DOI:10.16524/j.45-1002.2015.12.008

计算历史学：大数据时代的历史研究

马建强

[摘　要] 随着互联网、计算机技术的发展，大数据时代对历史研究的冲击在所难免，历史学也面临挑战与机遇并存的命运。当前内容庞大、功能多样的数据库与日俱增，历史研究所面对的数据环境越来越健全。面对这样的学术环境，一些学者提出了对历史学发展的思考，一些学者则凭借大数据时代的独特环境，开展了一些新的史学研究实践。大数据时代中的历史研究是一条正在探索的道路，计算历史学可能会成为历史研究发展的一个趋向。

[关键词] 计算历史学；大数据时代；历史研究；数据库

[作者简介] 马建强，湖北大学中国思想文化史研究所博士研究生，武汉大学社会发展研究院大数据与计算社会科学研究中心跨学科团队研究人员，湖北　武汉　430062

[中图分类号] C919：K02　　[文献标识码] A　　[文章编号] 1004－4434(2015)12－0099－07

19世纪，当科学劲风鼓鼓吹袭时，传统史学便已被吹开了科学化的航帆。20世纪的史学受到科学的冲击前所未有，可谓巨大、全面、彻底，几有颠覆传统史学根基的势头，使得传统史学无法独立自存，开始自觉依傍科学这个阔气的后台。传统史学一方面改良自身的基因缺陷，另一方面吸收其他学科包括自然科学的优秀基因，通过积极的自我调适，达到“科学性”，最终凤凰涅槃一般地生存下来。今天虽然历史学通常不被视作“科学”，但其“科学化”转变以后形成的学科特色和学科价值已经被广泛接受和认同。经历生死存亡的20世纪，历史学似乎暂时坐稳了自己的学科地位，然而它所面临的冲击、挑战却始终未绝。回顾21世纪刚刚过去的十几年，我们可以惊愕地发现，历史学正面临着全新的、势头更猛烈的、速度更快的科学巨浪的冲击。计算机科学、互联网技术以及由此带来的“大数据”便是这一波巨浪的代表。面对冲击，史学研究者应该有更多的理性思考，在理论与实践两方面积极探索史学未来的发展趋向，这对历史学适应时代潮流获得崭新生命有着重要意义。本文试图梳理当前史学界应对新环境作出的积极回应，并探讨“大数据时代”历史学发展可能的走向以及史家应具有的态度和付出的努力。

一、挑战与机遇：“大数据时代”下的历史学命运

“大数据”(big data)概念诞生未久，是一个新兴事物。它伴随着信息技术产业和互联网行业制造的巨量数据而出现。目前人们对于大数据的探索才刚刚开始，对于它的定义也莫衷一是。维基百科这样定义：“大数据，或称巨量数据、海量数据、大资料，指的是所涉及的数据量规模巨大到无法通过人工，在合理时间内达到截取、管理、处理、并整理成为人类所能解读的形式的信息。”[1]百度百科如此定义：“大数据，是指无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的数据集合。”[2]维克托·迈尔-舍恩伯格及肯尼斯·库克耶所著的《大数据时代：生活、工作与思维的大变革》一书认为：大数据是指不用随机分析法（抽样调查）这样的捷径，而采用所有数据进行分析处理。当前人们对于大数据特征的认识被总结为“5V”，即Volume（大体量）、Velocity（时效性）、Variety（多样性）、Value（大价值）、Veracity（真实性）。因此人们认为大数据只有运用云计算服务器，才能进入可运行的轨道，得到处理并实现价值。Web2.0时代，数据究竟有多大？据统计，2013年中

国产生的数据总量超过0.8ZB(1 ZB = 1,024 EB = 1,048,576 PB=1,099,511,627,776GB),是以万亿GB为单位的量。同时2013年产生的数据总量是2012年的两倍,相当于2009年全球数据总量。预计到2020年,中国产生的数据总量将超过8.5ZB,相当于2013年的10倍。然而中国每年新增数据量还不到美国的十分之一[3](P21)。这个统计体现了大数据不仅巨量,而且增长速度极快,国与国的分布也很不均匀。这样巨量数据的产生是人类过去生产方式、生产能力所不能实现的,大数据开启了一次重大的时代转型,是数千年未有之大变局。因此近几年信息技术行业开始用"大数据时代"来称呼我们当前的这个时代。

