数字技术与史学观念

——中国历史数据库与史学理念方法关系探析*

申　斌　杨培娜

【提要】 历史数据库是数字技术与史学观念相互激荡的产物，它既是史学研究方法的技术载体，又改进、创新着研究方法。典藏检索型数据库的研发基于传统实证史学理念和文献考据方法，量化分析型数据库的设计则植根于历史研究的社会科学化与统计学方法，而数据库又使这两种史学方法得到改进，发挥出最大效能。近年来，数字人文这一技术革命催生的数据库，给史学方法创新提供了可能，并且与注重文本脉络与历史脉络交错关系的当下史学趋势暗合。

【关键词】 史学理念　史学方法　历史数据库　数字人文

引　言

20 世纪 80 年代以降，信息技术在社会生活中的应用日益普遍，也日益改变着学术研究方式。近年来，数字人文更成为国内外学界的热门话题。①

数字技术在史学中的应用，主要体现在资料保存、公共展示、辅助教学、学术研究等多个层面。项洁从信息科技角度将数字人文与历史学研究的既有合作方式分为检索与计量两种，并指出基于数字人文理念和技术的数据库，可以给历史研究提供一个探究的环境和平台，而不仅是检索结果或者数据。② 历史数据库是数字技术与史学观念相互激荡的产物，它既是史学研究方法的技术载体，又改进、创新着研究方法。在借鉴项洁分类的基础上，本文从内容与功能角度，将以资料保存和学术研究为目的的中国史数据库划分为三类：典藏检索型数据库、量化分析型数据库和数字人文研究平台。③ 文章通过分析这三类数据库的设计理念与近代以来中国史学研究方法之间的关联，探讨数字技术与史学融合的问题，着重讨论：数据库的设计受到何种史学理念的影响？设计、使用历史数据库应从史学研究方法那里借鉴什么经验？数据库给既有史学方法又可能带来怎样的改变与补充？

*　本文是全国优秀博士学位论文作者专项资金资助项目（项目编号：201211）的阶段性成果。

① 关于数字人文的发展历程、我国史学领域对信息化进程的回响，参见王晓光《“数字人文”的产生、发展与前沿》，载全国高校社会科学科研管理研究会组编《方法创新与哲学社会科学发展》，武汉大学出版社 2010 年版，第 207—221 页；项洁、涂丰恩《什么是数字人文》，载项洁编《从保存到创造：开启数位人文研究》，台湾大学出版中心 2011 年版，第 9—28 页；牟振宇《数字历史的兴起：西方史学中的书写新趋势》，《史学理论研究》2015 年第 3 期；王旭东《信息史学的研究概要：定义、学术脉络和理论建构的主要内容》，载陈启能编《国际史学研究论丛》，社会科学文献出版社 2015 年版，第 275—295 页。

② 项洁、翁稷安《数位人文和历史研究》，载项洁编《数位人文在历史学研究的应用》，台湾大学出版中心 2011 年版，第 11—20 页。

③ 近日，赵思渊从构建方式角度将历史文献数据库分为数字化、数据化与文本挖掘三种形态。参见赵思渊《地方历史文献的数字化、数据化与文本挖掘》，《清史研究》2016 年第 4 期。

一、传统实证史学与典藏检索型数据库

中国史学有着悠久的文献考据与史实考辨传统。北宋司马光的《资治通鉴考异》和南宋王应麟的《困学纪闻》标志着我国史学考据走向成熟。及至清代，从顾炎武到钱大昕，传统实证史学技法达到顶峰。[①] 以此为基础，20 世纪 20 年代胡适倡导的整理国故运动和傅斯年在中研院史语所开展的事业，通过引进欧洲近代实证史学方法，完成了以研究方法论为重心的第二次史学革命。[②] 傅斯年强调史学便是史料学，史料学方法核心是比较不同史料，求得近真与头绪。[③] 他一方面主张无限扩充史料范围，将考古材料、档案纳入史学视野；另一方面，明确提出要改读书为“找东西”。[④] 在传统时期，读书就是学问，而读书要依靠《四库全书总目》、《书目答问》等目录学著作指示门径。而在傅斯年倡导的现代史学研究观念下，寻找材料成了治史的首要工作，于是传统目录学知识就显得远远不够了，新的研究方式亟需相应的基础史料整理工作和辅助工具的支撑。胡适、傅斯年的总体理念虽然未必为多数学人认同，但是科学地收集、整理史料无疑已经成为当时史学界的共识。陈垣在 1929 年《中国史料的整理》演讲中就提出要“改良读书的方法，整理研究的材料，使以最经济的时间得最高的效能”。[⑤]

