數位人文研究叢書 2
Series on Digital Humanities

數位人文研究的新視野：基礎與想像

項潔 編

New Eyes for Discovery:

Foundations and Imaginations of Digital Humanities

臺大出版中心
NATIONAL TAIWAN UNIVERSITY PRESS

國家圖書館出版品預行編目資料

數位人文研究的新視野：基礎與想像／項潔編. --
初版. -- 臺北市：臺大出版中心出版：臺大發行,
2011.11
面；　公分. --(數位人文研究叢書；2)

ISBN 978-986-03-0165-6 (精裝)

1.人文學 2.文獻數位化 3. 數位科技

119.029　　100024129

數位人文研究叢書2
Series on Digital Humanities

數位人文研究的新視野：基礎與想像

New Eyes for Discovery: Foundations and Imaginations of Digital Humanities

策　劃　國立臺灣大學數位典藏研究發展中心
叢書主編　項　潔
叢書編輯　陳怡君　蔡炯民

總　監　項　潔
總 編 輯　湯世鑄
責任編輯　游紫玲
編輯協力　方誼　李竺頤
封面設計　楊啟巽
內頁編排　極翔企業有限公司

發 行 人　李嗣涔
發 行 所　國立臺灣大學
出 版 者　國立臺灣大學出版中心
法律顧問　賴文智律師
印　刷　中原造像股份有限公司
出版年月　2011年11月
版　次　初版
定　價　新臺幣400元整

展 售 處　國立臺灣大學出版中心
臺北市10617羅斯福路四段1 號
電話：(02)2365-9286　傳真：(02)2363-6905
臺北市10087思源街18號澄思樓1 樓
電話：(02)3366-3991~3 轉18　傳真：(02)3366-9986
http://www.press.ntu.edu.tw E-mail: ntuprs@ntu.edu.tw
國家書店松江門市　電話：(02)2518-0207
臺北市10485松江路209號1樓
國家網路書店 http://www.govbooks.com.tw

GPN：1010003789
ISBN：978-986-03-0165-6

數位人文研究叢書 2
Series on Digital Humanities

數位人文研究的新視野：基礎與想像

項潔 編

New Eyes for Discovery:

Foundations and Imaginations of Digital Humanities

序

從全球各國數位典藏發展的歷史來看，臺灣開始的不算太早，但也不算晚。暫且不論個別學校、單位或個人的研究和努力，從國科會在 1998 年以「數位博物館計畫」開始投入國家資源，有系統地發展數位典藏來算，到現在也已將近 15 年了。在這段不算短的時間裡，我國投入了大量經費，也將大量的文化資產數位化，其中最重要，從 2002 年開始執行的「數位典藏國家型計畫」，更產生了指標性的作用，讓「數位典藏」在臺灣成為一個大眾語言內的詞彙。

數位典藏國家型計畫的成功，至少有一部分歸功於許多資深且傑出的研究人員，尤其是人文學者，不計報酬地全心投入。但是從大約 2005 年開始，我們漸漸發覺這些優秀的學者在他們本身的研究中，似乎並沒有充分利用他們花了大量精力數位化的檔案；這令我感到困惑，因為這樣不是事倍功半嗎？為什麼不能將數位化工作和本身的研究結合在一起呢？觀察到這個現象後，我開始去探討它的環節。我發現其中的一個重要原因是數位典藏系統的設計往往沒有考慮到使用者——尤其是研究者——的需求，以致於一直到現在絕大多數的研究者還是認為檢索系統只能幫他們找資料罷了，而不能幫忙整理或分析資料。這是很可惜的，因為許多資訊技術已經十分成熟，如果能夠與數位檔案結合並有效地運用在這些系統中，應對人文研究產生非常大的助益。在更進一步的探討後，我發現這個問題並不是臺灣獨有的，在國外亦有學者思考這個問題，而且已有一個研究社群，那便是「數位人文」。

用最簡單的話來講，「數位人文」就是結合大量數位材料，運用資訊科技，來從事人文研究。顧名思義，這是一項跨領域整合的工作。要達到這個目標，除了要有大量高品質的數位資源可供使用外，更需要人文學者與資訊學者密切的互動與合作。有鑑於此，國立臺灣大學數位典藏研究發展中心從 2009 年開始舉辦每年一度的「數位典藏與數位人文國際研討會」，希望藉由這個會議，不但能夠介紹世界最新的研究成果，並能提供國內外人文與資訊學者交流的機會，讓臺灣的數位人文研究能夠遍地開花，並與國際接軌。

