用大规模量化历史数据库检验中国的长期代际遗传

梁 晨
（南京大学 历史学院，南京 210046）

一、量化数据库与“求是型学术”

“求真”是历史学和自然科学共同的最高原则，历史研究应以客观事实为基础，而非价值观，这是中西方史学家长久以来共同追求的核心守则。在中国历史研究中，位列二十四史之首的《史记》，之所以能被誉为“史家之绝唱”，很大程度上是因为司马迁“不虚美、不隐恶”，将“秉笔直书”作为史家之最高准则。在西方史学研究中，兰克被尊为“近代史学之父”，他“据事直书”，不做任何价值判断的学术追求，因此被视为引领西方史学研究进入近代的标志性人物。然而，16 世纪“科学革命”以后，随着近代学科体系的建立，自然科学和人文与社会科学逐渐分化，自然科学研究着眼于发现，人文与社会科学研究强调解释。从学术研究发展的历程来看，自然科学研究逐渐转化成“求是型学术”，人文与社会科学研究则归于“解释型学术”①。

近年来，量化历史数据库显示出为“发现”和“解释”搭起桥梁的作用。近 50 年里，尤其是 21 世纪初，随着大数据技术的发展，历史研究的方法也有所更新，量化历史数据库的构建开始冲破“解释”的桎梏，向着“探索新知”迈进。所谓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。

中国有着丰富的历史文献，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。第一类是历代户籍材料；第二类是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；第三类是官员铨选材料，隋唐以来，考试（考核）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—康文林研究团队（两位目前均是香港科技大学人文与社会科学学院教授，以下简称李—康团队）一直是大规模微观量化历史数据库构建与研究的倡导者和践行者。大规模个人层面微观数据库的构建与研究是近些年的学术新动向，新的中国个人层面微观历史数据库的构建，对推进东亚社会经济史的研究与认识有着重要影响。李—康团队依靠以档案登记为主的史料，围绕着四个主题独立又彼此关联的个人层面微观历史数据库，进行建设和研究。这四个数据库分别是近代以来的土地财产、大学生、职业技术人员和官员群体个人层面历史数据库，数据跨度超越 200 年（1800－2000），包含超过 200 万人次的个人生命历程。其中有约

① “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，参见 Ernest. L. Boyer，*Scholarship Reconsidered: Priorities of the Professoriate. Princeton*，N. J: Carnegie Foundation for the Advancement of Teaching，1990.

100 万人次为独立的农村农民数据，100 万为城市中官员、职业技术群体和大学生数据。

这四个主题可以帮助我们研究和理解，在长期以来的中国社会中，谁获得财富，谁获得教育，谁获得职业，谁获得权力；将其串联起来，又有助于重新认识中国的长期“代际遗传（inter－generational transmission）”问题。所谓代际遗传，亦称代际传递，是社会科学家借用的生物学术语，意指社会地位、资源禀赋等在家庭前后代际间的影响或传递，是分析社会流动性和理解社会不平等形成的关键。这四个主题数据库，可以进一步帮助学界从多维度和长时间两方面推进代际遗传问题研究：从过去以追寻财富遗传为主扩展到理解教育和职业的代际传递；从过去强调血缘、宗族对后代的影响，到更直接地探讨官职（职位）的代际传递。同时，从 1800 年到 2000 年约 200 年的时间跨度，可以将对代际关系的观察从两代扩展到三代甚至更多代。

二、多主题数据和多维度代际遗传研究

李—康团队通过构建量化历史数据库来理解中国长期代际遗传项目，最初是从构建财富数据库起步的，试图以此理解近二百年来中国社会财富分配与获得状况。

1. 谁获得财富？

贫富差距大、社会不平等是发达国家和发展中国家所面临的共同现实困境，也是社会经济史研究的核心议题。在展开前，有必要澄清我们为何将主要的研究对象落脚于家庭财富而非收入。第一，一般而言，进入近现代以来，各个国家对于家庭财富都有相对完善以及统一的登记制度，这保证了历史材料的可获得性。第二，在进行经济社会史的跨国比较研究时，财富相较收入更具可比性。由于各国历史传统与制度等原因，收入的种类、形式更为复杂。第三，即便仅仅聚焦于中国，由于收入分配制度的巨大变化，在长时段内进行收入比较也较为困难。