陈垣《中国史料的整理》演讲，很好地概括了纸本时代史料整理和工具书编纂工作的内容。无论民国时期还是 1949 年之后的大陆或台湾，基础史料整理工作大体都可以在陈垣开列的框架中找到。[⑥] 这也说明，这些史料整理要求并非陈垣个人意见，而是反映了现代史学研究的普遍需求。具体而言，史料整理和工具书编纂的进步主要包括：第一，古籍目录、索引大量编制，为研究者寻找图书、材料提供了线索。传统目录学重在“辨章学术，考镜源流”，以内容提要为主，附及版本。而此时科学整理史料思路下的目录索引编纂，更多是接受了欧美图书馆学传统下以检索为核心的书籍编目方式。这一转换最典型的体现，就是洪业主持下的哈佛燕京学社引得编纂处及其后继者中法汉学研究所通检组所进行的重要古籍索引编制。[⑦] 如陈垣所言，编制重要书籍索引的目的就是要让“什么人都可以不读史籍而能利用它们”。第二，古籍被以翻刻、影印、排印、断句、标点、校勘等方式整理出版，不但极大增加了古籍复本量，方便学者获取，更重要的是提供了易于理解的可靠文献版本，便于学者集中精力于自己感兴趣的问题。第三，提供基本历史事实的工具书（图书馆学中称为“事实型工具书”）被大量编纂。这类工具书以详实的史料考辨、史实考证为基础，将学术研究的权威成果以词

① 罗炳良《从宋代考据史学到清代实证史学的发展》，载罗炳良《宋史瞥识》，北京师范大学出版社 2011 年版，第 248—250 页。

② 黄进兴指出，清代考据学因符合史料考订、科学方法而成为引介西学的桥梁。王汎森则认为中国近代有三次史学革命：梁启超重在厘定“什么是历史”，胡适、傅斯年重在明确“如何研究历史”，马克思主义史学重在解决“怎样解释历史”。参见黄进兴《中国近代史学的双重危机》，载黄进兴《历史主义与历史理论》，陕西师范大学出版社 2002 年版，第 272—305 页；王汎森《晚清的政治概念与“新史学”》，载王汎森《近代中国的史家与史学》，复旦大学出版社 2010 年版，第 1—28 页。

③ 傅斯年《历史语言研究所工作之旨趣》，载傅斯年《民族与古代中国史》，河北教育出版社 2002 年版，第 467—478 页（初版于 1928 年）；傅斯年《史学方法导论》，上海古籍出版社 2011 年（原为 20 世纪 30 年代初北京大学讲义）。关于傅斯年史学与史料学思想的阐释，参见桑兵《傅斯年“史学只是史料学”再析》，《近代史研究》2007 年第 5 期。

④ 王汎森《什么可以成为历史证据——近代中国新旧史料观点的冲突》，载王汎森《近代中国的史家与史学》，第 103—139 页。

⑤ 陈垣《中国史料的整理》，载吴泽编《陈垣史学论著选》，上海人民出版社 1981 年版，第 244—252 页。

⑥ 如台湾的明实录校勘、内阁大库档案整理，大陆的二十四史和资治通鉴点校，中国历史地图集编绘。

⑦ 索引包括两种，一是针对人名、地名、书名、事物名等专门事项的，称为引得（index）；另一种是原文逐字索引，称作堪靠灯（Concordance）。参见黄永年《古籍整理概论》，上海书店 2013 年版，第 144—150 页。

典、表格、地图等形式呈现出来，使后来者可以高效地直接处理自己的研究主题，而不必总为枝蔓问题分散精力。第四，探索、完善了档案整理方法。如果说古籍整理尚有目录版本校勘等学问传统可以借鉴，那么作为新史料的公私档案整理则可谓无章可循了。20 世纪 20 年代初罗振玉等所进行的早期档案整理，仅是个别专题史料摘编，而非全面系统的整理。经过北大、故宫、史语所、一史馆等机构学人的探索，逐步确立了以档案原存机构为系统划分全宗分类和以档号、时间、职衔、责任者、事由、文种等为基本著录项的原则，有助于研究者从浩如烟海的档案文献中寻找自己需要的资料。①

典藏检索型数据库最初都是以既有纸本目录、索引、工具书为模板设计的，可以说是给传统研究技法披上了新的技术外衣。② 这种工作可以分为如下几类：

第一，图书馆、档案馆的古籍、档案数字化编目。简言之，就是把传统的卡片目录或纸本目录上登载的图书、档案著录项目作为元数据字段录入计算机，形成书目数据库和档案著录数据库，提高了检索效率。尤其是在档案文献的检索与内容揭示方面，数据库表现出得天独厚的优势，解决了纸本时代的诸多两难问题。比如原来档案编目面临的一个问题是如何处理全宗原则与内容揭示之间的矛盾，为了方便读者利用，只能根据每一个著录项各自编制一套卡片或纸本目录；而在数据库环境下，可以根据内容、类型（文种）、地域（归户）、时间（朝年）、责任者、职衔等著录项目方便地对档案自动进行聚类分析和关联分析，实现了分类与索引功能的融合，全面揭示文献内容和形式的多种属性。

