改革开放史研究如何应对大数据时代的新挑战

潘　娜

随着大数据浪潮的深卷，人类社会正在加速进入计算型和智能型社会。就当前的科学研究而言，大数据不仅广泛应用于自然科学领域，社会科学中的许多研究领域也逐渐转变为数据密集型学科，包括历史学。纵观改革开放近40年的历史，其中的半部已然是数字信息技术驱动我国经济社会创新发展的历史。数据不仅是记录和测量这段历史的工具，更是继续创造这段历史的“基础设施”。在大数据时代，数据收集与数据挖掘技术已经发生了翻天覆地的变化。甚至继大数据之后，信息科学领域对元数据（metadata）的认知与挖掘正在快速加强人类应对复杂数据世界的能力。这一趋势使量化历史研究迎来了前所未有的黄金时代。但遗憾的是，我们的思维方式和研究方法还没有跟上这种改变。当前，在探讨改革开放史学科建设和方法体系建设时，大数据已经成为一个无法回避的话题。

一、大数据时代改革开放史研究面临哪些新挑战

大数据是计算机和互联网技术发展到一定阶段的必然产物。然而，大数据的影响已经远远超出了信息技术领域，正在深刻改变着整个人类历史的发展逻辑和发展方向。在这种情况下，改革开放史研究的外部条件和历史本体均发生了质的变化，传统的历史资料、研究范式、学科发展正在一步步受到冲击与挑战。

首先，大数据时代改革开放史研究资料巨量化、多源化、数据化的新挑战。一是巨量化。与年代较远的历史相比，改革开放史研究的史料瓶颈并不在于稀缺，反而在其巨量特征上。正如有学者指出的那样，虽然改革开放史研究的机密档案多未解密，但是通过公开渠道披露信息的速度极快、数量极大[①]。这实际上加重了研究者收集史料和应用史料的难度。二是多源化。随着互联网和智能移动终端的普及，历史资料的来源越来越广泛，保存的介质也更加多元，文本、图片、音频、视频等各种载体的历史资料丰富多样。传统的党史国史研究往往较多依赖解密档案和权威部门文献这些集中呈现历史事实的“直接史料”。而在大数据条件下，原本鲜少作为史料使用的历史信息，一旦数量达到一定规模形成完整信息链或相关性信息网，便可以作为更为客观的历史资料反映历史事实，但处理这些海量“新史料”需要借助新的研究方法和分析工具。三是数据化。改革开放进程中越来越多的历史信息直接以数据的形式产生和存储，非数据形式存储的历史信息数据化的速度在不断加快，特别是国内图书档案管理机构和教学研究机构，甚至研究者个人都在大规模、系统性地将珍藏史料进行数据转化。数据化意味着计算机可识别、可检索、可复制、可计算，这将从根本上改变史料应用的逻辑和效能。

其次，大数据时代改革开放史研究面临“范式失灵”的危机。每一个学科经过一段时期的发展和积累，都会形成本学科独特的内在结构和研究模式。例如，改革开放史研究的学术共同体在史观、史料、史学理论、研究手段、技术标准等各方面所达成的基本遵循，构成了该学科的研究范式。研究范式往往具有稳定性，但如果研究对象或研究条件发生巨大变迁则会导致“范式失灵”，即传统研究范式不能提供解决新问题、应对新挑战的科学方法和理论预设。在大数据的冲击下，改革开放史研究的范式失灵突出体现在两个方面。一是传统研究方法缺少量化研究流程的缺陷被放大，定性研究或抽

① 章百家：《积极开展改革开放史研究》，《中共党史研究》2009年第1期。

样统计方法正在受到大样本量化研究的挑战。传统历史研究方法以具体档案或文献为支撑，依靠研究者的经验判断和理论思辨得出结论，不管被大数据证实或证伪，其科学性和规范性显然都难有说服力，难免被诟病为“史观学派”。二是传统的解释性研究难以揭示历史规律的弊端更加显露。党史国史领域已有的改革开放史研究成果较多聚焦于对改革开放缘起、分期、主线等问题的学理探讨以及对改革开放进程中某些争论的思想交锋，而对波澜壮阔的历史实践和历史规律性问题研究则较为薄弱。由于改革开放史与现实发展紧密相接，改革开放实践的全局性、规律性问题恰恰是党和国家现实决策亟须汲取的历史经验。在大数据条件下，挖掘隐藏在海量历史资料中的规律性问题在技术上已经具备可行性，而传统学科范式缺乏对技术发展的积极回应，局限性日益突出。

