

人工智能驱动的历史研究,恰如数码相机时代的绘画艺术一样,并非零和博弈。从写实的角度看,绘画技艺显然敌不过哪怕最低端的数码相机;但是,在艺术创意的维度上,绘画艺术依然有极强的生命力。历史研究同样如此。人工智能“加持”的历史学家的确能够实现智能增强,但因循传统路径的学者也并不会走投无路,两者可以在不同维度上找到自己的定位。

这种乐观主义的底气来自于人文主义的价值,也正是包括历史学家在内的人文主义者的核心竞争力所在。实际上,人工智能尽管是当下最尖端的科技,但它依然无法撼动历史研究的一些基本要素。英国历史学家爱德华·霍列特·卡尔(Edward Hallett Carr)在反思历史研究客观性时,提醒同行要有更广阔的视野。按照卡尔的说法,客观的历史学家不是单纯地寻找事实,<sup>①</sup>还意味着“有能力将自己的视野投射到未来,从而对过去有更深刻、更持久的洞察力,这一点是那些将视野完全局限于自己眼前境况的历史学家无法企及的”。<sup>②</sup>在卡尔写作的时代,数字技术并没有进入人文学者的日常,人工智能被应用于历史研究更是闻所未闻,但在他为历史学家预设的宏伟气象中,要有“朝向理解未来”的目标,人工智能为历史研究带来了宏大视角,历史学家也要学会驯服人工智能这个“大杀器”,<sup>③</sup>强有力的手段就是人文性。

## 从理论倡导到研究实践:数字史学未来走向刍议

梁 晨 (南京大学历史学院教授)

近年来,数字技术的迅猛发展催生出一系列数字化分析工具——各类文本、关系网络和地理信息分析工具不断出新,能自主挖掘材料并形成具有创意文本或视频的大语言模型、文生视频模型更是横空出世,数字技术似乎已能介入从史料收集、材料分析到文本写作和史实再现等全过程的史学研究环节。史学界对依靠数字技术推进历史研究的期待与日俱增,学术热度也愈发凸显:仅2022年,“人工智能、大数据与历史学融合发展”入选中国历史研究院评选的“中国历史学十大研究热点”、“数字人文”正式成为教育部批准的本科专业、北京大学历史学系开始招收“人工智能与历史学”方向的博士研究生。

不过,在目前中国数字史学甚至数字人文学界,学理层面的分析和对未来的展望较多,而基础设施建设和实证研究均很不足。整个领域给人以理论倡导者多、付诸实践者少的感受。与此同时,许多历史学者对数字技术是否真地能够有助于历史学研究心存疑虑。例如,成一农认为,历史学的终极目标是有目的、有人性地“解释”和“预测未来”,数字技术对此没有多少帮助。<sup>④</sup>不管是否认同成一农的看法,总体而言,史学界对数字史学尚持观望之态,很少有研究者下场一试是不争的事实。一份2023年的中国数字人文学术体系调查研究指出,当前中国数字人文存在的首要问题是“学科发展

① Richard Evans, *In Defence of History*, Granta Books, 2000, p. 225.

② Edward Carr, *What is History?*, Penguin Books, 1987, p. 123.

③ “大杀器”当然只是一个比喻,但人工智能具有潜在的风险是有目共睹的事实。人工智能的训练需要消耗大量化石燃料,带来了更多温室效应的排放,破坏了环境;如果人工智能被别有用心的人用来开发核武器、生物武器,对于人类世界则是切实的威胁。所以,驯服人工智能,不是杞人忧天,而是非常迫切的需求,参见 Guglielmo Tamburrini, “Artificial Intelligence and Large-Scale Threats to Humanity”, in Hannes Werthner et al., *Introduction to Digital Humanism*, Springer, 2024, pp. 241 – 254.

④ 成一农:《抛弃人性的历史学没有存在价值——“大数据”“数字人文”以及历史地理信息系统在历史研究中的价值》,《清华大学学报》2021年第1期。

不平衡”：“信息资源管理学科在中国数字人文版图中占据了较大的优势”，语言、文学、历史、艺术等理应是数字人文研究主体的学科却“只有屈指可数的资深学者和部分中青年研究者在扛起大梁”，“在身体力行投入数字人文研究之外，还不得不对更多的同行鼓与呼，证明数字人文存在的价值与意义”。<sup>①</sup> 国际数字人文学界的专业研究尽管比国内要活跃，但也有一个明显的现象：“数字人文研究基本等同于数字文学研究”，<sup>②</sup>数字史学的活跃度不高，成绩不够突出。因此，在数字技术和数据库建设都有一定积累的今天，数字史学界的同仁或许应努力从理论家转变为行动者，尤其要多开展既能突破传统研究，又能与之对话的数字史学研究，以此更好地融入和推动史学研究，在彰显自身价值的同时，获取应有的发展空间。

