131 26

古籍校讀工具「中文文獻處理系統」的設計

The Design of a Chinese Document Processor and Its Application to Ancient Document

謝清俊 Ching-chun Hsieh 中央研究院資訊科學研究所研究員
莊德明 Der-ming Juang 中央研究院資訊科學研究所助理

摘　　要

本文探討計算機如何處理古籍，問題涉及：古籍之版本、注疏、以及今人所加之註釋等在計算機中的表達、檢索、對映、參照、和應用等。文中採用《心經》為例，說明設計的理念和所發展的軟體工具「中文文獻處理系統」（CDP）。

CDP是一個通用的工具，其應用之對象並不限於處理古籍，然而在本文中僅針對古籍特有的性質和問題而提出討論。CDP可視為一種超文件(Hypertext)系統，它包含文件本身，文件之間相連的關系，以及檢索和瀏覽的功能等。文件間的關係是透過「知識結構」相連的。知識結構種類繁多，本文僅舉類似分類的樹狀結構為例加以說明。

此系統目前已可處理古籍各版本間之對映、注疏和相關資料與原文間之參照，以及原文內容之標誌等。CDP尚在開發之中，本文所報告者僅為初步之進展。

Abstract

A hypertext tool, the Chinese Document Processor (CDP) and its application to ancient document are presented in this paper. The design of CDP will be explained around the themes of handling various versions of original document and their related explanatory documents.

CDP provides a friendly interactive way of doing 'content markup' of documents by linking the designated text string to generic knowledge structures. In this paper, for simplicity and clarity, we selected the《Heart Sutra》(心經) as the original document and its Ker-wen (科文, a skeleton of the content) as the generic knowledge structure to illustrate the techniques of representing various versions, providing matching, browsing, and retrieving of ancient documents.

中國古籍整理研究出版現代化國際會議
一九九五年七月廿二至廿四

古籍校讀工具「中文文獻處理系統」的設計

壹、前言

本文以佛教經典《心經》的諸版本及相關的注疏文件為例，探討計算機如何處理古籍的問題。討論的內容包括：古籍的各種版本，它的古今注疏釋譯等在計算機中如何表達，如何瀏覽、檢索、參照、以及應用等問題。

為此，我們開發了一套軟體工具「中文文獻處理系統」(Chinese Document Processor, CDP)來完成上述的任務。CDP是一個通用工具，處理的對象並不限於古文件。它是一具有超文件(hypertext)性質的工具，以文件的字串和某種知識結構間關聯的關係作為超文件的鍵(links)，把文件內容與相關知識連接成一網路，以提供瀏覽、對映、檢索、參照等功能。CDP的雛型已完成，然而它的功能和結構並未完備，尚在改良和發展中。

選擇《心經》為例是為了簡潔。《心經》最小的版本才260字，注疏釋譯甚多，且有極佳的內容綱要（即關於它的內容的知識結構）甚合適作為本文之題例。此外，佛經的結構較一般文章有規律，它們有共同的結構，而顯示這結構和其內容綱要的文字稱為它的科文（或科判）。換言之，科文是一個樹狀結構的經文提要，比目錄還詳細，以往主要用於方便解釋或檢索經文。在計算機裡，科文正可用做經文的內容、主題、或文句等的檢索和比對。若為一般古籍比照佛經編一科文（或編一較詳細的目錄），則此古籍即可適用於本研究提出的系統。

本文用CDP對治古籍主要從兩個功能上著想，其一是從版本問題下手，以校讎為重點。在漢學研究的領域裡，無論是考據之學或是義理之學，版本問題都居首要，是故本研究應該對古籍之電子化具有積極的意義。其次是古籍的研讀。在此包括古籍及其相關資料之匯集，內容之鉤稽對映，瀏覽檢索等。根據這些功能需求，建立超文件式的全文資料庫是自然的想法。根據此二主要功能，本文將CDP釋名為古籍校讀工具。

貳、古籍之校讎

對治不同版本之古籍，古有校讎，今曰校書。校讎方法大致包涵了：逸書蒐輯，眞偽辨別，底本互勘，群籍鉤稽，篇第審定，和目錄論次等工作【註一】。這些工作，原來都是人去做的，是故目前所有的方法也都是因人而設，因人而寫。如果要計算機來幫忙，首要之事便是要分辨和規劃：那些事適合給計算機做，那些事還是該由人做，以及人和機器如何相互配合等。

