历史学堂

HISTORICAL STUDIES WITH BIG DATA

大数据时代的历史研究

舒健 主编

上海译文出版社

HISTORICAL STUDIES WITH BIG DATA

大数据时代的历史研究

舒健 主编

上海译文出版社

图书在版编目(CIP)数据

大数据时代的历史研究/舒健主编. —上海：上海译文出版社，2018.1
(历史学堂)
ISBN 978-7-5327-7623-8

Ⅰ.①大… Ⅱ.①舒… Ⅲ.①史学—文集 Ⅳ.①K0-53

中国版本图书馆 CIP 数据核字(2017)第 206631 号

本书由上海文化发展基金会图书出版专项基金资助出版

图字：09-2014-119 号

大数据时代的历史研究
主编/舒　健
责任编辑/钟　瑾　装帧设计/半和创意

上海世纪出版股份有限公司
译文出版社出版
网址：www.yiwen.com.cn
上海世纪出版股份有限公司发行中心发行
200001 上海福建中路 193 号 www.ewen.co
上海信老印刷厂印刷

开本 890×1240　1/32　印张 14　插页 2　字数 270,000
2018 年 1 月第 1 版　2018 年 1 月第 1 次印刷

ISBN 978-7-5327-7623-8/K·258
定价：68.00 元

拥抱“大数据”对历史研究的挑战

（代序）

一

梁启超先生在谈到史料对于史学发展的重要性时，把史料喻为“史之组织细胞”，认为“史料不具或不确，则无复史之可言”。傅斯年先生也强调，若要不断地使史学发展、层累，就必须不断地收集资料、挖掘资料。近代以来，研究者一直孜孜不倦地扩充史料来源，扩大研究范围，“上穷碧落下黄泉，动手动脚找材料”即体现了对获取研究史料的重视。随着科技的进步，尤其大数据时代的来临，史料的获取、存储、交流、再生产等诸多环节发生深刻变化，一个突出的表现就是数据库的蓬勃发展。英国著名学者魏根深（Endymion Wilkinson）在其著作《中国历史研究手册》（以下简称《手册》）的弁言中介绍了《手册》的四个重要目标，其中之一就是“突出利用电子资源对中国历史资料进行传播、归类和分析，这些资料上起商代甲骨文下到最近的中外学术研究”，还提到《手册》选择了225个数据库（此外还有数以百计的原始资料及参考著作的电子版），并强调“中国历史和考古研究领域在过去的十年中飞速变化，数据化亦改变了人们研究的方法，本书试图跟上这些新的发展”。

就国内而言，2010年以来，仅国家社科基金支持的、以数据库建设为核心的文史研究项目就不下70项，这两年来增长尤甚。资料的数字化改变了历史研究的资料来源，数字资源的采集、加工和处理对研究成果的取

得作用日益显著，如何理解历史学在大数据时代下的发展就成了一个重要的课题。

首先，我们来回顾一下国内史学类数据库的发展历程。尽管数据库技术滥觞于20世纪60年代末，但是直到80年代，国内的人文学科才逐渐开展数据库的建设工作。就史学类数据库的建设来看，迄今大体可以分为三个阶段：

第一阶段主要为数字化制作、整理阶段，重点在于资料的输入与整理，制作成电子光盘。如在1985年，台湾“中研院”历史语言研究所便启动了“汉籍电子文献资料库”的建设工作，内容包括“二十五史”“十三经”以及超过2 000万字的台湾史料、1 000万字的大正藏、道藏、清代经世文编等大型类书、丛书，收入典籍达460多种，计4亿多字。此阶段使用的范围更多限于电脑单机，改变了知识的存储形式，体现了携带方便、易于查询等优势，完成了从旧时的汗牛充栋到如今大容量可移动介质的重大转变。

