

※「數位人文研究的個案與反思」專輯※

華文學界的數位人文探索： 一種「史前史」的觀察角度

徐力恆^{*}

數位人文研究在臺灣、中國大陸和香港正方興未艾，掀起了學術風潮。但是，數位人文在這些地區的學界其實絕非新事物。自一九八〇年代以來，三地都有學者處理數位科技和人文學術碰撞，而產生的新學術議題。此等努力之學術意義，不只於電子化和數位資料（又稱「數據」，以下直接以「資料」名之）的積累，但卻長久沒有被放進「數位人文」(digital humanities) 的學術範式來看待；自一九八〇年代至該範式在二〇〇九年前後出現後，逐漸變得流行，三地和世界數位人文學術之間的相互影響和交流，也未見有學者進行專門的檢討。有鑑於此，本文將採取比過去數位人文研究更長時段的眼光，務求從學術史角度考察數位人文的這段「史前史」(prehistory)，亦即一九八〇年代至二〇〇九年的時段。本文提出，當下華文世界的數位人文研究風潮絕非憑空而生，近來數位技術雖已遠比上述階段發達，但數位人文史前時期的探索實際上產生了一種路徑依賴 (path dependence)，基本決定了當前數位人文研究的取向，是一段重要的數位研究學術史。

在做這樣的探討時，不得不提到近來已有學者提倡對類似的數位人文史前史問題進行反思。例如戴安德 (Anatoly Detwyler) 近來著文，聚焦一九二〇年代清華大學梁啟超 (1873-1929) 和魏聚賢 (1898-1990) 對歷史統計學的思考，認為那段學術往事為「現代中國學術研究中的信息管理和數據分析的更具系統性的歷史提出了可能性」¹。同時，他認為那段歷史「構成了數字〔數位〕人文史前史的史前史……挑

^{*} 徐力恆，香港城市大學中文及歷史學系助理教授。

¹ 戴安德 (Anatoly Detwyler) 著，林太平譯：〈數字人文及其史前史〉，《澳門理工學報》，2019 第 4 期，頁 148。

戰了認為當代數字分析是獨特的或無前例的觀點」²。如果我們據此認定那段學術史是當下華文學界數位人文研究的遠因，那麼本文要探討的就是其近因。兩者不同之處在於，戴安德討論的遠因和當下的數位人文研究之間沒有「直接線性關係」，本文討論的近因卻直接塑造了當下數位人文對中國材料進行研究的大致圖景。由於數位人文研究是近年人文學界研究熱點之一，除了戴安德以外，林富士、舒健、李友仁 (Paul Vierthaler)、陳靜、邱偉雲在內的學界先進，都在近年撰述回顧關於中文材料的數位人文歷程和狀況³。此外，二〇一八年的美國亞洲研究學會年會曾有小組以「數據的過去，數據的當下：資訊史的中國敘述／中國敘述作為資訊史」(Data Past, Data Present: Chinese Narratives of/as Information History) 作為主題，和本文的關懷也息息相關。在該小組裏，陳靜曾以〈數位人文在中國的史前史〉(“A Prehistory of Digital Humanities in China”) 為題發表論文⁴。陳氏和筆者也在後來繼續發展這一思路，合著一文勾勒數位人文在華文學界四十年來的發展，務求從比較整全的角度分析數位人文在華文學界的進程及前景⁵。以上研究論著都啟發了筆者對本文的思考和寫作。

在展開討論前得先簡要說明為何要做這種探討。在筆者看來，數位人文本來就不是一個能輕易地辨識其開首和起源的一種研究範式。早在數位人文流行以前的二十多年，學者已利用電腦來做人文課題之探究，這意味著我們思考數位科技及人文學科相交而碰撞的各種可能時，有必要打開更寬廣視野，考慮學者們曾經面對的各種可能和選擇的路徑。對筆者來說，其中一種可行方法，是追溯前輩學者們在「數位人文」這一名詞出現以前所做的學術嘗試，考察其對當下數位人文研究的影響。

² 同前註。

³ 林富士主編：《「數位人文學」白皮書》(臺北：中央研究院數位文化中心，2017年)；舒健：〈擁抱「大數據」對歷史研究的挑戰〉，舒健編：《大數據時代的歷史研究》(上海：上海譯文出版社，2017年)，頁1-11；李友仁：〈北美與西歐的數字人文中國研究狀況論析〉，《山東社會科學》，2018年第7期，頁54-58+63；陳靜：〈當下中國「數字人文」研究狀況及意義〉，《山東社會科學》，2018年第7期，頁59-63；邱偉雲：〈我國臺灣數字人文研究進程(2009-2017)〉，《圖書館論壇》2020年，網路首發，頁1-11。

⁴ Jing Chen, “A Prehistory of Digital Humanities in China” (paper presented at the 2018 Annual Conference of the Association for Asian Studies, Washington D.C., March 22-25, 2018).

⁵ Jing Chen and Lik Hang Tsui, “Debating and Developing Digital Humanities in China: New or Old?” under review.

有論者認為，由於和日新月異的數位科技密切相關，數位人文研究通常宣示某種前瞻性 (forward-looking) 和遠見，甚至帶有一點預言的意味⁶。由是，數位人文研究往往著眼未來的可能，而不在過去；然而，這樣往往會讓人忘懷當下的歷史 (history of the present) 中一些重要但不起眼的碎片，而接受科技、數位研究的軌跡所展示之必然性。數位人文這一範疇雖然「新穎」，卻是有歷史縱深的。例如，即使是“computers”這個詞本身的用法，也絕非一直指我們今天習以為常的數位化「電腦」。在數位化電腦大行其道以前，該詞其實主要指人力操作、人手演算的「計算員」，包括天文觀測站、電腦公司、學術項目僱用的計算人員。有了她們在這方面的參與，才讓科學家確信，發展大規模運算是極具學術價值的，而逐漸發展出當下諸多讓人炫目的數位技術⁷。因此，筆者希望在本文反其道而行，對數位人文的早期發展（而不是它的未來）做一點回顧，以求梳理數位人文在華文學界的學術史，甚至是藉此探討這段學術史為當下數位人文學術造就的深刻影響。實際上，已有學者在做類似對西方數位人文學術的回顧，例如，對人文計算 (humanities computing) 的溯源、對關鍵學者的訪談及對相關訪問內容的反思等⁸。

以下討論將盡量兼顧華文學術界各地情形，尤其是臺灣、中國大陸和香港——雖然本文局限在三地學界的情形，但其中涉及的頭緒已經極多，筆者難免受到見識所限，掛一漏萬；限於篇幅，也往往只能略舉梗概。為何要專門討論華文學界的數

⁶ Patrik Svensson, “Envisioning the Digital Humanities,” *Digital Humanities Quarterly* 6.1 (2012), <http://www.digitalhumanities.org/dhq/vol/6/1/000112/000112.html>, 我在別處曾根據這一點解釋，中國歷史學界何以對數位人文研究仍有疏離感和懷疑，參見 Lik Hang Tsui, “Charting the Emergence of the Digital Humanities in China,” in *Chinese Culture in the 21st Century and Its Global Dimensions: Comparative and Interdisciplinary Perspectives*, eds. Kelly Kar Yue Chan and Garfield Chi Sum Lau (Singapore: Springer, 2020), pp. 211-214。

⁷ David Alan Grier, *When Computers Were Human* (Princeton, NJ: Princeton University Press, 2005); Martin Campbell-Kelly, William Aspray, Nathan Ensmenger, Jeffrey R. Yost, *Computer: A History of the Information Machine*, 3rd ed. (Boulder, CO: Westview Press, 2014), pp. 3-19. 關於電腦運算的發展史概略，可參見 Paul E. Ceruzzi, *Computing: A Concise History* (Cambridge, MA: MIT Press, 2012)。這裏說「她們」而非「他們」是因為許多計算員都是女性。

⁸ Susan Hockey, “The History of Humanities Computing,” in *A Companion to Digital Humanities*, eds. Susan Schreibman, Ray Siemens, and John Unsworth (Oxford: Blackwell Publishing, 2004), pp. 3-19; Julianne Nyhan and Andrew Flinn, *Computation and the Humanities: Towards an Oral History of Digital Humanities* (Cham, Switzerland: Springer International Publishing, 2016).

