解析地方志资源大数据特性 管窥地方志“计算”研究新领域

林　浩

提　要：广义而言，拥有丰富资料内容的地方志资源既是一个资料库也是一个体量巨大的数据库，通过简单的比较就不难发现，传统的地方志具备一定的大数据属性。在大数据视野下，数据成为一种重要资源，地方志资源所蕴含的大数据特性如能善加运用，不仅将使方志文化更好融入大数据的时代洪流，而且能为地方志研究发展带来新的变革，甚至能将学科研究导入崭新的“计算”领域。

关键词：地方志　大数据　计算

多年以前，文艺青年中间流行一句话：学文科的只讲美丑不讲对错。这虽是玩笑话，却也说明一个规律，人文领域重感性思维，天马行空不拘一格，而理工科则更多理性思考，多缜密演算、推理求证。大数据时代的到来正逐步打破这种界限，无论是文艺圈还是理工界都必然融入数据的洪流。拿历史学来说，2014 年 6 月 15 日的《上海书评》上发表过一篇关于大数据和历史学的文章，署名尼克。文章在呼吁重视大数据发展的同时也对传统的历史研究吐槽不断，甚至犀利地指出“给中国做历史的提个醒，大部分的中国哲学家翻译水平已经被谷歌或百度翻译器反超，历史学家要是再不上进，也快没饭了”。此文一出自然遭到拍砖无数，但笔者认为，文中观点虽有偏颇，大数据对历史学研究带来的诸多影响却不容忽视。地方志是历史学的一个重要组成部分，在大数据视野下，可以看到地方志资源“天赋异禀”，具备诸多大数据特性，这些属性的有效发掘和利用将会为地方志的创新发展乃至学科建设带来新的变革。以下先对地方志资源的大数据特性做一解析。

一　地方志资源大数据特性解析

根据国务院 2015 年颁布的《促进大数据发展行动纲要》的定义，大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合，正快速发展为对数量巨大、来源分散、格式多样的数据进行采集、存储和关联分析，从中发现新知识、创造新价值、提升新能力的新一代信息技术和服务业态。① 在目前业界较为通行的观点里，大数据具备 4V 特性：Volume（大量）、Veracity（真实性）、Variety（多样）、Velocity（高速）。做一些简单的分析比较就不难发现，地方志资源的很多属性与大数据的 4V 特性相似度很高。

大数据的第一个明显特征是体量巨大。公元前 3 世纪，希腊时代最著名的图书馆——亚历山大图书馆收集了当时所能收集到的所有书写作品，可以代表当时世界上其所能收集到的知识量。但当数字数据洪流席卷全世界之后，每个人可获得的数据量已相当于亚历山大图书馆存储数据总量的 320 倍。② 地方志是一种资料性著述，其收录的各种资料可被视作基础数据，而地方志本身

① 参见董西成：《大数据技术体系详解：原理、架构与实践》，机械工业出版社，2018 年，第 2 页。

② 参见马海祥：《详解大数据的 4 个基本特征》，2014 年 9 月 12 日，https://sanwen8.cn/p/1besjEx.html。

是一个海量地情历史资料的数据库。据不完全统计，目前现存旧志有 8000 余种、10 万余卷，约占我国现存古籍的 1/10。国务院办公厅 2015 年印发的《全国地方志事业发展规划纲要（2015—2020 年）》指出：“目前首轮修志结束，第二轮修志进入关键时期，已出版省、市、县三级地方志书 7000 多部，行业志、部门志、军事志、武警志、专题志、乡镇（街道）志、村（社区）志等 2 万多部，地方综合年鉴 1900 多种、1.5 万多部，专业年鉴 1000 多种、7000 多部，以及大量地情文献。这些与旧志及其整理成果，共同构成了一座以国情地情为主要内容并不断丰富的地方志资源宝库。”此外，地方志资料内容丰富、涵盖面广，国务院 2006 年颁布的《地方志工作条例》明确指出，“地方志书，是指全面系统地记述本行政区域自然、政治、经济、文化和社会的历史与现状的资料性文献”。地方志书所记述的主体内容都是相关领域的实体资料，其资料含量可以用浩如烟海来形容。

