

互动与共生:数字人文与史学研究

——第二届“北京大学数字人文论坛”综述

□朱本军* 聂 华

摘要 北京大学图书馆和北京大学人文社会科学院联合举办的第二届“北京大学数字人文论坛”,以“互动与共生:数字人文与史学研究”为主题,围绕“数字史学与量化史学”“数字工具 GIS 与史学研究”“史料与历史文献数字整理”和“史学研究数字环境与网络基础设施”四个方面进行了深入讨论,对高校图书馆实践数字人文提出了一些新主张和新看法。

关键词 数字人文 会议综述 北京大学数字人文论坛

分类号 G250

DOI 10.16603/j.issn1002-1027.2017.04.003

数字人文的核心目标是利用数字技术有效地提出和解决人文问题,其对历史学相关研究的促进已初见成效。2017年5月,北京大学图书馆联合北京大学人文社会科学院主办的第二届“北京大学数字人文论坛”,探讨了数字人文与史学研究的互动与共生关系。本次会议注册参会人数277人,其中教学科研人员138人、学生106人(本科生占20.75%、硕士生占43.4%、博士生占35.85%)、企业界代表33人,年龄分布在18—60岁之间(18—25岁占23.47%、26—30岁占22.74%、31—40岁占28.88%、41—50岁占16.25%、51—60岁占8.3%),参会人员所从事的学科领域主要集中在人文学科(44.77%)、图书情报与文献学(40.43%)和信息技术学科(9.74%)。在数字技术促进史学研究及史学研究越来越依赖数字技术的互动过程中,图书馆应当承担什么样的角色和任务?笔者试图从图书馆的角度对此次会议的主要内容进行综述和反思,与图书馆界同行共享。

1 数字人文与史学研究议题

历史学是一个跨时空且包罗万象的庞大领域,其主要内容是对客观世界已发生的一切物质运动发展过程进行探索和研究。尽管所谓的“历史”从大类上可以分为自然史和人类社会史,但当前国内外数

字人文关注的主流领域主要还是人类社会史。为突出本次会议“互动与共生”的主题,在对北京大学及国内与历史学研究相关科研院所的研究领域进行调研的基础上,会议选定了“史学研究”“历史地理”“历史文献”和“史学网络基础设施”四个数字人文可快速介入且有一定实践案例的议题。

• 史学研究议题,探讨如何利用数字技术促进对古代史、近现代史、世界史或区域国别史、专门史(如经济史、文化史、军事史等)等方面的研究;

• 历史地理议题,探讨数字技术(特别是地理信息系统、遥感技术)如何促进传统沿革地理、政区地理研究,如何进行考古和地理空间分析等,提出或解决新的史学问题;

• 历史文献议题,探讨利用数字技术将古舆图、古籍、拓片、地方文献、特色文献等的数字对象化、光学识别(OCR)、标引、可视化与文本挖掘,促进史学研究;

• 史学网络基础设施议题,探讨如何构建历史研究的数字化环境,方便开展史学研究。

在嘉宾的报告中,我们确实看到了“数字人文”和“史学研究”两个领域的交融,很好地诠释了“互动与共生”的会议主题。

* 通讯作者:朱本军,0000-0001-6240-7015, bjzhu@pku.edu.cn。

2 互动与共生视野下的数字人文与史学研究

2.1 作为传统史学研究补充的数字史学与量化史学

数字人文落脚在史学研究领域,究竟改变了“史学研究”什么?从本届论坛的讨论来看,不仅仅只是为传统史学研究提供数字技术及衍生品以辅助解决传统史学问题,更重要的是衍生出了一些超越传统史学研究边界和内容的新领域——“数字史学”和“量化史学”(有时也称“计量史学”)。

(1) 史学研究中引入社会科学研究方法。

尽管重视个案解释的历史学与重视整体理解的社会科学在研究方法上存在分野,但是实际情况是人类遗留下来的史料并没有划分彼此,而是杂糅在一起的。香港科技大学康文林(Cameron D. Campbell)教授在题为“历史大数据开启社会科学新认识:李康研究团队理念”的报告中认为,通过历史大数据和社会科学研究方法的结合,不仅可以衔接两个领域,而且可以超越传统历史学或社会科学,提供一些新认识。在其报告中,“历史大数据”被定义为融合多种属性、适合定量分析和可视化展示的高质量结构化数据,其目的不是用于文献检索,而是用来做量化分析和可视化展示,像香港科技大学李中清、康文林团队(以下简称“李康团队”)所主持的“中国多代人口数据库”^[1]“精英教育”^[2]和“清代官员”(缙绅录及相关资料数据库)^[3]等项目都是这个思路。他认为历史大数据由于是对已经发生数据的集中结构化,其相对于当代社会科学数据,在长期性、大规模、细节性、可连接性,和进行国际比较或按地域进行比较的潜力方面更有优势。

