史籍自動化計畫
中文資訊處理部份第二期 研究報告

中文全文處理系統的設計與製作

研究人員：謝清俊 丁之侃 王苑華
葉權榮 舒啓洲 童淑芬
林 晰 王芳華

中華民國 七十五 年 九 月
中央研究院 計算中心

第一章　中文全文處理系統簡介

史籍自動化計畫是一個長期的學術研究計畫，其目的在探索計算機應用於文史工作的可行性。在此計畫之前，國內尚無類似的計畫。所以，這計畫算是一種新的嚐試。

此計畫始於民國七十三年七月，以一年為一期。在第一期中，我們做了個很好的示範，將部份食貨志的資料以全文形式建立了機讀檔案，並裝配了相當完備的鍵語查詢功能。詳細的報告請參閱<史籍自動化食貨志輸入電腦第一年總報告>〔1〕。

本文是史籍自動化的第二期報告。在此期中，主要的工作是將正史(廿四史)全部的食貨誌建立完整的檔案，並將第一期的系統作些改進，包括：

(一)建立文獻自由詞查詢(free-term search)的系統，並與第一期所採用的控制詞彙查詢(Controlled vocalurary search)作一比較。

(二)建立文獻處理之邏輯模式。在第一期中，處理文獻的程式是把文獻的一些內容和處理的步驟混合在一起寫的。因此，這種程式結構會使得程式依靠該原始文獻，也就是說，對今後不同的文獻，程式必須重寫。這樣的做法將使處理文獻的成本和各種資源的投入不堪負荷。在本期中，我們將文獻結構的共同部份，經實驗分離出來，將之以抽象的邏輯架構寫成程式。如此一來，處理文獻的各種步驟，就可依照此邏輯架構進行，不必再對原文獻產生依類性。這樣的做法，可使得結構類似的文獻享用同一個程式。可大幅降低文獻處理的成本和投入的各種資源。

(三)在上一期中，無目錄查詢，本期中增此項目。

(四)在上一期中，所有的系統在個人電腦中。因此，有下列缺點：

- 反應速率慢
- 容量小
- 只可供一人使用

在本期中，我們將之改變到32位元的工作站上，除改善了它的反應速率和容量外，同時可供八人共同使用，並可透過通訊電路，接裝遠程查詢終端機。

(五)在上一期中，無應用的程式。換言之，當資料查到了，就已結束。因為全部文獻已在

計算機中，若不將之充份利用，十分可惜。故在本期中，將作一示範性質之卡系系統，以為下一期之參考。

(六)訂立了文獻登錄的全文標誌規則，對文獻登錄工作有統一及制式化的效果。

(七)建立了文獻在電腦中表達方式的模式及其相關的處理工作程序，包括：

- 文獻主要元素之認別方法
- 文獻結構化程序
- 由文獻結構做查詢工作的程序。
- 以自由詞查詢為基礎的控制詞彙查詢方式。
- 各種人機交往的畫面管理方式，使得此畫面較第一期更具彈性。

本期的工作，雖較第一期有上述之改進，然而仍有未盡理想之處。例如：報表及列印系統並非完善，人與機器交往上並非盡美等。這些問題，並無學術上或技術上之困難，只是工作量較大較煩雜而已。在設計一個真正實用的系統時，以上問題均可改進。

在此計劃中，有關計算機的工作分配如下：

- 丁之侃、王苑華：將第一期之成果改在 MICRO-VAX-II 機器上，另加英文的介面和自由詞查詢功能。此外並協調此計劃之輸入與校對工作，參與設計之討論。(本文第二章)
- 舒啓洲、童淑芬：建檔模式(本文第三章)與文獻結構模式(本文第四章)之設計。
- 林　晰：人機介面與查詢檢索系統之設計(本文第五章)。
- 葉權榮：各種報表之產生程式，包括正文列印、索引表、目錄表等；並協助林晰寫螢幕控制部份之程式。
- 王芳華：卡系部份(本文第六章)。
- 謝娟娟：使用說明書之撰寫(另冊)，輸入打字及校對工作之安排與督導。