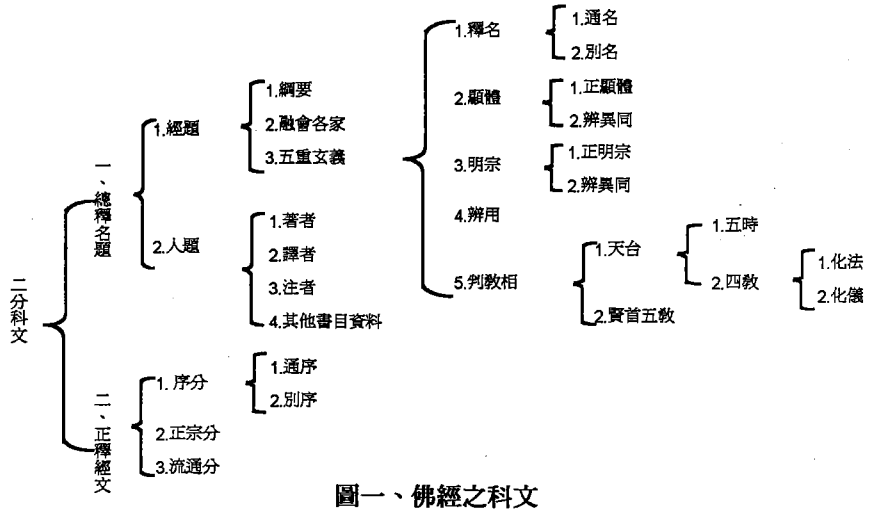
在逸書蒐集方面，主要還是靠人。待人蒐集到書之後，計算機可協助編立書目，並可將該書原文製成機器可閱讀之電子形式，除了準備給自己用以外尚可供同行參考。在本研究中假設所輯之書均已製成機讀形式。至於眞偽辨別方面，還是靠人；計算機或可提供快速之比對與相關文件之互參功能，來協助人做眞偽之辨，然而這些功能卻已包涵在底本互勘之中。

底本之互勘，據張舜徽引陳垣《元典章校補釋例・第四十三〈校法四例〉》所用者四端：【註二】

一、校對法：以同書之祖本或別本對讀
二、本校法：以本書前後互證
三、他校法：以他書校本書
四、理校法：無古本可據、或數本互異無所適從時用之。

以上四法，均涉及同書或不同之書（或不同之版本）相關資料間之鉤稽比對，如果設計得當，計算機於此應可發揮極大之作用。至於群籍鉤稽之條，事實上已含在理校法之中。至於篇第審定和目錄論次之主要工作，主要還是靠人，計算機可協助作目錄或篇第之整理工作。

在佛典中有目錄作用者為該經典之科文，亦稱科判。由於佛典有一致之編輯體例，故一般之佛典可共用一個科文，請參照圖一之說明。



圖一、佛經之科文

圖一之科文中，經題和人題部份即含有該經典之部份目錄資料。每部經典均可整理出總釋名題中之經題和人題、以及正釋經文中之序分，正宗分，以及流通分等部份。各經典重要的正文，全在正宗分之內。【註三】（至於正宗分之結構請參閱下文中之說明）。科文之主要用途在方便解釋或檢索經文，其結構式極適合計算機之處理，它不僅提供可相互比對之處所及範圍，更可作內容或主題之檢索及參考，甚至可協助作詞彙的整理以及文句中訛、衍、缺、脫等之校勘。

參、體例之整理和在計算機中之表達

關於各版本間互異之實際情況，有衍文、偽體、倒置、脫落、誤改、誤解、誤增、誤刪，以及簡策錯亂、篇章顛倒等多種現象，俞樾在《古書疑義舉例》中將他校書經驗歸納為37例，可稱通例【註四】。依此37例觀之，其分劃頗細，分劃之原則多涉及導致錯誤的原因。這些原因涉及文章的內容，以「衍例」為例，即劃分為「兩字義同者」、「兩字形似者」、「涉上下文者」、及「涉注文者」等四例。這些原因涉及許多語文知識和對文章的理解，均非目前之計算機技術所易於處理者，故需將這些通例建構在計算機中作為一種通類的知識結構(Generic Knowledge)，再以標誌工具供使用者聯繫文章中文字相關內容，才能順利處理這些通例知識。

然而，若從注記之常例而觀之，張舜徽僅舉出十種作為常見之情況，請參照表一【註五】。例如其第一條：「凡文字有不同者，可注云：『某，一本作某』。」這種注記實是側重文字差異之形式，而非如俞樾之注重其內容。計算機是長於處理「形式」者，是故計算機可以協助做注記常例之比對。至於如何將常例之形式詮釋為內容的關係，則目前恐仍須如上靠人力而為之了。

