·讨论与评议·

量化数据库与历史研究*

梁 晨 董 浩 李中清

摘 要： 历史学在今日仍被视为传统学科之一，但非历史学者借助大规模历史量化数据库进行自然和社会科学研究并取得重要成果的现象却有不断增长之势。不同于传统计量史学和以经济学为代表的社会科学学科所注重的“解释型学术”，量化数据库推动的是“求是型学术”，即通过统计分析从大规模系统数据中挖掘新事实、产生新认识。中国历史中长期存在规模宏大的户籍登记、土地分配和科考记录等系统材料以及海量文字记录，很多已得到收集和整理，有利于构建大规模数据库并开展定量研究。这种新的研究范式不仅有助于历史学科自身发展，更能够促进跨学科、跨国界的学术交流与融合，并为全面深入认识中国社会历史特征、平衡东西方学术发展作出贡献。

关键词： 长时段 大规模量化数据 定量分析 史学发展

一、前言：量化数据库研究的兴起与研究范式调整

自然科学和人文社会科学学术传统在16世纪开始的“科学革命”（Scientific Revolution）后出现“大分流”：自然科学追求对“未知”事物的发现和认识，人文社会科学注重对“已知”现象的理解和解释。根据美国教育学家欧内斯特·博耶（Ernest L. Boyer）对学术类型的分类，前者更符合“求是型学术”（Scholarship of Discovery），后者偏重于“解释型学术”（Scholarship of Integration [or Interpretation]）。①

然而，20世纪90年代以来，一种将大批量历史材料数据库化，并依靠定量分析揭示其中隐含的史实、检验和发展历史认识与经验的新方法逐渐在国际学术界流行起来，很多成果与著作产生重大学术和社会影响。例如，美国家谱学会保存的历史长时段人口资料对20世纪90年代遗

* 本文受国家社科基金青年项目（项目号10-CZS-023）、香港研究资助局GRF项目（640613，16400114和16400714）资助。承蒙外审专家提出宝贵修改意见，谨致谢忱！

① Ernest L. Boyer, *Scholarship Reconsidered: Priorities of the Professoriate*, Princeton, N. J.: The Carnegie Foundation for the Advancement of Teaching, 1990. “求是型学术”和“解释型学术”为作者根据英文定义进行的意译。

传学家和医学家在乳腺癌基因遗传方面的重大研究突破有很大推动作用。① 法国经济学家托马斯·皮凯蒂（Thomas Piketty）《二十一世纪资本论》（*Capital in the Twenty-First Century*）一书依据多国 20 世纪国民账户、收入、财产与纳税等多种系统历史数据，研究资本主义社会不平等的长期演化。该书曾高居英文畅销书排行榜首并引发热烈的学术讨论。② 美国经济学协会主席克劳蒂亚·高丁（Claudia Goldin）与《经济学季刊》（*Quarterly Journal of Economics*）主编、前美国劳工部首席经济学家劳伦斯·凯兹（Lawrence Katz）合著的《教育与技术的竞赛》（*The Race between Education and Technology*）基于近一个世纪以来美国教育、职业和收入的个人层面微观数据讨论美国经济不平等的历史脉络和技术革新、教育进步对收入分配结构的影响。③ 这些成就在推动全球学术进步的同时，也彰显了大规模量化数据库在结合历史资料系统研究人类社会长期变化与延续等大问题上的作用。由此可见，人文社会科学研究同样可以贡献“求是型学术”。

量化数据库研究是统指各种搜寻能够涵盖一定地域范围、具有一定时间跨度的整体性大规模个人或其他微观层面信息的系统（一手）资料，并将这些资料按照一定数据格式进行电子化，构建成适用于统计分析软件的量化数据库并进行定量研究的方法。量化数据库研究多以“大数据”为基础，关注材料的系统性和可量化数据平台的构建，重视对长时段、大规模记录中的各种人口和社会行为进行统计描述及彼此间相互关联的分析，以此揭示隐藏在“大人口”（Big Population）中的历史过程与规律。相较于传统定量研究，这种方法对数理统计分析技术要求不高，很多时候只需要描述性统计分析和比较研究即可，大大方便了对复杂定量分析方法认识有限的普通学者对数据的理解和运用。同时，“大数据”本身往往涵盖相对完整的“大人口”的多种“长时期”的具体信息，这不仅使学者可以依托数据库理解个体是如何被宏观社会环境所规范和影响的，也可以理解这些微观、个体行为又如何集合起来塑造和改变宏观社会进程，甚至可以用来检验“中观”社会群体，如家族、邻里社区等在宏观环境和微观行为互动下的特点和作用。传统定量研究则倾向于将研究问题和具体操作抽象化，倚重数理模型和逻辑推导，强调复杂统计方法的运用，但往往不够重视实证材料或数据本身。因此，量化数据库研究是一种更为基础和宽泛的研究思路和方法，它既能够丰富、完善我们对微观人类历史和行为的认识，还有助于构建更为可靠的宏大叙事，促进我们对人类社会发展规律的进一步认识。

20 世纪 90 年代中期以后，利用历史资料进行量化数据库构建与研究逐渐成为国际学术界关注的一股“热潮”。以当前国际上最有影响的五大历史量化数据库为例，建成初期学界对它们的利用、研究很有限，但进入 20 世纪 90 年代，尤其是 1995 年以后，情况发生巨大转变（参见下图）。2006—2010 年的五年间，索引五个数据库④的新增学术发表成果已达 2360 余篇。尽管这

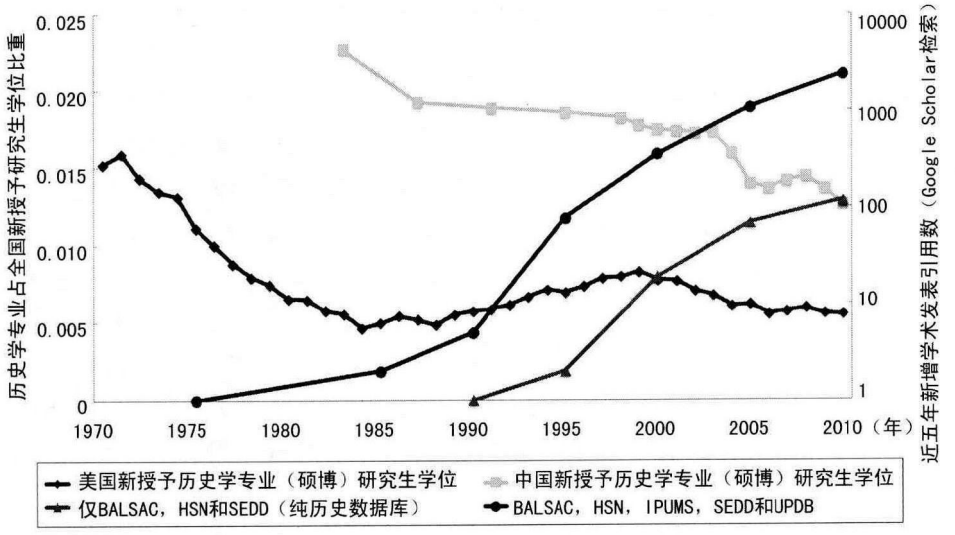
① Kevin Davies and Michael White, *Breakthrough: The Race to Find the Breast Cancer Gene*, New York: John Wiley & Sons Inc., 1995.

② Thomas Piketty, *Capital in the Twenty-First Century*, trans. Arthur Goldhammer, Cambridge: Belknap Press of Harvard University Press, 2014. 图书所获奖项详见 http://www.hup.harvard.edu/catalog.php?isbn=9780674430006.

③ Claudia Goldin and Lawrence Katz, *The Race between Education and Technology*, Cambridge: Belknap Press of Harvard University Press, 2010.

