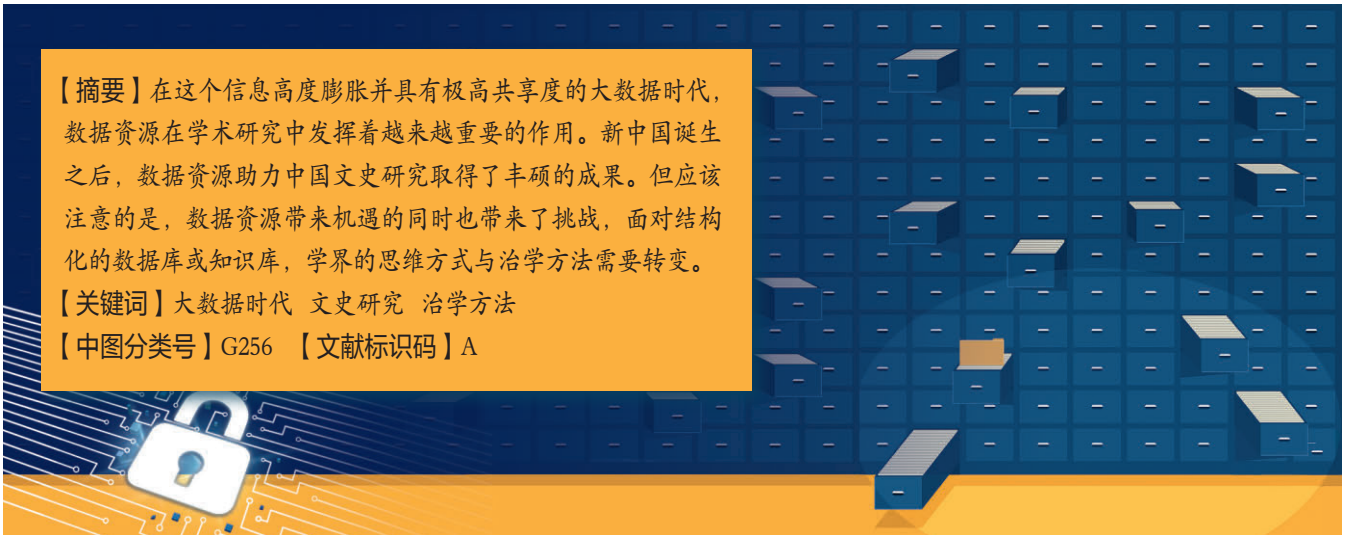
大数据时代下的文史研究方法

刘 源　罗 兵



【摘要】在这个信息高度膨胀并具有极高共享度的大数据时代，数据资源在学术研究中发挥着越来越重要的作用。新中国诞生之后，数据资源助力中国文史研究取得了丰硕的成果。但应该注意的是，数据资源带来机遇的同时也带来了挑战，面对结构化的数据库或知识库，学界的思维方式与治学方法需要转变。

【关键词】大数据时代　文史研究　治学方法

【中图分类号】G256　【文献标识码】A

在印刷术发明以前，书籍文献弥足珍贵，人们获取知识及其他信息的路径除了人与人之间的口耳相传，便只有阅读文字。相比于人与人之间的传授，阅读已经足以突破时间和空间的限制，让人的学习方式更加自主便捷，获取的知识也更加真实可信。然而，无论是天灾还是人祸，都可能对书籍的保存和流通造成巨大损害，知识载体的损毁甚至可能直接带来知识本身的消亡，有感于此，先贤常通过对古书的编辑来将大量的信息与知识整合在一起，这可以视作古人“数据库”建设的初步尝试。