第二，借助扫描、数码拍照，将纸张转化为数码图像文件，形成了对古籍、档案、报刊的图像文件进行存储和检索的图像资料库。最初是图书馆、档案馆为了保护古籍，减少对古籍原书调阅损毁而采取的措施，是古籍影印和缩微胶卷的自然延伸，仅供读者馆内使用。后来出现了商业或公益的古籍图像数据库，读者足不出户就可坐拥书城，进一步改变了文献获取方式，解决了文献稀缺性问题，提高了史料开放程度。

第三，通过 OCR 技术与人工核对的结合，全文检索资料库诞生了。③ 全文数据库一举取得了此前各类古籍索引梦寐以求的效果，既可以用于检索人名、地名、书名、事物名等专门事项，起到引得（index）的作用，又可以用于检索语料，起到堪靠灯（Concordance）的作用。数据库反过来成为编制索引的工具。④ 透过全文检索，研究者不但可以发现依据目录书发现不了的史料源，而且极大地提高了史料获取效率。

第四，事实型工具书被做成数据库、软件或插件。有的还与全文资料库配套使用，如 THDL（台湾数位历史图书馆）提供了包括中西历日期对照查询、清代官职表、苏州码转换器、度量衡单位换算系统在内一系列研究工具集。关于纸本时代史料整理方式、工具书与典藏检索型数据库的对应沿革关系，表示如下：

---

① 关于民国时期对清宫档案整理的探索，参见张会超《民国时期明清档案整理研究》，世界图书出版公司 2011 年版。共和国时期对档案整理取得的成绩，参见单士魁《中国第一历史档案馆》，《历史档案》1981 年第 1 期。

② 关于古籍数字资源的早期发展，参见杨朝霞《古籍数字资源述略》，《大学图书馆学报》2000 年第 3 期。

③ 关于中文编码等全文数据库的技术实现问题，参见谢清俊、林晰《中央研究院古籍全文资料库的发展概要》，《中文计算语言学期刊》第 2 卷第 1 期，1997 年；朱岩《谈古籍数位化》，载澳门图书馆暨资讯管理协会编《两岸三地古籍与地方文献》，澳门图书馆暨资讯管理协会 2002 年版，第 143—149 页。

④ 香港中文大学刘殿爵主持的先秦两汉魏晋南北朝一切传世古籍逐字索引丛刊，就是建立在汉达文库这一全文资料库基础之上的。

纸本史料整理方式、工具书与典藏检索型数据库沿革关系对照表

类别		纸本举例	数据库举例
书目	图书馆馆藏版本目录	《北京图书馆古籍善本书目》	国图 OPAC
	联合馆藏目录	《中国古籍善本书目》	高校古文献资源库
卷目篇名目录		《清代文集篇目分类索引》	明人文集联合目录及篇目索引资料库
索引	引得(index)	《二十四史地名索引》	汉籍全文资料库
	堪靠灯(Concordance)	《十三经索引》	中国基本古籍库
古籍、档案影印、缩微胶卷		《四部丛刊》《续修四库全书》	国图、一史馆馆内善本、档案电子图像阅览用数据库 四库全书图像版 国图数字方志
卷次篇名总目、分段分节、古籍标点、校勘		中华书局点校本“二十四史”	中华经典古籍库
档案编目		《内阁大库现存清代汉文黄册目录》、《清内阁旧藏汉文黄册联合目录》	中研院史语所内阁大库档案
报刊卷期篇目		《全国主要期刊资料索引》	全国报刊索引数据库
事实型工具书	历法	《中西回史日历》	两千年中西历转换工具
	地图	《中国历史地图集》	《中国历史地图集》软件
	地名	《中国历史地名大词典》	
	人物	《中国历代人名大词典》	人名权威资料检索系统

正因为此类数据库是按照传统研究习惯设计开发的,所以它们迅速得到了史学界的热烈欢迎,[①] 即便有批评意见,也是在肯定的大前提之下。[②] 全文数据库很快就展现出了其对史学研究的巨大功效,以致有学者提出了 e 考据的概念。[③] 但必须看到,这类数据库只是提供史料或者可供查考的基本史实,就研究方式而言,只是人工阅读被辅以电子检索,缩短了“找史料”的时间而已。要充分发挥这些数据库的作用,还需要使用者有足够的目录学、文献学知识,能够从史源学角度对检索到的史料作出取舍判断、恰当使用。校勘等传统治史技艺在数据库辅助下得以更加便利地运用。[④] 中国古代史领域较早深入接触计算机、数据库和网络的学者陈爽近期即指出,技术更新没有带来学术思维革命,数字化时代处理史料需要重温传统史学技艺。[⑤] 所以,与其说典藏检索数据库改变了旧有史学研究方式,还不如说借助新技术手段,把传统史学研究技法发挥到了极致。

