当下中国“数字人文”研究状况及意义

陈 静

(南京大学 高研院数字人文创研中心,江苏 南京 210093)

[摘要] 2009 年以来,“数字人文”作为一个舶来概念进入中国学界。如果说“数字人文”是一套提出、重新定义和回答学术问题的新智能办法,用以回答已经存在的人文议题或提出新的议题,那么在“数字人文”概念被翻译到中国之前,中国早已有了数字人文,主要出现在计算语言学、历史地理信息系统、学术专题数据库、图书馆或者商业主导的数据库/档案库等。当下数字人文研究的意义体现在:为人文研究提供新的数字研究路径与方法,从“基础数据”的层面实现真正的跨学科协同,以“实践性”的方式塑造新一代数字人文学者等。

[关键词] 数字人文;知识生产转型;数字人文学者;实践性

[中图分类号]I0-05 [文献标识码]A [文章编号]1003-4145[2018]07-0059-05

一、中国“数字人文”研究的缘起

作为一个舶来概念,“数字人文”真正进入中国学界,受到广泛关注,是近十年间。2009 年,武汉大学的王晓光教授在“2009 年教育部人文社会科学研究方法创新论坛”上发表了名为《“数字人文”的产生、发展与前沿》的论文。此文随后发表在“科学网”(2009 年 12 月 3 日),获得了上万的在线点击及多次学术引用。同一年,台湾大学举办了第一届“数位典藏和数位人文”会议,提出将数字人文与数位典藏放在同样重要的地位。2011 年以后,越来越多的以“数字人文”为主题的论文出现在学术期刊上。但这并不意味着,在 2009 年之前中国就没有数字人文。我们将“数字人文”定义为一套提出、重新定义和回答学术问题的办法,即人文学者可以利用数字技术(尤其是电脑数据库)作为外在的工具,来回答他们过去已经提出的学术问题,或者是受到数字技术和思维的影响而提出的新课题,甚至产生新的研究范式。①实际上,在“数字人文”概念被翻译到中国之前,中国早已有了数字人文实践。那么,我们在当下强调“数字人文”的意义在哪里?中国的数字人文在已有的学术图景中又遭遇了什么样的挑战?本文以 2009 年作为中国数字人文的一个分水岭,称此前为前史,此后为当下,展开关于中国数字人文当代意义的讨论。

中国第一篇有据可查且被广泛接受的介绍数字人文的文章是由武汉大学的王晓光教授在 2009 年发表的。在台湾地区,2009 年同样是一个起点。据台湾数字人文博士邱伟云观察,“台湾的数字人文学则应以 2009 年台湾大学所举办的第一届数字典藏与数字人文国际研讨会为起点。观察第一届及其之后历届的会议主旨、征稿议题、发表篇目等,可以看见 2009—2012 这四年乃是台湾数字人文学发展的奠基期,这一时期最大的特色,即是有一从数字典藏到数字人文学发展的转向”②。海峡两岸学者在 2009 年的不谋而合,看似偶然,但实则顺势而为。③王晓光教授就提到“中国的与西方的人文研究相比,大陆的人文学者对计算机技术

收稿日期:2018-06-08

作者简介:陈 静(1981—),女,甘肃天水人,文学博士,南京大学高研院数字人文创研中心副教授、硕士生导师,主要研究方向为文化与媒介研究、数字人文。

基金项目:本文系国家社会科学基金重大项目“西方美学经典及其在中国传播接受的比较文献学研究”(项目编号:17ZDA021)的阶段性成果。

①徐力恒、陈静:《“数字人文”浪潮来袭,倡导之余仍要警惕过分乐观》,《社会科学报》2017 年 8 月 26 日。

②邱伟云:《台湾数字人文研究综述(2009—2017)》,即将发表。

③2007 年 6 月,台湾中正大学人文研究中心协同台湾“中研院”人文中心 GIS 专题组和中正大学历史学系举办了地理资讯系统与人文研究研讨会。2009 年 1 月,台湾政治大学文学院身体与文明研究中心、历史学系和地政学系联合召开了 2009 人文地理资讯系统研讨会。