### 一、从跨学科到本学科

师出需有名，名正言才顺。开展行动或进行实证研究，首先应明确研究所属学科、应采用的方法等基本问题，这关系到研究的目的和应当遵循的学术规范。实际上，尽管近些年量化数据库建设、史料数字工具开发等已有一定的进展，但数字史学家们对什么是数字史学，如何定位数字史学的研究目标和学科属性等根本性问题仍有颇多不同意见和难分彼此的暧昧态度。有学者指出，无论数字人文还是数字史学，都强调探索和应用当下前沿的数字技术和设备，而技术与设备的变革又特别迅捷，这使得数字史学或数字人文有些漂浮不定，难以被简单定义，不仅自身的定义不甚明确，甚至应被看成一套研究方法还是某个“专门学科”也存在较多争议。<sup>③</sup>

不可否认，跨学科是数字史学最重要的学术特征，是其学术活力的源泉。作为综合采用多种学科方法开展研究的新领域，数字史学当然应强调多学科合作，推动包括网络信息、图书档案、数据分析、出版与学术评价等在内的多个专业和领域进入史学问题的研究。目前中国大学中的数字人文研究，往往依托于建立跨学科的数字人文中心，这类研究中心一般不归入某个传统学科与院系，这一情况在很大程度上说明了数字人文研究需要跨越现有的学科体系。有学者认为，数字人文不仅与语言、历史等传统人文学科相关联，同时也要与某种技术基础设置相关联，因而在某种程度上与现有的高等教育体系不相容。<sup>④</sup> 不过，这可能既不能说明传统的历史、语言等学科体制的失败，也不足以说明跨学科的数字人文研究可以没有学科或者必须去学科。仅就语言逻辑而言，若是没有本学科，又何来跨学科？尽管数字史学的学科界限相对含混，新技术条件下的学科理念也会松动，但跨学科的前提仍然暗含学科立场。

“学科”是17世纪科学革命后逐渐形成的历史产物，在它出现之前，各类人文艺术的研究与创作早已存在。人文学术因研究对象的复杂性和方法的主观性，并不能完全熨帖地套入后设的“学科”概念中，这是一直以来的现象。人文学科本质上都是跨学科的，数字史学并不具备独一无二的特殊性。对历史研究而言，自1929年法国年鉴学派成立，整体史就被抬上舞台。要研究整体的历史，自然要

① 夏翠娟、祁天姣、徐碧珊：《中国数字人文学术体系构建考察——基于实践项目的内容分析和文献研究》，《数字人文研究》2023年第4期。

② Adam Crymble, *Technology and the Historian: Transformations in the Digital Age*, University of Illinois Press, 2021, p. 30.

③ 苏珊·施莱布曼、雷·西门子、约翰·安斯沃斯：《新数字人文导论》，尹倩、肖爽译，北京大学出版社2023年版，第1、165页。

④ 苏珊·施莱布曼、雷·西门子、约翰·安斯沃斯：《新数字人文导论》，第593页。

涉及诸多对象,如经济、法律、科学等,由此历史学从研究对象到研究方法都需要跨学科。此外,现代学科普遍需依附于高度制度化的高等教育与科研体系以获得资助、开展研究,这些制度环境具有很强的稳定性,所以有学者认为学科“仍将是当代大学的一股强大力量”,虽然各种跨学科的研究领域不断兴起,但学科依然会“幸存”下去。<sup>①</sup>实际上,学科之所以稳定,就在于它拥有明确的研究对象、规范成熟的理论方法及与之相关的职业场域。很多学科碰撞下产生的新研究领域,在孵化阶段一般需要创新的体制机构来容纳它,但随着发展成熟并逐步形成稳定的研究范式,又需要明确地融入所对应的主流学科。

总的来说,学者们普遍认同数字史学是技术与史学的结合,是使用数字技术方法开展史学研究——包括数字化研究对象和技术化研究方法的采用。数字史学确实具有很强的技术属性,但绝不仅仅是一套工具,而是史学研究的新思维和新拓展。量化统计、空间与网络分析等数字史学所擅长的部分恰是传统史学较薄弱处。正因如此,不管采用形态多新颖的材料还是使用多先进的技术手段,数字史学所要研究或解决的依然是史学问题。数字史学必须对话传统史学和融入传统史学,才能在真正丰富和提升传统史学研究的同时,完成自己的价值追求。换句话说,虽然数字史学具有很强的跨学科特征,是科学与人文交融的产物,但跨学科是对拓展学科边界的努力,是针对学科过往难以解决的学术问题的一种新尝试,而不是寻求“大帐篷”式的简单化的学科综合。简而言之,数字史学是历史学之一种,历史学是数字史学的本位学科。

当然,强调数字史学是史学,并非要矮化数字技术的价值或狭隘地认为数字史学只是史学研究对数字技术的借用。实际上,数字史料平台和数字史学研究成果已经为诸多社会科学甚至自然科学提供了学术资源或用武之地,应当给予数字史学必要的灵活性与兼容度,以更好发挥其学术价值。但这只是数字史学跨学科建设的目的而非不明确学科定位的理由。跨学科的数字史学不是没有学科,而应是基于数字化史料,采用数字技术方法回应史学研究问题的史学新领域。如果不明确学科定位,就无法明确要实现怎样的研究目标和遵循何种的研究规范。当下不断出现的“不明所以”或充满“炫技”的“数字史学”研究,不公布底层数据,从过程到结论都无从验证的“数字史学”研究等乱象,<sup>②</sup>都是目标与规范不明确的表现。这些问题的存在,不仅影响了传统学科对数字史学的观感与接纳,更阻碍了传统与革新之间的对话和相互砥砺。受此现象影响,很多传统史学家对所谓能力巨大的技术方法形成了一种“人力愈多而天理愈蔽”的糟糕印象,<sup>③</sup>学科跨得越远,学术对话越没有可能。由此,无论从融入当下的学术体系,还是实现研究成果的学术价值等角度考量,学者们都需要为具有很强跨学科属性的数字史学明确其史学定位。

