# 大数据思维与数字人文的加值应用
## ——传统文化数据库发展的新趋势

王　洁（武汉大学文学院）

**摘　要：**文章认为利用大数据研究与传播中国传统文化是大势所趋，并阐述其目前已经过了数字典藏与数字人文两个阶段的发展。中国传统文化相关数据库日新月异，这些数据库被广泛应用于专业的文史研究，其中一些还通过加值应用于传统文化的大众传播。最后，提出在坚持大数据思维，继续积极建设专业、完备、可以加值的数据库的同时，须警惕科技对学术研究与人文思维的负面影响，并且不能忽视相关法律法规的建设。

**关键词：**数据库；大数据；古籍；传播
**中图分类号：**G250.74；G203　　**文献标志码：**A　　**文章编号：**1005－8214(2018)05－0104－05

# Big Data Thinking and Value-Added Application of Digital Humanities
## ——The New Trend of Traditional Culture Database Development

Wang Jie

**Abstract:** This article considers the dissemination of Chinese traditional culture by applying Big Data to be the general trend, and expounds the dissemination process of digital collections and digital humanities. Chinese traditional culture database is changing with each passing day, which is widely applied to the research of professional literature and history, as well as the mass communication of traditional culture by added-value. Finally, it suggests us be alert to the negative impact of technology on academic research and humanistic thinking while adhering to Big Data thinking and continue to actively build professional, complete and value-added databases. Meanwhile, we cannot ignore the construction of relevant laws and regulations.

**Keywords:** Database; Big Data; Historical Books; Dissemination

### 1　利用大数据研究与传播中国传统文化是大势所趋

大数据为中国传统文化的研究和传播带来了巨大的便利。利用数据库所创造的新资源，人们可以突破原有时间和空间的局限，在短时间内查找更多的资料；同时，数据库也为研究者提供了新方法、新思路，为中国传统文化的爱好者提供了新渠道、新平台。利用数据库，研究者可以开辟新的研究领域、进行更科学的研究，传统文化爱好者也可以更全面地了解传统文化、进行传统文化的二次传播。

中国古典文学是隶属于中国传统文化中的一门学科，充分占有史料是进行中国古典文学研究的基础。正如傅璇琮先生所说："古代包括经史子集中的典籍，都与文学史料有关。"[1] 那么，如何占有这么多的史料就是一个问题。在古代，研究古典文学时能够"读万卷书"就算是充分占有史料了；在今天，情况却发生了变化，一个人读万卷书是远远不够的。同样，在今天，中国古典文学爱好者们也不再满足于"熟读唐诗三百首"，想要满足他们的需求，只讲唐诗三百首也是不够的。

最近 20 年，随着数字化的发展，史料也出现了"大爆炸"，即可以利用的史料急剧增加。我国是历史文献大国，现存的古代文献资料数量巨大，今人的研究成果也层出不穷。仅唐宋文学作品及相关的史书、地理、笔记、目录版本学著作等就数以万计。这些资料有着非常重要的史料价值，但在过去，仅有部分专业研究人员会接触这些文献中的一部分。2009 年开

[基金项目] 本文系 2012 年国家社会科学基金重大项目"唐宋文学编年系地信息平台建设"（项目编号：12&ZD154）的阶段性成果。

始，王兆鹏教授领导的《唐宋文学编年地图》研发团队利用专业的考据编年方法将这些文献中的信息整合并数字化，最后以编年地图的方式呈现。2016 年年底，《唐宋文学编年地图》正式上线，使得这些晦涩驳杂的文献能够以最直观的方式系统地展现在大众面前。除《唐宋文学编年地图》外，许多古籍数字化技术公司都在大力开展文献数字化工作，所涉及的文献数量都很惊人。