此外，協助打字及校對的小姐們有：謝娟娟、嚴婉如、黃燕勤、李淑玲、阮可卿、林惠英、曾維芳、陳芊羽等八位，藉此一角表仲謝之意。在行政工作方面，顧秋芬小姐、顏純真小姐幫助頗多，使得工作進行順利亦值一記也。

壹、系統概述

設計計算機裡的中文全文處理系統是本計劃的重點之一。此系統分為建檔作業和全文處理作業兩大部份。建檔作業的目的是將原始文獻輸入計算機，使之成機讀檔案，然後再分析其文獻結構，將之轉換為結構化的全文檔案以供全文處理之需。所以這一部份可說是建立系統的準備工作，而它所產生的全文結構檔案群是可以供大家共享的。

全文處理部份是使用者的主要伙伴，可獨立作業，包括查詢系統以及應用程式兩部份。查詢系統可提供①文獻結構，②自由詞，以及③控制詞等三種查詢方式。目前的應用程式部份有列印以及一個示範性質的卡系系統。

中文全文處理系統的結構方塊圖如圖1.1與圖 1.2。以下各節當對各部份分別說明。

一、建檔作業

建檔作業分為二部份。一為登錄作業，其目的在將原始文獻依一定的標誌規則鍵入計算機中建立機讀式的原始文獻檔案，如圖 1.1。

全文標誌規則是本期所發展出來的成品，請參考本章第參節。在 ISO 的標準中，有所謂的標誌語言(Mark-up Language)〔2〕。標誌規則之目的與標誌語言者略同：都是在資料輸入時期將原文的結構部份加以標誌。這樣做的目的是使得文獻結構分析時可以由原始文獻檔案中認出主要的文獻元素(以下稱文素，見本章第貳節定義部份)，而使文獻能完全結構化。

由於在我國正史資料中無英文字母及符號，所以在中文資訊交換碼未在計算機中建立以前，暫用英文字母及符號代替文獻控制用之符號。目前的標誌規則較 ISO 之標誌語言簡單甚多。然而，它已可將食貨志中所有的文素完全識別出來了。

建檔作業的第二部份是將原始文獻檔案結構化。所謂結構化，就是將原始文獻中所含的資訊依其特性加以區分並予以系統化。在一般文獻中的資訊可分為三類：

(一)文書資訊

原始文獻中文字、圖形、表格、或公式等之內容部份所含的資訊稱為文書資訊。其構成之檔案稱為文書檔案(請參閱第貳節之定義)。文書檔案又分為：文書字串檔案、圖表檔案、以及文書結構檔案等三部份，請參閱圖 1.1。此部份與原始文獻之編排及其表達媒體無關。

文書字串檔案中只含有原始文獻之字串以及在一段內之必要控制符號。在本系統中，最小的文素是段落。是故在段落內之結構或索引不再預先分析。當查詢時，在現場再以程式由段落中分析出所欲得之資料。這是較經濟有效的做法〔3〕。

圖表檔案中包括文獻中的圖形、表格、及公式等資料。由於本期內無此類資料，故此部份未予設計，只暫留有擴充之餘地。

文書結構檔是本系統設計的中心思想之一。它是一個抽象的樹狀結構，再以附著屬性的描述來表達文獻的實體結構部份。如此一來，此程式就可適用於類似結構的文獻，而不局限於食貨志。

