論漢學研究用中文資料庫的開發

On the development of Chinese language Data-bases for sinology studies

謝清俊，中央研究院 資訊科學研究所 研究員

兼 中央研究院 計算中心 主任

摘 要

近三、四年來，中央研究院計算中心陸續開發了一些中文資料庫以協助本院學者從事漢學研究。這些資料庫包括了本文介紹的二個全文資料庫，及八個傳統式的欄位資料庫；其中一般支援性的資料庫有五個，專業性的五個，分別是：

專業性的：
- 漢代墓葬資料處理系統
- 戶籍資料處理系統
- 台灣土著語言資料庫
- 土地申告書資料庫
- 清代內閣大庫索隱

一般性的：
- 台灣省博碩士論文資料庫
- 中研院研究人員著作資料庫
- 電子辭典
- 二十五史全文資料庫
- 電子卡系資料庫

本文將就以上發展的經驗，探討建立漢學研究用中文資料庫的問題，並針對這些問題提供一些建議，以為建立漢學研究用資料庫者參考。這些問題分為三類：有關設計和製作的問題，有關運作的問題，以及關於使用人的問題。此外，本文將從更廣的角度來討論資料庫對漢學研究可能的利用，以及未來的展望等。

壹、前言

大家都知道計算機是處理資料的利器。然而，由於語文上的障礙，未經「加工改造」的計算機是無法直接處理中文資料的。為了使計算機有能力處理中文資料，國內的學術界和工商業界經過十七年的努力，終於使中文資料處理的問題，獲得若干舒解。譬如，中文的輸入、顯示、列印等已不再是嚴重的問題，而中文的檔案管理、資料庫運作、文書處理、編輯、排版、程式撰寫、傳送,.....等等，目前都可用計算機做到了。雖然上列的這些功能並非完善，甚至有些系統還有些小毛病，然而，我們已經擁有計算機處理中文資料的環境，則是不爭的事實，而利用計算機助長漢學研究的時機，已然成熟。

近三年來，中央研究院（以下簡稱本院）計算中心（以下簡稱本中心）陸續開發了些中文資料庫以協助學者從事漢學研究。這些資料庫有的相當單純，並不需太多人力開發，例如清代內閣大庫索隱的系統，花一個人工作一、二個月就夠了。有的相當複雜；經過三年多的時間，投入廿餘人年，還未全部完成。如戶籍資料處理系統、電子辭典，廿五史全文資料庫等是。有的可用些現成的軟體開發，例如下列的八個傳統欄位式的資料庫；有的卻需要連軟體工具一起開發，如全文資料庫。有的是專業性的資料庫，這是專門為某研究計劃開發的；有的卻是通用性的資料庫，它們不限於支援特定的研究計劃開發。開發這些資料庫的經驗彌足珍貴。在本文中，挑選了本中心開發的十個資料庫作為例子，希望借此討論公開我們的經驗，以為有志利用計算機作漢學研究者參考。

本文的結構安排如下：

壹． 前言

貳． 各資料庫簡介

參． 關於資料庫的設計與製作上的考慮

肆． 關於資料庫運作上的一些問題

伍． 使用者發生的困難

陸． 漢學研究的新工具

柒． 建議與結語

在「各資料庫簡介」這一章裡，介紹各個資料庫建立的目的、資料的性質與數量、功能、特性、優缺點、和一些附加的說明．以下三章，將分別依資料庫的製作(development)，運作(operation)，及使用 (use)三個角度，提供我們的心得與經驗。在「漢學研究的新工具」一章裡，將談談目前的計算機技術對漢學研究還可能有那些助益。最後一章則以對發展漢學研究資料庫一些重要原則性建議作為本文的總結。

貳、各資料庫簡介

資料是研究的基礎。漢學研究面對著的是浩瀚的資料。自古以來，這浩瀚的資料一直構成漢學研究的嚴酷挑戰。對學者而言，這本是個皓首窮經的艱辛歷程，而如今，利用計算機處理中文的技術，可能帶來突破此障礙的契機：創造一個斬新的研究形態，使得學者免於從事大量的資料處理工作，而能夠將精力集中在創意的研究工作上。

資料庫是目前計算機中最適合管理大批資料的工具。舉凡資料之組織、儲存、擷取、搜尋、比對、印製、增刪、更新等等，皆可由資料庫為之。若將資料庫中查詢到的資料，以適當的應用軟體處理，則可以做些更複雜的事情，譬如：統計、分析、計算、推理、編排....等等。由此觀之，資料庫系統對漢學研究是十分有幫助的。基於這樣的理念，本中心先後開發了不少為支援漢學研究用的資料庫，茲將這些資料庫分別介紹如下。

一、漢代墓葬資料處理系統

此系統共蒐集3040個漢代古墓的記錄報告資料，大多是1986年以前發表的，其中豎穴墓1366筆，磚室墓1674筆。每筆資料含一般性資料、墓結構、裝飾與銘文、棺結構、槨結構、和隨葬品等六類，合計超過四百項。每類資料變化甚大，譬如隨葬品再分為63種，總計達十餘萬件，平均每墓約有30種隨葬品。

由於資料分項甚細，此系統可以查詢和計算甚為細微的項目，這是它的特色。此外，它還可以做簡單的統計和百分比計算等。例如，它可以做年代統計、地理區統計、或某時某地區未擾墓中各類隨葬品器物件數之比值等等。

此系統是本中心早期發展的系統之一，當時中心成立不久，受人力、設備和經驗之限制，缺點難免。譬如使用者介面設計不是很好，反應速率也嫌慢。在發展過程中，遇到的最嚴重困難是專業上的溝通：包括資料的分類問題和系統功能上的界定。此外，資料的校對相當費人力，查詢畫面之設計，因也欠缺軟體工具而不是很賞心悅目。這些問題在後繼發展其他資料庫的過程中，曾獲得若干程度上的改善。然而它們仍然是發展各個資料庫時，要設法解決的問題。

二、戶籍資料處理系統

這是為協助人類學量化研究所發展的系統。此系統利用戶籍資料，從事1906年後在台灣地區客家和閩南社區「家庭與婚姻」方面的研究。至目前為止此系統已支持了三個研究計劃：在1985年國科會資助的「家庭與婚姻」研究中，用以分析婚姻與生育率的問題；在1986年美國Rockefeller 基金會資助的計劃中，則以婚姻類型及婦女生育率為變項與台北海山資料比較，用以驗証Freud和 Edward Westermarker 的近親禁忌理論；在1987年國科會資助的「出生排行序與家庭結構動態」研究中，則以出生排行序為自變項，分家的時間、收養的決策、婚姻的類型、出生率、離婚率、及遷徙為應變項，以探討其因果關係；進而比較閩、客社區之差異與特質，並驗証 Schachter和Caplow所建構的聯盟理論是否可以概推到中國社會。這些研究之結果當可增加對中國家庭成員之間互助互動關係之了解，並提供其他社會學科對孝道、社會化、政治態度等研究之參考。