大数据的另一个特性是真实性。数据的真实性是制定正确决策的基础，代表数据的质量，将直接影响分析和预测的准确性和有效性。不是所有数据都具有可靠性，不同数据源的质量千差万别，在数据的覆盖面、精确度等方面存在着巨大差异。① 追求高质量数据是一项重要的大数据要求和挑战，在互联网领域，各种数据信息良莠不齐甚至真假难辨，甄别筛选这些数据信息往往要耗费大量人力物力。地方志资料在真实性上具有得天独厚的优势。入志资料真实可靠是编修地方志的基本要求，志书所收资料绝大多数来自政府或机关、企事业单位的档案、地方文献、实地勘察数据等第一手资料，真实可靠程度很高，这也是地方志书权威性的重要保证。同时，地方志资料的稀缺性特点也很突出。因为志书所记内容多来自档案文献等第一手材料，许多资料数据从其他渠道难以获取，这样的稀缺性更体现出地方志资源难以替代的高应用价值。

大数据的多样性主要是指数据来自多种数据源，数据种类和格式日渐丰富，包括数字、文字、图片、语音、视频、地理位置信息等。② 从地方志特性的角度看，地方志资料的多样性首先体现在其地方百科全书的特性上。地方志综合记述一个地方自然和社会发展变化的基本面貌，其内容上至天文下至地理，涵盖自然、政治、经济、文化等方方面面，资料来源广泛，覆盖面广，综合性强。在数据类型上，传统的地方志多以文本形式出现，其中包含大量涉及地情地貌的关联数据，图片也较丰富，视频、音频等数据类型较少。随着影像方志等新兴方志形式的出现，地方志承载的资料类型也在不断丰富。

大数据的高速特性一方面是指数据产生快，另一方面是指数据处理快。③ 有学者曾戏称历史学是变化最慢的学科，客观讲地方志书的生成速度与飞速更新的网络数据相比并不具备优势，这有其自身学科发展的特点和规律。现阶段可以看到地方志的发展速度在不断加快，但如何更好顺应时代发展要求、提高方志资源更新速度和利用效率，想必也是未来学科发展难以回避的课题。

以上是将地方志的一些基本属性与大数据的 4V 特性做一个简单的对照分析，不难看出，古老的地方志与新兴的大数据之间存在诸多共同点，地方志与生俱来呈现出较为显著的大数据特征。

---

① 参见董欣、［美］戴夫士·斯里瓦斯塔瓦著，王秋月等译：《大数据集成》，机械工业出版社，2017 年，第 14 页。

② 参见孙静主编：《大数据引爆新的价值点》，清华大学出版社，2018 年，第 10 页。

③ 参见陈明：《大数据技术概论》，中国铁道出版社，2019 年，第 11 页。

二　加速地方志资源数字化，构建“全方志数据平台”

要发挥和运用好地方志资源的大数据特性，使其融入大数据的时代洪流，一个基础条件是地方志资源的数字化。中国地方志指导小组原常务副组长李培林在第一次全国地方志工作经验交流会暨2017年全国地方志机构主任工作会议上的讲话中指出，要推动地方志“从单一纸媒体志向广泛运用数字媒体志转变”。地方志从传统纸媒转化为数字媒介后，将具备可不断复制、多地存储、异地共享等优势，这是实现地方志信息化的必然步骤。有资料显示，地方志的数据资源库建设在一些地区和机构已有发展并具备一定规模，比如在2002年，国家图书馆已启动数字方志项目，全国各地规模较大的地方志网站多数也建有地方志资源数据库，其内容主要包括省市县三级志书、年鉴和地情资料的数据资源，一些网站还收录有部分旧志资源和地方志理论研究成果，这些数据资源的总量应能达到数百亿字。① 但这些数字化成果离大数据的运用仍有一定差距。

在大数据视野下，各种分散的地方志数据资源应得到有效整合，建设“全方志数据平台”将是一项体量巨大但又不可或缺的工作。“全方志数据平台”的建设首先需要量的积累，对国内甚至是流传到国外的地方志资源做数字化处理并加以统一整合利用，同时，需要对新生的地方志资料进行扩充，当量的积累达到一定规模，大数据运算的数学模型和算法分析可以发挥应有的功能，经过分析处理的新解构数据又会不断充实原有的数据库。“全方志数据平台”将成为一个不断自我充实的数据体系。② 当“全方志数据平台”具备一定规模，将极大提升资料占有的广度，大数据运算处理技术得以充分运用，相关编研工作的精细化水平将大幅度提升，方志文化在编纂、开发、利用、研究等领域将呈现出新样态。