在此系統中，凡目錄和正文之查詢，以及蒐尋範圍之設定，甚至於鍵語之索引指標均以此文書結構檔為對象。所以，文書結構檔案是本查詢系統之樞紐。

(二)編排有關的資訊

這部份資訊集合起來就構成了圖 1.1 中的編排檔案，此檔案中包含的是原始文獻編排的規格，例如有：

①橫向或直向之編排
②頁次
③字形
④字體
⑤留空(空字、空行、空頁、留空格)，等等

在目前的系統中，只有做①和②兩項，餘容後繼。

(三)書目與其他屬性資訊

這部份資訊為一般圖書系統中常運用者，它可由標誌方法或經由語文處理自原文中取出。在本期中，由於未用到此部份，故未予設計，只留下擴充之餘地。

由以上之敘述可知，在本期中文獻結構化處理已做的是將原始文獻中的字串部份、文書結構部份、與編排結構部份予以認別、分離，並加以系統化。關於認別的作法，是將文書結構之特徵－－以全部食貨志為對象－－以制式的ＢＮＦ描述法表示之。再根據此ＢＮＦ之文法寫成編譯程式，來認別所有的文素和文書架構。關於食貨志之ＢＮＦ表示請閱表三，而編譯器之設計請參閱第四章文獻結構模式。此部份的設計是相當成功的。

在資料登錄作業時，是分別以頁為一單元，又在文獻顯現或列印時，其橫直安排可隨使用者選定(若不選定則取橫式)，是故目前編排檔案之製做並不困難。

關於此三檔案之資料結構詳如第三章建檔模式。在第三章中，關於存在磁碟中之檔案及置於主記憶體中之檔案均有詳細之描述。

二、全文處理作業

這一部份是可以獨立作業的，因為凡經過建檔作業部份製作的機讀式結構化文獻檔案可以被其他的全文處理作業分享。能夠這麼做的理由主要是建檔作業的標準化與結構化所致。目前的全文處理作業包括三大部份，請參閱圖 1.2。茲將各部份分述如下。

(一)系統監理及執行程式

第一個是一個系統監理及執行程式。它是管理全文處理系統的主宰。它包括一個命令解譯程式(Command Intrepreter)和人機介面相連接。當使用者透過人機介面下達操作命令時，命令即經此解譯程式分析、分解、而轉換成由執行程式可以選擇之巨集程式(Macro Calls)來執行。監理程式還負責系統之控制以及任務分派之工作。

(二)查詢系統

第二部份是全文查詢的系統。此系統提供文書結構、自由詞、及控制詞(或稱鍵語)等

三種查詢的方式。文書結構查詢是經由文書結構檔及編排檔直接查詢。文書結構檔將提供所查文獻之位址，再經由文書字串檔內讀出原文資料。自由詞查詢(free-term Serach)是指使用者可任意指定一詞或一串詞作為查詢的鍵語。在此情形下，系統將查詢範圍內之各段原文讀入工作檔中，再用匹配程式去查閱該自由詞。鍵語(Key word)查詢的途徑和自由語不同。它將以鍵語為鑰，到索引檔中取得地址去讀出原文資料。

為了節省資源以及避免錯誤之查詢命令，系統設計有限制查詢範圍之功能(Proximity Functions)。限制的方式可經由文書結構檔限制卷、冊、章、節等，亦可經由編排檔限制頁次，或以二者之組合為之。此外，此程式尚可於來執行查詢之前估計查詢所需之時間，給使用者在查詢前再作一選擇。

這是第一次以文書結構為主體的方式所設計的查詢作業。理論上雖然很優美，但是在此期中由於時間之限制，功能設想或有不週之處。當可於下期中改進之。

(三)應用程式

第三部份是應用程式部份。應用程式可有很多，存在應用程式館中。理論上來說，查詢與應用是可以密切結合的，而結合的方式應該採用共存程式(Co-routine)之形態。然而在本期中，應用程式只是示範性質，故取其簡易方法，而以檢索系統之副程式形態設計。這樣做法是有缺點的，就是不能自由地往復於查詢與應用二者之間。此點當於下期中再改進。

