•新聞學研究• 第四十二集
民79年1月 頁 123-148

新聞研究的新工具：
新聞全文資料庫與新聞索引資料庫*

謝瀛春 馬立君**

《本文節要》

本文主要在介紹一種適用於新聞內容分析的研究工具，並詳細說明其發展的方法與步驟，以供中文新聞實務及研究之用。此研究工具是利用電腦來處理新聞資訊；可將新聞內容以全文處理方式貯存起來，然後透過索引資料庫進行各類新聞的查詢、統計。

新聞全文資料庫與新聞索引資料庫的發展成功，不只有助於新聞內容分析在量與質上的突破，難得的是此研究工具是為中文的新聞資訊而設計，研究者期望，此研究工具對新聞實務界及學術界的工作，能助一臂之力。

壹、前言

有關新聞媒介內容分析(News media content analysis)的研究，一向是新聞學的重點之一。但是，由於新聞資訊的數量與類別既龐雜又難以計算，許多研究者在使用內容

＊本文改寫自馬立君的碩士論文「我國主要報紙使用字彙之內容分析——以新聞全文資料庫及新聞索引資料庫為研究工具」的第四章。鑑於此研究工具的首創開發成功，特別改寫介紹給國內新聞學界，並特別感謝中央研究院計算中心謝清俊主任及諸位工作人員鼎力指導、協助開發此二資料庫。本資料庫中各類別名稱、權威檔名稱及其定義，由作者（謝瀛春、馬立君）共同發展界定，如需引用，請註明出處以尊重原創者的創作權。

＊＊本文作者謝瀛春為政治大學新聞系副教授；馬立君為輔仁大學大眾傳播研究所碩士。

分析法分析新聞資訊時，不得不抽取較少的樣本（如一年中的一週）或僅做某一或某些類別（如政經新聞）的分析。在過去新聞學的研究中，都是沿襲上述方式了解新聞內容，嚴格而言，這種研究結果並不能全面了解新聞的整體表現(performance)。

慶幸的是，由於電腦的應用而漸漸可彌補前述人工分類、統計的限制，而使內容分析法的理論可以近乎完美的實踐。「新聞全文資料庫與新聞索引資料庫」就是在電腦的協助下，嚐試將兩類新聞（經濟與犯罪）以全文(full-text)方式輸入，並以新聞屬性的類目爲查詢(searching)依據，透過電腦作分類、統計，而完成內容分析的量化工作。如此，不只可以將全年所有的報紙內容輸入電腦，更可以隨研究者的旨趣、需要，查詢、分析新聞內容。

其實，全文資料庫的研究開發在國外已有相當歲月了，但是有關新聞方面開發成功的並不多，頗具代表性的當推紐約時報。但仍屬資料貯存之用，是否有爲研究者研究之用而設計的？本文作者尚不得而知。而本文介紹的兩種資料庫則是首次開發的中文新聞系統。

本研究工具嚐試使用電腦協助新聞全文文獻的處理、貯存及統計工作，開發「新聞全文資料庫與新聞索引資料庫」，是一項綜合新聞學、文字學及電腦科學的科際整合工作。

這兩個資料庫：一是以中央研究院計算中心所開發的中文全文處理系統爲依據，發展成適合新聞資料特性的新聞全文資料庫，以階層(hierarchy)的方式貯存資料，是基本性質的描述（如報別、年、月、日、版次等）；另一個則是以謝瀛春（李瞻等，民75）所提出的新聞屬性類目爲基礎，並綜合其他新聞學者的相關意見，發展成新聞索引資料庫，以表格的方式貯存屬性資料，是一種細部屬性的規畫（如新聞的來源，新聞內容的性質等）。茲分述如下：

貳、新聞全文資料庫簡介

全文資料庫(full-text data-base)是指將文獻的全部原文以盡量忠於原來形式的電子方式，逐字存入電腦的資料庫中，並建立各種檢索或查詢的方法，以便使用者能透過電腦網路或相關的電子通訊設備，及時在線上檢索到文獻裡每個句子中的每個字，或進一步做統計、分析、整理等應用之工作的系統（徐惠文，民77）。

傳統資料庫所處理的對象是格式化的資料(formatted data)，它必須遵循一定的規

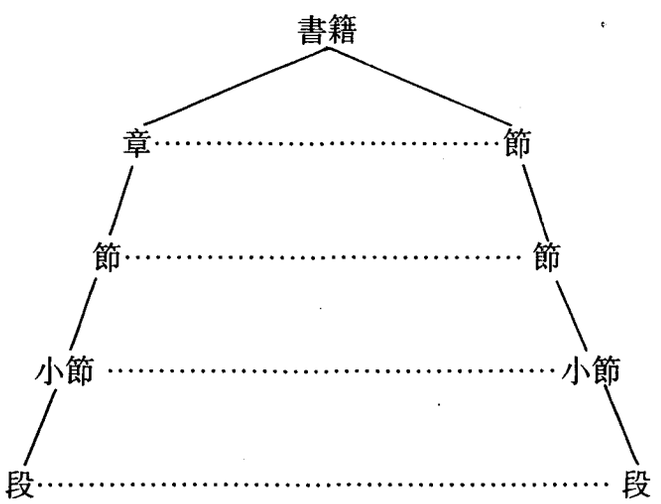
則描述資料的各項特性；而全文資料庫則沒有固定格式，是依據全文資料的結構要素，來組織資料（曾士熊，民77）。兩者在處理資料的本質差異，使全文資料庫的結構與傳統資料庫的結構大不相同。

一、新聞全文資料庫的結構

新聞全文資料庫是全文資料庫的一種，所處理的對象是全文資料。一般文獻中所蘊含的訊息，大致可區分爲下列四種（謝清俊，民75）：

㈠正文資料。包括：⑴文字描述的內容——即主體。⑵輔助說明資料——包括夾雜在敍述文字中的表格、圖案、照片等。

㈡文獻結構的資料。以書籍爲例，它包括目錄中所列的資料，如卷、册、章、節、小節、乃至於段落。換言之，在處理書籍全文資料時，「段落」是最小的單位。（如圖一）。



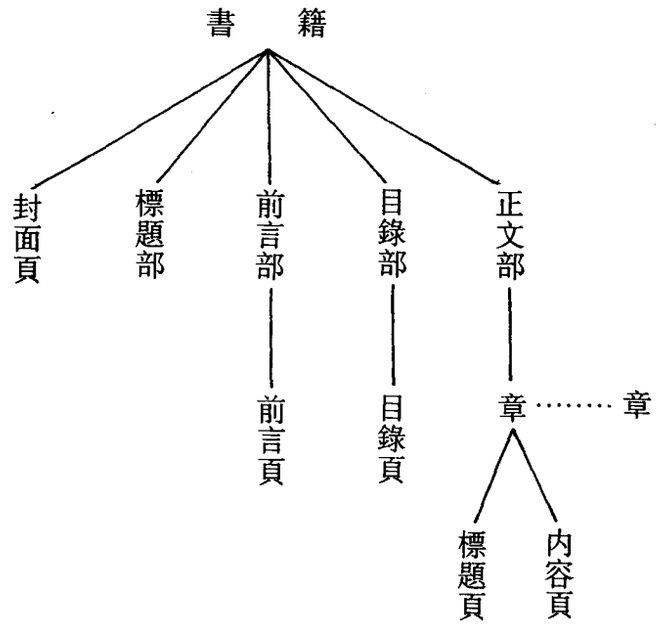
圖一：文獻結構資料圖（取材自曾士熊，民77：258）

㈢文獻編排的資料。即文獻表現在各媒體（如紙張、螢光幕等）上的編排情形，例如：頁次、橫式或直式的排版方式，字體大小變化等（如圖二）。

㈣文獻的屬性資料。如書籍上版權頁上的訊息。

由前述可知，一般文獻無論就文獻結構或排版結構，大致是一種階層式(hierarchy)的樹狀結構。

理想上，不需外加任何符號或結構指令，將原始資料直接輸入電腦，並可完整表達上述文獻訊息的方式最簡便。但由於查詢的需求，以及目前電腦對中文語意結構的認知



圖二：文獻編排結構資料圖（取材自曾士熊，民77：259）

不足，電腦並不能以不加符號的方式來精確表達文獻的訊息。因此，必須仰賴下列兩種表現的方式：（曾士熊，民77）

㈠內涵表現(implicit representation)