依資料性質來分，此系統可分為日據時代戶籍資料(從1906至1945)，和民國時代之資料(自1945迄今)兩部份。日據時代資料含竹北、峨眉、大內三個鄉共10338戶。光復後之資料有竹北、峨眉二個鄉,共約7000戶。每戶之資料為樹狀結構；即每戶下分為個人資料，而個人資料下再分事件記載資料。亦即基本記錄(record) 分為戶、人、事三種，並分開儲存。在日據資料中，共計111672筆人的記錄，113227筆事的記錄。至於民國資料的數量，則由於資料仍在校對中，詳細數目待查。

此系統除供查詢外，曾完成若干人類學研究用的應用程式。發展應用程式的方法和程序是 :此系統先產生有關生育、養育、婚姻、職業、教育、家庭型態、出生排行等等中介檔案，以便利用簡單的統計程式可由上列的中介檔案裡得到單項或多項之因果分析。而在因果分析上，則利用統計或人工智慧中的推理技術來得到欲求之結果。此資料庫雖然專為人類學研究而設計，

但由於上述的結構甚具彈性，甚易轉換為其他人文、社會研究之用途，此為其特色。

此系統主要的困難在應付原始資料之錯誤情形以及資料涵意混淆的情形。是故配備有專為校驗資料的程式。例如：稱謂中之" 妹",可為妹妹、弟之妻、或養妹。究竟代表那一個？則需在程式裡經過一大串的推理程序始可做正確之判斷。利用人工智慧技術解決此類問題亦為此系統之特色。

三、台灣土著語言資料庫

這是配合台灣土著語言研究所發展的工具，其主要目的在整理資料。原計劃在1986年二月獲國科會之資助至1987年六月結束，此為全程計劃之第一期。之後由於主持人李壬癸教授兼理行政工作忙碌，第二期計劃迄今尚未展開。

此系統蒐集之資料皆為李教授數十年工作之心血，其中包括五大項：*(1)*台灣土著語言詞彙比較表、*(2)*個別方言詞彙檔案五種、*(3)*同源詞檔案、*(4)*台灣土著語言論著目錄，以及*(5)*其他檔案。

在*(1)*項中，共計有1373筆資料,每筆資料包括仍然存活的14種台灣土著語言，30多種方言、以及日據時代記錄的五種平埔族語言。詞形中含有古語的擬式 (reconstructed forms) 。 每筆資料佔1818 bytes，共計約佔 2.5 Mbytes的儲存空間。

在*(2)*項中，集有瀕臨完全消滅的邵語、噶瑪蘭語、和巴斯海語,以及卑南王方言和泰雅語汶水方言。合計2734筆，共佔約0.5Mbytes。每筆資料含序號、語言符號、及中英文語意說明等等。

在*(3)*項中，蒐集了業經認定了的同源詞目1022筆。然而,此資料仍將隨研究工作之進展而陸續增加。每筆包括各層次古語擬式：古南島語、西部南島語、古台灣土著語、南台灣語、古魯凱及古泰雅語等。

在*(4)*項中,包括中、日、英、荷、德、法、西等國發表的相關著作之書目資料。由於語文關係，此部份資料在IBM5550 計算機中建檔。

在*(5)*項中,包括泰雅語群男女性語形比較表，魯凱語茂林方言肯定與否

定詞變化表，和台灣土著語言的對音關係及其對古南島語的反映等資料。

這個系統至目前僅止於資料整理的應用而已，在應用程式的開發上尚未開始。並由於開發的工程人員對語言學了解不夠，溝通又復不足，以致資料及檔案結構之設計並非完善。在第二期中，或將重新定義資料結構，以期建立較完善的系統。

四、二十五史全文資料庫

這是一個通用目的(general-purpose)的全文資料庫，其目的在將正史資料之原文納入資料庫，以利索檢與分析。目前納入資料庫的有史記、漢書、後漢書、三國志，以及廿四史食貨志匯編等，共計約七百萬字。

此資料庫是利用本中心自己發展的中文文獻處理系統 (Chinesetext processor 簡稱 CTP) 為工具(資料庫管理系統)所作成的。它有兩種檢索的方式，其一是利用詳細至段落的目錄，稱為「文獻結構」的路徑來檢索原文。其次是利用字串比對技術所做的「自由詞檢索」(free-term search)。此方法可檢索任意數目與長度的詞彙或字串(註1)。CTP 是一個處理中文文獻檢索通用的工具，其使用並不限於史書。在其他的計劃裡，CTP 亦曾用於處理三民主義講稿原文、四書、及一些教育與文化的論文。此系統除了做檢索以外，還可以做些語文統計 (con-cordance)，做各種形式的列印，以及和電子卡系資料庫合用來製作研究用的卡片(note-card)。目前，簡單的語文統計可以在線上及時做，複雜的就要以批次作業來做(以免等待太久)。至於卡系，則是另一個協助文史研究用的資料庫系統，且留在下文中介紹。

這個系統開發時投入的人力和時間較其他系統多，在此系統之前，尚無中文全文系統出現。 開發的過程中遇到不少學術研究的問題。譬如:文獻的結構問題，文獻標誌 (mark-up) 的問題，文獻在計算機中如何表達的問題，以及如何檢索中文全文資料等等。

在製作過程中，資料校對頗為嚴格，執行得甚為徹底，是故在資料庫中幾無錯字。研究人員可以放心使用。人機介面亦經精心設計，文獻可以隨使用人挑撰呈現直式或橫式的畫面，而每次畫面亦可隨使用人選擇呈現一段、一頁或是將畫面充滿文字資料。此外有「裁文」功能，使用人可以用游標選擇螢幕上任一段文字(無論直排、橫排)留作後用。「裁文」也是電子卡系資料庫標準介面，是故裁文後得到一段資料可以經系統自動送到卡系中建立研

