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

北京大学 历史系 北京 100871

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》标志着中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。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。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。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"。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,新的研究方式亟须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。

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。这种工作可以分为如下几类:其一,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。其二,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。其三,通过 OCR 技术与人工核对的结合,全文检索资料库诞生了。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。

正因为此类数据库是沿着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,即便有批评意见,也是在肯定的大前提之下。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具有范式转化的意义,其研究理念背景是 20 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,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 世纪年代和 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。

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史学风格的产物。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 世纪六七十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。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,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,方便学者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。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。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。

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、史实考证等繁琐的历史研究第一层面的工作,而直接在第二层面即历史事实分析上开展研究。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)提取数据所做出的研究,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。

再次,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,历史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）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。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。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，而是协助学者发现议题、开展研究的工作环境。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。

借助文本挖掘，我们可以揭示史料所处的多重脉络，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，因此要蒐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。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。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。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。

《史学理论研究》2017 年第 2 期，约 11000 字