要充分了解中国社会不平等机制，就必须深刻思考财富分配中的国家角色，以及财产权利与政治权力的互动与角力。李—康团队收集并建构的财富信息数据库，包括中国多世代人口数据库—双城部分（China Multi－Generational Panel Dataset，Shuangcheng，CMGPD－SC），中国土地改革数据库—双城部分（China Land Reform Dataset － Shuangcheng，CLRD－SC），以及中国四清阶级成分数据库（China Siqing Social Class Dataset，CSSCD）。这三个子数据库涵盖了从 19 世纪到 20 世纪中期 150 年内，国家在不同的区域、社群内对财富分配进行主导、干预，以建构不平等与平等的历史过程，为从基层出发理解中国社会的财产权利与政治权力提供了可能。

中国多世代人口数据库—双城部分（CMGPD－SC）[①]，提供了 1866－1926 年间生活在中国东北地区今黑龙江省双城县，基于人口统计学与社会经济特征的长时期的个体、家庭及其他相关方面的当地居民信息。数据库中包含了于 19 世纪初期移民至双城地区的超过 100 000 位隶属于八旗组织的居民的 130 多万条记录[②]。

这些旗人于 1815 年至 1838 年期间在清朝政府的组织和安排下移民至双城地区。当时，双城地区是一大片无人定居的草场。1815 年，清政府为了缓解北京地区旗人人口压力造成的财政困难，提出了将 3 000 户京旗从京城迁至这个边远地区的计划，通过为他们提供土地来替代政府月度、年度的财政补贴。CMGPD－SC 数据库所基于的原始历史材料，主要来源于双城

① CMGPD－SC 已经公开，读者可通过以下链接获取，https://www.icpsr.umich.edu/icpsrweb/DSDR/studies/35292。CMGPD－SC 的准备工作和通过 ICPSR DSDR 的公开发行是在 National Institutes of Health，Eunice Kennedy Shriver National Institute of Child Health and Human Development（NICHD）的资金支持下完成的，立项编号为 R01 HD070985，“Multi－generational Demographic and Landholding Data：CMGPD－SC Public Release”。

② CMGPD－SC 数据中的人口记录转录自保存在辽宁省档案馆的八旗人口户口册，读者可通过美国犹他州家谱学协会获取其中的部分电子数据。

地区的旗人户口册以及土地登记册。双城地区的旗人户口册由当地旗人政府编制和维护。每个户口册的条目首先依据村庄进行分类，然后是各族、各户。每个条目首先记录户主的信息：其迁出地、民族、原属旗、职业、姓名、年龄，以及其他自上次登记后的重要人口信息变动情况。随后户口册记录了户主其他直系亲属（妻子、儿女、父母）的信息，包括与户主的关系、姓名、年龄、职业等，以及其他与户主共同居住的亲属的相关信息。户口册中对于所有家庭成员的索引标识都是根据他们与家庭主要成年男性的关系来确定的。由于人口登记在土地分配中非常重要，双城地区的政府每年都对户口册进行更新。在每年的十一月，双城当地政府都会编制更新过的户口册，送至行省审阅。清朝规定旗人人口登记册应每三年更新一次，双城地区的旗人户口册是我们目前所找到的唯一一个每年一次编制更新的户口册。因此，与其他地区的三年一更新的户口册相比，双城旗人人口登记册在细节和完整度上都拥有较高的质量。

中国土地改革数据库—双城部分（CLRD－SC）①，收录了黑龙江省双城县在1946－1948年土地改革过程中所登录的涵盖全县70 000户、370 000人的农村人口和11 000户、50 000人的城关镇居民的阶级划分表，以及以个人为单位的被分地人登记表，被斗争人员登记表，在逃地主登记表与被处置人犯登记表。CLRD－SC辅之以李—康团队成员收集的其他围绕该县土地改革的多层级的质性档案。中国1946年到1952年间在全国范围内发起的土地改革运动是人类历史上规模最大的国家对私人财富的干预与调拨行为之一，这一数据库为研究土地改革中的国家行为与社会实践提供了基础。