第二阶段始于20世纪90年代中后期，在互联网兴起后，各类数据库的建设和上线层出不穷。如中国知网、超星图书馆的上线为论文和图书的查找提供了巨大的便利。就专业的历史类数据库的开发而言，尤为突出的是北京爱如生公司研发制作的“中国基本古籍库”，该库分4个子库、20个大类、100个细目，精选了先秦至民国的历代重要典籍，总计收书1万余种，全文超过17亿字。其他一些主要的全文数据库，如书同文古籍数据库、中华经典古籍库也广受欢迎。全文数据库比较直观，就像把一本纸书变成电子文本，然后我们在这里面抓取所需文本，其中原理跟搜索引擎颇为类似，而且也不需要什么培训就可以上手使用。此外还有一些专题类型的数据库，如中国金石总录数据库、东方杂志全文数据库等也已陆续上线。除了相对商业化的文献数字化机构，还有各公共图书馆、高等院校等

都有为数不少的古典文献数据库，如中国国家图书馆的电子文献也比较丰富，有数字善本、甲骨世界、碑帖菁华、敦煌遗珍、西夏文碎金等经典文库，另有其他各色常见的古典文献。

第三阶段主要为关系性数据库的出现和建设。关系性数据库注重利用数据库内在内容再发现并分析文本间的关系，帮助研究者理解人物、文献的脉络，这是近年来人文研究类数据库的发展趋向。目前在史学界有较大影响力的是哈佛大学燕京学社、台湾“中研院”史语所、北京大学中古史研究中心合作的“中国历代人物传记资料库”（China Biographical Database Project，简称 CBDB）。CBDB 早在 20 世纪 80 年代由美国宾州大学郝若贝教授（Robert Hartwell）开始建立，1996 年郝教授去世，将其捐赠给哈佛大学燕京学社。2005 年，上述三家单位开始共同开发，截至 2016 年 5 月数据库共收录约 370 726 人的传记资料。CBDB 相较于一些企业开发的全文数据库来说，在数据结构上更加复杂、精细，可以进行地理空间、社会网络分析、群体划分、统计分析等研究，但需要多种软件的结合使用（如 QGIS、PAJEK 等），甚至需要进行专门培训。由项洁教授领导、台湾大学数位人文研究中心开发的台湾数字历史图书馆（简称 THDL）也是此方面的杰出代表。[①] THDL 中提供词频分析、上下手契关联分析、人物相关性分析等不同功能，还可以部分地实现契约空间分布的展示。THDL 提出了数据库建设的理念，即数据库的主要功能是为研究者

① 项洁、陈诗沛、杜协昌：《台湾古契书全文资料库的建置》，第三届台湾古文书与历史研究学术研讨会，逢甲大学历史与文物管理研究所，2009 年 3 月 14 日，第 1—19 页。该数据库主要收录台湾地区契约文书及台湾总督府档案，台湾大学数字人文研究中心并不拥有这些资料的版权，而是以授权复制或者录入为电子文本的形式建立这一数据库的知识产权。

提供研究环境并帮助研究者发现问题，而非仅仅是储存与检索。[①] 另外，上海交通大学建设的《中国地方历史文献数据库》基于针对性设计的元数据结构提供交叉导航、数据统计等多种功能，这些功能不仅可以帮助研究者找到自己所需的文献，更可能帮助研究者发现新的研究议题。

可以预见的是，未来的数据库肯定会朝着开放性、共享性、可视化方向去发展。大规模的数据化历史资料平台建设，为整个学术界提供了更丰富灵活而有效的研究资源，而且随着海量的图书、报纸、期刊、照片、绘本、乐曲、视频等人文资料的数字化，更多的专题性数据库或以独立的形式或以合作的方式在如火如荼的建设中，诸如华东师范大学沈志华教授主导的关于冷战档案数据库的建设、上海大学陶飞亚教授主导的关于汉语基督教书目文献数据库的建设都各有特色，值得关注。它们不仅是历史研究的重要基础，也越来越被其他相关学科所看重。

二

各种类型数据库的广泛使用，极大突破了研究者获取资料的局限性，带来了研究上的一些显著变化。[②] 仅从材料的获取角度而言，各类数据库

---

① 涂丰恩、杜协昌、陈诗沛、何浩洋、项洁：《当资讯科技遇到史料——台湾历史数位图书馆中的未解问题》，载项洁：《数位人文研究的新视野：基础与想象》，台湾大学出版中心 2011 年版，第 21—44 页；项洁、翁稷安：《数位人文和历史研究》，载《数位人文在历史学研究的应用》，台湾大学出版中心 2011 年版，第 11—20 页。

