後設資料與內容標誌—
淺談數位化的文章和其語意處理

謝清俊 ***951221***

壹、文章的生態變遷

人人都知道，數位化改變了溝通(communication)的生態。溝通的生態既然改變了，擔任溝通中介的文字紀錄，其生態也必然風行草偃。觀察文字紀錄或文章的生態變遷，首先想到的可能是網際網路上「輕薄短小」的文章；或許是迎合青少年味口「圖多字少，膚淺花俏」的圖文夾雜；也可能是有關閱讀習慣和認知行為變遷的研究報告。關於這些，已有許多研究和專文報導，本文不談這些，僅從數位資訊的情境，來看文章生態的改變。這是件既有意義、又有趣的事。

首先看一個事實：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構成某種資料結構，以利往後的處理和運用。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。這些文章情境的信息，是依管理文章的需要或配合文章的應用情況而決定。比方說，作者、寫作日期、出版處、出版日期、發行機構……等皆是文章情境的信息，而這些也是讀者、書商、圖書館，或建構資料庫常用到的。談到這裡，引發了一個文章範疇的問題：所謂文章，該不該包含傳統文章外的情境信息？這就對傳統文章的概念產生了疑惑，並考慮修正文章的操作定義(operational definition)，以適應工作情境的變遷。

· 文章與情境信息

我們用紙筆寫文章，已經歷了二千多年，早就熟知文章就只是文章，不須顧及情境的信息；整理情境信息是讀者或是後人的事。然而，數位化以來，越來越將文章與其情境信息結為一體的趨勢。撇開數位化工作不談，諸如：以前出書不必提供書目資料，現在則必需提供；現在投稿需附檢索詞、摘要和作者的信息等，這都是將文章與情境信息合為一體的作法。這做法是為讀者、管理者、媒體工作者和後人設想，有了這些情境信息使大家都方便。然而，對作者而言，文章就不只是文章本身了，還需提供必需的情境信息。換言之，作者的工作增加了：只作文章本身是不夠的，必需和情境信息一起打包，文章才算完成。

最初遇到情境(context)一詞，很可能是在探索字詞語意的時候；當我們不明白文章中某詞語的意義時，最直接的方法是看看該字詞的前後文。此時，「前後文」就是該詞語在這篇文章中最直接相關的情境。如果參考了前後文，仍然弄不清楚該詞語的意義，那麼，可以看看前一段或後一段、前一節或後一節、前一章或後一章，乃至整篇文章、整本書，這些也都是該詞語所處的情境。若看完了整篇文章、整本書，還是不明白，進一步可參閱其他相關的著作，或是找些作者生平、成書時的時代背景和作者當時的遭遇、相關的事件等。這些也還是情境。所以情境信息有許多種，小到可以僅僅是前後文、

作者的識別信息、寫作時的意圖、身心狀態等；大則可能包括文章寫作時的社會情境、政治情境、文化情境。

為什麼要參考這麼多各式各樣的情境？答案很簡單：為的是要了解語意，想弄清楚作者真正想要傳達的信息；因為意義是隨著情境而有所不同、而變化的。這就是所謂的「意隨境轉」。一般而言，科學作品的情境較狹窄，只需了解其基礎的科學知識便已足夠；可是人文作品的情境卻如前述的那麼複雜。比方說，為了解牛頓的力學公式：**F**(力) = **M** (質量) x **A** (加速度)，我們不需了解牛頓的生平，也無需知道那時他是不是和萊不尼茲爭執些什麼；可是，閱十三經、佛經，看詩詞、《古文觀止》，讀《紅樓夢》、《戰爭與和平》等，就不能忽略作者的生平和成書時的時代背景和作者著書時的境遇。

· **數位化文章的雙語結構**

當文章存入電腦時，如前述，必要的情境資料也隨之存入。文章和情境資料本身當然是用我們日用的語言(自然語言)寫的，然而情境資料在電腦中的結構以及它與文章彼此之間的關係，就不能也用自然語言描述，因為電腦沒有能力和人一樣了解自然語言；如果電腦不了解，就沒有辦法為我們做事。再者，這些情境信息必需不分國家、種族，甚至於不分電腦機種和軟體版本等，都要能夠處理。因此，必需要用一種在幕後支撐性質的通用後設語言，即描述自然語言的人工語言(artificial language)，來向電腦說明情境資料的結構以及它與文章彼此之間的關係。