当前,大数据似乎已经以其不可阻挡的强势冲击着各个领域,展露出它在很多实用领域的优势,体现了大数据的控制、主宰力量。历史学恐怕也不可能侥幸逃离这一波冲击。广义而言,数据之于史学等同于史料。史料是史学得以存在与发展的基础,是史学家借以获取史学认知、判断的依据,这是中外史家的普遍共识。"史料者,历史知识之唯一源泉也"[4](P176),"近代的历史学只是史料学"[5](P3),这样类似的言论见诸中外史家者屡见不鲜。笔者以为,大数据时代的冲击对于历史学来说应该既是挑战也是机遇,而且首先是挑战。

每当人类历史面临全新事物挑战之时,有的热情拥抱新事物,以积极的心态参与到新事物的建设之中;也有的对新事物采取鸵鸟政策,埋首不问时间流转、宇宙变迁;还有始终视新事物为"奇技淫巧""变乱纲常"的妖孽,势不两立,大加挞伐。近100年前,当国人对"新文化"充满争议的时候,钱玄同如此形容"残废颓败"的守旧老人:"似乎不必请他们享受新文明的幸福,尝新生活的趣味;因为他们的心理,只知道牢守那笨拙迂腐的东西,见了迅速捷便的东西,便要'气得三尸神炸,七窍生烟','狗血喷头'的骂我们改了他的老样子。"[6](P27)历史学在面对大数据时代的挑战时,首先应该调整姿态,应对挑战,适应新的研究环境,既不盲目拒斥,也不一味趋新,在新旧传承、互动中探索,在探索的同时找到新的发展机遇。

早在大数据概念还没有产生广泛影响时①,华文学界已有先行者在探索计算机科学与互联网时代的史学研究方面迈出了步伐。就21世纪而言,具有代表性的首先应属台湾清华大学的黄一农教授,他在2005年出版的《两头蛇:明末清初的第一代天主教徒》一书中率先提出了"e考据"的概念并进行了成功的研究尝试。他认为,"一个有条件孕育'e-考据学派'的时代或许已悄然到临","大量史籍被整理重印,再加上网络和电子资料库的蓬勃发展,深感史学研究已晋入一前人所无法想象的新局,益发决定要面对这项自我挑战"[7](P64)。这体现了黄教授对学术研究新环境的敏锐性观察和在史学方法论建设上的自觉。2008年旅台学人金观涛、刘青峰夫妇出版了《观念史研究:中国现代重要政治术语的形成》一书。书中作者提出"数据库方法"[8](P5-7),并以"中国近现代思想史专业数据库(1830-1930)"研究中国近现代观念演变。2009年台湾大学数位典藏研究发展中心举办了第一届"数位典藏与数位人文"国际研讨会,提出"数位人文"的概念:"指的是那些唯有借助数位科技方能进行的人文研究,反过来讲,数位人文的研究,即是企图寻找在前数位时代中难以观察的现象、无法想象的议题与无法进行的研究。"[9](P11)同年辽宁大学历史学院教授焦润明提出了"网络史学"的概念:"是历史学在信息时代的一种新存在形式"[10]。

在大数据观念逐渐兴起的同时,史学界对互联网时代的史学发展问题的思考也更加深入广泛。2011年《史学理论研究》杂志刊发"互联网与史学观念变革"笔谈②。2013年《甘肃社会科学》以"信息转向:新世纪的历史学在召唤"组织专稿两篇,其中学者周兵提出了"数字史学"(Digtal History)概念:"是指运用数字媒体和工具展开的历史学实践、演示、分析和研究。"[11]2014年暑期黄一农教授在浙江大学主持的"第二届e考据与文史研习营"结束不久,上海《文汇报》"文汇学人"刊发专题文章《当乾嘉学派遇上互联网》介绍"e考据"方法及研究成果[12]。2015年《史学月刊》第1期刊发"计算机技术与史学研究形态笔谈"③。2015年12月4日至6日,上海大学也成功举办国内首次以"大数据

①"大数据"概念诞生于1980年,著名未来学家托夫勒在《第三次浪潮》一书中将"大数据"称为"第三次浪潮的华彩乐章"。2008年9月《自然》杂志推出名为"大数据"的封面专栏。而直到2009年开始"大数据"才成为互联网技术行业中的热门词汇。