的应用研究并不算落后",但存在不足,研究方法和教学手段较为陈旧,"面对人文社会科学研究方法创新的需要,将国外数字人文研究的内容、方向和前沿集中介绍给国内的人文社会学者以加快我国人文学科研究范式的升级和转型已经显得十分必要"。① 王晓光教授将"数字人文"介绍到中国学界,是出于基于内在需求的自觉和一种对国外数字人文的借鉴意愿,类似的想法也被台湾数字人文先驱、台湾大学资讯工程系特聘教授项洁描述过:

> 1995 年我开始规划并执行台湾大学的台湾史料与藏品的数位典藏工作,这也是在技术上,将新的科技媒体与传统类型史料结合的开始。这项工作进行十年后,我们累积了相当数量的高品质的数位史料,但是我的不安也越来越深。我开始思考,到底如何才能运用资讯科技,在庞大的数位史料基础上从事历史学的学术研究。闭门造车一年多后,才发现在国际已经隐隐约约有一个类似的学问浮现,这就是"数位人文"。近十多年来,我找到了越来越多志同道合的朋友,大家均是被数位人文所隐含的可能性深深吸引,也做了不少相关的研究工作。②

可见,"数字人文"被翻译、介绍到中国,并不能证明数字人文是一个由西方发展起来、被引进到中国的学科;相反,是在一个历史趋势下,从自身的研究需求出发,意识到在数字时代必然而为的一种学术研究转型,而这场转型恰恰呼应了西方语境中"数字人文"浪潮的兴起。过去几年间,数字人文浪潮在中国发展迅猛,台湾连续几年举办"数字典藏和数字人文"会议,大陆有关数字人文的会议越来越多,更有小型研讨会和工作坊,相关论文也在学术期刊和大众媒体上频频发表。数字人文研究在非西方语境中发展的特殊意义、学术价值和面临的挑战也成为学者们越来越关注的问题。如果要展开讨论当下语境中的数字人文,有必要进行一个回顾性说明,以探讨作为一种知识生产转型的数字人文并非仅仅是名称上的创新,更是一种在新的数字语境中的自觉选择。

二、前期中国"数字人文"实践

在"数字人文"这个概念进入中国之前,国家机构、高校、图书馆、研究者及商业公司已经在关注数字转向过程中所带来的知识生产问题。但在学术研究中,数字化资料和数据库依然被认为是一种资料的提供方式而非知识生产本身。

最早以数字方式来处理中文文本的,是计算语言学。中国在 20 世纪下半期开展了相关的研究,例如 1976 年武汉大学语言自动处理研究组利用计算机统计老舍《骆驼祥子》的字频。从 1979 年到 1983 年,有 4 个大型的现代汉语语料库项目在中国大陆发展成型:武汉大学的汉语现代文学作品语料库(1979 年,527 万字)、北京航空航天大学的现代汉语语料库(1983 年,2000 万字)、北京师范大学的中学语文教材语料库(1983 年,106.8 万字)和北京语言学院的现代汉语词频统计语料库(1983 年,182 万字)。这些项目以高校为依托,以现代汉语语料为对象。1991 年,国家语言文字工作委员会启动了国家语料库,推动包括语法、句法、语义和语用在内的现代汉语语法的研究。2003 年,由国家 973 项目经费资助,中国中文信息学会语言资源建设和管理工作委员会发起了"中文语言资源联盟"(Chinese Linguistic Data Consortium,CLDC),推动中文信息处理。③