④ 这五大数据库分别是美国整合公共微观数据库（Integrated Public Use Microdata Series，简称 IPUMS）、加拿大巴尔扎克人口数据库（BALSAC Population Database，简称 BALSAC）、荷兰历史人口样本数据库（Historical Sample of the Netherlands，简称 HSN）、瑞典斯堪尼亚经济人口数据库（Scanian Economic Demographic Database，简称 SEDD）和美国犹他人口数据库（Utah Population Database，简称 UPDB）。

些学术成果中有很大一部分来自 IPUMS 所包含的当代人口统计数据，但如果只统计三个纯历史微观数据库，即 BALSAC、HSN 和 SEDD，其贡献的新增学术成果在 2006—2010 年五年间也达到 117 篇，且近 20 年的增速与五个数据库发表成果数量增长趋势几乎一致。需要注意的是，下图中的右侧纵坐标已经过指数转换，而非传统线性坐标，可见这些大规模量化数据库对学术研究的贡献是呈几何级数增长的。



中美两国应届历史学专业研究生（硕士及博士）学位比重与主要量化数据库研究引用量变迁图

数据来源：Google Scholar 搜索结果（截至 2014 年 1 月）；美国国家教育统计中心《教育统计简报 2012》（*Digest of Education Statistics 2012*）表 310 和 364，http://nces.ed.gov/programs/digest/ 2012menu _ tables.asp；《中国教育年鉴》（1982—1984、1988、1992、1996 及 1998—2012 年），北京：人民教育出版社（分别出版于 1985、1989、1993、1997 及 1999—2013 年）。

值得注意的是，目前历史学界似乎对量化数据库这种科学化的方法在研究和学科建设上所具有的重要价值缺乏认识。积极参与到这股“热潮”中来，将历史量化数据库作为新的工具与资料开展研究的更多是社会科学和自然科学学者。历史量化数据库日益凸显的学术价值与历史学者及整个学科在这一新兴学术浪潮中远非充分参与，形成一种对比强烈的“内冷外热”现象。

一方面，重视构建大规模史料数据库已经成为国内外史学界共识，国内一些学术单位与学者也尝试建立数个重要的大型电子史料库，但史学家对数据库建设的认知大多还停留在资料永久保存阶段，以文献和数据检索为主要目的。数据库的可量化研究以及由此带来的方法转变还未曾涉及，这也造成一些大规模历史数据库被构建却无法得到很好利用的尴尬。① 量化数据库是分析大规模史料的重要方法之一，能为历史学者解决这种“尴尬”提供帮助。

另一方面，由于教学和研究方法与技术发展潮流脱节，传统史学越来越难以吸引年轻人，已经是一个持续多年且国际化的状况。以哈佛大学为例，50 年来以人文学科为主业的学生比例

① 2013 年 8 月，教育部社会科学委员会历史学学部在哈尔滨召开年度工作会议，主要讨论历史数据库的构建与研究问题。参见教育部社会科学委员会历史学学部：《2013 年度工作会议纪要》，2013 年 10 月，第 2—9 页。

已从近40%下降到20%，全美这一比例则从14%下降到7%。① 作为历史学者的最直接来源，如图所示，20世纪70年代以来，中美两国应届历史学专业研究生比重呈现明显下降趋势，这间接反映了历史课程与研究方法的“落伍”和不够实用。

本文着重通过介绍国际学界历史大数据库构建与研究的状况，分析探讨这种新研究方法对历史研究的价值，历史学者面临的挑战与诸多自身比较优势，以及开展相关教学的必要性。

二、国际大型量化历史数据库的创建与运用

量化数据库研究方法是近60年来计算技术发展的结果。第二次世界大战后，计算机逐渐深入到人类生活的各个方面，海量电子原始数据（Raw Data）也逐渐在全球积累起来，人类正步入“大数据”时代。② 这些数据是分析、研究人类多种行为的重要依据。但面对庞杂的信息数据，不仅人眼（人脑）无能为力，甚至早期的统计分析工具也无法胜任，③ 这促使统计分析技术向“数据挖掘”方向发展。利用“数据挖掘”，研究者可以对复杂的“大数据”进行定量分析，从中有效挖掘隐藏的现象与规律，总结经验模式。④ 20世纪中期以后，计算机辅助下的定量分析逐渐成为国际学术研究中的一股新风潮，许多大型量化数据库得以构建并服务于学界。⑤ 1962年，以美国密歇根大学为基地成立的跨大学政治和社会校际联合数据库（Inter-university Consortium for Political and Social Research，简称ICPSR），联合全世界600多个成员机构，存储超过17000多种调查资料，包括全球各地各种社会调查，其中美国自身资料包括各州选举投票资料、军队名册、遗嘱、遗嘱查验与税收记录和美国联邦调查局案卷资料等，是目前世界上最大的社会科学数据中心之一，对经济学、政治学、社会学、人口学以及法学等学科研究具有重要价值。⑥ 1968年，德意志联邦共和国成立了德国社会科学信息中心，负责收集各种社会科学信息并建设成专业数据库，供学术界以及公众使用。目前，该机构的数据库涉及社会学、心理学、人口学、政治学、历史学和经济学等多个领域，是欧洲最重要的数据库之一。⑦

定量研究的风气也渗透进历史研究。20世纪60年代起计量史学在欧美学界曾盛行一时。⑧

---

① Russell A. Berman, “Humanist: Heal Thyself,” *The Chronicle of Higher Education*, June 10, 2013, http://chronicle.com/blogs/conversation/2013/06/10/humanist-heal-thyself/.

② Jonathan Shaw, “Why ‘Big Data’ is a Big Deal: Information Sciences Promises to Change the World,” *Harvard Magazine*, no.3, 2014, pp. 30-35, 74-75.

③ William J. Frawley, Gregory Piatetsky-Shapiro and Christopher J. Matheus, “Knowledge Discovery in Databases: An Overview,” *AI Magazine*, vol. 13, no. 3, 1992, pp.57-70.

④ Onno Boonstra, Leen Breure and Peter Doorn, *Past, Present and Future of Historical Information Science*, Helsinki: Edita, 2006, p.57.

⑤ 计量方法在历史研究中运用的发展历程，参见孙圣民：《历史计量学五十年——经济学和史学范式的冲突、融合与发展》，《中国社会科学》2009年第4期。

⑥ ICPSR的详细介绍，参见其官方网站，http://www.icpsr.umich.edu.

⑦ 德国社会信息中心的情况，参见蔡莹：《德国社会信息中心数据库建设经验》，《中国信息界》2005年第5期；国际社会科学界对数据库的使用与研究，参见埃尔温·K. 朔伊希：《社会科学数据服务的历史与观念》，王星译，《国际社会科学杂志（中文版）》2004年第3期。

⑧ 美国新社会史代表人物史蒂芬·塞恩斯特罗姆（Stephen Thernstrom）的《另一些波士顿人》以及新经济史代表人物福格尔和恩格尔曼的经典著作《十字架上的岁月：美国黑人奴隶制经济学》都是这方面的代表，而法国史学界则集中将计算机计算技术运用于历史人口学研究。参见Onno Boonstra, Leen Breure and Peter Doorn, *Past, Present and Future of Historical Information Science*, p.20.

20世纪80年代中期起，英、美两国先后成立国际性历史与计算学会。① 一些国际学者开始尝试建设小型个人历史专题数据库。到20世纪90年代，一些重要的大型、超大型量化史学数据库构建成功并对学界开放使用，引起国际学界极大关注。美国明尼苏达大学人口中心创建的整合公共微观数据库是其中最有影响者之一。微观数据（Microdata）指社会抽样调查和人口普查等包含的个人层面（Individual-level）信息，如性别、年龄、婚姻、家庭状况、职业和出身等。微观数据在世界各国广泛存在，内容和形式高度一致，适合连缀成超大数据库，进行国际比较，是新社会史和经济史研究的关键。历史学家是除统计机构外较早使用微观数据进行研究的学者。② 1998年起，IPUMS首席科学家罗伯特·麦凯（Robert McCaa）先后说服100多个国家的统计机构与IPUMS合作，将各自数据库连接起来并免费用于学术研究。目前，IPUMS包含19世纪以来多个国家（包括中国第三次［1982］、第四次［1990］人口普查）的微观数据。现在，IPUMS数据还在不断增长，其中最显著的是从18世纪开始到20世纪中期，以数字抄本为基础的历史人口微观数据。到2018年，IPUMS数据记录总量将扩展到20亿人次。③

