文學文獻的數位化問題

謝清俊
中央研究院資訊科學研究所、語言學研究所

主席、各位女士、各位先生：

很高興能夠參加這樣一個文學與資訊科技的研討會。去年，鄭院士錦全在香港倡議並準備主辦這個會議的時候，我就感慨良多。當時剛好碰到非典肺炎事件沒辦法舉行，所以展延到現在，在元智舉行。

我已經退休了；兩年前從中央研究院退休，退休後又替數位典藏國家型計畫服務了兩年，今年八月一日又退了下來，真正卸下了所有的專職工作。退休以前曾服務了三十六年，回過頭來看這三十六年，我的第一個研究計畫是「怎麼用電腦來處理漢字」，時間是在 1969 年。這一路走過來，經歷了電腦用在科學計算、管理財務、軍事、醫療、產業生產和自動化、用在圖書館、博物館等，在各行各業 ...。如果電腦算一種污染的話，那最後污染的一塊淨土就是藝術跟文學。所以看到開這個會，我心裡很感慨，我這一輩子做電腦化的事情，在退休的時候碰到最後這一塊淨土。今天我的演講，不想談技術性的東西，我本來想的題目是「文學數位化的意義」，後來一想，我不懂文學，所以只能談文學文獻的數位化，今天主要想和各位談的大概就是如此，包括：為什麼文學需要數位化？數位化有哪些重點？又該注意哪些事？等。

壹、美學、文學和資訊

文學跟美學大概是最晚用計算機的兩個領域，文學比美學還要晚。在很多美術館的展出裡，已經可以看到用電腦做的美學作品、作聲光的控制，以及作一些意象上的呈現，文學到現在還沒有把電腦用到這樣的程度。因為文學跟藝術、美學有相當多的共通之處，所以首先就讓我借用美學的一些理論，來談一下文學的數位化。

一、符號學美學

讓我們以蘇珊・郎格（*Susanne K. Langer, 1895-1982*）的符號學的美學（或稱符號論美學）作為引子。符號論美學在二十世紀的五十年代主導美國的美學思潮，它把文化看成由符號所組成，在人文主義跟科學主義方面都有較大的包容性。[1] 文化符號可分爲兩類：一類是推理符號（是內涵概括確定的理性符號，它既可以翻譯，也可以被分解、推理；如語言符號。）一類是表象符號（是非理性的、完整獨特不能被分解的，具豐富含義的情感意象；

[1] 對此有興趣的讀者請參考唐孝祥、袁忠、唐更華編著之《美學基礎教程》華南理工大學出版，1998 初版（2002 重印），第四章西方美學發展歷程，第 149 頁。

如藝術符號）。藝術可作爲一種傳播的文化符號，文學也可以這樣看。

我們曉得電腦的能力很大，但經常感到困惑的問題是：電腦能作什麼事？不能作什麼事？這個問題如果問電腦專家，他回答會很複雜，他會告訴你：「問題的解法如果複雜到某某程度以上，計算機就沒辦法作。」這是從問題的複雜程度來看的。如果從美學或哲學的角度來看的話，其實非常單純：「電腦只能處理形式，它不會直接處理問題的內容。」這裡所說的形式及內容都是美學裡面的辭彙。

因為電腦可以直接處理形式，所以任何美學、文學，只要把它看成一種符號，不管是推理符號、表象符號，不管它是文字還是色彩，只要是一種形式，電腦就可以設法處理。這是為什麼電腦可以踏進文學或美學領域的基本理由。

二、資訊的界說

資訊是很有趣的。有一篇文章，它說：「什麼是 Information 已經困擾了我們五十五年，至今沒有一個行業可以把什麼是 Information 搞清楚。[2]」2000 年回頭算五

[2] 請參考：A.D. Madden, A Definition of Information, Aslib Proceedings vol. 52, No.9, p.343-, 2000

十五年是 1946 年，1946 年是賓州大學作出第一部電子計算機的那一年，也就是有了電腦和 information processing 以後，到現在為止，各學科到現在始終搞不清楚 information 是什麼。在我們現在要討論「文學的數位化」時，這個問題又出現了。試想：若是我們大家對什麼是 Information 的概念、想法都不一致、沒有共識的話，怎麼來做文學文獻的數位化呢？所以，我們很需要給 information 下一個簡單而通用的界定，作為文學文獻數位化的前題。

