

加快“数字化”向“数据化”转变

——“大数据”、“云计算”理论与古典文学研究

郑永晓

内容提要 在浩如烟海的古籍资料中发掘有价值的信息和知识一直是人文科学和社会科学领域面临的重要挑战。应对这一挑战的艰巨性伴随古籍数字化进程和数据库技术的发展而有所缓解。然而与日新月异的 IT 技术和互联网发展相比,古籍数字化及相关应用尚处于初级阶段,文献“数据”本身的价值及应用尚未得到足够的重视,近年来兴起的“大数据”、“云计算”理论及相关应用为网络信息技术深度应用于传统人文学科开启了黎明之门,有可能对未来的学术尤其是需要处理大量文献的古典文学研究产生重要影响。

关键词 大数据 云计算 古籍数字化 古典文学研究

现任牛津大学网络学院教授维克托指出:“大数据已经撼动了世界的方方面面,从商业科技到医疗、政府、教育、经济、人文以及社会的其他各个领域。”^①在国内,大数据在互联网、物联网、移动通讯、电子商务等领域风生水起,颇具声势。2012年5月,由科技部发起的香山科学会议组织了以“大数据科学与工程——一门新兴的交叉学科?”为专题的学术研讨会。有学者认为:“数据科学作为一个以大数据为研究对象,横跨信息科学、社会科学、网络科学、系统科学、心理学、经济学等诸多领域的新兴交叉学科方向正在逐步形成。”^②

然则,何谓大数据?维克托所言“方方面面”能够涵盖我们的传统文史研究吗?这一新兴技术与相关思想是否能为古典文学研究带来新的启示?

一 “大数据”、“云计算”与“互联网思维”

所谓“大数据”(Big Data),维基百科的解释是:“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信息。”“由巨型数据集组成,这些数据集大小常超出人类在可接受时间下的收集、应用、管理和处理能力。”^③伴随近年来互联网、物联网、云计算等技术的迅猛发展,网络间尤其是移动互联网中的各种应用层出不穷,引发了数据规模的爆炸式增长,从而形成了大数据。

大数据的概念在近几年间不断发展、丰富,与传统数据工程相比较,大数据具备所谓 5V 特征,

① [英] 维克托·迈尔-舍恩伯格、肯尼思·库克耶著,盛杨燕、周涛译《大数据时代——生活、工作与思维的变革》,浙江人民出版社 2013 年版,第 15 页。

② 程学旗、王元卓、靳小龙《网络大数据计算技术与应用综述》,《科研信息化技术与应用》第 4 卷第 6 期(2013 年)。

③ <http://zh.wikipedia.org/wiki/大数据>。

即：①Volume，数据规模从GB而TB而PB，甚而开始以EB和ZB来计算①。②Variety，数据类型繁多，包括结构化数据、半结构化数据及非结构化数据，尤其是近年来个性化的非结构化数据呈几何级增长。③Velocity，数据的产生和处理速度按秒计算。④Veracity，数据真伪杂陈，良莠互见。⑤Value，数据量大而价值密度低。

鉴于大数据所具有的这些特性，如何从纷繁复杂的数据中提取所需的精华将考验人类的智慧，于是，业界专家又提出了“云计算”。

所谓“云计算”（Cloud Computing），最早是由Google首席执行官埃里克·施密特（Eric Emerson Schmidt）在2006年的搜索引擎大会上提出的。其基本含义是一种基于互联网的计算方式，将庞大的计算处理程序自动分拆成若干个较小的子程序，再由多部服务器组成的庞大系统联合进行搜索、计算、分析，并将处理结果瞬间反馈给用户。云计算具有超大规模、虚拟化、高扩展性等特征。

大数据与云计算相辅相成，被视为一枚硬币的正反两面。大数据着眼于“数据”，即内容，重在信息资源；云计算着眼于“计算”，重在数据挖掘和分析计算。没有云计算，则大数据再丰富，也如镜花水月，无从发挥其效用；没有大数据，则云计算再强大，也终难有用武之地。可以说，云计算是发掘“数据”价值，征服“数据”海洋的重要工具。

