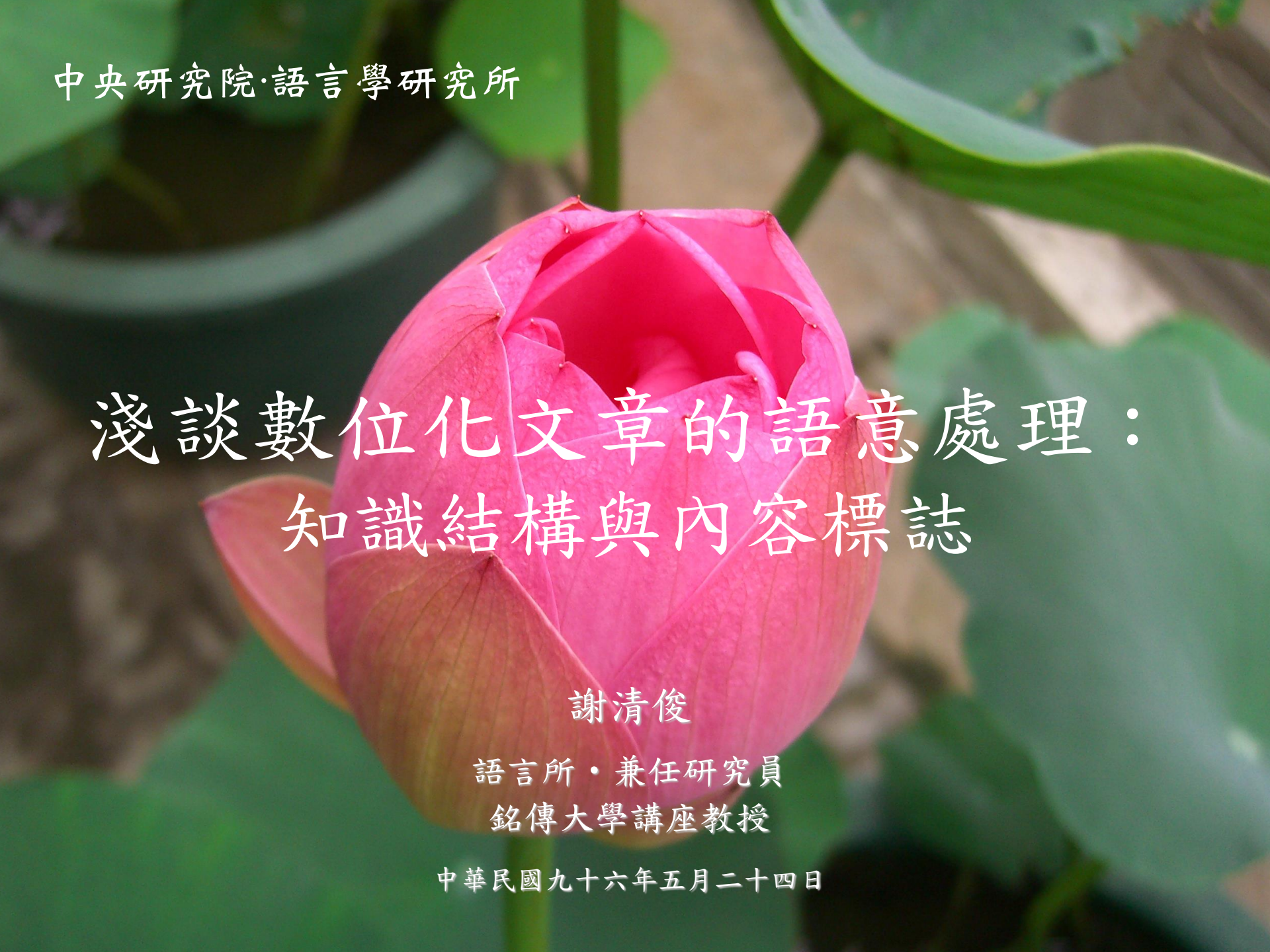
中央研究院·語言學研究所



淺談數位化文章的語意處理：知識結構與內容標誌

謝清俊

語言所・兼任研究員

銘傳大學講座教授

中華民國九十六年五月二十四日

大　綱

一、從數位化看文章生態的變遷

二、文章背景的表達：後設資料

三、文章意義的表達：內容標誌

四、演示：《心經》版本與內容標誌

五、結語

從數位化看文章生態的變遷

- 數位化改變了溝通的生態，擔任溝通中介的文字紀錄或文章，其生態也必然風行草偃。如：
  - 網際網路上「輕薄短小」的文章。
  - 迎合青少年味口「圖多字少，膚淺花俏」的圖文夾雜。
  - 有關閱讀習慣和認知行為變遷的研究報告。
  - 電腦中數位化的文章。
    - 引起文章定義（界定）和範疇的問題。

文章定義與範疇的問題

- 文章經數位化存在電腦中時，只存文章的信息是絕對不夠的，必需把一些有關背景的情境資料也存起來，並與文章作適當的連接。所以，我們習以為常的文章，到了電腦裡就必需含蓋文章情境的某些信息，並且要和文章構成一個整體。
  - 互為文本(Inter-textuality) 理論
    - ***Julia Kristeva***
  - 文章內容與外界的聯繫
    - ***hyperlinks***

表現內容的手法與語文的應用

- 呈現 (presentation)
  - 輸出至個種周邊設備, 如螢幕、列印設備…
- 內容之外化
  - 標點、句讀、標題、章節段落…
  - 位置、字體、色彩、加網加邊、美工加工…
  - 標誌 (markup)與內容標誌 (content markup)
    - 通用結構 (DTD)
      - 版面的、結構的、語文的、內容上的……
    - 標籤集 (tag set)
  - 後設語言(meta-language)和後設資料 (metadata)

數位化文章的界定

數位化的文章，該不該包含傳統文章外的情境信息？

- 將文章與情境信息合為一體的作法：
  - 文章與其後設資料
    - 作者、寫作日期，出版處，發行人……