除计算语言学外,另一个常常与人文研究结合、被认为是"数字人文"的技术和领域是地理信息系统与历史地理信息系统(Historical/ Geographical Information System)。其中可以作为例子的是台湾"中研院"的"中华文明之时空基础架构"(Chinese Civilization in Time and Space,CCTS)和台湾文化历史地图(Taiwan History and Culture in Time and Space,THCTS)④,复旦大学与哈佛大学合作的"禹贡"(CHGIS)⑤,中南民族大学文学与新闻传播学院王兆鹏与"搜韵网"合作的"唐宋文学编年地图平台"⑥。这些项目以地理系统为依

①王晓光:《"数字人文"的产生、发展与前沿》,载《方法创新与哲学社会科学发展》,武汉大学出版社 2010 年版。

②项洁:《一个台湾数位人文学者的贺词》,"零壹 Lab",最后登录时间:2016-10-10。

③"中文语言资源联盟",http://www.chineseldc.org/cldcTest.html,最后登录时间:2018-06-04,22:55。

④中华文明之时空基础架构"(Chinese Civilization in Time and Space,CCTS),http://ccts.sinica.edu.tw/ ;台湾文化历史地图(Taiwan History and Culture in Time and Space,THCTS),http://thcts.sinica.edu.tw/;最后登录时间 2018-06-04, 22:58。

⑤禹贡,http://yugong.fudan.edu.cn/views/chgis_index.php? list=Y&tpid=700,最后登录时间:2018-06-04, 22:59。

⑥唐宋文学编年地图平台,http://sou-yun.com/poetlifemap.html,最后登录时间:2018-06-04, 23:11。

托,人文学家参与其中,试图以地理框架来落实历史文本信息,从而以新的时空观来审视中国历史与文化。

此外,还有一些研究型的学术数据库,提供全文数据库和基本的搜索功能,以便学者能开展相关的研究。比如:北京大学中文系开发的全唐(宋)诗分析系统(the TangSong Poem Project)、先在香港中文大学后迁至台湾政治大学的"中国近现代思想史研究专业数据库(1830—1930)"。它们的出现体现了学者在研究中的需求,也隐含着对当时已有的数据库的一种补充性批判。这种自觉性可以从金观涛、刘青峰两位老师自1997年以来在香港中文大学建立的"中国近现代思想史研究专业数据库(1830—1930)"①及基于该数据库开展的研究中略见一斑。1997年,金、刘尚未接触到数字人文概念,就启动了一个名为"特定现代中文政治概念形式的量化研究"的项目,意图对新文化运动期间最具代表性的12个中文期刊杂志中的文章进行量化统计和分析。在这个过程中,金、刘两位老师意识到现代重要政治观念的研究开展是可以通过对更大范围内的文本进行检索和分析来进行的,由此开展了持续20年的数据库开发和研究工作,在2008年出版了《观念史研究:中国现代重要政治术语的形成》。他们在台湾政治大学开始使用数字人文方法,开展以关键词列句为中心的观念史研究,明确地与"量化历史"划清了关系。②

从1990年代开始,国家各大图书馆,以及一些商业公司开展了大量以数字化为基础的档案库/数据库建设。比如,上海图书馆的晚清期刊全文数据库(1833—1911)和民国时期期刊全文数据库(1911—1949,1—10辑)。它们利用上海图书馆的民国文献资料,建立了两个具有影响力的数据库。资料库建设更多是从图书馆的角度出发,建立数据库,遵循档案原真性原则,呈现给读者的还是以编目为框架的结构化数据呈现。在这个数字化和编目的过程中,文字识别并没有做到全文检索,只是有限地从数字图像中提取了文献信息数据。对于该数据库的用户而言,数据库本身提供的检索能力有限,其最重要的意义在于作为一种可在线浏览的文献呈现方式,使用户得以看到作为证据的文献的存在,而非深入地利用文本进行数据挖掘。这造成了早期图书馆数据库与研究导向的数据库之间的差别。

