江日昇	49
4AA	093	斯未信齋雜錄	1冊	120	徐宗幹	49
4AB	094	劍花室詩集	1冊	152	連橫	49
4AC	107	臺風雜記	1冊	62	佐倉孫三	50
4AD	119	諸蕃志	1冊	106	趙汝适	50
4AE	122	使署閒情	1冊	140	六十七	50
4AF	129	臺海見聞錄	1冊	68	董天工	50
4AG	131	李文忠公選集	5冊	806	李鴻章	50
4AH	154	明季荷蘭人侵據澎湖殘檔	1冊	64		51
4AI	155	清初海疆圖說	1冊	122		51
4AJ	161	臺灣語典	1冊	108	連橫	52
4AK	162	臺灣三字經	1冊	52	王石鵬	51
4AL	166	雅言	1冊	130	連橫	52
4AM	171	淡水廳築城案卷	1冊	119		52
4AN	198	清季外交史料選集	3冊	376		53
4AO	206	戴案紀略	1冊	62	蔡青筠	53
4AP	207	陳清端公年譜	1冊	114	丁宗洛	53
4AQ	208	雅堂文集	2冊	306	連橫	53
4AR	211	臺灣旅行記	1冊	110	諸家	54
4AS	214	清稗類鈔選錄	1冊	132	徐珂	54
4AT	220	碑傳選集	4冊	606	諸家	55
4AU	221	清史講義選錄	1冊	92	汪榮寶	55
4AV	222	臺灣兵備手抄	1冊	66		55
4AW	223	續碑傳選集	2冊	260	諸家	55
4AX	283	重修臺灣各建築圖說	1冊	80	蔣元樞	59
4AY	287	使琉球錄三種	2冊	290	諸家	59
4AZ	292	清代琉球紀錄集輯	2冊	282		60
4BA	293	琉球國志略	2冊	337	周煌	60
4BB	298	臺灣霧峰林氏族譜	2冊	386		60
4BC	299	清代琉球紀錄續輯	1冊	219		61

表二之三 文哲所進行中的資料庫（總字數：約4,941,000字）

一、道藏計畫：約 1,891,000 字

(一) 已完成標點、標誌工作，未建檔： 370,000 字

1.無上祕要 300,728 字

2.三洞珠囊 67,620 字

3.上清道類事相 1,700 字

(二) 二校已完成，正進行標點工作： 977,000 字

1.真誥 136,814 字

2.雲笈七籤 840,256 字

(三) 待校對、標點： 544,000 字

1.道樞 251,951 字

2.道教義樞 35,274 字

3.上清經派(六朝前) 257,017 字

(四)正在進行輸入(字數不詳)

1.正統道藏第五十六冊

(上清經派--六朝之後)

二、姚際恒著作集：(標誌中，約 950,000 字)

1.詩經通論

2.古文尚書通論輯本、禮記通論輯本(上)

3.禮記通論輯本(下)

4.春秋通論

5.古今僞書考

6.好古堂書目、好古堂家藏書畫記、續收書畫奇物記

三、劉宗周全集：(輸入中：約 1,100,000 字)

四、泉翁大全集：(輸入中：約 1,000,000 字)

表二之四 古漢語文獻進行中的資料庫（總字數：約650,000 字）

一、竺法護譯作(擬輸入)

no.291	《佛說如來興顯經》（《大正藏》第十卷）	約 26,250 字
no.292	《度世品經》（《大正藏》第十卷）	約 77,000 字
no.310	《大寶積經（三）密跡金剛力士會》（《大正藏》第十卷）	約 66,500 字
no.310	《大寶積經（四七）寶髻菩薩會》（《大正藏》第十卷）	約 26,250 字
no.323	《郁迦羅越菩薩行經》（《大正藏》第十二卷）	約 14,000 字
no.324	《幻士仁賢經》（《大正藏》第十二卷）	約 10,500 字
no.338	《佛說離垢施女經》（《大正藏》第十二卷）	約 14,000 字
no.342	《佛說如幻三昧經》（《大正藏》第十二卷）	約 33,250 字
no.381	《等集眾德三昧經》（《大正藏》第十二卷）	約 26,250 字
no.399	《寶女所問經》（《大正藏》第十三卷）	約 36,750 字
no.401	《佛說無言童子經》（《大正藏》第十三卷）	約 24,500 字
no.425	《賢劫經》（《大正藏》第十四卷）	約 112,000 字
no.433	《佛說寶網經》（《大正藏》第十四卷）	約 23,250 字
no.459	《佛說文殊悔過經》（《大正藏》第十四卷）	約 12,250 字
no.585	《持心梵天問經》（《大正藏》第十五卷）	約 56,000 字
no.606	《修行道地經》（《大正藏》第十五卷）	約 85,750 字