直接將一羣特定的符號或符號串（又稱爲界標，delimiters）加入全文資料裡，藉以區隔兩筆資料。例如在英文語句中，空白是用來分隔英文詞彙的界標。

內涵表現方式的優點是資料檔的組織簡單；缺點則是搜尋資料時速率太慢。

㈡外顯表現(explicit representation)

以外加的階層式樹狀結構，來顯示全文資料中的文獻結構及排版結構。

此法的優點是資料的搜尋快速：缺點是資料的組織較複雜，且資料的維護較麻煩。

中央研究院計算中心的中文全文處理系統，其資料描述的摸式是一種階層式的樹狀結構。爲了兼顧資料搜尋快速及組織簡單，該系統兼採內涵與外顯兩種表現方式，在全文資料段落以上（包含段落）的文獻結構，使用外加樹狀結構的外顯表現方式（如：書、章、節、小節與段落的區分等）；至於段落以下（不包含段落）的文獻或排版結構，則採用內涵表現法，直接在全文資料中加入符號來區分（如內文、標題與註解的區分等）。

雖然，新聞文獻並不像書籍文獻具有章、節、段……等明顯的階層式結構，但是新

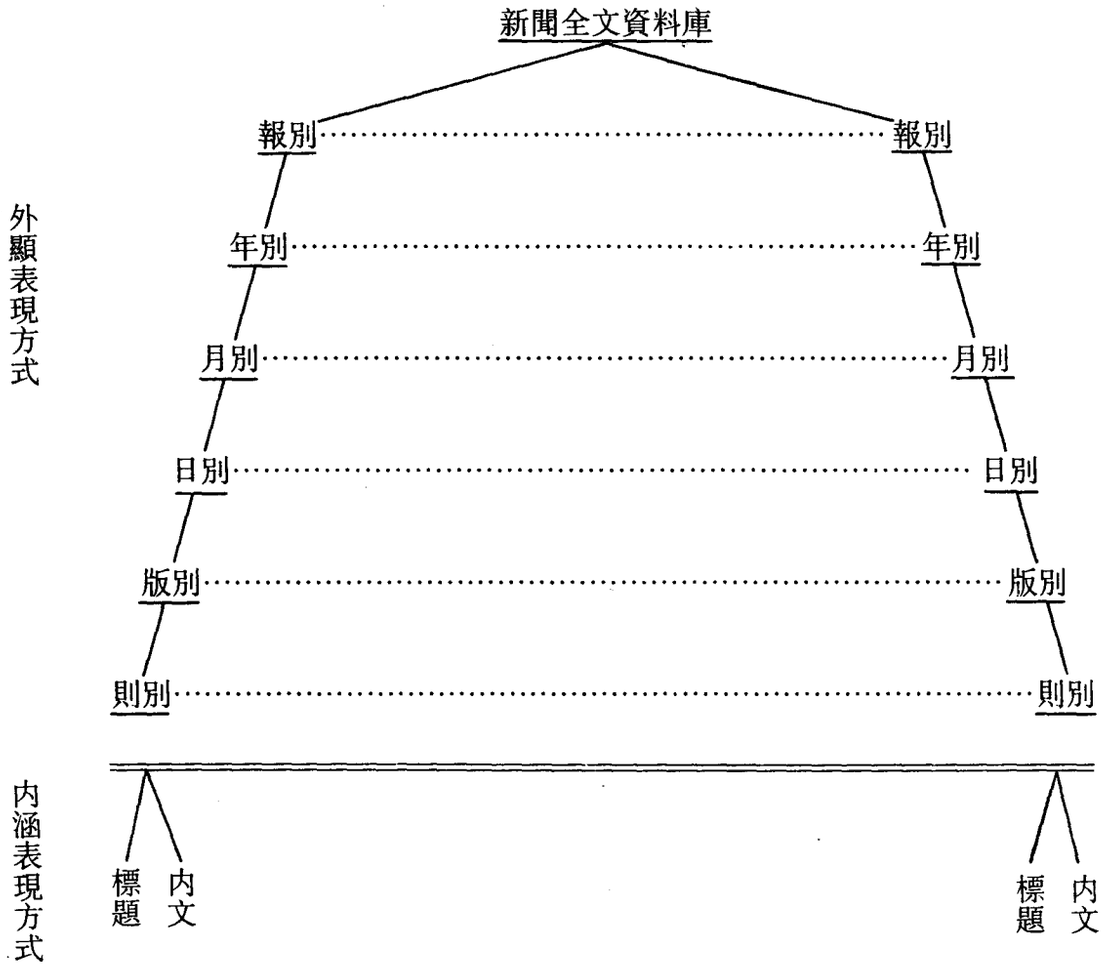
聞資料的基本性質，如報別、年別、月別、日別、則別……等，仍然符合階層式的特徵。

此外，由於中文報紙排版方式複雜，要想在全文資料庫中以實際中文報紙版面的方式呈現，技術上仍待克服。本新聞全文資料庫，將以「則」爲呈現新聞排版結構的基本單元。

基本上，本新聞全文資料庫是依循中央研究院計算中心中文全文處理系統結構，採用階層的方式，來描述新聞資料的基本性質，至於新聞資料的細部屬性（多半是人爲分類，如經濟與犯罪新聞等），將在下一節新聞索引資料庫中，詳加討論。

根據研究目的，本資料庫亦採用外顯與內涵混合表現方式。

㈠外顯表現部分，以「則」爲最小的文獻結構單位，由上至下分下列數層：



圖三：新聞全文資料庫樹狀結構圖

第一層，新聞全文資料庫。

第二層，報別：以資料所屬報紙區分。

第三層，年別：以資料時間區分。

第四層，月別：以資料時間區分。

第五層，日別：以資料時間區分。

第六層，版別：以資料版次區分。

第七層，則別：以新聞則數區分。

㈡內涵表現部分

在每則新聞的全文資料中，使用標誌符號來區分標題與內文（參見圖三）。有關標誌符號的使用將在下節中詳細討論。

二、建檔作業流程

新聞全文資料庫的建檔作業流程分爲下列兩個程序：㈠原始資料的登錄作業。㈡原始文獻檔案結構化作業。茲分述如下：

㈠原始資料的登錄作業。

原始資料的登錄作業，是指將原始資料轉變爲「電腦可以直接處理的形式」的文獻，可分爲下列幾個步驟：

1. 資料的整理

爲了方便電腦處理全文資料起見，必須兼用外顯及內涵兩種表現方式，一方面外加階層式樹狀結構；另一方面則依一定的標誌規則，在全文資料中加入適當的標誌符號(mark-up symbols)，供電腦識別文獻。