大量散落在国外的历史文献中，也包含很多重要史料。在过去的 20 年里，一些机构开始建设域外汉籍数据库。如哈佛燕京图书馆耗时 10 年，将馆藏的 4,200 部中文善本特藏全部数字化；由中、英、法、俄、日等多国合作共建的“国际敦煌项目”（IDP）专属数据库，于 1994 年开始筹建，目前数据仍在持续上传。像这样的史料，对中国传统文化研究者和传播者来说极为重要。这些数据库一旦建成，国内学者将看到一大批全新的资料。国内的王兆鹏教授早已看到了这一点，积极与海外相关机构取得联系，充分吸收和利用了域外汉籍资料。

史料急剧增加带来的结果是，无人能自夸读遍天下书，也不会有人止步于唐诗三百首。即使一位学者或文学爱好者过目成诵，要读遍《唐宋文学编年地图》数据库中涉及的书，也是很难做到的。数据库的建立必然引起研究和传播方法的革命，这在大数据不断发展的今天，显得尤为突出。

## 2　如何使用数据库研究中国传统文化

经过学界 20 多年的努力，历史文献的数字化已经取得了很大进展，建成了一些可供传统文化相关专业使用的数据库。建设者们也在积极推进这些数据库的高级化和专门化。中国传统文化资料大数据的形成，为传统文化的研究带来了极大的便利。许多传统文化研究者虽然认同这一点，却又觉得无从下手。其实，数据库在中国传统文化研究中的运用方法有很多种，研究者可以依据自己研究的课题选择不同的方法，以充分利用丰富的大数据资源。

中国传统文化数据库建设之初，是将历史文献资料通过摄影、扫描、影音拍摄、全文输入等方式转化为数字文件（Digital Records），再经过元数据（Metadata）的后台整理，供读者使用。前文提到的哈佛燕京图书馆中文善本特藏资源库和国际敦煌项目专属数据库都属此类。这类数据库实际上是传统图书馆的数字化，被称作数字典藏（Digital Archive）或数字博物馆（Digital Museum）。数字典藏虽然有强大的存储和检索功能，但对使用者来说，其中的史料都是信息碎片，很难被充分使用。后来兴起的数字人文（Digital Humanities）克服了数字典藏的弊端，即通过资料探勘（Data Mining）将数字典藏所汇集的无数史料进行处理、分析，展示出史料内部隐含的关系。它不仅能实现资料检索，而且能为研究者提供一个观察某历史时期具体事件发生的环境，呈现出相关史料之间的时空联系，弥补了史料本身存在方式的不足。在使用这类数据库时，只要输入一个检索词，搜索结果就会形成一个意义的集合。通常情况下，这些数据库还会在每条搜索结果的后面标注史料来源，为研究者观察史料之间的联系提供方便。总之，借助数据库提供的这些方法和便利，研究者能更好地进行人文研究。

中国传统文化的研究者可以根据自己的研究课题使用不同的数据库。对许多研究者而言，最常见的方法是使用数字典藏数据库，从海量的数字文件中检索出自己需要的史料，然后对这些史料进行筛选和甄别，以便得到最可靠的史料，进而找出这些史料之间的深层联系。如有学者谈到，自己为撰写某位历史人物的年谱，用了三个月的时间，查阅了与之相关的所有史料，从而发现了许多不为人知的文献，并以此为依据得出若干新结论。正是得益于有检索功能的数字典藏数据库，学者们才能如此高效地查阅到前代学者尚未穷尽的史料。否则，依靠纯人工的方法，用十几年工夫也未必能遍览这些文献。

研究者们还可以使用更高级、更专业的数字人文数据库进行学术研究。经过学界的不懈努力，现已建成一些可供传统文化专业研究使用的数字人文数据库。与数字典藏不同，数字人文是由人文计算（Humanities Computing）发展而来的。因此，数字人文数据库除了具有数字典藏汇集了大量史料的基本特征外，还可以实现史料探勘视觉化、量化分析、建立 3D 模型、数字原生（Bom-Digital）论文、动态环境制作、自造实境空间等功能。即这些数据库中的史料信息依据某种人文理念被整合加工过，可以全方位清晰呈现并被直接使用。以现今发展较为成熟的《唐宋文学编年地图》为例，经过近 10 年的努力，该数据库现已收入唐宋时期大部分重要作家及作品的相关史料，并不断参照今人的研究成果予以更新。它通过资料探勘技术处理、分析数字文件信息，将这些数据信息结构化，数据结构框架由时间、人物、地点、事件四部分构成，最终用立体、多元的地图把古代文学人