三、美學中資訊的界說

資訊是什麼？資訊就是『所知表現在媒介上的形式』[3]，這兒的所知指的是心智上的一些活動，包括理性的和感性的都在內。從這個定義來看的話，美學以及文學作品可以視為是「人類感情表現在媒介上的形式」。因為它是一種形式，所以電腦可以處理它。

我們再回頭看看蘇珊・郎格（*Susanne K. Langer, 1895-1982*）講的美學的問題，她說：「藝術是表現人類

[3] 講者已使用此定義多年，請參照〈談資訊的定義與性質〉謝清俊，「資訊科技與社會轉型學術研討會」中研院社會學籌備處，Dec. 1996

情感的符號形式。」就是因為藝術是表示人類情感的符號形式，所以電腦可以處理藝術作品，因為藝術是一種形式，所以可以將它數位化（digitalize）。文化符號包括語言符號，所以文學的作品是可以透過電腦處理的。

然而，為了數位化文學的作品，，我們必須要先建立對資訊的共識。比方說，如果松岡榮志教授所想的 information 和我所想的 information 是兩回事的話，我們怎麼合作呢？所以在文學裡談數位化，第一件事需要有共識：「資訊是什麼？」

依『資訊是所知表現在媒介上的形式』這個定義，可以概括蘇珊・郎格所說：「藝術是表現人類情感的符號形式。」這是因為：所知包含感性的 、理性的心智作用，其中感性的所知中又包含了人類的情感。所以，什麼是美學作品或藝術作品的資訊呢？就是『人類感情表現在媒介上的形式』。什麼是文學資訊呢？也可以仿照上述的方法處理，只是文學所用的表現系統是文字、口語，其作品的形式則多為各種文體的出版品，較藝術作品更有法度。

然而，在上述的討論中，還需要特別注意一點，那就是「媒介」。為什麼現在要用電腦來處理文學的資料？

是因為媒介的關係。此所謂的媒介包含媒介材料、依媒介材料所設計的工具，以及用這個工具所發展出來的技術。媒介有何影響呢？這就需要從媒介與文化發展的關係來談起了。

貳、媒介之與文化發展

我們都知道人類歷史的發展，經濟學家和社會學家喜歡將它分成遊牧時期、農業時期，工業時期、後工業時期等。這是從經濟的角度，從勞動服務的角度，來看這些問題。但是真正從文化的角度來看的話，最重要的功能就是傳播行為（communication）；如果沒有傳播的話，根本沒有社群也沒有文化。從傳播的角度來看，我們引麥克魯漢說的文化發展的三個時期[4]：

一、口語文明：就是文字還沒發明的時期。此時的文明典型的就是口耳相傳；如，菲律賓到紐西蘭這一帶南洋的波里尼西亞人（Polynesians），有兩千多個種族，都沒有文字。這個文明都生活在一種優遊自在的環境裡面，是相當可愛的文明。麥克魯漢稱之為聽覺文明，因為文化的傳承以聽覺為主。

二、文字文明：有了文字以後就發展成所謂的大文明。世界上的四大文明主要的特徵就是很早期就有相當成熟的文字。有文字可以將事情紀錄下來，就可以『百官以治，萬民以察』。不要小看記下來，如果沒有文字的話，世界就如同孔夫子講過的話：『逝者如斯夫，不舍晝夜』，時間的概念是線性的；說過的話，忘了就不算數。一旦文字記下來後時空的關係改變了，可以仔細推敲以前作的事情對不對，累積的知識怎麼整理，深度文明的過程從這裡開始。麥克魯漢稱之為視覺文明，文化的傳承以視覺為主。

三、多媒體文明：當電子媒介出現之後，引發了多媒體的表達方式，因此，文明的傳承不再被視覺主宰，又回復到口語文明時的類似現象—五官各司其是，這便是多媒體文明不同於文字文明的一個顯著特徵。

這三個時期，可以看是媒介轉變而引起的。例如，從有文字以後產生了素養問題（literacy）；要會讀別人寫的作品，要會用文字跟別人溝通。這是因為媒介和表現系統改變的