商业数据库在近20年的发展丰富了数据库的数量和种类。其中堪举为例的是两项中国古籍数字化工程:"四库全书"和"中国基本古籍库"。文渊阁四库全书的电子版由香港迪志文化出版公司推出。在传统中国的大型丛书中,《四库全书》是第一套被数字化的,但就研究者而言,其编辑过程经过审查,内容有删除或者修改的现象发生,这造成了研究者在使用上的障碍。自2001年开始,由北京大学等高校与北京爱如生公司合作建立的"中国基本古籍库",号称囊括上万本中国古籍、超过17亿字的全文。这些大型商业数据库在数字化方面起到了基础性的作用,但因为各自商业利益的需求和数字版权的缺陷,使得商业型数据库存在着发展无规划、内容重复、数据不规范、数据质量参差不齐、文本数据挖掘不够、用户使用体验差的问题。就中国近现代报纸而言,广告基本数字资源的获取并不便利。这一方面是因为中国近现代报纸的数量非常庞大,其保存地也相对比较分散,这就造成了学者在研究的时候获取相应的资源不方便;另一方面是因为,尽管中国及国外很多机构,比如图书馆和一些商业公司对报纸进行了商业化,但这些数据库大部分是收费的。这些数据库的建设主要针对的是报刊上的新闻及评论文章,对广告的内容加工和信息提炼不很充分,大部分都只有广告中的一行字,没有对具有研究价值的图像等作进一步的分析。

三、中国"数字人文"机构建设及研究进展

2011年,武汉大学成立了中国第一家数字人文研究中心。③ 2012年,在台湾大学前图书馆馆长项洁教授的带领下,台湾大学正式成立"数位人文研究中心",并陆续建立了11个数据库,包含超过600万笔元数据、近3000万张影像、近4亿字全文,及数百小时影音资料。④ 台湾大学发起的"数位典藏与数位人文"会议召集亚洲地区乃至全世界对中文数字人文研究感兴趣的学者,每年在台湾相聚,成为亚洲地区最大的数字人文国际会议。此后武汉大学、台湾政治大学文学院、香港公开大学、南京大学等也纷纷成立相关的数字人文研究机构。尽管各个机构有大有小,有实有虚,但从体制上予以数字人文以认可,确是推广数字人文最切实

①中国近现代思想史研究专业数据库(1830—1930),http://www.cuhk.edu.hk/ics/rcccc/database/,最后登录时间:2018-06-04, 23:15。

②金观涛、刘青峰:《就观念史研究再答张仲民先生》,《南方都市报》,http://news.gd.sina.com.cn/news/2010/09/19/1002985.html , 发布时间:2010年9月19日,最后登录时间:2018-06-04, 23:17。

③"武汉大学数字人文研究中心",http://dh.whu.edu.cn/dh/web/index.html,最后登录时间:2018-06-04,23: 19。

④"台湾大学数位人文研究中心",http://digital.ntu.edu.tw/introduction.jsp,最后登录时间:2018-06-04,23: 21。

的举措。

相应地,在近十年间,有关中文文本的数字人文研究项目纷纷凸显出来。比如由哈佛大学、台湾"中研院"和北京大学共同开发的"中国历代人物传记资料库"(CBDB)。这是一个已经运作超过十年的国际合作项目,它的目标在于系统地收录中国历史上所有重要的传记资料,并将数据开放供学术研究之用。截至2016年,它共收录超过37万人的自7至19世纪的传记资料。它的数据既可在线查询,又可以下载,供用户离线使用。研究者可以利用其中提供的大数据,进行相对复杂的查询和分析。除了用作研究历史人物的参考资料之外,还可作统计分析、地理空间分析与社会网络分析之用,为中国史研究引入新视角。从2016年起,这个数据库项目在中国连续举办了不少推广活动,向学界介绍其资料特点和用法。