## 二、从发现材料到创造材料

既然数字史学是史学之一种,开展数字史学研究当然应遵循史学研究的根本要求。史学研究的具

① 乔·莫兰:《跨学科:人文学科的诞生、危机与未来》,陈后亮、宁艺阳译,南京大学出版社2023年版,第212—213页。

② 或是对此问题的反思与回应,2024年度北美社会科学化历史研究协会(Social Science History Association)的年会主题便是“Trust and Distrust of Historical Sources in the Digital Age”。参见 [https://ssha.org/files/2024\\_SSHA\\_CFP.pdf](https://ssha.org/files/2024_SSHA_CFP.pdf) [2024-03-05]

③ 王守仁:《王阳明全集》上册,上海古籍出版社1992年版,第28页。



体方法灵活且多样,但终极追求或根本精神是明确且唯一的——“求真”。尽管在人类创造出时空穿梭技术之前,简单绝对的真实再现并无可能,但在科学精神的指引下尽可能客观探求真实则是可能的,这也就形成了今日史学客观化研究的基本路径——必须凭借史料,才能呈现或阐释历史。因此,科班训练出身的历史学者,自入门起就需谨记历史研究的不二法门——“上穷碧落下黄泉,动手动脚找材料”。<sup>①</sup> 史学研究必须围绕史料展开,“搜集、鉴定和整理史料的方法,分析和综合史料的方法,以及编写史学论著的方法,是一切历史研究的基本方法”,而“史料是否完备和正确,直接关系到历史研究成果的大小和好坏”。<sup>②</sup> 早在1889年,美国历史学家查尔斯·亚当斯(Charles K. Adams)就指出:“目前,世界上没有任何完备的历史教学不是建立在……德国的研讨班一开始设定的细致、准确和细微地检查史料的基础之上。”<sup>③</sup> 直到今日,中国历史专业的大学生们反复学习、奉为圭臬的学科入门经典著作,如梁启超的《中国历史研究法》、杨鸿烈的《历史研究法》、严耕望的《治史三书》等,无不在反复教导从业者应该如何正确地搜集、鉴定和整理史料,就是因为唯此方能做出规范和有学术价值的研究。

通过史料去认识历史方能客观和符合科学精神,是因为整理和鉴别后的史料,特别是原始史料,被认为是真实可靠且并非为研究而被保存下来的,具有客观性和独立性,从而能够“言之有据”。因此,当爱德华·霍列特·卡尔(Edward Hallett Carr)在《历史是什么?》中指出历史是“历史学家与历史事实之间连续不断的、互为作用的过程,就是现在与过去之间永无休止的对话”时,埃尔顿(G. R. Elton)随即反驳道,历史学的独特之处在于它重视严谨的学术研究和关注原始资料,“真理战胜一切”。<sup>④</sup> 历史学者对史料,尤其是原始史料充满了尊敬。与一般的社会科学可以根据研究需要,通过调查、访谈等设计进而创造出研究材料不同,史料主要是过往人类生活过程中被遗留和保存下来的文献与实物,它的形成与当下研究无关,这既使得史料具有超越其他研究材料的客观性,但也限制了历史学者的作为——历史学者只能不断发现和挖掘新史料,不能够发明或创造新史料。傅斯年在主持中央研究院历史语言研究所时高度重视发掘新史料,曾费巨资购买清代内阁大库档案用于研究。但在研读后,他却未能从这批被抱以厚望的原始史料里获得什么重要的学术发现,因而颇为失望。李济则调侃他:“什么叫重要发现?难道说先生希望在这批档案里找出满清没有入关的证据吗?”<sup>⑤</sup> 傅斯年听闻此言不禁大笑,因为他建设历史语言研究所的理念之一便是要“一分材料出一分货,十分材料出十分货,没有材料便不出货”,<sup>⑥</sup> 历史学者只能通过发现史料去再现历史,而不能随意发明以求出奇出新。

数字时代,历史学者却能够借助数字技术对史料进行“改头换面”般的改造,形成原本史料中不存在的数据集合。刘萍认为数字技术带来了“史料革命”,计算机及数字化技术“推动历史资料在存储、检索、传递和分析、处理等方面发生重大变革”。<sup>⑦</sup> 尹媛萍指出,互联网时代的社会生活将“改变

① 此处化用傅斯年“上穷碧落下黄泉,动手动脚找东西”,参见傅斯年:《历史语言研究所工作之旨趣》,《国立中央研究院历史语言研究所集刊》第一本第一分,1928年10月。