另與中山大學合作之明清文獻語料庫（1997年度）及與香港中文大學合作之「北宋五子語料庫」書目均待最後確定

參、相關技術的開發

中文全文檢索系統(CTP/FTMS)是在UNIX作業系統下開發的，歷經多次版本的修訂。目前我們推荐使用[表三]的規格來裝設此系統製作的資料庫和端末設備。

CTP的文件擷取方式(access method)非常單純，主要用的技術是字串比對(string matching)。比對時對字串的長度和數目沒有任何限制，而它能將設定範圍內所有文件中匹配的字串全部找出來並加索引。比對時對字串的內容亦無限制，只要是系統字碼表中（我們用BIG-5）任何的字(character)都行。所以，理論上英文和標點符號等等也是可以的。只是英文的詞(word)有種種字尾及自身不規則的變化，我們未加處理，是故對英文檢索的能力甚有限。比對的速度是經精心設計的。目前在486(33MHZ)中平均每秒約可處理25萬字。這是連做索引等工作全部一起計算的「單位時間內的產量」(throughput)。若在SUN工作站中，則可達每秒70萬字以上。

由於字串比對的速度極爲優越，整個系統可以不必預作索引，要查詢時現場做索引也不慢，且用過即可捨棄。因此，節省了大量空間，並省去系統更新資料時必須重作索引的麻煩事。但是，本系統也並非不可預加索引，預加索引後查詢的速度可不太受文獻大小的影響，查閱全部二十五史的反應時間在二秒左右，然而預加索引卻必須付出儲存空間及資料更新時維護的工作代價。索引、排除字集（Not in this word list）和邏輯關係(或OR,且AND，且非AND NOT)，以及網路功能等是由CTN進化到FTMS後所增的功能。此系統現有WWW(World Wide Web)版的檢索界面(1994年底完成)，可由本院的主畫面（WWW Home Page）切入，地址（URL）爲http://www.sinica.edu.tw

CTP/FTMS 保存了原始文件的版面訊息，如頁碼、行次。也保存了原始文件的文件結構，如標題、篇、章、節、小節、段落、註解等等。爲了要使計算機能了解這些訊息，在打字校對好的資料中，必須加上些標誌以使計算機能認得出上述的部件。有些標誌是在打字時，就訂在打字規則中的，如換行及頁碼。有些可以用程式偵察出來作自動標誌的，例如一般的段落。但是有些卻需要在打校之後用人工加上去，如篇、章等。因此打校好的資料，須要再經過「標誌」這一層工作，才可以輸入到FTMS去建構資料庫。輸入到FTMS後就一切都自動了，FTMS的輸出就是讀者可以立刻使用的全文資料庫。本系統的＜標誌規則＞，請參照[表四]。

由上觀之，要做資料庫，其實步驟甚單純，即在選定文獻後，經過繕打、校對、標誌後，即可上FTMS測試，若無毛病，資料庫即已大功告成。但是，從工作細節來說，仍有許多「經驗之談」(know how)是不吃過虧不會注意的，茲將這些經驗之談詳述如下兩小節中，其中最麻煩的是新造字的管理，這是目前字碼設計不良的後果。

表三、中文全文檢索系統裝置環境

中央研究院計算中心
1996/5/20 修訂

一. 主機部分

*作業系統: 各型 UNIX。中文全文檢索系統曾裝設於下列 UNIX 環境中

作業系統	主機
SunOS 4.x, SunOS 5.2 以上	Sun Workstation
AIX	IBM RS6000
SCO UNIX	PC 486 級以上
MITUX	同上
UNIX SVR4.2 for the Intel Architecture	同上

*硬碟容量: 視全文資料庫的大小而定, 如二十五史佔用 160M。此外檢索時宜有 40M以上的工作空間。

*資料庫安裝方式:

1) 經由 Internet: 如果主機已連上 Internet, 且檔案傳輸的效率良好,可經由網路安裝於硬碟上。

2) 使用磁帶: DAT(4mm)磁帶、8mm 磁帶或 150M cartridge 磁帶。

*備份裝置:

資料庫安裝後宜定期備分。常用的備分工具包含磁帶、可讀寫光碟、或其他硬碟等。

*Word Wild Web(WWW) server: 如果想經由 WWW 瀏覽器(browser)如 Netscape 進行檢索,需要這個環境, 供安裝 WWW 版檢索程式之用。

*系統印表機: 與主機直接相連的印表機, 或與主機經網路互通的印表機伺服器, 具備BIG5 中文字型, 且能擴增新字。若欲節省經費, 可予省略, 而以與使用者 PC 連接的 local 印表機替代。

二. 使用者設備

1. WWW 瀏覽器:

如果主機上裝有 WWW 版檢索程式, 使用者可利用 WWW 瀏覽器來檢索, 如NetScape、Webexplorer等, 它最好有表格(table)功能。WWW 瀏覽器所在的視窗(window)環境,要具備中文能力, 中文碼設爲大五碼(BIG5)。由於操作程序簡易, 一學即會, 建議儘量採用此法。中文全文資料庫(如二十五史), 往往包含爲數可觀的造字,初次使用前最好依指示安裝造字檔。不過目前只提供 CWindows 的造字檔, 假如未安裝 CWindows,無法顯示造字。這個問題有待克服。有意試用者, 可由中央研究院的首頁(home page)進入, URL 是 http://www.sinica.edu.tw。連上首頁後, 選「資料庫」,再選「中文全文檢索系統」即可。

2. 終端機(terminal): 將 PC 模擬成終端機連接主機, 並執行文字版檢索程式,這是傳統的作法。所需設備如下:

*IBM 相容 PC

*DOS

*連線

. 已架設 Ethernet: 網路卡與中文連線軟體, 如交大中文版 NCSA。

. 未架設 Ethernet: RS232 port 與相關連線軟體, 如 MS-Kermit。

*倚天或相容的中文系統(如國喬)

. 造字近四千五百字, 字型分爲 24x24 與 16x15 兩類, 前者字體美觀正確,後者則是由前者經倚天的工具轉換而來, 字體差強人意。目前可顯示 24x24字型的中文系統已非常普遍, 再配合 super VGA 顯示器即可。

. 如果使用者經由網路連上主機, 可自行由檢索系統取得造字檔,安裝於個人的中文系統中。否則必須設法提供造字磁片給使用者。

3.可配置 B4(132 bytes) local 印表機

表四、標誌符號簡表

1995/4月製表

1. 層級結構

~b	文素(textelement)起始。	~l	基本文素(段落)起始。
~e	文素終結。	~B	文章結構文素起始。
~E	文章結構文素終結。	~p<page>	第<page>頁開始。頁碼可爲雙階如3-2。
~d<dir>	被連結的檔案位於目錄<dir>。	~f<file>	連結檔案<file>。

2. 基本排版單元

\s	最常見的排版單元,排版方式可以設定。	\U<x>	使用者自訂的排版單元,<x>爲種類。
\h	標題。		
\ul	條列:各行分立,不得接合, 如將詩一句印於一行。	\ui	條項:除各行分立外, 行裡再以空白區隔諸句或諸項。
\um	居中:一行的文字內容位於行的中間。	\ur	偏下:一行的文字內容位於行的下側。

3. 表格

\tb	表格起始兼第一列起始,無表頭。	\th	表格起始兼第一列起始,有表頭,並視其爲第一列。
\tr	列的起始兼列的第一欄起始。	\td	欄的起始。欄的內容由基本排版單元組成。
\te	表格終止。		

4. 圈引

\qn,\Qn	夾注起迄。	\qf,\Qf	小字起迄。
\qa,\Qa	補文起迄。	\qd,\Qd	贅文起迄。

一、 資料建製管理

資料建製包括繕打、校對、標誌與測試四個步驟, 循序漸進。這並不表示完成所有的繕打才全面展開校對; 完成所有的校對才開始標誌。否則既會降低工作的效率, 也會礙及逐時推出部分成品的機會。除了繕打之外, 應該將全部的資料分割成若干單元, 如將一本古書按卷分工之類。每個單元由專人依校對、標誌與測試的作業順序, 一氣呵成。而不同的人可以對不同的單元平行作業, 最後加以組合,即爲完整的成品。即使只有一人, 前述的方式仍然適用。