究卡片。

在列印方面，可以印檢索之引得，印檢索列的各段落原文,以及印含有檢索詞彙之原文子句(兩標點符號之間的字串)等等之選擇。

註1:目前的限制是至多1000個檢索詞,所有詞彙的總長不得超過10000 bytes，限制可方便的重行訂定。

五、土地申告書資料庫

這是為協助台灣社會經濟史研究，特別是清未及日據時代土地開墾型態及租佃關係方面所發展的系統。資料的來源是根據日據初期台灣總督府為了重新分配土地所有權所製作的調查:「土地申告書」。這是讓老百姓報告的文件，其內容包括調查前土地所有的情形和調查後的土地分配狀況。然而，本省大部份地方政府所存的土地申告書均不幸焚毀，目前留下僅存的新竹地區部份文獻共計 472冊，約25萬筆資料。這些珍貴的資料至 1988年9月，已完成鍵入的工作，目前在校對中，整理後的資料每筆約有83項。

此系統之設計將仿戶籍資料處理系統，估計在發展上問題不多，全部資料量粗估約有15MBytes。

六、清代內閣大庫檔案索隱

這是配合明清檔案整理計劃所做的索引資料庫。全部檔案資料仍在整理中，並未詳細清點過，故粗估約有廿萬筆，每筆有九項資料，都是簡單的索引。

由於只是單純的索引，是故以PC/AT為工具，利用 dBASE Ⅲ plus做資料庫，舉凡資料之增減、修改、排序、引得建立等皆利用原資料庫之指令，所費人力不多，操作亦無困難. 目前已在使用中。

七、台灣省博碩士論文資料庫

此資料庫目前已建立了 23111筆1974年至1984年台灣省發表的博碩士論文資料，以及 1985 年博士論文資料。其中 1980 與 1981 兩年的工學院論文有論文摘要。每筆資料有 14 項，包括學生及指導教授姓名、論文題目、學位別、學校及院、所、系別、年度、及摘要指引等。

此資料庫提供八種檢索點(如上列)，有簡單的組合邏輯及字串比對功能

以提供查詢服務。此外，可作些列印功能。

這是一個簡單的資料庫，沒有索引典亦無權威檔控制，所以檢索時問題很多。例如，中文英文混合之查詢就必須分別做。此外，資料之蒐集是一大問題，這是無法使資料更新至當年度的原因。目前每年約增加三千多筆資料，對此資料庫的維護而言，是個不小的負擔。

八、中研院研究人員著作資料庫

此資料庫目前已登錄了 6530 筆資料，每筆資料包括十項，它們是：作者、所別、著作種、著作標題、發表時間、研究範疇、關鍵字等等。有查詢、統計、列印等功能．查詢的方式有四種； 游標定位、字詞檢索、 字串比對 、及全部項目之簡速查詢。 在使用者介面上，此系統的設計算是考慮較周延的。在統計方面，它可依全院、所別、或個人，作某時間內，某種著作之統計。

這也是個簡單的系統，它仍然有蒐集資料不易的困難。在計算機技術上，此系統有幾個特色：資料結構和檔案結構嚴謹，故易於擴充新欄位；利用 proximity 技術及 ISAM 特性,增加了利用 INFORMIX 資料庫管理系統查詢之速度；再者，在游標查詢作業時，不需鍵入任何中文字便可完成查詢作業。這些特色值得作發展新系統時參考。

九、電子辭典（國語日報辭典加語法及構詞資料）

這是一個全文資料庫。它將國語日報辭典全部資料建立成資料庫中的一部份，另外加上本中心與工研院電子所合作的「詞知識庫計劃」中對辭性分析的資料部份，結合為整個的資料庫。此資料庫不僅是一個通用的電子辭典，它也是支援語文學研究及人工智慧領域中對自然語文處理研究的重要工具。目前，史語所黃居仁先生與資訊所陳克健先生合作的「中文語法模式的研究」，以及本中心和工研院電子所合作的「中文語句剖析計劃」都用到這個電子辭典。除此之外，在發展此系統之初，亦和發展 CTP 一樣,希望借此計劃發展一種能專門處理條列式全文資料的軟體工具。此軟體工具之雛型，亦隨此資料庫順利完成。換而言之， 它不只可處理國語日報辭典， 也可以處理其他的辭典。

此詞典目前有四萬目詞，每筆詞除了國語日報辭典上的資料以外，還有

其詞類、構詞、屬性、用法、及一些對應指標等資料。在功能方面，有查詢、維護（修改、新增、刪除等）及說明（類似help）等，並可任意於詞彙、發音、詞意、詞類等欄位做一個或一個以上條件之自由詞檢索。

十、電子卡系資料庫

這是一個較特殊的資料庫，因為它沒有固定的資料！它所處理的是屬於使用者私人的卡系檔案。每用到此系統時，使用人應先載入已有的卡片，用畢後應將卡片資料載至軟性磁碟片上攜回。它的功能就像一般研究用的卡系。然而，它可以單獨使用，或是和其他系統配合使用。目前開放使用的，只是配合廿五史全文資料庫使用之版本。單獨使用的版本近期間當可開放。

每張卡片資料最多佔 1KBytes 空間，共有 13 項資料，包括： 資料內容、序號、出處、及八種分類項、及關鍵字欄位。 此系統主要的功能有四； *(1)*卡片之管理：如傳輸，使用權設定、複製、消除；*(2)*卡片建立:包括自行建立或與其他系統合用(資料內容由其他系統轉來）；*(3)*查詢:包括依分類查詢，或填單式查詢(關鍵字查詢部份尚未完成）；以及*(4)*列印。

在卡片中「資料內容」部份，可用 UNIX 下的 Vi 編輯器做中英文的文字處理。「資料內容」之全文查詢功能尚在開發中。

各資料庫較詳細的技術資料，請參考表一與表二。

除了以上的資料庫以外，本院還有幾個資料庫亦與漢學研究有關，它們是：史語所鄭秋豫副研究員發展的中文語音資料庫,這是為了作語音分析(包括：聲韻學中的時域，time domain，和頻域，**frequency domain**的分析)，以及作語音學習及接受(perception)研究用的；本院圖書館作業自動化計劃所擬中西文的圖書書目資料庫；史語所策劃的善本書影像全文與書目資料庫(與中央圖書館與本中心合作發展的)；以及本院公眾資料庫等等。這些資料庫除語音資料庫文發展已具規模外，皆費甚多時間在策劃上，目前雖已開始發展，但進度甚少，故不在此討論。然而這些資料庫均對漢學研究的資料處理上，或訊息溝通上，相當有幫助。