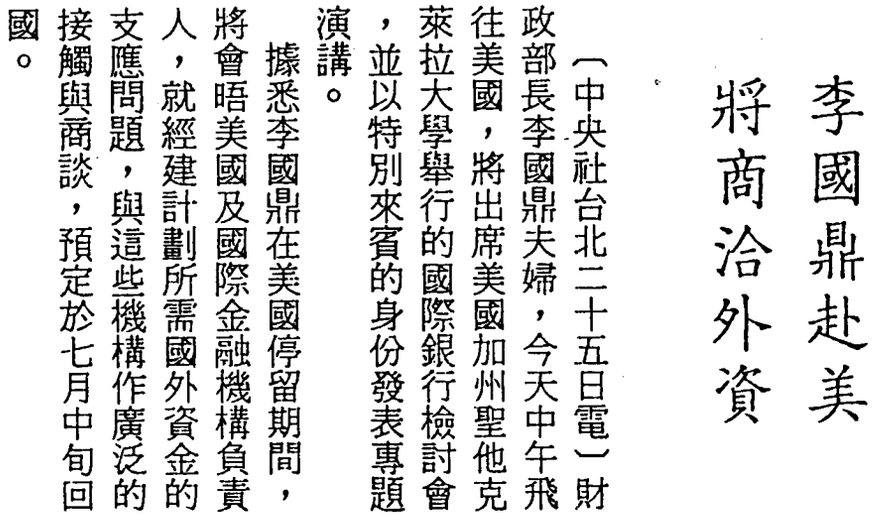
首先在外顯表現方面，先找出有意義的最小區分單元，依照研究目的，確定樹狀結構的階層數，把資料依其特性逐層歸類、排比，形成一個足以顯示文獻排版及文獻結構的樹狀結構，它的形狀就好比一株倒立的樹，因此又稱爲結構文獻樹。

以本研究爲例，是以「則」爲有意義的最小區分單元，結構文獻樹的階層由上至下分別爲：新聞全文資料庫、報別、年別、月別、日別、版別、則別共計七層。如例一（圖四），原載於中國時報61年6月26日第二版的新聞，則在歸類時，依次併入中國時報、61年、6月、26日，第二版、第n則。其餘則依此類推，逐步形成一株結構文獻樹。

完成歸類後，每一則新聞都包括一組絕對與多組相對序號及一個路徑。絕對序號是指，該則新聞在整個新聞全文資料庫中的順序；相對序號是指，該則新聞在某一特定範圍內的順序；而路徑則是指，該則新聞在結構文獻樹上歸類的途徑。

在建檔作業的過程中，電腦就是靠這兩組序號來辨認資料的位置，並以路徑來表示該則新聞在結構文獻樹上的位置。如例一（圖四），經排列得絕對序號爲18，即在所有483則新聞樣本中，排名第18。相對序號則依層數不同而有不同，如就報別而言，該則新聞屬於中國時報，相對序號爲1，若就版次而言，該則在中國時報61年6月26日第二版，因此，相對序號爲2。該則新聞經逐層歸類後，其路徑爲/1.1.1.1.2.1.4,其中「/」表根目錄，「.」是分隔單元，經由數字的對應的關係，即可了解該則新聞在結構文獻樹上的位置。

此外，在有意義的最小區分單元下，常可區分更小的單位（如則以下可分標題、內文、註解等），爲避免外顯樹狀結構太過龐雜，此時，就必須靠內涵表現法的協助，在新聞全文資料中，加入標誌符號，作爲區分。



李國鼎赴美將商洽外資

（中央社台北二十五日電）財政部長李國鼎夫婦，今天中午飛往美國，將出席美國加州聖他克萊拉大學舉行的國際銀行檢討會，並以特別來賓的身份發表專題演講。

據悉李國鼎在美國停留期間，將會晤美國及國際金融機構負責人，就經建計劃所需國外資金的支應問題，與這些機構作廣泛的接觸與商談，預定於七月中旬回國。

圖四：例一原始資料圖

圖四經加入標誌符號後，形成下面的形式：

8
1
bv
lh
李國鼎赴美
將商洽外資

e

lp

（中央社台北二十五日電）財政部長李國鼎夫婦，今天中午飛往美國，將出席美國加州聖他克萊拉大學舉行的國際銀行檢討會，並以特別來賓的身份發表專題演講。

據悉李國鼎在美國停留期間，將會晤美國及國際金融機構負責人，就經建計畫所需國外資金的支應問題，與這些機構作廣泛的接觸與商談，預定於七月中旬回國。

e

e

圖五：例一原始資料標誌後簡圖

圖五的例子中，lh代表新聞結構中的標題部分，利用標誌語法剖析器（簡稱parser）剖析後，即可賦予這段文字實體意義——表示是這則新聞的標題。同理，bv表示下啓一則新聞，lp則是指進行資料處理時，將符號lp與e之間的內容，視爲一個完整的單元。

一般而言，標誌規則所產生出來的語言是一種制式語言(Formal Language)，這種制式語言可能是程序性的(procedured type)，也可能是描述性的(descriptive type)（丁之侃，民77）。前者如例一中的lp（見圖五），指示資料處理的程序。後者則如例一中的lh（見圖五），描述新聞標題的特性。各標誌符號所具有的特定涵義及詳細的標誌規則，請參考中央研究院中文全文處理系統——建檔作業系統使用手冊（舒啓洲，民78）。

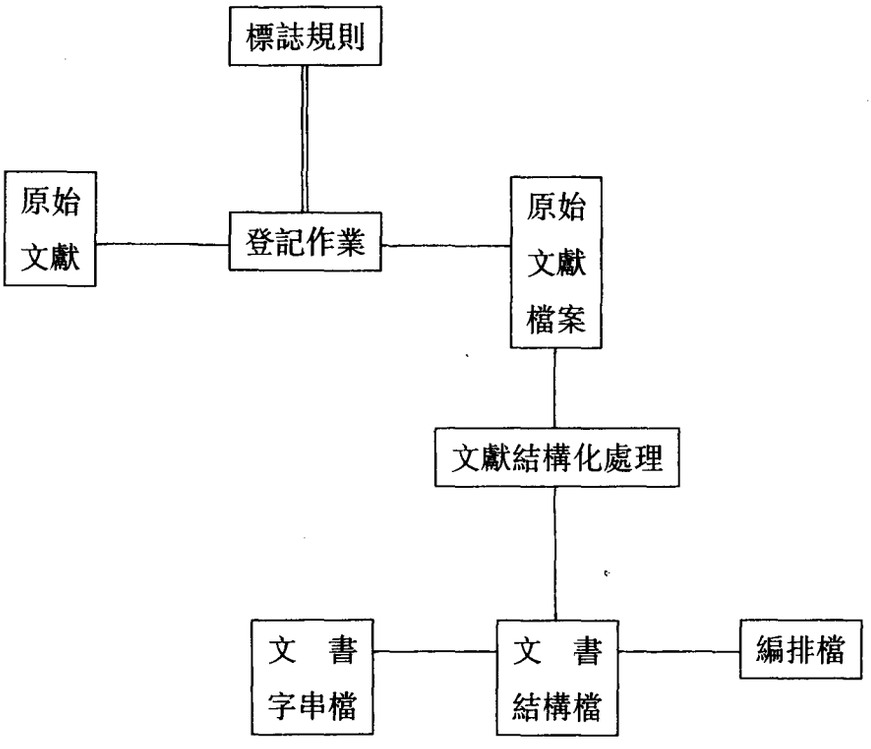
2. 資料的輸入與校對

定成原始資料的標誌程序後，即可交由資料輸入人員鍵入電腦，並經校對後，形成電腦可以處理的文獻。然後利用文書編輯程式，以機讀式原始文獻檔案備用。

㈡原始文獻檔案的結構化作業

原始文獻檔案的結構化，是指將原始資料所含的文獻訊息（如文獻編排、文獻結構……等訊息），在電腦中作有系統的整理和表達。

此部分由中文全文處理系統的建檔作業系統全權處理。詳細的步驟及規則，請參考中文全文處理系統——建檔作業系統使用手冊（舒啓洲，民78）。經程式處理後所產生的結構文獻檔案，配合新聞索引資料庫的應用，即可展開進一步的檢索、文字統計等工作。整個建檔作業的流程如圖六所示。