後設語言，如 SGML(標準標誌語言)、HTML(超文本標誌語言)、XML(可延伸標的誌語言)以及標誌實務 TEI (Text Encoding Initiative)等規範，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。所以，數位化文章的結構，可如下表一所示。從下表知，文章的結構在電腦中產生了根本的改變，成為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文章與外界的各種關係。

表一：數位化文章的雙語結構

<table>
<tr><th colspan="2">數化之文章</th><th>使用之表現系統</th></tr>
<tr><td colspan="2">文章本身</td><td>自然語言</td></tr>
<tr><td rowspan="3">文章
與
外界
的
關係</td><td>情境描述（metadata）</td><td rowspan="3">後設語言</td></tr>
<tr><td>參照聯繫（hyperlinks）</td></tr>
<tr><td>內容詮注（content markup）</td></tr>
</table>

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。自後工業社會以來，人們會越來越依賴電腦才能做事，著書立說、發表文章也逃不出這個魔咒。

貳、後設資料與傳統標誌 (mark up)

目前數位文章做得最多、最惹人注目的情境資料，是後設資料(metadata)，它是熟知的一種電腦標誌形式。電腦標誌與漢語文獻傳統的標誌是各異其趣的。本節先討論後設資料的概念和性質，再略述一般的電腦標誌，其次回顧傳統的標誌。

一、後設資料

不太了解數位化的人常常以為後設資料只有一種。這就錯了！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，記者撰寫的新聞稿，對記者而言，有一種著述用的後設資料；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，在實務上會要求獨特者能彼此互通，重複者需彼此一致。

再者，人世間的事情常有變化，且人們通常要求電腦做的事情越來越多，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。一般人不容易想像，後設資料竟有這麼複雜，它的負擔竟有這麼重。

後設資料既然如此複雜，就不是電腦中常用的欄位結構可以處理得了的。所以，以前用欄位結構的形式描述後設資料，現在都改用後設語言，也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。

・後設資料的迷思

也許是數位化風行以來後設資料太惹人注目，以致有人認為：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；也有人引據國外的文章，說後設資料就是「資料的資料」。有了洋人這樣的說法，許多人更認定了：「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，它將原本清晰的事理攪模糊了。

「資料的資料」這句話，前一個「資料」指文物數位化所呈現的形態，後一個「資料」則指：我們要利用「文物數位化呈現的資料」時，必需搭配的一些背景資料；而這背景資料就稱作「後設資料」。稱「後設」的原因是：它不是位居前線（第一線）的主要資料，僅僅是作後勤（第二線）支援的資料❶。

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」犯了同樣的錯誤。這錯誤將於下文說明。

・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，以及欄位（或標籤）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」用這樣自我參照(self-reference❷)的邏輯概念去理解。

既然說：不是所有的「資料的資料」都是後設資料，那麼目前的後設資料都是些什麼？簡言之，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，新聞有新聞的後設資料，玉器、青銅器、畫作、雕刻.....等都有各自的後設資料。

・後設資料的特徵和它的侷限

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。首先，它適合敘述某類文物的共同現象(共相）。目前後設資料中的標籤(或欄位)，像是作者、年代、地點、收藏者、檢索詞、摘要.....都是一些共相的名目。既是共相的敘述，就無法顧及個別現象(別相）。所以，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。了解到這裡，就不難明白「內容的標誌」何指。內容標誌要照顧的正是後設資料無法觸及的─關於文物個別內容描述的這一部份。

其次，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，不涉及文本內容的理解、感受、比較、批評，以及詮釋等。所以，後設資料是可以由具技術專業人士查訪、考證的；但是，它不可以作詮釋❸。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。

二、電腦中的標誌 (markup)