  - 書籍：以前出書不必提供書目資料，現在則必需提供
  - 投稿需附檢索詞、摘要、作者的信息等
  - ……

這就對文章的界定產生了疑惑，使文章的概念變得和以前不一樣了

- 在電腦中只有文章本身是不夠的，必需和情境信息一起打包，文章才算完成。

情境（context）

- 許多人以為情境只是「上下文」，其實「上下文」是context 在語言修辭「情境」下的意義。
- 情境對文章而言，可泛指文章作成時所有相關的背景，包括：
  - 與其他文章相關的背景
  - 作者相關的背景
    - 如：作者生平、成書時間、著作時的身心狀態……
  - 時代相關的背景
    - 如：政治背景、經濟背景、軍事背景……
      - 是承平還是戰亂、天災人禍、社會重大事件……
  - 文化相關的背景……

情境（context）

- 一旦作品完成，情境信息即已固定，且恆久不變。
  - 此所以文物為文化之記錄。
- 意義是依情境而定的。情境既已固定，則作者創作的原意亦隨之固定。
  - 但閱聽者可依閱聽的情境作詮釋。
    - 『作者已死』
      - 羅蘭・巴特（釋義學）
    - 詮釋是一種創新。
- 若無法描述情境，則無法真正處理文章的意義。

情境的處理

- 因為情境信息必需不分國家、種族，甚至於不分電腦機種都要能夠處理，所以需要用一種電腦會處理的通用人工語言(artificial language)，也就是後設語言，來描述。
- 後設語言不僅僅可以描述情境信息，文章內容的注疏、註釋，以及文章之間彼此的參照，甚至於文章內容與實物之間的聯繫關係等，也都可以用後設語言描述 。

數位化文章的表達

- 文章的結構在電腦中產生了根本的改變：變成以自然語言和後設語言相輔表達的雙重結構：
  - 以自然語言寫文章本身
  - 以後設語言描述數位化的文章與外界的各種關係。

數位化文章信息之表達

數化之文章		表現系統
文章本身		自然語言
文章與外界的關係	情境描述(*metadata*)	後設語言
	參照聯繫(*hyperlinks*)	
	內容詮注(*content markup*)	