圖六：新聞全文資料庫檔流程圖（取材自舒啓洲，民78：10）

叁、新聞索引資料庫簡介

新聞資料除了具備前節所討論的基本性質（如新聞所屬的報別、版次、年、月、日等）外，還有許多其他屬性（如新聞的內容性質、文體、來源等）。通常這些新聞細部屬性間並不是階層式的關係，例如某一內容類別（如經濟類）的新聞，可能分屬不同的新聞來源（如本報訊、美聯社……等）；同樣道理，某一新聞來源的新聞，也可能分屬不同的內容類別。因此，以處理全文資料爲目的的新聞全文資料庫，並不完全能處理這些新聞屬性資料。此時，就必須發展一套結構及描述資料摸式截然不同的新聞索引資料庫，來輔助新聞全文資料庫的使用。

一、新聞索引資料庫的結構

㈠新聞索引資料庫的特性

新聞索引資料庫，顧名思義，就像書目索引一樣，提供新聞細部屬性的資訊。爲了精確表達新聞屬性間多對多的關係，新聞索引資料庫採用關聯式(relational)的資料模式，具有下列特性：

1. 處理的對象是格式化的資料

即原始資料必須照一定的規則，加以歸類處理。以表格的方式來描述資料的特性（如表一）：

表一：表格式資料（取材自曾士熊，民77：256）

學　號	姓名	體重	身高	考績
760101	張三	78	160	乙
760102	李四	63	182	甲
760103	趙五	59	177	甲
760104	孫六	54	178	甲

2. 控制詞彙的查詢方式

控制詞彙查詢(controlled vocaburary search），即使用一套既定的詞彙來檢索資料的方式。

使用這種查詢方式必須涉及權威檔(authority file）與索引典(thesarus)的問題（謝清俊，民75）。權威檔是指系統中授權的「合法」詞以及其同義詞，相當於上述的控制詞彙，至於索引典，則是指詞彙間一些語意關係的檔案，它包括廣義詞，狹義詞，以及特定的關係詞等。

以本研究爲例，在來源類別中，以「合衆社」一詞來代表新聞來源是「合衆國際社」的新聞，查詢時即必須查詢「合衆社」而非「合衆國際社」，此處，「合衆社」即是權威檔中的控制詞彙；而「合衆國際社」則是索引檔案。

由前述可知，須先對新聞細部屬性，有一通盤了解，並整理相關詞彙，給予明確的區分與定義，才能發展出理想的新聞索引資料庫，而進一步奠定推動新聞資料自動化的基礎。

㈡新聞索引資料庫的類目及定義

在衆多的新聞學文獻中，新聞屬性幾乎是每本相關論著中必談的主題，然而新聞屬

性的劃分，往往因爲涉及人爲主觀的判斷，不但分類粗疏不一，類別也常混淆不清，始終無法形成一個標準化、統一的新聞屬性類別，這不僅造成相關研究的不便，更成爲推動新聞資料自動化的絆腳石。

本研究從整合科際研究的角度出發，不以適應個別傳播研究的需求爲滿足，廣泛蒐集新聞屬性的相關文獻，並結合電腦科技的運用，試圖發展一個能涵蓋新聞文獻一般屬性的新聞索引資料庫，一方面輔助本次研究的進行，另一方面提供後續整合傳播與電腦科技研究的參考。

本新聞索引全文資料庫是以傳播學者謝瀛春（李瞻等，民75）所提出的新聞屬性爲主要依據，並參酌其他相關文獻（錢震，民56；賀照禮，民58；程之行，民57；徐詠平，民60；戴華山，民69）發展而成，有五大類型，共計30個新聞屬性類別。

以下是新聞索引資料庫的類別及定義

甲、基本屬性類目

在此類目下，部分新聞屬性訊息必須與新聞全文資料庫的訊息重複，用來聯結這兩個資料庫。

1. 則數

即新聞資料的編號。如第 9 則新聞，則數爲 9。

2. 路徑

即新聞資料在新聞全文資料庫中位置。如例一，它的路徑是：/1.1.1.1.2.1.4。參看圖七。

3. 報別

即新聞資料所刊載的報紙名稱。如聯合報、中時晚報。

4. 日期

即新聞資料所刊載的日期。如例一，它的日期就是 610626。

5. 版次

即新聞資料所刊載的報紙版面。

6. 內容

即依據報紙的組成要件區分，包括下列九項：

(1)報頭。

(2)刊頭。

(3)總新聞：包括新聞、言論等所有新聞的總新聞版面，簡稱總新聞。

(4)總廣告：包括分類廣告、非分類廣告等所有廣告的總廣告版面，簡稱總廣告。

(5)副刊：除了總廣告版面、總新聞版面等正刊之外，具有文藝性或知識趣味性的部分稱爲副刊，如中國時報的「人間」。

(6)專刊：將一部分具有專門性質的副刊題材予以個別處理，充分發揮其題材特性的版面，如中國時報的「民俗」專刊等。

(7)特刊：報紙爲發動一項社會活動，或紀念一個重大節日而特別增加的版面。如各報的國慶日特刊。

(8)圖片：刊載的照片、繪圖、漫畫、表格等。

(9)提要：爲吸引讀者注意力所製作的新聞提要與一週大事的新聞提要。

乙、總新聞版與副刊屬性類目

總新聞版與副刊的內容常是新聞學研究的重點，因此特別針對這些報紙內容的特性，再予以分類。

7. 內容類別

指報紙的新聞及言論部分，依其報導的內容性質所做的分類。包括下列十六項：

(1)政治新聞，包括下列三種情形：

a. 報導國際社會的成員、政治人物、國際形勢、國際性的政治、外交會議、國際組織、國與國的關係等。

b. 中央政府的措施，諸如選舉事務和外交動向、中央政府的機構及其運作、政黨的歷史和策略，以及政要的背景、習慣和前途等。

c. 有關省縣（市）民政、建設新聞等。

(2)經濟新聞：包括農、林、漁、牧、工、礦、商業、金融、貨幣、物價、運輸交通、貿易外滙、美援、財政、賦稅、觀光、保險、勞工及工商團體的活動新聞等。

(3)社會新聞：包括報導私人集會、聯誼餐會、宗教活動、遊藝會、慈善會、訂婚結婚、慶祝大會、義賣會、好人好事、展覽、剪綵、落成、悼念、祝嘏新聞等。

(4)犯罪新聞：報導犯罪事實與犯罪行爲的新聞，犯罪事實與行爲包括：a. 對人的妨害；b. 對住所的妨害；c. 對財產的妨害；d. 對公共福利的妨害；e. 對司法機關的妨害；f. 對國家安全的妨害。這裡所說的犯罪新聞，相當於一般所說的法院新聞或司法新聞。