把文章數位化存入電腦時，除了需準備後設資料外，還需要對文章本身做標誌。常做的標誌包括文章的篇章段落結構、圖、表、注釋、參考文獻、版面、乃至超連接(hyperlinks)等呈現的形式，用標誌語言的標籤(tag)把這些**形式**標清楚讓電腦知道，以便電腦能幫我們後續的處理。

這些標誌是直接為電腦做的，其間接的受益者才是使用電腦的人。這種做法是典型的人機合作：電腦無法理解的事由人來做(做標誌)，人們不耐煩做的例行事務交給電腦執行，相輔相成。

目前為電腦做文章標誌，絕大部份還只做到如上所述；對文章內容(content)的標誌，則幾乎還沒有開始嚐試。

三、傳統的標誌 (mark up)

現在，很多人的電子郵件不喜歡用標點符號。行文遇到需要分斷處，就乾脆「換行」。如此一來，信件的形式好像新詩，似乎別有風韻，可是依其內容看來，就顯得散漫、支離。這也許是「換行」的動作比輸入標點符號方便許多的緣故；圖個自己的方便嘛，那管讀信人的感覺。本小節將回顧兩種傳統的標誌：標點符號和句讀，至於另一種內容標誌「眉批」則從略，因其結構略同於注釋。

・ 標點符號

傳統的文言文大多沒有標點符號❹。現行的標點符號是五四運動推行白話文後訂定的。兩千多年來，文言文雖然不用標點符號，並未使讀者誤解了文章的內容。這固然是讀者有比較高的理解能力使然─理解文章時能正確的斷詞、斷句，也是作者使用文字的能力高強的緣故─要高到寫出的文句只允許有一種斷詞、斷句和理解的可能。這是寫文言文最基本的功夫─要能寫出這樣的文章才可避免語意混淆，才能派上用場。

古人為什麼不使用標點符號？是他們不夠聰明？不致如此吧！筆者曾對整個二十五史的全文做過統計：標點符號之間的平均字串長度約 4.5 個字。這個數字可以這麼理解：如果記錄的版面不變，使用標點符號將使該版面的記錄減少約 18.3%；以一篇文章而言，增加標點符號要增加約 22.2%的版面。換言之，一個人的藏書若需 4.5 個書櫃存放，那麼用了標點符號後就要 5.5 個書櫃；如果一個書簡的重量是 4.5 斤，用了標點符號後就成了 5.5 斤。這也許是古人沒有用標點符號的緣故吧。

有了標點符號的確方便不少；至少免除了斷詞、斷句和理解的困擾，使讀寫文章能力的門檻降低許多，讀寫能力也就更為普及。可是這門檻的降低，也意味著寫作能力和理解能力的降低。若無標點符號，讀者需經斷詞、斷句的分析，有時還要參照該文章的情境，才能了解文章的內容。這種使用漢語的本事，在使用標點符號後逐漸式微。

標點符號也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。所以標點符號是內容標誌的一種，有將部份文章內容由隱晦轉為外顯的功用。

· 句讀

古文雖然不用標點符號，然而有另一套常用的標誌系統：句讀。句讀不是作斷詞、斷句用的，它的主要用途是作文章內容的標誌─標明文中之美辭、佳句、警句，或文中之不佳處、敗筆等；對詩詞韻文，也有用於標示韻腳和朗誦時的間歇者。句讀的作用也是使文章的內容外顯。

標點符號或句讀這類的文章標誌，都是設計來幫助讀者理解文章內容的❺，它們的作用有二：對讀者而言，它能將文章隱晦的內容外顯，方便讀者理解文章；對作者或專家而言，則可將他們對文章的理解記錄下來留作後用。

當然，除了標點符號和句讀外，還有其他的設施也幫助文章內容的外顯，如分段、字體變化、版面安排等。然而這些已超出了本文討論的範圍，故從略。

參、電腦標誌與傳統標誌

電腦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。電腦標誌以標誌文章呈現的形式為主，傳統標誌則側重在文章內容的標誌。比較這兩種文章標誌，雖然顯得南轅北轍，然而也顯示出合則兩利的前景。例如，電腦一直對文章內容的處理力有未逮，而漢語文獻的標誌卻正提供了一種處理內容的方法；目前的電腦一遇到語意問題就一愁莫展，漢語文獻的標誌則顯示，文章的語意