表一、張舜徽所舉之十種常用的注記體例

一、凡文字有不同者，可注云：『某，一本作某。』(或具體寫明版本名稱)
二、凡脫一字者，可注云：『某本某下有某字。』
三、凡脫二字以上者，可注云：『某本某下有某某幾字。』
四、凡文字明知已誤者，可注云：『某當作某。』
五、凡文字不能即定其誤者，可注云：『某疑當作某。』
六、凡衍一字者，可注云：『某本無某字。』
七、凡衍二字以上者，可注云：『某本某字下無某某幾字。』
八、字倒而可通者，可注云：『某本某某二字互乙。』
九、字倒而不可通者，可注云：『某本作某某。』
十、文句前後倒置者，可注云：『某本某句在某句下。』

如用計算機來處理注記之常例，則處理之規律（即其形式之變化）仍可再減。對於計算機而言，一個版本之文章，其機讀形式就是一字串（如下節說明）。若比較兩版本之異同，實即比照兩字串之同異。因此，以計算機處理字串的指令來看，只要有：增（insert）、刪（delete）、取代（replace）及互換（swap）等四者，就足夠能表示兩字串在其形式上不同之處了。若再能用移（move）或抄（copy）指令將相同之字串自一處移至他處，那麼，對張舜徽所舉之注記常例，除無法區分八與九兩例之外，均可由計算機自動處理了，而區分八與九兩例之事若經使用者協助做標誌，則全部都可以用計算機來做了。

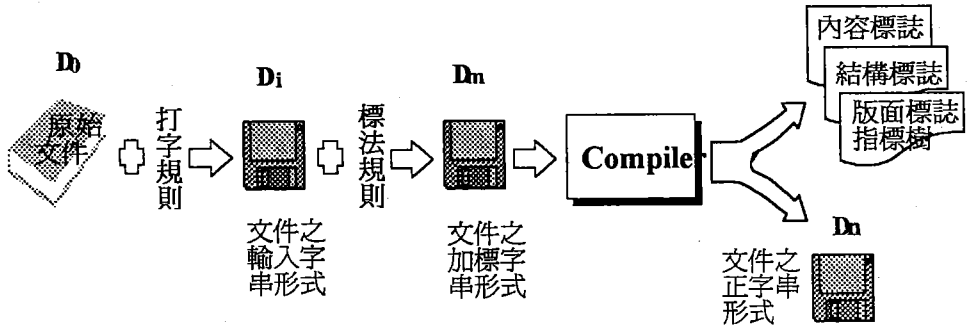
肆、文件在計算機中之表達

一、文件之字串形式

原始文件（以D表示，下文仿此）經打字後即變成該文件之一種字串形式，此字串形式之文件，其文字部份和原始文件相同，然而其文字間或夾有標誌符號，以載明原始文件之一些版面特徵，諸如頁次、夾注等等。換言之，在輸入時，為保留原書之版面訊息，吾人常加入表示版面結構之標誌。此形式之字串稱為原始文件的「輸入字串」（Di）。

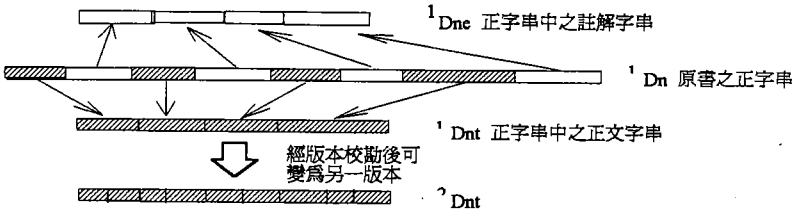
為了使計算機能明了原始文件的內容結構，如目錄或科文等，或是欲使計算機能識別一些與文件內容有關的文句，如檢索詞或句讀等等，便須在Di中加入一些特殊的標誌符號，即加入結構標誌與內容標誌，以資認別【註六】。加過這些標誌後的文件字串稱為「加標字串」（Dm），而此Dm即可讀入計算機中處理。

設若計算機中有一編譯器（Compiler）可以解讀上述之各種標誌，則在編譯器處理過Dm後，可產生三種樹狀結構的標誌檔案。這些檔案除保有原標誌的信息外，還有該標誌指向它在Dn中位置的指標（pointer）。而Dn則是將Dm中所有標誌均剝除後，所得到的字串，稱之為「正字串」。這些關係如圖二所示。