大数据和云计算绝非可以单纯地理解为使用了一个先进的计算方法以处理更多的数据而已。事实上，大数据和云计算的出现正在或即将改变我们的思维方式，对于我们重新认识世界提供了更为科学的方式。

例如，大数据时代需要处理的数据如此之多，速度要求如此之快，则有可能造成我们不再热衷于追求细节的精确度而是注重于事物的发展趋势，并在宏观层面较之以往展现出更深刻的洞察力和预见力。

传统数据库要求数据高度精确并且按预设的规则排列，这种注重精确性的关系数据库自有其存在的理由而且还将在特定领域存在相当长的时间。但是注重精确性显然是信息缺乏时代的产物。因为数据量小，所以要求精确，也能够做到精确。然而，纷繁复杂的自然和社会现象往往并非小数据所能涵盖，况且小数据即使都是真实的，也有可能得出以偏概全的结论。俗语所言“一叶障目，不见泰山”、“只见树木，不见森林”就是这个道理。微软资深数据库设计家派特·赫兰德（Pat Helland）2011年指出：“我们再也不能假装活在一个整齐划一的世界里。”（“We can no longer pretend to live in a clean world.”②）伴随云计算的逐步成熟，从小数据过渡到大数据是必然趋势。而在大数据称雄的数据海洋中，精确的结构化数据只占极少部分，大量非结构化数据成为有待开采的金矿。而要处理大数据，就必须一定程度上接受不精确性。因此，我们需要放弃传统的追求确凿无疑的思维方式，放弃对一些局部或细节真实性的追求，转而追求对概率和趋势的认知。纷繁而小有瑕疵的大数据所得出的结论较之无瑕疵的小数据得出的结论更为可靠和科学。

事物或现象之间的关系复杂，存在着各种各样的可能性，例如因果关系、相关关系、共变关系、反变关系等。严格来讲，很多事情的因果关系被完全证实几乎是不可能的，只能说，两者之间可能存在着因果关系。当两类现象在发展变化的方向或大小方面存在一定联系时，我们视之为相关关系。因果关系实际上是一种特殊的相关关系。在研究相关关系的基础上，可进而研究因果关系，相关关系为研究因果关系奠定了基础。

在小数据时代，无论是因果关系还是相关关系，很多都基于理论上的假设，然后再进行验证。这种“大胆假设，小心求证”当然也促进了科学的发展。但是，基于假设的论证有可能受主客观因素的

① 按字节（byte）计，1GB = 10^9 ，1TB = 10^{12} ，1PB = 10^{15} ，1EB = 10^{18} ，1ZB = 10^{21} 。

② <https://queue.acm.org/detail.cfm?id=1988603>。

限制而出现偏差。而在大数据背景下,当数据点以极大的幅度增长时,则极有可能会观察到许多在小数据环境中很难观察到的相关关系,且不受偏见或先入为主等因素的影响。因此,基于大数据的相关关系分析,必然取代基于假想的方法。

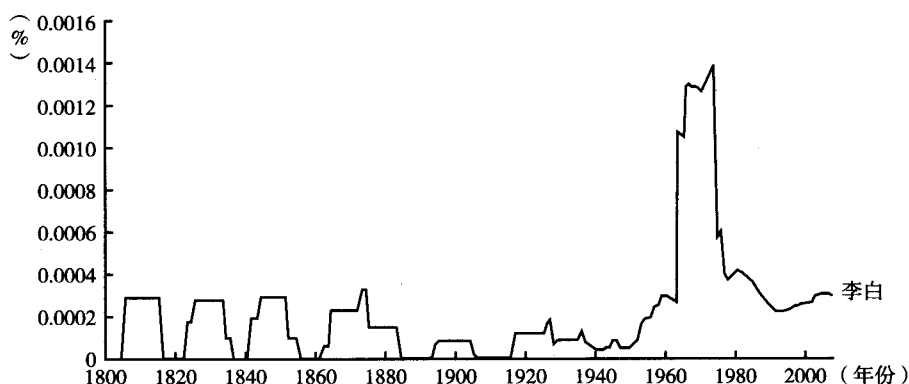
大数据的出现代表着人类认识世界的方式在发生着某些重要的改变。如果我们把世界看作是由信息构成的海洋,利用现代化手段去收集、统计、分析这些信息,无疑为我们提供了一个此前从未有过的审视世界的视角,可以帮助我们更全面、深入地探索这个复杂多变的世界。