四清阶级成分数据库（CSSCD）②，收录了1965－1966年间山西省、河北省、内蒙古自治区以及广东省19个县25 000个家庭、近十万成人的回溯性数据，为研究20世纪上半叶中国的社会、经济及政治环境提供了坚实的历史资料。该数据库所采用的史料来源于1965－1966年社会主义教育运动期间四清运动工作队记载的《阶级成分登记表》。登记表正面为各户家庭信息，记载了每户从1946年土地改革前到1966年的经济状况、户主的社会关系、三代家史；背面为家庭成员简况，记载了该户中15岁以上个人的社会人口信息，包括性别、年龄、民族、宗教、文化程度、职业、与户主关系，以及户主从祖父辈起的三代家史。

这三个数据库涵盖了从19世纪到20世纪中期的150年内，国家在不同社群中试图建构不平等与平等的制度历史。其中，CMGPD－SC与CLRD－SC所记录的不同政府在同一地区、同一社群建立不平等与平等制度的实践，我们得以窥见历史的延续性与断点之处；CSSCD的样本涵盖四个省份，分处华北和华南，通过区域间的比较，可以极大地拓展我们对从20世纪上半叶到农村集体化阶段的乡村社会财产制度变迁的理解。

2. 谁获得教育？

教育是另外一项确定个人社会地位的重要指标，它虽然不像财富那样可以直接遗传，但几乎所有的研究都表明，父辈、祖辈等所形成的家庭文化背景、家长职业乃至家庭财富状况等都会直接影响后代的教育获得，因此教育获得水平、教育获得的内容等在很多社会也有遗传性

① 本部分关于CLRD－SC的介绍及文字基于李—康团队成员倪志宏（Matthew Noellert）的博士论文及书稿（Noellert，Matthew，Power over Property：The Politics of Land Reform in China，1946－1948）。

② CSSCD介绍及用户手册，详见Noellert，Matthew，Xing Long，and James Lee，"The CSSCD－RCCSH User Guide：An Introduction to the China Siqing 四清（Four Cleanups）Social Class Dataset"；LI Xiangning，Wang Yuesheng，Noellert，Matthew，and James Lee，"The CSSCD－HB User Guide：An Introduction to the China Siqing 四清（Four Cleanups）Social Class Dataset"；Noellert，Matthew，Li Xiangning，Hao Xiaowen，and James Lee，"The CSSCD－SX User Guide：An Introduction to the China Siqing 四清（Four Cleanups）Social Class Dataset"。

特征。

李—康研究团队从20世纪80年代中期起，便开始注意收集和整理中国教育精英的个人层面系统性历史档案资料，以期能深入、准确地掌握中国教育精英群体的长期变迁。历经30年的努力和多方合作，目前已初步掌握近300年来中国教育精英社会来源的主要信息（如家庭背景、地理分布等）并开展了量化数据库的构建和分析工作。这些信息规模庞大、记录系统，且较为连续，涵盖清代几乎全部进士和官员以及大部分举人、贡生（1644－1911）、大部分“中华民国”时期大学生（1912－1949）以及中华人民共和国时期两所国家与省级精英大学的全部大学生（1950－2008）。该数据库的三部分数据来源大致如下：清代信息目前以缙绅录及各种科举数据（如题名录等）为主，几乎能涵盖所有清代进士和官员、部分的举人和贡生等，同时我们已经和厦门大学刘海峰、郑若玲教授团队合作，利用他们的长期研究成果，逐步系统构建清代分省举人数据库；民国信息主要是全国各档案馆所藏民国时期大学生学籍材料；中华人民共和国时期以北京大学、苏州大学两校毕业生及校友调查资料为主，同时目前正在山西大学、华中师范大学和上海交通大学进行同样的资料输入。其中“民国大学生数据库”是李—康团队自主构建和研究的，2010年以来由李—康研究团队成员梁晨主导，基于学生个人层面的量化数据库，是对“北京大学—苏州大学学生量化数据库”及其研究的延伸。

该数据库项目的主要目标是根据民国时期各大学记载的学生记录，主要是入学时填写的信息，来构建量化数据库，并以此研究民国时期的教育获得、不平等和教育精英群体特征。教育精英的社会来源包括多个维度：地域来源、家庭来源、系统来源，分别对应数据库中学生的籍贯和家庭住址、家长职业和来源中学。数据库中丰富、全面、庞大的学生社会来源信息，是量化研究教育精英社会来源与社会流动难得的材料。

