古籍校讀工具「中文文獻處理系統」的回顧與展望

莊德明

中研院臺史所研究助技師

derming@gate.sinica.edu.tw

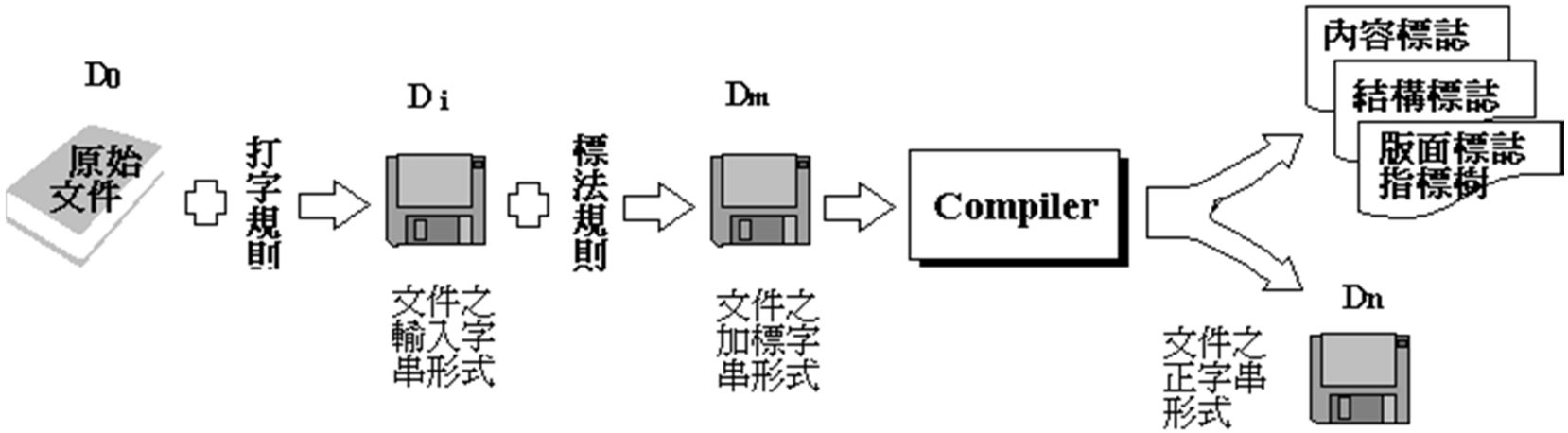
中文文獻處理系統1.0

- 開發時間：1994-1995
- 作業系統：Windows 3.1/95/98/2000/Xp/7(32位元)
- 參考文獻：
  - 古籍校讀工具「中文文獻處理系統」的設計 (http://cdp.sinica.edu.tw/paper/1995/19950722_1.htm)
  - 電子佛典中處理中文版本的方法 (http://cdp.sinica.edu.tw/paper/1994/19940406_1.htm)

中文文獻處理系統的特色

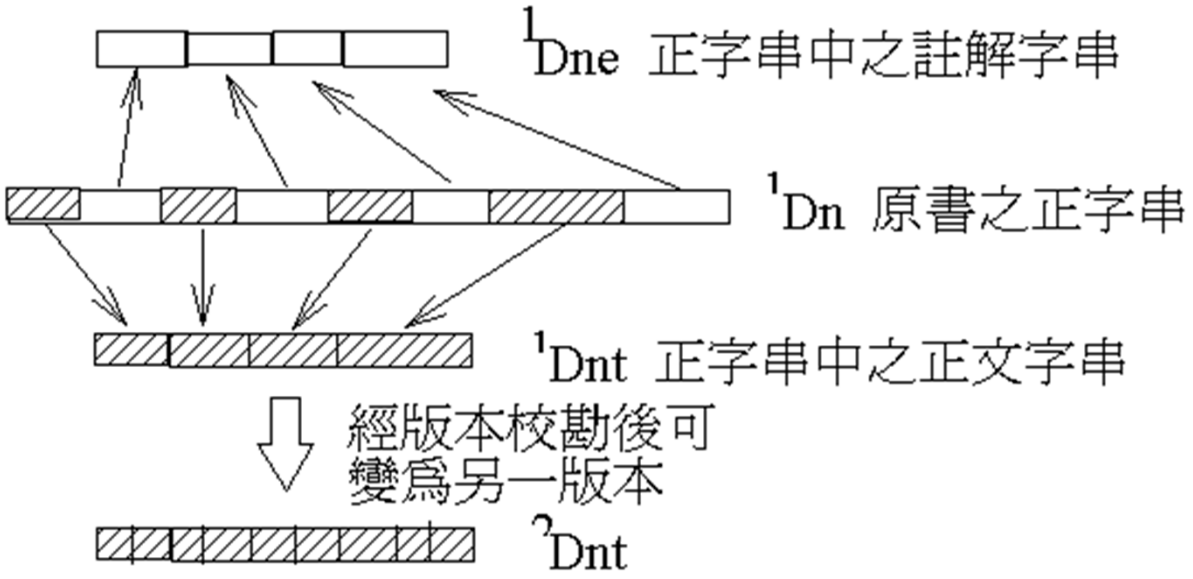
- 是個通用工具，處理的對象並不限於古籍。
- 可處理超文件(hypertext)
- 以文件的字串和某種知識結構間關聯的關係作為超文件的連結(links)
- 把文件內容與相關知識連接成一網路，以提供瀏覽、對映、檢索、參照等功能。