位人文探索呢⁹？因為和自然科學相比，人文學科和部分社會科學學科（尤其是其中量化特點不明顯的學術領域）深深浸淫在語言之中，極受語言的特點影響——數位人文自然也是如此。以主題模型 (topic modeling) 這一數位人文分析方法為例，其基礎為斷詞 (word segmentation) 技術。可是，斷詞一旦用於中國文史研究時，馬上就面對頗多挑戰。例如，當機器閱讀「中華民國」這四個中文字時，它可以把那理解為四個詞，也可以理解成兩個各為兩個字的詞，或是一個四字詞。要讓機器獲得這方面準確判斷的能力，需要人文學者善用他們的專業知識介入。這意味著語言的特點其實左右著所有涉及人文學科的數位研究。而具體就古漢語與現代漢語而言，它們的語言特點也為組織和分析數位材料方面帶來不少技術挑戰，例如，漢字的數量與差異性增加了從文獻中進行數位挖掘，提取中文文本的難度¹⁰。

早在「數位人文」被提出沒多久的二〇〇八年，就有學者在〈數位人文宣言〉(“A Digital Humanities Manifesto”)中提倡這種新的研究範疇應該生成「全球導向」的知識¹¹；後一年的〈數位人文宣言 2.0〉(“Digital Humanities Manifesto 2.0”)也再次明確提到當時「研究社群具備的全球特質」是數位人文所「擁抱」的¹²。然而，哪怕是十多年後，當下的數位人文學界，尤其是以文本為研究對象的數位人文研究，仍大體是以西方和英語世界主導的¹³。加上大部分國際數位人文研究者並不專門處理中文材料，華文學者面對的這方面挑戰未必是他們熟知的，考察非西方和非拉丁字母

⁹ 由於譯法不一，本文一律將“digital humanities”統稱為「數位人文」，但中國大陸一般譯為「數字人文」，香港一般譯為「數碼人文」，三地對此用詞習慣不同，對其他電腦術語的用法也不同，以下不一一贅述。

¹⁰ 參見王宏甦、徐力恆、包弼德：〈服務於中國歷史研究的網路基礎設施：對相關探索的建議和展望〉，《數位典藏與數位人文》，將刊。

¹¹ “A Digital Humanities Manifesto,” (UCLA Humanities Technology [HumTech]), <http://manifesto.humanities.ucla.edu/2008/12/15/digital-humanities-manifesto/>.

¹² “The Digital Humanities Manifesto 2.0,” (UCLA Humanities Technology [HumTech]), p. 5, https://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf.

¹³ Alex Gil and Élika Ortega, “Global Outlooks in Digital Humanities: Multilingual Practices and Minimal Computing,” in *Doing Digital Humanities: Practice, Training, Research*, eds. Constance Crompton, Richard J. Lane, and Raymond George Siemens (New York: Routledge, 2016), pp. 22-34; Simon Mahony, “Cultural Diversity and the Digital Humanities,” *Fudan Journal of the Humanities and Social Sciences* 11.3 (2018): 371-388; Simon Mahony, “Language and the Subjects of Study in the Digital Humanities,” *Comparative Literature in China* (中國比較文學) 1 (2020): 96-113.

語言文獻的數位實踐之研究仍然很少，相關討論也未得到足夠關注。有鑑於此，本文討論數位人文的史前史時，聚焦在華文學界對中文材料的開發和研究，尤其是針對文本資訊方面的探索——數位人文資訊當然不限於文本資訊，另外如地理資訊等也非常重要，相關探索在華文學界也有深厚底蘊，但不是本文探討的重點¹⁴。筆者選擇文本資訊的研究作為重點，而不是面面俱到地剖析數位人文的一切維度，是希望由此突出語言在數位人文研究中的重要性，同時藉以討論中文文獻研究的獨特處境。

*

*

*

*

華文學界開展對中國典籍的數位整理和研究，實際上不算很晚。一九七〇年代已經出現了可以處理中文的電腦，尤其是美國和日本公司開發的電腦。至少從中文的編碼來看，一九八〇年由「國字整理小組」編製的「中文資訊交換碼」(Chinese Character Code for Information Exchange) 包含了教育部公布的四八〇八個常用字¹⁵。一九八六年，經濟部中央標準局公布了「通用漢字標準交換碼」，共含一三〇五一字。其後各種內碼出現，一九八〇年代臺灣業界出現「萬碼奔騰」的局面¹⁶。至於中國大陸方面，官方在一九七四年開始了簡稱「七四八工程」的「漢字信息處理系統工程」，希望解決漢字作為電腦資訊的處理問題。一九八〇年，中國國家標準總局

¹⁴ 空間資訊的數位處理是數位人文重要的一脈傳統，尤其是歷史 GIS。換個角度看，和專門處理文本材料的人文學者相比，空間人文學 (spatial humanities) 學者通常對數位技術的運用是毫不陌生的。關於歷史 GIS，可以參見 Ian Gregory and Paul S. Ell, *Historical GIS: Technologies, Methodologies, and Scholarship* (Cambridge: Cambridge University Press, 2009)。華文學界的情形可以參見李宗信、顧雅文：〈近二十年來應用歷史地理資訊系統的回顧與展望：以臺灣區域史研究為例〉，《臺灣史研究》第 21 卷第 2 期（2014 年 6 月），頁 167-196；林富士總策畫，曹銘宗撰稿：《遠見與承擔：中研院數位人文發展史》（臺北：中央研究院數位文化中心，2017 年）；林敬智：〈地理資訊系統 (GIS) 在史學研究中的應用〉，蔣竹山編：《當代歷史學新趨勢》（臺北：聯經出版事業公司，2018 年），頁 487-538；張萍：〈地理信息系統 (GIS) 與中國歷史研究〉，《史學理論研究》，2018 年第 2 期，頁 35-47；趙耀龍、巢子豪：〈歷史 GIS 的研究現狀和發展趨勢〉，《地球信息科學學報》，2020 年第 5 期，頁 929-944。

¹⁵ 參見謝清俊、黃克東：《國字整理小組十年》（臺北：資訊應用國字整理小組，1989 年）；林富士：《「數位人文學」白皮書》，頁 35-36。

¹⁶ 參見 [https://zh.wikipedia.org/zh-hk/ 中文標準交換碼](https://zh.wikipedia.org/zh-hk/中文標準交換碼)。

發布《信息交換用漢字編碼字符集》，其中包含了六七六三個漢字的編碼字符，一九八一年開始實施，滿足了電腦處理漢字的基本需求¹⁷。也就是說，從漢字編碼的角度看，一九八〇年代的臺灣和中國大陸都具備了以電腦處理中文材料的技術，開發了相應標準。這是各學科進行中文相關的數位開發和研究的必要條件。