1. 繕打

繕打前應瀏覽原始資料, 擬訂繕打規範。原則上按版面登錄,使輸入者藉直覺高速作業。然而爲了方便標誌, 甚或以程式協助標誌, 必要時可對常見的版面格式訂定特殊的繕打方式。如有造字檔, 繕打時務必裝設。

繕打規範通常包括下列事項:

- 擬訂資料分割、檔案命名方式。
- 一律橫打, 採橫式引號。
- 規定其他標點符號打法。規定各種排版單元起始空白數, 原則上按書籍版面留起始空白。
- 加頁碼標誌。
- 夾注起迄處標 \qn 與 \Qn。一行裡假如同時有大字原文及小字雙行夾注, 按行文順序打成一行。原始資料中的 \ 打成 \\。
- 含文字的表格原則上依行文先後順序打, 不理版面因素。
- 缺字打成●(A1B4)。
- 建立異體字替換表。

2. 校對

由於傳統人工校對的效率遠低於繕打, 且適當的校對人員也比繕打人員難求, 建議一式資料由不同人員繕打兩分, 初校時用程式比對兩分資料的差異, 僅就差異處加以修訂。兩位輸入者在相同位置犯相同錯誤的機會很小, 初校往往能迅速過濾大部分的錯誤。

校對時, 逐一記錄打不出來的字及其出處, 成爲缺字表。然後藉程式的輔助, 由缺字表篩出新字表。切勿純賴人工操作, 否則不僅浪費時間, 而且錯誤頻仍。把新字表交給造字管理小組做後續處理, 該小組會回覆一分核校過的新字表。受制於造字空間有限，依據各新字的的字頻, 決定處理方式:作爲公用造字、專屬造字、代以既存的異體字或捨棄。專屬造字至多 471 個。捨棄的字除外, 儘量查出諸字的注音。公用造字由造字管理小組統一製作，專屬造字通常由資料庫建製單位自製。

造字檔完成後, 據缺字表補字。事前最好複製資料檔, 供修補後校對之用。如果還有不造的字, 應自缺字表勾選其出處, 供使用者參考, 兼日後補正之依據。

3．標誌與測試

標誌前瀏覽資料，訂定標誌通例。實際作業時若遇到例外狀況，隨時討論增修。各個檔案先校對再標誌；先標文件層級結構，再標基本文素裡的排版單元。標層級結構時，逐層而下，完成一層後，再進行下一層。標誌最好用編輯器直接進行，不要先行紙上作業。標完，即刻試將此檔建成資料庫，檢查是否正確。

由於校對甫成時，對資料最是熟捻，針對限量資料，宜由同一人連續校對、標誌及測試，期能獲致較高效率。此亦可增加工作的變化，及定時推出部分成品。

4．資料庫開放

開放時宜附文件，說明緣起、原典版本、選擇依據、作業經過(人力、時程等)、電子版體例、造字表、異體字替換表、缺字表、校勘記、未含的資料等。

資料庫中若有專屬造字，稱爲專屬資料庫，必須配合特定的專屬造字檔，才不會導致誤讀。專屬資料庫的開放程度，及相關造字檔的散播，由持有單位決定。

正式開放的資料庫，只含公用造字。專屬造字一律取代成●。使用者的造字檔只要有正確的公用造字，不論含專屬造字與否，都不至於產生誤解。這樣的資料庫稱爲公用資料庫。資料庫完成備分後，原始資料檔便可刪除。