此外,还有为数不少的研究和电子化项目,许多国家社科基金项目资助学者建设各种专题数据库。然而,不少学者还是觉得无从入手学习数字人文的最新动态,认为各个学术机构还可以投放更多资源,让研究者学会如何在研究中利用新的数字化工具。比如,一般人文学者通常都熟悉在全文数据库进行关键词检索,但对于其他可以用于研究的计算机工具还是很陌生。例如,要把自己搜集到的数据以GIS方法画一张电子地图,就不是很多人能够做到的。所以,推动数字人文的发展,与其停留在讨论数字人文的理念,或介绍众多数据库和电子资源,不如注重实践更有意义,例如培养制作可视化的技能,或传授对数据进行分析、操作、解读等技能。

数据的获取和开放程度也是中国数字人文面临的另一大挑战。以中国古代典籍为例,数学化材料的获得远远不足。各类古籍数据库有许多,但数据共享的做法仍然非常罕见。许多数据库都以商业模式运营,必须得到学术机构和研究者订购,才能生存。这样,它们的数据肯定不会完全开放。这对不同数字资源之间的协作造成一定障碍。对于费用高昂的数据库,不少学校不能负担,也是另一大难题。虽然如此,还是有一些机构希望推动开放数据的做法。例如上海图书馆建立了开放数据平台,以关联数据(linked data)的方式发布一些各个机构、项目都可调用的数据。同时,又创办了应用开发竞赛,开放了其馆藏家谱文献信息和内容信息,鼓励参加者有创意地利用数据,从而发挥资源的最大价值。① 类似活动无疑有利于推广数字项目,让更多人了解数字人文的理念和成果。

四、当下中国"数字人文"研究的意义

伴随数字人文在国内的日益热门,也有不少学者提出疑问:数字图书馆、数字档案馆、数字标准化、计算语言学、GIS、HGIS,这些国内已经有学者做了很多年了,现在专门提"数字人文"有什么意义?"数字人文"强调的是面对尚未完成的数字革命中的知识生产方式转型,推动面向未来的知识体系及方法的建构,其回应的是大数据时代基于学者导向(research oriented)的研究需求与基于资源共享的网络基础设施建设(cyber infrastructure),其建设的是面向数字原生代人类的认知方式系统与路径。

首先,数字人文提供了数字时代的新的研究路径与方法。比如,目前被使用最为广泛的"词频分析"。从技术处理上看,中文与英文的词频统计是同一模式:列出所有文章中出现的词汇,再统计其次数。但进行实际操作时,就有很多不同,英文需要处理同一词汇的语法变形,而中文需要处理"断词",可以运用自然语言处理(Natural Language Processing)和统计学方法进行断词。依据词频统计所做的研究,不仅仅可以做风格研究,而且可以从更大的范围内开展思想史的研究。另外存在一种数字人文研究方法的可能性,是关于系统性发现大量资料内隐含的内部关系的,是比分词更进一步的数据挖掘或者文本挖掘技术。这类技术在商业应用中已经较为多见,比如用以分析顾客的消费行为来进行购买推送。在中文的文本研究中,项洁教授开展的"类书"研究是比较具有代表性的案例。除此之外,数位人文研究中还较为普及的研究就是人际网络研究,前文提到的CBDB近年来基于历史文献数据,开展了大量的社交网络研究。

其次,数字人文从"基础数据"的层面,实现真正的跨学科协同合作,并从方法和路径的层面打通自然科学、应用工程、社会科学、人文科学和艺术的综合研究,使得研究者从自身的学科立场出发,得以扩展到其他领域,并能以"问题导向"出发,与其他学者协同研究,实现研究层面的资源最大共享化、分析方法的最大通约化和知识内容的最大综合化。近年来基于互联网的数字人文社群讨论和传播,显得非常融洽且富有活力。

①"上海图书馆开放数据平台",http://data.library.sh.cn/,最后登录时间:2018-06-04,23:23。

许多关于数字人文的学术交流和讨论已经通过非传统的渠道进行,并受到众多学者的关注,逐渐形成一种跨领域、跨专业、跨地区和跨平台的学术共同体。

第三,数字人文将科学严格的系统性、明晰性和方法的规范性带入人文研究领域。这是在不可逆的数字技术所构成的人文研究的基础条件和环境中所作出的必然回应。数字人文近年来的"数据/算法驱动"尽管存在"技术黑箱化"支配下的盲目乐观/悲观主义,即简单地将数字人文等同于算法或者数据,或者将数字技术的能力夸大到可以迅速地、高效地解决一切人类世界问题;然而,数字技术的高度渗入化和大数据的发展确实已经为人文研究提出了新的挑战,而这需要一种新的知识生产范式的介入。