物事件关系呈现出来。通过该数据库对数字文件的探勘、分析和视觉化，研究者既可以对文学作品开展深入研究，又可以得到相关历史人物、事件的时空分布及复杂的作家作品关系网络。值得一提的是，这也是迄今为止完全由我国本土独立开发的、专业性最强、影响最大的古典文学数据库。其中的史料也可广泛应用于哲学、历史、语言学、民俗等中国传统文化相关学科的研究中。

除此之外，具备条件的中国传统文化研究者还可以根据研究的需要，建立课题组专用数据库，依此展开相关研究。如有语言学研究团队把汉江流域方言的音频信息建成一个数据库，开展汉江流域语言文字的传承与流变研究。依据这些数据库得出的结论，已经开始受到学界的重视。

以上这些都表明，中国传统文化研究者不仅应当使用相关数据库，而且可以使用好这些数据库。

### 3　数据库的加值应用与中国传统文化的传播

从 20 世纪 40 年代末，Roberto Busa 用人文计算建立 St. Thomas Aquinas 的著作索引数据库“Index Thornisticus”[2] 开始，到 2005 年澳、加、日和欧洲各国数字人文联盟（Alliance of Digital Humanities Organizations，ADHO）成立，大数据日渐成为人文研究的一种趋势，甚至一种运动。在这种趋势下，《四库全书》《中国基本古籍库》《唐宋文学编年地图》等一批中国传统文化数据库相继建成，并全部可以应用于学术研究。其中能实现加值应用（Value-added Applications）的数据库并不多见。通过对数据库加值应用来传播传统文化，不仅意味着传统文化网络传播方式的革新，更是将数字学术（Digital Scholarship）引入大众传媒的有益尝试。《中国基本古籍库》和《中国历代人物传记资料库》就属于未实现加值应用的数据库，而《3D 实景莫高窟》和《唐宋文学编年地图》就属于实现加值应用的数据库（见图 1）。

网络时代文化传播的最初形态是将传统媒介网络化，推出视频音频节目和纸质文献文章的网络版。在此基础上，搭建多种受众自主选择的平台。如喜马拉雅 FM 就是一个音频节目的手机选择平台，微信公众号则是网络文章的微信选择平台。传统媒介的网络化作用很大，对许多文化传播者来说，这也是最简单的方法。即通过在自媒体（We Media）上写文章、录视频，将自己对传统文化的认识以文字、视频或音频的方式记录下来，然后发布到网络上，以电脑或手机客户端的形式供受众选择性阅读和转发。很多上班族谈