②该笔谈共刊发了李剑鸣的《网络史学的神话与实际》、王晴佳的《互联网的普及与历史观念的变化》、王加丰的《互联网资料的史料价值》、马勇的《"自媒体时代"的历史研究与史学表达》、王旭东的《信息化的历史学:基于互联网驱动的史学变革》、刘军的《互联网与社会平等》6篇文章。

③该笔谈共刊发乔治忠的《历史研究电子资源运用的兴利除弊》、王子今的《"史实"与计算机"利器"》、王文涛的《信息时代的文献阅读与史料检索》、陈爽的《回归传统,浅谈数字化时代的史料处理与运用》、周祥森的《空间转向:电子传媒技术与当代史学形态》5篇文章。

时代下的历史研究”为议题的国际学术研讨会。

大数据时代的冲击，究竟能够给史学带来什么新的机遇，史学界提出了很多有价值的思考。正如黄一农教授所言，“e 考据其实就是文科对于大数据的一个回应”[12]，其实何止“e 考据”，包括“数据库方法”“数位人文”“网络史学”“数字史学”等等在内的这些概念建设都应当看作史学界为回应大数据时代对于史学研究机遇前景的思考与探索，是人文主义与科学精神的再次碰撞。这些思考和探索的核心都离不开计算机、互联网以及“数据”或“数据库”（数据集合）。站在今天回望历史，早在 1949 年世界上第一台计算机诞生刚刚 3 年，意大利学者罗伯托·布萨便开始使用计算机对神学家托马斯著作中的字词“进行大规模的处理，包括每个字的用法、位置，大概就预示了未来史学研究与计算机的不可分离。而 1960 年代末法国年鉴学派史家勒华·杜拉里预言，“将来的历史学家一定要是电脑程序员，否则，就不足以成为历史学家”[11]，虽然这番预言所针对并不是今天史学所面对的局面，但从当前来看这一预言无疑正在一步步走向现实。

二、建设与尝试：历史学研究中的“大”数据运用

面对大数据时代对史学的冲击，史学界应该并且已然进行了一些勇敢的建设与尝试。所谓的“建设”是以积极心态为营造更好的数据环境而进行的建设；所谓的“尝试”是在大数据时代的环境下进行史学研究的新尝试，主要是利用海量的网络数据以及规模较大的“数据库”进行。

在数据环境的建设方面，台湾地区是先行者，最先开始探索以实现全文检索为目标的古籍数字化。早在 1985 年，台湾“中央研究院”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“超过两千万字的台湾史料、一千万字的大正藏”、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①。值得一提的是，据笔者对黄一农教授访谈所知，早在 1987 年黄教授便使用该资料库中的“二十五史”部分研究天文史的议题，并有了对文史环境改变的最初体会。1999 年香港迪志文化出版公司出版“文渊阁四库全书”电子版，该数据库以超过 7 亿字的规模成为当时最大的数据库②。进入 21 世纪，以全文检索为基础的数据库发展迅猛。台湾雕龙中国古籍全文检索数据库起始于 2001 年，在 2013 年时已声称收入古籍文献约 20000 多种，近 25 亿字，且以每年新增 5000 种文献 10 亿字的速度递增，数年后将成为全球第一的超大型中国古籍全文检索数据库③。

大陆方面在数据环境建设的方面起步晚于港台，但是近年来成果显著。在古籍数字化方面成就最为突出的是北京爱如生公司。2001 年该公司与北京大学刘俊文教授合作，研发制作“中国基本古籍库”，该库分 4 个子库、20 个大类、100 个细目，精选先秦至民国历代重要典籍，总计收书 1 万种，单库全文超过 17 亿字。目前爱如生公司已陆续推出包括中国近代报刊库、中国方志库、中国谱牒库、中国类书库等在内的大型数据库 14 个；包括四库系列、别集丛编系列、历代碑志系列、地方文献系列等在内的 9 个系列共 82 个专题数据库；包括明清实录、永乐大典、四部丛刊等在内的数字丛书库 50 个。另外还有“原文影像版数字原典”产品 8 个、“全文检索版拇指数据库”9 类 1000 个产品④。由北京时代瀚堂科技有限公司推出的《瀚堂典藏》，分为古籍数据库、近代报刊、民国文献大全三大主体部分。全库共包含有 15000 多种古籍，25000 种民国报纸期刊，近 4000 万条记录，汉字总量超过 40 亿⑤。近年来湖南青苹果数据中心有限公司提出创建“华文报刊文献数据库”计划，将从清朝嘉庆年间至今两百年的 4000 种报刊中挑选十分之一进行数字化，形成拥有 4000 亿汉字和 4 亿篇文章的海量历史文献库⑥。