民国大学生量化数据库尽管未能包括民国时期所有大学及学院，但已经涵盖四所最大和最重要的“国立”大学，分别是北平“国立”清华大学、上海“国立”交通大学、杭州“国立”浙江大学和广州“国立”中山大学，亦包括多所精英的教会大学，分别是上海圣约翰大学、沪江大学、苏州东吴大学、杭州之江大学以及南京金陵大学和金陵女子文理学院；在私立大学方面，数据库亦收录了一度被誉为上海滩学生数量最多的私立大同大学。但需要指出的是，由于资料的保存等问题，被收录的各个学校在数量和连贯性上有较大的差别。截止到2019年1月，量化数据库已经涵盖民国时期全国共27所专科以上学校学生信息，其中包括128 242名学生共157 160条个人信息。目前，仍有燕京大学、铭贤学院、“国立”沈阳医学院、辽东学院、英士大学的学生信息正在录入中，数据库建设完成后，有望囊括民国时期33所学校近18万大学生的微观层面数据。

3. 谁获得工作？

中国劳动力量化数据库（The China Workforce Database，CWFD）是收集民国时期和中华人民共和国成立后中国新兴职业群体信息的数据库，是李—康研究团队这两年新启动的数据库构建和研究项目。该数据库初步拟定包含的职业群体类型主要有大学教职员、医生、工程师、会计和律师等。建成后，将包括超过100 000名专业人士的个人微观记录。同时，该数据库具有和中国大学生量化数据库匹配连接的巨大潜力。

我们把数据库分为两个阶段，分别是民国时期和中华人民共和国成立后：中国劳动力量化数据库—“中华民国”（CWFD－ROC）和中国劳动力量化数据库—中华人民共和国（CWFD－PRC）。中国劳动力量化数据库包括五个子数据库：中国会计师数据库、中国工程师数据库、中国法律从业者数据库、中国医生数据库和中国大学教职员数据库。

4、谁获得权力?

官员群体及选任制度事关一个社会权力的分配与结构，尽管从现代社会的角度看，官员也是职业之一种，但至少在中国社会，官员或权力历来是影响社会发展之最核心要素，相对于一般职业有其特殊性，因此需单独进行考虑。李—康研究团队的中国社会经济史研究一直将官员群体作为一大重点。

缙绅录是记录职官的职掌、姓名、出身、籍贯、字号等基本情况的专书，最早可以追溯到南宋的《班朝录》，在明朝已颇为流行。如果从发行者的角度来区分，缙绅录可分为官刻本和坊刻本，官刻本一般是吏部进呈的本子，坊刻本则是由京城的书铺出版的；如果从内容来看，还可以分为文缙绅和武缙绅。如果按照发行的时间区分，缙绅录每季推新，分为春、夏、秋、冬四部，如此循环往复。清代保留至今的缙绅录文献规模浩大，提供了连续性的官员记录，是建立清代官员群体的大规模历史量化数据库的理想史料。

现存的清代缙绅录文献规模庞大，仅清华大学图书馆出版的《清代缙绅录集成》中已收录的便有 200 余种。每种缙绅录几乎覆盖了当时所有的文职官员和部分八旗驻防官员，内容简洁、条目化、全覆盖，并且具有持续性的特征，非常适合建立长时段、大样本的量化数据库。缙绅录为学界提供了极为系统的个人层面微观数据，同时这些数据每季度出版，具有极好的历史连续性，而且至少几乎完全包含了 1760 – 1911 年间所有清政府官员的信息。李—康研究团队在《清代缙绅录集成》的基础上借助互联网对海内外各大图书馆所藏缙绅录进行了更为全面的查询和整理，目前已搜索到 2 700 余种，其中包括大量重复和季节不明的版本。如果按照年代和季节排除重本，目前能够确定至少现存 402 种不同时间点的缙绅录。

目前，李—康研究团队已经将清华大学图书馆出版的《清代缙绅录集成》基本输入了量化数据库。这部分数据总计约 280 万条记录，涵盖近 30 万名官员。同时，研究团队也正在积极输入其他来源的缙绅录，最终预计数据库将达到 500 万条记录规模，对应约 40 多万名官员。同时由于缙绅录规模庞大，信息较为复杂，自 2018 年起，李—康研究团队还与中国人民大学清史研究所达成共同开放和研究的协议，李—康研究团队将通过人大清史所的平台，对中国学界免费开放缙绅录数据库。其中首批数据预计在 2019 年底开放，数据涵盖 1900 – 1912 年。