表一、 各專業資料庫系統介紹

項次 \ 項目 \ 資料庫系統名稱	漢代墓葬資料處理系統	台灣土著語言資料庫	戶籍資料處理系統	土地申告書資料庫	清代內閣大庫索隱
1.相關所別 2.相關研究計劃	史語所 「漢代墓葬研究」	史語所 「台灣土著語言資料自動化」	民族所 「家庭與婚姻」方面的一系列研究	三民所 「清代竹塹地區漢人聚落發展與土地租佃關係」	史語所 明清檔案整理
3.主持人及研究人員	蒲慕州	李壬癸	莊英章 Wolf	張炎憲	張偉仁
4.開始發展日期	1985.11	1986.2	1986.4	1987.4	1988.3
5.目前狀態	已完成 使用中	已完成 使用中	使用中 擴充中	發展中	使用中
6.重點敘述	目前尚無更新計劃	此計劃第一期作已於1987.6完成 目前等第二期計劃	日據時代資料已於1987.12完成。目前在擴充民國時代資料，預計1988.12完成。	資料輸入已完成，系統設計估計在1989.4月完成第二期工作。	系統單純 1988.6啓用
7.資料量估計	3040筆 400項/筆	五大項資料總共約佔五MB	戶籍資料： 18000筆 個人資料： 約20萬筆 事件資料： 約20萬筆	約25萬筆資料，每筆83項 總共15MB	總數約為20萬筆，每筆9項資料，均為數目字
8. 工作環境 機器 軟體	3B/2 UNIX INFORMIX C-ISAM 中文系統	(1)3B/15 (2)IBM5550 (1)UNIX INFORMIX C (2)用5550之多國語文系統	3B/15 UNIX INFORMIX C 中文系統	3B/15 UNIX INFORMIX C 中文系統	IBM PC/AT MS-DOS DBASE Ⅲ PLUS

表二、通用資料庫一覽

資料庫系統名稱	廿五史全文資料庫	電子卡系資料庫	電子辭典	台灣省博碩士論文資料庫	中研院研究人員著作資料庫
開始發展日期	1984.7	1986.9	1986.10	1987.1	1987.12
目前狀況	1988.6完成開放使用	1988.6完成開放使用	1988.4完成開放使用	1988.5完成開放使用	1988.6完成開放使用
資料狀況	目前有史記、漢書、後漢書、三國志、廿四史食貨志等，共計約 700萬字	處理使用者之私人卡片檔案，使用前要將已建立之卡片載入至系統中使用後則載出至碟片中帶回	約有4萬目詞	已建立1974」資料及1986博士論文23111筆，其中1980～81工學院資料有摘要。	已登錄6530筆，正在更新中
資料敘述	預計在1990.6完成全部廿五史的資料	每張卡片 1KB，其中包括出處、內容及八項檢索分類等13項資料	含國語日報詞典全部資料。對每一個詞另加詞類及語法分析資料，有：詞素、詞類、屬性、說明及對應指標等。	每筆資料有14項，均可供單項或組合方式查詢。資料尚陸續在建檔中。	每筆資料有十項，可供單項或組合方式查詢。目前資料蒐集尚不完整，在陸續更新之中。
功能敘述	有文獻結構查詢，自由詞查詢，字詞統計，及列印引得及原文之功能。並且可與電子卡系資料庫聯合使用。	卡片管理：傳輸、使用權更改、消除、列看、複製等。 卡片建立：包括自CPT 建檔與自行建檔。 查詢：分類查詢、表單式查詢及列印。	查詢、修改、新增、刪除等維護功能 自由詞檢索（在詞意說明中）及相關之統計、列表功能。	線上交談式查詢有邏輯表示及字串比對功能，另可列印各種資料	線上交談式查詢，可用游標依所別、作者、研究範疇、發表時間、著作別、關鍵詞查詢。此外作者姓名、研究範疇及關鍵字詞可用詞直接檢索，作者姓名及標題可用字串比對查詢
工作環境	3B/15, 中文系統, UNIX			C,3B/15, UNIX,INFORMIX/ESQL	

參、關於資料庫的設計與製作上的考慮

資料庫的價值依其所典藏的資料而定。若資料有怪古不易的價值，則此資料庫是傳世工作，其價值一如史料，所不同者只是記載的媒體改變而已。傳統的資料多為文書形態，轉換至資料庫中的資料是機器可以閱讀的電磁形態(以下簡稱「機讀」形態)。由於機讀形態的資料計算機可以處理，我們就能夠撰寫程式(如資料庫)來命令計算機幫我們做種種資料處理的事情。所以，資料庫的價值亦依其處理資料的功能而定。

決定製作資料庫之初，首需決定收集什麼樣的資料以及要這些資料做什麼樣的事情(資料庫的功能)。這兩件事都不是計算機工程人員能夠決定的。因此，在我們發展漢學資料庫時，都尊奉漢學學者為主持人，在發展過程中亦要求漢學學者積極的參與。畢竟，開發漢學資料庫是跨越科際的繁重建設工程，而漢學資料庫只是漢學研究的一種自動化工具而已。

本章以下，將就我們的經驗，舉出若干重要的理念，作為製作漢學資料庫工作之參考。

一、品質

當資訊由一種記載的媒體轉換至另一種媒體時，都會失真。失真的情況有三：一是會失去些原有的訊息，其次是在轉換過程中會引入些雜訊。當然，製作資料庫時理想的情況是沒有失真，然而這幾乎是不可能的。是故在製作之初，必需沿所有的原始資料，逐項檢討資料經轉換可能的失真情況，並訂定製作時對失真情形要求的規格。

研究用的資料庫，對失真的要求是非常嚴格的，否則就會影響到利用這些資料庫所做的研究工作的品質。譬如，對於文字記載的資料，研究人員的要求是一個字都不能錯，這種情形雖然很難做到，我們還是朝此目標努力。為達成此品質要求，校對的工作非常重要。以史籍資料庫而言，採三人五校制，即校對必需經過三個不同的人作五次校對。這使得校對的工作量約相當於輸入的工作量。

古書中有許多字的寫法是與目前的字不太相同的，這也會造成校對上的困擾。例如，古書「者」字中有一點，而現在的「標準字體」中卻無。

原始資料的版面格式在轉換後是無法完全保留的，充其量只能存其「神似」而已。在史籍資料庫中，保留的較多，無論在螢幕上或列印中，都保有原始資料中每行間的關係。而在電子詞典中，就略去了原字典的面貌，只將其資料呈現在螢幕上。

以上所舉的例子都是全文資料庫。由於全文資料庫處理文獻的原始全文，故失真狀況的考慮較多。對於欄位性質的資料庫，由於處理理的是表格化的資料，都是已經過人工整理的，故沒有版面失真的問題，然而每個欄位內記載的資料，仍然可能由於編碼、裁減、標準化等等過程而失去了些細微的寓意。例如，在地籍資料庫中，就曾發生這種現象，而地籍資料庫中每筆土地申告書的資料，就與原土地申告書中之資料不盡相同。此現象，在戶籍資料庫中亦有，然其程度不若前者。