IPUMS等基于人口普查的微观数据库尽管具有地理、人口覆盖面广的优势，但这类数据往往只能反映一个或几个时点的横截面（Cross-Sectional）微观数据，无法对个人进行跨时点的连续追踪。与之相对的是基于历史户籍、族谱、教会记录等长时段连续记录构建的长时段纵贯（Longitudinal）数据。虽然这些数据库往往只能覆盖一些地区，但可以连续观察这些地区居住人口几十甚至几百年的行为活动，成为研究历史社会发展与个人行为互动的重要材料。目前全球有多个公开或半公开的大型历史微观数据库，除上图涉及的5个数据库外，还有加拿大历史人口计划数据库（Le Programme de Recherche en Démographie Historique，简称PRDH）、瑞典乌米亚人口数据库（Umea Demographic Database，简称DDB）等。它们普遍涵盖一个或几个地区个人层面的大规模人口信息，除人口事件外，往往还包含一定的个人或家庭层面的社会经济信息。

由于这类西方长时段微观历史数据库的材料来源和结构具有一定的相似性，本文着重介绍四个纳入上图计算的数据库。BALSAC涵盖自17世纪以来第一代欧洲定居者至当代的加拿大魁北克地区约500万人口，通过对该地区以婚姻证书为主的人口事件记录进行电子化转换，并对个人记录进行人际、代际连接，重建亲属网络和谱系。④ HSN是一个对荷兰全国人口具有统计意义的代表性人口样本数据库。研究者通过对全国人口出生记录进行随机抽样，将1812至1922年间在荷兰出生的78000人纳入其中，并尝试查找和连接相关个人婚姻证书和死亡证书。另外，除了这些“分散”的人口事件记录之外，近年来荷兰研究者还尝试录入人口户籍资料信息，以

① 1986年国际历史与计算学会（AHC）成立，1989年牛津大学出版社发起创办《历史与计算》杂志（*History and Computing*，参见 Robert J. Morris, “*History and Computing* a New Magazine,” *Historical Social Research*, vol. 15, no. 1, pp. 118-120）。随后美国也成立历史与计算学会，并于1996年出版同名期刊（http://quod.lib.umich.edu/j/jahc/）。这些学会的重要宗旨是让历史与计算科学更好、更有效地结合，以推动历史教学、研究和历史作品的写作。

② 20世纪30年代后期，弗兰克·劳伦斯·奥斯利（Frank Lawrence Owsley）使用电子计算技术分析人口普查数据以研究南北战争前美国南方的社会结构。

③ IPUMS及其所含历史微观数据的介绍，参见 Patricia Kelly Hall, Robert McCaa and Gunnar Thorvaldsen, eds., *Handbook of International Historical Microdata for Population Research*, Twin Cities: Minnesota Population Center, 2000; Steven Ruggles, “Big Microdata for Population Research,” *Demography*, vol. 51, 2014, pp. 287-297.

④ 加拿大巴尔扎克人口数据库的信息，可访问 http://balsac.uqac.ca/english/.

期较“连续”地观察每个人的生命历程。① SEDD资料来源相较BALSAC和HSN更为复杂。数据库不但包含9个教区的人口出生、婚姻和死亡证书，用于构建个人层面的人口事件和家庭关系，还与相应教区的“人头税”册以及教会问答测试记录连接，记录较为详细的动态家庭构成和社会经济信息。其公开数据记录了1829年至1968年间5个教区共108000人，非公开部分连续记录至2011年。② UPDB数据来源是这几个数据中最为丰富的，包含的信息也最为全面。与其他人口数据库类似，生育、婚姻、死亡等人口事件证书和由摩门教会记录的家族谱系资料是UPDB的基础。研究者将数据库与美国人口普查数据、犹他州选民登记资料、犹他州驾照信息、医院出诊和手术记录、癌症登记资料、离婚登记、社会保险死亡登记等多种资料连接，形成一个包含11代730万人共1900万条记录的大型微观数据库。③ 这些数据库不仅有助于学者深入理解大规模人口历史，亦可为更广义的社会科学以及遗传学、医学和其他自然科学研究作出巨大贡献，为开发珍贵历史资料的巨大潜能创造条件。

最近十几年里，一些侧重东亚研究的历史学者和研究团体意识到，西方学术界构建量化历史数据库使用的相关人口、社区原始材料，在中国或东亚地区很早便广泛存在。一些东亚地区的历史人口微观数据库由此陆续得以建立。如美国学者李中清（James Z. Lee）、康文林（Cameron Campbell）从20世纪80年代起，花费20多年时间，建立基于八旗户口册和清代皇室族谱资料的中国多代人口系列数据库（China Multi-Generational Panel Data Series，简称CMGPD）。④ 该系列数据库包含辽宁、双城和皇族三个子数据库，其中前两个已经在ICPSR网站上对全球学界免费公开。辽宁数据库涵盖1749—1909年间辽东地区26万人的150余万条记录。双城数据库涵盖1866—1913年间黑龙江双城县10万人的130余万条记录，并尝试与不同时段的家户地亩资料相连接。这些大规模、长时段的微观历史数据包含丰富的人口和社会经济信息，具有时间上的深度和空间上的广度，对人口统计学、家与家族、亲属关系、社会分层与流动、卫生健康等多个学术研究领域有重要价值。⑤ 东亚其他国家和地区也存在大量类似的户口册资料，如日本的户口册、韩国的户籍大帐、台湾日据时期户籍资料等。目前这些相应数据库都在各国研究者的努力以及国际学界的合作下不断完善，对整个东亚历史人口和社会变迁的研究产生积极影响，更为与西方社会进行长时段和微观层面的比较研究提供更多便利。⑥

大规模历史微观数据库不仅成为历史研究的重要基础，也越来越为其他相关学科学者所看重，成为学术研究的重要推动力。利用Google Scholar的搜索功能，统计五大国际微观量化数据库和中国多代人口数据库—辽宁部分在不同学科领域内研究使用情况，可以看出，所有数据库在社会学、经济学、人口学、政治学、生物学和健康学等非历史学领域都有普遍运用。

---

① 荷兰历史人口样本数据库的信息，可访问 http://www.iisg.nl/hsn/.

② 有关瑞典斯堪尼亚经济人口数据库的信息，可访问 http://www.ed.lu.se/databases/sedd.

③ 有关美国犹他人口数据库的信息，可访问 www.hci.utah.edu/groups/ppr/.

④ 有关中国多代人口数据库的介绍、使用指南、数据下载等，可访问 http://www.icpsr.umich.edu/icpsrweb/ICPSR/series/265.

⑤ James Z. Lee, Cameron Campbell and Shuang Chen, *China Multi-Generational Panel Dataset, Liaoning (CMGPD-LN) 1749-1909: User Guide*, Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2010; Hongbo Wang et al., *China Multi-Generational Panel Dataset, Shuangcheng (CMGPD-SC) 1866-1913: User Guide*, Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2013.

⑥ 有关这些东亚微观历史人口数据库的介绍和对比研究，参见 Hao Dong et al., “New Sources for Comparative Social Science: Historical Population Panel Data from East Asia,” *Demography* (forthcoming).