(5)災禍新聞：包括報導天災和人禍的新聞。天災指颱風、地震、山崩等天然意外災害，人禍指車禍、火災、墜機、沈船、溺斃等。

(6)科技新聞：包括有關自然科學和技術的新聞等。如醫藥、保健、太空技術、物理、化學、天文、氣象、地質、生物、電機、礦冶等。

(7)文教新聞：包括有關文化與教育的一切報導。如教育界人事動向、教育政策、教育問題、考試消息、書評、文化活動、杏壇芬芳錄、出版動態等。

(8)體育新聞：報導田賽、徑賽、球類比賽、游泳、划船、拳擊、摔角、爬山、打獵、賽車、野營、騎馬新聞等。

(9)軍事新聞：包括戰爭、鎮暴、備戰、裁軍、軍事會議、軍事訪問、軍事機構、軍事學校、基地、武器、軍事動員、軍用物品調配、敵軍與友軍動態新聞等。

(10)交通新聞：包括交通建設、運輸活動、國際航運、與航線調整新聞等。

(11)醫藥新聞：包括疾病防治、醫療方法、藥物介紹、公共衛生、醫學器材、醫藥管制與醫院活動新聞等。

(12)影劇新聞：包括電影、電視明星、電影、電視節目介紹與批評等。

(13)藝術新聞：包括畫展、音樂演奏會、國劇、雕塑、棋奕新聞等。

(14)人情味新聞：如拾金不昧、義犬護主、車上產子、離亂重逢等充滿人情趣味的新聞，簡稱作人情味。

(15)大陸新聞：有關大陸的政治、經濟、社會等大陸動態的新聞。

(16)其他：無法歸於上述項目者。

8. 文體類別，包括下列八項：

(1)新聞：描述事實的報導文字。

(2)言論：表達意見的文字。

(3)小說：刊載於報紙副刊或專刊上的小說，包括文藝小說及武俠小說等，也包括連載小說、非連載小說及極短篇等。

(4)詩：刊載於報紙副刊、專刊或特刊的詩。

(5)散文：刊登於報紙副刊、專刊或特刊的散文。

(6)雜記：刊登於報紙副刊、專刊或特刊的雜記、雜俎感言等。

(7)更正啓事：報社更正錯誤的文字，包括更正啓事、道歉啓事等。

(8)其他：無法歸於上述項目者。

9. 新聞類別，包括下列二項：

(1)純淨新聞：註明本報訊或各通訊社電稿的純淨新聞。

(2)非純淨新聞：純淨新聞以外的新聞，包括署名的特寫、專題報導、集體採訪等新聞。

10.非純淨新聞類別，包括下列六項：

(1)特寫：記者針對某一新聞有關的人事物，作資料的蒐集和深入的描述，使讀者對此新聞有更深刻的體認，又稱爲「特稿」，均以「特寫」代表。

(2)專訪：記者針對某一新聞中的特定對象（可以是人或事），作個別專訪，尤其注重人的訪問。即使對事，亦是從人的口中說出，而不重資料的補充。刊登時，通常註明「本報記者×××專訪」。

(3)專題報導：記者針對某一特定問題，作專門探討與分析的報導。刊登時，通常註明「本報記者×××專題報導」。

(4)集體採訪：多位記者聯合採訪後所作的報導，刊登時，通常註明「本報記者×××、×××、×××集體採訪」或「本報記者集體採訪」。

(5)綜合報導：爲方便讀者對凌亂、片斷的現象能夠有一完整、有條理的了解，將零星的報導予以綜合後，再報導出來的新聞稱爲綜合報導。通常刊登時，註明「本報綜合報導」或「本報記者綜合報導」。