② 吴泽:《史学概论》,安徽教育出版社2000年版,第8、158页。

③ 林·亨特:《历史学为什么重要》,李果译,北京大学出版社2020年版,第55页。

④ 乔·莫兰:《跨学科:人文学科的诞生、危机与未来》,第133页。

⑤ 李济:《傅孟真先生领导的历史语言研究所——几个基本观念及几件重要工作的回顾》,张光直主编:《李济文集》第5卷,上海人民出版社2006年版,第168页。

⑥ 傅斯年:《历史语言研究所工作之旨趣》,《国立中央研究院历史语言研究所集刊》第一本第一分,1928年10月。

⑦ 刘萍:《“史料革命”:近十年来的史料学研究及反思》,《北方论丛》2021年第5期。

史料的形态、影响现有的学术生产机制、缔造全新的历史学家”。<sup>①</sup> 数字技术能改变的绝不只是史料形态,更可以通过各类史料的组合或联结,深度匹配史料中的各项信息,形成原始史料本不存在的信息集或数字平台,进而为过往缺乏史料或人力所不逮的研究问题提供可能。信息学科针对关系型数据库 (relational databases) 的数据整合与分析,提出可以进行数据的内外两种连接,而对于数字化后的史料,同样也可以进行内外连接,形成新的更能服务研究问题的数据平台。史料的“内连接”,指即便是文献型的数字史料库,技术方法也能够接受研究者的指令,在其中收集研究所用的信息并形成数据集合,供研究者使用。“外连接”则指技术也可以突破单一的史料库,在各种不同的数据库中收集信息,构建分析材料。这种连接,常常能够帮助研究者获取多代际、长时段和不同地区的可比较信息,为分析群体的历时变迁等提供难得的资料平台。

香港科技大学李中清、康文林团队建设的“中国多世代人口数据库(双城)”是“内连接”形成新史料平台的典型。<sup>②</sup> 黑龙江双城地区清代八旗政府不仅每年都以村庄为单位,按户开展人口登记,同时每五年还以村庄为单位,对土地的所有者、类型、面积及部分土地的位置等信息都进行了记录。该团队首先确认了七个年份 23 个土地登记册记录的 19609 条地块和地块拥有者的信息,再将这些年份的土地所有权信息和同年或相邻年份的人口登记信息匹配,其中 13155 条户与人口记录能够借助电脑整合技术和婚姻关系信息在户口册中定位出来。它提供了 1866—1913 年生活在今黑龙江省双城地区,超过十万名隶属八旗组织的居民个体、家庭及其他相关方面的一百三十多万条记录,成为研究该地区历史上代际遗传 (inter-generational transmission) 问题极好的史料平台。<sup>③</sup>

笔者领导开展的“民国清华留美生职业数据库”或可为“外连接”提供示例。该项目首先依靠的还是传统结构性史料,即 1937 年《清华同学录》和此后历年清华公费考选生名单,建立完整的民国清华留美生名录;再根据 1917—1948 年间多种《清华同学录》,尽可能拓展每一位留美生的职业信息;然后利用 python 语言编写爬虫程序,在各类开放文献数据库中检索每位留美生信息,<sup>④</sup>收集以后进行人工比对,确认各项信息一致后输入数据库。<sup>⑤</sup> 数字技术使得在海量文献数据库中进行史料连结成为可能。爬虫程序 24 小时不停歇地在各数据平台按照给定姓名查找信息,先后自动检索到 11529 条潜在信息。通过人工比对,确认其中来自 3928 种不同材料的 9112 条确为研究所需,占比 79.04%。机器+人工的路径高速且有效地实现了研究目标——1356 位,约 96% 的留美生由此具备至少一项职业信息,相比《清华同学录》只有六成学生的职业记录是一个很大提高。有 61 位留美学生没有采集到职业信息,多数可能是无职业可采集,其中有 30 位因各种原因,或亡于留美期间,或亡

① 尹媛萍:《未来已来:互联网历史学蠡测》,《中国社会科学评价》2021 年第 1 期。

② 读者可通过以下网页链接获取该数据库: <https://www.icpsr.umich.edu/icpsrweb/DSDR/studies/35292>。数据库的详细介绍和使用指南,可参见该数据库网页“Data & Documentation”栏“Download”中的“User Guide: Wang, Hongbo, Shuang Chen, Hao Dong, Matt Noellert, Cameron Campbell, and James Z Lee, 2013, China Multi-Generational Panel Dataset, Shuangcheng (CMGPD-SC) 1866 - 1913”。

③ 所谓代际遗传,亦称代际传递,是社会科学家从生物学领域中借来的术语,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,是理解社会流动或社会不平等的关键所在。

④ 信息收集的主要文献数据库,包括上海图书馆“全国报刊索引数据库”、中国知网“中国工具书网络出版总库”、“读秀学术搜索”“抗日战争与近代中日关系文献数据平台”、台北“中研院”近代史研究所“近代人物资讯整合系统”和 East View 公司“晚清和民国时期中文报纸集”等电子文献库。