二、历史研究的社会科学化与量化分析型数据库

如果说典藏检索型数据库带来了传统治学方式的增强升级版,那么量化分析型数据库则很可能

① 乔治忠:《历史研究电子资源运用的兴利除弊》,《史学月刊》2015 年第 1 期。
② 陈尚君:《〈中国基本古籍库〉初感受》,《东方早报》2009 年 8 月 9 日。
③ 黄一农:《两头蛇》,上海古籍出版社 2006 年版,第 64—65 页。
④ 何志华:《古籍校雠机读模式初探——兼论中国文化研究所汉达文库的另类功能》,载罗凤珠编《语言、文学与资讯》,(新竹)清华大学出版社 2004 年版,第 401—422 页。
⑤ 陈爽:《回归传统:浅谈数字化时代的史料处理与运用》,《史学月刊》2015 年第 1 期。

具有范式转化的意义，[①]其研究理念背景，是二十世纪前半期经济学学者运用统计学方法开展历史研究，以及20世纪后半期史学的社会科学化。

梁启超在提倡新史学的同年，就曾撰写《中国史上人口之统计》(1902年)这种试图以统计表形式梳理历史变迁大势的文章。[②] 但真正依照社会科学研究规范、利用统计学方法进行大规模史料整理和分析研究的，当推20世纪30年代北平社会调查所经济史组学人的工作。中国现代经济史学诞生于20世纪20—30年代之交的中国社会史论战，但当时只是引入若干概念对中国社会发展进行定性判断，缺乏基于史料的实证分析，更不必说量化分析了。要对中国社会经济整体性结构与趋势作出判断，就不能采取举例式分析，而需要借助大量原始史料，进行全面细密的定量研究。

这一时期，内阁大库档案的发现恰好为展开此类研究提供了可能，但海量史料又给传统史料利用方式提出了挑战。1930年开始，汤象龙、梁方仲、罗玉东、刘隽、郑友揆、千家驹、巫宝三等一批出身经济学、供职于北平社会调查所(后并入中央研究院社会科学研究所)的学者，开始大规模整理、抄录清代档案。[③] 因为他们是带着研究问题开展档案整理工作的，所以其档案整理方法与前述从陈垣到一史馆的整理方法不同。具体而言，就是结合清代的财政经济制度，拟定出待研究的题目，如地丁、厘金、关税、漕粮等，再围绕这些主题分工进行档案抄录工作，编出不同的专题史料汇编。

相应于他们重视对财政经济问题系统量化分析的研究理念，其工作最大的特点就表现为采取统计表形式对档案中的财政经济数据加以整理。根据研究问题需要和史料特性，舍弃原始档案中繁冗的文字表述，合理设计统计项目，采取集体工作方式从原始档案中提取有效数据，利用统计方法编制成一目了然的表格，这可谓在海量档案史料利用方式上的一个创新。正是以这些工作为基础，才产生了罗玉东的《中国厘金史》、汤象龙的《清季五十年关税收入及其用途》等专著。他们围绕“清季九十年全国粮价之变迁”这一主题，从故宫博物院所藏粮价清单中抄录数据编成粮价表，部分资料先为王业键的清代粮价资料库所采用，近年来更是全部排印出版，嘉惠学林。[④]

其实，除粮价外，中国社科院经济研究所图书馆还藏有题为《清代黄册》的293巨册清代财政数据统计表。20世纪50—60年代，梁方仲继续进行《中国历代户口、田地、田赋统计》的编纂工作。中国科学院经济研究所在其承担的中国近代经济史参考资料编研工作中，也率先出版《中国近代经济史统计资料选辑》。这均可视作20世纪30年代北平社会调查所开创性工作的直接后继成果。

而且，这一思路也被自然科学研究所采用。20世纪50年代，为配合社会主义建设，亟需掌握我国各地历史上的自然灾害情况作为参考。以地震资料为例，中国科学院地震工作委员会历史组首先摘取文献中与地震相关的史料原文，编出了《中国地震资料年表》(1956)这样的专题史料汇编，进而将其中的文字描述转化为震级，时间换算为公历，并且推定震中位置，编成《中国地震目录》(1960)这种专题史

---

① 梁晨、董浩、李中清《量化数据库与历史研究》，《历史研究》2015年第2期。关于量化历史研究，参见钱学森，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；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；李伯重《史料与量化：量化方法在史学研究中的运用讨论之一》，《清华大学学报》2015年第4期；彭凯翔《历史视野下中国经济的长期变迁——近年中国经济史之计量研究综述》，《经济研究》2015年第5期。

② 梁启超《梁启超全集》第四卷，北京出版社1999年版，第900—905页。

③ 关于这一群体的学术活动及其意义，参见陈峰《两极之间的新史学：关于史学研究会的学术史考察》，《近代史研究》2006年第1期。

④ 参见王砚峰《清代道光至宣统间粮价资料概述——以中国社科院经济所图书馆馆藏为中心》，《中国经济史研究》2007年第2期；罗畅《两套清代粮价数据资料的比较与使用》，《近代史研究》2012年第5期。