本期示範之應用程式是卡系系統。也就是一個電子化的讀書研究用的卡系(Note-card system)。電子卡系比人工卡系更有彈性。譬如卡片之大小、欄位之大小、索引之多寡及種類，均可不必受紙張大小之限制。此外，計算機裡的索引及查詢較卡片有更優越的性能，這些均使電子卡系更顯得有效用。目前之電子卡系可以將查詢到的文獻部份原文自動地抄到卡片上，並加註出處，更可將原文之索引，自動移到卡片上。還可任由使用人加上心得及私人用之索引，在第六章裡有卡系的示範和一些卡系的例子。

除上述的各項系統以外，尚有一些系統公用程式。它們是列印程式和螢幕的映像處理程式。此系統的人機介面是採用選單式(Manue)的，對於使用者而言，這是打入中文最少的方式。在使用手冊中，有許多查詢的例題，並有詳細的操作說明。

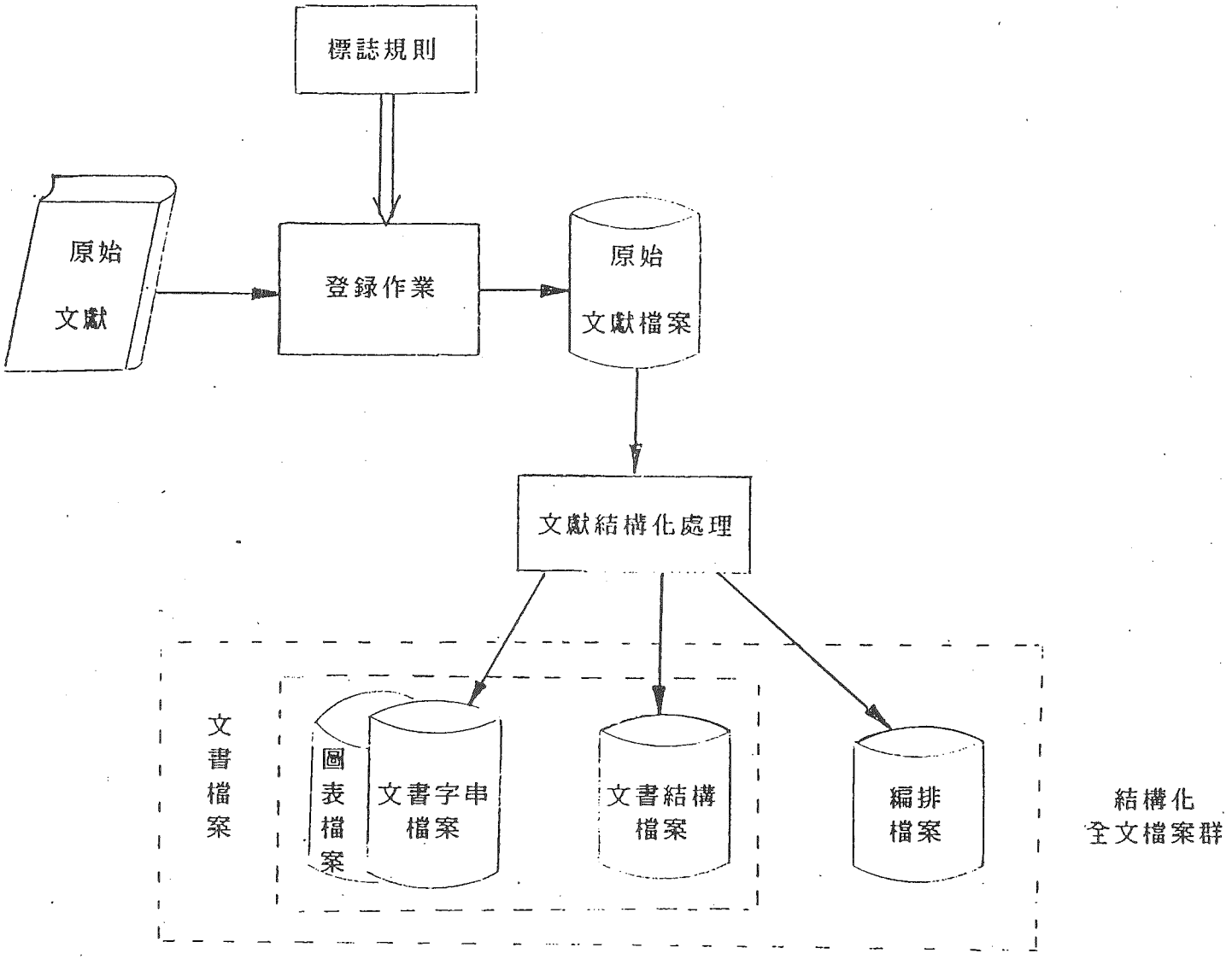

圖 1.1　全文處理系統之建檔作業部份方塊圖

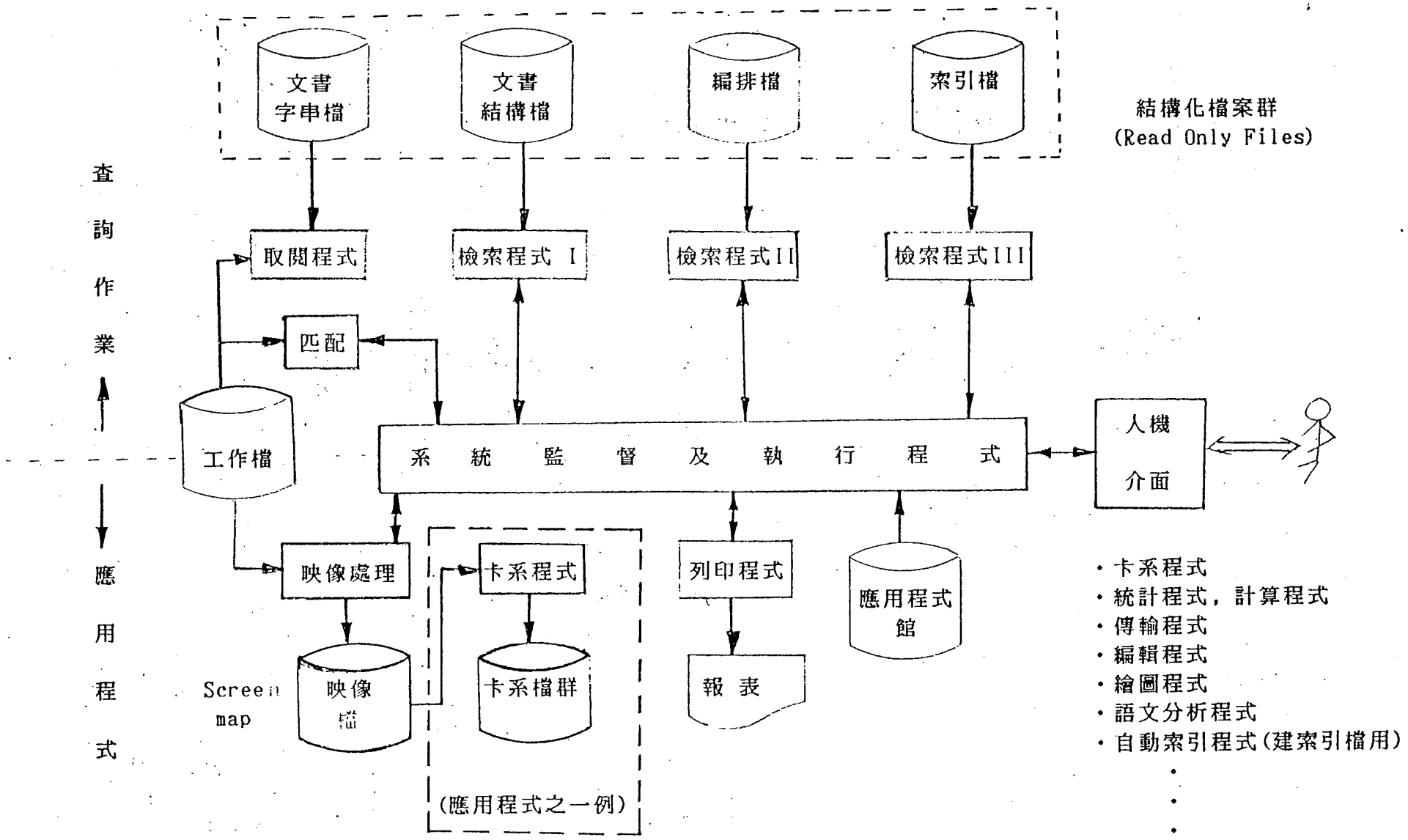

圖 1.2　全文處理系統之查詢及應用部份結構方塊圖

貳、定 義

為了釐清名詞可能產生的混淆，在本研究中將重要名詞予以嚴謹之定義。

定義一：全文(text)

根據國際標準組織(ISO)文獻，其定義的原文為：

Text is all information for human comprehension that is intended for presentation in a two-dimensional form. Text consists of symbols, phrases , or sentences in natual or artificial languages, figures, formulas and tables [1, p.13]