《數位人文研究叢書》即是這個年度會議產出的一項成品。叢書中的每一篇文章均在研討會中發表過，再經修改及至少兩位審查人的審查通過。在此特別感謝臺大數典中心的蔡炯民博士、陳怡君小姐與全體同仁對本叢書投注的心力。我們希望透過這個系列的叢書，提升國內學界對數位人文的認知，並激發進一步的研究。

項潔

2011 年 9 月於臺大

Preface

Taiwan started its cultural digitization effort in earnest in 1998, when the National Science Council initiated the Digital Museum Program. Since then the government has invested a large amount of money and has digitized a significant portion of Taiwan's cultural heritage. Indeed, the effort has been so successful that 數位典藏 , the Chinese translation of "digital archives", has become a household term in Taiwan.

The success of the digitization enterprise, in particular the National Digital Archives Program, is at least partly due to the altruistic devotion of many senior scholars. However, around 2005 I started to notice that many of these scholars do not seem to utilize the outcome of their digitization work in their own research. This phenomenon baffled me. Since an important purpose of digitization is to make materials more accessible, why couldn't they use the fruits of their own hard work to expedite their own research? After some studies, I observed that at least one reason is that the digital archive systems are not designed with the need of the users, especially researchers, in mind. Thus most people still think that a retrieval system can do just that, retrieve, but not to help them organize, observe, and explore what has been retrieved. This is a pity, because combining information technology with massive high quality digital objects should provide tremendous opportunities for humanities research. I then learned that this phenomenon is not unique to Taiwan. Indeed, many scholars in the world have been thinking about this challenge, and the community is called Digital Humanities.

To put it in the simplest term, digital humanities is humanities research with the help of digital materials and information technology, with an emphasis on the type of research that cannot be accomplished otherwise. It is interdisciplinary by nature. In 2009, the Research Center for Digital Humanities of the National Taiwan University started an annual International Conference on Digital Archives and Digital Humanities to serve as a forum for researchers from different disciplines and different parts of the world to showcase their work and exchange ideas. This series, the Series on Digital Humanities, is a product of this effort. All papers in the volumes have been presented in one of the conferences. The final version is then submitted and went through rigorous reviews. We hope that the Series will promote digital humanities and stimulate further research.

Jieh Hsiang
September, 2011
National Taiwan University

目　錄
Contents

Part II 工具的進步
Technologies Forward

Part III 現象的探索
Ground Truth

國立臺灣大學數位典藏研究發展中心於2010年舉辦第二屆數位典藏與數位人文國際研討會，共有二十八篇論文發表。經會議中的熱烈討論與交流，由作者參酌會中所獲得之回饋意見進行文章修改，並經十二位初審委員與十二位複審委員的匿名審查；經過嚴謹的審查程序與作者修正，最後擇取十八篇優秀論文，並將其分成上下兩冊出版，每冊各收錄九篇文章。本書即為其中的第一冊。

導論——關於數位人文的思考：理論與方法

項潔 *、翁稷安 **

摘 要

數位人文於今日已逐漸受到重視，在不斷研發其各種可能性的同時，也必須進一步去思索數位人文作為一個學術門類所該有的規範。本文回顧了數位人文發展的歷史，並特別介紹了金觀濤對數位人文的討論。金觀濤從哲學的角度出發，去討論了數位人文作為知識的方法論，是十分值得重視的洞見，但同時卻也有著無法迴避的質疑。我們認為作為一實作性很強的學門，只由抽象的角度去理解數位人文的方法論是不夠的，也必須從實踐中獲得。我們指出一個以研究為取向的系統在數位資料運用上所扮演的關鍵角色，唯有建立這樣一個功能強大的系統，使用者才能更自由地依自己研究所需，去觀察史料，建立、發挖出史料間的脈絡，開展出自己的論述。也唯有以一個研究取向的系統為基礎，研究者和資料之間的關係才會真正被改變，這才是我們思考數位人文方法論，乃至其未來發展的起點。