圖二、文件之各種字串形式和文件在計算機中的表達示意圖

在本文所討論之文件，都假設已是Dn的形式，而科文的結構則是上述的那種樹狀結構檔案。【註七】如果Dn只是一部佛經的原文，那麼問題就很單純了，我們可以直接探討多版本的表達與彼此間對映的方式。如果Dn中夾有注疏、校勘、評論、句讀等等不是本文的註解，則我們可以將該經之原文和註解分開，即分Dn為兩個字串Dnt和Dne；其中 Dnt表示Dn中之正文部份，即「正文字串」，而 Dne表示 Dn中註解部份,即「註解字串」,如圖三所示。由於Dn，Dnt或Dne的內容或彼此的關係都是不允許更改的,我們仍然可以得到原經文之字串Dnt去做多版本之表達。故為方便計，下文中我們直接用Dn來討論多版本表達的方法。



圖三、經書中正文字串和註解字串之映射關係，斜線部份表示正文，空白部份表詳註解文字

二、版本間的對映

在此，讓我們用《般若波羅蜜多心經》（簡稱《心經》）的版本作例子來說明各版本的表達與彼此間對映的方式。依周止菴的蒐集，《心經》現存共有十五個不同的版本；其中有八譯本由梵譯漢而得，有二譯本為梵文之音譯，有五譯本則是由梵譯為藏，再由藏譯為漢。【註八】今取鳩摩羅什及玄奘兩譯本為例，原文如表二中所示。

表二：《心經》兩版本原文

玄奘版	鳩摩羅什版
觀自在菩薩。行深般若波羅蜜多時。照見五蘊皆空。度一切苦厄。舍利子。色不異空。空不異色。色即是空。空即是色。受。想。行。識。亦復如是。舍利子。是諸法空相。不生。不滅。不垢。不淨。不增。不減。是故空中無色。無受。想。行。識。無眼。耳。鼻。舌。身。意。無色。聲。香。味。觸。法。無眼界。乃至無意識界。無無明。亦無無明盡。乃至無老死。亦無老死盡。無苦。集。滅。道。無智。亦無得。以無所得故。菩提薩埵。依般若波羅蜜多故。心無罣礙。無罣礙故。無有恐怖。遠離顛倒夢想。究竟涅槃。三世諸佛。依般若波羅蜜多故。得阿耨多羅三藐三菩提。故知般若波羅蜜多。是大神咒。是大明咒。是無上咒。是無等等咒。能除一切苦。眞實不虛。故說般若波羅蜜多咒。即說咒曰。揭諦。揭諦。波羅揭諦。波羅僧揭諦。菩提薩婆訶。	觀世音菩薩。行深般若波羅蜜時。照見五陰空。度一切苦厄。舍利弗。色空故。無惱壞相。受空故。無受相。想空故。無知相。行空故。無作相。識空故。無覺相。何以故。舍利弗。非色異空。非空異色。色即是空。空即是色。受。想。行。識亦復如是。舍利弗。是諸法空相。不生。不滅。不垢。不淨。不增。不減。是空法非過去。非未來。非現在。是故空中無色。無受。想。行。識。無眼。耳。鼻。舌。身。意。無色。聲。香。味。觸。法。無眼界。乃至無意識界。無無明。亦無無明盡。乃至無老死。亦無老死盡。無苦。集。滅。道。無智。亦無得。以無所得故。菩薩依般若波羅蜜故。心無罣礙。無罣礙故。無有恐怖。遠離一切顛倒夢想苦惱。究竟涅槃。三世諸佛。依般若波羅蜜故。得阿耨多羅三藐三菩提。故知般若波羅蜜。是大明咒。是無上明咒。是無等等明咒。能除一切苦。眞實不虛。故說般若波羅蜜咒。即說咒曰。揭帝。揭帝。波羅揭帝。波羅僧揭帝。菩提僧莎呵。

在表二中的經文，其句讀符號參夾在正文中，這是為了讀者易於閱讀的關係。在計算機的經文正字串中，並無句讀符號，是故在計字數或算計字的位置時，勿將句讀符號計入。

如果將此二版本分別存在計算機中，而不表明此二版本間對應的關係，我們認為是無意義的，因為計算機在此情況下實難對它們做內容之比對和做任何進一步的處理。表明版本間對應關係的方法有兩種，其一是依據內容之分段作版本間比對的準繩，例如用科文。我們可以把對映於同一種科文節點的各版本文字來做比較，在下文中，我們將以實例對這個方法做說明。其

次，是用前述之字串運作指令，將版一本改變為另一版本來觀察此二版本之間的關係。茲將鳩摩羅什版改變為玄奘版的程序列如表三中以為參考：

表三：將鳩摩羅什版《心經》轉換爲玄奘版《心經》之程序

1、斷詞取代 〔鳩<'觀世音菩薩'>,'觀自在菩薩'；
鳩<'般若波羅蜜'>,'般若波羅蜜多'；
鳩<'舍利弗'>,'舍利子'；
鳩<'五陰'>,'五蘊'〕

2、取代 〔鳩<18,1>,'皆空'；鳩<64,4>,'色不異空'；
鳩<68,4>,'空不異色'；鳩<86,2>,'菩提薩埵'；
鳩<301,3>,'莎婆訶'〕