各数据库分学科引用情况表（Google Scholar 检索，截至 2014 年 1 月）①

	IPUMS	HSN	SEDD	CMGPD	UPDB	BALSAC
总　数	4680	180	69	13	1460	53
历史学	3520	172	66	13	1300	49
社会学	2100	102	25	8	91	21
经济学	4370	161	58	12	303	26
人口学	3580	158	50	11	659	41
政治学	429	7	1	2	17	4
生物学	890	62	24	4	812	32
健康学	2960	109	48	8	1310	35

总的来说，大规模量化历史数据库的建立为整个学界提供了更为丰富、灵活而有效的研究资源。IPUMS 的专家们宣称，对于他们的数据库，研究者只需要一个稳定的网络链接和一台个人电脑就可以在任意地点展开研究。②对于历史学家来说，构建大规模历史数据库并采用定量研究方法，必然促使他们从传统的文献解释研究模式向信息数据收集、数据挖掘、数据库建设和记录分析与写作模式转变。这种研究方式的大转变意义重大。

三、量化历史数据库的学术价值

量化历史数据库是计算技术和结构数据的组合，这种源于方法的更新带来史学研究形式的改进。巴勒克拉夫认为 20 世纪后半叶以来，历史学向其他社会科学寻求新方法，突破旧史观和研究框架已成必然。这使得社会科学在两个层面上促进了历史学的新发展：一是应用更加广泛的社会科学思维范畴；二是新的定量方法的运用。这些影响最终推动了历史研究重心的大转移，从特殊转向一般，从个别事件转向普遍过程，从叙事研究模式转向结构分析模式。③ 量化数据库研究模式及其在历史学中的应用，在有效扩展研究史料范围和转变研究驱动模式的同时，帮助历史学家从更普遍、更基层的角度发现隐藏的史实与规律，开展更具普遍意义的国际比较，实

① 1. 数据依据 Google Scholar 搜索引擎，通过使用“　”功能搜索包含各数据库全称的公开学术发表。例如，搜索“Integrated Public Use Microdata Series”以获得 IPUMS 相关条目数目。2. 分学科引用情况数据由搜索数据库全称加学科名获得。如搜索“Integrated Public Use Microdata Series” history 以获得引用 IPUMS 数据库的学术发表成果中与历史学相关的条目。这种简洁的搜索方式，可以比较不同学科的引用分布并在数据库之间进行比较。但不可避免的是，存在很小一部分搜索结果包含学科名关键字但并非该学科相关文献。我们假定这些“噪音”检索条目在搜索不同学科关键字时出现的概率相当，所以对不同学科的引用分布趋势没有很大影响。3. 需要注意的是，一篇学术文献往往涉及多个领域。同一文献可在多个领域搜索，表格内的数据并没有相互排他性。4. 为统一检索标准和结果，表格中 CMGPD 相关数据仅包含 2010 年在 ICPSR 网站使用“China Multi-Generational Panel Data-Liaoning”名称公开数据后的引用文献，这些只占基于辽宁多代人口数据库发表的学术研究成果很小一部分。CMGPD 的另一个子数据库双城多代人口数据库（China Multi-Generational Panel Data-Shuangcheng）已于 2014 年在 ICPSR 网站免费公开，但公开时间较短，没有纳入表格。

② Patricia Kelly Hall, Robert McCaa and Gunnar Thorvaldsen, “Introduction: International Historical Microdata—A New Resource for Research and Planning,” in Patricia Kelly Hall, Robert McCaa and Gunnar Thorvaldsen, eds., *Handbook of International Historical Microdata for Population Research*, pp. 3-10.

③ 杰弗里·巴勒克拉夫：《当代史学主要趋势》，杨豫译，上海：上海译文出版社，1987 年，第 74、77 页。

现史学研究的突破与转变。

（一）扩大史料范围和史学研究基础

量化数据库方法对分析大规模的系统性、连续性历史材料非常有效，在扩展几乎所有史学研究门类材料范围的同时，为克服史料繁芜提供重要思路，可使史学家们从传统的“选精”与“集粹”研究方法中超脱出来，对“大人口”、“长时段”的整体史学研究颇有帮助。大规模、同类型的群体性记录广泛存在于历史材料中，但过去往往被历史学家忽略或无法运用。历史因果关系是“多种事物互为因果的复杂关系”，普通逻辑思维本就“难济事”，① 个人层面的微观材料又繁芜庞杂，如果没有科学的研究方法几乎无法开展系统性整体研究，史料的学术价值无法被开发。就国际史学来说，警察记录、税收记录、人口调查资料、遗嘱、市政委员会的诉讼、教堂登记、新闻报纸、选举统计和人口史领域内的家族谱与人口登记材料等均是在量化历史数据库建立后，首次为学者系统利用，对研究选举与政治史、商业周期史以及历史人口学和家族等重大问题起到重要作用。② 近些年笔者一直从事的中国大学生学籍卡数据库构建及研究亦如此。学籍卡是近现代教育史上常见材料，但在建立量化数据库之前，一直不能为中国史学界深入利用，通常只为佐证某位名人的求学经历等。学籍卡数据库的建立，为分析整个大学生群体的社会来源与社会流动等学术问题提供了可能。③

中国历史文献中类似材料非常丰富，量化数据库的构建与研究可以有效发挥它们的学术价值。笔者认为传统官方文献中至少有三类非常适合数据库化和定量研究。第一，历代户籍材料。中国约自战国时期已有户籍制度，秦统一后逐渐形成规模，经三国及南北朝时期的调整，隋唐后已经非常完备。这个过程中，户籍登记材料得以累积并逐渐完善。明代以来的黄册更是蔚为壮观。这些连续长达两千多年的户口材料是数据库与量化研究的绝好素材。第二，与户籍材料相伴随同样历史悠久的土地及财产占有与分配登记材料。西周以来，中国土地赋役制度经历井田制、屯田制、均田制、府兵制、均税法、一条鞭法、摊丁入亩到现代土地改革多次调整，但每个新制度下，对土地数量丈量、归属权益的明确以及相关赋税情况都有庞大记录，这些重要材料非常适合数据库化研究。第三，自隋唐以来，考试（考核）就成为中国社会选拔精英人才的重要方式，历代皆有数量惊人的科考或官员铨选材料，这些材料历时长，系统化程度高，是不可多得的量化数据库素材。李中清—康文林研究组目前正在收集此类教育与官员考核史料，以构建两个新的全国性量化历史数据库。一个是由李中清、梁晨负责的民国大学生学籍信息数据库。目前已收集、输入近10万民国大学生的学籍信息，对民国大学生社会来源问题研究很有帮助。另一个是康文林负责的清代《缙绅录》资料的数据库化和量化分析。目前已确定《缙绅录》涉及50万官员超过260万条个人记录，对系统分析清代官员人际网络和职位波动意义重大。以上三方面材料及相关研究问题是中国历史研究的重点与热点。以这些材料来构建大型量化历史数据库，对重大历史选题研究的突破和为教师提供教学素材等都非常有利。此外，民间材料

---

① 吴承明：《经济史理论与实证：吴承明文集》，刘兰兮整理，杭州：浙江大学出版社，2012年，第339页。

② Heinrich Best, “Technology or Methodology? Quantitative Historical Research in Germany,” *Computer and Humanities*, vol.25, no.2/3, 1991, pp.163-171.

③ 相关成果参见梁晨、李中清等：《无声的革命：北京大学与苏州大学学生社会来源研究（1952—2002）》，《中国社会科学》2012年第1期；梁晨、张浩、李中清等：《无声的革命：北京大学、苏州大学学生社会来源研究（1949—2002）》，北京：三联书店，2013年。

如家族谱、商业机构账册等都是适于构建量化数据库的重要资料。

大数据库通常涵盖某一范围内所有参与者或构成者的状况，量化方法又能系统、细致地研究或描述不同规模群体的多种信息。一些多变量分析方法还能同时比较多个因素与结果的相互关系，或者在考虑到结果与其他因素相互关联的情况下，估计特定因素与结果的统计相关性。这不仅能帮助研究者更深入理解各种因素的变化与彼此关系，还能兼顾所有个体的影响与权重，很大程度上避免选择资料时的疏漏与偏废，可弥补史学家惯用的“选精”与“集粹”研究方法之缺陷。李伯重认为，“选精”与“集粹”研究方法，前者是“从有关材料中选取一两种据信是最重要或最有‘代表性’的，以此为据来概括全面”；后者是在研究“一个较长时期或一个较大地区中的重大历史现象时，将与此有关的各种史料尽量搜寻出来加以取舍，从中挑选出若干最重要或最有代表性的，集中到一起，合成一个全面性证据，以求勾画历史现象的全貌”。他认为这两种方法是史学家惯常使用的，但在本质上并无大异，其特点都是通过从史料中选取具有代表性的例证推导出结论。研究者常将“某一或某些例证所反映的现象普遍化”，从而可能丧失真实性，导致研究结果具有一定不可靠性，需要加以改进。① 量化数据库研究在处理海量材料上的能力，在一定程度上可以解决史学研究中“史料不可能竭泽而渔，甚至难以把握边际”的困难，避免“或分门别类，缩短战线；或随意比附，看朱成碧”的研究困境。②