最早以數位方式處理中文的是計算語言學，比其他人文學領域都更早。王士元教授早於一九六六年就在加州大學柏克萊分校提出了建立漢語資料庫的構想，後一年開始成立，把各時期漢語音韻學和各種漢語方言資料輸入電腦，建立了“(A Chinese Dialect) Dictionary on Computer (DOC)”計畫¹⁸。臺灣於一九八八年舉辦了首次計算語言學研討會，又在一九九〇年成立中華民國計算語言學學會，並在一九九六年創辦了 *International Journal of Computational Linguistics and Chinese Language Processing* 學刊(《中文計算語言學期刊》)¹⁹。至於中國大陸學界，學者自一九七〇年代開始他們的探索，相關研究計畫幾乎都以現代漢語語料為對象，建立了幾個百萬字級別的語料庫。許多這些語料庫後來都成為二〇〇三年由中國中文信息學會語言資源建設和管理工作委員會發起的「中文語言資源聯盟」的成員²⁰。有了這些鋪墊，針對漢語的語言學就成了一個「對電腦運算相對比較歡迎」的學科，這和其他人文學科是不太一樣的²¹。

在數位人文的史前時代，中國典籍的數位化中最有影響力的，當屬「二十五史」和《四庫全書》，甚至可以把它們當作華文學界的「谷歌圖書」(Google Books)來看待——即使其中的巨量資料(或譯「大數據」)未盡完整，也已包羅萬有²²。自

¹⁷ 參見 <https://zh.wikipedia.org/zh-hk/> 中文資訊處理。墨磊寧(Thomas Mullaney)將出版的專書也會處理這些話題：*The Chinese Computer: A Global History of the Information Age, Part II* (Cambridge, MA: MIT Press, forthcoming)。從產業和創新角度出發的討論，可參見 Qiwen Lu, *China's Leap into the Information Age: Innovation and Organization in the Computer Industry* (Oxford: Oxford University Press, 2000)。

¹⁸ 李王癸：〈百年來的語言學〉，《臺灣語文研究》第7卷第1期(2012年3月)，頁19。

¹⁹ 參見 http://www.aclclp.org.tw/index_c.php。

²⁰ <http://www.chineseldc.org/cldcTest.html>。參見陳靜：〈當下中國「數字人文」研究狀況及意義〉，頁60。

²¹ David M. Berry and Anders Fagerjord, *Digital Humanities: Knowledge and Critique in a Digital Age* (Cambridge: Polity Press, 2017), p. 29.

²² 關於「谷歌圖書」，可以參見 Robert Darnton, *The Case for Books: Past, Present, and Future* (New York: PublicAffairs, 2009)，中譯本見羅伯特·達恩頓著，熊祥譯：《閱讀的未來》(北京：中信

一九八四年起，中央研究院歷史語言研究所與計算中心合作進行文獻的數位化，推動「史籍自動化計畫」。首先以宋以前的正史〈食貨志〉為數位化對象，之後則處理「二十五史」的全文，於一九九〇年完成了「二十五史全文資料庫」，後來又擴充為「漢籍電子文獻資料庫」（簡稱「漢籍」），納入更多典籍資料。這個資料庫的建置大大便利了研究者對史料的搜集，建置的資料標準嚴謹，又連帶促進了計算語言學對古代漢語的處理，是人文計算在華文學界的一個重要體現²³。

至於《四庫全書》的數位化，香港的迪志文化出版有限公司於一九九九年分別與上海人民出版社、香港中文大學在中國大陸和香港發行《四庫全書》的電子光碟版，由迪志文化出版有限公司、書同文電腦技術開發有限公司承辦開發製作。這個計畫自一九九七年六月迪志文化出版有限公司成立時即啟動開發。電子版以現藏國立故宮博物院的文淵閣本《四庫全書》為底本，參與技術開發的除了迪志文化出版有限公司、書同文電腦技術開發有限公司以外，還有清華大學計算機系負責 OCR 引擎的開發，北大方正電子有限公司負責建立專用字庫，和微軟公司（北京）研究開發中心負責平臺技術方面的援助，所以本身是香港和中國大陸合作的成果²⁴。高達八億漢字的這種規模的數位化計畫，不只開啟了典籍的系統數位化進程，推進了學界對中國典籍的數位利用和商業開發，還積累了大批以數位資料形式存在、可供分析研究的語料。根據這些語料，書同文電腦技術開發有限公司對三萬餘個漢字的使用頻率等做統計，出版參考書《古籍漢字字頻統計》，按《康熙字典》字頭統計了《文淵閣四庫全書》和《四部叢刊》內容的字頻²⁵。

在其他更大型的資料庫出現之前，「漢籍」和《四庫全書》兩個全文資料庫可說是研究中國傳統文史的標準配備。這些工作所起到積累數位資料 (data

出版社，2011 年）。從文化角度分析「谷歌圖書」等大規模數位化計畫的，參見 Nanna Bonde Thylstrup, *The Politics of Mass Digitization* (Cambridge, MA: MIT Press, 2019)。

²³ 網址為：<http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm>。參見黃寬重、劉增貴：〈中央研究院人文計算的回顧與前瞻〉，《漢學研究通訊》第 17 卷第 2 期（1998 年 5 月），頁 145-168；林富士總策畫，曹銘宗撰稿：《遠見與承擔：中研院數位人文發展史》，頁 23-34。

²⁴ 錢華：〈余志明為《四庫全書》冒險〉，《文匯報》（香港），2004 年 11 月 24 日，<http://paper.wenweipo.com/2004/11/24/RW0411240001.htm>；<https://www.chiculture.net/0704/html/d01/0704d32.html>。

²⁵ 北京書同文數字化技術有限公司編：《古籍漢字字頻統計》（北京：商務印書館，2008 年）。

accumulation)的作用，為後來的數位人文時代奠定了重要基礎²⁶。不過，學者越發意識到各種古籍的《四庫全書》本存在局限，不宜對其數位版過分依賴；而類似資料庫，由於長期通過固定用戶界面才能使用，限制了對文獻的靈活取用，在深度利用文本的各種數位科技越來越進步和流行之後，就顯得用途相對單一了²⁷。在這種狀況下，後來的各個典籍數位化計畫遠遠超出了「二十五史」和《四庫全書》的規模，境況到了二〇一〇年代尤其大不一樣，使用的資料庫和工具形態更加多元化，可供學者搜尋、擷取資料的數位文獻來源也繼續擴大²⁸。

「漢籍」和《四庫全書》以外，同一時期值得注意的中國典籍數位化探索，是中國社會科學院學者所做的研究。在錢鍾書(1910-1998)先生的支持下，中國社會科學院文學研究所在一九八五年設立「計算機室」，開展了中國典籍的數位化和研究工作。後來，中國社會科學院又把「古典文獻的計算機處理」定為重點項目。在計算機室成立以後，處理和出版了《論語》、《全唐詩》、《永樂大典》、《全唐文》、《十三經》、《水經注》、《大藏經》等典籍的索引或整理本，也處理了《全唐詩》、《先秦魏晉南北朝詩》、《全上古三代秦漢三國六朝文》、《十三經》、《諸子集成》、《全唐文》、《全宋詞》、《全金元詞》等多部古籍的數位化，並完成了「諸子集成數據庫」和「全唐詩速檢系統」等系統。同時，為了處理這些典籍使用的眾多漢字，他們以《中華大字典》為依據，建立了一套四萬五千字的漢字字庫，在一九八八年基本解決了整理古籍所要求的字庫²⁹。幾位熟知情況的學者回憶裏，都提到錢鍾書先生的女兒錢媛(1937-1997)在一九七八至八〇年訪學英國蘭卡斯特大學(Lancaster University)以後，向錢先生介紹了英國學者用電腦儲存及分