今取此定義，並根據〔2〕，將圖案(figures)引申為包括一切以「圖形」表達之資料，如：圖畫(pictures)、插圖(illustration)、照片(photograph)等，以符實用之需。經譯後之定義為：

全文是以自然語言或人工語言中的符號(symbols)，片語(phrases)，和句子(sentences)，以及圖形(仍用 figures 一詞)，公式(formalas)及表格(tables)等所組成的二度空間的形式。它用以表達任何人類能理解的資訊(information)。

此定義之前半段說明全文的組成，後半段則為其功能。

text 之譯文亦經考慮。text 之原意有正文、本文、或原文之義，但它並不一定包含圖形、公式、表格等在內。ISO 對 text 之定義是指 full-text (亦譯作全文)。是故在一般文獻中之 text 與 full-text 兩詞常有交替運用之情況。在本定義中，為避免混淆，將 text 譯為全文，以求其文義一致耳。

定義二：文素(text element)

文素是指全文中的一部份實體(entity)。文素是以它的功能來界定的，不以它的形式或內容來定義。為稱呼上的便利，以語文表現的文素稱為自然文素，而經過整理後之形式如：圖形、公式、表格等稱為人工文素。又，在定義一中指明之構成元素稱之為基本文素(basic text elements)，以基本文素組成的更高層次元素像段、節、

章、標題、跋、序等，用以表現全文結構稱之為結構文素(architectonic text elements)。