以上所举仅是能够实现全文检索的大型综合数据库，除此以外还有大量规模较小的全文数据库，如书同文古籍数据库、中华经典古籍库；或专题数据库，如中国金石总录数据库、东方杂志全文

①参见成果网站：http://hanji.sinica.edu.tw/，2015-05-22 日。

②2002 年台湾启动“数位典藏国家型科技计划”，2008 年与“数位学习国家型科技计划”结合，形成“数位典藏与数位学习国家型科技计划”。参见成果网站：http://digitalarchives.tw/。该计划包含档案、图片、古籍、影音等多种类型的台湾地区学术、收藏机构的资料、藏品，并不以全文检索的方式实现。

③参见成果网站：http://www.diaolong.net，2014-11-13。

④参见成果网站：http://www.er07.com/，2015-11-20。

⑤参见成果网站：http://www.hytung.cn/，2015-11-23。

⑥参见成果网站：http://www.huiawenku.cn/index.html，2015-09-24。

数据库;以及不能实现全文检索的大型数据库,如“大成故纸堆”系列数据库、晚清期刊全文数据库(1833-1910)、民国期刊全文数据库(1911-1949)、中美百万册数字图书馆、国家图书馆民国图书、民国期刊数据库、读秀学术搜索等等①。另外在企业行为之外,史学界也对数据建设进行了探讨。2013年8月,教育部社会科学委员会历史学学部年度会议进行了“历史资料的整理、研究和数字化建设”的专题研讨,有赵毅、桑兵、钱乘旦、曹树基、常建华、沈志华、葛剑雄、李剑鸣等15位史学家作了专题发言。2010年以来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有近70项,其中隶属于“中国历史”学科门类的重大项目有6项、重点项目1项、其他类别2项②。

虽然目前数据建设还未臻成熟,但是史学界一方面已经认识到了建立相关专业数据库的重要性,同时也意识到数据库对推动研究的促进作用。伴随着日益丰富的数据环境,有一些史家利用数据库或创建数据库展开新的研究尝试,获得史学研究的新突破或开创了新领域,涌现出一些代表性的成果。

首先,谈谈黄一农教授提出的“e考据”。自2005年以来,黄教授始终号召并实践着这种“大数据时代”的文史研究方式。在笔者对黄教授的访谈中,黄教授提出“e考据”并不仅仅是一种研究方法,并且还应该是一种融通数位与传统的研究态度。“e考据”是在e时代作考据,而并非只是用e的方法作考据。以“e考据”的学术方法和学术态度,2010年黄教授从原本非常熟悉的科学史、中西文明交流史跨入了被认为已遭遇研究困境的“红学”这个陌生的领地。但是仅仅5年时间,黄教授从第一次完整阅读红楼梦开始,深入探索并在“红学”领域取得了一系列令人瞩目的成果,出版了第一部红学专著《二重奏:红学与清史的对话》。黄教授的研究为原本被认为已无多少新材料会出现的“红学”挖掘出一批过去不为人知的真实可靠的新史料,并填补诸多历史细节的隙缝,使得“红学”与“清史”之间的隐秘联系被彰显出来。这本著作既是“清史”与“红学”的“二重奏”,也是数位与传统的“二重奏”,是一部充分展现“e考据”典范的力作。

第二,在文学史研究领域,以武汉大学王兆鹏教授为代表的团队,自2005年开始尝试以数据计量分析唐诗名篇的影响力,并陆续扩充数据、完善统计方法。于2011年出版《唐诗排行榜》一书,对外公布了该团队研究成果的第四个版本。著名的文学史家傅璇琮先生评价该研究说:“这是一部既有传统深厚理论依据,又处处洋溢着现代学术新意的著作。这部著作从传播和接受的角度,依诗作影响深度和广度的标准对有唐三百年间的诗歌第一次进行了令人信服的排行,这种研究方式和文本呈现,无论在理论拓展还是实践创新方面,都具有开创性意义。”[14](P235)考察王兆鹏教授团队研究的内在理路,其学理依据仍然是文学史研究中的传播、接受理论,而在方法上则是利用了新时代才能实现的依托于数据库的计量分析。虽然该研究也遭到来自各方对于数据量、计算方式等的质疑,但是我们也应该看到,在文学史研究领域中,古典文学数字化与定量研究这个议题逐渐被更多的文学史研究者关注、认同并加入其中[15]。学者更愿意提出一些建设性的意见和可以开拓的新领域。