²⁶ 1990年代就開始開發的其他資料庫還有不少，參見陳靜：〈當下中國「數字人文」研究狀況及意義〉，頁61。

²⁷ 例如：可以參見李裕民：〈論《四庫全書》文淵閣本的缺陷——以宋代文獻為中心〉，《安徽師範大學學報》，2013年第2期，頁156-163；Hilde De Weerd, "Isn't the Siku quanshu Enough? Reflections on the Impact of New Digital Tools for Classical Chinese," *Communication and Empire: Chinese Empires in Comparative Perspective*, 2014, <http://chinese-empires.eu/blog/isnt-the-siku-quanshu-enough-reflections-on-the-impact-of-new-digital-tools-for-classical-chinese/>。

²⁸ 參見項潔、涂豐恩：〈導論——什麼是數位人文〉，項潔編：《從保存到創造：開啟數位人文研究》（臺北：臺灣大學出版中心，2011年），頁16-17。

²⁹ 鄭永曉：〈錢鍾書與中國社科院古代典籍數字化工程〉，《山東社會科學》，2019年第6期，頁33。

析莎士比亞作品的情況所取得之成果，因而啟發了他利用電腦研究中國典籍的想法³⁰。確實，翻查相關成果，英美學界以電腦輔助對莎士比亞作品之語彙索引和統計(concordances)，早在一九六〇年代末已經開始陸續編制出版，後來成為莎士比亞研究的基本參考書³¹。不知道是否受這種思路影響，中國社會科學院這方面做的成果也多是對典籍文獻的索引出版物，為學者提供了尤其是定量研究的參照，包括對古籍中字詞的分析。參與的學者也多次強調中國文獻中的定量問題，例如唐詩的數量化研究³²。雖然他們沒有使用「人文計算」的標籤，但以當今的眼光看，完全可以歸到該範疇之中。

不論如何，錢鍾書先生指導和支持下的這個團隊及相關成果，已經發展出後來數位人文研究者的一些基本認識。以下在一九九二年發表的文字可說頗具代表性：

電腦可以幫助人腦，但還不能代替人腦……我想，作為一個認真使用電腦對中國古典文學作品進行研究的人，既不會為電腦的發展和進步而昏昏，認定電腦能代替人腦，有了電腦就有了一切；更不會為電腦的難於駕馭而茫茫，放棄實實在在、長期艱苦的努力和奮鬥。**實踐證明，能幫助人腦的電腦需要人的更多幫助。**

據說，以上引文的最後一句是錢先生審改文章時加上的，從今天目光看來，反映了以電腦進行數位化和處理的人文數據，往往需要人文學者的人工介入這一重要數位人文現象。這種認知既相當超前，又確實建立在該團隊的大量數位化工作之上。然而，從學術影響來講，也有學者提出反思，指出中國社會科學院計算機室開發的成果「都是以公開出版物的形式即紙質版向社會公布，固然有利於知識產權保護，但是龐大的數據庫本身除個別時段曾被少量學者試用外，其廬山真面目學界一直未能

³⁰ 同前註，頁 31-32。

³¹ Marvin Spevack, *A Complete and Systematic Concordance to the Works of Shakespeare*, 9 vols. (Hildesheim: G. Olms, 1968-1980); T. H. Howard-Hill, *Oxford Shakespeare Concordances*, 36 vols. (Oxford: Clarendon Press, 1969-1973). 前者是基於英語單詞的現代拼法，後者是基於舊拼法。關於這些成果的編纂及相關技術，參見 B. H. Rudall and T. N. Corns, *Computers and Literature: A Practical Guide* (Cambridge, MA: ABACUS Press, 1987), pp. 59-78。莎士比亞的相關數位研究近況可參見 Janelle Jenstad, Mark Kaethler, and Jennifer Roberts-Smith eds., *Shakespeare's Language in Digital Media: Old Words, New Tools* (London: Routledge, 2018)。

³² 例如：田奕：〈電腦裏的唐詩〉，《文學遺產》，1992 年第 5 期，頁 125-126；〈古籍整理與研究的電腦化〉，《中國文化》，1994 年第 9 期，頁 85-86。

得見……該數字工程的主體產品——那些建設多年的數據庫，一直未能被廣大學界和社會所使用，著實可惜」³³。也正因为這個原因，這些成果的學術影響以及對後來數位人文研究的直接作用，似乎遠不如「漢籍」和《四庫全書》的數位版，不無可惜。

除了中國社會科學院團隊的工作，早期運用電腦科技研究中國典籍的一大熱點，是對《紅樓夢》的研究。在電腦時代以前，固然已經出現了通過人工計算，對《紅樓夢》前八十回和後四十回的用詞進行研究、比較的成果。到了一九八〇年代，一系列學者憑藉電腦科技，對此議題及相關問題進行研究和討論，一時頗為熱鬧。尤其是身處美國的陳炳藻以電腦進行計算，提出《紅樓夢》一百二十回皆為曹雪芹所作以後，引起頗多討論，不少中國大陸和臺灣學者都針對該問題提出過觀點³⁴。這些討論大多通過統計文本的用詞，來進行風格的計量分析 (stylistic analysis)，以處理《紅樓夢》不同部分之作者身分問題。這個議題成為熱點，除了因為作者問題本身就是紅學的焦點議題以外，無疑也因為這種研究應用了數位科技，而引起較大迴響。對於《紅樓夢》的數位人文研究，自此不絕如縷。

從上文所論已可見，香港沒有自外於華文學界這種數位研究的學術趨勢。除了商業公司對《四庫全書》的數位化工作以外，大學的研究者也從事中文材料的開發。自一九八八年以來，香港中文大學中國文化研究所分階段建立了「漢達古文獻資料庫」。一九八九年起，由劉殿爵 (1921-2010) 教授帶領的人員著手建立「漢達文庫」，把先秦兩漢至魏晉南北朝之一切傳世文獻輸入電腦，製作資料庫，於一九九六年完成。資料庫人員利用資料庫編纂古籍逐字索引，陸續出版³⁵。根據機構對建庫理念的介紹，「在建立資料庫的過程中，研究所並非僅僅把文獻資料從『書本形態』，轉變為『電腦可讀形態』，而是一次有系統的傳世文獻整理工作」³⁶。除了傳

³³ 鄭永曉：〈錢鍾書與中國社科院古代典籍數字化工程〉，頁 39。

³⁴ 相關回顧參見羅鳳珠：〈在網際網路建立漢學研究環境的重要性及可行性——就中國文學而論〉，《漢學研究通訊》第 16 卷第 1 期（1997 年 2 月），頁 3-9。陳炳藻的相關觀點見 Bing C. Chan, *The Authorship of the Dream of the Red Chamber: Based on a Computerized Statistical Study of Its Vocabulary* (Hong Kong: Joint Publishing Co., 1986)。

³⁵ 介紹參見 <http://www.cuhk.edu.hk/ics/rccat/database.html>。

³⁶ 參見 <http://www.cuhk.edu.hk/ics/rccat/database.html>。何志華、潘銘基：〈資料庫之利用與學術研究〉，《國文天地》第 25 卷第 4 期（2009 年 9 月），頁 49-56。