⑤ 为方便核对资料和检索采用史料信息,研究团队为每一位留美生建立了单独的备查文件,专门记录其职业信息等依据的史料来源和相关信息。

于归国后不久;有三位留学期间已有精神问题。不能确认是否有工作但未采集到信息的,只有26人,约占全部留美生的1.8%。此外,511位留美生追踪到10个以上职业信息,占全部留美生的36%;追踪到三个以上职业信息的留美生1172位,占全部留美生的83%。

职业生涯是个人“长时段”的“大历史”,研究者必须尽可能多地获取不同时间节点信息,形成连贯的数据链,才能形成动态和有说服力的研究。但是,面对有一定规模的群体,完全依靠人工来获取较全面的数据链,难度很大。过往历史学者收集、整理和分析史料的工具有限,所能依靠的往往只有人工,尽管对清华留美生的研究已经很多,但群体性的系统和连贯的职业信息平台的缺失,使得群体性职业发展研究一直空白。该数据库的建设,不仅为动态研究清华留美生学有所成后的职业发展状况提供可能,也为突破单一结构性史料的记载局限,依靠技术手段收集和分析群体多节点关键性信息,提供了机会和范例。

总的来说,对历史数据信息进行各种连接的最终目的,是希望将存储于各种数据库中的历史信息从静态推向动态,实现“数据追踪”。如,利用多种材料或不同数据库的互相连接,可以连缀起某个人和某些群体不同时间节点的信息,形成一种动态化的“追踪数据”,这不仅改变了传统史料的形态,更直接形成了原本不存在的研究数据,可谓发明新材料。因此,构建数据库进行史学研究,其优点并不只是“广”而“全”地使用史料,更在于拥有“专”且“深”的新分析数据。

最后,同传统史学研究需要对史料进行谨慎考证一样,数字化的史料处理与利用同样必须对史料进行考证。任玉雪指出,在微观层面,“量化数据库的考据从选择史料的时候就已经开始了”,“和所有的描述性史学研究一样,首先要对文献进行考据,辨别真伪”,“史学考据的学术规范,都是建立量化数据库必须遵循的规则”。但在总体层面,“大规模数据的考据与个体数据的考据有一些差异”,“对于描述性史学来说,如果某个数据或人物的记录存在错误,那么这个史料可能完全没有价值了,但对于系统性的大规模数据来说,存在少量的错误是在所难免的,需要从整体上考察史料的价值”。<sup>①</sup>

### 三、从解释历史到发现历史

针对史料的种种努力,目的自然在于推动史学研究。史学当如何研究,学者们的看法差异很大。黄侃曾说:中国学问的方法,不在于发现,而在于发明。<sup>②</sup>他认为,挖掘新史料、新事实的方法是“发现”之法;通过探索既定材料与事实,揭示客观规律,创造新理论的方法是“发明”之法。<sup>③</sup>有学者认为,这反映了黄侃与王国维治学的差异。黄氏主发明之学,主张用传统方法处理新、旧史料,善于从常见史料中发掘出新东西,重视旧材料重于新材料;王氏主发现之学,提倡“二重证据法”,主张用新方法处理新材料,重视新材料胜过旧材料。<sup>④</sup>我们或可借用黄侃的句式,提出数字史学的目标与价值:既在于发明新史料,更在于发现新史实。史料无所谓新旧,重要的是能够根据研究题目,借助数

① 任玉雪:《量化数据库的构建、数据清理与史学考据方法》,“史学青年工作坊16·数字人文与史学研究视野的拓展”,会议论文,上海交通大学,2019年12月7日。

② 吉川幸次郎:《我的留学记》,钱婉约译,光明日报出版社1999年版,第94页。

③ 杨晓、黄亚栋:《黄侃“在于发明,不在于发现”治学原则的创新意蕴》,《辽宁师范大学学报》2014年第5期。

④ 司马朝军、王文晖:《黄侃年谱》,湖北人民出版社2005年版,第254页。



字技术创造性地挖掘、组合成新的数据化史料平台,再通过分析此类数据平台,发现诸多传统史料本身没有直接提供、传统阅读方法也无法获知的新史实。这样的史实发现,自然会帮助历史学者形成新的史学认识和学术理论。