後設語言

- 有些學者認為，後設語言越來越重要：往後的年輕人除了母語外，其次最重要的不是外語，而是後設語言；

因為它是超越國界、不分種族、文化，能跨越時空、讓你充份表達思想、意念的人類共同工具 。

後設資料

後設資料的認知

- 有人引據國外的文章，說後設資料就是「資料的資料」。
  - 有了這樣的說法，許多人便認定：

    「除了文物數位化的本身之外，所有其他的資料都屬後設資料」。
- 其實，這樣的認知是有問題的！

後設資料的認知

- 說後設資料是「資料的資料」，只是為了闡明後設資料這個概念的性質，並不是將後設資料定義為「資料的資料」。
  - 因為，後設資料固然是「資料的資料」，可是並不是所有的「資料的資料」都是後設資料 。
  - 將後設資料界定為「資料的資料」這種認知，與「不吃豬肉的都是回教徒」犯了同樣的邏輯錯誤。

後設資料的範疇

- 現行的任何後設資料，其
  - 表達的方式
  - 訂定的規格，以及
  - 標籤 (tag) 或欄位 (field) 的選擇和數目…等

  都限制了後設資料的範疇。
- 這很明顯表示：
  - 不是所有的「資料的資料」都是後設資料。要明白數位化的後設資料，不能把資料二分為資料和「資料的資料」這樣籠統的概念去理解。

時下後設資料的性質

- 目前的後設資料都是為了某類文物訂定的。
  - 比方說，書目資料是一般書籍的後設資料，玉器、青銅器、畫作、雕刻……等都有各自的後設資料。
- 後設資料既然是描述「某類」文物的資料，那麼就有它的特徵和它的侷限。
  - 它適合敘述某類文物的共同現象（共相）
  - 既是共相的敘述就無法顧及個別現象（別相）
    - 後設資料充其量只能摘錄文本的一部份，而無法深入觸及文本的內容。

時下後設資料的性質

- 一般而言後設資料敘述的多屬事實、屬性這類較客觀可考的資料(共相)，不涉及文本內容的理解、感受、比較、批評，以及詮釋等(別相)

所以：

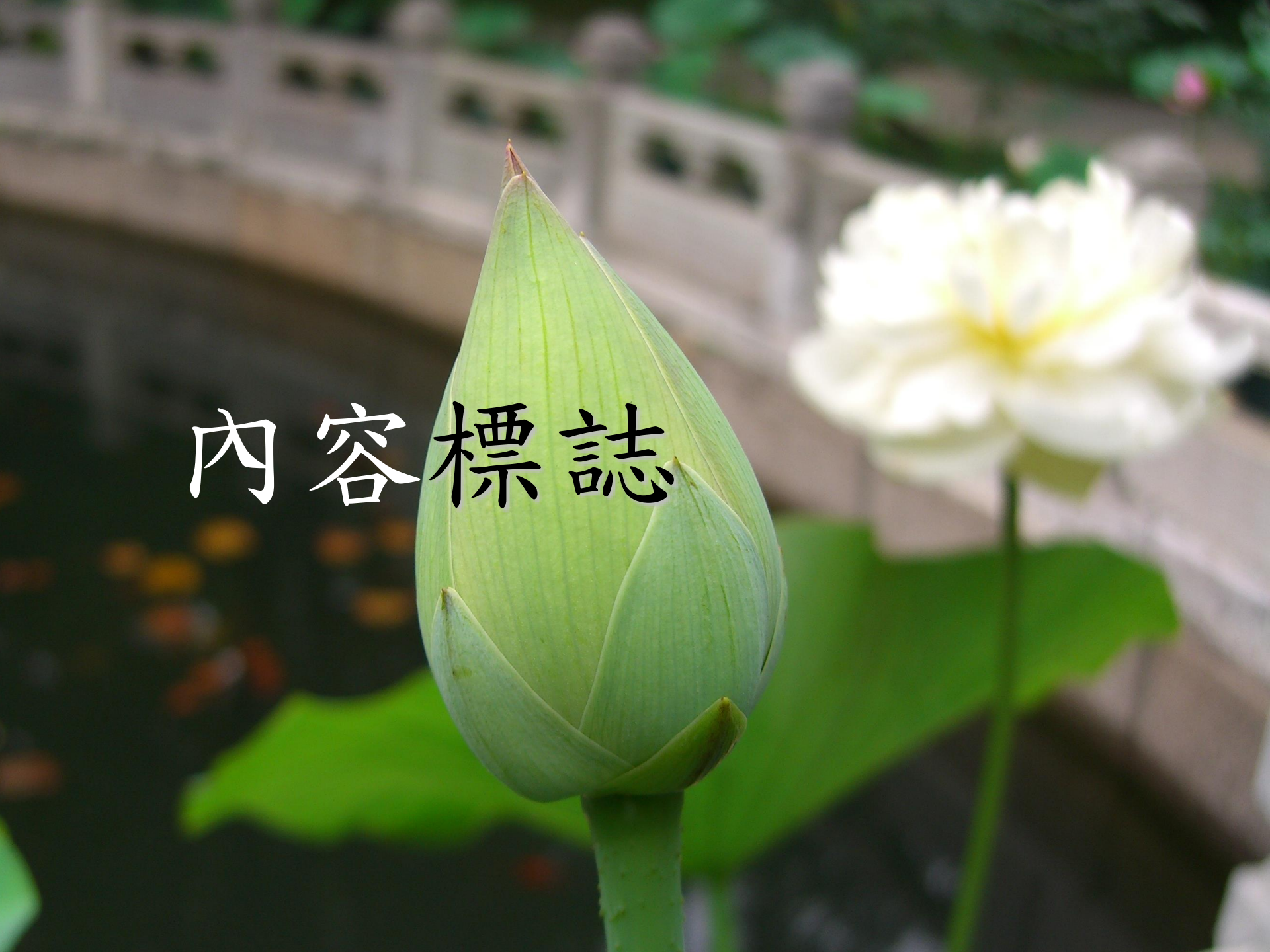
- 後設資料是可以由具技術專業人士查訪、考證
- 但是，它不可以作詮釋。
  - 比方說，我們可以考證《紅樓夢》的作者是誰，卻不能詮釋《紅樓夢》的作者是誰。