* 國立臺灣大學資訊工程學系特聘教授。
** 國立臺灣大學數位典藏研究發展中心碩士後研究人員。

Introduction
Pondering on Digital Humanities: Foundation and Methodology

Jieh Hsiang *, Chi-an Weng **

Abstract

As Digital Humanities gains popularity and extends its reach into more domains of humanities, it becomes imperative to consider issues facing Digital Humanities as it emerges into a new discipline. This paper gives a very brief overview of Digital Humanities before discussing in greater length of the article by Professor Jin Guan-Tao, who provides a treatise of the foundation of Digital Humanities from an epistemological point of view. We feel that as a discipline originated from application, Digital Humanities' foundation cannot be fully grasped from an abstract angle. We point out the crucial role played by a software (retrieval, presentational) system in the utilization of a digital archive. Only with a user-oriented, dynamic system design, can a scholar fully observe, discover, and explore the contexts explicitly or implicitly embedded in the digital materials. Furthermore, a system developed for humanities research purposes can claim to truly satisfy research needs if it is developed with a close collaboration between its users (humanities researchers) and builders (computer scientists). Thus, such a system will fundamentally change the relationship between the humanists, the user, the information technologists, the designer, and digital archives, the content. Putting system into consideration is, in our view, an important starting point for the methodology for Digital Humanities, and is a future direction of its development.

* Distinguished Professor, Department of Computer Science and Information Engineering, National Taiwan University.

** Research Associate, Research Center for Digital Humanities, National Taiwan University.

一、前言

近年來，「數位人文」已成為一眾人矚目、蓬勃發展的領域，相關的研究成果無論是專業論文、數位典藏或資料庫的建置等等皆如雨後春筍般地出現，並開始有各種大型計畫的推動，這樣的榮景皆說明了數位人文所引起的重視。在本書所收錄的諸篇文章中，看似各自獨立的命題，其實有著共通的關切，即希望從理論、方法以及應用等各個層次上，替作為一新興學術範疇的「數位人文」尋找其在研究上的優勢，乃至更進一步為這個新興領域建立起一套規範（discipline）。誠如我們所不斷強調的，數位人文作為一個發展中的新領域，應該以一個開放、包容的定義，取代太過明確、狹義的界定，如此才能夠包含各式的可能和想像，擴大數位人文所蘊藏的能量和潛力。（項潔、涂豐恩，2011）但在不斷擴大可能的同時，也必須持續對學科內部已發展的諸項議題深化的思考，在花團錦簇的榮景之外，進行扎根的工作。換句話說，作為在初始階段的數位人文，必須既是「狐狸」也是「刺蝟」，要能「博」也能「約」，誠如中國古代儒者劉宗周所言：「博而不約，俗學也；約而不博，異端也。」數位人文也必須在這兩端的發展間取得平衡，否則將難以面對來自其他領域研究者合法性的質疑。

二、數位人文的發展歷程

數位人文和所有的學術門類一樣，皆非憑空而生，而是長久摸索、蘊釀之後的結果。當我們在強調其「新興」的一面時，不能忽略它其實有其自身發展的源頭和積累。作為一個長期關注於資訊技術和人文研究兩者間關係的學者，Susan M. Hockey 將數位人文的起點溯源至上世紀「人文計算」（Humanities Computing）的發展，於 1949 年 Roberto Busa 神父開始使用電腦對神學家 St. Thomas Aquinas 著作內的字詞進行大規模的處理，包括每個字的用法、位置，以類似索引的方式，試圖釐清其規則，花了三十年的時間，終於將成果出版。以此為起點，開啟了以電腦自動整理文本內容的研究方式，以電腦為主要的工具，用計量的方式對文學作品的文本進行分析，來討論書寫的風格。最有名的研究應用，是用這種分析寫作習慣的方式，來判斷莎士比亞的作品何者為真，何者為他人託名之作。除了文學之外，在歷史學的領域，也有學者開始應用資訊科技面對大量資料的優勢，進行統計和量化的研究，開啟所謂的「計量史學」。

進入 1970 年代乃至 80 年代中期，資訊技術和人文研究間的關係從早期的摸索進入了整合期，隨著電腦日益的普及，越來越多研究者和機構開始思考電腦可以為人文研究和教學帶來什麼樣的變化？相關的教學組織、研究機構紛紛成立，投注大量的時間與心力於其中，這些組織成為數位和人文結合的重要支柱，開始進行相關