世文獻外，後來也把簡帛、甲骨文、金文、類書等資料納入資料庫。基於資料庫記載，中國文化研究所發展研究先秦兩漢引錄經籍和先秦兩漢互見文獻的研究計畫³⁷。資料庫後來成為「中國古籍研究中心」，二〇一一年命名為「劉殿爵中國古籍研究中心」(D. C. Lau Research Centre for Chinese Ancient Texts)，資料庫所包含的傳世及出土文獻迄今共八千萬餘字³⁸。此外，香港中文大學「人文電算研究中心」(Research Centre for Humanities Computing)於一九九三年由關子尹創立，前身為一九九二年成立的人文電算與人文方法研究室。這研究室和後來的研究中心呼應了「人文計算」(香港一般把“humanities computing”譯為「人文電算」)的學術取向，採用這個名稱的中心和學位課程，在一九九〇年代不只在華文學界少見，在西方世界也是非常罕見的³⁹。自一九九九年以來，人文電算研究中心先後開發《林語堂當代漢英詞典》網絡版、《粵語審音配詞字庫》、《香港、大陸、臺灣——跨地區、跨年代漢語常用字頻統計》、《漢語多功能字庫》等網路資源，俱為漢語的基本參考工具⁴⁰。這幾種資源的功能同時兼顧香港、臺灣和中國大陸的用戶，頗能反映研究中心的定位——其中《香港、大陸、臺灣——跨地區、跨年代漢語常用字頻統計》就是源於針對港、臺、陸語料字頻進行調研的書面研究報告進行重組和數位化，方便學者搜尋和利用⁴¹。類似工具書和資料的數位化，使得相關資訊的查找便利得多——在後來的數位人文年代，於研究工作中參考這些學術工具書之數位化整理，對人文學者來說，已經是習以為常的事了。

在數位人文範式發揮影響以前，長年和華文學界進行交流的海外漢學界，尤其是圖書館學者也有從事數位開發，同樣為後來學界對中文典籍的數位研究奠定了基礎。其中，主要包括由普林斯頓的艾思仁(Sören Edgren)帶頭的「中文善本圖書項目」(Chinese Rare Books Project)，針對一七九六年以前出現的中文善本圖書，開發了北美各校館藏的聯合目錄。項目的先導計畫在一九八九年開始，和中國大陸的

³⁷ 參見 <https://humanum.arts.cuhk.edu.hk/Lexis/chifreq/manual.html>。

³⁸ 梁德華：〈「漢達文庫」與古籍研究〉，譚國根、梁慕靈、黃自鴻編：《數碼時代的中國人文學科研究》(臺北：秀威資訊科技公司，2018年)，頁62-64。

³⁹ 該中心網址為：<http://humanum.arts.cuhk.edu.hk/>。Hockey, “The History of Humanities Computing,” p. 16.

⁴⁰ 這裏的介紹參考了 <https://humanum.arts.cuhk.edu.hk/Lexis/lexi-mf/>。

⁴¹ 參見 <https://humanum.arts.cuhk.edu.hk/Lexis/chifreq/>。

圖書館員合作，為普林斯頓大學和哥倫比亞大學的中文古籍館藏做了編目。隨後，項目在一九九〇年代持續進行，一九九六年開始，由普林斯頓大學東亞研究系統籌主持。學者們在創建這套聯合目錄的過程中，制訂了一套機器可以識別的編目標準格式，對日後的館藏數位化和發布有重要的作用。項目在二〇〇〇年發表了《中文善本書機讀目錄編目規則》(*Cataloging Guidelines for Creating Chinese Rare Book Records in Machine-Readable Form*)，調和了英語世界的編目標準和中文傳統古籍的慣例，方便項目成員為北美的古籍進行標準劃一的處理和編目。使用者除了可以利用這部目錄進行聯合搜索，還可以快速檢閱同一部古籍位於不同圖書館的各種版本資訊⁴²。儘管這個項目在北美各個大學已停止，並在二〇一一年轉移到北京的國家圖書館，但生成的數據已然併入通行的世界性聯合目錄 WorldCat 中⁴³。從今天的數位人文眼光看來，這種工作自然屬於數位基礎設施的核心部分，是任何更深入的數位研究的前提⁴⁴。

華文學界自一九八〇年代始，對中國文獻的數位化投入了大量心血。上文論及的幾個重要數位資源項目，都為華文學界的全文資料庫奠定了基礎。影響所及，以關鍵詞對材料進行全文搜尋，已經成為人文學者普遍習用的數位功能。甚至說「現在做研究『離不開』全文檢索、資料庫的使用」⁴⁵，一點都不誇張。有了這些數位資料的積累，關鍵字詞的搜尋功能為學者搜集資料帶來極大便利，也由此得到眾多中國文史領域學者之注意和使用。搜尋功能的易用性和數位文獻規模之大，讓人文學者一般對全文資料庫比較能接受。這種使用資料庫及相關搜尋工具的方式，也同時成為他們接觸數位研究最重要且固定的模式，尤其在二〇〇〇年代。但是，這種模式也似乎造成了數位研究方式上的路徑依賴，學者容易受制於慣性，停留在關鍵字詞的搜尋功能，而沒有很大的動機使用其他各類數位工具。換言之，關鍵字詞的搜

⁴² 項目的歷史可參見 <https://www.oclc.org/research/archive/projects/chineserarebooks.html>。關於電子的館藏聯合目錄之學術意義，尤其從數位人文研究方面而言，參見王宏魁、徐力恆、包弼德：〈服務於中國歷史研究的網路基礎設施：對相關探索的建議和展望〉一文。

⁴³ 參見 <http://worldcat.org/identities/lccn-no2016148800/>；李友仁：〈北美與西歐的數字人文中國研究狀況論析〉，頁 55。

⁴⁴ 參見王宏魁、徐力恆、包弼德：〈服務於中國歷史研究的網路基礎設施：對相關探索的建議和展望〉一文。

⁴⁵ 祝平次：〈數位工具與文史研究〉，《歷史學柑仔店》，2018 年，<https://kamatiam.org/> 數位工具與文史研究。

尋功能雖然便利，卻蓋過了其他數位人文研究方法可被發揮、發掘的可能性⁴⁶。對文獻進行數位化、開發索引工具、發展字詞的檢索，固然是一種整理知識的方式，但正如項潔、翁稷安總結的那樣：

檢索只不過是索引功能的強化；其邏輯仍是傳統的，並沒有因應數位化時代的前進，而有新的變化。如是，必然造成系統功能的弱化，系統成為史料數位化過程的附加物……關鍵字詞的檢索……該是一個起點而非終點，一個以研究為導向的系統當在繼續鞏固檢索與保存的優勢外，開拓更多的可能，不單只是被動的提供查找，而要能具有主動開創議題的能力。⁴⁷

近來已有數位人文學者提醒，對數位工具的特點和限制需要進行反思，因為它們無疑對學者的研究思路和詮釋方式都有巨大影響。這種反思的範疇已不限於對文獻版本的考察，例如指出某古籍的《四庫全書》本存在問題，不應作為數位化的（唯一）底本；而是應該連帶涉及研究方法層次的思考，作為一種對數位工具的批評反思 (digital tool criticism)⁴⁸。類似的反思，諸多對數位研究工具和資料庫進行反省，甚至質疑的臺灣和中國大陸學者都已有詳論，這裏不再贅述⁴⁹。

值得注意的是，雖然全文資料庫在這段時期得到長足發展，相關的國際編碼標準卻沒有相應地流行起來，例如 TEI「文本編碼規範」(Text Encoding Initiative，又譯「文本編碼倡議」或「文本編碼組織後設資料標誌標準」。TEI 的概念在一九八七年就開始出現，一九九四年公布了完整的標準說明，為文本的電子描述提供了標準的處理方式，至今已包含五百八十種元素的描述標準，在國際數位人文學界影響巨大。利用這種標準的規則進行標記，可以把包括史料在內的各種人文數據轉化為

⁴⁶ 祝平次甚至認為：「數位人文的發展並沒有離開全文檢索的地步多遠……有不少的單位以及個別學者投入不少的精力與時間，貢獻於這個『領域』的發展，可惜這些貢獻並沒有造成有明顯影響力的累積。而臺灣的數位人文之於臺灣的人文學界，目前可以說是一種可有可無的存在。」參同前註。

⁴⁷ 項潔、翁稷安：〈數位人文和歷史研究〉，項潔編：《數位人文在歷史學研究的應用》（臺北：臺灣大學出版中心，2011年），頁18。

⁴⁸ Marijn Koolen, Jasmijn van Gorp, and Jacco van Ossenbruggen, "Toward a Model for Digital Tool Criticism: Reflection as Integrative Practice," *Digital Scholarship in the Humanities* 34.2 (2019): 368-385.