再次，非专业领域的“数据治史”挑战改革开放史学科发展的地位和前景。当前，统计学、经济学、新闻传播学、情报学、管理学等社会科学的许多领域已经建立了大数据的研究体系和挖掘平台，产出的研究成果也深入改革开放史研究的广阔领域。例如，清华大学公共管理学院建立“政府文献中心”并研制开发“政府文献信息系统”供学者进行量化研究①；北京师范大学新闻与传播学院与大数据平台公司合作建立视频大数据挖掘研究基地，形成开放性平台吸引国内外学者进行视频大数据挖掘②。在非专业学科积极主动将大数据内化为学科发展新动力的形势下，党史国史领域的改革开放史研究对大数据的反应相对迟滞。虽然80年代以来随着西方计量史学研究方法的引入，人口史、经济史、军事史、社会史等领域产生了一些关于改革开放历史阶段的计量研究成果，但量化研究远未形成主流。此外，在互联网开放平台上，各种真伪难辨的“历史档案”“历史回忆”“历史数据”往往居心叵测地“抹杀”改革开放的历史成就。而反驳历史虚无主义，仅靠逻辑之辩或个别史料回击，恐怕也难以体现学术研究的专业性、严肃性和权威性。总之，在大数据条件下，非专业领域计算历史、检验历史的技术能力越来越强，党史国史学科开展改革开放史研究的专业优势在逐渐减弱，“专家治史”越来越深地受到“数据治史”的挑战，研究成果影响力式微、学科发展边缘化的风险日益严峻。

二、改革开放史研究如何应对大数据时代的新挑战

从积极的角度来看，大数据冲击是“倒逼”改革开放史夯实量化研究基础，逐步趋向于科学研究范式的宝贵契机。事实上，对历史全面进行量化研究并不是新理念。早在20年代，梁启超就提出了“历史统计学”的概念，并设想将统计学应用到全部史学研究当中③。时至今日，这一史学理想终于具备了时代条件。当前，利用大数据开展改革开放史量化研究，既在技术上具备可能性，又在政策上进入窗口期，亟须在史料基础、范式转换、人才结构等层面提出应对大数据挑战的可行路径。

第一，呼吁改革开放以来的政府数据实现全面开放共享。没有历史数据的全面开放共享，改革开放史的量化研究只能是空谈。与传统计量史学“抽样样本+数学模型”的研究方法不同，大数据逻辑是全样本逻辑，需要尽可能穷尽历史数据。而改革开放以来高质量的历史数据大部分都集中在政府部门和行业机构，不是

① 有学者基于这一系统，对1978年至2013年的科技政策文献进行数据挖掘，在联合行文部门之间行文数量的异常变化中，发现了科技主管部门与其他部门的合作与冲突关系，而政府部门间冲突往往是隐蔽化的，不通过大样本量化研究很难证实。参见黄荣 《政策文献量化研究》，科学出版社，2016年，第202、215—217页。

② 《北师大—蓝鹰视频大数据挖掘研究示范基地成立将探索校企合作良好范式》，《北京师范大学校报》2017年3月15日。

③ 梁启超认为：“欲知历史真相，决不能单看台面上几个大人物几桩大事件便算完结；最要的是看出全个社会的活动变化。全个社会的活动变化，要集积起来比较一番才能看见。往往有很小的事，平常人绝不注意者，一旦把他同类的全搜集起来，分别部居一研究，便可以发现出极新奇的现象而且发明出极有价值的原则。”《梁启超全集》第14卷，北京出版社，1999年，第4045页。这一设想实际上就是大数据时代的计算特征即全样本，可见我国史学研究在理念上是十分超前的。虽然这一超越时代条件的设想在实践中并没有实现，但梁启超史学研究的量化思维和全景思维直到今天仍然有其前沿启示性。