3、刪 〔鳩<24,37>；鳩<108,12>；鳩<210,2>；鳩<216,2>；
鳩<257,1>；鳩<263,1>；

4、增 〔鳩<249,->,'是大神咒'〕

茲將此程序之義意說明如後：「斷詞取代」的指令是要求計算機先在文件的正字串中認出某一字串，如例中之'觀世音菩薩'...'五陰'等，並以另一字串，即如'觀自在菩薩'...'五蘊'去分別替入。在此指令中，「鳩」表示鳩摩羅什本《心經》正字串之檔案名稱，如：鳩<'舍利弗'>即表示在鳩摩羅什本《心經》中任何位置上凡是有「舍利弗」字樣者，餘仿此。

「取代」指令是指某正字串中某一位置之字串，以後列之字串取代之。例如：「鳩<86,2>,'菩提薩埵'」表示鳩本《心經》中第86個字開始長度為2之字串（即'菩薩'），用'菩提薩埵'取代。而刪〔鳩<24,37>〕則表示鳩本《心經》中第24個字起刪37個字。增〔鳩<249,->,'是大神咒'〕則表示在鳩本《心經》第249個字之前（'－'表示之前，'＋'表示之後）加入'是大神咒'四字。

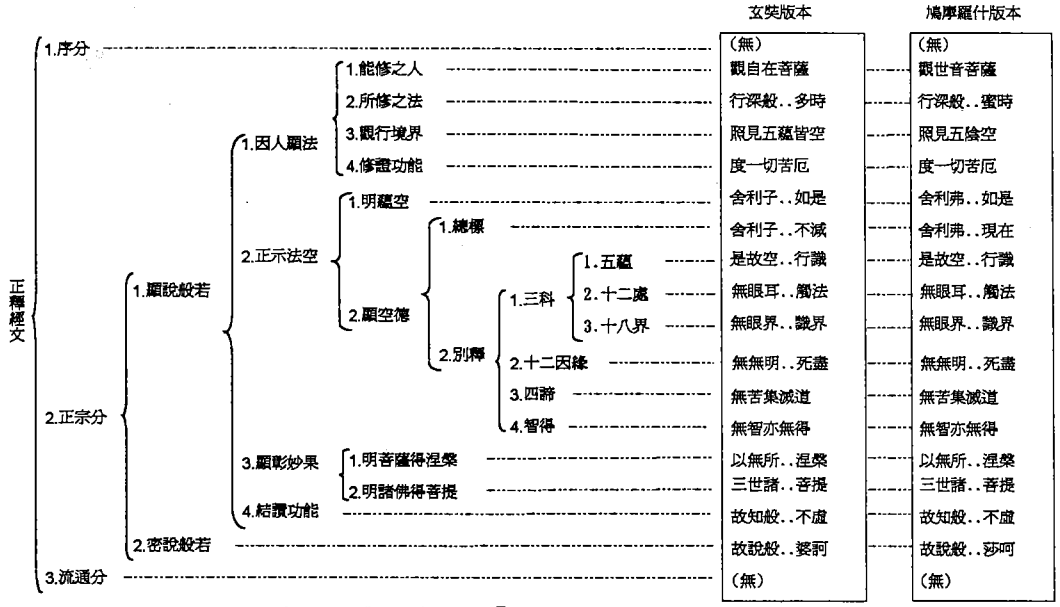
用這種方法，可以十分簡易地將一個版本轉換為另一個版本。是故計算機中若存有一版本之正字串，以及存有將其轉換為另一版本之程序時，則無必要保存第二個版本之正字串。依此類推，則無論有多少版本，只要保存一個版本，並保留這個版本轉換為各其他版本之程序即可。

由於這些版本的文字都是不容改動的，所以版本間換的程序亦可自動地從一個方向的程序產生反方向轉換的程序。例如，有了鳩本轉換為玄奘本的程序，則可自動地得自玄奘本轉換為鳩本的程序。也正因此，若有多版本的情形，各版本間直接轉換的程序，皆可自動獲得【註九】。

其次，在上例中，字串間的轉變均以詞為單位。這是因為考量到使用上給使用者方便的關係。如：以「舍利子」取代「舍利弗」而不以「子」取代「弗」，又如：以「皆空」取代「空」而不是在「空」之前加入一「皆」字等。這樣做的好處是可以產生詞彙的對照表如「舍利子」之於「舍利弗」，產生語意的對照表，如「非色異空」之於「色不異空」....等等。這些性質將大有助於文件內容之比對和檢索。