刘向《别录》叙述左氏学传承云：“左丘明授曾申，申授吴起，起授其子期，期授楚人铎椒，铎椒作《抄撮》八卷，授虞卿作《抄撮》九卷；授荀卿；荀卿授张仓。”由此可见，“抄撮之学的立意，是为了在较短的时间内，花费较少的精力，而得以对一部著作进行整体的简要性把握。这一时期的此类著作采取何种体制，是完全忠实于原文还是有所发挥，今已无法考知，但这种删繁节要，便于观览的编纂出发点，是非常值得重视的。”吴均《五总志》：“唐李商隐为文，多检阅书史，鳞次堆集左右，时谓为獭祭鱼。”辛文房在《唐才子传》也记载：“商隐工诗，为文瑰迈奇古，辞隐事难。及从楚学，俪偶长短，而繁缛过之。每属缀，多检阅书册，左右鳞次，号‘獭祭鱼’。”李商隐为诗为文，都以用典绵密著称，而支撑他的如此不厌其烦的隐词比事的獭祭行为，难免成了人们津津乐道的话题，从中我们不难觉察出一种取向，即对知识的占有量和剪裁程度是人们作诗作文内容丰富与否的必要条件。秦观也提到了自己在成年记忆力衰退之后，感慨检索文献之费时，故而不免依样画葫芦。观古可知，古人在建构自己的知识体系的同时，就已经有意识地对既有知识做减法（所谓“抄撮”之学），以便于记忆和检索，又对其做加法（所谓“杂学”著述），借杂糅所学所见以属词比事。可见，早在电子数据库诞生以前，中国古代学者已经有了类似的活动，其目的和功用，与现在并无二致。

数据资源助力新中国文史研究取得巨大成就

随着信息时代的到来，文史研究可以利用的技术手段拓宽，材料获取途径的多样性增强。数字网络以及移动通信技术的不断进步，使得学界得以应用计算机和互联网对徽州文书以及甲骨文等原始文史研究材料进行更有效且有深度的辨析。20 世纪 50 年代，语言学家迈克尔·文特里斯（Michael Ventis）成功地破译了线形文字 B。麻省理工学院和谷歌人工智能实验室的研究人员提出的神经网络算法，实现了古希腊迈锡尼文明时期线形文字 B 的自动翻译，把 67.3% 的线形文字 B 同源词翻译成了希腊语。人工智能、核技术和物理化学技术等前沿技术的应用，使得原始材料被更有效地甄别和解释，从而拓宽了

材料广度，也促进了新材料的发现。除此之外，信息时代带来的材料的重组、学科的交叉、信息文化的兴起，也促进了新材料的发掘、分析及更新。随着计算机硬件能力的不断提升，加之数据资源的持续累积，以大数据为核心逻辑的智能应用革命开始逐步影响人类的日常生活。在大数据技术的帮助下，人们可以利用崭新的视角来实时、多角度、全方位地掌握事物的发展规律，并更好地预测未来，进而为生产和社会活动提供海量而优质的决策。所以，信息文化的快速发展，使得整体文化环境发生转变，新材料不断出现、新材料整合速度不断提高，客观上促进了数据资源的累积，文史研究方法也因此发生变化。

党的十一届三中全会后，在解放思想、实事求是路线的引领下，文史学界不断开拓创新，中国古代文史研究焕发出崭新的生命力。随着我国对外开放的深化，国家经济实力日益增强，中国古代文史研究取得了丰硕的成果，具体表现在学科构建、人才培养、成果出版、国际交流等方面。中国古代文史研究不断向全方位、多角度、深层次发展，我国文史工作者在科学系统地借鉴并融合古今中外优秀研究理论和方法的基础上，不断整合完善现有资料，积极探索新的文献和考古材料，许多海内外罕见文献因此得以整理并出版。以敦煌吐鲁番文书、甲骨文、徽州文书、悬泉置简帛以及众多民间文书为代表的新出文献，夯实了我国古代文史领域的研究基础，丰富了研究内容，拓宽了研究的深度和广度。与此同时，文献古籍的数字化也被提上日程，科研单位和各大高校纷纷上线数据库项目，催生交叉学科研究方法，文史领域治学与数字化时代同步推进的趋势日益明显。