文件在計算機中之表達



- D_0：原始文件
- D_i：輸入字串。文字部份和原始文件相同，但加入標誌符號以載明原始文件頁次等版面特徵。
- D_m：加標字串。於D_i加入目錄或科文等結構標誌、檢索詞或句讀等內容標誌。
- D_n：正字串。將D_m中所有標誌剝除後，所得到的字串。

文件在計算機中之表達(續)



- D_n：正字串
- D_{nt}：正文字串，表示D_n中之正文部份。
- D_{ne}：註解字串，表示D_n中之註解部份。

多版本的表達與對映—以心經為例

玄奘版

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
能除一切苦真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

鳩摩羅什版

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真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

多版本的表達與對映—以心經為例(續)

- 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。
- 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。

多版本的表達與對映—以心經為例(續)

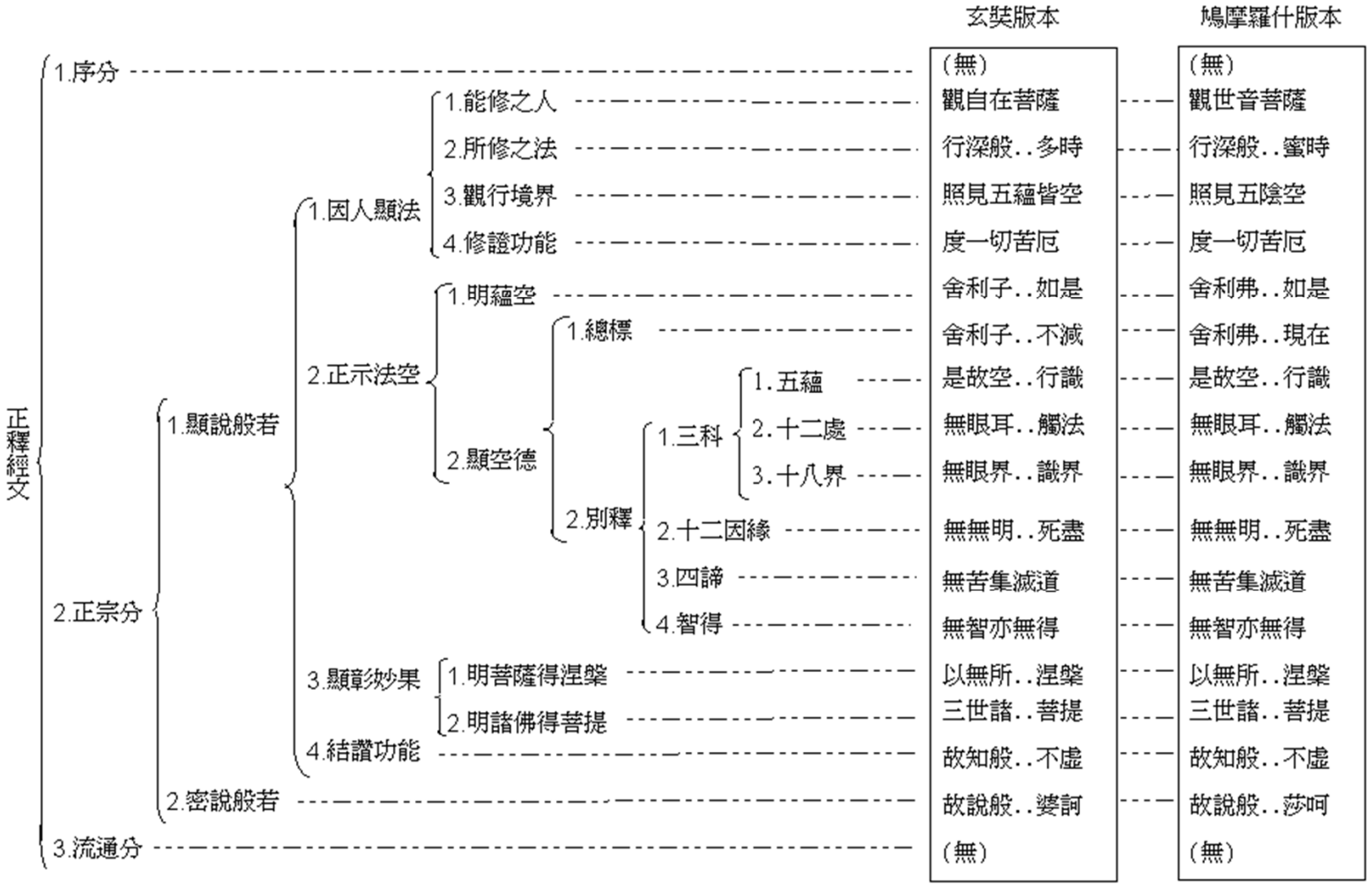
- 將鳩摩羅什版《心經》轉換為玄奘版《心經》之程序

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”
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”
刪	鳩＜24,37＞；鳩＜108,12＞；鳩＜210,2＞；鳩＜216,2＞； 鳩＜257,1＞；鳩＜263,1＞
增	鳩＜249,->,“是大神咒”