（二）善于发现隐藏史实，改进研究驱动模式

量化研究的一个重要优势是，能够发现靠传统文献阅读无法发现的隐藏在历史资料堆中的史实。与传统计量史学强调高级统计方法和侧重因果推论不同，大规模微观量化数据库最根本的优势在于，可以帮助研究者通过简单的统计分析，得出基础的描述性统计结果、发现不同因素间基本的相互关系。当然，通过巧妙的研究设计和逻辑推导，研究者同样可以基于这些“大数据”得出因果推论。同时，这种基于量化数据库的基础性统计描述有很大兼容性，对系统的、可归类的文字信息也有很好的分析能力，这使得它为帮助史学家发现传统文献阅读无法察觉到的历史因素的相关性，对话已有的学术理论和观点以及进行更为复杂的研究分析提供了基础。

以大规模数据为基础的量化研究还能较好纠正研究的主观性，实现研究从常见的理论或问题驱动转向数据或经验驱动，在发现真实历史的同时修正过去的讹误。③ 我们强调的构建量化数据库应是一个整理、开发历史材料的客观过程。设计合理的量化数据库，应首先完整体现史料所记录的原始信息而非研究者对数据的主观判断和改动，将主观处理和理解数据库的操作空间留给数据库使用者，并在设计上为数据库使用者依据当前数据生成需要的新变量提供可能。不同于以存档和检索为目的的文史资料数据库，构建量化数据库的初衷在于方便研究者直接对数据进行量化分析。量化数据库所包含的信息既可以直接来自史料中的数字记录，如人口、税赋、田亩等，也可以对文本资料如传记、名录等进行系统编码。量化数据库的形式往往是一个或多个数据表格。虽然具体数据结构可能因材料类型、内容及其他特点而不同，但绝大部分数据表格的设计都是每列包含一个因素或类别（即变量）的信息，每行代表一个最低层次的研究个体，例如某人、户等。对这些数据库的量化分析，实际上就是通过统计软件对大规模研究个体（行）

① 李伯重：《理论、方法、发展趋势：中国经济史研究新探》，北京：清华大学出版社，2002年，第110—121页。

② 桑兵：《晚近史的史料边际与史学的整体性——兼论相关史料的编辑出版》，《历史研究》2008年第4期。

③ Heinrich Best, “Technology or Methodology? Quantitative Historical Research in Germany,” pp.163-171.

在不同因素（列）之间相互关系的统计分析。这种量化分析本身并不依赖任何单个或正反方事件参与者的叙述，相对简单、明确的数量关系也可以避免表达上的主观与刻意，可以更好避免研究者基于自身认知或经验所形成的预设观点对研究结论的影响，减少研究时的“先入为主”，并方便重复验证，得出的规律性现象或结论也更可靠。因此，只有保证数据库最大程度地反映历史材料的原有信息，才可能保证分析结果的客观性和完整性。

构建量化数据库尤其是大规模微观量化数据库能够更好地推动历史研究，其原因主要在于，相比传统史学的“选精”、“集粹”或是新经济史研究中常见的集合数据（Aggregate Data），量化数据库能够提供更加全面、系统和微观的信息。在获取历史事实方面，仅仅依靠对各变量最简单的描述性统计，如频数、平均值、标准差、比例、列连表等，量化数据库就可以提供许多集合数据或个案研究无法提供的信息和视角。这一点在结合历史数据往往覆盖长时段的特性时作用尤为明显。更重要的是，这些简单的统计描述仅反映数据库的客观情况，不同的数据库使用者可以如自然科学研究一样重复检验。目前许多定量历史研究在解释历史现象方面依靠自然实验（Natural Experiment）的设定，借鉴现代计量经济学如工具变量（Instrumental Variable）、双重差分法（Difference in Difference）、断点回归（Regression Discontinuity）、倾向值匹配（Propensity Score Matching）等前沿技术进行因果推断。① 这些统计方法的运用往往需要对变量之间的关系和分布引入许多较强的假设。尽管很多假设并非直接针对数据的局限性，但是有限的数据往往限制了研究者对假设合理性的检验，从而增加了研究结论受研究者选择方法的主观性影响的风险。而大规模量化数据库的构建，无疑可以提供更多微观、有效的信息，提高统计功效，为大规模历史数据与复杂统计模型分析的结合提供条件。

（三）为国际比较和多学科合作创造精准的数据基础

国际比较和多学科合作研究是很多大规模历史研究项目追求的重要目标之一，但很多历史现象本身蕴涵复杂的信息，直接进行跨国、跨文化比较研究困难重重。李伯重指出，比较是史学（特别是现代史学）的基础，但在进行比较研究时需认真考虑研究对象是否具有可比较性，有无一致的时空范围以及是否具备合适的比较标准，而数据的优点使得量化比较成为比较史学中最成功的部分。② 量化数据库则致力于从文献史料中系统抽取相对明确的信息，如传记履历或人口调查材料中的生卒年份、职业身份、财富收入等，通过编码处理供计算机分析。这些信息基本都有普遍通用的含义，适合进行跨地区、跨国界比较。加之定量方法的客观性与可比性，这削弱了材料与结论的意识形态和国别属性，可以保证国际比较的事实基础可靠，比较的结果相对准确，容易形成共识。

由多国社会科学家组织开展的欧亚人口和家庭史项目（Eurasian Population and Family History Project，简称EAP），可视为近些年来利用大规模微观数据研究进行国际比较的范例。

---

① 有关计量分析与历史相关研究结合部分具体实例的中文讨论，参见孙圣民关于Robert Fogel，Douglass North和Daron Acemoglu等人研究的介绍。（孙圣民：《历史计量学五十年——经济学和史学范式的冲突、融合与发展》，《中国社会科学》2009年第4期）近年来基于微观历史数据定量研究发表论文众多但较分散，很多都出现在经济学、政治学、社会学、历史学等学科国际主流甚至顶级期刊。读者亦可通过Claude Diebolt对近年来发表在计量史学专门期刊 *Cliometrica* 上论文的总结，集中获取一些研究实例。（Claude Diebolt, “Where Are We Now in Cliometrics?” *Historical Social Research*, vol. 37, no. 4 (142), 2012, pp. 309-326）

② 李伯重：《量化史学中的比较研究》，http://history.orientalpatek.com/newsdetail.aspx?id=955.

1994年以来，来自欧洲、美国、日本和中国等国家的人口史学家、社会学家和经济学家等，将比利时、中国、意大利、瑞典和日本等国100个村庄的个人层面微观数据分别建成量化数据库，通过构建标准统一、便于比较的统计模型，分析死亡、出生、婚姻等多种人口行为与社会环境的互动。该系列研究成果集中关注中西方社会人类行为的空间差异及其与人类意志的关系，挑战了现有传统研究下的某些经典宏大理论，对人口学、历史学及社会学等学科发展具有重要意义。①

量化历史数据库的构建还能为社会科学家提供强大且灵活的研究资源，推动历史学与其他社会科学的优势整合。传统史学研究，一般通过文献资料收集和学者个人化解读，梳理出历史史实。这些史实通常可以成为其他学科的认识基础与判断依据，但由于研究方式不一致，大部分描述性的历史材料或结论，较难直接应用于其他学科相对系统和量化的研究方法，进而阻碍历史学与其他学科的融合。通过对史料整合，构建量化数据库，历史材料转变成可直接适用于定量分析研究的数据，成为历史学和其他学科均可直接分析利用的一手材料，而不是难以“消化”的描述性史实，为实现历史学科研究方法的多样化，以及与其他社会科学甚至自然科学的交流互动提供基础，提升历史学的学术价值。

此外，量化历史数据库大多向学界开放，这意味着会有更多不同学科的研究者和研究方法投入历史材料的解读和分析中，丰富历史研究的形式和成果，形成与传统历史学家“单打独斗”不同的局面。如中国多代人口数据库—辽宁部分，早期主要由研究组成员及其合作者使用，但随着数据材料于2010年对学界在线开放，已有越来越多数据计划成员（李中清—康文林研究组）以外的学者开始使用这些数据进行研究，成果数量甚至超过了李中清—康文林研究组本身。这只是各大公开量化数据库对学界贡献的一个缩影。量化数据库的构建与公开，对学术发展的贡献显然已超出单个学者或研究组的能力范围。