在譯詞考慮方面，雖然 element 有單位之含意，亦有將之譯為「位」者，如語位，然而，將 text element 譯為「文位」恐易生意義上之混淆，故選用文素一詞作為譯名。

定義三：文書(Document)

根據 ISO 文獻，其定義之原文如下：

A document is defined as a material reproduction of the auther's thoughts. The document's prime objective is to transmit communicate, and store these thoughts as accurately as possible, regardlass of the medium used for these concepts. [1, p.5]

根據此定義，文書與表現的媒體無關，而著重在它的內容。因此，在我們的定義中它也不含編排上的資料。換而言之，當我們指文書時，著重在它的內容，無涉於它的版本，印刷編排等。文書的結構常以結構文素來表示。

今取用以上之定義，經譯後為：

文書是以物質形態表達的作者的思想。其主要目的是用以傳送、溝通，或儲存這些思想。文書的表達宜盡量忠實，而與表現的媒體與編排的格式無關。

定義四：文獻

文獻是指一群全文資料的集合。有時全文與文獻兩詞可以交替運用，然而在語意上全文著重在指資料的內容結構，文獻則著重在指資料整體表現的特性，是一個集體名詞。例如：從設計程式的立場來看廿四史是全文，因為其觀察著重在全文之內容與結構，然而從史籍立場來看，廿四史是史籍中的一種(或一類)文獻，此看法，意在指廿四史之外在集體特性。