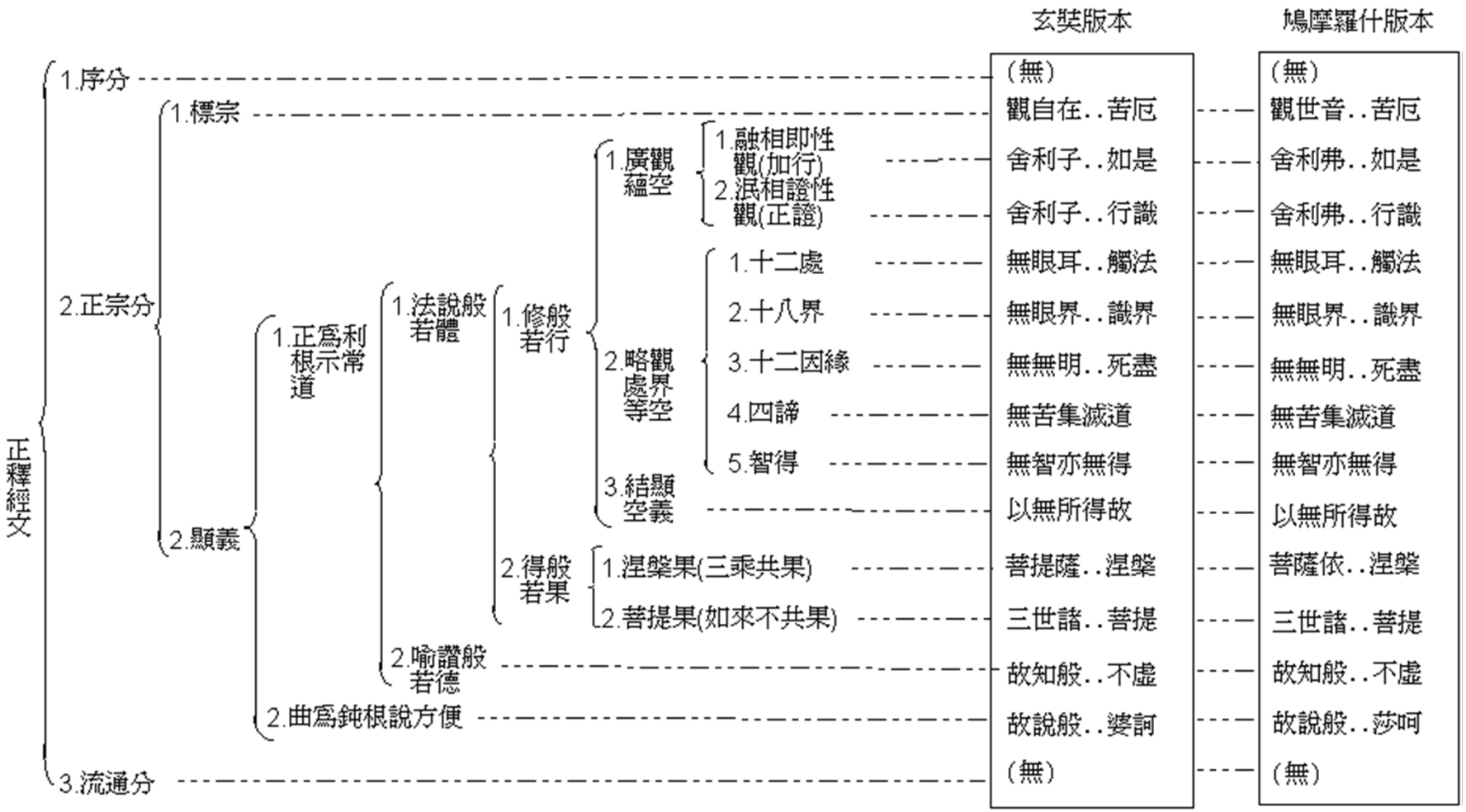
知識結構之利用 — 以佛經為例

- 佛經的內容結構可用科文來解析
- 科文即是一種知識結構
- 科文可用數學中之樹狀結構來表達
- 佛經中所用的科文通常不只一個
- 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。
- 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）。

知識結構之利用 — 以心經為例(續)



知識結構之利用 — 以心經為例(續)



科文的同步標誌問題