時下後設資料的性質

- 後設資料的內容是依應用的目的而異，一件數位文物的後設資料可以有許多種。例如新聞稿：
  - 對記者而言，有一種後設資料；
  - 對報社來說，同一則新聞有編輯用的、管理用的，甚至於是與其他通訊社交流用的各種後設資料。
  - 這些後設資料之間，會有些重複，但也有獨特之處。
- 所以，應用時會要求後設資料的獨特者能彼此互通，重複者需彼此一致。

時下後設資料的性質

- 人世間的事情常有變化，所以後設資料不會是固定不變的，它會與時遷移，需要花很多力氣更新、維護和保養。
- 後設資料既然如此複雜，就不是電腦常用的欄位結構可以處理的。所以，描述後設資料現在都用後設語言(meta-language)，如：HTML、XML。
- **只有語言才有能力描述後設資料的種種規格和後設資料之間的相容關係，以符合應用的需求。**


內容標誌

內容標誌 (content markup)

- 內容標誌要照顧的正是後設資料無法觸及的一關於文物個別內容描述的這一部份。
- 以文章而言，對文本內容的理解(解釋)、感受比較、批評、詮釋等，正是內容標誌的主要工作。
  - 這些工作觸及人文、歷史、社會、美學、哲學…等學門的核心問題, 需要真正了解內容的專業人士為之。

內容標誌

內容標誌，無論作理解(解釋)、感受、比較、批評或詮釋，均觸及一個人文方面最根本的問題—意義(meaning)和了解(understanding)。

- 這是認知科學、語言學、記號學等近幾年來致力研究的重點，也是電腦迄今未能處理的痛處。
- 內容標誌正是為了解決這個困局而設：
  - 一個人機合作的構想：
    - 由人(專業人士)負責意義和了解的部份，
    - 再由機器來處理其餘的工作。