的研究計畫，發行期刊與舉辦研究會。但數位人文真正突破性發展應為 2000 年前後十年，個人電腦的普及與網際網路的出現，使數位和人文的結合提升至另一層次，數位化典藏、資料庫的建立成為可能，如雨後春筍般，世界各地都開始致力於將文物、檔案等資料數位化，置於網路珍藏的行列。各式各樣的資料庫開始出現，多數的歷史文獻被掃描、拍照，製成了與原件極為相近的、存真度高的數位影像，並由專人撰寫補充描述文獻背景的 metadata（詮釋資料）；亦有以文獻內容的全文打字的方式，進行數位化。（Hockey, 2004）凡此，都對人文研究環境和知識取得帶來了巨大的改變，資料庫的檢索與應用逐漸成為研究者在研究過程必要的環節；而綜合人文計算和數位典藏所累積的經驗和能量，也開啟了數位人文的新頁。

三、數位人文理論基礎的探索：從知識論的角度

首先，就數位人文的理論基礎而言，本書所收錄的金觀濤〈數位人文研究的理論基礎〉一文，是篇很特殊的文章，不同於目前多數的討論都是從數位人文自身發展的角度出發，去討論數位人文該如何界定與看待，金文改從哲學的視界，試圖從知識論的角度，去討論人文科學的本質是什麼，而數位科技又能從中扮演什麼樣的角色。或因為金觀濤作為一個長期鑽研於思想史和觀念史之中的研究者，所以才使他能有那麼獨特的視角，從另一個層次探索數位人文的理論。

金觀濤首先指出人文知識和自然科學、社會科學的知識不同，並無客觀性的原則，而是基於他人經驗在研究者心中的可重演性。因此數位分析在人文研究中所扮演的角色也不同於其在自然、社會科學中的應用，不是針對數據統計（事實分析），而是在處理歷史文獻和各式各樣的文本；又因為文本直接涉及了人文研究的核心，所以數位科技在人文研究的重要性勢必將遠大於它在自然、社會科學中的比重。為了適應各類人文研究的需求，不同的研究皆需要有適合、能與之對應的電腦數據庫的建立；金教授認為在這樣的條件下，隨著 IT 技術對文本深度挖掘技術的開展，將會出現一門稱之為數位人文學的新學科。金教授回顧了由 R. G. Collingwood 到 Reinhart Koselleck，一直到中國當代觀念形成之研究，勾勒出數位人文學在觀念史研究中的起源、發展過程，以及數據庫統計之分析、文本深度挖掘和數位分析技術在人文研究中所扮演的意義。金教授在最後總結道，數位方法和人文研究的結合，猶如科學假說與實驗配合，將形成人文研究中長程的二階（second order）反思視野。

當然，作為一個還未定型、充滿各種可能的學術領域，數位人文的理論亦尚未定論，金觀濤的論述是重要的參考，開啟了一個可能的思索方向，畢竟從抽象的角度去思考數位人文在知識光譜上的定位，是十分難能可貴的；一個成熟的學術門類也必須具備本體論和知識論上的嚴謹規範的定義，才能禁得起考驗。但這並不等

於這樣的洞見即是最後的定論。一方面，即便如金觀濤所言，數位人文在觀念史研究上有著明顯優勢，可是並不等同於數位人文全部的潛能，從歷史研究角度而言，不論是政治史、社會史、文化史乃至生活史等等，一個資料完備、功能強大的資料庫，無疑都能帶來很大的幫助。更重要地，觀念在歷史的進程扮演著很重要的角色，但人類的發展不單僅依靠著觀念的作用；觀念的演變歷程該是研究者審視過去時的重點，但不應是唯一；作為以人文世界為核心關懷的「數位人文」，自然亦不該以此自限。