第三,以金观涛、刘青峰的《观念史研究》一书为代表的数据库关键词词频统计、语义分析与观念史研究。作者借助于内容达一亿两千万字的“中国近现代思想史专业数据库(1830-1930)”进行观念演变的探讨,并将这种方法称之为“以包含关键词例句为中心的数据库方法”[8](P1)。作者认为这种研究得以展开的前提便是“历史文献向数码化的方向发展”,“原则上讲,研究者可以通过建立包括过去所有文献的专业数据库,采用数据挖掘方法,把表达某一观念所用过的一切关键词找出来,再通过核心关键词的意义统计分析来揭示观念的起源和演变”[8](P5)。这种数据库方法将观念史从思想史的附庸中解放出来,获得了独立的生命,也避免了过去以核心人物、经典为本为中心的思想史研究的局限。观念史的研究更能够体现思想发展的一般性特征,使思想史成为可以检验的。这种可检验性当然取决于数据库与计算机的数据挖掘能力。但是作者也承认,在整个研究过程中,数据库与计算机并非是唯一的全程参与者,“最重要的仍是研究者能否有效地利用挖掘出的大量数据,结

①分别参见成果网站:http://guji.unihan.com.cn/,2015-11-23;http://www.zhbc.com.cn/shownews.asp?id=2349,2015-11-23;http://jsk.ch5000.cn/,2015-11-23;http://cpem.cp.com.cn/,2015-11-23;http://www.dachengdata.com/,2015-11-23;http://www.cnbksy.com/shlib_tsdc/index.do,2015-11-23;http://www.cadal.zju.edu.cn/,2015-11-23;http://mylib.nlc.cn/web/guest/minguotushu,2015-11-23;http://mylib.nlc.cn/web/guest/minguoqikan,2015-11-23;http://www.duxiu.com,2015-11-23。

②该统计依据“国家社科基金项目数据库”,统计包含“中国历史”“中国文学”“语言学”“民族问题研究”“图书馆、情报与文献学”等学科门类。参见 http://fz.people.com.cn/yangshuo/skyg/sk/。

合历史背景和文本结构分析，概括出某一时代某一普遍观念的理想类型，这依然是思想史研究的基本方法”[8](P6-7)。

第四，以李中清、梁晨为代表的研究团队以“量化史学”的方法和“群体史学”的眼光进行中国教育精英研究。2013 年两人曾出版《无声的革命：北京大学、苏州大学学生社会来源研究（1949-2002）》一书。在今年 11 月 7 日的北京论坛史学分论坛上，李中清教授以《中国教育精英四段论》为题首次向国内外听众介绍了这项研究，认为：“1865-1905 年，即清政府废除科举之前，超过 70%的教育精英是官员子弟，来自全国各地的‘绅士’阶层；1906-1952 年，超过 60%的教育精英是地方专业人士和商人子弟，尤其是江南和珠三角地区；1953-1993 年，约超过 40%的教育精英是来自全国的无产阶级工农子弟；1994-2014 年，超过 50%的教育精英来自各地区的有产家庭，与特定的重点高中。”该研究依托于李中清、康文林领衔的“基于个人层面的、从 1760 年至今中国教育精英社会与地区来源的数据库”[16]。这项研究使笔者联想到潘光旦的《近代苏州的人才》、张仲礼的《中国绅士》、何炳棣的《明清社会史论》三部著作，它们都利用了大量的历史数据和统计计量，具有典范意义。然而相较于今天计算机所能够处理的数据而言，这些数据都只能算是小数据。

第五，由哈佛大学燕京学社、台湾“中研院”史语所、北京大学中古史研究中心合作的“中国历史人物传记资料库”（China Biographical Database Project 简称 CBDB）及基于此数据库的相关研究。当前该数据库还在持续建设之中，截至 2015 年 4 月数据库共收录约 360000 人的传记资料，这些人物主要出自 7-19 世纪，目前数据库正在收录更多的明清两代人物传记资料。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。研发者将历史事件转化为结构化数据，数据架构由人物、亲属、非亲属社会关系、社会区分、入仕途径、宦历、地址、著述等部分构成。通过这种结构化数据的提取、分析，研究者可以据此对历史人物进行群体研究，能够得到相关人物、事件的空间分布以及复杂的社会关系网络。相对于一般的数据库，该数据库可以实现更深层次的数据挖掘。同时也提供了一个计算机处理语义复杂的汉语文言文文本的示范，使得长时段的量化研究、空间分布研究可以实现，并从社会经济史领域扩展到政治史甚至是思想史领域的研究中，对于开启未来研究新方向很具启示意义。