參、全文標誌規則

此為本期計劃裡資料登錄時用的標準作業規範，茲詳列之，以為參考。

壹．通則

一、凡需利用本中心發展的「中文全文處理系統」(CTPS)來處理的文獻，都必須嚴格依照本規則的規定做資料登錄的工作。若否，CTPS將產生錯誤而無法正常工作。

二、依照本規則登錄的文獻中不能有外國文字、字母、及阿拉伯數字。本規則亦不能用之於圖形、表格、公式等。故凡有上列資料之文獻，不宜直接以本規則處理。

三、文獻登錄時以頁為單位，亦即每頁建一檔案，此檔案之名稱為：「xy.src」之格式，其中x為原始文獻之代碼，y為以數字表示之頁碼，而src表示原始檔案之類別，文獻之代碼由系統工程師設定，不可任意取用。

四、資料登錄時，應盡量依照原稿之版面排列形式。凡有空頁、空行、空白之處均須比照登錄，除在細則中有規定者外，不得任意改變版面之形式。

五、所有文字及符號之鍵入，除細則中另有規定者外，均取全形。

六、關於此標註規則之實務詳如細則。凡對此規則之使用有任何疑問時，請立刻與系統工程師連絡，不可以隨自己意思選擇處理之方式。

貳．細則

一、檔案結構

檔案之第一行應先以阿拉伯數字打入頁次(即通則三中之y)，然後按換行鍵，自第二行輸入正文。

二、正文

(1)若為空白頁，則其正文以 "a" 表示之。

(2)正文前至少有二個全形空白(ㄅㄅ)。

三、書、卷

⑴書卷起處加 "d"，迄處加 "e"。

⑵書卷之標題起處加 "f"，迄處加 "g"。

⑶標題群之間以一個或一個以上之全形空白隔開。

⑷每個標題之起處加 "l"，迄處不加特別標誌。

四、字形大小

小字起處加 "★"，迄處加 "■"。

五、註釋及校勘

⑴註釋起處加 "h"，迄處加 "i"。

⑵校勘起處加 "j"，迄處加 "k"。

⑶註釋和校勘同時出現處，校勘出處加 "*"。

六、標誌表

上述之標誌，綜合如下表，可作速查之用。

標誌 / 文素	起	迄
空白頁	a	—
卷（書）	d	e
卷標題	f	g
小　字	★	■
正　文	ﬠ	—
註　釋 (ordered list)	h	i
校　勘	j	k
區別註釋校勘	※	—
小標題 (un-ordered list)	l	—

參考資料

〔1〕毛漢光等，〈史籍自動化食貨志輸入電腦第一年總報告〉，（台北市：中央研究計算中心，民國七十四年七月）。

〔2〕ISO Draft International Standard DIS 8879., <Information Processing—Text and Office Systems—Standard Generalized Markup Language(SGML), ISO/TC 97> , International Organization for Standardization, Geneva, Switzerland, Oct, 1985.

〔3〕Arno J. H. M. Peels, Norbert J. M. Janssen, and Wop Nawijn Twente University of Technology, "Document Architecture and Text Formatting", <ACM Transactions on Office Information Systems>, Vol.3, No.4 (Oct. 1985), pp.347—369.

〔4〕R. Furuta, Scofield J. and Shaw A. "Document formatting systems : Survey, concepts and issues", <ACM Computer Surveys>, Vol.14, No.3 (Sept. 1982), pp.417—472.

〔5〕C. Faloutsos "Access Methods for Text" <ACM Computing Surveys>, Vol.17, No.1, (Mar. 1985), pp.49-73