三、用量化数据库与数字人文推进历史教研

上述四个数据库从不同角度形成了我们对中国长期代际遗传的新理解。举例来说，通过“民国大学生学籍数据库”和科举数据库、当代大学生数据库的连接计算，我们得出了一些初步但可能较为重要的实证发现：在过去一个半世纪中，中国教育精英在社会和地理来源等方面出现过多次剧烈的阶段性、结构性转变。这些转变大致可分为四个阶段：1865 – 1905 年为第一个阶段，官员子弟在教育精英中几乎形成了垄断；1906 – 1952 年是第二个阶段，商人和专业技术人士等近现代新职业群体代替传统官员群体，在教育精英者家长职业中占据明显优势；1953 – 1993 年是第三个阶段，工农或无产者子弟成为新的教育精英优势群体；1994 – 2014 年是第四个阶段，有产者和工农无产者子弟混杂，但有产者子弟的优势逐渐显现①。这四个阶段与晚清的国家新政、民初的资产阶级建国、1949 年以后的社会主义改造以及 20 世纪 80 年代开始的市场化经济改革等重大社会革命和转型阶段相互交织。相对于现代西方主流社会，中国教育精英的来源更为多样和多变，精英群体的构成异质性较强。理解近现代中国的发展路

① 梁晨、董浩、任韵竹、李中清《江山代有人才出——中国教育精英的来源与转变（1865 – 2014）》，《社会学研究》2017 年第 3 期。

径，除了显而易见的重大社会革命这条主线，我们不应忽视中国精英群体尤其是教育精英在这段历史中所经历的转型。

当前，李中清、任韵竹和梁晨正在围绕“民国大学生数据库”进行《中国近代知识阶层的形成与来源》一书的写作。依靠民国大学生学籍数据库中 26 所高校近 10 万名大学生的个人数据，我们发现新式教育在中国建立后，很多学生在大学甚至中小学时就需要离开出生地和家庭，进入城市，且多数时间在学校而非家庭。校园从空间上为民国大学生提供了个人生活和独立思考的私人天地。这种空间能够让个人和家庭及其所在区域相对隔离，使得学生可以自由选择自己的所学科目、毕业出路、宗教和政治信仰与人生追求。因此，尽管科举家族在明清时期具有重要地位，但通过考察民国大学生数据库中的学生家长，我们发现新式教育下的成功者更多以家庭而非家族为单位出现，这一点在从事学术教育的家庭中表现尤其明显，这说明了民国时期家长对学生价值观的影响远大于家族。

以上工作表明，量化历史数据库的构建，不仅是一种国际史学界的新潮流，更对深化史学研究工作具有重要价值。量化历史数据库的构建与研究不仅可以看成是数字人文研究的重要组成，而且得益于数字人文的发展，其学术价值也不断提升。量化数据库的构建离不开大规模史料的掌握，特别是系统性史料。数字化史学的发展，带来了史料的大规模数字化；数字化后的史料，不仅给开放、查阅带来了便利，更为研究利用提供了方便。同时，一些数字人文技术还能帮助研究者直接将电子史料转化为所需格式，大大便利了系统性输入。最新开发的一些技术甚至可以帮助研究者从海量电子化史料中，智能辨识和抓取所需信息，进行数据库构建和研究分析。尽管这些方法还需要不断试验和调整，但这无疑表明了量化数据库建设与研究的用武之地会越来越广阔。

基于石仓文书的清代物价数据库建设①

蒋　勤

（上海交通大学 历史系，上海 200240）

一、民间文书中物价数据库建设的意义

当今学界热议“数字人文”，史学界也不例外。文献全文库、人物履历库、量化数据库及相应的内容分析、社会网络分析、计量分析方法，讨论最多。本文侧重量化数据库建设，以分享利用清代民间文书建立物价数据库的经验，为学界系统利用新史料抛砖引玉。

社会经济史研究以可靠的史料整理工作为基础。近年来，明清民间文书的收集、整理与研究迈入了新阶段。民间文书收藏的数量巨大，多家机构收藏规模达十万件以上。按“归户”原则整理文书成为共识，图文并录的出版方式开始流行，降低了研究者的释读难度。2015 年，国内首个大型契约文书数据库②上线，为数字人文技术的介入提供了条件。

① 本文是上海市社会科学规划课题（2018BLS003）阶段性成果。

② 《中国地方历史文献数据库》（http://dfwx.datahistory.cn/）。