三　发挥方志资源大数据特性，拓展地方志“计算”研究新领域

最早洞见大数据时代发展趋势的数据科学家舍恩伯格在其《大数据时代》一书中指出，一切皆可量化。③ 伴随信息技术的高速发展，数据已成为一种基础性资源，国务院2015年颁布的《促进大数据发展行动纲要》正式将发展大数据提升为国家战略。④ 在大数据视野下，地方志资源的大数据特性如能得到妥善开发利用，将会为实现地方志事业的大局化、社会化、信息化、全国化、国际化提供强大助力，还会为未来地方志学科发展抢占制高点，同时带来良好的社会效应、文化效应乃至经济效应。地方志资源大数据特性的功能运用可渐进推演出三个层次：

（一）数据源价值的开发利用。在大数据时代，地方志资源首先要体现其应有的数据源价值。方志文化包含海量的资料数据、资料真实度高、类型多样，是一个不断更新丰富的历史、地情资料数据源。现阶段，很多企业机构正为获取数据不遗余力。2014年10月，马云爆出金句：“我们是通过卖东西收集数据，数据是阿里最值钱的财富。”在一次演讲中，他直言阿里巴巴公司本质上是一家数据公司：“我们做淘宝的目的不是为了卖货，而是获得所有零售的数据和制造

① 参见申小红：《对地方志数字化、信息化建设的几点认识》，《新疆地方志》2013年第4期。

② 参见吴玲：《大数据时代历史学研究若干趋势》，《北方论丛》2015年第5期。

③ 参见［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》，浙江人民出版社，2013年，第97页。

④ 参见董西成：《大数据技术体系详解：原理、架构与实践》，机械工业出版社，2018年，第2页。

业的数据；我们做阿里小微金服的目的，是建立信用体系；我们做物流不是为了送包裹，而是把这些数据合在一起。”① 与之相悖，地方志坐拥一座数据宝库却大多被束之高阁乏人问津，这表明地方志资源的数据源价值有待充分发掘和利用。

举一个可供借鉴的例子，早在 2002 年，当时的谷歌公司还没有多大名气，但他们已经启动 GOOGLEPRINT 项目，要把全世界的数字图书馆项目统一起来。2010 年，谷歌已经扫描 1500 万册书，到 2013 年这一数字增长到 3000 万册。② 一方面，大量书籍资料通过这个数据库被重新激活，进入人们的视野；另一方面，越来越多的科学研究正有赖于这一数据资源，许多以往难以解决的课题因为得到其庞大的数据支撑而变得可行。地方志的数据源价值同样可以通过类似的方式得到开发和利用，结合“全方志数据平台”建设和大数据技术的运用，许多地方志资源即便尘封千年同样能重新闪耀熠熠光芒。

（二）融入“云平台”实现方志资源广泛共享。长期以来，地方志资源知晓率低、参与率低、使用率低的状况客观存在，而地方志资源大数据特性的发掘和运用很可能使这种状况得到极大改观。让地方志数据资源融入“云平台”是实现广泛共享的一种有效手段。根据百度百科的定义，“云”是网络、互联网的一种比喻说法，“云计算”则是基于互联网的相关服务的增加、使用和交付模式，通常涉及通过互联网来提供动态、易扩展且经常是虚拟化的资源，云计算甚至可以达到每秒 10 万亿次的运算能力。大数据离不开云处理，云处理为大数据提供弹性可拓展的基础设备，是产生大数据的平台之一。云平台顾名思义，这种平台允许开发者们或是将写好的程序放在云里运行，或是使用云里提供的服务，或二者皆是。

融入云平台需要一定的网络信息技术支撑，但并非高不可攀、遥不可及，现在许多地区都在构建各类云平台，实现与这些云平台的有效对接已有不少成功先例。比如，上海市 2016 年 3 月正式建立起“文化上海云”平台，该平台汇聚上海市 16 个区县的文化馆、图书馆、美术馆、文化活动中心，以及上海博物馆、上海当代艺术博物馆等 529 个公共文化场馆的公共文化服务资源，每年提供近 3000 万人次的活动订单。上海嘉定区依据文化云提供的大数据改变粗糙的考核、管理模式，其公共文化活动平均上座率突破 84%，场所设施利用率突破 90%。③ 在贵州省，贵州地方志信息网于 2016 年 12 月正式在中国·贵州政府门户网站云平台上线，地方志的成果发布、省情地情介绍获得更广泛的关注，地方志资源能够更有效地为社会和公众提供服务。