例：內容標誌與意義表達

- 標點符號改變了文章內容的表現方式：
  - **無標點符號的文章內容較為隱晦(implicit)—需經分析、理解的過程才能窺見原意。有了標點符號，則內容較外顯(explicit)，諸如：**
    - **私名號的使用已明顯的標出姓名或機構名稱，減少了斷詞的工作**
    - **句點、逗點、分號等則已將斷句標明。**
- 標點符號將部份文章內容由隱晦轉為外顯，這就是一個「意義表達」的例子 。

內容標誌與意義表達

古文雖然不用現代的標點符號，然而有另一套常用的標誌系統：句讀 。

- 句讀主要用途是作文章內容的標誌：
  - 標明文中之美辭、佳句、警句，或文中之不佳之處等。
  - 對詩詞韻文，也有用於標示韻腳和朗誦時的間歇者。

標點符號或句讀這類的標誌，都是設計來幫助讀者理解文章內容的；它也幫做標誌人，把他們對文章的理解用標誌記錄下來。

時下的文章標誌 與 內容標誌

- 時下電腦中的文章標誌是近年來西方資訊科技發展出來的，而漢語文獻的標誌，則是我們先人留下的智慧。
  - 目前的電腦標誌多對文章的外形作標誌。
    - 通用結構 (DTD)
      - 版面的、結構的、語文辭彙的……
      - 標籤集 (tag set)
  - 傳統漢語文獻的標誌則側重於文章內容的、意義上的標誌。

意義處理的問題

- 近年來，計算語言學和人工智能均致力於處理意義的研究，也取得一些成果。例如，詞網(word net)、主題圖(topic map)、知識本體結構(ontology)等。
  - **它們將詞彙間的關係在電腦中作了適當的表達(representation)，並構成資料庫和研發為數位工具。**
  - **詞彙間的關係是語意中的一種，將它數位化，對意義的處理是有助益。**
    - **可是助益有限，並沒有突破性的進展。原因是囿於形式和內容(意義)是一對一前提，因此，並無能力處理意義的癥結—多義問題(ambiguity)。**

意義處理的問題

- 多義問題，簡單說，就是當一種形式可能對應到好幾種意義時，如何作正確選擇的問題。此即「義隨境轉」語意隨情境而轉移的現象。
  - 例如，作數目字時，「十、拾」通用，可是情境變為「路不拾遺」時，就不可以作「路不十遺」。
  - 人面對多義或義隨境轉問題並無太大難色，所有的自然語言都有濃厚的義隨境轉色彩，因為人多半了解情境，對「意義」會作適當的「了解」。所以，電腦處理意義問題的先決條件，是要會表達情境。可是目前學界在這方面的努力，還沒有顯著的成績。


容標誌之例《心經》

《心經》內容與意義的標誌例舉

- 心經各版本與其科文間內容的對應
  - 給人用的標誌介面。
  - 標誌的彈性：
    - 無固定的標籤集。
  - 內容與結構(ontology)間的對應：
    - 對本文的詮釋。
  - 版本間內容的對應：
    - 對本文呈現的不同形式。
  - 結構間的對應：
    - 對本文不同的詮釋。
  - ……

《心經》內容與意義的標誌例舉

- 以上展現的內容處理，都沒有用到檢索常用的詞語聯繫(morphological linking) 結構。
- 以上的展現可呈現有個別差異的「別相」
- 請注意意義聯繫間的彈性：
  - 跨語文
  - 跨作者
  - 跨版本
  - 知識結構的延伸
  - 各種意義之間的聯繫關係……