只要研究用的資料不失真，資料庫仍是可信賴的、可用的。然而，製作資料庫工程浩大，絶少只為一個研究計劃的目標而設計，因為沒有幾個研究計劃在經費上允許如此奢華的投入。然而，預估以後研究工作可能需要用到那些資料並不是簡單的事，因此，我們在設計時的原則是：把原始資料視為史料，盡可能保留各種原有資訊，即使現在不用到也不輕言揚棄，以保持轉換後資料之完整性與真實性。換言之，製作漢學研究用的資料庫，工作人員應有治史的胸懷與氣度。

從反面來說，目前太多的工程人員是求急功近利的，這種態度不能苟同。他們把當時計劃中用不到的資訊悉數刪去，這麼做的結果是此資料庫對目前的計劃可能救一時之急，然而對其他研究計劃則將可能變為毫無價值；目前的計劃結束了，資料庫也壽終正寢，沒有再留下的必要。這種做法，一如搭蓋工寮，而不是建百年之計的宮室。長此以往，總共的投資會更大，效果反而降低。

二、資料之整理與登錄

以目前的設備來說，從鍵盤輸入中文資料仍是最可行的方式。然而，這種方式需用大量的人工，成為許多資料庫建立時的障礙。

為克服上述的障礙，本中心成立之初便組織了資料輸入小組，希以專業的安排來解決問題。有許多單位將資料輸入工作外包給專業的公司去做。然而，當我們考慮本院工作環境後，還是決定成立資料輸入小組。理由是這樣

的：

(1) 研究的資料特殊，有許多均為原始資料。若依外包的要求，先設計輸入格式，再用人將原始資料了解後填入輸入格式中，再予鍵入，則勢必花許多人力來建立輸入之表格。這個工作對許多研究計劃而言，是負担不起的。所以，安排具水準的資料登錄小組隨時向研究人員請教並做校對，是較可行的辦法。根據我們的經驗，對輸入人員作短期訓練後，直接輸入資料，而免除建立輸入格式的工作，是可行的。

(2) 漢學資料中有許多字是在目前計算機中没有的，在輸入過程中會產生些新造字，這些新字無法和本中心使用的計算機溝通，勢必重做一次。若自己輸入，則一次造字後在中心的計算機中可永久享用這些字。例如：根據我們的統計，在二十五史全文資料庫建立的途中，史記、漢書、後漢書、三國誌這四部書，只用了8322個字，然而卻有1300餘字在現有字集以外。這麼多自己造的字，將使外包輸入工作意外的困難。

(3) 中心為了推動全院辦公室業務之自動化，有些公文、信件、報告、論文、資料等等勢必無法樣樣外包，自己總要打的。

(4) 中心有責任訓練各所工作人員作中文輸入、文字處理、排版、及使用各種辦公室軟體系統之任務，因此，成立此小組可代訓相關人員。

基於上述的理由，我們成立了輸入小組。然而這個小組的成員並非只作中打的工作。他們在技術層面的工作甚多；譬如，他們應該會用些常用的編輯程式 (editors)，文字處理程式(word processors)，小型排版程式 (desk top publishing systems)， 資料傳輸程式 (transmission or networking softwares)等等。換言之，他們兼負有排版、編輯、校對、傳輸等等工作。

每位新進的資料登錄人員先要經過一個星期的中文輸入的訓練，然後做試用人員，試用期間為12週。試用期滿後，平均每人每分鐘可鍵入40字以上（倉頡輸入法）。 為了顧及工作人員的健康，每人每日在螢光幕前的工作以四小時為限(工作45分鐘休息15分鐘之間隔)。因此，每人每日約可輸入一萬字，此外還有些多餘時間做些校對或雜事。若每月以25工作天計，每人每月均可輸入25萬字。若考慮其他工作，包括校對，改錯、編排等等，其工作量約等於輸入的工作量。換言之，若組成十人小組，每月可以負吳120萬字的輸入、校對、改正、編排、傳輸等工作。

三、工程實務上的考慮

有了正確的理念和體認之後，精良的工程技術是保証品質的另一個關鍵。目前，一般的工程技術人員人文素養普遍不夠，他們不了解人文社會科學方面研究的方法和過程，更不必談研究的重點和精神了。尤有甚者，他們對人文的價值觀相當淡漠，在雙方溝通上會產生極大的隔閡。曾有過這麼個例子：在發展某資料庫的初期，有位負責的工程人員和主持計劃的研究人員談得不歡而散。工程人員認為研究人員的要求是没有價值的莫明其妙，而研究人員則認為該工程人員簡直是不學無術的黑手黨。事後，換了位工程人員，重新來過，終於雨過天晴，事情也進行的順利成功。那位出問題的工程人員，坦白講，從工程角度來說是很優秀的，甚至比後者強，然而，卻無法與漢學者達到共識。所以，發展漢學研究用資料庫，在工程人員的遴選上，要重視其人文素養。反過來說，若漢學學者能夠了解些計算機或資料庫，對雙方合作之事亦有助益。

工程人員的人文素養不夠還會發生另一種問題，那就是在工程實務過程中，會扭曲資訊的內容，遺漏資訊的內涵，並且無法充分利用計算機的性能，把資料庫設計好。關鍵之處在設計計算機內之資料結構（最易使資料失真處），檔案結構，或是資料庫之檢索方式。在設計之初，使雙方充份溝通，互相學習，完全了解所做的事，似乎是避免不幸後果唯一的途徑。否則系統會出現一改再改的情形，最重時，甚至需要全部重新來過。所以，不只是主其事的人員，凡是參與工作的人，均應加強溝通。

資料結構是構成資料庫的根本，不可心存「以後再改」的想法。改資料庫的資料結構就是改整個資料庫，什麼都要從頭來過。關於應用程式，多附屬在資料庫上，改變他不會搖動資料庫的本體，所以應用程式的改變，遠没有改資料結構嚴重。換而言之，資料庫可以有修改或增加應用程式的彈性，這點對於研究而言，是很好的。

肆、關於資料庫的運作上的一些問題

當資料庫開始使用之後，便進入了運作(operation)的階段。在運作階段裡，最主要的問題是應付一些「改變」：大的改變如機器改變機種與型號，操作系統或資料庫軟體版本的更新，或是功能之增加等。而日常的改變如，資料的增加、刪除、更新等。以上這些工作大多可納入「維護」的工作範圍，它們分別包括在機器、軟體及資料的維護工作內。關於屬於計算機專業的