实参考书,作为研究地震分布规律的材料。这种思路为后来地理学、环境灾害史研究者所沿用。①

与 20 世纪前半期的量化历史研究主要表现为经济学者从事历史研究不同,20 世纪后半叶中国的计量史学则是历史学社会科学化与问题导向的史学风格产物。因应于 20 世纪 50 年代开始于欧美的史学社会科学化潮流,台湾地区 20 世纪 60—70 年代社会经济史和定量分析盛行,王业键主持建设的清代粮价资料库及其系列研究,刘翠溶利用族谱进行的历史人口学研究可为代表。② 20 世纪 80 年代计量史学被介绍到大陆,陈春声的《市场机制与社会变迁:18 世纪广东米价分析》和复旦大学历史地理研究所集体编纂的《中国人口史》即为翘楚。

在此转变过程中,数据库的出现为这类研究方法的大规模应用提供了利器。数据库不但可迅速进行各种统计运算,而且可以方便地进行不同变量之间的相关性分析,极大地提高了工作效率。早期的这类数据库,如王业键主持的清代粮价资料库,尚且可以看出较深的统计表的痕迹,随着研究问题的变化和数据库技术的发展,李中清团队开发的 CMGPD(中国多代人口系列数据库)、复旦大学历史地理研究所开发的中国人口地理信息系统(Chinese Population GIS, CPGIS)等数据库的功能日益多样化。量化分析型数据库给历史研究带来的改变可以概括如下:

第一,在史料整理层面,它可以有效地处理政府档案、民间文书等文本结构高度格式化且具有同质性的海量史料中记载的历史事实,③比如粮价单、契约、人口调查、缙绅录、科举题名录、学籍卡等,将其转化为结构化、数量化的信息。

第二,在研究层面,结构化、量化的信息表达形式简单明了,方便进行统计分析,便于史学家利用海量史料,也便于非史学研究者参与历史研究。而且,由于量化历史数据库提供的是数据而非文字描述,所以只需要将变量等极少数词语进行翻译,量化历史数据库就可以为不懂原始史料语言的研究者所利用,这极大地方便了从事长时间跨度、跨文化、跨地域、跨语言的研究。

第三,运用大规模数据统计分析,学者可以发现数据统计与传统记述性史料不同的历史面向,或者不同数据系统之间的差异,进而以此为起点,提出新的学术问题。例如王业键通过对清乾隆时期粮价的统计分析,发现清代官书中言之凿凿的"乾隆十三年米贵问题"其实很难成立,陈春声、刘志伟由此提出应该关注当时官员们的经济观念。④ 再如彭凯翔通过中国利率史数据库发现,刑科题本和民间契约两类史料中记载的同一地区利率变动趋势不同,前者起伏较大而后者更为平缓。他并没有止步于讨论孰是孰非,而是指出刑科题本中记载的利率变动可以作为反映政府利率管制强度的指标。⑤ 沿着彭凯翔的思路,量化分析型数据库不但在"数据系统"考证方面实现了传统史学考据无法企及的目标,而且还蕴含着分析文本脉络和历史脉络彼此交互关系的可能性,而这正与 20 世纪 70 年代以来国际史学趋向和数字人文理念暗合。如果没有大规模量化数据库,前述研究都是不可能的。

正因为此类数据库对历史研究具有范式性改变的可能,所以研发、使用时需要注意的问题也就比典藏检索型数据库更为深刻。反过来,这其中也蕴含着丰富既有史学研究方法论的契机。

---

① 从 20 世纪 80 年代的《中国近五百年旱涝分布图集》,到近年来方修琦、夏明方等自然地理学、环境史学者对气候变迁、灾害的数据库建设,都明显可以看出思路的延续。

② 刘翠溶《明清时期家族人口与社会经济变迁》,"中央研究院经济研究所"1992 年版。

③ 构建量化历史数据库的前提条件是:第一,史料文本是格式化的,可以准确、方便地提取史实信息;第二,被提取的数据也是结构化。

④ 陈春声、刘志伟《贡赋、市场与物质生活》,《清华大学学报》2010 年第 5 期。

⑤ 陈志武、彭凯翔、袁为鹏《清初至二十世纪前期中国利率史初探——基于中国利率史数据库(1660—2000)的考察》,《清史研究》2016 年第 4 期。

首先，由于粮价单、契约、账本等史料的原始性，常常让人们以为此类史料记载的内容就是历史事实（这一理解确有其合理性）。汇集这些事实的量化历史数据库，可以使研究者跳过史料收集、考辨等繁琐的历史研究第一层面的工作，而直接在第二层面开展研究，即历史事实分析上。但是，无论史料源多么原始，量化数据库中的数据都是有待考证和阐释其数字意义的史料，或者说只是历史上某一种观察视角下看到的“事实”。尤其是目前的量化历史数据库，无论其所包含数据量多么大，一般都是以单一类型史料为数据源搭建的，或为粮价单，或为科举题名录、缙绅录等。而某一类型史料的生成过程是具有选择性的，所以依据从某一类型史料（粮价单、登科录、学籍卡）提取数据所作出的研究，其实隐含着由于史料类型的局限而导致系统性偏差的风险。单靠数据“量”的扩充，量化历史数据库还是不能避免“集精选粹”的问题。必须利用来自不同性质史料（政府档案、民间文书、时人文集笔记）的数据互相参证，才能更大限度地保证结论的稳妥。