② 关于数据库的使用对历史研究的影响从 2011 年曹树基就开始涉及，见曹树基《数据库：历史研究的社会科学化》一文，其中讨论了数据库使历史研究社会科学化的问题，参见《中国社会科学学术前沿（2010—2011）》，社会科学文献出版社 2011 年版，第 171—177 页。另梁晨、董浩、李中清的《量化数据库与历史研究》一文也涉及了量化数据库对历史研究的改变，参见《历史研究》2015 年第 2 期。

的使用令各种类型史料得以进入研究视野，接近“竭泽而渔”的地步，同时也扩展了史学研究的材料范围，方便了多时段、多角度的研究。研究人员足不出户就可以查到各地的藏书状况及学习资源，国内外开放的网络数据资源使知识获取更为便捷迅速，这在十多年前几乎是无法想象的。

尽管各类数据库的建设和使用如火如荼，但就目前而言，国内文史类相关数据库仍然存在诸多不足。首先，从材料的把握度上看，以“大数据”为代表的数据资源来源更加广泛，信息质量参差不齐。文献数字化也是各自为政，由于版权及产权的原因，数字化文献分散于不同公司、不同研究机构中，而且重复建设的现象严重，不但功能单一，数据往往只是某个类别、某一专题，数据分散以致难以实现多元化及整体化的研究对比与分析。如《申报》电子版，至少存在湖南青苹果数据公司版、北京爱如生公司版和台湾中国近代报刊数据库等三种检索系统。数字化古籍文献整合势在必行，打破数据库建设“小、散、乱”[①]、各自为政的局面，已经形成学术界共识。其次，少数民族古籍数字化过程中也存在着诸多问题。我国历史为多民族共同造就，除了汉语材料之外，还保留了众多少数民族古籍，少数民族古籍的多样性和模式化与数字化过程之间存在矛盾，数字化过程的安全问题、字库不健全问题、著录困难等解决起来难度也颇大。另外，中文光学字符识别软件（Optical Character Recognition，OCR）的开发不够，中文识别软件的开发相对还是较为滞后，亟须技术创新和突破；国内古籍数据与境外汉籍数据库的打通融合；同时当前史学类数据库的建设存在重复化和个人市场开发不充分、学商合作不充分等一些问题，也亟待改进创新。

① 杜晓勤：《国学大数据时代来了》，《光明日报》2013年9月16日第15版。

尽管存在不足，但数据库建设对于人文学科愈益重要，数据、文献等信息加工方式的改变使得人文学科的建构知识流程与研究方式、方法也出现重大改变，可以说数据库的使用和发展正深刻地改变着史学研究。①

第一，体现在历史研究的内容。由于海量的资料涌现使得史学研究如同技术领域一样，需要不断地推进数据分析的能力。在诸多人文学科中，这种能力对注重文献的历史学研究而言尤为重要。面对大数据，研究者在处理分析数据的过程中将会充分发挥以往的研究模式与个性化、学术性的经验优势，以创造出不同于以往的认知对象的方式。这将大大拓展人对于数据材料的应用水平，为创新提供更多的可能。

第二，从研究的主体形式来看，之前的史学研究更多是个人的工作，个体的活动和兴趣爱好的体现。而数据库的建构以及对庞大数据的处理，仅仅靠一个人很难完成，需要数名研究者形成合力才能做到，历史研究中的搭建史学学科团队现象日益突出。在香港科技大学李中清团队参与研究的南京大学梁晨就认为："鉴于我们研究组的自身经验，组成由不同学科背景的研究成员密切互动的专项研究团队，实现跨学科合作，是一条切实可行和较有成效的道路。"② 这也是目前国外学者应对大数据变化的方式——注重团队工作。或许是传统思维的限制，国内人文学科的学术团队建设较弱，但也已开始迎头赶上，部分高校建立了学术研究中心，如北京大学成立量化历史研究所，以面对大数据所带来的史学新革命。

---

① 当然大数据的研究并非局限在数据库上面，数据库的使用仅仅是运用工具的一种变革，更多的改变将会体现在研究思路、方法乃至研究范式的变化上，因行文关系，在此重点谈论了数据库的建设。