是可以處理。這些差異正是二者的互補的明證。本節將比較電腦標誌與傳統標誌的異同。

· 數位化與內容標誌

對文本內容的理解(解釋)、感受、比較、批評、詮釋，以及文本與外界的種種關聯等，正是內容標誌的主要工作。內容標誌可說是了無止境、極為深廣的工作。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來做。此所謂核心問題是指這些學門的本體論、認知論(知識論)、方法論，以及研究方法等，例如近代的現象學、釋義學(詮釋學)。

再者，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。迎著內容標誌而來的正是意義和了解這個問題。是故內容標誌的研究可作為電腦處理意義的基礎，其重要不言可喻。

・電腦與語意處理（意義與了解）

能了解萬事萬物的意義是人獨有的特質。如果人和電腦比本事，似乎唯一人確能超越電腦的，就是能了解事物的意義。比方說，電腦只會做 1+1=2，它不知道為什麼一加一等於二，更不懂一加一等於二的意義。電腦對它所做的事，全都不明白「為什麼」，也都不「了解」這些事情的「意義」。

其實，不僅電腦如此，所有的科學工具、儀器、程序莫不如此。對「意義」和「了解」進行探索和研究的，大多屬人文和軟科學(soft science)的學科，如哲學、美學、現象學、詮釋學、語言學、修辭學、記號學、認知科學、腦神經學等。傳統的硬科學(hard science)早已將心智的功能排除在科學之外，所以科學的工具(如數學、邏輯)、儀器、程序等，都不會涉及意義或了解的問題。例如，數學、統計、邏輯的推演可以很複雜，可以導出前人未見的結果，然而，只有人能夠了解推演結果的意義。能了解事物意義的這個特質，給予人存在的理由，也在社會日益物化、科學化、去人性化(dehumanize)中，維繫住人的尊嚴。

如此說來，電腦似乎無法做「了解」和處理「意義」的工作了，是嗎？其實，不盡然。要言之，電腦可以做一部份相關的工作，但不是全部。要談這個問題，可追溯到十九世紀中期對機器智慧(machine intelligence)問題的探討。

當時，自十八世紀起，歐洲各國因應航海和發展海權的需要，爭相製作高準確度、高正確性的對數表。1830 年左右，機械的計算器(calculator)大放異彩，成功的協助對數表的製作，令許多學者大為驚奇，譽之為有「智慧」的機器❻。可是，機器沒有心靈，為什麼會有智慧？為什麼有些機器有智慧，而有些卻無？機器智慧又是什麼？機器智慧能發展到什麼程度？……這些問題對學者和政客都是巨大的衝擊；於是，展開了一連串關於機器智慧的探討。

現在看來，當時的學者沒能了解智慧型工作(像四則運算)都包含形式和內容(意義)兩部份。以四則運算為例，機器無法了解其意義(內容)，卻無礙於機器處理其形式。再

者，四則運算的形式和意義是一對一的關係，這情形使得機器處理形式的情況，就如同在處理意義。形式處理好了，意義問題也隨之解決。這真是「歪打正著」，也可說是「狐假虎威」。

從語言學來看上述的問題：語言現象分為語法(syntax，語言的形式)和語意(semantics，語言的內容)兩部份。如果形式和意義的關係是一對一，那麼其語意問題（內容問題）就已化約為語法問題（形式問題）。換言之，語意問題沒有了，它已併入語法問題中，當電腦解決了語法問題，也就「順便」解決了對應的語意問題。然而，此時電腦還是不「了解」該形式（語法）的「意義」，當然，也不了解被它解決的語意問題。所以，嚴格說，電腦本質上是不會直接處理意義問題的。