其次，数据是在特定的结构下产生的，不了解数据产生背后的结构与制度，就无法对数据进行合理阐释，也不知道将数据置于怎样的结构模型中进行实证分析。所以，设定分析变量、确定元数据标准时，首先必须充分考虑数据产生的政治经济社会结构与制度。在进行长时间跨度、跨地域研究时，如何处理结构性因素更值得深思。①

最后，提取数据的过程也是对数据去差异化，将数据与其所在文献史料脉络剥离，变成可以进行统计分析的同质性数字的过程，这个过程必然伴随着信息的流失。尽管这是要处理海量数据必须付出的代价，但在设计元数据标准时，还是要对史料文本做充分研读，确保被剥离的信息并非是严重影响到所欲分析变量关系的要素。

因此，数据库的设计研发需要社会科学家与史学家通力合作，在研究问题设计、概念界定、变量设定、史料（数据源）选择与提取方案等方面做周密思考，确保所设定变量涵盖了意欲分析问题的主要关联因素，所选定史料（数据）确实可以回答提出的问题而不存在严重系统性偏差，并且对史料（数据源）本身的形成背景（何时何地什么人出于何种目的怎样制造出来的，其中所记载数据的生成机制）、流传脉络（史料的固有系统性、完整性是否被破坏）、文献特性、数据特性以及数据提取方案做出详细说明、举例阐释，并且提示利用者使用该类数据进行分析时的限度（数据源本身有哪些局限，可以说明什么问题，不可以说明什么问题，应对研究结论做怎样的限定）。

三、数字人文与中国史学发展的新阶段

如果说前两种类型数据库的诞生，都是因应于史学研究需求，承袭纸本时代既有研究理念和方法而发展出来的功能相对单一的技术工具；那么哈佛大学、北京大学、台湾“中研院”合作的中国历代人物传记资料库（CBDB）和台湾大学项洁主持的台湾历史数位图书馆（THDL）②等数据库则可以说是由技术革命催生的、主要是基于数字人文理念设计出来的，具有反向刺激、推动史学研究观念和方法创新的可能性。数字人文理念认为，数字技术不仅可以提供保存资料的典藏手段和寻找资料的检

① 目前量化数据库运用最多也最成功的跨国比较领域，是与人为的制度设置关系最松弛的历史人口学。处理不同结构下数据的难度，可以 20 世纪下半叶国民账户体系（SNA）和物质产品平衡表体系（MPS）下的经济数据之难以相互比较为例。

② 参见项洁、陈诗沛、杜协昌《台湾古契约文书全文数据库的建置》，载《第三届台湾古文书与历史研究学术研讨会论文集》，逢甲大学出版社 2009 年版，第 243—269 页。

索工具，还可以协助研究者重新组织、分析资料，提供一个探索环境，成为一种人文学研究方式。下面就按照产生途径的不同，分别概述具有数字人文理念的中国历史数据库的情况。

第一类是经由典藏检索型数据库中的事实工具数据库功能扩展而来的。这样的数据库不再是单纯的史实检索工具，而是可以对知识进行重新组织的分析工具。如与原来的人名权威资料检索系统相比，中国历代人物传记资料库（CBDB）就可以从不同角度重组人物信息，不仅可以进行群体传记学的统计分析，还可以进行空间分析与社会关系网络分析。

第二类是经由全文数据库的功能扩展而来的。伴随着计算语言学的发展，自然语言处理技术（语义计算、文本挖掘）的进步，机器可以为研究者快速呈现出史料间的多重脉络或整体意义，帮助我们观察到隐藏于海量史料背后值得深入研究的现象，提供讨论的基础或可能的新角度，而不只是实现数据库设计者预定的功能。[①] 例如利用 bi-gram 词频统计，可让计算机迅速自动处理全文，既节省人力，又避免了研究者先入为主的干预，其最后呈现出来的结果常具有意想不到的相关性和延展性。

不过，数据库自动计算出的结果，只是呈现某种现象，而不能、也不该直接导向某一结论，现象背后的意义，需要人文学者的研究来阐释。[②] 这时候，数据库就不再只是一个检索工具或根据预设元数据回答特定问题的数据源，[③]而是协助学者发现议题、开展研究的工作环境。例如，陈诗沛利用《台湾总督府抄录契约》数据库绘制出契约的时间和空间分布，发现其空间分布与一般认知的台湾开发史顺序不同，这就提示研究者一方面去检视《总督府抄录契约》史料本身在形成中是否存在系统性偏差、流传中是否发生遗失以致信息失衡，另一方面去重新检讨关于台湾土地开发空间进程的成说。如果没有数据库帮助，单凭传统阅读方式，很难迅速识别出上述海量文献的总体特征，发现新问题。[④] 这种减少预设、半自动分析关联性的设计理念，与基于云计算等数据挖掘手段的“大数据归纳法”颇为相似，而与“问题——假设——收集数据——统计分析验证”的传统科学研究思路不同。在某种程度上，传统的经验归纳法借助数字人文技术和海量文献史料在更高层次上回归了。