[4] 此處並未直接引用麥克魯漢所用的名詞，僅依劃分文明的內容說明。請參照 Marshall McLuhan《Understanding Media》McGraw-Hill, 1964

緣故。當多媒體文明出現時，也產生了素養問題，那就是近三十年來陸續出現的電腦素養 、資訊素養 、網路素養 … 等，而素養造成的社會問題便稱為數位落差（digital divide）。現在上網路的小朋友不喜歡看只有文字的文章，喜歡看動畫、漫畫，喜歡短篇不喜長篇，喜歡用聽的，喜歡有配樂 … 這表示人的五官：眼、耳、鼻、舌、身又開始在新的文明裡面展現了其應該有的地位，新的文明不再被眼睛主宰（dominate）。

參、媒介發展的進程

從口語發展到文字，從文字發展到多媒體，我們可以看到影響文明的進展有兩個重要的因素，第一是媒介，第二是表現系統。所謂媒介，就是傳播使用的材料；口語時期用的媒介空氣音波，表現系統便是口語。當文字使用時，便有了書寫的系統，有了各式各樣文體的文章，也就形成了書寫的表現系統。這書寫的表現系統使用了一些新的媒介，包括：筆、墨、紙、硯等物質。物質媒介的使用甚至可以早推到美索不達米亞的泥版、埃及用的 papier、印度用的貝頁等。在文字文明的時候 ,用的是各式各樣的物質媒材，從竹片、木片、絲帛、石材，一直到紙張以後的種種發展。

文字文明可以說是物質媒材的文明；我們要紀錄事情、物件，紀錄情感，都要耗用物質，用過以後的媒材就就破壞掉、不勘再用了。這是物質媒材時代很重要的特色。

但是，到了多媒體文明的時候，媒材改變了：從物質媒材轉變為能量的媒材。各位想想現在用的磁碟、隨身碟，上面可以存放幾百個 mega-bytes 的資料，當沒有存資料的時候它是那麼重，當載滿資料的時侯還是一樣重。因為它不再是用物質作為傳播媒介，而是用能量作媒介，能量沒有物質障礙。所謂物質障礙，佔空間、有重量、會破碎、耗損、…。所以一旦媒材由物質變成能量時，就突破了我們常識下認識的時空，如網路的無遠弗屆、瞬息即至。從人類學、從文化方面、從傳播的角度來看的話，由物質媒材演變到能量媒材的這個大躍進帶來的影響，比從口語到文字的更為重要、深遠。

口語時期的表達系統是口語，約定俗成即可。文字便講究許多，可以紀錄的很細，比口語細很多。但是口語文字皆有限制，文字不適合描述動作太細膩的事情，所以怎麼用文字描述游泳、騎腳踏車，讀者永遠學不會；文字沒辦法描述複雜的圖像，所以外國人說 a picture worth thousand words。文字沒辦法描述物質的本質，只能當一個指標告訴你物質是什麼，比方說能夠用文字說清楚什麼是椅子、什

麼是桌子、什麼是水嗎？連科學家也說不清楚。在符號學裡講文字只是一個指標（signifier），以電腦來講，字符也只是一個指標（pointer）。

讓我們再回頭看看從口語文明到文字文明的轉變。四大文明裡面沒有一個記清楚了怎麼從口語文明到文字文明。這也許是因為剛剛開始有文字，不太可能把轉變記載下來。不過很有幸的是，佛教裡有這樣的資料：釋迦摩尼在公元前約 600 年的時候，傳了四十九年的道，過世後其弟子們將他所說的道整理成文字紀錄，這個過程有詳細的紀錄，此轉換，從釋迦摩尼死後到把整個佛經紀錄下，來花了約 500 年。佛教有極大的宗教熱忱來做這件事，所以作的非常精確、仔細。這表示從口語文明到文字文明的時候，也就是說要把口語媒介改成文字的媒介，要重新把已往之所知表現一遍，再現一遍。