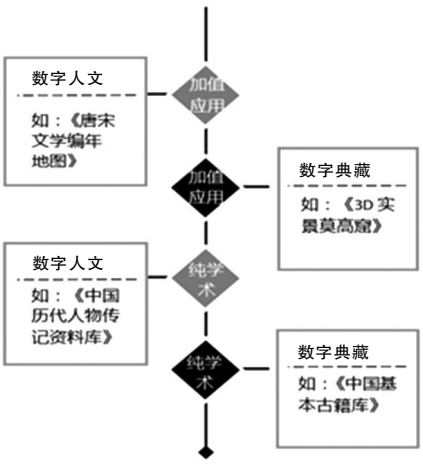


图 1　数字人文与数字典藏加值应用举例

到，自己利用开车时间用手机软件收听历史文化类讲座，一个月的时间已经对《资治通鉴》的基本内容有了一定了解。他之所以能在没时间翻书的情况下了解到这么多历史文化知识，就是得益于网络化传播方式。否则，终日忙忙碌碌，又想用传统文化知识充实自己，是很难做到的。然而，对受众来说，这些传统文化知识只是换了一种呈现方式，依然难以直接、迅速地应用于生产生活实际。加值应用的数据库就克服了这种网络原始传播方式的弊端，即建立的数据库不仅可以展示丰富的历史文化知识，而且还可以根据受众的即时需求选择性提供其可能需要的传统文化信息。许多数据库还可以呈现出这些信息之间的某种趣味性或实用性联系，以弥补网络时代原始文化传播方式的不足。在使用这些数据库时，只要输入一个检索词，搜索结果就会是一个与使用者需求高度匹配的传统文化知识集，并为使用者提供各种角度来了解其中的关联，从而更直接、更迅速地满足使用者的求知欲或生产生活需求。总之，加值应用的数据库使传统文化更好地应用于生产生活实际，革新了网络文化传播方式。

无论是数字典藏还是数字人文都可以通过加值应用，向大众传播传统文化。中国传统文化的传播者可以根据自己的行业需求，采用不同的方法来选用不同的数据库。对许多传统文化传播者来说，最简单的方法就是用检索的方式从海量的数字典藏文献中搜寻自己需要的资料，然后将这些资料整合成一个能满足传播需求的资料集，再通过电子媒体展示出来，从而达到传播的目的。如应用数据库《3D 实景莫高窟》传播敦煌壁画文化时，就是通过检索的方式，将所需的

敦煌壁画原貌整合起来，直接呈现给观者，很好地解决了敦煌文化传播中的地域限制和洞窟保护问题。再如，台北县某小学中年级任课教师就将若干艺术与人文专业数位典藏网站的内容选择性整合，编成可通过电子媒体向学生展示的“数位典藏网站内容教材”，实现了教学方式的更新。在教学活动结束后，教师们对学生学习成效进行了测验。测试结果显示，相较于“传统纸本教材”，“数位典藏网站内容教材”在某些课程上，传播效果显著。更重要的是，随后发放的《学生学习兴趣调查接受问卷》中显示，“学生觉得上艺术与人文领域课，使用数位典藏网站教材是有趣的，而且喜欢老师用数位典藏网站来教授艺术和人文领域课程及期望以后还能学习数位典藏网站内容课程。”[3] 由此可见，数字典藏比传统纸本更能提高接受者的兴趣，在某些领域，还能收到更好的传播效果。