第三类是多数据库整合形成的。数字人文的前提是存在大量可计算的基础数据对象，如数字、自由文本、格式化数据、图像、声音等，并且实现了数字化存储。目前不同数据库积累的历史基础数据对象已经足够多了，但尚未实现数据之间、数据库之间的有效通联整合。目前学术界的尝试有两种做法。

一是以地理信息系统（GIS）为平台整合多个专题数据库的数据，如中国历史地理信息系统（CHGIS）、中华文明时空基础架构（CCTS）、台湾历史文化地图（THCTS）等就整合进越来越多含有空间信息的专题数据。厦门大学郑振满设计的莆田历史人文地理信息系统，则是以 GIS 为平台整合文献（民间文献、地方档案、书籍）与田野调查资料（实物、建筑、仪式、音声），构成一个跨越史料文类、主题、数据类型的数字人文系统，也可以说是一个时空史料综合体。

另一做法则是通过应用程序编程接口（Application Programming Interface）技术，实现不同数据库之间、数据库与互联网资讯之间的通联。CBDB 之空间分析功能的实现就是建立在与 CHGIS 对接整合基础上的，而牛津大学魏希德（Hilde De Weerdt）与何浩洋开发的 MARKUS 系统在文本标记基础上

① 项洁、翁稷安《数位人文和历史研究》，第 18—19 页。由于自然语言处理技术应用于古汉语语料难度很大，所以全文数据库向具有文本挖掘功能的数据库转型过程中，目前出现了以中华经典古籍库为代表的一种折中型数据库，即采取文本标记方式，将人名、篇目、事件、地名、职官、纪年等专名单独分类标引，构建专项资料库，并且实现人名、地名异称关联检索。

② 项洁、涂丰恩《什么是数位人文》，第 19 页。

③ 项洁、翁稷安《数位人文和历史研究》，第 19 页。

④ 陈诗沛《资讯技术与历史文献分析》，台湾大学资讯工程学系博士学位论文，2011 年，第 78、87 页。

将不同词语分别与 CBDB、网络词典链接。

这一类数据库虽是由数字人文理念催生的，但其技术所支持的分析理路实与 20 世纪 70 年代以来国际史学界的转向暗合。经历后现代主义洗礼后，欧美史学研究的钟摆开始从科学一端向人文一端回归。史学家不再只把史料看作历史事实的载体，对史料的考察也不仅满足于考证其真伪和记载可靠性，而是更多地思考作为文本的不同史料自身形成与流传所蕴藏的社会文化意涵与过程，探究史料的文本脉络与社会历史事实建构之间的复杂关联，在某些领域这种分析甚至被作为主要研究课题。①

借助文本挖掘，我们可以揭示史料所处的多重脉络，②最重要的是有可能重建文本生成的社会脉络。比如梳理一件政务从起始到结束的完整行政流程，是了解官僚行政中资讯流通、探究政府运作机制的重要途径。但清代行政文书归档特点以及后世整理方式的多样性，围绕同一件公务的相关文书分散在不同全宗、目录之下且彼此相隔数月的情况绝不鲜见，因此要搜集围绕特定政务在皇帝、官员间往来讨论的全部行政文书并不容易。陈诗沛利用行政文书中的引用关系编制程序，不但从《明清台湾行政档案》数据库中提取出“左宗棠参李彤恩”事件相关的 23 件文书，而且半自动地生成了其引用关系图，揭示了详细的政务程序。③ 再如项洁团队在古地契关系自动重建问题上的探索已经取得显著成绩，使分析上下手契等文本联系成为可能。④ 赵思渊对中人、代笔所代表的信用机制与交易类型关系的分析，也可视作通过揭示文本脉络进而分析社会机制的典型作品。⑤ 数字人文理念下对多重脉络的寻绎，与目前史学注重文本脉络与历史脉络交互关系的转向可谓异曲同调。

数字人文方兴未艾，还没有充分展现出其对史学的根本性影响。这一方面是由于技术与研究积累不够，但更重要的是，目前历史学所研究的时代并非数字时代，作为数字人文处理对象的可计算的基础数据对象是我们通过扫描、OCR、编目等方式在研究过程中制造出来的。制造研究对象的过程本身已经是在某一研究理念下进行的了，要想依靠这样的数据彻底颠覆既有社会科学、史学认知，可能性自然不大。但是，当我们生活的这个时代——数字时代——成为历史研究的对象时，我们将面对着大量数字原生数据（born-digital data），而网络成为信息保存与传播的主要介质（这可能是未来主要的史料形态），我们要致力于探讨的社会事实首先可能表现为虚拟现实（virtual reality），历史学方法论或将会有一场根本性变革。这场范式性转换中，数字人文将大显身手，甚至促成科际整合，开创历史研究的新形态。⑥