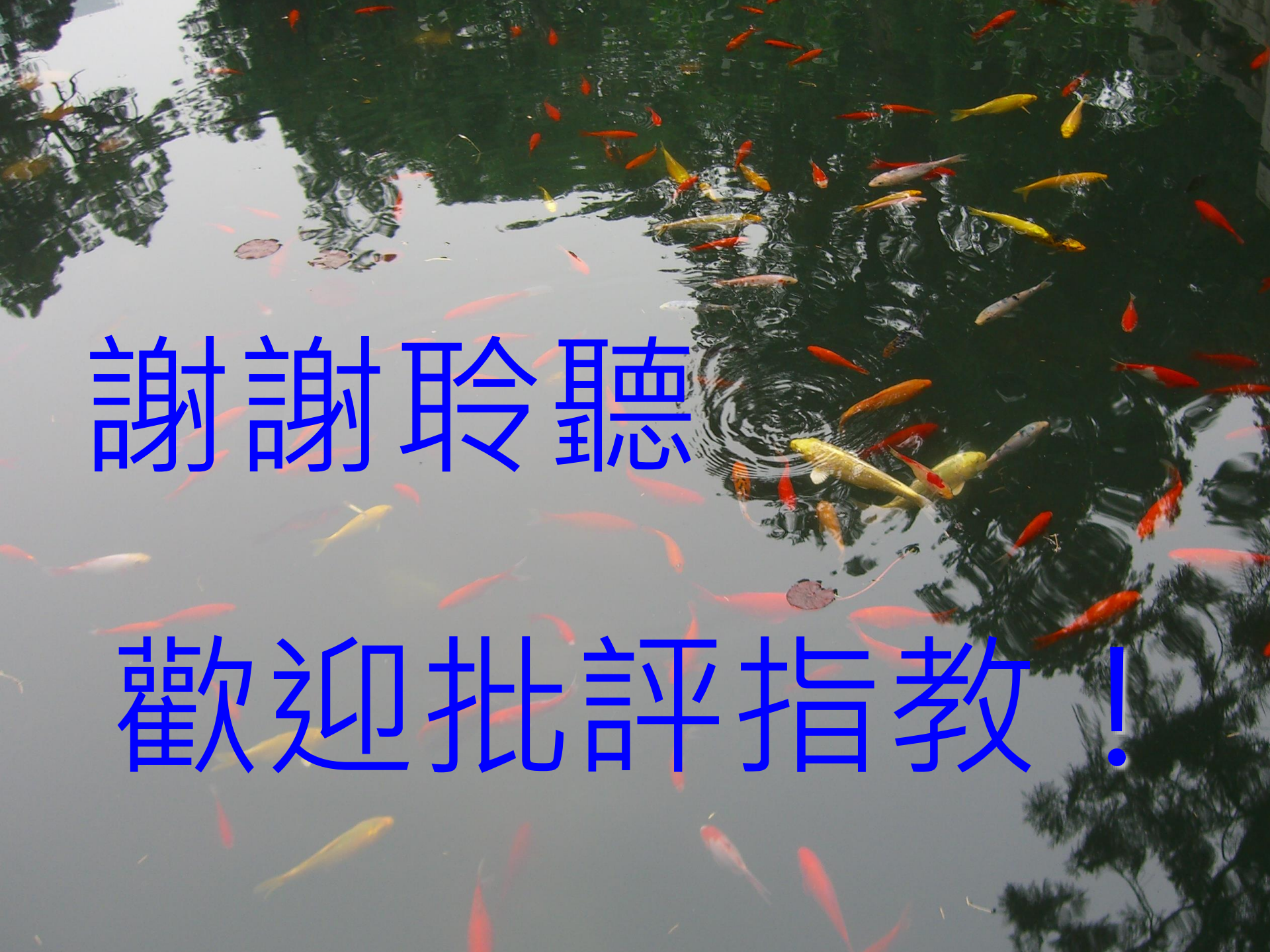

結　語

後設資料和內容標誌

- 後設資料和內容標誌兩者都是不可缺的，彼此並不相互排擠，它們是兩種類型完全不一樣的工作，可以相輔相成、相得益彰。
- 若認為：除了文物數位化的本身之外，所有其他的資料都屬後設資料，那麼就犯了不可原諒的大錯—它扼殺了內容標誌生存的空間。

意義處理的問題

- 意義(meaning)是可以用電腦處理的。
  - 內容標誌即文章意義的標誌。
  - 從《心經》的例子可知：情境在電腦中可以表達。據此可以設法處理多義或歧義的問題。
- 未來，電腦可能以兩種方式來處理意義問題：
  - 其一是將所有的多義關係轉化為單義的語法關係。例如建立「常識庫」讓電腦能辨識「情境」。
  - 其次是與人合作，以人機共建的系統來做「了解」和處理「意義」問題。
    - 這就是內容標誌要做的事。


謝謝聆聽
歡迎批評指教！