(6)其他：無法歸於上述項目者。

11.言論類別：

指包括社論、短評、專欄、專論、讀者投書等項目的意見性文字。包括下列六項：

(1)社論：報社或雜誌社表明其總主筆或領導人意見的文章。即明顯註明「社論」、「社評」者。

(2)短評：文字具有固定形式或名稱，經常針對不同問題分析、解釋，含有議論、批評者。通常篇幅不多，至多五、六百字。例如聯合報的「黑白集」。

(3)專欄：由學者專家或記者署名，定期在報紙固定的版面位置發表的意見性文字。

(4)專論：報社不定期邀請社會上有名望的學者專家，對不同的問題，就學術的觀點，署名發表的意見。常標明「專論」或「星期專論」。

(5)讀者投書：註明「讀者投書」或「來函照登」等的讀者來信。

(6)其他：未能歸於上述項目者。

丙、廣告類別

12.廣告類別

依照報紙廣告的刊登格式，可分爲下列兩項：

(1)分類廣告：依照性質分類排列的小廣告，又稱小廣告，此處以「分類廣告」代表。

(2)非分類廣告：分類廣告以外的廣告。包括商業廣告、機關公告、啓事廣告等非分類廣告。

13.非分類廣告類別

分類廣告以外的廣告，依照其內容可區分爲下列六項：

(1)商業廣告：以促銷、增加廣告主商業利益爲目的的廣告，包括一般商標廣告、商品廣告及商店廣告等，約占總廣告量50%以上。

(2)機關公告：政府各級機關的公告。

(3)啓事廣告：社團或私人刊登的通告或啓事，如律師聲明廣告、私人遺失啓事、婚喪啓事等。

(4)公益廣告：以建立政府或私人機構良好的公共關係，以服務社會的方式來爭取羣衆好感的廣告。如杜邦公司在報紙上刊登呼籲注意環境污染的廣告。

(5)工商服務廣告：提供各產業新發明、動態或消費資訊等的廣告，通常刊登在工商服務版。

(6)其他：未能歸於上述項目者。

14.廣告性質類別

依廣告的性質，可分爲下列二項：

(1)廣告新聞化：以新聞形式呈現的廣告。

(2)非廣告新聞化：非以新聞形式呈現的廣告。

15.色彩類別

指廣告的色彩，可分爲下列二項：

(1)彩色

(2)黑白

丁、來源類目

16.署名類別

指新聞資料依照是否署名，可區分爲下列二項：

(1)署名

(2)無署名

17.作者名

指新聞資料上作者的名字。如撰寫特稿的記者名，撰寫專論的學者專家名等。

18.來源類別

指新聞資料的出處，可分爲下列十二項：

(1)本報訊：冠以「本報訊」字樣者。

(2)本報：社論或短評等無署名的新聞評論，來源以「本報」表示。

(3)本報記者：由報社記者署名撰寫的文字，來源以本報記者代表。

(4)學者專家：由具有某一領域專長的學者或專家署名撰寫的文字，來源以學者專家表之。

(5)中央社：來源註明「中央社」者。

(6)美聯社：來源註明「美聯社」者。

(7)合衆社：來源註明「合衆社」或「合衆國際社」者。

(8)路透社：來源註明「路透社」者。

(9)法新社：來源註明「法新社」者。

(10)塔斯社：來源註明「塔斯社」者。

(11)外電：來源註明「綜合外電」者。

(12)其他：未能歸於上述項目者。

戊、編輯屬性類目

指新聞資料經由編輯處理後所產生的屬性，計有下列幾個類別：

19.文字類別

指文字的性質，包括下列二項：

(1)標題：位於新聞之前，類似文章題目或摘要，用以提示新聞要點，吸引讀者的文字，通常以不同的字體及較大的字號出現。

(2)內文：不包括標題的文字。

20.標題性質類別

指依照標題的結構性質所形成的類別。包括下列七項：

(1)主題：表現出新聞中最主要事件的文字，通常主題用的標題字號最大，最醒目、最突出。

(2)子題：又稱「說明題」，是主題的補充說明，子題位置在主題之後，所用的標題字字號小於主題。

(3)引題：即位於主題之前，具有導引功能的標題者，所用的字號小於主題。由於引題所在的位置猶如臉上的眉毛，所以又稱「眉題」。

(4)副題：在一條較複雜的新聞中，表達次要事實的標題文字稱爲副題，副題字號小於主題字號。通常位於主題之後、子題之前，也有置於子題之後。

(5)分題：在處理長稿時（如元首文告等），用以調節美化版面的一組標題。通常字號較主要標題小。

(6)插題：又稱：「小標題」。用以美化版面或區隔長文，所使用的簡單字句，可以是

原文中的句子，或加大每段的第一個字。

(7)其他：未能歸於上述項目者。

21.標題格式類別

指標題登載的格式，依照標題排列的方式來區分，包括下列四項：

(1)直題：直式標題。

(2)橫題：橫式標題。

(3)文包題：標題位於文章中央，通常以闢欄處理。

(4)其他：未能歸於上述項目者。

依照標題所佔欄數，可區分爲下列八項：

(1)一欄題。

(2)二欄題。

(3)三欄題。

(4)四欄題。

(5)五欄題。

(6)六欄題。

(7)七欄題。

(8)八欄題。

九欄以上的標題較少見，在此不詳列，研究者可依實際狀況分類登錄。

22.字體類別

指報紙文字所使用的字體，包括下列五項：

(1)楷體：又稱「楷書」。

(2)黑體：又稱「方體」、「方頭」、「等線體」、「方黑體」。

(3)宋體：又稱「老宋」或「明體」。

(4)長宋。

(5)其他：未能歸於上述項目者。

23.字號類別

指報紙文字字體的大小，鉛字印刷以「號」爲計算單位；照相打字以「級數」爲計算單位，研究者可依實際的號數或級數登錄。

24.位置

指新聞資料在版面上的位置。以報紙長與寬的二分之一爲基準，分爲右上、左上、右下、左下四部分，該則新聞資料篇幅一半以上所在的部位，即爲其在版面上的位置。

包括下列四項：

(1)右上。

(2)左上。

(3)右下。

(4)左下。

25.欄高

指新聞資料所佔欄位的高度，依實際欄高（公分或吋）登錄。

26.欄數

指新聞資料所佔欄位的多寡，依實際欄數登錄。

27.篇幅

指新聞資料所佔的面積，依實際測量數字登錄。

28.編輯格式類別

指新聞資料是否經過闢欄的編輯格式的處理。包括下列二項：

(1)闢欄：在版面上闢出一角來刊登新聞、特寫或專論。

(2)未闢欄。

29.關鍵字

指經研究者定義而與原文內容有關的字或詞。關鍵字查詢可以讓研究者直接以文獻內容的線索來檢索資料。例如將所有有關俞國華的新聞資料的關鍵字定義爲「俞國華」，則檢索「俞國華」，即可查出所有有關俞國華的新聞資料。爲因應不同的檢索需要，此類別不採嚴格規定，可由研究者依需要自行定義檢索。

30.備註

此類別爲預留的欄位，供後續研究者開發新的屬性類目時使用。

二、建檔作業流程

新聞索引資料庫的建檔作業流程可分爲下列兩個程序：㈠新聞索引資料庫的系統設定；㈡登錄作業。茲分述如下：

㈠新聞索引資料庫的系統設定

本資料庫是依據現成的套裝軟體INFORMIX系統架構發展而成。INFORMIX系統可以同時處理格式化的中、英文資料，對於資料的欄位、輸入格式均有詳細的規定，且須使用一套近似英文語法的結構化查詢語言（SQL,Structured Query Language)做資料查詢，而查詢的內容則是經過標準化、統一的控制詞彙。

爲了輸入及查詢的方便起見，本研究一方面儘量以簡約的方式統一各索引類目的名

稱，形成本資料庫控制詞彙的權威檔與索引典，並據此，定義各類別的欄位及輸入格式。另一方面將類別名稱轉換爲英文名稱，以搭配結構化查詢語言的運用。（詳細說明請參考下節）新聞索引資料庫類別名稱、欄位及輸入格式對照表如下：

表二：新聞索引資料庫類別名稱、欄位及輸入格式對照表

新聞屬性類別	英文名稱	所佔欄位	輸入格式
1. 則數	n-item	系統設定	數字
2. 路徑	n-path	20	文字
3. 報別	np-type	10	文字
4. 日期	n-date	6	文字
5. 版次	n-page	系統設定	數字
6. 內容	n-content	6	文字
7. 內容類別	c-type	6	文字
8. 文體類別	w-type	8	文字
9. 新聞類別	n-type	6	文字
10.非純淨新聞類別	npn-type	8	文字
11.言論類別	e-type	8	文字
12.廣告類別	a-type	6	文字
13.非分類廣告類別	nca-type	12	文字
14.廣告性質願別	ad-j	12	文字
15.色彩類別	n-color	4	文字
16.署名類別	by-line	6	文字
17.作者名	n-author	36	文字
18.來源名	n-source	24	文字
19.文字類別	n-format	6	文字
20.標題性質類別	h-type	4	文字
21.標題格式類別	h-style	14	文字
22.字體類別	wp-type	8	文字
23.字號類別	wm-type	系統設定	數字

24.位置	n-area	4	文字
25.欄高	c-height	系統設定	數字
26.欄數	n-column	系統設定	數字
27.篇幅	n-space	系統設定	數字
28.編輯格式類別	n-frame	6	文字
29.關鍵字	key-word	12	文字
30.備註	n-note	60	文字

㈡登錄作業

新聞索引資料庫的登錄作業，是指將原始資料按定義好的類目、欄位依序歸類，形成表格式的資料，再輸入電腦中。可分爲下列幾個步驟：

1. 製定歸類原則及歸類表格

首先，依照新聞屬性類別的定義，製定歸類原則，詳細記載在登錄簿上。

其次，爲方便登錄作業，應製定歸類表格（如表三所示）備用。

表三：歸類表格

1. 則數＿＿＿＿＿＿＿＿
2. 路徑＿＿＿＿＿＿＿＿
3. 報別＿＿＿＿＿＿＿＿
4. 日期＿＿＿＿＿＿＿＿
5. 版次＿＿＿＿＿＿＿＿
6. 內容＿＿＿＿＿＿＿＿
7. 內容類別＿＿＿＿＿＿＿＿
8. 文體類別＿＿＿＿＿＿＿＿
9. 新聞類別＿＿＿＿＿＿＿＿
10. 非純淨新聞類別＿＿＿＿＿＿＿＿
11. 言論類別＿＿＿＿＿＿＿＿
12. 廣告類別＿＿＿＿＿＿＿＿
13. 非分類別＿＿＿＿＿＿＿＿
14. 廣告性質類別＿＿＿＿＿＿＿＿
15. 色彩類別＿＿＿＿＿＿＿＿
16. 署名類別＿＿＿＿＿＿＿＿
17. 作者名＿＿＿＿＿＿＿＿
18. 來源類別＿＿＿＿＿＿＿＿
19. 文字類別＿＿＿＿＿＿＿＿
20. 標題性質類別＿＿＿＿＿＿＿＿
21. 標題格式類別＿＿＿＿＿＿＿＿
22. 字體類別＿＿＿＿＿＿＿＿
23. 字號類別＿＿＿＿＿＿＿＿
24. 位置＿＿＿＿＿＿＿＿
25. 欄高＿＿＿＿＿＿＿＿
26. 欄數＿＿＿＿＿＿＿＿
27. 篇幅＿＿＿＿＿＿＿＿
28. 編輯格式類別＿＿＿＿＿＿＿＿
29. 關鍵字＿＿＿＿＿＿＿＿
30. 備註＿＿＿＿＿＿＿＿