- 佛經中所用的科文通常不只一個，必須利用同步標誌(Concurrent Markup)來處理。
- 心經科文示例：

釋印順版	<標宗>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。</標宗>
周止菴版	<因人顯法><能修之人>觀自在菩薩</能修之人>，<所修之法>行深般若波羅蜜多時< /所修之法 >，<觀行境界>照見五蘊皆空</觀行境界>，<修證功能>度一切苦厄</修證功能>。</因人顯法>

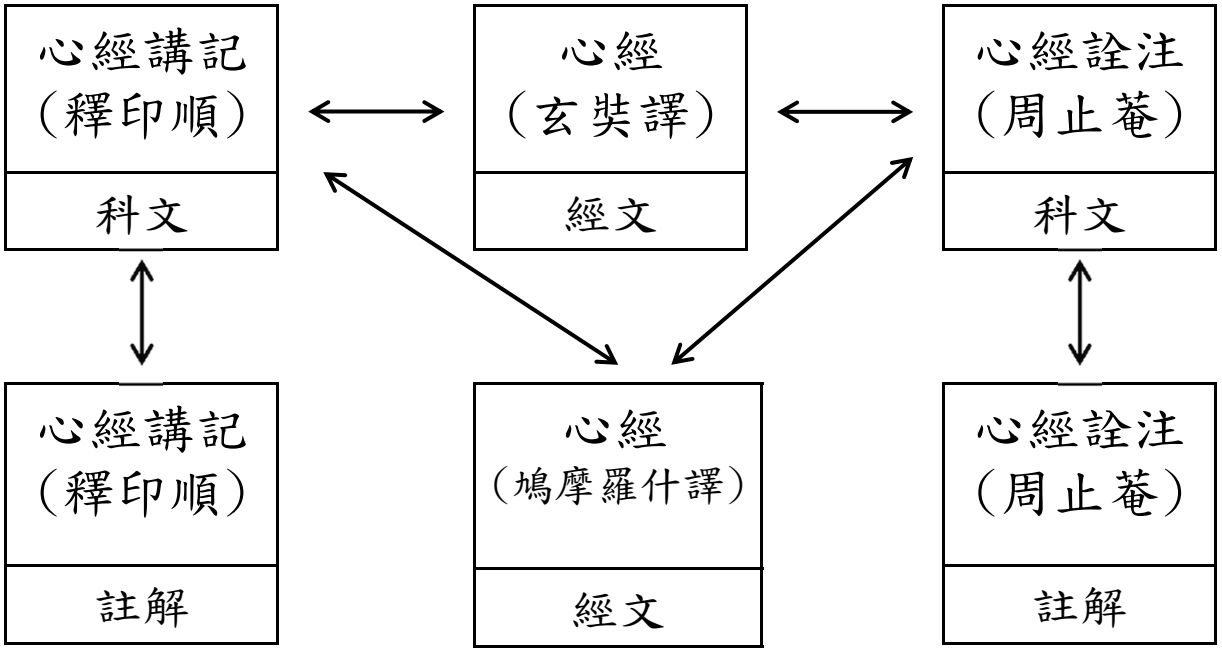
科文的同步標誌問題(續)