四、量化数据库运用中的挑战与机遇

构建数据库和量化分析的研究方法是史学研究“现代化”、“科学化”的一个重要途径，但现代方法与古老学科的结合可能存在很多困境与挑战。巴勒克拉夫认为，历史学者往往有一种“心理障碍”，对新研究方法不容易接受。② 还有学者认为，相较于其他学科，计算机技术在历史研究中的应用状况并不好，这不是因为计算机技术不够发达，而是历史学家没有学会使用新技

① 该计划的重大研究成果已由美国麻省理工学院（MIT）出版社集结成三册系列专著出版，分别是 Tommy Bengtsson et al., *Life under Pressure: Mortality and Living Standards in Europe and Asia, 1700-1900*, Cambridge: MIT Press, 2004; Noriko Tsuya et al., *Prudence and Pressure: Reproduction and Human Agency in Europe and Asia, 1700-1900*, Cambridge: MIT Press 2010; Christer Lundh et al., *Similarity in Difference: Marriage in Europe and Asia, 1700-1900*, Cambridge: MIT Press, 2014. 有关国际学界对前两本著作学术贡献的评价，参见 Ronald Lee and Richard H. Steckel, “Life under Pressure: An Appreciation and Appraisal (Review of *Life under Pressure: Mortality and Living Standards in Europe and Asia, 1700-1900*, by Tommy Bengtsson et al.),” *Historical Methods*, vol.39, no.4, 2006, pp.171-176; Jack A. Goldstone, “Prudence and Pressure: Everywhere (Review of *Prudence and Pressure: Reproduction and Human Agency in Europe and Asia, 1700-1900*, by Noriko Tsuya et al.),” *Historical Methods*, vol.44, no.4, 2011, pp.181-184.

② 杰弗里·巴勒克拉夫：《当代史学主要趋势》，第333页。

术，甚至对新方法抱有很强的偏见和反感。① 因此，即便定量分析明显为“描写大人口群的历史提供了巨大机会”，但“很多从事流行文化与社会研究的历史学家对这一方法却非常消极”。② 这些论断未必完全正确，但历史学家运用社会科学化的研究方法确实存在很多挑战。

在中国，这种挑战似乎更严峻。欧美不少大学早已设立专门机构，研究计算机技术在人文学中的应用，并依托其来培训学生，打造未来学者。英国普利茅斯大学以第二次世界大战后的英国选举数据为例，训练历史专业的学生掌握基础数据分析方法；③ 牛津大学在很多人文领域进行计算机研究和拓展，包括人文计算中心、牛津文献档案馆、人文中心等；格拉斯哥大学有人文高级技术与信息中心，鼓励使用信息技术开展人文研究和教学。美国几十所大学设有专门机构研究量化史学方法。普林斯顿大学、罗格斯大学以及弗吉尼亚大学设立人文电子文献中心。加拿大多伦多大学等也有类似机构。④ 国内在此方面却几乎还是一片空白。钱学森先生30年前就呼吁中国史学界要加强对学生的计算机技术教育，开设系统科学、电子计算机和高等数学等课程，培养定量研究专门人才，以能运用系统的科学方法进行历史研究。⑤ 但这一倡议并未引起重视。而自中学就实行的文理分科和历史等传统文科对科学方法教育、应用的漠视使未来史学家在掌握新方法上遇到更多困难。最近几年，一些外国学者开始尝试以暑期班的形式在国内大学开设量化史学课程，如2011年起，康文林、李中清在上海交通大学开设“中国多代人口数据库暑期学校”，陈志武于2013年起在清华大学组织“量化历史研究班”。⑥ 他们希望将大数据量化分析方法教授给国内史学新秀。但在授课过程中，主讲教师意识到国内年轻史学工作者对史料有较深的理解和掌握，但对基本统计学理论与方法则比较陌生，使得教学效果大打折扣。

当下的多数历史学者，不仅在掌握量化数据研究技术与方法上存在很大困难，接受与适应这种社会科学化的研究组织与管理模式也不容易。构建量化数据库通常包括数据采集、数据分类、数据编码、数据存储、数据信息挖掘和定量分析等多个环节，数据库建成后还可能需要数据管理和维护等多种工作。相对于以数据为中心的“科学化”、“电子化”研究方式，传统史学研究显得有些手工艺式的陈旧。⑦ 历史学者从事研究时多是“单打独斗”，而构建大规模、量化史学数据库并对其进行分析与研究，通常需要构建起一个研究团队，由多学科专业人员合作参与。这种社会科学化的研究，无论对经费还是组织管理都有相当要求。另外，量化历史数据库要发挥更大学术价值，开放是非常关键的。李中清—康文林研究组的中国多代人口数据库、包

① Onno Boonstra, Leen Breure and Peter Doorn, *Past, Present and Future of Historical Information Science*, p.16.

② Pat Hudson, *History by Numbers: An Introduction to Quantitative Approaches*, London: Arnold, 2000, p.7.

③ Paul Lambe, “An Introduction to Quantitative Research Methods in History,” *History and Computing*, vol.6, no.2, 2003, http://hdl.handle.net/2027/spo.3310410.0006.205.

④ Onno Boonstra, Leen Breure and Peter Doorn, *Past, Present and Future of Historical Information Science*, p.10.

⑤ 钱学森、沈大德、吴廷嘉：《用系统科学方法使历史科学定量化》，《历史研究》1986年第4期。

⑥ “中国多代人口数据暑期学校”主要教学内容，参见上海交通大学中国东北历史与社会研究中心官网相关介绍，http://ishnc.sjtu.edu.cn//do/list.php?fid=51；“量化历史研究班”基本情况，参见清华大学社会科学学院官网关于该班2014年第二届招生宣传海报，http://www.tsinghua.edu.cn/publish/sss/8028/20140408080501602515533/222.pdf.

⑦ Heinrich Best, “Technology or Methodology? Quantitative Historical Research in Germany,” pp.163-171.

弼德领导下的哈佛大学中国历史人物传记数据库（China Biographical Database Project，简称CBDB）① 以及耶鲁大学自1949年开始建立的人类关系地域文件库（Human Relations Area Files，简称HRAF）② 等都实现了在线公开，方便全球学者利用，这与过往那种将珍贵史料“敝帚自珍”的做法也完全不同。

不过，面对“大数据”时代量化研究的大趋势，历史学者绝非“赤手空拳”，只能消极被动接受转变，而是有其独到的比较优势。历史学者掌握的众多史料、丰富的历史知识以及考据等研究方法等对量化研究历史资料来说都是必须的。量化数据库方法要在历史乃至社会科学研究领域发挥更大价值，历史学者的作用不可或缺。实际上，尽管使用数据进行分析的多为非历史学者，但前文介绍的社会科学最重要的公开数据整合中心之一ICPSR和几个重要量化历史数据库IPUMS、HSN、SEDD、CMGPD的项目发起人或领导者都获得历史学博士学位。这说明历史学者不仅能够参与，而且对于这些国际主要量化数据库项目的成功有不可替代的作用。③

历史学者在研究中认识和处理史料的宝贵经验，是构建量化数据库和进行后续分析不可或缺的。量化数据库的构建与研究，必须依托于统一的制度性定义或标准化结构的信息材料才能完成，但历史材料的丰富性、多样性和复杂性成为构建系统、直观的量化数据库的障碍。如郭松义曾分别研究清代在京山西商人和司法审判中私通行为，前者有136宗样本，后者有403宗案例，从社会科学角度来看体量并不大，但这些个案分散在“乾隆朝刑科体本”、“宗人府来文”、“内务府来文”、“八旗都统衙门档案”、“刑法部档案”以及《刑案汇览》、《刑案汇览续编》等众多官方文献以及如《资政新书》、《樊山政书》等海量个人文献中。对缺乏史学训练和长期研究积累的非历史学者来说，了解这些史料并找到所需研究个案是非常困难的。④ 又如，前文提到的SEDD数据库，整合了出生、婚姻、死亡三类登记册以及税册和教会考试册等多种材料。如果没有对各种历史材料的深入理解，很可能会在整合不同资料构建量化数据库的过程中产生种种问题。构建量化数据库并非单纯将历史材料“电子化”，而是需要凭借历史学者对原始材料的深刻理解，创造性地进行归类和整合。

在选定和整理好原始史料后，对史料中包含的具体信息的分类与编码依然需要依赖历史学者的专业知识。由于时代背景错综复杂、史料建立和涵盖的时间长短不一，史料记录的内容可能不一致或不完整，且难以今日的常识直接理解。此外，单一历史材料也可能包含多种类型和层次的信息，比如人口户籍材料不仅包含个人信息，也可能包含家户成员关系和土地、财产构成等信息，往往需要细致、全面地理解和辨别。因此，将史料记载的复杂信息灵活妥当地分类并设计变量编码方式并非简单依靠电脑技术或其他模版即可完成。这种史料的复杂性一直是历史研究的难点，同时也是历史学者学习、训练和研究的重点。历史学者对史料文献中各种信息的真伪与具体含义的把握优于一般的非历史学者。历史学者在长期训练、研究中积累起来的专

① 参见该数据库官网：http://isites.harvard.edu/icb/icb.do?keyword=k35201&pageid=icb.page145374.