當媒介由物質媒介轉換為能源媒介的時候，需不需要把所有文字文明的記錄重新表現一遍呢？這是個值得大家思考的問題，我以為非作不可，現在不作二年後即將後悔，現在不作二十年後、二百年後還是要做！我們這一代有幸生在這三千年才碰到一次的機會裡，也就是面臨從文字文明到多媒體文明、從物質媒介到能源的媒介的轉變關頭，做數位化是我們的權利，也是時代賦予我們的機會和責任。

從口語文明到文字文明是將口語改變成文字，很單純的，因為只涉級語文。但從文字文明到多媒體文明的時候，就不是這麼單純了，是需要把所有的文物通通數位化，改為多媒體方式表現，這是多媒體涉及的不僅僅是語文一項的緣故。博物館的數位化正說明了這種可能。

肆、業已數化的文獻

若問：現在世界上的文明怎麼留下來，可能會聯想到American Memory計畫所展現的或是數位圖書館，數位博物館的表現。為什麼從圖書館可以擴及博物館？因為多媒體已經不是光靠文字，也不只靠眼睛，數位化將所有的文物都變成0跟1來表達，所以現在我們整個數位化的風潮，波及所有典藏的單位、所有典藏的東西。文獻是文學中最早數化的；歷年來已經數化的文獻有很多，如二十五史、大藏經、辭典、全唐詩、全宋詞、四庫全書等等，一直到台灣最近努力的國家數位典藏。所以，文學文獻的數位化，雖然這是第一次的學術會議，事實上，這個工作已經默默進行了很長一段時間。也許是因為大家對資訊業者的定義不了解，大家並不認為文學家是資訊業者，所以，在此之前，從來沒有舉辦過從文學的角度來談數位化的學術會議。

根據聯合國的定義，文學家是十足的資訊業者（information occupation）。這是因為，所有處理 information 為主的行業都是資訊行業（information Professional），其中絕對包括文學家在內。認識到這點，就會知道，如果要數位化，文學家應該是首當其衝。自從電腦開始處理文字的時候，就已經接觸到文學的外在形式了，雖然「打字、排版」這種工作只是文學邊緣的邊緣，但究竟關係著文學作品的呈現。接下來，讓我們談談文獻數位化的內涵、層次以及相關的重要問題。

伍、人文文獻數位化之層次

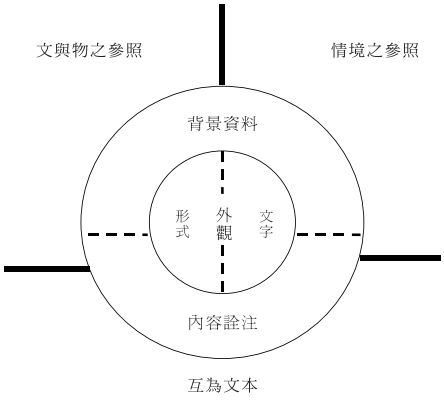
一個文件的數化可以分三個層次：

其一是文獻外觀的數位化：這部分大家最喜歡看，也最知道怎麼做，而且也問題最少。通常就是把文獻掃描，如文獻中的文字、版面、照片 … 等。這部分很容易作，大家看過的也多。大家對於外觀的數化，不論是打字輸入、掃描輸入、用 OCR 辨識，都耳熟能詳。外觀方面還有版面，比方說一些文學文獻原件還在，可能是善本書，可能是刻在竹簡上面的、寫在絲帛上面的，這些原來的版面、裝訂現象，都是屬於外觀，都是可以數化的。

其次是文獻背景資料的數位化：舉凡一份文件的書目、作者、譯者、注者、典藏相關資料、圖說、配音、配樂…等都屬此類。很多文件少了背景資料的說明和襯托，給人的感受差別很大。背景資料包含有時下流行的後設資料（metadata，或譯為元資料）

再次是文獻內容的數位化：文獻內容的數位化，目前所做的還不是很多。國內最早做內容的數位化的就是羅鳳珠老師。當初，他做紅樓夢網站時，就蒐集了許多紅樓夢的版本、舞台劇、電影，以及各種其他的相關的文物，例如:紅樓夢的版本 詮釋等問題。內容的數位化對數化工作而言，是最重要的；沒有這一部份，數化就失去了應有的深度和廣度。