2. 資料的格式化與資料的輸入

依照登錄簿上的歸類原則，將原始資料逐則逐項登錄在歸類表格上，一張表格代表一則新聞的屬性資料。如例一（詳見圖四），這則原載於中國時報61年6月26日第二版的經濟新聞，其在新聞全文資料庫的路徑爲/1.1.1.1.2.1.4。依其新聞屬性逐項歸類，並填入歸類表格（如表四）。然後再由輸入人員將歸類表格上的資料，逐張鍵入電腦中，經系統程式處理後，即完成建檔程序。

表四：新聞索引屬性歸類表

1. 則數	18	16. 署名類別	無署名
2. 路徑	/1.1.1.1.2.1.4	17. 作者名	
3. 報別	中國時報	18. 來源類別	中央社
4. 日期	610626	19. 文字類別	
5. 版次	2	20. 標題性質類別	
6. 內容	總新聞	21. 標題格式類別	
7. 內容類別	經濟	22. 字體類別	
8. 文體類別	新聞	23. 字號類別	
9. 新聞類別	純淨	24. 位置	
10. 非純淨新聞類別		25. 欄高	
11. 言論類別		26. 欄數	
12. 廣告類別		27. 篇幅	
13. 非分類別		28. 編輯格式類別	
14. 廣告性質類別		29. 關鍵字	
15. 色彩類別		30. 備註	

肆、新聞全文資料庫與新聞索引資料庫的應用：實例介紹——本研究的應用

爲進一步釐清新聞全文資料庫與新聞索引資料庫的建檔程序，特以本研究爲例，說明如下：

一、新聞全文資料庫部分

依研究目的，分別自聯合與中時61年及76年兩年的報紙中，抽出樣本共483則新聞。經初步資料整理後，即著手新聞全文資料庫的建立。以下列兩個程序進行：

㈠原始資料的登錄作業

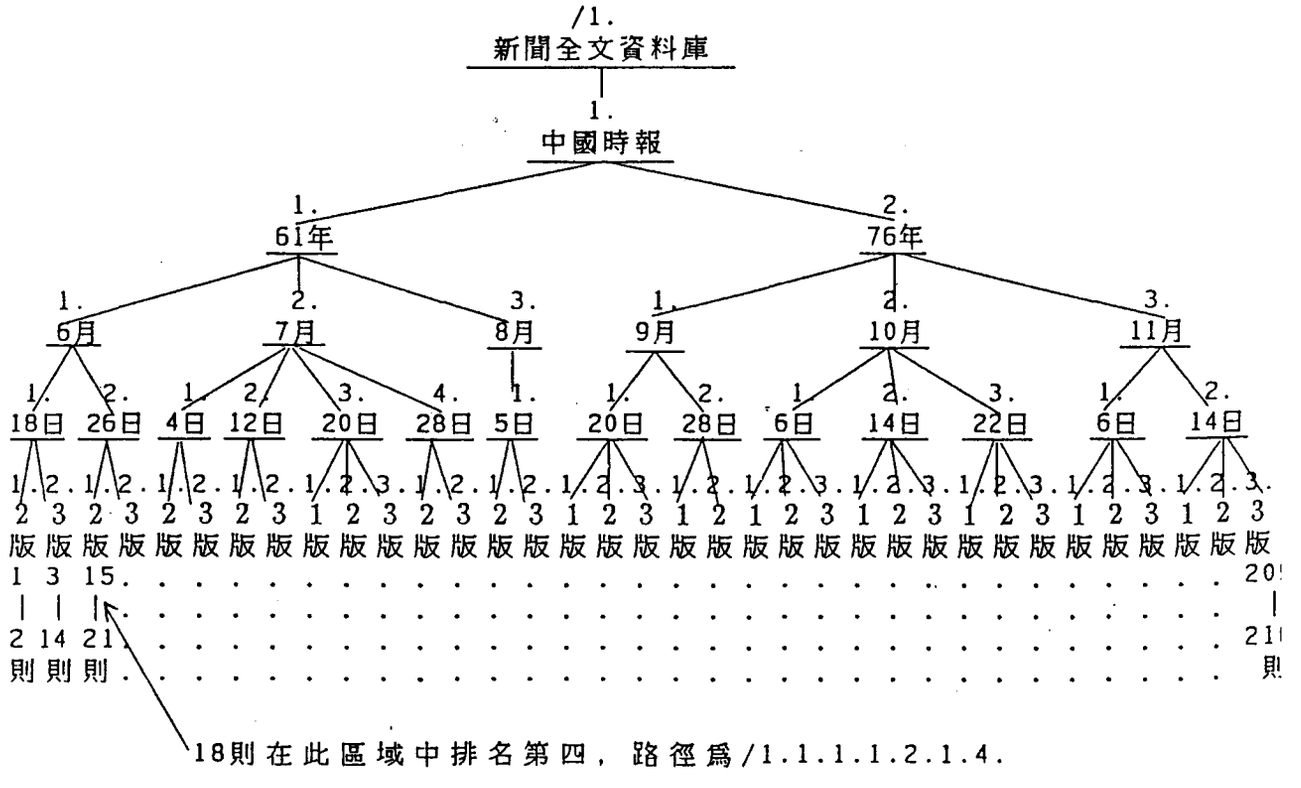
登錄作業又分爲資料的整理與資料的輸入與校對。

1. 資料的整理

甲、外顯表現部分

首先，找出有意義的最小區分單元（本研究中是「則」），然後依研究所需，將新聞資料的基本特性依階層式排列，依次是：第一層，新聞全文資料庫；第二層，報別；第三層，年別；第四層，月別；第五層，日別；第六層，版別；第七層，則別，共七層。隨後將483則新聞依其基本特性，逐則逐層歸類，並依序排列，形成一結構文獻樹。

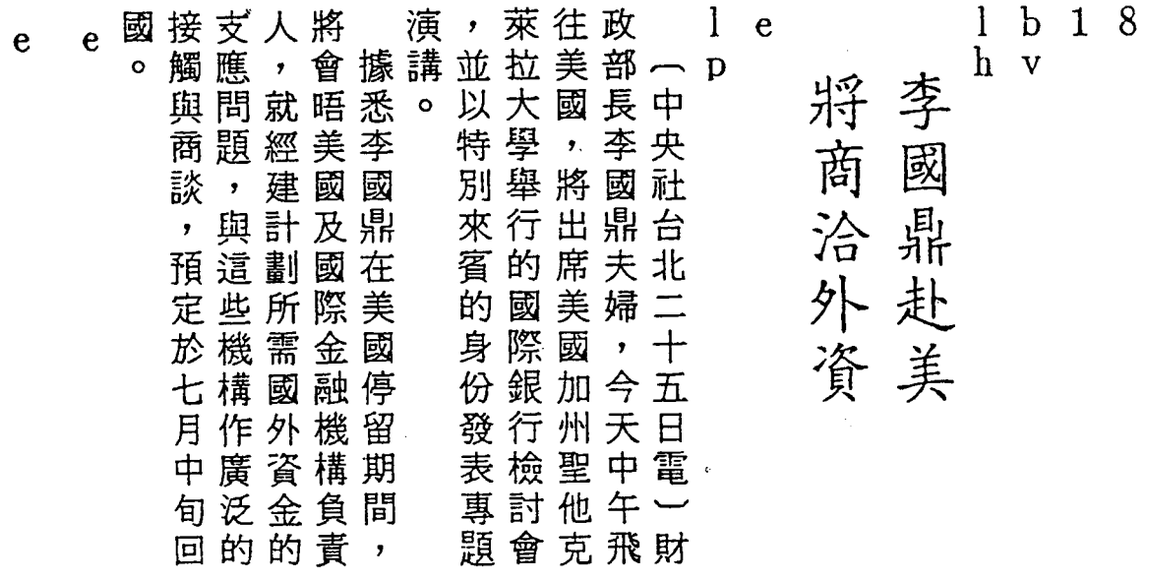
經歸類後，每則新聞均獲得絕對與相對兩種序號及一個路徑。如例一，經逐層歸類後，獲得絕對序號爲18（即在所有483則新聞中，排名第18）；路徑則爲/1.1.1.1.2.1.4, 參見圖七。