② 详情参见官网：http://www.yale.edu/hraf/.

③ ICPSR中心主任乔治·奥特（George Alter）、EAP计划领导者及CMGPD数据项目发起人李中清、HSN项目及欧洲历史人口数据网络（European Historical Population Sample Network）计划领导者凯斯·曼得麦克斯（Kees Mandermakers）、IPUMS项目主任斯蒂芬·鲁格斯（Steven Ruggles）、EAP领导者及SEDD项目主要负责人汤米·本特森（Tommy Bengtsson）均为历史学（或经济史学）博士。

④ 相关研究及材料状况分别参见郭松义：《清代北京的山西商人——根据136宗个人样本所作的分析》，《中国经济史研究》2008年第1期；《清代403宗民刑案例中的私通行为考察》，《历史研究》2000年第3期。

业历史知识是构建、研究量化历史数据库工作必不可少，甚至是至关重要的保证。

历史学家对当时社会、人文环境的认识比较全面和深入，在对各种历史文本信息的解读和对分析结果诠释方面，往往也拥有比较优势。随着数据挖掘和信息处理技术的发展，尤其是文本挖掘（Text Mining）技术的进步，使得分析人物传记、文学手稿等结构较为复杂的文本材料成为可能。例如哈佛大学学者让-贝普提斯特·米歇尔（Jean-Baptiste Michel）和耶瑞兹·列博曼·艾登（Erez Lieberman Aiden）领导的研究组于2011年在美国《科学》（*Science*）杂志上发表《基于百万电子化图书对文化的定量分析》（Quantitative Analysis of Culture Using Millions of Digitized Books），通过“谷歌图书计划”（Google Books Project）得以构建涵盖人类印刷出版物4%的电子文字数据库，并量化分析1800—2000年间英文语言应用所反映的文化趋势。通过词典编纂、语法进化、集体记忆、技术传播、名誉获得、文字审查、历史传染病学等方面的分析实例，他们很好地展示了如何通过构建大规模量化数据和应用简单统计方法（仅仅是对特定词语的出现频次统计）认识历史和发现新知。①

历史学家对具体史实的细致把握，往往能够弥补大规模定量分析注重整体推论但部分忽视或无法理解具体演变过程和机制的弊端。量化历史数据库的应用并非鼓励纯粹定量分析，而是需要传统史学方法与定量方法的互补。在量化数据库的分析过程中，社会科学化的定量分析优势在于对多变量相对作用的分析，以及对整体统计或因果关系的概括，然而对具体作用机制和社会、经济、制度环境的认识往往较为表面。而传统研究擅长结合丰富的相关史料，深入理解一些典型个案，这无疑对研究设计和具体分析助益良多。

更为重要的是，研究者需要有对特定历史背景的准确把握，才能选择合适的定量分析方法并正确解读分析结果。进行以量化数据库为基础的研究分析，研究者需要思考的首要问题是选取构建数据库的史料是否存在选择性偏误，即这些材料能够在多大程度上反映当时的社会现实，能够反映哪些特定人群、特定条件下的具体情况？只有对数据来源的选择性有充分认识，才可能避免错误解释分析结果或过分夸大结论的代表性。而这些必要的历史背景往往是通过传统文本分析获得的。因此，史学研究方法一方面可以深化对定量分析结果的理解和解释，另一方面也可以对定量分析结果和定性研究结果进行经验比较，通过多种研究方法的结合减少主观性风险的影响，提高对研究对象全面、深入的整体认识。“数量分析本身不是目的，只是认识的手段”，② 这些定性分析，对于以复杂的人类行为为研究主题的社会科学必不可少。对新一代历史学者来说，要将科学方法融入到传统研究中去，既不能因循守旧，也不可以盲目推崇新方法完全替代旧方法。

五、无限可能：量化数据库与“求是型学术”在中国

对史学研究来说，构建量化数据库是兼及史料整理和问题研究的重要工具，代表历史学未来发展的一个新方向。其尊重材料、注重探求事实本源的方法导向，与历史学的学术传统契合。然而，中国历史学者的研究和教学在此方面仍停滞不前。一方面，大量掌握在历史学者手中的

① Jean-Baptiste Michel et al., “Quantitative Analysis of Culture Using Millions of Digitized Books,” *Science*, vol.331, no.6014, 2011, pp.176-182.

② Dobrov Deopik et al., *Quantitative and Machine Methods of Processing Information*，转引自杰弗里·巴勒克拉夫：《当代史学主要趋势》，第337页。

重要人类经验数据资料可能难以被利用，历史学的学科功用受到限制；另一方面，历史专业的学生对新研究方法缺乏足够了解和学习，在数字时代的竞争力自然受到影响。如何缩小差距、转变思路，在发挥自身优势的同时实施教育改革，仍是历史学科建设亟须思考和解决的问题。

目前，历史学者从事量化数据库构建和研究的门槛已大大降低。近年来，一些暑期集训课程在推动中国量化历史分析教育方面取得很多进展，丰富的社会科学定量方法暑期培训项目为学生和教师短期内提升分析方法的修养提供可能。许多大学经管专业开设的发展经济学和经济史课程，也大量涉及量化历史研究的各种经典实例，为历史专业学生和学者深入学习相关理论与分析应用提供可能。而且，Excel 等大众化数据管理软件的普及与发展，对于数据库构建、数据管理和基本分析提供便利。SPSS、SAS、STATA、R 等专业统计软件在操作指令和交互界面方面也日趋人性化。

同时，社会科学的发展，为史学家和社会科学家的交流合作提供可能。① 跨学科合作交流或团队的建立，可以帮助历史学家突破技术壁垒。历史学家自身不可比拟的比较优势，更能够帮助他们有效融入新的研究潮流中，对正确构建和分析历史量化数据库意义重大。②中国史学界长期积累和正在进行的许多文献收集和整理项目，都具有构建大规模量化数据库的先天优势，预示着未来短期内大规模历史量化数据库及相关研究在中国蓬勃发展的广阔前景。

国内很多学术单位和图书档案管理单位都已经开始重视大规模、系统化历史材料的收集与整理，建立了一些文献保存或文献索引型数据库。如国家与地方档案馆的文献材料数码化工程和国家清史工程对清代文献的收集和整理；中国社会科学院近代史研究所对所藏约 19 万件珍贵档案的整理和电子扫描。各地高校历史研究机构则致力于大规模收集地方社会历史材料，部分也已进行电子化处理。如山西大学中国社会史研究中心收集、整理逾千万卷山西地方社会历史档案；中山大学历史人类学研究中心收集以闽粤为中心的地方家族谱和民间契约文书等；上海交通大学历史系开展以上海郊区、苏南、浙北地区为中心兼及中西部地区的“县级档案与契约文书的收集、整理与研究项目”，已汇集 8 省 50 县约 600 万页档案和 30 多万件民间契约；浙江大学地方文书与编纂中心则与浙江龙泉市档案馆合作，将该馆所藏晚清至民国时期总计 17333 卷宗，88 万余页的地方司法档案进行整理、电子化并出版。这些大规模史料收集与文献数据库的构建对材料的永久保存意义重大，大量分散在各地、各机构的重要史料得到了整合，一些数据库还支持检索功能。

虽然这类文献检索类数据库的保存格式和数据结构仍无法直接用于定量分析研究，但已经为构建量化数据库奠定坚实的材料基础，从而较快实现向量化数据库的转变，进而促进系统、深入地利用这些历史材料进行分析研究。可以预见，未来短期内我国历史学在上述材料相关研究领域将会有长足发展，不仅在国家与社会、民间契约、宗族行为与司法诉讼等社会历史研究方向取得重大进展，更会为相关社会科学和国际比较研究提供全新动力。

---

① Patricia Kelly Hall, Robert McCaa and Gunnar Thorvaldsen, “Introduction: International Historical Microdata—A New Resource for Research and Planning,” pp.3-10.