5．資料庫修改

資料庫完成後，蒐集使用者提出的指正，隨時更新。此外還可能受新公用造字的影響，必須修改，見＜中文造字管理＞。

二、中文造字管理

1．公用區與專屬區

Big5 碼造字空間有限，不能容納所有的新字，故將其割爲兩區，亦即公用區及專屬區。新字經過字頻統計後，選擇使用次數較高者，置於公用區，稱爲公用造字,其餘的捨棄不用。唯有些資料建製者，仍希望將這些字納入造字檔中，使電子資料完整無缺，遂有專屬區的設置。個別的資料建製者，得將排除於公用區之外的新字，安置於專屬區內，是爲專屬造字。不同的資料建製者，規畫專屬區的方式各異。專屬區無法容納的新字，應予割棄。公用造字由造字管理小組製作，專屬造字原則上由資料建製單位自製。造字檔如果只含公用造字，叫公用造字檔；假如又有專屬造字，便叫專屬造字檔。倚天系統的中文造字空間安排如下:

公用區:高位元組（high byte）	FA~FE, 8E~9D, 81~8D	
低位元組（low byte）	40~7E, A1~FE	(計34*157 = 5338字)
專屬區:高位元組（high byte）	9E~A0	
低位元組（low byte）	40~7E, A1~FE	(計3*157 = 471字)

2．造字程序

(1) 蒐集缺字: 建製資料時, 於校對過程中, 記錄每個打不出來的字之出處, 成爲缺字表。

(2) 初篩新字: 執行輔助程式, 逐一輸入缺字表裡每個字的倉頡碼, 由程式檢查是否和先前輸入的其他倉頡碼重複。假使未重覆, 這個字必然是新字, 將字型及程式所提供的字號登錄於新字表上。反之, 程式會報告與某號字的倉頡碼相同, 應據以核對新字表上該號字的字型, 判斷是否相同。若不同, 仍按發現新字的方式處理。初篩新字宜由單人爲 之, 使每個缺字的倉頡碼趨於一致, 不論其正確與否。

新字表格式：字號、字型: 初篩新字時，只填此二欄。
處理: 新字處置方式爲，新公用造字(公)、專屬造字(專)、代以異體字(異)或捨棄，四者之一。由造字管理小組決定是否爲新公用造字,如果不是，由資料建製單位決定處理方法。
注音: 處理方式決定後，由資料建製單位提供，捨棄的字除外。
異體字: 造字管理小組在複查新字時提供異體字或疑似異體字。
新字總表字號: 造字管理小組匯整各方新字後填寫。
已造: 檢查新公用造字和專屬造字是否已經造過，由造字管理小組填寫。

(3) 複查新字: 將新字表及新字篩檢程式所產生的檔案交給造字管理小組複檢。首先檢查該檔中的倉頡碼是否正確。如有更正情事, 應檢查更正後的倉頡碼是否和其他倉頡碼相同, 若相同則再查兩字字型相同否, 若又同,便將兩字字頻合計於一字處, 並將另一字自檔案及新字表刪去。接著裝妥公用造字檔, 啓動中文系統(所有的造字務必已設倉頡碼)。仔細輸入每個新字, 並刪除打得出來的字。由於各人拆解倉頡碼的方式, 在大同中偶見小異, 建議用倉頡簡易法詳細篩檢, 以免因錯用倉頡碼, 漏失了既存的字。簡易法較可能發現新字的(疑似)異體字, 請將之記於新字表異體字欄, 必要時可用來替代新字。

(4) 累計字頻: 造字管理小組裝妥新字篩檢檔, 此檔含有歷來發現的新字而未歸於公用造字者, 與其對應的新字表叫新字總表。執行輔助程式,一面累計新字表諸字字頻, 一面在新字表上標記新字總表字號。如找到從來未遇的字, 同時將其字型及字號登錄於新字總表。

(5) 決定處理方式: 按新字篩檢檔的字頻排序,選字頻最高者爲公用造字。把這些字由累計用新字篩檢檔及新字總表剔除,其餘的字由資料建製單位,研判處理方式。

(6) 造字: 依字形，實際製作新字。

(7) 修訂已完成資料庫: 其他已完成的資料庫, 可能受到新公用造字的影響。其新字表如果含這些新公用造字, 應適當加以處置 。

肆、進行中相關的研究和展望

CTP的設計始於1985年，第一版在1988年大功告成，並於6月間推出廿五史的前四史，即＜史記＞、＜漢書＞、＜後漢書＞、＜三國志＞供大家使用。自此之後，雖屢經改版並陸續改進和新增了不少功能，但是其主要結構未曾改變。目前國內已有十一個機構，國外也有七所大學安裝了本院製作的古籍全文資料庫，使用日益普遍，詳如[表五]，古漢語文獻語料庫雖未在院外裝設，然亦有國內外共八個單位獲准使用。由於計算機科學及信息處理技術的突飛猛進，這十多年來計算環境也變得和以往大不相同了，設法更新以迎合新設備、新技術固然是理所當爲，然而，也有更徹底的做法，即將整個系統重新再設計一次。在本節中，我們將報導些在這方面的進展，以展望未來的中文全文檢索的面貌。