② 梁晨、董浩：《必要与如何：基于历史资料的量化数据库构建与分析——以大学生学籍卡片资料为中心的讨论》，《社会》2015 年第 2 期。

第三，从历史研究的方法来看，量化史学的研究方兴未艾、持续发酵。定量方法的使用使得历史学的研究成果增加了定量的特征，增强了人文科学中的科学属性，这不仅使传统定性研究中的模糊认识得到量化验证而更趋于严谨和精确，而且还可以获得仅靠定性分析难以达到的认识，从而有效弥补了中国传统历史研究中定性分析不足的缺憾。运用计量史学方法还可以帮助研究者揭示历史规律，发现并预测历史趋势，明确强化史学研究的价值。

第四，从历史学与其他学科的关系来看，由于研究内容、方法所带来的变化，各种人文社会科学的互相依赖和渗透趋势将日益强化，历史学不断汲取政治学、经济学、社会学等学科的相关知识和方法，其研究观念亦渗透进其他学科之中，将会有更多的交叉学科和边缘学科出现，学科之间的边界有可能日益模糊，同时也将为专业研究者提供更为广阔的空间。如最早提出“e考据”概念的台湾清华大学的黄一农教授的研究，黄教授以“e考据”的学术方法和学术态度不断开拓研究领域，从天文研究跨越到科学史、中西文明交流史乃至红学领地。

第五，从历史书写的完成和传播来看，借助于各类数据库以及网络提供的材料，一批非职业历史学家异军突起，彻底打破了原有的史学生态，导致史学话语权向大众转移。无数个“当年明月”凭借着数字技术的便利，似乎正在发挥着比职业历史学家更大的影响力。[①] 随着资料获取的日益快捷，可以预见历史的书写、传播方式将会更加多元化，出现更多的争议。

---

① 朱绍杰：《专家解读国际历史科学大会：史学研究大变革》，http://culture.china.com/11170621/20150920/20429509_1.html。

由于各类数据库的广泛使用，丰富的信息量似乎大大减低了历史研究存储资料的麻烦，但也导致相关从业者要花费大量精力用来搜索、整理有效信息。计算机技术的进步和跨学科交流的频繁，使更多资料被数字化，并呈现在互联网上供研究者存取利用，但文献材料呈几何级增长，以往的阅读习惯、研究分析方法立刻显示出了不足，促使研究者必须充分借鉴、吸收数据科学领域在获取、分析数据方面的成果，以更好地推进学术研究视域中相关材料的把握能力。

由此，知识结构和数据处理能力变得十分重要，对于历史研究者的逻辑思维和数理分析能力提出了全新的要求，传统的学科思维受到挑战，计算机分析处理能力延伸为人文社会科学研究者科研素养的重要组成部分。随着“数字鸿沟”的不断拉大，广大学者除了要秉承传统的为学精神，还要不断学习新知识的获取及交流方法，使自身的学术研究不断适应时代的发展要求。

但是我们也要看到，如果过分依赖数据库或数据分析会导致碎片化、片面化，将不利于全面、完整、系统地了解历史。将作为工具的“数据库”与传统的史学训练结合起来，虽然能减少翻阅故纸堆的麻烦，但不应忽视传统史料在整全性、连续性、批判性方面的优点，切勿出现“史料的尽量扩充与不读二十四史”的局面。此外，理解历史还需要理论的指导，需要借助于抽象思维和理性判断。价值判断是中国史学的优良传统，否则就无法“以史为鉴”。历史学的首要任务是探求历史的真相，史料考证和文本诠释都十分重要。作实证分析时，要把所论事物或行为置于其所处的具体历史背景下；作规范分析时，则可用今天的价值观评论其当时的得失，以及它们对后人的潜在效应，并揭示其历史局限性，这仅仅依靠书库是做不到的，也是不能替代的。