人文学术应重发明而非发现,也在一定程度上反映了长久以来对人文与科学研究之差别的刻板认识——人文学科的研究重在通过解释已知现象来探寻规律,是一种解释型学术;科学研究的重点是通过发现未知新物质,获取新规律,是一种求是型学术。从16世纪“科学革命”到19世纪近代学科体系逐步形成,科学与人文两大学科在体系不断完善的同时,畛域亦不断加深。1959年,英国化学家查尔斯·珀西·斯诺(Charles Percy Snow)以《两种文化》(*The Two Cultures*)为题,深刻地指出了人文与科学看似无法弥补的隔阂:“科学研究面对未来,而人文研究则是对于过去的回应。”<sup>①</sup>人文研究重解释和主观,科学研究重发现和客观的标签也不断被强化,并影响到各自学科自我标准的认定。德国历史哲学家威廉·狄尔泰(Wilhelm Dilthey)认为,世界永远不可能被客观和彻底地认识,而只能被解释。<sup>②</sup>但是,作为研究人类过往的历史学,虽归属人文学科,但求真不仅是其最高原则,也是学科的唯一标准,先实事再求是的理路和科学研究有相通处。前贤已指出,历史学是具有人文和社会科学多维面向的,不宜简单地归类为人文或科学。何兆武认为,历史学有两个层次,一个是对史实的认知,一个是对史实的理解和诠释,前者是科学的,后者是人文的。历史学是一种人文知识(*geisteswissenschaft*),而不是自然科学(*naturwissenschaft*)意义上的那种科学,虽然性质有别,但两者应该有共同的科学规范、纪律或准则。<sup>③</sup>因此,尽管几乎所有的学科都要做解释工作,但历史学必须以求真为标准,解释不能仅建立在理论、价值观和政治立场之上,必须基于客观和广泛的史料掌握、严格的史料考订和规范的史料解读。茅海建认为,中国近现代史研究长期处于“主题先行”的发展轨迹中。很多结论尽管影响广泛,多年来学者们不断对其进行阐释,但其“所依托的史实却是在匆忙中搭建,根基并不是很深”。<sup>④</sup>所以,考订史实,发现史实依然是历史研究的重要基础。甚至可以说,从浩瀚的史料中不断发现新史实,是历史研究更为重要的工作,是其作为基础文科的价值所在。

通过构建数字史料平台,数字史学在发现新知上完全能够大有作为。20世纪90年代初,科学家们已经完成了以计算机为工具进行数据库分析和知识发现(*knowledge discovery*)的基础技术工作,在逐步开发出各种数据挖掘工具的同时,<sup>⑤</sup>证明该技术方法对多个学科的研究都非常有效,他们相信21世纪会是这种研究方法大爆发的时代。<sup>⑥</sup>今天,以量化数据库构建和研究为代表的数字史学,确实在发现新史实方面取得很多突破,大量隐匿于史料之中的历史过往得以被呈现。量化历史数据库的构建与研究不仅可以被视为数字人文研究的重要组成,更得益于数字人文的发展,其学术价值也不断得以提升。一方面,量化数据库研究需要使用统计分析、GIS系统等数字化的研究手段和分析

① Charles Percy Snow, *The Two Cultures*, Cambridge University Press, 1959, p. 6.

② 乔·莫兰:《跨学科:人文学科的诞生、危机与未来》,第132页。

③ 何兆武:《对历史学的若干反思》,《史学理论研究》1996年第2期。

④ 茅海建:《戊戌变法史事考》,生活·读书·新知三联书店2005年版,“自序”,第1页。

⑤ 有关数据挖掘技术的发展历程和研究应用,参见 Jiawei Han、Micheline Kamber:《数据挖掘:概念与技术》,范明、孟小峰译,机械工业出版社2007年版。

⑥ William J. Frawley, “Gregory Piatetsky-Shapiro, and Christopher J. Matheus, Knowledge Discovery in Databases: An Overview”, *AI Magazine*, Vol. 13, No. 3, 1992.

工具,与当前的数字人文研究在技术方法和研究理念上有很多共通处。另一方面,当前数字人文虽然已经发展到包含文本挖掘、社会网络分析、GIS系统研究及可视化呈现等多个方面,但无论文本挖掘、语意分析还是社会网络分析,其理念都是通过技术手段,以电脑代替人脑,处理繁复、甚至有一定规律的数据。比如,文本挖掘的核心理念是词语计数(counting words),即计算一个特定词语在某个文档中出现的频率。在词语计数之上,学者们也注意到词语与词语之间的表达结构、词语在语境中的意义等问题,并逐渐发展出词组频率分析和语境关键词(keyword-in-context)分析等技术方法。至于社会网络分析,目前常用Gephi等分析软件,将研究对象抽象成点,彼此关系由连接的边来表示,同时边可以被赋予方向和权重。当研究对象具备一定规模后,软件就会呈现出复杂的点线关系与结构图。社会关系网络图常常呈现多个大小不等的交往中心点,从而为历史学者理解复杂群体中的关系脉络和核心节点提供帮助。可以发现,量化分析同样是这两种技术手段的核心,是数字史学的重要基础。亚当·克林布尔(Adam Crymble)就直截了当地指出,对数字史学最合理的描述应当是定量研究与人文计算结合的产物。<sup>①</sup>

依靠量化分析上文提及的“中国多世代人口数据库(双城)”数据库,美国艾奥瓦大学历史系的陈爽撰写了《国家资助的不平等:八旗制度与中国东北的社会阶层化》(*State-Sponsored Inequality: The Banner System and Social Stratification in Northeast China*)一书,<sup>②</sup>梳理了土地财富在双城社会中的分配状况和流转的历史,进而发现该地区清代国家主导的社会工程直接介入了财富分配,并在与地方实践的互动下,最终形成持久的不平等的社会经济。该书出版后受到学界的关注和肯定,荣获2017年度美国图书馆协会颁发的卓越学术著作奖(Winner of the 2017 Choice Award for Outstanding Academic Title, sponsored by the American Library Association)。此外,笔者曾负责建设包含33所高校近十二万名大学生的民国大学生量化数据库。分析数据库中民国大学生地理来源,发现其高度集中于江浙、广东、福建等近代开风气之先的极少数省份,大多数省份的大学生数量竟可以忽略不计,与清代科举成功者的地理来源相差巨大。清代的人才选拔是朝廷统一控制下的“天下选才”模式的结果,而民国大学的招生主要由各大学自行负责,国家的统一调控极少,这在导致省际间精英人才供给严重失衡的同时,也留给学者很多理论探讨的空间。<sup>③</sup>