- 本系統的科文和經文個別儲存，並以經文子字串的位址來記錄科文標誌，同時解決科文的同步標誌問題。
- 心經科文示例：

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
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

釋印順版	標宗(1,25)
周止菴版	因人顯法(1,25)、能修之人(1,5)、所修之法(6.14)、觀行境界(15,20)、修證功能(21,25)

文件與知識結構的連結


心經講記
(釋印順)
科文
心經
(玄奘譯)
經文
心經詮注
(周止菴)
科文
心經講記
(釋印順)
註解
心經
(鳩摩羅什譯)
經文
心經詮注
(周止菴)
註解

中文文獻處理系統2.0

- 開發時間：2014
- 作業系統： Windows xp/7/8(32及64位元)
- 系統仍在開發中，新增的功能主要增加「語意屬性」的知識結構。
- 下載版功能略同於1.0版，下載網址：http://xiaoxue.iis.sinica.edu.tw/download/cdp.htm

語意屬性

- 目前系統能處理的知識結構，只限於科文類的樹狀結構。
- 從形式上來看，科文類的結構特點在於每個節點只能對應同一份文件中的某個子字串。
- 目前擬再增加另一類知識結構，這類結構雖然也是樹狀結構，但是每個節點可以對應同一份文件中的多個子字串。
- 科文類的樹狀結構可用來描述文件內容的前後關係或整體結構，新增的知識結構比較像是文件內容的意義分類，例如人名、地名等。

語意屬性(續)

- 科文類的結構可暫稱「文章結構」，新增的結構可暫稱「語意屬性」。
- 同一份文件可有多個文章結構，也可有多種語意屬性。
- 語意屬性如同文章結構，都是個別儲存，並以文件子字串的位址來記錄標誌。
- 語意屬性的進階應用，仍在探索中。

幾點觀察

- 小(眾)型 全文資料庫的時代即將到來，但須解決資料庫間的整合及連結問題。
- 文本意義的探究逐漸受到重視
- 內容標誌的比重將會增加
- 社群網站將導引小型全文資料庫的發展，尤其是佛學社群網站。

作法之一

- 開發中文文獻處理系統的網頁版
- 網頁版須優先考量權限管理
- 資料的版權問題須先釐清
- 儘可能採用自由軟體的開發模式

作法之二

- 「中文文獻處理系統」可作為個人端的資料整理工具
- 伺服器端的資料庫整合，小學堂文字學資料庫(http://xiaoxue.iis.sinica.edu.tw/)的架構可為選項之一。
- 小學堂文字學資料庫由甲骨文、金文、楚系簡帛文字、小篆、秦系簡牘文字、上古音、中古音、官話、晉語、吳語、徽語、贛語、湘語、閩語、粵語、平話、客語等資料庫組成；各資料庫除可互相連結外，也可獨立使用。
- 例如可研發心經資料庫，各版本的心經除可互相連結外，也可獨立使用。

結語

- 「中文文獻處理系統」開發時間雖已二十年，但以文件子字串的位址來記錄知識結構的標誌，進而將文件內容與相關知識連接成一網路的概念，仍有獨到之處。
- 古籍全文資料庫的建置雖然日益完整，但仍不足以處理文本的意義，這也是本系統重新開發的契機。
- 這二十年來，網際網路的普及與社群網站的崛起，有助於此系統的推廣與成長。