在这些研究中，“e 考据”融通数位与传统，综合使用各种互联网数据、数据库以及传统文史研究方法来开拓研究新局。其他几种大都依赖于某一专业数据库的建设，是基于专业数据库对原有研究议题或新的研究领域所展开的新尝试。实际上，数据环境的建设与史学研究的尝试两方面是紧密相关、不可分离的。建设和尝试围绕同一个核心即“大数据时代下”的史学研究，都依赖于互联网、计算机等技术与设备，建设是尝试得以展开的前提和基础，尝试又为建设积累经验教训，并进一步指导建设的前进方向。这两者应该始终保持有序互动、共同推进。

三、传承与开创：“大数据时代”与历史学的前瞻

第一，大数据时代带来历史学方法论预流与范式突破。1930 年，现代著名史家陈寅恪在为陈垣《敦煌劫余录》所写的序中提出了一个著名的观点。他说：“一时代之学术，必有其新材料与新问题。取用此材料以研究问题，则为时代之新潮流。治学之士，得预此潮流者，谓之预流。其未得预者，谓之未入流。此古今学术史之通义，非彼闭门造车之徒，所能同喻者也。”[17](P266)陈寅恪从新的学术材料的发掘以及由此产生的新问题来前瞻学术的发展趋向，认为进入这个时代新潮流的学术称之为“预流”。伴随着计算机、互联网技术的发展，大量的数据库层出不穷，历史存留的文献也正在被夜以继日地数字化，我们明显感受到了未来文献载体数字化的这种强劲趋势。大数据时代伴随着新的文献载体，史学研究的新方法论也正在形成。借用陈寅恪的“预流”观，我们可以发现，大数据时代下史学方法论的新潮流也正在成型，今天的文史学界正在经历一场由技术革新带来的方法论预流。

1962 年，美国科学哲学家托马斯·库恩在《科学革命的结构》一书中系统提出范式理论。范式通常是一套学术共同体共同遵守的研究体系，它是当时一切研究的显著模式并为后来研究发展提供空间。当范式发生突破，便出现科学革命，导致探讨的问题发生转移，确定合理问题及解决问题的标准发生转移，改变了思维方式、研究对象并引发相关重要问题的争论[18](P5)。借库恩的“范式理论”来理解历史学的学科前瞻，可以认为大数据时代利用计算机、互联网以及大型数据库来获取史料、挖

掘分析史料信息的一套思维和方法也将成为史学研究的一种新范式。这种范式的形成将会带来全新的学术问题、学术理念、学术思维、学术视野以及学术方法、学术形态。从一定程度上说，大数据时代正是历史学范式突破的一个契机。

第二，计算历史学可能成为大数据时代史学的新趋向。在社会学领域，罗玮、罗教讲的《新计算社会学：大数据时代的社会学研究》一文将新计算社会学（new computational sociology）这一概念介绍给了中国学者，产生了广泛的学术影响。作者认为："新计算社会学是当代社会学界借助计算机、互联网与人工智能技术等现代科技手段，利用大数据、新方法来获取数据与分析数据，从而研究与解释社会的一种新的范式或思维方式。"[19]在中国历史学领域，1922 年梁启超在东南大学史地学会作了"历史统计学"的演讲，提出"历史统计学"的概念："历史统计学，是用统计学的法则，拿数目字来整理史料推论史迹。"[20](P4045)1935 年商务印书馆出版了史学家卫聚贤的《历史统计学》一书。西方 1950 年代产生了"计量史学"的概念，并逐渐影响中国史学界。近年来大数据时代冲击下的史学界也产生了"e 考据""数字史学"等思考，但是目前中国史学界还鲜有对 "计算历史学"（Computational History）这一概念的自觉认识与建设。