② 孙圣民指出，20 世纪 80 年代末 90 年代初，当国内历史学家开始关注历史计量学研究时，系统的西方经济学教学在国内才刚刚起步。历史学家缺乏合格的合作对象，即掌握现代经济理论和分析工具的经济学者。目前国内经济学教学和研究水平正逐步与国际接轨，上述条件和时机都已经具备，当务之急是为史学家和经济学家提供必要的交流和合作平台，如创办协会和年会、开办杂志或专栏等，通过双方合作撰写高质量的研究论文并参与国际历史计量学活动，提升国内历史计量学的研究水平。（孙圣民：《历史计量学五十年——经济学和史学范式的冲突、融合与发展》，《中国社会科学》2009 年第 4 期）

整合历史数据与当代数据的思路，无疑为进一步通过大规模量化数据库促进跨学科、跨时段、跨地域全面认识人类社会与自我，提供了无限可能。而在实际操作层面，各种数据公开、数据管理、数据安全和数据分析方面的具体问题也不容忽视。前文讨论的ICPSR中心以及IPUMS和UPDB等数据项目就是成功范例。它们不仅为许多学科学术研究提供整合历史与当代数据的新资源，更在实践中总结出大规模量化数据库管理方面的宝贵经验，涉及明确原始数据所有权、签署用户保密协议、隐去数据内个人隐私信息、限制用户接触和分析数据的方法与层级等方面的许多具体实践。

另外，卢森堡收入研究跨国数据中心（LIS Cross-National Data Center）除能如IPUMS等数据项目提供在线生成统计表的简单分析功能以外，还支持数据使用者自己编写统计软件代码，以远程分析保存在中心主服务器上的数据方式替代使用者自行下载、保存和分析数据的传统数据分享方式，从根本上提高了对公开数据的安全和个人隐私的保障。这些都为今后我们在构建、管理、公开和使用大规模量化数据库时，如何结合互联网技术最大限度公开数据、方便研究分析但又保证数据安全和个人隐私等敏感信息安全方面，提供重要的参考模板。

我们认为，提倡依托于大规模量化数据库的学术研究，对推动整个人文社会科学的“求是型学术”、真正认识中国社会和历史发展的特征、平衡东西方学术见解意义重大。人文社会科学注重“解释型学术”的传统，使得各学科在资料收集方面往往不断确认“已知”而轻视探索“未知”。历史学科的“选精”、“集粹”如此，当前社会学科中较为普遍和“科学”的统计抽样调查等方式亦是如此。尽管在数据收集的功效和人口层面的代表性等技术层面，统计抽样社会调查有其得天独厚的优势，但由于西方理论范式先行，我们在中国实施抽样调查项目往往仍需要大量借助西方已有研究经验，通过研究直觉提出问题和设计问卷时，难免会受到西方“已知”事实和理论的影响。诚然许多重要研究成果确认或修正了东西方“共性”的认识，但我们也在一定程度上丧失了认识中国“未知”特质、完善和平衡国际学界对不同社会差异根本认识的可能。不论是历史还是当代，基于档案等人口层面系统记录的大规模量化数据库，无疑成为弥补当前这一研究缺憾的最佳选择。这种以注重材料、探求事实为先的“求是型学术”，必定会促进我国历史学科乃至整个人文社会科学平衡、健康、全面地发展，也会为国际学术进步提供丰富的“中国经验”。

〔作者梁晨，南京大学中华民国史研究中心副教授。南京 210093；董浩，香港科技大学人文与社会科学学院社会科学部博士研究生；李中清，香港科技大学人文与社会科学学院讲座教授〕

（责任编辑：武雪彬 责任编审：路育松）

time, this involved various limitations and disadvantages.

Performing the Mystery Play of the Succession: The Rituals of Succession to the Throne in Ancient Egypt

Guo Zilin(85)

In the past, Western scholars tended to study Ancient Egyptian rituals of succession to the throne within the framework of ritual theory and from the perspective of religious studies. To gain an accurate appreciation of the practical significance of this ritual, however, we need to examine the enthronement and coronation rituals using a combination of the methods of anthropological and historical analysis. The two rituals were generally held separately at specific times in accordance with specific procedures. The basic meaning of the enthronement ceremony was that the successor to the throne was now king and was possessed of kingly authority. The coronation ceremony signified acceptance of the king's enthronement and his assumption of kingly power. The two ceremonies developed from many elements including kingly power, social and cultural concepts, and associated practices, as well as agricultural production activities. Successfully combining both the sacred and the profane, they used ritual spectacles, symbolic objects, relief carvings and tomb inscriptions to set up a sacred relationship between the king and the gods that would sanctify his position and rule. Furthermore, they established the various powers of the king throughout his realm and strengthened national identity and the cohesion of social forces. They thus played a very important role in maintaining and extending ancient Egypt and kingly rule. The survival of Ancient Egyptian kingship for nearly 3000 years was closely bound up with the performance and dissemination of these two rituals.

On the Reform of the German Empire and the Construction of the Nation State in the Late Middle Ages

Wang Yaping(101)

In Western Europe, from the 15th century on, England and France had successively implemented parliamentary monarchies which provided the essential political conditions for the construction of the nation state. In the German Empire, however, the state system and the regional economic center of the Empire that had been developing since the 14th century weakened the centralization of royal (imperial) power, forming a many-headed "polycracy." Outside powers like England, France and the Catholic Church seized the opportunity to interfere in German affairs, a situation that was a great impediment to Germany's becoming a nation state. At the end of the 15th century, the German emperor strove to reinforce the supreme authority of the empire by adopting reform measures such as levying unified taxes and establishing imperial courts of law. Although the reform of the empire did not achieve the desired political objectives, it was a crucial step in Germany's progress toward being a modern nation state.

Discussion and Remarks

Quantitative Databases and Historical Research Liang Chen, Dong Hao and Li Zhongqing(113)

Historiography is still regarded today as one of the traditional disciplines. However, more and more non-historians engaged in natural and social science research are turning to large-scale quantitative historical databases, with important results. Unlike traditional quantitative

historiography and the "interpretive research" stressed by the social sciences, particularly economics, what quantitative databases promote is "truth-seeking scholarship," i.e., unearthing new facts and gaining new understanding from a mass of systemic data by means of statistical analysis. Chinese history has long had a huge amount of systemic material on household registration, land distribution and the imperial examination system, as well as an immense number of written records. Many of them have been collected and organized, facilitating the construction of large-scale databases and the development of quantitative research. This new research paradigm not only encourages the development of the discipline of history itself, but also promotes inter-disciplinary and cross-border academic communication and integration; it thus contributes to a comprehensive and in-depth understanding of the distinctive features of Chinese social history as well as to balanced Eastern and Western scholarly development.

Over the last century and more, an inertial discourse system has developed in research on North China villages. Mutually reinforcing studies, experience and theories on changing village traditions, seen in the light of the *longue durée*, demonstrate that Lengshuigou Village, with its fine traditions and convenient location near the provincial capital, has had growing contact with external markets. The village's distinctive flavor is increasingly evident, and it has a relatively high standard of economic and socio-cultural development; yet it is frequently criticized in past research as closed, conservative, poor and backward. Modernity-dominated rural research forms an inertial discourse system that disrupts the regionalism, continuity and integrity of social history development and fails to understand or misunderstands village history and the nature of the village. To resolve this modern discourse dilemma, we must maintain a correct historical view, and, in particular, carry out our investigations from the viewpoint of the *longue durée*, putting a value on historical integrity and continuity and highlighting the important role of grassroots social practice in historical development.

Historical Notes