⁴⁹ 我曾分析這些學者的一部分觀點，尤其是2017年以來提出的看法。參見 Tsui, "Charting the Emergence of the Digital Humanities in China," pp. 211-214。

可供電腦進行數位分析的資訊⁵⁰。在臺灣學界，佛典數位典藏處理對 TEI 的運用是最成熟的，源於一九九〇年代的後半期⁵¹。不過在那之前，華文學界的其他學術項目所受到 TEI 的影響並不很大⁵²。這似乎一方面是因為漢字的電腦編碼本身經歷了一段發展歷程，另一方面因為這時數位化處理的中文文獻，絕大多數是有印刷本存在的文獻；不像西方數位人文研究所針對的數位版本 (digital editions)，往往以稿本、抄本為基礎，對詳盡地描述文獻的各種物質上、形制上的特徵有很大的需求。當然，到了二〇一〇年代甚至是當下，隨著針對中文文獻的文本標記工具越來越流行，不同平臺都具備了這種功能，華文學界已經比過去更重視 TEI 及相關標準所包含的學術思路⁵³。甚至有漢學研究者專門以 TEI 的文本標記思路為例，指出數位人文不只可以打開和過去研究截然不同的思路，即使是延續過去思路和「傳統」人文學術的研究步驟，如文獻學方面的研究，利用 TEI 等標準和數位人文方法，也對人文學術大有助益⁵⁴。

有的數位標準比較有普適性，可以適應中文材料，如 TEI 標準；面對某些特定問題，卻未必能直接採用國際的標準，例如對中國傳記資料的整理。面對後一種情況的時候，學者就得自己設計適用的規矩準繩來進行數位開發。其中一個突出的例子是「中國歷代人物傳記資料庫」(China Biographical Database, CBDB)。CBDB 最初是由美國的中國史學者郝若貝 (Robert M. Hartwell, 1932-1996) 在一九七〇至九〇年代開發的一個資料庫。郝氏從一九七〇年代中期即開始發展其中國歷史人物數據

⁵⁰ 該標準的中文介紹，參見魯·伯納·麥克·蘇寶麥昆、馬德偉著，謝筱琳、黃章寧譯：《TEI 使用指南——運用 TEI 處理中文文獻》（臺北：數位典藏與數位學習國家型科技計畫——拓展臺灣數位典藏計畫，2009 年）。TEI 倡議的網站見：<http://www.tei-c.org/index.xml>。

⁵¹ 杜正民：〈推薦序〉，同前註，頁 10-14。

⁵² 陳靜：〈複數的數字人文——比較視野下的中國數字人文〉，《中國比較文學》，2019 年第 4 期，頁 22。

⁵³ 關於數位化過程中標記的作用，簡明介紹可參見王雅萍等：《數位化工作流程指南：文字資料》（臺北：數位典藏與數位學習國家型科技計畫——拓展臺灣數位典藏計畫，2012 年）。運用標記的其中一個例子，可以參見魏希德著，徐力恆譯：〈唐宋史研究中的數字化語文學〉，《唐宋歷史評論》第 3 輯（北京：社會科學文獻出版社，2017 年），頁 3-19；Tsui, “Charting the Emergence of the Digital Humanities in China,” pp. 206-208.

⁵⁴ Scott Paul McGinnis, “DH and Humanities Workflows: Two Case Studies,” in *Doing More Digital Humanities*, eds. Constance Crompton, Richard J. Lane, and Ray Siemens (Abingdon, UK: Routledge, 2020), pp. 58-69, esp. 62-67.

研究，並在九〇年代發表論文，對這一研究方法進行了理論探討和實證研究，把他的數據開放給學者使用⁵⁵。他的典範性研究論文〈750-1550年間中國的人口、政治及社會轉型〉對唐中期到明朝中葉間士人入仕方式、家族形態和地區經濟升降等方面問題，提出了宏觀解釋，對歐美學界的中古中國研究產生了重大影響。在這種跨朝代的大規模人物群體分析中，他明確提及「群體傳記學」(prosopography，或譯人群學)的研究取徑⁵⁶。群體傳記學興起於一九二〇至三〇年代對歐洲歷史的研究，它指的是「通過對一群人物生平的集體性研究，來探討他們共同的背景特徵」，採用的具體方法是「先建立一個研究範圍，然後詢問一組系統的問題：關於出生與死亡、婚姻與家庭、社會出身與其所繼承的經濟地位、居住地、教育、個人財富的數量與來源、職業、宗教、仕歷等等。之後將這一範圍內所有個人的各種信息對比、組合，並探析其重要的變量。研究者會分析這些信息的內在相關性，以及它們與其他類型的行為與活動的相關性」⁵⁷。簡言之，群體傳記學的主要注意力不在單個人物的生平，而在群體特徵或其內在聯繫。這樣的一種研究角度，在二〇世紀下半葉數位資料庫技術迅速發展之下，大批儲存和分析人物資料比以前便利，群體傳記學也因而獲得了新的關注和推進⁵⁸。採用這種方法具代表性的研究計畫，包括牛津大學歷史系的「群體傳記學研究組」(Unit for Prosopographical Research)，也包括各個專門領域的計畫——類似的群體人物研究計畫仍在不斷出現、發展中⁵⁹。但是，由於用在

⁵⁵ Peter K. Bol, "The Late Robert M. Hartwell 'Chinese Historical Studies, Ltd.' Software Project," Projects at Harvard, https://projects.iq.harvard.edu/files/cbdb/files/the_late_robert_m._hartwell_chinese_historical_studies_ltd_software_project.pdf.

⁵⁶ Robert M. Hartwell, "Demographic, Political, and Social Transformations of China, 750-1550," *Harvard Journal of Asiatic Studies* 42.2 (1982): 365-442; Robert M. Hartwell, "A Computer-Based Comprehensive Analysis of Medieval Chinese Social and Economic History," in *Characters and Computers*, eds. Victor H. Mair and Yongquan Liu (Amsterdam: IOS Press, 1991), pp. 89-121. 前文中譯本見易素梅、林小異等譯：〈750-1550年中國的人口、政治與社會轉型〉，單國鈺編：《當代西方漢學研究集萃·中古史卷》（上海：上海古籍出版社，2012年），頁175-246。

⁵⁷ Lawrence Stone, "Prosopography," *Daedalus* 100.1 (1971): 46.