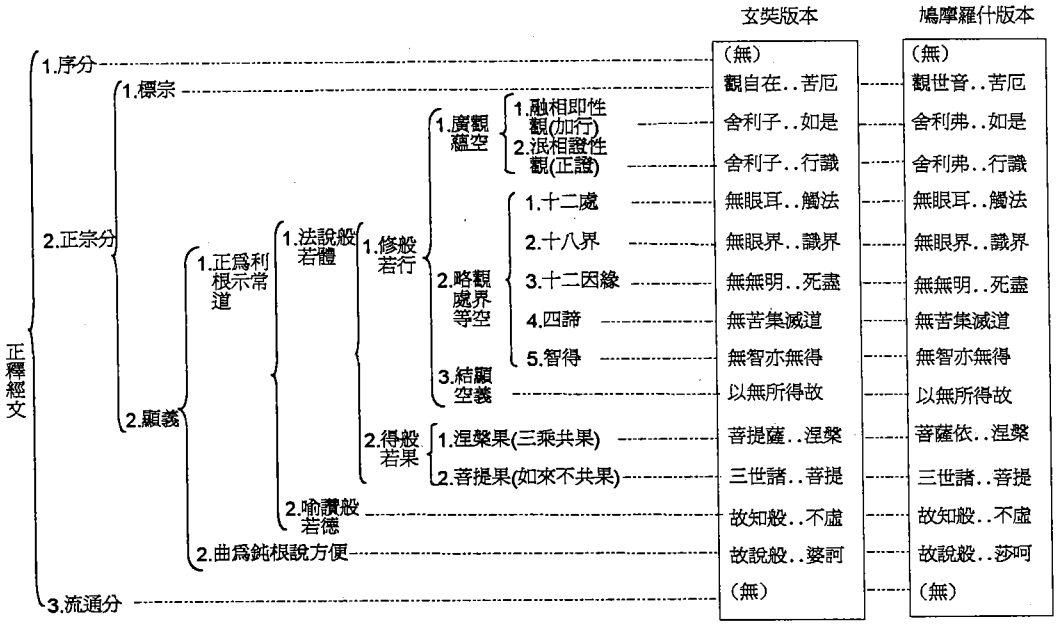
伍、知識結構之利用

前文已談到，所有的佛經內容結構可共用圖一中的科文來解析。此所謂的科文，即是一知識結構。原則上，根據人工智慧領域的發展知識結構種類甚多，本文所採之科文其結構單純，為數學中之樹狀結構。在CDP中並不排除用其他知識結構，例如語意網路等，這類的擴充日後當視需要一一補足，目前，則以樹狀結構代表各式知識結構來做說明。不同的佛經其正文不同之處，則主要呈現在「正宗分」之內；然而，若是異版的佛經，則可共用同一個「正宗分」，讓我們用周止庵和釋印順的心經科文的正宗分來說明。



圖四：《心經》的「正宗分」科文（周止庵）

依周止庵《心經》的「正宗分」科文如圖四所示。在圖四中，正宗分的樹狀結構提要已將鳩本《心經》和玄奘本《心經》剖析得一覽無遺。佛經中所用的科文通常不只一個，圖五則用釋印順的《心經》科文來解析。透過科文，我們可以找出相對應的經文；透過經文，我們可以找出相對應的科文。至於經文之注疏、校勘、評述也可以透過科文而形成文獻間的內容關係網（即構成 Hypertext，並可提供上述各書中註解文字之相互參照）。對照周止庵和釋印順的《心經》科文，可以發現彼此間並非一一對應，而且對照的經文也有不同。如周止庵科文中的「總標」，對照玄奘本為「舍利子。是諸法空相。不生。不滅。不垢。不淨。不增。不減。」；而釋印順科文中的「泯相證性觀(正證)」，對照玄奘本則為「舍利子。是諸法空相。不生。不滅。不垢。不淨。不增。不減。是故空中無色。無受・想・行・識。」。所以任何一個《心經》版本，雖然都可以用任何一個《心經》科文解析，但必須透過一個特定的科文，以找出由該科文衍生而來的注疏。