此外，雖然金觀濤和劉青峰兩位學者，很努力的強調其研究和計量史學的不同，卻很難化解其他研究者對此點的質疑與不安。面對這樣的詰問，他們作出如下的回覆：「我們所作的觀念史研究不能與計量史學混為一談。我們自己的定位是以關鍵字為中心的觀念史研究，這是典型的人文學科，只不過引進了資料庫方法。……我們強調了資料庫在人文研究中只有輔助作用，它為研究者提供了極大的便利，也提出了更高的要求。它只是在對關鍵字的使用情況和類型分析這一素材搜集和整理環節上提供了工具，而研究者在此基礎上，要以人文學科的基本範式和自己的研究素養來分析這些資料。只要不存偏見，閱讀我們相關論文，就可以看出，論點的得出，並不完全依靠統計。實際上，統計只是思想史分析的一項工具。否則，按一下電腦，列一大堆圖表，就可以交差，研究也不必做了。簡而言之，以關鍵字為中心的觀念史研究並不是計量史學，它本質上利用電腦資料庫這一工具，令其服務於人文研究，套用一句學者的評價，它是數位人文研究的開始。」（金觀濤、劉青峰，2010）但卻還是不能說服質疑者，畢竟在面對大量的統計和圖表下，很難不令人直覺地將其與計量史學加以聯想。

這樣的挑戰並不是金、劉二人所獨有，在很大的程度上，是現階段許多數位人文技術於史學應用時很容易會面臨的懷疑；因為資訊技術的長項本來就在於大規模資料的處理上，其結果的呈現很自然地會和統計與圖表相結合。是以，對於這樣的質問，一時之間很難輕易的加以回答。不過我們必須強調，計量史學的失敗並不等同於統計方法於人們看待過去時毫無用武之地，正好相反，計量史學的失敗並不在於統計本身，而是在於過度迷信統計，失去了人文的眼光和判斷；這正好是數位人文的起點，數位工具肩負起資料保存和處理的功效，但要能真正產生意義和影響，還是需要人文思維和專業的投入，兩者的有機結合，才能避免重蹈計量史學的覆轍，才能開展出數位人文的獨特氣象，金、劉二人的答辯絕非只是避實就虛的遁詞而已。誠如我們再三強調的數位人文不同於人文計算，絕不迷信數字可以解決所有人文問題。

再者，金、劉二人的回應無法說服質疑者的另一個關鍵，在於這是一個典型「有待來者」的難題，因為所謂數位和人文之間的「有機融合」是個十分抽象、微妙

的概念，僅從理論的面向回應，也難以一言道盡，取得理解的共識。唯有持續不間斷繳出實績，讓研究成果自己為自己辯護，才是最好的解答。在下文中我們將具體的從「系統」出發，提供方法論上一個實質的關鍵環節。

四、作為方法論核心的「使用系統」

從實作的角度去回應對數位人文的質疑，所涉及的面向很多，其中「使用系統」扮演著十分關鍵的角色。多數人對數位人文的印象往往停留在「數位典藏」或史料數位化的階段，這其實是無可厚非的，一來因為數位典藏和數位資料庫，無論量或質，這幾年來有著大幅度的成長，成為大多數人最常接觸、運用的數位人文成果；二來使用資訊和網路科技來保存資料，也最貼近一般人對數位知識所能有的認知，很自然會成為對數位人文的理解，也成為人們對數位人文想像的極限。所以，對於系統的思考，便很自然地從如何保存內容的角度去思考，所收的資料是否正確，相關的紀錄是否皆能收錄，成為人們在評斷一個系統好壞時最主要的著眼點。

內容的考量當然是重要的，這是任何數位人文研究的起點，必須要有正確而完備的數位典藏和數位資料庫作為基礎，資訊和人文研究才能有更進一步合作的可能。可是，這應該是起點而非終點，是部分而非唯一，單從內容保存的角度去思考系統的構成，會形成某種限制，壓縮系統在研究上所能發揮的強大功能。換言之，我們必須跳脫傳統紙本時代保存檔案等相關資料的邏輯，才能開展出在數位時代系統對人文研究所能有的貢獻。

在傳統資料整理過程中，內容的品質和整理分類決定了一切，因為一旦出版成冊便難以再有更改的空間。以這樣的邏輯去建置資料檢索系統，包括目前市面上常見的搜尋引擎、圖書館自動化系統，所重強調內容的求準率（precision）和求全率（recall），以及在檢索搜尋的便利，以此作評斷好壞的標準。然而，這樣的設計背後的假設並未納入文件間的關聯性，也因此才會產生數位化資料會帶來紙本脈絡破壞的印象。