尽管历史学家强调史料考订,依靠第一手材料还原历史,但大量的史实并不能简单地经由史料“不言自明”(speak for themselves),历史学者必须依靠各种方法进行文本解读。开展各种数字化解读,可以用完全不同的形式组织和使用材料,从而发现很多传统文献阅读所无法呈现的新史实。以有冲击力的新发现史实为基础,进而再理解或诠释之,先实事再求是,本就是历史学应有的研究路径,也即桑兵认为的,历史虽然是自然科学和艺术的结合,但历史研究要“以实证虚,诚为要言哉”。<sup>④</sup>历史学解释的发展、进步,必须以史料的新发现、新解读为起点。德国马克斯·普朗克科学史研究所的薛凤(Dagmar Schäfer)就认为:“历史学通过数字人文重组和呈现,其中很重要的一环就是对史料资源的数据分析和结构分析,最大程度解析了史料的产生原因和背景”,“把历史数据转到一个新的

① Adam Crymble, *Technology and the Historian: Transformations in the Digital Age*, pp. 44 - 45.

② Chen Shuang, *State-Sponsored Inequality: The Banner System and Social Stratification in Northeast China*, Stanford University Press, 2017.

③ 具体研究参见梁晨、任韵竹、李中清:《民国大学生地理来源量化考析》,《近代史研究》2021年第3期。

④ 桑兵:《傅斯年“史学只是史料学”再析》,《近代史研究》2007年第5期。



结构或做成一个新的数据库时,我们也同时会对该历史事件有一个新的了解,发掘出新的意义”,“像从文本中提取数据,以新的方法排列并重组数据,这些重组和展示工具定义了整个时代,就是 new database construction 的时代。”<sup>①</sup>

数字史学的首要目的在于发现新史实,这使得它虽然重视量化方法,但与经济学、社会学等学科的量化工作存在重要差别。社会科学的量化分析,主要目的是证明不同因素之间的因果关系。作为社会学创立者之一的埃米尔·涂尔干(Émile Durkheim),其雄心壮志之一,便是用社会因果关系的概念把整个人文社会科学统一在自己的“权杖”下。<sup>②</sup> 经济学家同样依赖数据证明因果问题,并建立起计量经济学(Metrics)。<sup>③</sup> 过往的计量史学也曾受此影响,强调复杂、严格的统计方法和分析模型的使用,这使得部分研究曲解或误用史料,研究结论常看似“高深”,却实则“无根”。今天的数字史学,尽管工具选项更多,分析能力更强,但重要的并不是借助数字工具得到什么问题的结果,而是要“产生问题”,<sup>④</sup>即发现新史实。因此,数字史学所要开展的研究计算,更多应是描述性的,而非探索性或验证性的。通过数据计算发现各类现象后应回到史料,回到社会环境、制度以及文本语境中去进行理解,而不是继续单纯地依靠数理方法,人为限定各类干扰变量以解答问题。

### 余论:人机合作,赢得未来

将统计技术引入历史研究可能是现代史学与技术最早的一次结合。这个结合过程长期且反复,只有在计算机技术足够发展后才出现了越来越多的重要行动者和经典著作。傅斯年早就指出史学工作的核心有两个,一曰到处找新材料,二曰用新方法(科学付给之工具)来整理材料,以努力达成“聪明的考证”。<sup>⑤</sup> 梁启超可能是中国最早倡导用科学方法来“聪明”地研究历史的学者。1922年,他在东南大学的演讲中就提出“用统计学的法则,拿数目字来整理史料,推论史迹”,提倡研究历代战乱统计表、异族同化人物表、地方统治离合表、历代著述统计表、历代水旱统计表等。<sup>⑥</sup> 其后,他的学生卫聚贤不仅在大学讲授历史统计学课程,还编写了《历史统计学》教材。<sup>⑦</sup> 不过,或许是计算机技术尚未出现的原因,这些努力效果不大,投身其中的行动者更是寥寥。

随着计算机的出现和计算能力的不断进步,量化史学研究随之取得很大进步。法国年鉴学派第三代代表人物皮埃尔·肖努(Pierre Chaunu)认为,计算机的运用推动了系列史和计量史的发展,成为历史学“此前和此后的分界线,它把过去的研究统统归入考古学的领域”。埃马纽埃尔·勒华拉杜里(Emmanuel Le Roy Ladurie)更是豪迈地预言:“明天的历史学家都将是程序设计员,否则就不再是