研究者个人或个别研究单位收集史料自建数据库就能完成的，需要得到国家层面的支持。当前，世界各国都在加快政府数据开放以助力经济社会和科学研究的创新发展。以美国为例，美国政府的数据开放网站发布的数据集已经超过19.4万个①。当前，我国正在加快透明政府建设的步伐。2015年9月，国务院印发《促进大数据发展行动纲要》，着力解决我国各级政府数据“不愿开放共享”“不敢开放共享”“不会开放共享”② 的问题，明确提出到2018年底建成国家政府数据统一开放平台。借这一政策窗口期，改革开放史专业领域应进一步呼吁将政府历史数据开放纳入国家大数据战略，并提出全面清理和开放政府历史数据的实施方案。除了保密期档案和敏感部门数据，应将改革开放以来大量沉积在各级政府部门和公共机构的历史数据进行数据库化建设，并在一定范围内全面开放共享，不宜联网获取的数据也应建立完整的信息资源目录以备研究者申请利用。

第二，充分利用元数据建构改革开放史量化研究的分布式数据关联平台。在数据开放的基础上，数据关联是更为重要的基础建设。实际上，近年来国内相关研究单位和图书档案管理机构基本上都在大规模、系统性地将库藏档案和文献资料进行数据库建设，如“中国共产党思想理论资源数据”“中国共产党历史文库”“人民数据库”“文献研究室资料数据库”“国家图书馆海外中国问题研究资料”“当代中国研究所自建数据库”以及大量由高校和社会研究机构建设的历史档案资料数据库等。当前亟须将分散在全国各地的数据库进行平台一体化整合，形成数据库联盟，这是大数据量化研究的重要基础准备。在技术层面，元数据的挖掘和运用提供了数据库整合的解决路径。例如，欧洲数字图书馆和美国数字公共图书馆都开发了独有的元数据模式，将来自于数千个文化遗产机构（包括图书馆、档案馆、博物馆等）的资料进行平台性整合与在线共享。而这两个机构并不存放这些巨量资料，只是作为“枢纽”为用户提供在线搜索和利用这些资料的操作平台。改革开放史的量化研究亟须建立数据库联盟，以分布式数据共享平台的模式实现高质量的政府数据、行业数据、文献数据一体化整合，从而扩大样本规模，加强多源数据的相关性研究。当前，建立数据库联盟的难点不在技术层面，关键在于学术共同体的合作意识和意愿，能否以开放的心态实现改革开放历史数据的共享共建。

第三，改革开放史量化研究的范式转换要与大数据挖掘相结合。在做好基础数据准备的同时，更为重要的是如何形成善用数据的研究范式。这就需要改革开放史研究的学术共同体能够对本学科逐渐转向数据密集型学科的发展趋势有清醒认识，从而达成推进量化研究主流化的共识。研究者要能够突破传统研究范式的思维惯性和流程缺陷，从学科规范的角度将量化研究内化为研究流程中一个不可或缺的步骤，更要将大数据挖掘作为检验研究结论、发现历史事实、探寻历史规律的基本方法。在此基础上，逐步建构历史数据科学的研究框架和体系。需要特别强调的是，强化数据挖掘的基础作用并不是否定史学研究者的主体性，而是为研究者抓住改革开放过程中的大线索、大脉络、大节奏提供客观的判断依据。正如有学者指出的那样，“数据挖掘始于数据”的观念十分错误，数据的挖掘始于要解决的问题，只有弄清解决什么问题，才知道需要什么样的数据，才知道选择何种数据源③。当前，党史国史学科的研究重心正在整体向改革开放史转移，很多重点难点问题尚未找到研究突破口，（下转第53页）

① 2017年6月12日，笔者在美国政府数据开放网站以“China”为关键词进行搜索，可以找到1822个中国主题的数据集，涉及领域广泛。其中，美国国家航空航天局地球观测系统数据与信息系统数据中心单独设立了一个“中国数据集”，11个子集主要呈现了中国20世纪八九十年代的基础地理、人口、农业、医疗、行政区划等方面内容，为研究者和公众提供了观察中国改革开放以来经济社会快速发展的数据信息。详见美国政府数据开放网站 https://catalog.data.gov/dataset?q=CHINA&sort=score+desc%2C+name+asc&ext_location=&ext_bbox=&ext_prev_extent=-183.515625%2C-30.751277776257812%2C-17.578125%2C72.81607371878991&page=1.