笔者认为，"计算历史学"应该与"新计算社会学"相似，可能成为超越"计量史学"的大数据时代下的史学发展新趋向。"计算历史学"所能够实现的前提是计算机科学、互联网、大数据以及人工智能技术等在历史学研究方面的有效利用。史学界对于"计算历史学"的认识与建设也会伴随着大数据时代下相关技术的进步、数据的完善、研究的推进而不断深化。"计算"最终将远远超越"统计""计量"，体现出人类借助于技术而实现的对历史文本、信息、数据更强大的挖掘、分析能力，弥补人脑在面对庞大信息时搜集、分析上的自然局限。正如上文所述，当前借助于"大数据"的一些历史研究新尝试所示，通过丰富的互联网资源，建设庞大、精准甚至结构化的数据库，能够让历史研究者为进入研究困境的学科开创新局，能够处理过去无法处理的学术议题，能够获得过去人类自身认识局限所不能够认识到的问题，也能够启发研究者开拓更多的新研究空间。

第三，研究者的主体地位与温故知新的学术态度仍然重要。中国传统文史学界将"博雅"视为一个崇高理想，"博雅"实际体现的是人对史料的吸收记忆范围之广，运用处理能力之强。钱钟书以《管锥编》《谈艺录》两部经典著述成为 20 世纪文史学界"博雅"的典范。在今天有人质疑钱钟书的价值，认为钱锺书无非是一个"电脑数据库"。不过吊诡的是，因"博雅"而被称为"电脑数据库"的钱钟书在 1984 年便开始倡导将计算机技术引入古典文献的搜集、疏证和整理中来，并且规划指导了"中国古典数字工程"[21](P237-244)。钱钟书非常注重计算机技术在文史研究中的运用，但同时也认为："实践证明，能帮助人的计算机需要人的更多的帮助。"[22]作为一个具有深厚文史积淀的前辈学人，钱钟书超前而又辩证地提出了对未来文史领域中人与计算机技术之间关系的思考。

未来计算历史学得以飞跃发展的一个关键应该是人工智能技术的进步，人工智能技术一定程度上也可以认为是针对人与计算机关系的探索。罗凤珠女士是台湾地区较早关注计算机与文史研究领域的一位学者。她在 1987 年曾访问当时信息科学领域的张仲陶教授，文史领域的周何教授、毛汉光教授、王邦雄教授、王熙元教授，发表了《探一探文史数据自动化的路》一文。张仲陶教授认为，"不要问计算机能做什么，而是问你要计算机做什么"；毛汉光教授认为"在文史自动化的过程中，成败的关键在文史界，不在计算机界"；王邦雄教授认为"文史自动化不能失去人的主导地位，计算机毕竟不是人，无法做创发性的工作"[23]。这些与钱钟书看法相似的关于人与计算机关系的思考，说明在技术面前研究者的主体地位的重要性，这对史学界来说仍然有着指导性意义。

"计算历史学"作为大数据时代中历史研究的思维和范式，研究者在探索的过程中既要注重开创也要注重传承，应该有"温故知新"的学术态度。所谓的"故"既包含传统研究的学术方法和学术积累，也包括大数据时代下陆续开展的种种史学研究的新尝试所积累的经验与教训。所谓的"新"则是不断发展的计算机技术、互联网技术、人工智能技术，以及与日俱增并不断系统、完善、精确的数据环境，以及在此基础上的新问题、新思维、新视野，它是永远面向未来开放发展的。在充分温故的前提之下，不断地知新，不断地积累经验、教训进行再创造，使"故"与"新"之间保持一种健康有序的互动、动态和谐的传承。

大数据时代的历史研究没有特别的捷径，需要史学工作者的勤勉与努力，严谨厚重仍然是历史学的特点。研究者在面对新的学术环境时必须有方法

论更新的自觉和勇气，也必须有全新的历史思维和问题意识，大数据时代既带来了研究的便利，也给研究者施加了新的研究压力。计算机能够帮助人，但同时它帮助人的能力更需要通过人的帮助来不断提高。面对新环境更好地发挥人脑的主动性、创造性，引导计算机、互联网、人工智能技术配合历史研究发展，积极地面对并建设历史研究所需要的数据环境，更是这一代历史学者的使命。