国际公共史学会主席、意大利欧洲大学研究院教授瑟奇·诺里特(Serge Noiret)在第22届国际历史科学大会上指出:“数字历史,不仅带来新资料、新工具,也将产生新的历史学家”,为此,他强调史学家的责任在于“抵制由于新媒体广泛传播导致历史知识的稀释以及扭曲,最好的办法就是历史学家自己学会运用这些新技术,以新的方式触摸过去,感知现在,了解未来”。历史学者的价值和作用恰恰在这方面能够实现,相信新时代的历史研究将会出现变革和新局面。

三

为了迎接和面对大数据时代对历史研究的挑战,讨论史学发展之可能性,2015年12月,在上海大学中国史高原学科的资助下,《中国史研究》杂志社与上海大学历史系联合主办的“传承与开启:大数据时代下的历史研究”国际学术研讨会在上海召开。来自中国大陆、中国香港以及日本、韩国等国家和地区的48所院校及研究机构、期刊社的110余位专家学者齐聚一堂,围绕“数据库的建设与使用”“大数据与史学研究”“大数据在历史研究中的技术问题”“大数据时代下的人文关怀”等议题展开了深入的探讨。整体而言,本次会议有三个主要特点:一是跨学科研究现象明显,大数据时代的历史研究与其他学科发生不可避免的联系,打破了学科间的壁垒,拓展了研究领域;历史学科内部的藩篱也在逐步破除,大数据研究下长时段的视野为此提供了研究的工具和手段,同时也对研究者提出了更多的要求。二是重视对话和交流,与会学者在分组发言与评议中,就研究视角、研究方法、史料搜集和运用、学术观点等进行坦诚交流,不同论点的交流使本次会议更具价值。三是参与度高,知名学者、大量青年教师、硕博研究生都对本次会议表示了极大的热情,表明大数据时代的到来

对学界尤其是青年学者产生了深刻的影响，显示出此领域的蓬勃生机，新的研究范式将会在大数据时代获得进展。

本书就是此次会议论文合集，需要强调的是，本次会议得到学界的很多关注。香港科技大学李中清教授团队以及山西大学行龙教授团队、四川大学徐君团队都出席参与本次会议，对大数据结合历史的研究做了有典范的讲演，浙江大学陈红民教授由于时间冲突，最后未能成行，甚为遗憾。由于篇幅、主题以及版权并尊重各位参会学者的意愿，本书并没有收录全部参会的论文，不足之处，恳请大家的谅解和包容。此外，本书还收录了徐力恒博士 2016 年 5 月在上海大学的讲演内容，在录音整理的基础上，徐博士又重新校对、修订，定名为《数字人文时代的关系型数据库：中国历代人物传记资料库（CBDB）的应用》，在推动数字人文的研究方面，徐博士团队的付出有目共睹。

在本次会议筹备的前期和举办过程中，得到了学界的广泛关注和帮助。《中国史研究》杂志社的主编彭卫研究员发来致辞，并在后期的出版中给予了很大帮助；张彤、陈奕玲、苏辉等学者在百忙中专程赴会，参与了方方面面的工作，付出了辛苦和努力。中国社会科学院的刘正寅、乌云高娃、陈晓伟等学者也提出了很多建议，给予了很多的帮助。上海大学发展规划处、上海大学文学院以及历史系的领导和同仁们也提供了诸多指导和帮助，如学科带头人陶飞亚教授、文学院院长张勇安教授、发展规划处田立君教授；历史系宁镇疆主任、陈勇、郭红、肖清和、杨雄威、杨卫华、朱虹等，由衷感谢，此处不再一一罗列。总之，作为国内第一本关于大数据与史学研究的论文集，本书凝结了众多的期待与厚望。

令人欣喜的是，本次会议之后，北京、天津等地高校和研究机构相继举办了以大数据与历史研究为主题的学术研讨会，相关议题的重要性不断

得以彰显。大数据与历史的结合将会产生怎样的局面，在史学研究乃至课程设计上又会发生怎样的变革。我们这届学人何其有幸，能见证大数据时代的到来，成为历史研究的参与者和变革者。希与诸君共勉，一起迎接挑战！

上海大学历史系　舒健

2017 年 5 月

目录

第一部分　大数据与史学研究

第二部分　数据库介绍和建设

第三部分　大数据与历史研究结合案例

第四部分　存在的问题和新的领域