維護問題，限於篇幅在此從略。在本章以下，我們將分別討論:(1)關於中文語文的問題，這是涉及系統軟體和硬體維護的問題,(2)資料之取得與安全的問題，這是資料維護的問題，和(3)資料庫在管理上的一些問題。

一、中文語文的問題

目前市面上的中文系統，雖曰有國家標準 :「中文通用標準交換碼」然而，這個標準卻不是為人文社會學科設計的。 它只是一個「商用」貨品的標準，處理一些訂單、收據之流的事尚可，若是用以處理國家級的檔案，如警政、戶政等等系統，便會產生嚴重的字數不足現象，更不必談用以處理數千年文化資料了。因此，在製作漢學用資料庫時，計算機技術上必須要能克服中文計算機產品中字數不足的問題。

那麼，目前的產品究竟會缺少多少字? 嚴重到什麼程度呢? 根據我們的經驗，可用下例說明。在製作廿五史全文資料庫時，曾做過如表三中使用字數的統計。

表三、全文資料庫用字統計

書名	總共字數	使用字集字數	欠缺字數
史記	1,334,359	6,280	595
漢書	1,855,857	6,752	802
廿四史食貨志	539,112	4,466	248
小　計	3,729,328	8,322	1,233

在此採用的系統是宏碁公司的天龍中文系統，共計有一萬八千字，這個字集包涵了「通用標準交換碼」的一萬三千餘字，且較之已多出近五千字。然而，竟有1233字不在字集之中，若以原書所用的字集8322字計算，約有近14.8%的字是系統無法提供的。這是相當嚴重的現象。

為解決此問題，臨時的辦法是要求廠商儘量提供 「使用人造字」的空間。天龍系統可提供2000字，雖然目前勉強可用，然而我們認為造字的空間還是不夠。可是宏碁表示，這已經是目前天龍系統的極限了，無法再增加。即使可再加造字空間，也不是永久解決的方法，因為一旦造字，就使系統難與他人分享資源，再者，系統本身亦產生維護上困難。譬如，凡造一新字時，必須更新所有的終端機內的字形檔案和造字的輸入碼。因此，徹底解決的

辦法，還是要採用另一更大的字集，如美國的國家標準 EACC 碼或是中文資訊交換碼 (CCCII)。好在目前市面上已有幾種CCCII終端機應市，亦有PC上可以模擬 CCCII 終端機的軟體。相信這個問題，近期內可望解決。

另一個與語文相關的問題是資料中有多國語文混合的問題。例如：圖書系統面對著中、日、英....等等多種文字的典藏；又如研究人員發表著作時，時而中文，時而外文。這些系統，至少是中英文混合，若加上日文或其他語文，則更形複雜。因此，這些資料庫需要一個能支援多國語文處理的環境(如操作系統或資料庫管理系統)。目前市面上的系統，可支援部份中英文混合資料處理的功能，有些可以支援英日文，可是沒有一個系統能同時支援混合中、日、英及其他外語的資料處理。因此，在檢索、分類、排序、比對等等處理工作都會發生問題。這些問題不是不能解決，而是目前沒有現成的軟體可處理這些問題，必需要自己設計程式來處理這些事。例如，在書目或全文系統中要自己建立中英文的權威控制，就是一個好的例子。這樣的情況是非常不經濟的，也影響到資源的分享。因為上述的這些問題，都是計算機系統應該提供的，不必要每個系統自己發展。要解決這個問題可能還要一段時日，要等商用的多語言的操作系統發展成熟，並且，一些常用的軟體工具，如資料庫管理系統(dbms)，亦能處理多國語文資料時，方能徹底解決。

此外，目前中英混合的系統發展得還不是很成熟，經常會產生一些毛病(bugs)而使系統無法工作。這現象最常發生在資料傳輸時，其次是應用程式裡，偶爾在系統程式中也有。發生此現象時，只好重新來過，並請廠商設法解決。

二、資料維護

有些資料庫的資料是不允許使用人更改的，如廿五史全文資料庫。這種資料庫資料維護的工作是由系統管理人員以成批作業方式更新，問題較少。然而，有些研究用的資料庫，蒐集的是研究採集的資料或研究的結果，如台灣土著語言資料庫；清代內閣大庫索隱等，這類資料庫更新的資料並非是適合成批處理的大批資料，而是每日工作的累積，適合由研究室中以線上作業的方式更新。然而這種方式操作會使原資料庫呈現無保護狀態，出毛病的機會很大。譬如：操作時一不小心很容易造成已有資料的損失，亦可能不小心改錯了資料，類似的問題很多，會造成相當大的困擾。要做好資料線上更新時對原資料的保護不是件容易的事。 據我們的經驗，最好是不要做線上資料更新,如無必要，仍以定時成批作業為宜，不只較為安全，亦較易控制資料的品質(可以經過嚴格校對後，再成批更新)。

有些資料庫資料的來源是可以完全控制的，如漢代墓葬資料庫等。這些資料庫的更新問題不大。可是，有些資料庫的資料蒐集卻難以控制，如中研院研究人員著作資料庫等是，通常，研究人員發表了著作，沒有直接的管道告訴本中心，因此，非等到下個事件發生問題時，才能知道。雖然，定期的追討資料是通用的方法，可是依我們的經驗是效果極差！若是沒有行政當局強有力的要求和鞭策，縱使有些誘因，例如免費提供列印個人履歷或研究著作清單的服務等，也是沒有顯著的功用。在推動辦公室自動化的研究資料中顯示，若無辦公室主管的強烈支持（不止是精神上的，必須要有實際的行動），辦公室自動化的努力極可能失敗。這類資料庫的情形也正是如此，因為資料的蒐集涉及到辦公室中（研究室中）主要的成員。所以這一類資料庫的發展，極需注意解決資料維護的問題，否則就可能像博碩士論文資料庫一樣，至今尚欠缺近三年的資料無法補齊，嚴重影響實用的效果。

三、使用權的問題

資料像物質、人力、土地、時間、金錢、和能源等等一樣，是一種社會的資源。我們的社會還沒有建立任何分享資訊的法案，然而建立資料庫的人，缶必須面對這個問題。建立資料庫，首先應取得原始資料的版權或著作權，往後的問題有：誰擁有建好的資料庫？誰有權使用這些資料庫的資料？使用有沒有規範或限制？要不要收些費用?收的費用如何處理?.....等等一大籮筐的問題。 在沒有法規，沒有經驗的情形下，目前只能用一般的學術道德規範和良好行為的常識來處理這些事情。