圖七：新聞全文資料庫樹狀結構圖（中國時報部分）

乙、內涵表現部分

其次在「則」以下，還區分標題與內文兩類，因此，必須在原文中直接加上標誌符號，供電腦識別（如圖八）。



圖八：原始資料標誌後簡圖

2. 資料的輸入與校對

完成標誌的資料，即可由資料輸入人員，依格式逐字輸入，並經適當校對後備用（如圖九）。

18

bv

lh

李國鼎赴美

將商洽外資

e

lp

〔中央社台北二十五日電〕財政部長李國鼎夫婦，今天中午飛往美國，將出席美國加州聖他克萊拉大學舉行的國際銀行檢討會，並以特別來賓的身份發表專題演講。

據悉李國鼎在美國停留期間，將會晤美國及國際金融機構負責人，就經建計畫所需國外資金的支應問題，與這些機構作廣泛的接觸與商談，預定於七月中旬回國。

e

e

圖九：例一原始資料標誌後簡圖

㈡原始文獻檔案的結構化

登錄完成的資料（稱原始文獻檔案），經過建檔作業系統的處理，即可自動產生結構化的文獻檔案，完成建檔作業。

二、新聞索引資料庫部分

新聞索引資料庫的建立，分下列幾個程序：

㈠新聞索引資料庫系統的設定

依照本研究所提出的新聞索引類目，在INFORMIX資料庫系統下，逐類定義，有關各類目的英文名稱、欄位限制、輸入格式，請參閱表二。

㈡登錄作業

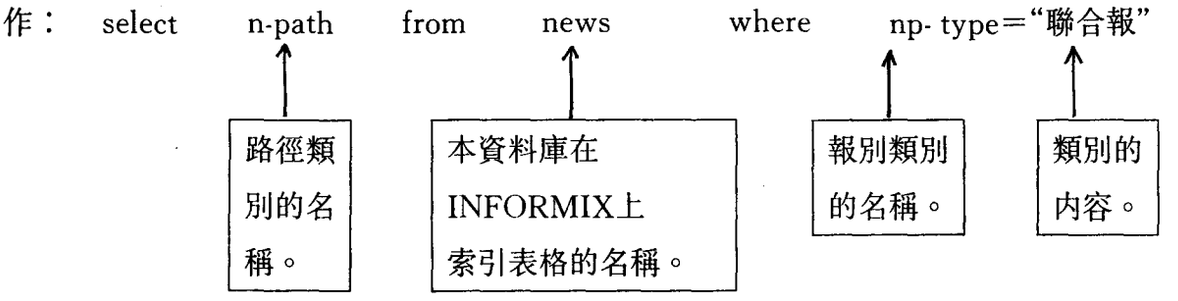
新聞索引資料庫系統設定後，即依照新聞索引類別的定義，製定明確的歸類原則，作歸類時參考。並依類目製成歸類表格，逐則逐張填寫，方便輸入人員輸入。

如例一，其路徑爲/1.1.1.1.2.1.4,內容類別爲經濟，文體類別是新聞，則在內容類別及文體類別上，分別記作經濟、新聞，其他類別亦以此類推，研究者可依所需擇項填寫（請參考表四）。

資料輸入人員依據填好的歸類表格，逐張輸入電腦（如表五），即完成建檔作業。

㈢資料的查詢與設計

如前所述，檢索新聞索引資料庫的資料時，必須使用一套近似英文語法的結構化查詢語言，以本研究而言，最常用的查詢語句爲「select類別名稱from表格名稱where類別名稱＝（類別內容）」。例如我們要選出所有報別是聯合報的路徑時，查詢語句應寫



表五：新聞索引資料庫的表格資料

1. 則數〔18〕
2. 路徑〔/1.1.1.1.2.1.4.〕
3. 報別〔聯合報〕
4. 日期〔610626〕
5. 版次〔2〕
6. 內容〔總新聞〕
7. 內容類別〔經濟〕
8. 文體類別〔新聞〕
9. 新聞類別〔純淨〕
10. 非純淨新聞類別〔　　〕
11. 言論類別〔　　〕
12. 廣告類別〔　　〕
13. 非分類廣告類別〔　　〕
14. 廣告性質類別〔　　〕
15. 色彩類別〔　　〕
16. 署名類別〔無署名〕
17. 作者名〔　　〕
48. 來源類別〔中央社〕
19. 文字類別〔　　〕
20. 標題性質類別〔　　〕
21. 標題格式類別〔　　〕
22. 字體類別〔　　〕
23. 字號類別〔　　〕
24. 位置〔　　〕
25. 欄高〔　　〕
26. 欄數〔　　〕
27. 篇幅〔　　〕
28. 編輯格式類別〔　　〕
29. 關鍵字〔　　〕
30. 備註〔　　〕

此外，本研究並使用中央研究院陸念慈所設計的文字統計系統，進行資料的統計，詳細介紹請參考中文文獻資料庫之文字統計系統套裝軟體使用手册（陸念慈，民78）。

伍、結　　語：

這兩個資料庫的建檔作業，對初學者而言可能起步階段進展緩慢。但是，如果能在此二資料庫的系統上使用，則免除了類目定義及欄位界定等繁瑣手續，使用者只需就新聞索引檔中符合研究或查詢所需的類別、項目輸入資料，即可在極短時間內獲得所需的資訊（包括頻次、總計等的統計結果），而且可以對新聞內容的類別、日期、版次、內容等有全貌的分析。

參考資料：

1. 丁之侃　民77　「史籍自動化——一個中文全文處理系統的實例」，科學月刊，19(4)：268～272。
2. 李瞻、祝基瀅、王石番、謝瀛春　民75　我國未來傳播政策之研究。　台北：行政院研究發展考核委員會。
3. 徐詠平　民60　新聞學概論。台北：中華。
4. 徐惠文　民77　「全文資料庫的發展與現況」，科學月刊，19(4)：248～251。
5. 陸念慈　民78　中文文獻資料庫之文字統計系統套裝軟體使用手册。台北：中央研究院計算中心。
6. 曾士熊　民77　「簡介傳統資料庫與全文資料庫」，科學月刊，19(4)：255～261。
7. 舒啓洲　民78　中文全文處理系統——建檔作業系統使用手册。台北：中央研究院計算中心。
8. 程之行　民57　新聞原論。台北：商務。
9. 賀照禮　民58　新聞學的理論與實際。台北：蘭台。
10. 錢震　民56　新聞論。台北：中央日報社。
11. 謝清俊　民75　「中文全文資料庫的設計與應用」，中央研究院計算中心通訊，2(22)：95～96。
12. 戴華山　民69　新聞學理論與實務。台北：學生。

■