此外，部分地方志工作机构还在尝试利用自身的方志资源构建独立的方志云平台，比如贵阳方志云平台计划在“十三五”期间建成投用。该平台将首先对贵阳市的方志、年鉴、旧志、期刊、地情书等资源进行整合，同时通过购买、共建、共享等方式，逐步完成对全省乃至全国以地方志书、年鉴为主体的文史类资源制作。此外，贵阳方志云还着力打造综合资源云平台、数据比对云平台、史实核查云平台、质量规范云平台、知识服务云平台、协同创新云平台和志能问答云平台 7 个下属云平台。该平台在“十三五”期间建成投用后，将形成全市、全省乃至全国地方志的网络资源覆盖，数据量将达 50TB。平台将以云计算为核心服务形式，提供定题跟踪、网上

① 潘旭涛、贺璞薇：《大数据最值钱马云爆“数据是阿里最值钱的财富”》，人民网，2015 年 4 月 18 日，http：//media. people. com. cn/n/2015/0418/c14677 - 26865653. html。

② 参见尼克：《计算历史学：大数据时代的读书》，《东方早报》2014 年 6 月 15 日，http：//money. 163. com/14/0615/09/9UP7BQL300253B0H. html。

③ 张熠：《“文化上海云”堪比“文化淘宝”》，《解放日报》2017 年 2 月 6 日，第 1 版。

专题信息、个性化链接、联机检索以及目录查询等公共文化服务，并可实现与其他数字图书馆资源库的关联检索和跨库连接。① 从贵阳方志云平台的建设规划中可以管窥到方志文化与云平台相融合的广阔蓝海。

（三）拓展地方志“计算”研究新领域。马克思曾说：“一种科学只有成功运用数学时，才能达到了真正完善的地步。”② 传统历史学偏重于定性分析，在定量分析上存着明显的局限。1922 年，梁启超曾倡导过“历史统计学”，20 世纪六七十年代“计量史学”又在欧美国家风靡一时，这些理念更多侧重运用数学方法对历史资料进行定量分析。在大数据时代，数据环境越来越健全，量化分析的基础和手段都更加丰富和完善，原有定量分析的视野和思维方式在不断被打破和颠覆。大数据视野下的计算要远远超过以往统计、计量的范畴，一方面，大数据时代，伴随着数据资源的不断丰富完善、各种分析处理技术的迅猛发展，可以实现对海量数据、信息的挖掘、综合和分析，弥补传统研究方式面对庞大信息时搜集、分析上的局限。③ 借助海量的数据资源，通过庞大、精准甚至解构化的数据库，能够认识以往分析方式所无法企及的领域，突破旧有研究的可视极限。另一方面，传统定性分析的格局同样在发生巨变，在大数据环境下，用大量数据作为导向，我们能获取远超前人的具有多维特征的、有代表性的数据，以此为基础的归纳、总结、预测等思维方式都会进入新层次，甚至伴随 AI（人工智能）分析体系的介入，定性研究的基础、手段和理念都将产生重大变革。两相结合，基于大数据计算的新“计算历史学”将进而打破定量分析与定性分析的界限，启发研究者拓展新的研究空间，开启“计算历史”新时代。

借助大数据的蓬勃发展，数据环境越来越完备，“计算历史学”必将成为历史研究的一个新趋势。先来看一个典型的例子，解析一个小型数据库就可能带来不一样的新视野和新成果：2015 年 2 月《习近平用典》出版后，人民网登载《大数据分析：习近平用典 300 句——读〈习近平用典〉》一文，通过梳理《习近平用典》，对其中习近平总书记曾引用的近 300 条典故做大数据分析，结果表明习近平总书记善于运用古代典籍、经典名句来阐述思想，他在讲话中引用次数最多的是源自儒学经典的名言，《论语》《礼记》《孟子》《荀子》《尚书》《二程集》等儒学经典著作都被多次引用；儒家经典之外，道、法、墨三家的经典语录也是习近平总书记讲话和文章中的高频词；而被引用典故最多的古代名人则是苏轼。④ 通过这样的大数据分析，可以清晰看到习近平总书记对中华传统文化的钟爱，也折射出他的一些治理思想和执政风格。