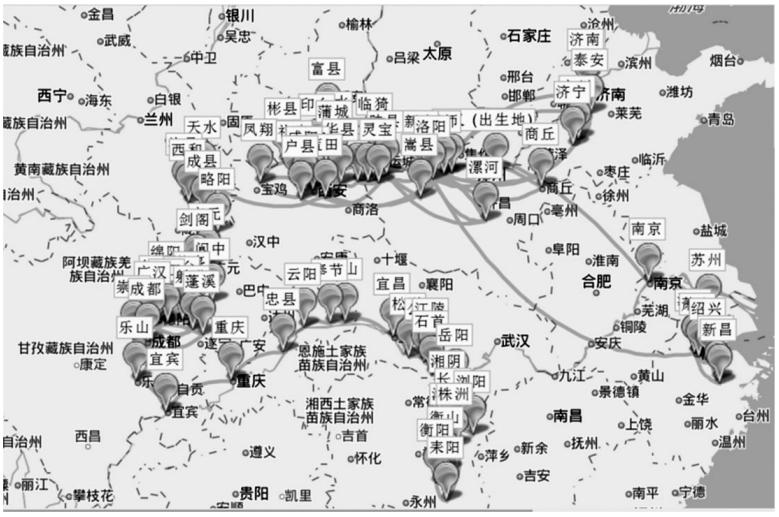


图 2 唐代诗人杜甫行迹图（引自《唐宋文学编年地图》官网）

文化传播者也可以选用更高级、更专业的数字人文加值应用来传播传统文化。传播者不再需要自行检索、整合单一的原始资料，而是可以直接利用这些数据库自动生成的知识集进行文化传播。通过加值应用，数字人文数据库中的史料摆脱了晦涩、陌生、单一的观感，以综合、多元、立体、平易近人的姿态化成一个个知识主题与大众接触，从而提升了文化古籍的附加价值。言及此处，不得不提数字人文加值应用的典范《唐宋文学编年地图》。经过多年专业的学术论证和技术研讨，到今天《唐宋文学编年地图》已经搭建成形式简洁、内容丰富的唐宋文学传播平台。它更加贴合受众生产生活实际，将古典文学知识浓缩到历史编年地图上，古典文学知识由古代地理、历史人物、历史故事、诗文集、行迹等部分构成。具体而言，它主要从三个维度上实现了唐宋文史知识的加值应用。

① 在人物维度上，只要输入一个历史人物的名称，该数据库就会自动生成一张地图，上面标注了该人物所有的仕宦经历、行迹、作品等。通过在人物维度上的加值应用，增强了史料的可读性和故事性，大众得以更全面地了解这些对中华文明有着巨大影响的历史名人。它提供的多元而丰富的素材，也可以直接应用于文艺创作、文化产品制作、文化旅游导览等众多领域。② 在地理维度上，近年来地方政府高唱的地域文化，其实就是我们身边环绕的地理人文。通过数字人文加值应用，这些地理素材能够更广泛地被发现、阅览，为大众所了解和使用。输入一个城市或景点名，相关的人物、故事和作品就会以年代的顺序呈现出来。它的迷人之处在于，将数千年的自然、人文、风俗、艺术等融为一体，沉淀在我们周遭的生活中，丰富着我们的生存体验。③ 在作品维度上，当一首诗被置于山川河流的实景当中，以故事的形式向我们展示它所富含的文化意蕴时，先民的精神、思想和生活方式也会生动地显现出来。在中国生活的人们，就会产生共识和投入感，并清楚地识别中国文化特质，对传统文化产生认同。以阿多诺为首的法兰克福学派曾极力批判文化工业（Cultural Industry），认为大量复制的文化商品（Culture Commodity）是一种低下的模仿。而今天借助专业的数字人文数据库来生产文化商品就是摆脱这种低下模仿的一种方式。现在已经有网友根据《唐宋文学编年地图》独辟蹊径，提出开辟“跟着驴友李白游湖北”主题旅行线路、建立“东坡美食”创意餐厅等想法。

其他行业的工作者也可以根据需要，建设与中国传统文化相关的加值应用数据库，以此为工具提高工作水平。如有些行业的从业人员把植物的图像信息搜集整理起来，建成一个数据库，并在数据库中录入与该植物相关的古典诗词，扫描识别植物的同时展示了植物的人文美学意义，以此开发成一款广泛应用于花卉零售行业的 APP“有形有色”。这样一些数据库创造的商业价值，已经受到了相关各界的重视。

以上这些都表明，中国传统文化相关数据库不仅能应用于学术研究，而且还可以通过加值应用服务于文化传播。因此，文化传播者应当使用好传统文化数据库。

**4 传统文化大数据在不断更新，需谨慎前行**

近些年国内传统文化的研究者和传播者，在使用大数据的认识和运作上都有了长足的进步。2014 年 8 月，在“中国英汉语比较研究会第 11 次全国学术研讨会暨 2014 年英汉语比较与翻译研究国际研讨会”上，学者们就“大数据时代如何更好地翻译中华文化经典”展开讨论；2014 年 11 月，有学者在“世界中医药学会联合会中医药传统知识保护研究专业委员会第二届学术年会暨中医药传统知识保护国际学术大会”上宣读了《我国传统医药非物质文化遗产名录数据分析》的论文；2015 年 11 月，在“第十届全国体育科学大会”上有学者提出了“大数据时代太极拳国际化推广”的问题；2015 年 12 月，在“传承与开启：大数据时代下的史学研究国际学术研讨会”上，专家们围绕“大数据在历史研究中的技术问题”展开研讨。然而也应看到，尽管这方面的工作已经取得了一些进展，但要更普遍、更理性地使用大数据研究和传播中国传统文化，还有很长的路要走。