表五・院外古籍全文資料庫

計算中心 1996. 12製表

裝設地點	資料庫	啓用日期
國史館	新清史	
中央圖書館	二十五史	1995. 2
清華大學歷史研究所	二十五史及其他[說明1]	1994. 11　1995. 518 啓用
東海大學文學院	二十五史及其他[說明1]	1994. 11
臺灣大學	二十五史等[說明2]	1995. 4
逢甲大學	二十五史及其他[說明1]	1995. 5
中興大學	二十五史	1995. 6
東吳大學	二十五史	1995. 9
臺灣師範大學	二十五史等[說明3]	1996. 6
中山大學	古漢語文獻語料庫(部份)	1996. 11 更新
成功大學	二十五史	1996. 11
英國倫敦大學亞學院	古漢語文獻語料庫	1993
美國華盛頓大學東亞圖書館	二十五史、十三經、文心雕龍	
美國哈佛大學燕京圖書館	二十五史	
德國海德堡大學漢學系	二十五史	
美國柏克萊加大東亞圖書館	二十五史、文心雕龍	1995. 1
香港科技大學	二十五史	1995. 10
香港中文大學	二十五史[說明4]及 古漢語文獻語料庫(部份)	1995. 10
日本東京大學	二十五史	1996. 5
美國史丹佛大學	古漢語文獻語料庫(部份)	

[說明 1] 其他包括太平御覽、唐會要、斷句十三經經文、藝文類聚、高僧傳、續高僧傳、弘明集及廣弘明集及廣弘明集，部分尚在編製之中。

[說明 2] 1996. 年 6 月 14 日加裝諸子與古籍十八種。

[說明 3] 另有諸子、古籍十八種及十三經。

[說明 4] 斷句十三經、莊子集注、老子、國語、列子等裝於 1995. 年 10 月。二十五史即將採購。

一、和關聯式資料庫的整合

全文檢索系統的一大缺點，就是處理文件的屬性時很不方便。例如，不能從作者找到其著作，只能檢查著作中的文句。有鑑於一般圖書檢索系統利用關聯式資料庫對屬性的檢索已發展得很成熟，試圖將二者合而爲一是很自然的想法。將二者結合有二種做法，其一是將全文資料檢索系統中的每一個文件元素都設一個關聯式的記錄(record)，再將這些記錄歸並成些關聯的資料庫，並提供這兩種資料庫之間彼此往返的

管道。其次是將關聯式資料庫中某些欄位(field)改成變動長度的全文式欄位，只在這些欄位下提供全文檢索功能。

這兩種做法都經過嚐試，前者結構太複雜，且沒有具體的計劃支持，是故沒有成功。後者已有成品，名爲文件檢索系統(簡稱DORE，即Document Retrieval System，本院計算中心發展)，利用它也做了些資料庫，如＜內閣大庫檔案索引＞約有五萬筆資料，又如＜歷史書目＞三萬七千筆，＜台灣考古學著作＞三千餘筆，及＜中國考古學資料庫＞五千餘筆等，目前使用得尙滿意。

依作者看來，前者雖未成功，但仍深具潛力，值得再試。後者功能受限制，目前雖成功，但前景並不樂觀。

二、擷取引擎的開發

雖然字串比對技術已運用得相當成熟，更新的擷取技術仍待開發。開發新的擷取引擎並不是完全取代已有的擷取方式，它們是相輔相成的。不同的擷取方式可滿足不同方式的查詢(query)要求。對使用者而言，查詢的變化越有彈性越好，而系統中裝設幾個不同的擷取引擎以伺候不同的查詢需求亦非難事。

一個利用字與字之間相鄰或相近的機率模式，再將字碼折疊編碼(super imposed coding)而呈現文件識別特徵(signature)以資擷取或檢索文件的系統發展得甚爲成功。這個系統的名稱是「尋易智慧型文件檢索系統」(簡稱尋易，CSmart)。