⁵⁸ 參見 Karen S. B. Keats-Rohan, ed., *Prosopography Approaches and Applications: A Handbook* (Oxford: Unit for Prosopographical Research, Linacre College, 2007)。

⁵⁹ 如「羅馬共和國的數位群體傳記學」(Digital Prosopography of the Roman Republic)、「古希臘羅馬銘刻記載人物」(Romans 1by1)、「盎格魯—薩克遜時期英格蘭群體傳記學」(Prosopography of Anglo-Saxon England)、「埃及中王國人名資料庫」(Persons and Names of the Middle Kingdom)、「拜

群體傳記學的數據，都是從具體的歷史環境中產生的，所以對數據所施加的模型化 (data modeling) 往往需要專門設計，以適應材料的特點⁶⁰。就 CBDB 的數據而言，數據模型必須考慮中國古代，尤其是中國中古時期社會的特點。

郝教授在去世前把載有二萬五千餘人的傳記資料，以及五千餘條書目的數據捐給了哈佛燕京學社 (Harvard-Yenching Institute)，後來由哈佛大學的費正清中國研究中心聯合其他機構管理，成為 CBDB 的核心基礎。作為 CBDB 的前身，郝教授搜集資料的工作開展於網路尚未普及、各種數位工具尚未開發的一九七〇至八〇年代，反映了數位人文史前時期的部分狀況——他的傳記資料編碼都以手動方式輸入，當時電子條目甚至不包含漢字，而是全以英文和中文拼寫來記錄。不少他當年留下的工作紀錄仍保存在 CBDB 相關人物的備註欄位中。從技術角度看，這雖然不無時代色彩，但當時的資料庫設計已經隱含數位人文理念。首先，所搜集的數據都建立在原始的史籍記載和研究論著之上，借助了人文學界研究中國文史的成果⁶¹。其次，所搜集的人物資訊可用以進行多樣的查詢和分析，並且靈活地導出到其他軟體，以便進行批量處理；或用不同方式來呈現，如統計表格、電子地圖等。第三，務求利用電腦資料庫的眾多記載，發揮巨量資料研究群體的優勢。因此，到了數位人文研究概念被提出而普遍為學界接納的二〇一〇年代以後，基於這類 CBDB 人

占庭帝國群體傳記學」(Prosopography of the Byzantine Empire)，還有關於伊斯蘭教 (Onomasticon Arabicum)、耆那教 (Jaina-Prosopography) 人物，以及關於澳大利亞 (Biographical Database of Australia)、日本 (Japanese Biographical Database)、新加坡 (Singapore Biographical Database)、臺灣 (Taiwan Biographical Database) 歷史人物等各個研究計畫。

⁶⁰ 關於數位人文研究中對數據的模型化，參見 Julia Flanders and Fotis Jannidis, "Data Modeling in a Digital Humanities Context: An Introduction," in *The Shape of Data in Digital Humanities: Modeling Texts and Text-based Resources*, eds. Julia Flanders and Fotis Jannidis (Abingdon, UK: Routledge, 2018), pp. 1-23。

⁶¹ 關於 CBDB 的數據化工作以及其中涉及的研究過程，參閱徐力恆：〈唐代人物大數據：中國歷代人物傳記資料庫 (CBDB) 和數位史學〉，譚國根、梁慕靈、黃自鴻編：《數碼時代的中國人文學科研究》，頁 121-139；Lik Hang Tsui and Hongsu Wang, "Semi-Automating the Transformation of Chinese Historical Records into Structured Biographical Data," in *Digital Humanities and Scholarly Research Trends in the Asia-Pacific*, eds. Shun-han Rebekah Wong, Haipeng Li, and Min Chou (Hershey, PA: IGI Global, 2019), pp. 228-246；胡斌、徐力恆：〈人物傳記研究數據庫：中國歷代人物傳記資料庫 (CBDB) 的探索〉，邱偉雲、陳靜編：《數字人文實踐：數據庫建設》（南京：南京大學出版社，出版中）。

物數據做出來的研究成果大大增加⁶²。

從本文採取的數位人文史前史觀察角度來看，北京大學的史睿從學術史和媒介的角度為 CBDB 的源流提出之定位是富有意義的，以下進行引述：

上世紀三〇年代開始，哈佛燕京學社在北京建立引得編纂處（今北京大學校內），這是現代學術史上極具關鍵性的大事件。當時不僅引得編纂處出版了系列古籍索引，還有中法漢學研究所的系列「通檢」，以及國立北平圖書館的《國學論文索引》、《清代文集篇目分類索引》、《石刻題跋索引》，稍後出版的有譚其驤主編的《中國歷史地圖集》。這些引得、通檢、索引、地圖看起來是普通的學術出版物，實際是中國學術現代性轉型的重要表現。此前的學術研究依賴個人的記憶、功力和視野，有了上述這些索引工具，便能聚合學術共同體的成就，形成有序有效的社會知識積累和傳播，中國從此跨入了現代學術的大門。中國學術界今天面臨的情況跟上世紀三〇年代引得編纂處時代的情況完全一樣，正處在向新學術轉變與跨越的門檻上，即利用新媒體、新技術實現更為深入新穎的知識重組，甚至代替學者的部分重複性和基礎性工作。而且有意思的是，……在燕京大學舊地正在實施的「中國歷代人物資料數據庫」項目（簡稱 CBDB），正如同新媒體時代的新「引得」事業。⁶³

筆者認為，史氏的評論值得注意的，不只是其關於 CBDB 例子的討論，而至少有以下三點可供闡發：第一，過去有的重大文獻整理項目是群體及中外合作，當下也是。如 CBDB 就是美國、臺灣和中國大陸的機構合作進行開發的資料庫，由哈佛大學費正清中國研究中心、北京大學中國古代史研究中心、中央研究院歷史語言研究所共同主持。而且它是集體協作的，聚合了幾代學者、多人的心血。就本文關注的焦點來看，跨國、跨地區學術合作，是數位人文出現以前已存在的突出現象。

第二，編纂大型的基本工具書，是一種知識重組的學術事業。放在數位的新媒體時代，同一工作的表現就是開發資料庫和數位工具，而這些表現當然和今天一般

⁶² 一般而言，學者對 CBDB 資料的運用，主要集中在幾個方面——除了群體傳記學的研究以外，還經常用於社會網路分析、地理空間分析、可視化之上。2010 年代出版的部分成果之介紹，參見 Lik Hang Tsui and Hongsu Wang, “Harvesting Big Biographical Data for Chinese History: The China Biographical Database (CBDB),” *Journal of Chinese History* 4.2 (2020): 505-511。

⁶³ 史睿：〈數字人文研究的發展趨勢〉，《文匯報》（上海）第 7 版，2017 年 8 月 25 日。

理解的數位人文研究密不可分。當然，除了史氏提及的 CBDB 項目，上文談到的香港中文大學劉殿爵中國古籍研究中心所開發的資料庫，其成果也繼承了這種「引得」的學術傳統。其實，考慮到數位人文研究起源的國際背景，語彙索引等工具書的編纂及數位開發，不只對中文材料的數位化有重要作用，在歐美學界也是如此。許多學者都熟知，人文計算甚至是數位人文的起源，可以追溯到義大利神父羅伯托·布薩 (Roberto Busa, 1913-2011) 在一九四九年開始使用電腦處理神學家阿奎那 (Thomas Aquinas, 1225-1274) 的全集，半自動化地生成其中中世紀拉丁文字詞的索引 *Index Thomisticus*。這項工作歷時數十年，至一九八〇年才完成，並曾得到 IBM 公司的協助⁶⁴。影響所及，尤其在一九六〇年代起，學者越發體會到語彙索引的價值。無論是研究《新約聖經》的希臘文內容、中古德語文獻、英國詩歌和莎劇，還是法語、荷蘭語文獻等，學者開始使用基於穿孔卡和磁帶儲存的電腦，對經典文本進行語彙索引、作者身分和風格學 (stylistics) 等文本分析研究工作。同時，各個領域的學者也越來越瞭解，電腦在其中可以扮演的重要角色，並積極嘗試運用技術來解決計算問題⁶⁵。