四、结　语

大数据时代的到来真切地改变着人类社会的方方面面，这种冲击也必然波及历史学研究。历史学研究在大数据时代遭遇新挑战的同时也面临全新的发展机遇，未来历史学是否能在这一波浪潮的冲击下乘风破浪，很大程度上取决于当代历史学者对时代的敏感性、对这一波冲击的认识以及是否具有方法论危机感和自我革新的勇气、自觉。当前数据建设的进程日益加快，越来越丰富、越来越多样的数据库为新的历史研究提供了新的环境和新的便利，在此基础上有一些学者对历史学的发展提出了颇有启发意义的思考，也有一些学者利用大数据时代的网络、数据环境开展研究，在打开研究新局面、开创研究新领域、提出研究新思维等方面作出了有益的尝试，为未来大数据时代史学研究提供了具有参考性的实践经验。大数据时代对于历史学来说是一个带来方法论“预流”与范式革命的时代，未来计算历史学可能成为大数据时代历史研究的一个发展趋向。但是在这一进程中，历史研究者既要以温故而知新的态度来对待数据建设与研究尝试，又必须充分发挥作为研究主体的能动性，协调好研究之中人与技术的关系。

[参考文献]

[1] 维基百科”大数据“词条[EB/OL]. https://zh.wikipedia.org/wiki/大数据.2015/10/23，2015－11－17.

[2] 百度百科”大数据“词条[EB/OL]. http://baike.baidu.com/subview/6954399/13647476.htm，2015－11－17.

[3] 郭为.一部精彩纷呈的时代杰作（推荐序二）[A].涂子沛.数据之巅：大数据革命，历史、现实与未来[C].北京：中信出版社，2014.

[4] 郎格诺瓦，瑟诺博司.史学原论[M].李思纯，译.上海：商务印书馆，1926.

[5] 傅斯年.历史语言研究所工作之旨趣[A].欧阳哲生.傅斯年全集：第三卷[C].长沙：湖南教育出版社，2003.

[6] 钱玄同.李大钊《新的！旧的！》的附言[A].钱玄同文集：第2卷[C].北京：中国人民大学出版社，1999.

[7] 黄一农.两头蛇：明末清初的第一代天主教徒[M].上海：上海古籍出版社，2006.

[8] 金观涛，刘青峰.观念史研究：中国现代重要政治术语的形成[M].北京：法律出版社，2009.

[9] 项洁，涂丰恩.导论——什么是数位人文[A].项洁，王泰升，等.从保存到创造：开启数位人文研究[C].台北：国立台湾大学出版中心，2011.

[10] 焦润明.网络史学论纲[J].史学理论研究，2009，(4).

[11] 周兵.历史学与新媒体：数字史学刍议[J].甘肃社会科学，2013，(5).

[12] 任思蕴，李纯一.当乾嘉学派遇上互联网[N].文汇报·文汇学人，2014－10－17.

[13] 项洁，翁稷安.导论——关于数位人文的思考：理论与方法[A].项洁编，金观涛，等.数位人文研究的新视野：基础与想象[C].台北：国立台湾大学出版中心，2011.

[14] 傅璇琮.唐诗有了排行榜之后——读唐诗排行榜[A].濡沫集[C].北京：北京联合出版公司，2013.

[15] 苗贵松，等.中国古典文学数字化进程中的定量研究和争鸣：兼论唐戴叔伦编年系地信息平台建设[EB/OL]. http://www.guoxue.com/?p =14705.2013/09/16，2015 －11－17.

[16] 彭珊珊. 专访李政道之子李中清：150 年来中国的精英出身什么家庭 [EB/OL].http://www.thepaper.cn/news－Detail_forward_1395229，2015－11－12.

[17] 陈寅恪.陈垣《敦煌劫余录》序[A].金明馆丛稿二编[C].北京：三联书店，2001.

[18] 托马斯·库恩.科学革命的结构[M].金吾伦，胡新和，译.北京：北京大学出版社，2012.

[19] 罗玮，罗教讲.新计算社会学：大数据时代的社会学研究[J].社会学研究，2015，(3).

[20] 梁启超.历史统计学[A].梁启超全集（第7册）[C].北京：北京出版社，1999.

[21] 胡小伟.钱锺书与电脑时代[A].丁伟志.钱钟书先生百年诞辰纪念文集[C].北京：三联书店，2010.

[22] 胡小伟.钱钟书与中国古籍数字化[N].人民日报，2011－01－13.

[23] 罗凤珠.引信息的“术”入文学的“心”——谈情感计算和语义研究在文史领域的应用[J].文学遗产，2009，(1).

[责任编辑：戴庆瑄]