在过去不久的2013年,“互联网思维”一词大行其道。至于“互联网思维”这一概念的确切含义,由于言人人殊,迄无确切定义,笔者不想在此讨论,但有一点是确定无疑的,即“互联网思维”在实际应用中是基于大数据和云计算的。

“大数据”时代的很多学科都将发生巨大甚至是本质性的变革和发展,进而影响人类的价值体系和知识体系,当然也影响到我们的学术研究。

二 “大数据”、“云计算”应用于学术研究的可能性

2010年,Google发布Ngram Viewer,该应用是基于庞大数据库Google Books^①开发的,其基本用途是通过输入字词,通过Ngram Viewer生成的趋势线,来观察这些词汇在不同年代出现的频率,借以了解不同年代社会文化的变迁。



上图是笔者利用该程序查询“李白”所得出的结果。这个结果的科学性当然有待于完善。一方面,Google Books在收录中文图书方面可能会有一定的局限性,数据尚不完整,我们不宜要求过高。另一方面,上图显示,“李白”在1970年代使用频率较高,显然与郭沫若《李白与杜甫》的出版及当时相关争论有关。证明此应用具有一定的科学性。而类似这样的统计分析依靠传统方法是难以实现的。

“Google”提供的这一应用基于近年来西方产生的一门新的学术“文化组学(culturomics)”,“culturomics”是“文化”与“基因组学”两个词的合并,其宗旨在于通过文本的定量分析来揭示人类行为和文化发展的趋势。哈佛大学生物学家埃略兹·利伯曼·埃顿(Erez Lieberman Aiden)和心理学家杰·B·米歇尔(Jean-Baptiste Michel)于2010年12月16日在《科学》杂志上发表《数字化图书的定量文化分析》一文,首次提出“文化组学”概念。他们认为,通过在海量数据中提取并分析某些词汇在图书文献中的增长、演变、消亡等趋势,有可能观察到大范围内文化特征的嬗变。美国学者

^① Google Books涵盖的具体图书数量因属商业秘密,Google秘不示人,但Google曾说每天可处理三千册书,至2007年时,已完成约七百万册图书的数字化。

丹·库汉 (Dan Cohen) 指出:“近年来,全球人文学科在数字化进程中得到了多方支持,并取得了诸多成就。如今‘文化组学’这一新兴术语既代表了大规模媒体数据库和其他文化数据的融合,也体现了人文学科领域学者积极与其他学科对话交流、努力实现数字化治学的愿望。”^①

大数据和云计算理论及其应用引起了国际学界对科学研究方法的重新审视。在科学研究史上,最早的科学研究为实验科学,可以追溯到古希腊,又称经验科学,代表人物如英国文艺复兴时期的哲学家弗朗西斯·培根 (Francis Bacon) 等,主张科学必须是实验的,归纳的,一切真理必须以大量确凿的事实为依据。与实验科学对应的是理论科学,使用演绎法以推究各种定律和定理为特征,17 世纪的英国物理学家艾萨克·牛顿 (Isaac Newton) 堪为代表。第三种为建立在模拟方法上的计算科学,在计算机上利用数学模型、定量分析等方法来解决科学问题。1982 年诺贝尔奖得主,美国理论物理学家肯尼斯·G·威尔逊 (Kenneth G. Wilson) 是这种研究范式的倡导者。而大数据和云计算的相关理论和应用有可能催生一种新的研究范式:第四范式 (The Fourth Paradigm)。其提出者为已故图灵奖得主吉姆·格雷 (Jim Gray),他把数据密集型科学从计算科学中单独区分出来以应对未来复杂性计算的挑战。具体解释可参见微软研究院编印《第四范式:基于密集数据的科学发现》(The Fourth Paradigm—Data-Intensive Scientific Discovery)^②。