一方面，在大数据成为一种趋势的今天，传统文化的研究者和传播者都要与时俱进、继续前行。① 尽管社会各界在纷纷建设专业、完备、方便使用的传统文化数据库，但由于中国传统文化相关学科的特殊性，目前数据库建设的整体水平还不高。已建成的数据库数量有限，能直接服务于学术研究和大众传播的数据库更是寥寥。而且，其中只有一部分有较好的研究和传播效果。② 使用数据库来研究和传播传统文化，目前还处于尝试阶段。有的学者过分依赖已有数据库中的信息，认为用数据库来从事学术研究，就不需要再查找原典了。王兆鹏教授就曾提到，大数据为古典文学研究提供了新方法，但不能完全替代传统的学术研究。因此完全忽视和过度迷信大数据都是不可取的。

另一方面，尽管要更好地利用数据库研究和传播传统文化还需要从业人员长时间的努力，但也不能盲目冒进。中国传统文化的研究者和传播者都应当看到大数据背后隐藏的问题，需谨慎前行。首先，应当注意到数据思维与学术思维的显著不同。大数据只认可相关关系，学术研究更多的是探寻因果关系。在使用大数据从事学术研究时，如果不能警惕数据思维的漏洞，将相关关系和因果关系相混淆，很容易就会得出荒谬的结论。其次，大数据时代决定研究质量和传播效果的主要是数据集的质量、数量和利用方式，而需要创新、思辨的人文思维变得相对容易。当大量数据以数据库研发人员希望的方式自动涌现时，不能以联想代替灵感、以形象代替形象思维、以数字编码代替人文创作，要警惕人文思维被架空，防止思想为技术所裹挟。最后，需要提到的是数字化工作的法律规范问题。世界知识产权组织政府间委员会（World Intellectual Property Organization-Intergovernmental Committee, WIPO-IGC）是现在国际公认的对传统文化资源保护最有成效的国际组织，但就传统文化数据库资源持有人（国）权益保护、惠益分享和强制公开等基本问题，尚处于讨论阶段。诸如智慧财产权的保护、文化资产的保存、数字化成果保护等已有的法律规范也是传统文化数据库开发者和使用者应当了解并遵守的。

大数据正在改变传统文化传承的现状和未来，但正如电脑不能完全取代人脑，数据库的使用也不可能取代文化研究者和传播者的主观思考、取代相关行业基本的理论和工作方法。想要更好地使用数据库传承中国传统文化，要求研究者和传播者充分发挥主导作用，利用数据库扩展和更新人文知识，而不是数字化复制已知的知识。惟其如此，自己才不会落后于时代，大数据才会不断完善，传统文化才能更好地传承。因此，传统文化工作者要转变观念，勇于应用大数据，将最古老的学科和最新的科技完美结合，把中华文明的传承带入新时代。

**[参考文献]**

[1] 傅璇琮. 应当重视古典文学的史料研究——中国古典文学史料研究丛书总序 [J]. 文学遗产，1997（2）：16—18.

[2] Burdick Anne，et al. Open Access eBook [M]. Cambridge：The MIT Press，2012：122—124.

[3] 林羿妏，林佳蓉. 台北市国小教师运用数位典藏素材融入教学之研究 [J]. 教学科技与媒体，2009（11）：219—225.

**[作者简介]** 王洁（1987—），女，武汉大学文学院中国古代文学博士研究生，研究方向：中国古代文学，历史语言学。

**[收稿日期]** 2017—09—13　　**[责任编辑]** 阎秋娟