圖五：《心經》的「正宗分」科文（釋印順）

陸、CDP之雛型

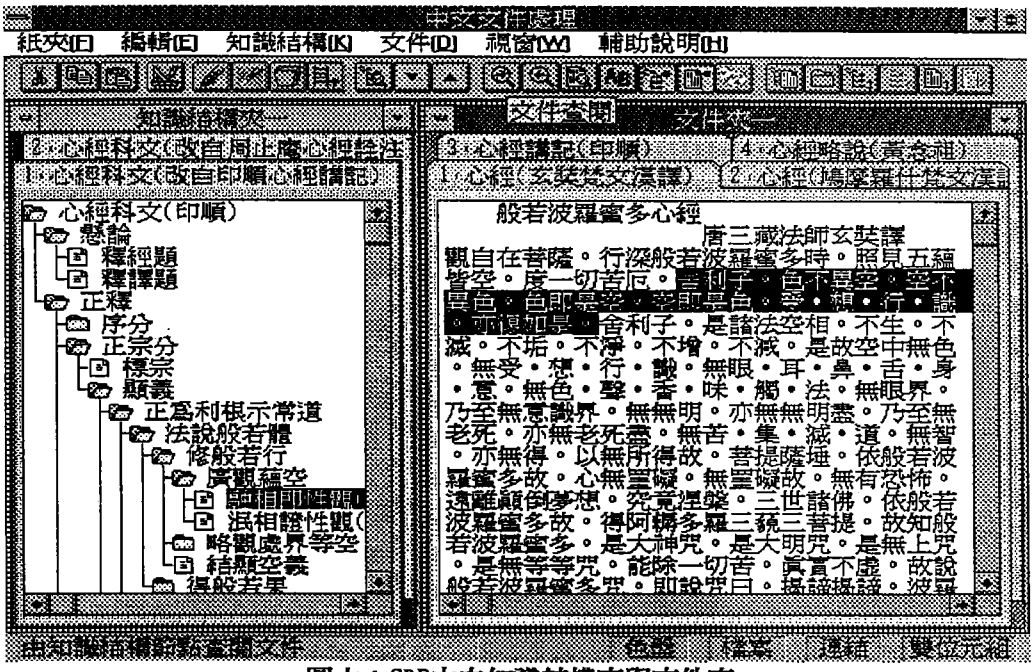
綜合參、肆兩節關於文件字串與科文的討論，可把古籍分析成字串與知識結構兩大部分【註十】。科文為樹狀結構式的知識結構，至於經、疏、鈔、解、講、演等則為不同類別的字串，這些字串於本系統中一概稱為文件。例如《般若波羅蜜多心經講記》可以分解成一個知識結構（科文）及兩個文件（經文與講解）。

為了處理這些知識結構與文件，CDP系統分為知識結構、文件與目錄三大部分（參考圖六、圖七）。知識結構夾與文件夾可存放選取的知識結構與文件，透過目錄夾則可選取、新增、刪除知識結構與文件，並可建立彼此間的關聯。

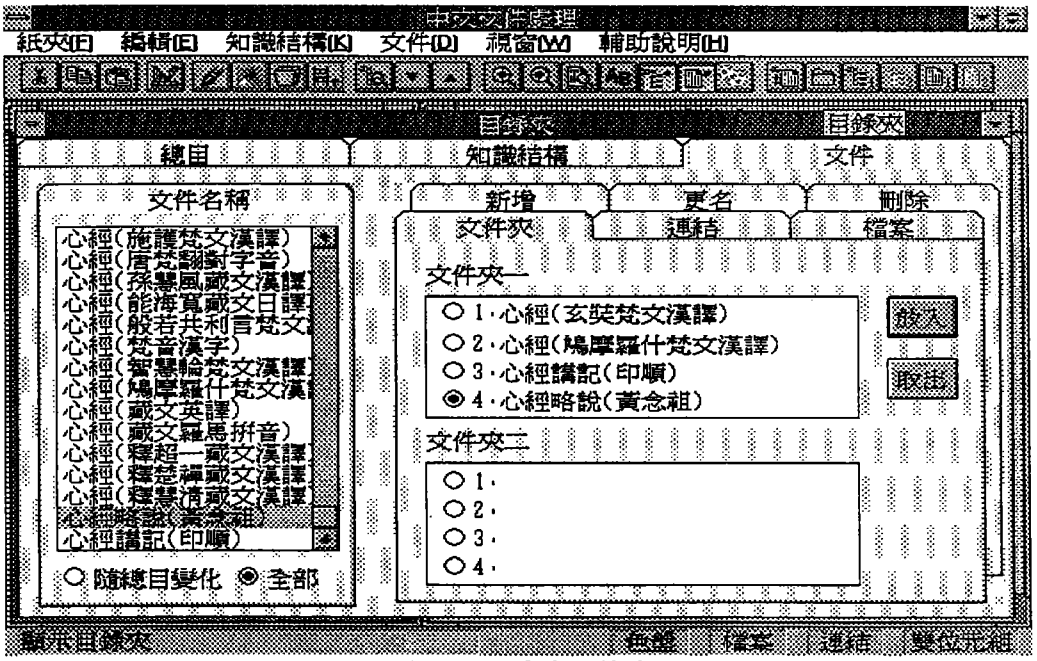
目前CDP中的文件只用了一種表示法，即為輸入字串的形式（如肆之一所述）。換言之，計算機內還沒有一套制式將文件各部份自動劃分的方法。在處理文件時目前是先將文件分為原文、注疏、釋譯等不同之文件，各個輸入系統內，再以知識結構將之鉤連。系統中還有一個編輯器，可新增、刪除及搬移樹狀結構的節點。至於文件字串的編輯，目前須利用系統外的編輯器編輯。一旦選取相關聯的樹狀結構與文件後，即可進行標誌或對照。所謂標誌即是描述樹狀結構節點和文件字串間的對應。