第四范式不仅是研究方法的变化,更是人类思维方式的重大变化。在这种研究范式中,研究者面对浩如烟海的数据,不再抽取少量的样本进行分析,而是把所有数据作为一个整体,利用数据挖掘、计算、分析等技术,直接从数据中探寻所需要的信息、知识和智慧。与其他研究范式颇为不同的是,这种研究甚至无需直接接触研究对象,而是把数据本身作为研究对象,通过数据去解释其背后纷纭复杂的世界。

与前述“文化组学”这样的理论及其应用相比,我们在学术理念和具体应用方面显然还有相当的差距。笔者以为,传统人文学科与信息技术的结合可以生发出很多学术增长点,尤其是大数据和云计算的相关理论为传统文史研究带来很多重要的启示,为传统学术注入了革命性的思维,具有划时代的意义。

自上个世纪初,林传甲为京师大学堂编写《中国文学史》讲义以来,一代又一代的学人撰写了数以百计的文学史专著和难以计数的专题研究著作,尤其是近三十余年来,学界的研究视野逐步扩大,研究领域涉及文学史研究的多个侧面,甚至对相当数量的二三流作家也有较为深入的研究,这对于推动这一学科的建立与繁荣无疑作出了很大贡献。但如果从“第四范式”的研究角度看,这些研究尽管具体方法和理论水平多有差异,但都可归类到实验科学和理论科学中。实验科学的不足在于无论列举多少证据,总有可能以偏概全;而理论科学的不足在于基于假设的探索往往因过于复杂而难以解决实际问题,比如关于文学史发展是否有规律可寻以及中国文学史的规律到底是什么等问题,数十年来众说纷纭,迄今未见共识。如果我们能在不久的将来把先秦以来的所有古籍都予以数字化,并引入大数据思维,则文学史研究即使不能出现划时代的进步,也能开辟出一片新的天地。

应用大数据理论和方法,从宏观角度而言,可以把历史上所有作家作品纳入统计分析的视野。所有作家的出生地、家族背景、家庭成员构成、求学、科举、游历、仕宦、爱好、作品数量、交游唱和情况、作品创作时地、文体构成比例、遣词用句习惯、时人和后人的相关评价、作品被选录情况等等按照预设要求瞬间以数据表的形式得到呈现。所得到的结果有可能暂时不能发现其背后隐藏的意义,也有可能发现使用传统方式永远都难以得到的结论。例如,关于创作与作家经济状况的关系,韩愈说:“欢愉之辞难工,而穷苦之言易好。”(《荆潭唱和诗序》) 欧阳修说:“(诗) 穷者而后工。”(《梅圣俞

① 张哲《“文化组学”用先进技术推动对史学的跨学科研究》,《中国社会科学报》2012 年 1 月 16 日第 257 期。

② Tony Hey 等编著,微软研究院 (Microsoft Research) 2009 年 10 月 1.1 版。

诗集序》)而张表臣不赞同此说,谓:“欧阳公、王荆公、苏东坡号能诗,三人者亦不贫贱,又岂碌碌者所可追及?然则谓诗能穷人者,固非矣,谓待穷者而后工,亦未是也。”(《珊瑚钩诗话》卷三)这两种观点显然都不难找出相当数量的例子证明其合理性。在我国文学史上,既有相当多的作家在其穷困潦倒时写出了很多优秀的作品,典型者如曹雪芹;也有不少生活条件充裕的作家创作成就很高,典型者如白居易。况且同一个作家在不同时期也可能有穷困与富足之别。因此这两种观点都代表了局部真理。但是根据现有的学术范式我们很难分析出在中国文学史的历史长河中,在有据可查的作家序列中,到底哪些作家适用这两种不同的理论?如果有基于所有数字化古文献的大数据支持,则不难发现哪种情况所占比率更高、在不同时期有哪些不同表现,并量化显示贫富程度对文学创作的影响。