② 单志广：《抓住“开放共享”这个关键》，《人民日报》2015年11月20日。

③ 陈潭等：《大数据时代的国家治理》，中国社会科学出版社，2015年，第49页。

今天的人们普遍认同市场经济，是经历几十年计划经济不成功的实践后得到的认识。依迄今的人类经验，市场经济是不可替代的，包罗万象的计划经济则是行不通的。

然而，市场经济也有多种模式，西方在政府干预与放任自由之间存在着一种周期性钟摆现象。20 世纪 20 年代末的全球性危机，使早期自由放任经济模式宣告终结，代之以凯恩斯主义和罗斯福新政的兴起，其核心是强调政府的全面干预。事实上，“自 20 世纪 30 年代起，整整有 40 年之久，支持纯粹自由市场经济学的知识分子都是孤立的少数。”[①]然而，1973 年至 1983 年，西方世界发生了以“滞胀”为特征的经济危机，凯恩斯主义成为众矢之的，以“撒切尔主义”与“里根经济学”为代表的新自由主义经济模式再度崛起。历史似乎已经宣告了新自由主义经济模式的最终胜利，但没有料到到 2007 年夏季，美国爆发了自 20 世纪 30 年代大萧条以来最严重的金融危机，并波及全球，西方国家中又出现了许多“反思资本主义”的声音[②]。这场危机对世界带来的震荡远没有结束，或者说才刚刚开始。资本全球化引发的广泛而深刻的不平等以及左、右民粹主义的兴起，对我们是另一种警示。它也再次提醒人们：当年思想界普遍关心的问题并没有永久地成为过去。中国现阶段的问题是如何继续推进市场化改革的问题。但是，在市场化改革的进程中，也决不可无视社会的公平正义问题。

（本文作者　华东师范大学中国当代史研究中心兼职研究员　上海　200241）

（责任编辑　王志刚）

① 〔英〕艾瑞克·霍布斯鲍姆著，贾世蘅译：《帝国的年代》，第 372—373 页。

② 参见萧冬连：《政府和市场关系的周期性钟摆现象》，《当代中国史研究》2016 年第 1 期。

（上接第 38 页）数据挖掘无疑是打破学科发展内在瓶颈的“他山之石”。

第四，通过改革开放史量化研究积极培养历史数据科学专业人才。学科发展的根本动力在人才，特别是创新型人才。史学研究者往往并不善于使用有一定技术门槛的统计方法和分析工具，这实际上是制约量化历史研究精深发展的最大瓶颈。有研究者考察已有计量历史研究成果，发现大多是频率分析、回归分析等基本统计方法，主成分分析、判别分析和聚类分析等高级统计方法在史学界很少有人用[①]。更别说在大数据时代，数据挖掘的多维性和复杂性更抬高了史学研究者开展改革开放史量化研究的进入门槛。短期内完成阶段性研究项目，可以委托专业大数据公司承接数据挖掘的基础工作，也可以与相关专业加强跨学科合作。但从长远来看，改革开放史量化研究的可持续发展，最终还是要培养本专业人才开展长期深入的研究，这就需要积极探索历史数据科学这一新的专业领域。当前，大数据挖掘刚刚起步，产学研合作培养数据科学专业人才也在探索之中。例如，2015 年，阿里云与八所高校合作开设“云数据与数据科学”专业方向，未来几年将逐步扩大数据科学教育的覆盖面。历史数据科学在史学领域和数据挖掘领域均处于学科探索的最前沿，应当通过开展改革开放史量化研究的契机，积极加强产学研合作，联合培养历史数据科学的专业人才。

（本文作者　中国社会科学院当代中国研究所助理研究员　北京　100009）

（责任编辑　吴志军）

① 陈争平：《大数据时代与经济史计量研究》，《中国经济史研究》2016 年第 6 期。