第四,数字人文以"实践性"的方式塑造了新一代数字人文学者。西方学者拉姆齐(Stephen Ramsay)提出数字人文学者必须具备写代码的能力(即使是在数字人文界,实际上也不是所有人都具备编码能力)。他所提出的广义数字人文实践者的概念,也值得我们借鉴。① 这样,就泛化了"数字"所指涉的范围,使得它不仅包括 XML 、XSLT、GIS、R、CSS 和 C 这样的编程语言,也包括利用软件开展相关研究,甚至开发软件。这就将使用软件来进行研究的学者、以软件来进行知识传播与管理(图书馆员等)以及发明软件的人(工程师)等都纳入了数字人文群体之内,为在更大范围内重新塑造新一代数字人文学者提供了一个很好的参考框架。尤其考虑到西方乃至中国大学近年来高度专业化、体制化和企业化的特点,强调具有"实践性"的数字人文群体有利于扩大学术生产的原动力、提升学术的多样性和促进学术研究的协作性。这也将促进新一代人文社科研究生的培养。我们不能被动地认为数字原生代一定或者自然而然地具备数字思维,事实上,他们也是需要培训和引导的,而这也正是数字人文具有广阔而光明的未来的可能性所在。青年一代将会比我们更加了解未来的数字社会,也更加需要掌握数字知识生产的基本思维、理论反思以及研究方法与工具。

(责任编辑:陆晓芳)

(上接第 58 页)要理解数字人文中国研究的当前状况,关键在于认识到开发数字工具的学者和机构不断增加的开放性。这些工具越来越易于掌握,为学者们提供的结果也越来越令人鼓舞。随着各大学开始提供更广泛的培训,以及更多材料得以数字化,这一潮流还会加速。当下,数字人文研究者在某种程度上仍然隔绝在自己的小天地里,但这些方法将会逐渐成为学者工具箱中的标准配置。当然,并非所有学者都需要在工作中使用量化分析或文本挖掘,但他们需要熟悉这些方法,并能够评价它们,正如他们面对那些更广为接受的方法时一样。中国研究正处于一个激动人心的时刻,而我们将不断从新的方法和模型中了解到更多中国历史和文化中的有趣内容。

四、结语

本文完全着眼于西方数字人文汉学研究中以 20 世纪 20 年代前的材料为对象的部分,主要原因之一在于我本人的专长领域,但更迫切的理由是:数字方法有赖于使用数字化研究材料,而对那些研究更现代材料的学者来说,做到这一点要困难得多。这主要是因为版权上的限制:公版材料的入手更加容易,而要取得 1925 年之后出现的材料的使用许可,难度远大于前者。尽管如此,数字人文现代中国研究领域中仍然出现了一些重要的成果,包括但不限于弗莱堡大学的毛泽东遗产项目正在开展的工作(Daniel Leese、Wang Baigulahu、Amanda Schuman 等)、戴安德(Anatoly Detwyler,关于 20 世纪 20 年代的科学与文学的研究)、苏真(Richard Jean So,现代中国文学)、郭旭光(Arunabh Ghosh,关于中华人民共和国初期的文献计量分析/文本挖掘)以及其他许多人的研究。

(责任编辑:陆晓芳)

①Stephen Ramsay , Geoffrey Rockwell, "Developing Things: Notes toward an Epistemology of Building in the Digital Humanities", *Debates in Digital Humanities*, University Of Minnesota Press, the online access link:http://dhdebates.gc.cuny.edu/debates/text/11, 2012.