近年来,有关文学与影响其发展的外部因素之关系的研究不乏热点,如文学与家族、科举、政治、经济、军事、地理环境、图书出版等等。这些研究所用的论证方式固然千差万别,但概括而言多类似于自然科学中的抽样分析,属于归纳法,或在归纳法所举证据的基础上再进行理论上的阐释。这些研究对于断代或文学群体的解释有重要意义,对于分析特定阶段的特定问题也颇有助益,在未来相当长的时期内这些研究还将存在下去。另一方面,我们也必须指出,这些研究都是基于小数据的研究。例如,文学史上经常有如像蒲松龄那样的作家参加科考数十年而不能成功的例子,科举考试对它们的人生和创作具有重要影响。我们研究单个作家或某个时期的文学现象时经常会关注科举因素。士人科举成功与否,对其经济、仕宦、心理、交游等产生多方面的影响,而这些因素都可能反映到其文学创作上,而且不同时代、不同地域、不同文化背景的士人所受到的影响力大小也肯定有差异。唐宋时期很多杰出的人士包括作家都多出自科举,而明清时期很多杰出的人士在科场上颇为失意。传统研究方法只能对某些个体、特定时期的作家群体、或者某个时代的科举与文学创作情况进行探讨,难以对所有作家与科举的关系进行探讨。而大数据的相关理论和方法为我们提供了这样一种可能性,即细致区分历代作家与科举的关系,诸如科举成功者与非成功者、成功者在考中进士前后的差异、不同朝代科举对士人影响的异同等等,这对于宏观而精确地分析科举与文学之关系显然大有裨益。

如前所述,基于大数据的思维特别注重事物间的相关关系,我们在分析文学与外部因素关系时,有可能发现其他此前我们从未注意的现象与文学的关系,果能如此,则其意义更远大于对已知的相关外部因素对文学影响的研究。

运用大数据理论也可以解决一些具体问题。例如用典的产生、发展、嬗变等,依靠传统方法当然也可以考辨,但是只能解决局部问题。而基于大数据的分析,则可以对历朝历代文学作品中的典故进行宏观而精准的分析。例如,根据用典数量和用典频率的统计分析,我们可以从一个侧面考察唐宋元明清诗歌的风格倾向。甚至对某个典故在不同时期的演化也能有更为全面的把握。例如,晋陶潜不为五斗米折腰事,受到历代文人的高度推崇,不断付诸吟咏。但是这一典故不仅在表述上有“五斗粮”、“五斗低腰”、“折腰”、“折腰禄”等形式上的演化与区别,而且在不同作品中用法和具体含义也不完全相同。如岑参《衙郡守还》诗“五斗米留人,东溪忆垂钓”,自述居官而怀念隐居的心情。李商隐《自贻》诗“谁将五斗米,拟换北窗风”,表示自己不愿就小吏之职。吴潜《水调歌头·题烟雨楼》“叹吾曹,缘五斗,尚迟留”,则借以自叹为生计尚未能弃官归隐。我们可以使用这种传统方法对少量作家使用这一典故的异同进行研究,却很难对历代使用这一典故的所有情况进行完整的分析。而基于大数据的方法处理这样的小事轻而易举。对这个典故的分析不仅可以看出典故的演化、使用频率,更可看出历代对陶渊明接受程度的异同。

推而广之,通过对作品遣词用句、用典、用韵等要素的分析,可以全面准确地分析不同时期作家之间的影响与接受情况。通过对包括诗话、词话但不限于这些文献的相关要素的统计分析,我们可以完整地构建文学批评史在范畴、观念等方面的递嬗。通过对句式、用词、情感意象等方面的全面统计分析,当能比较清晰地界定诗、词、曲等文体的区别。通过对意象选择、情感语汇等方面差异的分析,