目前，我們的做法是：避開著作權問題，不做商業行為，並盡可能地免費開放資料庫給學術研究或教學之用。這樣做法會減少許多行政上的困擾，然而並沒有解決問題。

建立一個資料庫的價值，就像是著書立說,其成果應該公諸於世,讓大家有機會分享。然而，有些資料庫的資料是私人蒐集發展的，經費又受到某單位的補助，這種情況就使得上述的關係更加複雜。資料庫建好了，主持計劃的人當然不希望他人搶著來分享，此間涉及利益的衝突，是很易理解的。然而，由長遠的社會利益看來，卻浪費了許多寶貴的資訊資源以及時效，延緩了「明天會更好」的發展腳步。

我們認為，應該建立一個共識：開發一研究用的資料庫，就如同發表一篇著作。如果此資料庫完全由個人開發，那麼他當然百分之百的擁有它，其

它人要用，應該付出合理的代價，就像買書要花錢一樣。若是資料庫的建立是經過許多人或單位的合作，那麼，應該研擬一套分配所有權的方法，像股份有限公司一樣，共同擁有它。對於要使用資料庫的人而言，應該沒有任何歧視的限制，就像是賣書不應限制對象一樣的開放，只要付出代價，就可使用它。

以上的想法，或許太理想化。執行起來可能會遇到各方的阻力，然而，我們認為這想法是合乎學術和社會道德的。在此特別呼籲大家重視這個問題，共同來建立有關的方法、規範和法律，以使我們人人都可分享所有資料庫的成果。

伍、使用者發生的困難

到目前，在本院使用資料庫的人，絕大部份是主持開發此資料庫計畫的研究人員。例外的情形有：成功大學黃競新副教授曾來本院利用廿五史全文資料庫協助作「殷商天文氣候彙考」及「殷商季風氣候彙考」二個研究計劃中查詢相關的記載資料；又如本院美國文化研究周碧娥研究員擬利用戶籍資料處理系統來協助日據時代「婦女地位與經濟發展」的研究計劃。像這樣的例子不多。使用率的偏低原因可能很多，在本章中，將由使用者的角度來檢討此一現象。

一、溝通的問題

由於本中心的這些資料庫都是最近才開放使用，在缺乏溝通管道的情形下，也許有許多人根本不知道有這回事。雖然，本中心有對院內發行的雙週刊通訊，但是這種程度的溝通顯然效果不夠。因此，如何加強這些科技新聞、科技資料對研究人員的溝通是當務之急。

二、使用人的困難

根據辦公室自動化研究的文獻顯示，推動使用計算機系統常會遇到下列的困難：

1、心理障礙

.不願學

. 不敢學
. 不屑學
. 不習慣用

2、行動上的障礙
. 系統太複雜
. 系統不好用，功能不夠
. 使用人計算機素養不夠
. 使用人操作技術不良

3、其他障礙
. 地區上的不方便
. 說明書寫的不好(技術文獻、使用文獻不良)
. 技術上有困難，無處求助
. 機器設備不良

這些問題我們都遇到過。也就是說，使用人所產生的困難和辦公室自動化系統遇到的情形是如出一轍。以上的這些問題，不是不能解決，而是要解決這些問題，必需配合行政、教育訓練、技術溝通、計算機技術以及技術性的服務等等各方面作長期的努力。本中心成立之初，即特別強調注意這些問題，然而囿於預算及員額，所能做的顯然不夠應付使用者的需要。依目前情勢看來，若在行政支援上無法突破此困境，只好以時間來換取品質與成效了。

陸、漢學研究的新工具

在以上的討論裡，我們將注意力集中在利用資料庫為自動化工具，來處理漢學資料的論點上。在這一章裡，我們將擴大視野；從漢學研究的資料、工具、方法等角度，來探討計算機科技對漢學研究可能的幫助。

全文處理技術的發展對漢學研究而言，是一項關鍵性的技術。在此之前，計算機多處理表格化的資料，對於漢學原始資料的處理並沒有適當的工具。然而，漢學資料能夠列成表格形式的不多，而且列表後會失去些訊息。因是之故，計算機在未能處理全文資料之前，對漢學研究的貢獻並不是很吸引人的。

除了文字敘述的原文以外，漢學資料的形成還有：表格、圖形、影像、語音等等的變化。目前，從計算機技術的角度來說，處理各型資料的技術都已達可以應用的階段。譬如：處理圖形影像的資料庫可用來處理甲骨文、各種拓本、繒本書或原始史籍檔案。這方面的應用實例目前雖然不多，然而這些技術將發展為能處理多媒體的漢學研究工具是可以預期的。如此將更擴大計算機在漢學研究上的領域。換言之，隨著計算機技術的進步及計算機價格的下降，計算機將可用來處理各種媒體表達的漢學資料。

在漢學的文史研究中用到許多工具書。在前文所舉的例子中，電子辭典、論文和著作資料庫等，可視為查考詞句和查考書籍、典籍等工具書的自動化工具。依此類推，目前計算機技術可以將許多文史研究用的工具書自動化。譬如：查考史事與人物的表、年表、曆表、年譜等；查考地理及歷代地名的方志、地理表、地圖等；查考典故的各種類書；查考政制的通志避諱書等，亦可為自動化的對象。

以上這些工具書的自動化，在觀念上，多半還是依循研究人員對這些工具書使用的傳統「方式」，如查詢、比對資料等等，作自動化的工具。若是我們更進一層，把這些工具書相關的知識，利用人工智慧技術製作為各種專業的知識庫，讓研究的人可以隨時獲得相關的知識來幫助研究，那就更提升了自動化的層面。譬如，若我們將曆法知識構成曆法的知識庫，就不止可查閱「某日相當於西曆那一年那一天」，「某皇帝有那些年號」，「在某時有多少政權？」等等問題，而是可以回答一些更具知識性的問題，譬如：「儒略曆是如何發展的？」，「十九年七閏是怎麼算的？」「某時考據的歷法誤差是怎麼算出來的？」等等問題。

工具書的自動化、傳統研究方法中部份工作的自動化、以及利用人工智慧發展智慧型的新工具－－知識庫，這些日後發展的必然趨勢。循此發展，當研究用的資料、工具，及一些研究例行工作，都變成電腦中的資料庫，應用軟體，和知識庫以後，將之整合，就可成為一個完整的文史研究工作站(work station)。這種發展的軌跡一如發展工程設計時用的工作站(CAD或CAE work station)，人文研究工作站可以提供一個相當完整且獨立的漢學研究環境。對漢學研究的工作將有劃時代的貢獻。再者，若利用通訊設備，更可將許多工作站和其他的電腦系統組成網路，充分達到研究工作的自動化、研究資源的共享、和研究工作充份溝通的更佳境界。