第三，如果說史氏的引文中，前半段所描述的是中國人文學者邁入現代學術的轉型期，那麼他所言今天「正處在向新學術轉變與跨越的門檻上」，即數位人文學

⁶⁴ 當然，如同一切起源故事一樣，一般認知中的這個故事不是毫無疑義的。戴安德也強調過這一點，參見戴安德：〈數字人文及其史前史〉，頁 144-145。借用王明珂的說法，人們對起源的敘述和建構往往會形成一種「魔力」。參見王明珂：〈起源的魔力及相關探討〉，《語言暨語言學》第 2 卷第 1 期（2001 年 1 月），頁 261-267。以布薩主持的項目作為數位人文研究的起源為例，布薩開始做的一些人文計算試驗，其實是基於類比式 (analog) 電腦和機械儀器，而非今天的數位化 (digital) 電腦。有學者因此認為，把數位人文的起源簡單地追溯到布薩的嘗試，未免有些時間錯置，經不起推敲，參見 Dean Irvine, "From Angel to Agile: The Business of the Digital Humanities," *Scholarly Research and Communication* 6.2 (2015): 2。布薩及其合作者到 1960 年代才用到數位化電腦，但即使是在那個時候，他們的項目仍大量使用人力和打孔卡，尤其是女性操作員，參見 Melissa Terras, "For Ada Lovelace Day Father Busa's Female Punch Card Operatives," *Melissa Terras: Adventures in Digital Cultural Heritage*, October 13, 2013, <https://melissaterras.org/2013/10/15/for-ada-lovelace-day-father-busas-female-punch-card-operatives/>。關於布薩上述項目早期歷史的深入研究，可參見 Steven E. Jones, *Roberto Busa, S.J., and the Emergence of Humanities Computing: The Priest and the Punched Cards* (New York: Routledge, 2016)。

⁶⁵ Hockey, "The History of Humanities Computing," pp. 4-7。中譯本見蘇珊·霍基著，葛劍鋼譯：〈人文計算的歷史〉，《文化研究》第 16 輯（2013 年秋），頁 173-193。

術範式所開創的人文研究。經過史前時期或遠或近的鋪墊準備，數位人文研究正式在華文學界浮出地表了。到了這個時候，所謂的「史前史」時代也快將結束，邁向數位人文時代了。

*

*

*

*

大約在二〇〇〇年代以後，個人電腦及許多應用軟體變得十分普遍，成為幾乎所有人文學者都能方便使用的設備。在研究中完全不用電腦，反而成了咄咄怪事。這也意味著人文學科正發生「數位轉向」(digital turn)。二〇〇〇到二〇一〇年，大致上是「人文計算」逐步被「數位人文」概念取代的時代。根據學者論述，數位人文和人文計算主要有以下三點不同：第一，數位人文研究得以使用的是龐大得多的數據；第二，數位人文強調的不只是「計算」，還積極追求應用其他研究手段；第三，學者進入這個領域的門檻變得更低，產生了比人文計算大得多的學術影響⁶⁶。其實，這些趨勢用以描述華文學界二〇〇〇到二〇一〇年前後的數位研究也是合適的。考慮到「數位人文」這一名稱在歐美，具體是在二〇〇四到二〇〇八年之間出現並流行，傳到臺灣和中國大陸其實沒有花太長的時間⁶⁷。臺灣學界自二〇〇九年起，每年舉辦「數位典藏與數位人文國際研討會」，成為學者交流及數位人文研究的重要平臺；同年，「數位人文」概念也進入了中國大陸學界。因此，華文學界的「數位人文」時代，可說是在二〇〇九年正式開始⁶⁸。適逢數位人文、巨量資料等概念在學界引起關注和討論，有更多學者對數位研究工具和資源感興趣，為人文研究中利用數位研究手段、分析數位資料帶來了重要契機。各種相關論著和研究計畫、資料庫，都在三地和海外漢學界發展起來。到了二〇一六年，臺灣數位人文學會成

⁶⁶ 關於「數位人文」概念，參見項潔、涂豐恩：〈導論——什麼是數位人文〉，頁 9-28；徐力恆、陳靜：〈我們為什麼需要數字人文〉，《社會科學報》第 5 版，2017 年 8 月 24 日。歐美學界對數位人文的定義有過諸多辯論，可參閱上述論文和 Melissa Terras, Julianne Nyhan, and Edward Vanhoutte, *Defining Digital Humanities: A Reader* (Farnham: Ashgate Publishing Limited, 2013)。Melissa Terras 等人所著一書中文版，將由南京大學出版社出版。

⁶⁷ Steven E. Jones, *The Emergence of the Digital Humanities* (Abingdon, UK: Routledge, 2013), pp. 4-9.

⁶⁸ 陳靜：〈當下中國「數字人文」研究狀況及意義〉，頁 59；Chen and Tsui, "Debating and Developing Digital Humanities in China: New or Old?".

立，隨後亦有《數位典藏與數位人文》學刊的創辦。至於中國大陸和香港方面，二〇一五年開始，包括上海大學、北京大學、南京大學、南開大學、（北京）清華大學、中央民族大學、香港公開大學、香港恆生大學等多所學校，都舉辦了與數位人文主題相關的會議，哈佛大學亦於二〇一八年發起在上海舉辦的「中國歷史研究的網絡基礎設施」國際研討會。這些都反映數位人文已經引起三地學界的諸多關注。至此，數位人文已經成為一股學術潮流，影響遍及各人文學科——在當下的數位時代、資訊化時代，如何利用新科技來處理、擴展人文學科的課題，無疑成為人文學者必須面對的挑戰⁶⁹。

就如上文所述，二〇一〇年代的數位人文歷程，是建立在過去二十餘年的數位探索之上的，尤其包括本文所提及史前階段的各種探索。而本文的分析反映，我們不能簡單地把數位人文視作外來的新研究範式，更不能把它在華文學界的歷程，簡化為一種歐美學術範式的移植過程。單以本文聚焦的問題來說，華文學界學者進行中文材料的數位開發和研究，往往因為文獻的語言，而面對和其他語言截然不同之挑戰。因為這個原因，不熟悉中文材料的國際數位人文學者，往往不瞭解華文學界都做過哪些探索，也因而可能簡單地把數位人文當作是一件在華文學界憑空降臨的新事物。

另外，這二十餘年的歷程也遠遠不像是一場「革命」或翻天覆地的變化，而是「常」(continuity)與「變」(change)同時並存，互為作用的一個複雜過程。當然，如果把華文學界的數位人文探索放到更廣闊的語境來看，例如資訊科技的歷史，我們能夠得出的基本認知也往往是類似的——它通常會讓人得出一種比較平和的看法，而不會讓人處處看到新發明和技術革命的影子⁷⁰。實際上，當我們充分考慮科技的歷史進程，所謂「數位化革命」、「資訊化革命」等說法往往只是一種誇飾，一種修辭，甚至是一種幻覺⁷¹。換言之，有的數位人文「先驅」所奠定的成果和學術

⁶⁹ 關於數位人文範式在華文學界出現的標誌和各種表徵，參見 Tsui, "Charting the Emergence of the Digital Humanities in China," pp. 204-211。

⁷⁰ 人類社會往往高估了科技新發明所造成之影響。對這進行反思的論著有很多，包括 David Edgerton, *The Shock of the Old: Technology and Global History since 1900* (New York: Oxford University Press, 2007)，中譯本為大衛·艾傑頓著，李尚仁譯：《老科技的全球史》（臺北：左岸文化，2016年）。

⁷¹ 對類似看法的論證，可以參見 Brian Winston, *Media Technology and Society: A History: From the*

線索，沒有得到「後浪」的呼應，戛然而止；而有的「新銳後進」所採取的數位研究路徑，卻又偶然地和「前輩」所做的隱然脗合，意外地達到傳承之果效。從數位人文的史前史到它本身的歷史，發展歷程往往是如此迂迴曲折，步伐是如此難以逆料。希望本文的分析，可以為我們思考這些數位人文「史前」時期的現象提供一點線索。

Telegraph to the Internet (London: Routledge, 1998)。關於「資訊革命」一詞從 1960 年代以來的流行的社會背景，參見 Alex Sayf Cummings, “Of Sorcerers and Thought Leaders: Marketing the Information Revolution in the 1960s,” *The Sixties* 9.1 (2016): 1-25。