此系統有許多優點。第一，它的索引深度是可調式的（基於機率），且空間需求小，搜尋速度極快（與原文件檔案大小無關），是故可處理巨量中文文件檔案，如數十億字的文獻。第二，它有近似自然語言的查詢功能，能做部份匹配 (partial match)，亦能做中英夾雜的查詢句，如查「宏碁ACER主機板量產」時，可將宏碁、ACER，主機板、量產等任意部份組合的文字都找出來。第三，它能和語音識別系統，如金聲三號，合併使用，直接以語音做查詢。第四，它可以在INTERNET的WWW上運作。這個系統目前已正式公開，讓大家免費試用【註六】。雖然該系統的機率是依白話文統計，我們利用江味農先生《金剛經講義》共約五十萬字作古文測試，成果尙令人滿意。若有古籍的語料庫，先校準文言文的頻率，此系統應可用於古籍的檢索及查詢。

尋易系統目前並沒有能力處理依標誌語言SGML所標誌的文獻，這是未來發展必須克服的。如果尋易能夠擴充至處理標誌結構的文件，那麼和全文檢索FTMS可嚐試著合而爲一，爲使用者提供更友善，更好的全文檢索服務。

【註六】「尋易智慧型中文資訊檢索系統」之技術資料，請參照簡立峰博士論文：

1.＜A Model-Based Signature File Approach for Full-text Retrieval of Chinese Document Databases.＞ Computer Processing of Chinese and Oriental Languages, 1995.

2.＜Fast and Quasi-Natural Language Search for Gigabytes of Chinese Texts.＞ ACM SIGIR95

3.＜尋易（Csmart）—A High Performance Chinese Document Retrieval System.＞ ICCPCOL95

4.＜適合大量中文文件全文檢索的文件索引及壓縮技術＞ ROCLING95

三、網路上的發展

由於網路的使用日益普遍，除「尋易」外，中文全文資料庫WWW版本也已推出使用。中文全文資料庫採用描述性標誌(descriptive markup)，WWW的 HTML 亦然，而前者遠較後者簡單，所以由前者轉換爲後者，通常非常直接。既然如此，爲何不直接採用後者? 主要的理由正是求簡單，因爲資料量動輒爲百、千萬字，標誌者的工作似乎愈單純愈好。另一個理由是中文古籍有某些編排習慣爲英文所無，我們可以設計一些簡易的標誌來保存這些特性。如果直接標上 HTML，可能發生不易逆推其原來特性，甚或失真而無以復原的現象。譬如當代排印的古籍，習慣以小字前後加上（與）的型式，代表贅文，亦即這些字爲原文所有，但被後人視作多餘。我們以 \qd 及 \Qd 包夾贅文，檢索時可以依據這對標誌剔除贅文，而顯示時則輕易將其轉變成HTML，獲致大家熟知的版面。

FTMS的標誌轉換爲HTML的標誌是直接而單純，主要的轉換規則如[表六]所示。

表六FTMS的標誌轉換爲HTML的對照表　　1996/5月

標誌性質	中文全文資料庫標誌	HTML
段開始	\s ,\U<x>	<p>
居中	\um	...
偏右	\ur	...
空行	*<newline>*	
頁開始	\p	<hr>
注文	\qn ... \Qn	 ...
大字	\qf ... \Qf	 ...
補文	\qa ... \Qa	[...]
贅文	\qd ... \Qd	(...)
表格開始(無表頭)	\tb	<table><tr>
表格開始(有表頭)	\th	<table><th>
表格一筆開始	\tr	<tr>
表格一欄開始	\td	<td>
表格結束	\te	</table>

說明：層級結構標誌部分, ~d 與 ~f 在全文資料庫建成後, 不復存在。~p 變成 \p,再轉成 HTML 的水平線。至於最重要的 ~b(或 ~l) 及 ~e 對子, 在資料庫建力過程中, 化爲樹狀(tree)資料結構。在 WWW 瀏覽器上查詢資料庫目錄時,根據目錄在樹狀結構上的相應位置, 即時產生目錄訊息, 其間的轉換, 已非直接。~B 與 ~E 對子係針對本院資訊所、史語所合製的古漢語語料庫而設, 目前未提供 HTML 轉換。