进入 21 世纪以来，我国文史研究者乘科学技术之东风，借助各类互联网信息技术手段，植根于中国历史实际，发现、整理和抢救了大量的文献古籍资料，文献和古籍的保护进程得以显著加快，古籍利用和保护之间的矛盾也得到了妥善的解决。近年来，以敦煌文献数字化和国际敦煌学、海外中华古籍合作保护以及“一带一路”邻国语言文字中汉字音的数字化整理等为代表的一批重点研究项目不断推进，通过目录汇编、图像 / 音频扫描、4D 数据库建设等工具手段，在全面保护存档既有资料的同时，有效地提高了文献内容和考古内容的质量，为未来文史研究领域的广度和深度提供了可靠的保障。这些成就，与新中国成立以来在文献数据资料领域持续不懈的探索整理，以及信息技术和数字化手段的有效助力，是分不开的。

大数据时代为文史研究带来的机遇与挑战

大数据时代，数据在我们的日常生活与学术研究领域发挥着越来越重要的作用，传统纸质文献越来越多地被数字化，各种形式的数据库层出不穷。数据作为研究成果的同时，其研究基础的地位也在不断被强调。具体到人文学科的研究，数字文献大致可以分成传统文献的数字影像和结构化的数据库。与数字文献相比，传统文献具有天然的劣势，除了传播方式单一、传播时间较长、保存传播成本较高等众所周知的原因以外，我们必须注意到：“旧媒体将知识分割于不同的物理载体之中，比如说这本书的知识很难与另一本书的同类知识关联，这种检索工具很难跟另一种检索工具互通，而学术研究则要求尽可能地联系各方知识，便于重新组合和运算。学者重组知识的能力越强，创造力也就越强。”大数据时代在减少文史研究所耗费的时间和物质成本的同时，使得学者可以高效选取材料进行组合和分析，材料获取效率增加。以往，学者为了查阅某一文献资料可能需要跨越大半个中国，准备许多证明材料，而现在足不出户便能查询到自己需要的材料。前人遍检群书而不得的内容，我们可能只需用几秒钟就可以得到答案，不会利用电子文献检索的学者则成了名副其实的“今之古人”。这使得文史研究从侧重获取新材料转变为侧重提出新问题，学术研究更具有效率性，为学科的深入探究提供了便利。

数据库的广泛使用，打破了学科之间的界线，拓宽了专门知识领域的边界。跨学科的知识链接，为新知识体系的出现架起桥梁，“国际数字人文机构联盟”和“数字人文中心网络”这两大人文研究数字联盟的出现，使人文科学和数字科学加深融合，例如促进了历史学科从解释性学科向求是性学科的转变，实现了学科价值的扩展。进而可以说，数据库的出现在不断拓宽文史研究角度的同时也能影响其研究价值的扩展。同时“人文计算”、复杂网络分析、大规模数据分析等研究方法的使用，虽然在一定程度上弱化了文史研究中的批判性与人文关怀，但却在某种程度上革新了文史研究的方式，从而使研究更具科学性。

数字文献的不足也是显而易见的，从文献的保存、阅读和检索来说，不同的数据库必然会展示出不同文字的准确率和检索的查全率、查准率，即使数据库的制作者精益求精并不断改进检索技术，其文本的准确率已经做到了与纸本文献不相上下，我们依然无法避免在检索“吴梅”时发现众多“吴梅村”相关

词条的情况，简而言之，数据库在无意识检索的层面可以速度惊人，却依然无法代替人类进行有意识的搜索。

从这个角度看来，大数据时代，我们更要警惕的是“方法论”的错位。前面已经提到，前人也构建过自己的“数据库”，虽然和如今的数字文献相比，它的规模无法同日而语，可恰恰是因为被人有意识地编纂，它的优势在于其内在的系统性和相互之间的关联性，“比如敦煌卷子中发现的很多小类书，像《孔子备问书》《随身宝》《太公家教》及《兔园册》等，它的包罗万象和排列秩序，其实可以反映当时知识的定型和简化”，这种系统性和关联性交织在一起，构成的内在的自足性正是这一时期图书的编纂者和阅读者“共识性”知识体系的反映，在这种“共识性”的知识、思想背景之下，同时代或之后的学者分享、传承彼此的知识与经验，他们对未知知识的检索的出发点源自于对既有知识的理解和掌握。如果我们不具备对“已知”的熟悉，而却偏偏执着于“未知”的汪洋，所面对的，将是极其危险的处境。