（作者申斌，北京大学历史学系博士研究生；邮编：100871；
作者杨培娜，中山大学历史学系副教授；邮编：510275）

（责任编辑：黄艳红）
（责任校对：张文涛）

① 例如娜塔莉·泽蒙·戴维斯《档案中的虚构：十六世纪法国的赦罪故事及故事的讲述者》，北京大学出版社 2015 年版。

② 项洁、翁稷安《多重脉络——数位档案之问题与挑战》，载项洁编《数位人文要义：寻找类型与轨迹》，台湾大学出版中心 2012 年版，第 25—59 页。

③ 陈诗沛《资讯技术与历史文献分析》，第 95—141 页。

④ 黄于鸣《台湾古地契关系自动重建之研究》，台湾大学资讯工程学研究所硕士学位论文，2009 年。

⑤ 赵思渊《19 世纪徽州乡村的土地市场、信用机制与关系网络》，《近代史研究》2015 年第 4 期。

⑥ 程美宝、刘志伟《数字化时代的历史学教育》，《中国高等教育》2000 年增刊；陈春声《真正的学术群体应该“脱俗”》，《开放时代》2016 年第 4 期。

others have perceived the "internal conflicts" in his ideas, it is necessary for us to examine his philosophy against the background of modern science and philosophy and consider them a theoretical system with consistency and uniformity. What Fu Sinian opposed was the theory of history in the sense of modern rationalism. Meanwhile, he pursued the deductivist theory of history on the ground of modern science. Thus, the author argues, Fu Sinian's idea of history was not modern Western positivism based on inductivism, but a new form of positivism that emphasizes historical contextualism and deductivism. In this sense, his ideas transcend the subject and object dichotomy. In sum, we should not merely consider Fu Sinian as an advocate of objective historiography; instead, we should examine his whole theoretical system, for it was key to our understanding of Fu Sinian's scholarship.

Gu Jiegang's "Discussion of Ancient History" Movement and Its Relationship with Developments in Western Sinology //*Li Chang-yin*

The development of scholarship in modern China formed a close relationship with developments in Western sinology. Gu Jiegang's launching of the "Discussion of Ancient History" movement in the 1920s was a representative example. Hu Shi and Gu Jiegang proposed the notion that "there was no history before the Eastern Zhou dynasty," which was indebted to the historical skepticism by Philip Van Ness Myers and Friedrich Hirth of the same period. The idea that "the Shang dynasty was still in the late Stone Age" advocated by Hu and Gu was also directly influenced by J. G. Andersson's *An Early Chinese Culture*. Conversely, Arthur W. Hummel played a key role in introducing the "Discussion of Ancient History" to Western academia. Paradoxically, while Berhhard Karlgren wrote the *On the Authenticity and Nature of the Tso Chuan* to refute Kang Youwei's reinterpretation of Confucian Classics, his work however became of value for Chinese scholars to reaffirm the value of New Text Confucianism and spear ahead the "Discussion of Ancient History" movement. In a word, if we would like to choose a saying to describe the relationship between modern Chinese historical scholarship and Western Sinology, the Chinese proverb, "the stone from other hills may serve to polish jade," may be an appropriate choice.

Writing Her-stories: The Rise and Development of Women's History in Africa //*Zheng Xiaoxia*

After many countries in Africa achieved independence and buoyed by the strong nationalist sentiment, African scholars pursued with passion the writing of African history to explore and rediscover the continent's past. Against the backdrop of great advances in "new historiography," African history and feminist movements in both Africa and around the world, women's history began to emerge across African countries from the late 1960s. During the 1970s, the field experienced a marked progress whereas as a whole it remained in the stage of infancy. Thanks to the advance of women's studies, the focus of women's history in Africa shifted from elite women to ordinary women in both urban and rural areas, including prostitutes, maids and servants, witchcrafts, laborers, slaves and farmers, etc. From the 1990s, gender studies also gained influence in Africa, women's studies in Africa have therefore also shown multi-perspectival and interdisciplinary characteristics.

Digital Technology and the Notion of History: A Study of the Relationship between Historical Database and Historical Methodology in China //*Shen Bin, Yang Peina*

The making of historical databases has merged the development of digital technology and historical research. They represent the technological improvement in historical methodology and also help the latter's innovation and expansion. The creation of digital archives and their searchable function extends the empiricist approach to historical study and textual criticism whereas the establishment of large quantitative databases is indebted to the interest in using social science and statistical methods in historical analysis. The development of digital technology has made improvement in both areas. In more recent years, thanks to the growing interest in digital humanities, it is possible to explore more innovations in historical methodology, which can help expand the emphasis of the new historiographical trend on analyzing both textual and historical contexts.