在地方志研究领域也应加快引入相关理念，尽早开启“计算”模式，这将为学科发展注入强劲的新生动力，开拓广阔的新领域。做一个简单类比，地方志是中华传统文化的重要组成部分，要证明方志文化在中华优秀传统文化中的可靠地位，详实的数据将具备强大说服力。前文曾提到，目前现存旧志有 8000 余种、10 万余卷，约占我国现存古籍的 1/10，这就是一个很令人信服的数据，但个别的数据缺乏全面性和系统性，无法形成有效的数据链。根据舍恩伯格在其

① 参见牛悦：《贵阳市方志云平台“十三五”期间建成》，《贵阳日报》网络版，2016 年 5 月 19 日，http：//www. gygov. gov. cn/art/2016/5/19/art_ 10683_ 931629. html。

② ［德］弗·梅林著，樊集译，持平校：《马克思传》，人民出版社，1965 年，第 871 页。

③ 参见马建强：《计算历史学：大数据时代下的历史研究》，《学术论坛》2015 年第 12 期。

④ 参见杜文明、蒋波：《大数据分析：习近平用典 300 句——读〈习近平用典〉》，人民网，2015 年 2 月 28 日，http：//theory. people. com. cn/n/2015/0228/c394175 - 26613538. html。

《大数据时代》一书中提出的理论，在大数据环境下，不再局限于抽取随机样本，而是全体数据分析；不再局限于探求因果关系，而是相关关系的互联；不再局限于追寻精确性，而是包容和解析混杂性。① 大数据全样本分析无疑比局部样本抽取得出的结果更具可信度，能够形成更为全面的数据体系；不必知晓“为什么”，知道“是什么”就已足够，数据自身给出的结论常颠覆以往的认知；不一味强调精确性，海量数据带来的混杂性能打开一扇通向未知领域的窗户。大数据会带来思维的变革，通过由此产生的新视界认知和新模式解构，方志文化在中华优秀传统文化中的地位必然可以得到更多维度的评判，获取更全面更令人信服的量化依据。

现代著名历史学家陈寅恪曾说：“一时代之学术，必有其新材料和新问题。取用此材料以研究问题，则为时代之新潮流。”② 大数据时代，定量分析与定性分析更高维度地紧密融合，泾渭分明的文、理科研究边界被不断突破。传统的地方志无论编纂还是研究，都不能在旧体式中故步自封。融入大数据，对已有的研究观念和方法做出适时调整，在创建数据库、资料数据解构、算法模型开发等方面构建新的方法论，拓展地方志“计算”研究新领域，必将给地方志的发展带来新的跨越。

（作者单位：贵州省地方志办公室）

本文责编：周　全

《曲阜市志（1991—2013）》出版发行

2019 年 6 月，曲阜市地方史志编纂委员会编纂的《曲阜市志（1991—2013）》由方志出版社出版发行。

《曲阜市志（1991—2013）》是 1993 年版《曲阜市志》的续志。该志除序、概述、大事件、人物、附录外，共设 34 编、247 章、922 节，有表格 204 个，图照 511 个，225.7 万字，1155 页。该志主要记述 1991—2013 年期间曲阜市自然、政治、经济、文化、社会等各方面的发展变化情况。在文化编之外，单独设立孔子文化节、文化遗产、儒学研究、文化产业等编，基本涵盖了曲阜文化的特殊性。

该志所记 1991—2013 年正是改革开放走向全面深化的关键时期，23 年间社会发展日新月异，人民生活蒸蒸日上，仅需客观记录，足以华章流彩，睹字馨香。志书不仅记录曲阜发展变迁的详实史料，还折射出传统文化创新发展的时代映像。是一部集资料性、思想性、科学性和权威性于一体的重要地情文献。全书记述全面、内容丰富、重点突出、特色鲜明，既是一部百科全书式的史料典籍，也是一部为社会各界提供翔实资料的大型工具书。

摘自：山东省情网

① 参见［英］维克托·迈尔—舍恩伯格，肯尼思·库克耶：《大数据时代》，第 27 页。

② 陈垣：《敦煌劫余录》，安徽大学出版社，2009 年 12 月，第 2 页。