当有助于厘清词体的婉约、豪放、质实、清空等词学概念的区别、厘清唐宋元明清等不同时期诗风的异同和演变^①。

在大数据和云计算出现之前,自然科学抑或人文社会科学,都主要依赖于抽样数据和局部数据,甚至在无法获得实证数据时只能依赖假设、经验、理论等去推测。这些基于经验、理论或抽样数据的学术研究和理论探讨在未来相当长的时间内还将继续发挥其应有的作用。但是,这种方法所得到的结论,有可能是扭曲的认识或假象,具有一定的局限性。而基于大数据思维和方法分析所得出的结论,在把握问题的实质和分析其发展趋势方面显然具有极大的优越性。大数据带来的更高意义上的科学性,使得少量不精确数据无碍于数据分析的科学性。

当然,任何事情都不可能是绝对的,基于大数据的方法也不可能解决所有问题。人文学科尤其是古典文学研究从某种意义上说是今人与古人的心灵对话,通过作品的阅读,欣赏作品的美感,体悟古人的心理活动,咀嚼、涵泳的功夫不可或缺,这是任何数据分析都不能替代的。

三 古籍数字化的发展方向

近二十年来,以互联网为核心的IT(Information Technology)获得快速发展。但是相对而言,以往的发展主要是在技术层面,即在“T”(Technology)层面,而对信息即“I”(Information)的重视则有待于提高。

根据现代知识体系形成理论,数据经过加工成为信息,信息经过系统化成为知识,而知识则是“智慧”和“思想”的渊藪。这就是著名的DIKW(Data-to-Information-to-Knowledge-to-Wisdom)理论。美国教育家米兰·瑟兰尼(Milan Zeleny)在1987年撰写的《管理支援系统:迈向整合知识管理》(Management Support Systems: Towards Integrated Knowledge Management)和管理思想家罗素·艾可夫(Russell L. Ackoff)在1989年撰写的《从数据到智慧》(“From Data to Wisdom”, Human Systems Management)对此有系统的论述。在这一链条中,人类的智慧是经由数据而信息而智慧这样一种层级递进的方式而产生。数据处于链条的基础位置。没有数据的处理,就没有后来的信息和知识,也就更不可能有高层级的智慧和思想。在大数据时代,数据的重要性更是得到了前所未有的显现。

基于这些理论,我们显然有必要重新审视古籍数字化的作用,探讨在人文学科加强文献型数据库建设并利用大数据理论和方法深化相关研究的必要性和紧迫性。

近二十年来,古籍数字化成果和数据库的建设对于推动人文学科的科学化颇有助益,为解决某些考据方面的疑难问题提供了极大的便利,并助推了“E-考据”等观念的产生^②。但令人遗憾的是,总体而言,古籍数字化的成果亦即各种古籍类数据库的功能仍停留在检索方便上,用户的观念仍然是把这些数据库当作方便查询某些词汇的工具,治学严谨的学者往往把在数据库里查询到的内容与纸质版书籍进行对照无误后才敢正式使用。书同文公司开发的电子版《四库全书》等往往采取电子文本与原版图像可以对照比勘的技术,就是为了满足学者的这一需求。这些诚然令人称道,但同时恰恰说明经过这么多年的数字化进程,我们仍然把数字化文献当作方便查询的“书”来使用,我们一直视书籍的内容为其核心价值,而不是把它们当作“数据”而深挖其潜在的各种价值。

显然,数据库的应用还处于相当初级的阶段,远远落后于网络技术及其所带来的观念变革。对此,有学者也在理论层面探讨了古籍数字化产品的深度开发和应用问题。如2009年发表的李铎《从检索到分析——

① 参见拙作《情感计算应用于古典诗词研究刍议》,《科研信息化技术与应用》2012年第4期。

② 参见黄一农《明末至澳门募葡兵的姜云龙小考:兼答熊熊先生对“e-考据”的批评》,(台湾)《近代史研究所集刊》第62期;黄一农《e-考据时代的新曹学研究:以曹振彦生平为例》,《中国社会科学》2011年第2期。

计算机知识服务的时代》、罗凤珠《引信息的“术”人文学的“心”——谈情感计算和语义研究在文史领域的应用》^①等论文，对于在文史领域如何深度应用IT领域的技术都有很好的思考。可惜这些颇具建设性的探索迄今未见在技术开发和实际应用中有大的进展。笔者以为，解决这一问题的关键在于观念上需要完成由“数字化”向“数据化”的转移，同时在技术上引入大数据和云计算的相关技术和理论。