柒、建議與結語

利用資料庫和相關的計算機技術來助長漢學研究，在國內是新的嚐試。根據本文的經驗，可以肯定資料庫對漢學研究的價值。然而，在華路藍縷之中，遭遇到的問題也很多。這些問題在前文中雖已略有敘述，且讓我們強調其重點，歸納為重要的建議，並為本文總結。

一、資料庫的價值就如同圖書和檔案一樣，其功能卻比圖書和檔案高出甚多。就所存的資料而言，資料庫是圖書和檔案的另一種版本，所不同的只是儲存的媒體而已。對於從事學術研究的學者而言，利用一個好的資料庫和利用原始素材具同樣的價值。資料庫亦如圖書和檔案一般，應該有良好的典藏制度永久保存，而且累積得越多，越有價值，越能提供好的研究環境。由另一角度來說，上述的這些觀念是推動漢學研究用資料庫的原動力。只有你我都有此體認，才有可能加速建立更多更好的資料庫，供我們分享。

二、將研究用的原始資料有系統的整理好，轉換成為機讀形態，是本文提到所有工具發展的基礎和必須的工作。這種工作是研究工作中重要的一環。我想，任何一位對研究工作有體認的人，都同意這論點。然而，我們卻遇到許多掌管研究的「官」，他們不認為資料的整理和轉換為機讀形式是研究工作的一部份。他們不支持這種工作，對中文研究的斬伐毀傷無以復加。這種「官」，根本不了解什麼是「研究」，卻掌理訂定國家研究的政策，和執行資助學術研究的令符，寧不以為怪事？改善這種不合理的狀態，只有靠大家爭取了。

三、資料庫的使用權應開放。(詳見本文第四章，第三節)。

四、發展資料庫宜採高品質的路線。所蒐集的資料在轉換為機讀形式時，宜注意其完整性與真實性。對於計算機使用的中文字集，現有的國家標準不足以為人文社會學科之用。應該規劃更完備的國家標準，勿使在社會進化之中傷害我們的語言文字和文化資產。

五、做好溝通工作是推廣使用資料庫或自動化工具的基礎。此所謂之溝通包括一般行政訊息的溝通和科學技術的溝通(science andtechnique communication)。後者的含意是廣義的，任何傳播科技知識的行為都包涵在內。

以上本文所呈現的，只是我們工作的經驗之談，歡迎批評指教，更盼對讀者有所助益。

誌　謝

本文感謝本中心的工作同仁們多年發展資料庫的合作與努力，使得本文得以寫成。亦感謝提供各資料庫詳細資料的林聯盟、劉忠全、曾士熊、杭極敬、何惠安、石玉祥、張孟元、李淑玲等各位小姐先生。又本文承顧秋芬小姐的行政協助與資料整理，以及安排打字排版，在此一併致謝意。對於和我們攜手合作的各位漢學研究先進，更是銘感五內，沒有他們的信任和指導，不可能有今日之成果。

參考資料

此處所引之參考資料，為本文介紹的各個資料庫的有關資料。由於文中參考之處甚多，故在本文中不列索引，敬希查照。

一、杭極敬著，漢代古墓資料處理簡介，
中央研究院計算中心通訊，第02卷第09期，p.39-p.40，1986年5月1日。

二、莊英章著，出生排行序與家庭結構動態----三個鄉鎮的比較研究，
中央研究院計算中心通訊，第03卷第01期，p.4-p.5，1987年1月1日。

三、林聯盟、王芳華、黃永坤等合著，PROBLEM SOLVING FOR ANALYSIS AND REASONING ABOUT ANTHROPOLOGY,
1988 ICCPCOL, Toronto, Canada. Aug. 29-Sep. 1, 1988.

四、張炎憲著，日據時代土地申告書資料的自動化，
中央研究院計算中心通訊，第04卷第13期，p.99-p.100，1988年7月1日。

五、杭極敬著，碩博士論文查詢系統簡介，
中央研究院計算中心通訊，第04卷第11期，p.84，1988年6月1日。

六、石玉祥著，本院研究人員著作查詢系統，
中央研究院計算中心通訊，第04卷第11期，p.85-p.86，1988年6月1日。

七、李壬癸著，台灣土著語言資料自動化的工作成果，
中央研究院計算中心通訊，第03卷第17期，p.16-p.120，1987年9月1日。

八、中央研究院歷史語言研究所與計算中心合作，文建會資助的史籍自動化計畫第一、二、三年工作報告。

九、謝清俊、丁之侃、談國蔭等合著，A NOTE-CARD SYSTEM WITH FULL-TEXT DATABASES,
ASIS'88 Atlanta, Georgia, Oct. 23-27, 1988.

十、曾士熊著，簡介本院公眾資料庫系統，
中央研究院計算中心通訊，第02卷第23期，p.99-p.100，1986年12月1日。

十一、曾士熊、張孟元、謝清俊、陳克健等合著,CED-A MACHINE READABLE CHINESE DICTIONARY,
First Pacific Conference New Information Technology For Library Information Professionals. Bangkok, June 16-18, 1987.

十二、謝清俊、丁之侃、王苑華、童淑芬、林晰、舒啓洲等合著，FULL-TEXT DATABASE FOR CHINESE HISTORY DOCUMENT,
1988 ICCPCOL, Toronto, Canada. Aug. 29-Sep. 1, 1988.

十三、曾士熊、楊鍵樵、謝清俊等合著，THE DOCUMENT REPRESENTATION AND A REFINED CHARACTER INVERSION METHOD FOR CHINESE TEXTUAL DATABASE,
1988 ICCPCOL, Toronto, Canada. Aug. 29-Sep. 1, 1988.

十四、曾士熊、張孟元、謝清俊、陳克健等合著，APPROACHES ON AN EXPERIMENTAL CHINESE ELECTRONIC DICTIONARY,
1988 ICCPCOL, Toronto, Canada. Aug. 29-Sep. 1, 1988.

十五、陳克健、查全淑等合著，THE DESIGN OF A CONCEPTUAL STRUCTURE AND ITS RELATION TO THE PARSING OF CHINESE SENTENCES,
1988 ICCPCOL, Toronto, Canada. Aug. 29-Sep. 1, 1988.

註：ICCPCOL:INTERNATIONAL CONFERENCE ON COMPUTER PROCESSING OF CHINESE AND ORIENTAL