數位化時代的整理邏輯則大不相同，數位化檔案無法單獨存在，必須與系統相結合，才能對檔案進行有組織的觀察與檢索。一個以研究為出發的數位檔案檢索系統其建置的基本邏輯，與前述搜索引擎相反，它預設了所收錄的檔案間必蘊藏著脈絡，而且還是一個開放、具有各種不同連結可能的多元脈絡，檔案檢索系統的任務便是盡量發掘文件間各種關聯，並建立觀察脈絡的環境，讓使用者可以自由地將檢索成果（query return）制定成一個有意義的文件集（sub-collection），並提供各種方法讓使用者去進行檔案文件間脈絡的觀察。數位化後資料以原子式的型態儲存，它便具備不受裝訂、出版所限制的可調整性，以譬喻形容的話，傳統資料庫系統就像

拼圖，研究取向的資料庫系統則像積木：一片片拼圖雖然看似好像可以分開，但一定還是要合在一起才有意義，而其拼湊的方法則只有一種，就是原來設計者所畫出的圖像；積木則大不相同，它可以依照使用者各種不同的想法和需要，而拼造出各式各樣組合。研究為導向的系統，便是能讓使用者可以更自由、更方便去做出各種符合其研究需要的組合。

是以，系統不當被視為檔案數位過程中的「附加物品」，而應當成為數位化作業中最重要的一環；任何一個系統的建置，都應依據檔案內容的特性，量身設計出所需要的系統。而這也才是數位化檔案最核心的方法論，也才能讓數位化檔案的研究潛能真正發揮。一套以研究為主要思考取向的系統，透過檔案檢索的設計，檔案豐富且多重的脈絡將可被發現、被呈現、被觀察。

要打造一個符合人文研究需要的系統平臺，必須從兩個層面加以考量，一個是技術層面，如果沒有足夠的資訊技術再多的理想只是空談，必須依賴資訊人員於技術面不斷進行拓展，以人文研究為導向的資料庫才會變成可能。本書第二部分所收錄的三篇文章，陳詩沛等所撰寫的〈《明清臺灣行政檔案》引用關係之重構〉、王昱鈞和蔡宗翰的〈歷史佛典文獻外來語借詞對辨識系統〉和謝育平的〈同位詞夾子：主題式分類詞庫萃取演算法〉都可以從這個角度加以理解。

陳詩沛等的〈《明清臺灣行政檔案》引用關係之重構〉一文，旨在介紹如何以資訊技術對 THDL 所藏的「明清兩代臺灣行政檔案」進行整理，明清兩代行政檔案是指中央及地方政府在處理政務過程中形成的官方文書，是重要的一手史料，但問題就在於龐大的數量與包羅萬象的內容，再加上存放的零散，造成使用上很大的不便。為了解決這樣的難題，他們試圖利用「引文」的概念，應用資訊科學中的「抄襲偵測」問題（plagiarism detection），意即自動偵測文章是否有抄襲、剽竊的技術對「引文」進行分析，將引用和被引用的文書析理出來，依引用關係將四散的文書脈絡集合起來，並繪製成圖表，希望能為相關研究提供更清楚、便利的理解方式。謝育平的「同位詞夾子」則在替「文字處理」這項數位典藏的最根本前置作業提出解答，他所設計的「同位詞夾子」，是個半自動主題式詞庫萃取演算法，利用中文中同類詞彙所具有的高度同位性，即同樣作為地名的兩個詞彙 A 和 B，理論上會出現在句子中的相同位置；由此開發出由五個部件（前文、前綴、中綴、後綴、後文），描述一個詞彙在文件某處的特徵，用以在文件中萃取該詞彙的同位詞。此外，並同時利用人工和機器，利用人工來保證精準率，利用機器速度來補足召回率，以達到極高的準確率與盡量高的召回率。這項技術在目前實際的使用上，已取得相當的成功。王昱鈞和蔡宗翰則試圖利用資訊科技處理古代歷史文獻中同源詞（cognate）及外來語的借詞（loanword），希望能為語源及跨地域文化交流的研究提供協助。他們設計出了一個包含詞彙抽取、候選詞過濾、語音相似度比對三大模組的系統，試圖