当文本成为数据，其用途便成倍放大，人可以阅读，机器可以分析。例如，面对一本《唐诗三百首》时，我们将其视为一本唐诗的普及读物，一本独具特色的，选录了很多脍炙人口的唐诗作品的优秀选本。但是如果我们把它当作数据交由计算机处理的话，则这些诗篇的情感分析、在历史上的演化、影响和被接受轨迹、自唐代以来被不同选本、类书等选录情况等等便会一目了然，甚至《唐诗三百首》本身自问世以来的阅读接受情况也会得到清晰的显现。

近年来，人文学科向实证的社会科学和自然科学靠拢的趋势有增无减。类似文史哲这样的传统人文学科，应顺势而为，在加快数字化进程的基础上，及早完成学术体系的转型。而大数据理论和方法的适时出现，为这一转型提供了难得的机遇。

经过近二十年的发展，我国古籍数字化和数据库建设取得了相当的成就。一方面，电子版《四库全书》、《四部丛刊》、《历代石刻史料汇编》、《十通》、《国学宝典》、《中国基本古籍库》、《古今图书集成》、《龙语瀚堂典籍数据库》等已经广为人知和使用。另一方面，网络上各种格式的电子书，如txt、doc、html、pdf、djvu以及知网、超星、方正等格式的文献也颇为繁富。这为使用大数据和云计算手段深度整合这些数据提供了重要的基础。

当然，我们也必须清醒地意识到，以中华文明历史之悠久，文献之宏博浩繁，现有的数字化文献还仅仅是所有文献的一部分，远不是全部。尚有相当数量的诗文集、方志、杂著、法帖等以稿本、抄本、胶片的形式存放在各图书馆、博物馆或私人藏书家篋中。大数据理论的基础是以全部而不是部分数据作为研究的对象，因此在未来若干年内尽可能完成所有古籍文献的数字化是将大数据理论应用于传统人文学科研究的前提。数字图书馆、数据库开发和拥有者、数字图书版权所有等都必须以开放的心态来顺应大数据发展的时代趋势。当然，数据拥有者的相关权利也应以适当的方式得到尊重。这个问题伴随技术进步和经济实力的壮大，必将最终实现。

而观念转换的难度更需要引起学界的注意。诚然，像古典文学研究这样的人文学科自有其独到的学术理论和学术方法，需要长久予以呵护。但网络技术对社会各项事物的影响甚至已经超出1995年尼古拉斯·尼葛洛庞帝(Nicholas Negroponte)在其《数字化生存》(Being Digital)一书中所作的预言，人们的生活、工作，包括科研都必然地受到这一技术变革的深刻影响，即使表面上与IT技术相距甚远的传统人文学科如古典文学研究也概莫能外。如能尽快转变观念，切实深入研究大数据、云计算相关理论和技术，汲取其适用于人文学科之精华，以互联网思维重新审视学科发展现状，必将有助于未来的古典文学研究产生新的生长点。可以毫不夸张地说，在网络时代，每一个人，每一本书，每一条文献，每一种思想都处于互联网的某个节点之中，伴随网络动态地演变和更替，并产生出新的信息和资源。维克托指出：“数据的真实价值就像漂浮在海洋中的冰山，第一眼只能看到冰山的一角，而绝大部分则隐藏在表面之下。”^②在大数据时代，数据是企业的战略资产，收集、运用数据的能力将成为企业的核心竞争力。在科研领域，数据的魅力导致科学研究呈现出数据密集和数据驱动的特点，数据分析成为第四范式。在人文学科，我们也有理由相信，一旦不再把古籍数据库视作仅仅可以查询字词出处的工具，而是把它们当作可分析的数据处理，必将为包括古典文学研究在内的传统人文学科带来革命性的变革，这一发展趋势目前看来只会不断深化而不会逆转。