目前幾乎所有的全文資料庫都已有WWW版本，然而由於這些資料庫智慧產權尚未釐清，所以僅提供本院內部使用。希望不久的將來能開放供全世界人士利用。

四、以網路上的字形資料庫解決字碼(缺字和造字)問題

在網路上的另一發展是利用網路強大的及時溝通能力，充份告知缺字和造字的字碼(Code) 和字體(Font)信息，並提供便利的字形資料檢索。這樣的安排可以使得使用者隨時掌握共用造字表中的詳細訊息，以達到缺字的資訊共享和交流的目的。

在前文中可知，交換碼中的字不夠用是個大問題。如果每個資料庫建構者都自已造字，那麼這些的字不僅會帶來使用上的不方便，造成衝碼，並且嚴重妨礙了資料的匯集和共享。因此，在沒有理想的交換碼可用之前，解決造字問題已是燃眉之急務。

目前，一個字形資料庫的雛型已經完成。它可利用字根檢索任何含有該字根的文字【註七】。這個字形資料庫將於近期內在本院內試用。相信它的使用不僅可解決上述的困難，對於建製全文資料庫時的資料登錄和造字管理工作，都將大幅化簡。

五、以內容爲檢索對象的研究

傳統的文件檢索，是藉詞彙或字串爲媒介來找到文件。然而有許多文件的內容相當隱晦，並沒有明顯的文詞可資識別。爲解決這種檢索上的問題和古文各版本經傳注疏等之相互參照關係，資訊所文獻處理實驗室也做了若干嚐試。目前這方面已有些初步的成績，例如，該實驗室已經建立了：（一）一個連接知識結構和文章的新超文件（hypertext）模式；（二）發展了新的標誌界面，使任何一個文件可輕易地建立多重的標誌(Multiple/concurrent togging)，而且使用者不必去記憶標誌的名稱；（三）知識結（knowledge chunk）的名稱可以作標示(tag)的名稱，這樣可以讓使用者充份活用此超文件中的知識結構來了解文獻、導讀文件和蒐尋文件等；（四）已成功的做了一個能處理多重版本間彼此對映參照的示範系統。這方面研究的前景是相當看好的。日後將會嚐試與FTMS合併，以提供多重檢索的功能【註八】。

從以上可知，全文檢索的下一代系統正在醞釀中，這些新的工具，配合中央研究院十餘年來，持續不斷地在電子古籍方面的努力，在不久的將來，利用電子計算機處理古籍的事務，應呈現嶄新的面貌。

誌　謝

本文所列之資料清單及其敘述，承蒙黃居仁、劉增貴、魏培泉、陳弱水各位先生之指正，謹此誌謝。打字、排版承張翠玲小姐鼎力完成，功不可沒。此外，黃佩燕小姐、黃淑齡小姐也協助整理表格資料，亦一併誌謝。

【註七】請參照

1.謝清俊、莊德明、張翠玲、許婉蓉＜中文字形資料庫的設計＞,第六屆中國文字學全國學術研討會，1995年4月29日，台中：中興大學中國文學系所、中國文字學會主辦。

2.謝清俊〈電子古籍中的缺字問題〉,第一屆中國文字學會學術討論會, 天津, 1996, 8

3.謝清俊〈漢字的字形與編碼〉,漢字字碼與資料庫國際研討會 京都・東京, 1996年10月4日

4.謝清俊〈A Descriptive Method for Re-engineering Hanzi Information Interchange Codes〉,漢字字碼與資料庫國際研討會 京都・東京, 1996年10月4日

【註八】關於這些研究，可參閱下列文件:

1.陳昭珍＜建立古籍多版本超文件之探討－以文心雕龍爲例＞中國古籍整理研究出版現代化國際會議,1995,7

2..謝清俊 莊德明＜古籍校讀工具「中文文獻處理系統」的設計＞中國古籍整理研究出版現代化國際會議,1995,7

3.王梅玲 郭怡雯 李明錦 張翠玲＜內容賞析資料庫初探＞中國古籍整理研究出版現代化國際會議,1995,7

4. 莊德明＜以心經爲例說明如何利用計算機處理佛經的多版本＞中華佛學研究所1995, 8