由於一個文件字串可能有兩個以上相關聯的知識結構，標誌時並不在文件字串中加入任何標誌符號，而只記錄對應字串的位置和長度。所以一個字串不只可對應到不同的知識結構，也可以對應到同一個知識結構的不同節

點。一個節點雖然可對應到不同的文件，但是對於一個文件，目前只能對應其中的一個字串，往後系統將會允許對應多個字串。當文件字串有所增減時，原先的標誌就會受到影響，本系統也能對受影響的標誌作等距離修正。



圖六：CDP中之知識結構夾與文件夾



圖七：CDP中之目錄夾

除了知識結構和文件字串的對照外，任何兩個字串也可比較異同。目前系統中尚未建立如表三中的轉換程序，而是即時計算出兩個字串的最大共同子字串(largest common substring)，而這個子字串即暫定為這兩個字串的相同部分。我們也曾經將本系統的標誌轉換成SGML（Standard Generalized Markup Language 即ISO 8879，1986）文件，但是這些轉換功能目前尚未和系統整合。【註十一】至於CDP其他的應用，請參考本會中陳昭珍及王梅玲等發表的文章。

本系統是在微軟中文視窗3.1(Microsoft Window Chinese Version 3.1)上，應用Visual Basic 3.0 撰寫而成。下一步的工作除了加掛中文全文搜尋引擎(search engine)以利全文檢索外，還會再發展新的知識結構以便對文件內容作更深入的分析。至於CDP的外觀和運作方式，則請參觀現場展示，茲不贅述。

柒、結語

本文呈現的，是一種處理古籍的方法。我們用的例子雖然是佛經，但是只要古籍有一較詳盡的目錄，或輯其主題、眉批、評語.....等於原目錄之內，即可構成一類似科文的結構，而即適用於本文所提出之系統。

由於本研究計劃還在進行中，計算機多版本文件的資料結構尚未完全定案，所以在本文中未作詳細之介紹。這部份留待以後再撰文報告。然而，多版本文件資料結構的設計考量，在本文已有詳細的討論和舉例，可資讀者指正。

【註一】 胡樸安、胡道靜，《校讎學》（台灣商務印書館，1990台二版,原書成於1931） 54-84頁

【註二】 張舜徽,《中國古代史籍校讀法》（上海古籍出版社 1986二刷） 181-182頁

【註三】 關於科文，在許多流通佛典內均有詮釋，本文仍根據： 1)周止菴《般若波羅蜜多心經詮注》（台北,華藏佛教圖書館印， 1992） 2)江味農《金剛經講義》（同前，1983） 3)弘一大師，《心經大意》（台灣，精進念佛會印）； 4）寶靜法師輯，《諦閑大師語錄》（台北，新文豐出版公司，1993）第251-269頁，等四書所歸納者；5）釋印順《般若經講記》（台灣，正聞出版社）

【註四】 同【註二】152~154頁

【註五】 同【註五】180~181頁

【註六】 張翠玲等，＜文件結構標誌手冊＞，中央研究院資訊科學研究所文件處理研究室第83001號技術報告，1994

【註七】 關於標誌（markup）部份，可參閱ISO 8879 SGML（1986）。關於文獻之表達方面，可參照謝清俊等＜談古籍之電子版本＞（海峽兩岸中國古籍整理研究現代化技術研討會論文集，北京，中文信息學會，1993）

【註八】 周止菴，《般若波羅蜜多心經詮注》（台北，華藏佛教圖書館印，1992),3-22頁。

【註九】 此證明甚簡單，但因省篇幅故從略。

【註十】 請參考陳昭珍博士論文《古籍超文件全文資料庫模式的探討》台灣大學 圖書館學研究所 1994年12月。

【註十一】 請參照謝清俊《電子佛典中處理方中文版本的方法》中研院 資訊所 1994年5月並發表於第五屆中日韓文獻處理會議（台北）。