① 均见《文学遗产》2009年第1期。

② 《大数据时代——生活、工作与思维的大变革》，第134页。

大数据和云计算标志着人类在认识世界的道路上又前进了一步,也为我们利用数据分析的方式探索古老的文明提供了一个绝佳的机会。在近现代史上,历次技术革命及其引起的思想变革,中国或者视若无睹,或者作为一个学习者。这次由大数据和云计算引起的变革中,我们与世界的差距最小。我们固然应该对传统的阅读和书写依然保持足够的敬畏,但是我们也应该珍视这次技术革命带来的机遇,在古籍数字化成果的基础上,加快由“数字化”向“数据化”的转变,借鉴新技术,拥抱新思维,努力开拓包括古典文学研究在内的人文科学研究新局面。

[作者简介] 郑永晓,中国社会科学院文学研究所研究员。出版过专著《黄庭坚年谱新编》等。

(责任编辑 石 雷)

· 学术信息 ·

“中华文学史科学学会古代文学史料研究分会 2014 年年会暨第四届出土文献与中国文学研究学术研讨会”召开

2014 年 8 月 8 日至 9 日,“中华文学史科学学会古代文学史料研究分会 2014 年年会暨第四届出土文献与中国文学研究学术研讨会”在山东蓬莱召开,中华文学史科学学会会长刘跃进教授,中华文学史科学学会古代文学史料研究分会会长郑杰文教授,中华文学史科学学会古代文学史料研究分会副会长蔡先金教授、周延良教授、贾三强教授、王德明教授出席会议。来自中国大陆、台湾地区和韩国等多所高校的九十余位专家学者参加了会议。中华文学史科学学会会长刘跃进先生在发言中首先强调文学史料研究具有重要价值,提出要对经典、对文献保持敬重之心,认为在古代文学研究领域谁避开了文学史料,这一领域也必将避开他;其次认为文学史料整理的目的贵在发现,整理文献是基础,要有辨伪存真的精神,将考证与立论相结合;最后鼓励学者对古代文学史料研究的前景持乐观态度,强调古代文学史料研究应当与当代社会的实际发展需要相结合,希望我们的研究能够为时代、为社会、为大众服务,做更有用的学问。

会议共收到论文一百零一篇,从内容上看大致可以分为三个方面:第一,出土文献与中国文学研究,这方面的论文有三十一篇,对清华简、上博简、北大简、郭店简等出土文献展开研究,其中以清华简为研究对象所占比例最大,有十篇,蔡先金教授所提交的四篇论文皆与清华简相关。姚小鸥、李文慧《清华简“视日”、“视辰”与先秦天命观》、俞艳庭等《权力话语与政治诗学——以清华简〈周公之琴舞〉为中心的讨论》也都是围绕清华简展开的研究。另外,郭丹《从郭店楚简看先秦时期对儒家六经功用的认识》、王洲明《〈上博诗论〉与〈毛诗序〉的研究》对上博简展开研究,俞林波《上古金文谱牒的文学试探》等对金文展开研究。张兵《〈战国纵横家书〉文学史料价值综述》则全面论述了《战国纵横家书》多方面的史料价值。第二,有关文学史料的刊刻、整理、考证等研究,这方面的论文有二十六篇,如程克雅《清胡绍煥〈昭明文选笺证〉方法与史料阐释探究——以“虫”“鱼”词汇为例》、范春义等《孔尚任旅晋诗文系年》、房锐《〈花间集〉编者赵崇祚考略》、贾三强《陕西古代文献集成编纂手记》、汪燕岗《论历史演义小说在清代的出版》等。第三,有关具体作品文学性以及其它相关的研究,有四十四篇,如柏俊才《平齐民的文学与文化成就》、汪春泓《论山水诗与陈郡谢氏之关系》、魏耕原《殷璠〈河岳英灵集〉诗学趋向与选诗取舍》等。此次会议有助于推动文学史科学、出土文献与中国文学相关研究的开展,同时对于提升相关学科学术水平和扩大学术影响力具有非常重要的意义。

(济南大学文学院 俞林波)