即使我们尽最大所能规避以“未知”检索“未知”的情况，却依然无法忽视数据（数据库）本身并不会说话的事实，面对同样的数据，对文献的分析和使用也是因人而异的，这种“横看成岭侧成峰”极有可能导致截然相反的结论。1980 年，美国威斯康辛大学陈炳藻先生在《红楼梦》讨论会上发表《从词汇统计论证红楼梦的作者》一文，通过统计《红楼梦》的词频，认定后四十回也出自曹氏，一时引起巨大反响，是继高本汉之后首次全方位运用电子检索和统计的手段对《红楼梦》进行研究，然而不久之后，中国学者陈大康先生同样用精密的统计方法得出与之相左的结论：《红楼梦》后四十回含有曹雪芹少量残稿，但并非是作者原作。由此可以看出，数据本身并不会说话，即使在大数据时代，单单靠先进的统计方法，并不是解决人文学科相关问题的“万能钥匙”。

大数据时代下文史研究的新路径

飞速发展的互联网信息技术，让我国的文史研究呈现出若干新趋势、新特点。国家的战略性规划，各级政府和相关单位的持续投入，以及优秀学术人才的积极参与，都为我国文史文献资源研究与建设的系统化、数字化、科学化打下了坚实的基础。利用大数据技术研究中国古代文史，对其本身与相关领域的学科建设和学术发展，具有极其显著的意义，这种意义尤其体现在研究范式与方法论的革新上。基于这样的理解，笔者认为，大数据时代下的文史研究方法，可以在以下三个方面有所创新：

一是解决单凭人力难以彻底解决的疑难问题。如中国古代文学中的周边国家意象与天朝朝贡体系以及中国古代对外交流关系的演化之间，是否存在联系？对此类问题来说，数据库是基础，文本分析技术是核心，需要通过定量统计分析，进行作品的辨伪、异文对照，解决修辞特色及风格题材的变迁等悬疑难决的问题。二是重新验证已有成说的史论。例如明代以李梦阳、何景明为代表的前七子，其诗文创作中是否落实了“文必秦汉，诗必盛唐”的主张？在大数据技术的帮助下，我们可以利用共词及语义分析、人物事件交杂等思路，尝试全新分析和检验诸如社团流派、人物关系、情节演进、阶段特征、历史影响等已有成说的问题。三是超越印象和定量分析层面，科学梳理文史中存在的特征、规律、关联性问题。例如白居易有诗近四千首，陆游有诗词近万首，它们的格局、题材、修辞、风格在不同历史时期发生过哪些变化？通过对某作家或某一类作品的深度学习，发挥其关联分析、文本比对等技术优势，挖掘到以往不曾注意到的迹象或线索，以期提高文学经典研究的科学性和可靠性。

现阶段的中国古代文史研究，在数据分析方面虽然已经起步，但多局限于文献数字化阶段。主要用于数据内容存放和管理的数据库仍然占据主流，而能够实现分析统计的关系型文史数据库仍然稀少。近些年，随着《中华经典古籍库》等数字化文献资源库的推出，数据库在文献检索功能方面已有较大的进步，但结构化的实现统计分析和知识再生、运用数字人文的分析工具和技术方法来研究古代文史等功能，仍处在尝试性阶段，未成规模，影响也不大。如何建设更为丰富、完善的数据库，如何使数据库功能更加人性化与科学化，如何让数据库在文史研究中发挥更加重要的作用，仍是有待解决的问题。未来，文史研究学界只要与时俱进，解放思想，将文史资源的发掘、考证、研究置于科学技术进步和文化繁荣的背景之下，充分调动各方面资源，就能更好地保护、开发和利用我国的文史资源，使文史研究始终与国家同发展，同时代共进步。

（作者分别为上海师范大学人文学院博士研究生；南京大学文学院博士研究生）

【参考文献】

①郭醒：《〈艺文类聚〉研究》，沈阳：辽海出版社，2010 年。

②史睿：《数字人文研究的发展趋势》，《文汇报》，2017 年 8 月 25 日。

③葛兆光：《思想史的写法——中国思想史导论》，上海：复旦大学出版社，2004 年。

责编 / 周小梨　美编 / 杨玲玲