① 此处所引内容为2020年6月北京论坛云端国际论坛系列之“数字人文视角下的中国历史研究”中薛凤教授所作的“新方法开辟新领域,历史研究中的数字人文”主题演讲。此论坛由北京大学数字人文研究中心组织,笔者为此次论坛的主持人之一。参见 <http://pkunews.pku.edu.cn/xwzh/28b8868a55e7407a9dce38b95134059d.htm> [2021-06-23]

② 弗朗索瓦·多斯:《碎片化的历史学:从〈年鉴〉到“新史学”》,马胜利译,北京大学出版社2008年版,第14页。

③ 乔舒亚·安格里斯特、约翰-斯特芬·皮施克:《精通计量:从原因到结果的探寻之旅》,郎金焕译,格致出版社2019年版,第1页。

④ 苏珊·施莱布曼、雷·西门子、约翰·安斯沃斯:《新数字人文导论》,第364页。

⑤ 傅斯年:《历史语言研究所工作之旨趣》,《国立中央研究院历史语言研究所集刊》第一本第一分,1928年10月。

⑥ 梁启超:《历史统计学》,《史地学报》第2卷第2期,1923年1月。

⑦ 卫聚贤:《历史统计学》,商务印书馆1934年版。

历史学家。”<sup>①</sup>尽管这些预判有将计算机吹捧成神谕之嫌,但至少20世纪60年代以来,量化史学确实取得很大进步,研究范围更是从以经济史为主,发展到历史地理、人口史、政治史和社会史各方面。仅就中国史而言,1962年何炳棣的《明清社会史论(1368—1911)》(*The Ladder of Success in Imperial China: Aspects of Social Mobility, 1368—1911*)、1988年毛汉光的《中国中古社会史论》、1999年李中清和王丰的《人类的四分之一:马尔萨斯的神话与中国的现实(1700—2000)》(*One Quarter of Humanity: Malthusian Mythology and Chinese Realities, 1700—2000*)及2014年谭凯(Nicolas Tackett)的《中古中国门阀大族的消亡》(*The Destruction of the Medieval Chinese Aristocracy*)等先后问世,在解决了诸多中国古代史中的硬骨头问题的同时,也分别成科举与社会流动、中国人口史和古代阶层变迁等方面的经典著作。

今天,史料数字化开展得如火如荼,随着可机读史料的飞速增加,依靠信息技术解读史料成为必然。在此过程中,“计量”“量化”“人文计算”等词正在逐步退出,数字化或数字史学随之兴起。<sup>②</sup>从“人文计算”到“数字人文”,不仅是简单地变换主语与修饰语,而是力图将这一领域的努力重点从“计算”转换到“人文”等具体学科领域,即:重要的不是研究者有多高明的数字技术,而是数字技术如何能帮助包括史学在内的人文学科提升研究。或者说,数字是修饰,人文或历史等学科才是关键。在此过程中,数字史学、数字人文等要获得真正的学术生命力,就必须行动起来,产生出诸多的经典作品。数字史学、数字人文不能只是一个技术的平台,而必须强调史学等人文学科的主体性。以信息资源管理为代表的技术学科介入数字人文的目的是“如何从资源、数据、技术、平台、方法等方面为数字人文提供基础设施”,<sup>③</sup>而人文学科如果没有相应的行动,基础设施建设的意义自然会被消解,数字史学的未来必然不乐观,甚至整个史学都有难以适应社会变化,无法保持“长青”的危险。

引入技术,当然不是将历史学装扮成“殷勤好客的女神”,以“新史学”的装束出现在公众面前,<sup>④</sup>而是当技术发展到人工智能时代,学科进步不仅需要学者与学者的人际合作,更需要学者与机器的人机合作。经济学家阿芙纳·格里菲(Avner Grief)在1997年就提出,随着拥有更强算力的计算机的出现,计量的潜力也会与日俱增。早期计量史学家所面临的技术限制将不复存在,学者们可以转向研究“更广泛的数据横截面、更长的时间序列,或者凭借以前的技术无法实现收集或分析的数据集”,这个预判与今日数字史学的功效和目标是如此的一致。<sup>⑤</sup>因此,如果说微观史是一台显微镜,<sup>⑥</sup>那么今日的数字史学很可能就是对撞机。更广泛和更长期的历史数据在这里不断碰撞与结合,会令历史学者观察到众多新史实和更好地平衡阐释与发现的研究范式,向着“求真”的目标不断迈进。如上,既是笔者对数字史学的期盼,也是与数字史学学人的共勉。

① 弗朗索瓦·多斯:《碎片化的历史学:从〈年鉴〉到“新史学”》,第173—174页。

② Adam Crymble, *Technology and the Historian: Transformations in the Digital Age*, pp. 23 - 27.

③ 夏翠娟、祁天姣、徐碧珊:《中国数字人文学术体系构建考察——基于实践项目的内容分析和文献研究》,《数字人文研究》2023年第4期。

④ 弗朗索瓦·多斯:《碎片化的历史学:从〈年鉴〉到“新史学”》,第176页。

⑤ Adam Crymble, *Technology and the Historian: Transformations in the Digital Age*, p. 26.

⑥ 乔瓦尼·莱维:《三十年后反思微观史》,尚洁译,《史学理论研究》2013年第4期。