文物數位化示意圖



一個文件的數位化，外觀是必要的成份；以文學文獻來講，主要是文字，以及版面所呈現的形式。外觀的數位化後，另有兩層次的數化:一是背景資料、一是內容的詮注。背景的資料是一些客觀的資料，此部份資料的對錯是可以考據的，考證的，但是不能隨意解釋。內容的詮注也需要有依據；可能是依某一個文獻、主義、理論、研究的方法…等作為詮注的立論所在。

以上只說明了一個文件數化的內涵。文件之間是有相當密切關聯的，這部份就涉及到文獻整體數化的問題了，請參照以下的討論。

陸、相關資料的參照

一個數位化文件，其相關資料的參照可以分三部分：

互為文本（*Inter-textuality*）：*Julia Kristeva* 說，任何文學史學之間沒有獨立存在的文字，這些文字一定跟其他的資料之間有某種關聯。[5] 如果把這些關聯找出來就是一個文學的知識庫（knowledge base）。文史資料跟其他各學科之間還有關係，這類不限領域文件之間關係都屬互為文本的

[5] 關於此說，建議可以參考葉嘉瑩著《清詞選講》三民書局，1996.08，第115頁中所說及其例。

參照。

情境的參照（*context*）：情境的參照，從背景來說可分為有關文化背景、社會背景、以及個人背景的情境；從傳播的角度來講，這些文學作品有作者情境及讀者情境。[6] 早期的口語傳播沒有這個問題，因為口語傳播是面對面，讀者跟作者處在同一個情境。當傳播或溝通的機制越來越進步，便幾乎把作者情境及讀者情境斬斷，成為越來越獨立而交集越來越少的狀態。從溝通的歷史來看，作者情境及讀者情境越分越遠。以前寫文章演講先要了解讀者是誰，聽眾是什麼水準。在網際網路下面，讀者情境已無法清楚認定，詮釋學家甚至說：一個作者做完作品後，作者已死，所有文章的意義是由讀者賦予他的。情境的參照是目前數化尚未接觸的領域，這些不同情境的詮釋資料怎麼匯集到電腦裡面去，是文學資料數位化面臨到較大的挑戰。

文跟物的參照：如紅樓夢中對食譜的參照、建築的參照等就是很好的例子這是人文跟自然的結合。這裡所要探討的是，文學相關的資料怎麼在計算機裡與其他學科的知識庫、數位圖書館、數位博物館等建立關係，以期相輔相成、

[6] 關於傳播學和資訊學對作者情境及讀者情境的觀點，可參考❷。關於詮釋學對情境的觀點，可參考葉嘉瑩著《詞學新詮》桂冠，2002.02，第37頁至42頁中所作之說明及其例。

相得益彰。

柒、結語

文學文獻的數位化是文學資料、文學知識的再整理，也是把傳統上文學所擁有的轉移到能量媒介上的一個大工程。現在開始作文學文獻的數位化，正是具有「天時」和「地利」，然而是否能「人和」則是成敗的關鍵。

人和的要點是要能充分溝通跟合作；這龐然的文學資料、遺產，要轉移個媒介物質，當然不是簡單的事，它比轉換佛陀說的 49 年的法，要複雜多了。試問：，我們累積了 5000 年的文化資產，要多久才能完成？要多少人合作？

早期數位化，失敗的很多。1984-85，已有好多從事文學作品數位化的工作，如紅樓夢、駱駝祥子等等。然而，因欠缺規範，即使是同一作品紅樓夢，使用的方法都不一致。這一批數位化的作品，很早就不見了，因為沒辦法使用。怎麼消除這個問題？要有一定的規範、章法。所以解決方法是，合作建立數位化的規範，以及各系統之間的相容互通。（我們不說標準，因為社會情境變化太快、 技術進步太快。一個標準的壽命太短，短到電腦硬體的壽命平均只有 18 個月，軟體平均 24 個月。其他的技術進步也快，傳統的標準

已受到嚴苛的挑戰。）

其次，數位化只有一次的機會，追求卓越的品質是存活的不二法門。數位化是爭千秋之事，不是爭一時之事。我想，我能用的時間已用盡，幸好所餘不多，就請大家參考我的投影片吧。最後，希望文學文獻數位化，就像投影片中的大樹一樣，成長、茁壯、永續千年。