若不能了解上面的分析，就會誤以為計算器具有了解四則運算的意義和處理四則運算的智慧。其實，它能做到的都屬形式的部份。

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)、語言知識的本體結構(linguistic ontology)等資料庫和工具的創立。能有這些成果的主要原因，應歸功於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；因為，所有這些研究的基本結構，還是和計算器做四則運算一樣，囿於形式和內容（意義）是一對一前提，並無能力處理意義的癥結─多義問題(ambiguity)。

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。比方說，當一種語言形式對應著二種或更多的語意，則此形式可能在某情境下它指向其中的一種意義，而在另一種情境下卻指向另一種意義。此即前文已說過的「義隨境轉」。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。二十多年前，有學者悲觀的指出：要電腦能處理意義是不可能的，除非能研製出「常識庫」。「常識庫」是「情境」的另一種說法，因為一個人判別情境常需要用到一生累積的常識，聚集眾人的常識才能構成「常識庫」。目前，「常識庫」依然是可望不可即。也正因為如此，人文學者常說：『「了解」須以整個生命的力量為之。』

未來，電腦可能以兩種方式來處理意義問題：其一是逐漸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。

結語

後設資料和內容標誌並不相互排擠，它們是兩種類型完全不一樣的工作。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。換言之，後設資料和內容標誌兩者都是不可缺的，且彼此相輔相成、相得益彰。

做文章內容的標誌，不是電腦工程師可以做的，需要了解文章內容的專家來做。這情形正好提供人文學者一個絕佳的機會加入文獻數位化的行列。如果人文學者能用標籤把他們的知識，也就是對文章的理解、真知灼見，表達給電腦知道，那麼，久而久之電腦將匯集大量的人文知識。果真如此，那麼，一種嶄新形式的人工智能（artificial intelligence）即將誕生。且讓我們拭目以待。

註：

❶ 後設一詞由「meta-」譯來。meta-有兩重意義，一為「形而上的」，一為「背後支撐」的。無論是那個含意，都是指居第二線，而不是居第一線的。

❷ 自我參照(self-reference)是指「自己」指回到「自己」的一種關係，在邏輯上自我參照的關係常產生矛盾或詭論。「資料的資料」還是資料，所以「資料的資料」與「資料」就構成了自我參照。這樣的說法並未釐清「資料的資料」與「資料」有什麼不同，故無法構成清晰的認知。

❸ 有人將 metadata 譯為「詮釋資料」。這是一個錯譯。

❹ 古文獻有標點符號者不多，且標點符號較時下使用者簡略。可參考：嚴文郁，《中國書籍簡史》，臺北市，臺灣商務印書館，民國八十一年十一月初版，圖四和圖五：明朝三色和五色套印的《楚辭》和《文心雕龍》樣張，其上即有句號和幾種句讀符號。

❺ 古書中幫助讀者理解文章內容的，不僅僅是此處所談的文章標誌，像注疏、校勘、眉批、批改文章的符號、乃至編輯符號等，也都可視為是幫助讀者理解文章內容的。

❻ 當時認為，算數的加減乘除四則運算，是智慧型的工作。起初，有些國家徵調了數學家參與對數表的計算，然而，表現得並不如所期，因計算常有錯誤。之後，使用計算器才順利的製出期待的對數表。四則運算居然有機器也能夠做，而且比數學家做得還要好，就不能不令當時的人驚奇。其實，要數學家做枯燥無聊的「煩」複計算，就像找畫家油漆牆壁，主事者實無知人之明。

❼

參考資料

一、謝清俊專欄，〈數位化的文章〉，國家數位典藏通訊第？？期，民國九十五年四月。網址：
二、謝清俊專欄，〈文章的標誌〉，國家數位典藏通訊第？？期，民國九十五年六月。網址：同

上

三、謝清俊專欄，〈後設資料與內容標誌〉，國家數位典藏通訊第？？期，民國九十五年七月。網址：同上

四、謝清俊專欄，〈意義與了解〉，國家數位典藏通訊第？？期，民國九十五年八月。網址：同上

五、謝清俊、莊德明，〈心經的多版本處理〉，民國八十四年八月