(2) 史学研究跨界到社会科学领域。

一般意义上的史学研究,主要是定性的个案分析和小样本定量分析,尽管史料的丰富性涵盖社会科学领域,但真正的历史学者很少去做社会科学(如经济学、金融学等)领域的研究。不过,随着大规模结构化史料数据库的出现,传统史学研究开始跨界到原先很少涉足的领域。中国人民大学的林展博士在“建立‘量化历史研究数据库’的若干思考”的报告中,用回归法分析“清代妻妾价格”,说明了跨历史学和社会科学研究的可能,并且诸如此类的研究思路已经形成一个新的领域——“量化史学”,来补充传统史学研究的不足。

2.2 数字工具与史学研究

20世纪90年代以来,地理信息系统(Geography Information System, GIS)已经成为处理和操纵时空数据最有力的工具。在众多数字技术及衍生品(如数字工具、数据库等)中,本届论坛选取GIS工具,分析它对史学研究各方面的促进作用。

(1) GIS促进传统历史地理研究。

“历史地理学”作为传统史学研究的一个重要分支领域,在顾颉刚、谭其骧、史念海、侯仁之等老一辈学人的倡导下,取得了不朽成绩。GIS作为一种对地球表面空间地理数据进行采集管理和分析显示的软件系统,运用在传统历史地理研究中,既可以辅助解决史地研究中的传统问题,对已成定论的问题提出质疑,又可发现新问题,并尝试回答新问题。北京大学图书馆朱本军副研究馆员以“GIS辅助战国历史地理研究”为题,分享其利用常见的三种GIS工具(Google 地形、Google 地球和 ARC GIS),辅助“战国古地名方位定点标绘”“战国诸侯交通线路勘研”“战国河流水系走向”和“地缘战略与军事地理教学”,解决了“武始”“磨笄山”“长平—丹水水系”等传统史地研究无法定位或定位偏颇的问题,发现公元前350年魏国疆域和交通线路问题,定位秦魏“女戟”——“軹道”——“南阳”交通线路,并对已定论的战国“高阙塞”位置提出新看法。认为数字时代要确定古地名方位,除了传统的史料推定、文物史料互证之外,还应利用GIS辅助发现问题,并尝试结合史料、出土文物和考古遗址,解答新发现的问题。

(2) GIS空间分析与跨学科交叉研究。

构建历史地理信息系统(Historical GIS)是一个十分基础的工作,它既可有效支持地理空间分析,又可有效支持跨学科交叉研究。国际上,通过GIS开展空间分析和跨学科交叉研究的机构不乏其例,美国斯坦福大学的空间与文本分析中心(Center for Spatial and Textual Analysis)^[4]就是这样一个集空间历史、人文设计和文学实验室为一体的跨学科交叉实验室。复旦大学历史地理研究所路伟东副教授的报告——“HGIS 20年:从数据生产到交叉研究”,对复旦大学与哈佛大学联合构建的中国历史地理数据库(CHGIS)^[5]进行了回顾,认为这种基础性工作把千年时间尺度下的沿革地理学、历史地图学和中国当代历史地理学中关涉地理信息数据的时间序列、隶属关系、继承关系标准化到了现代地理学的科

学体系中,为空间分析和跨学科交叉研究提供了便利。这种时间序列下的空间分析如何做跨学科研究?路伟东副教授以 CHGIS 基础地理信息成果为基础,举了两个例子:一是对存在解释分歧的晚清宣统人口调查数据进行人口重心建模和空间分析,可以验证其中一种解释的合理性^{[6][7]};二是对宣统时期 1000 人以上聚落与人口空间集聚的时空分析,有利于了解中国整体以及区域人口发展的脉络。

(3)GIS 模拟与史学新问题。

通过史料和 GIS 的结合,反向模拟一定时空序列下的事件经过,可以提出并回答一些传统史学难以涉足的新问题。陕西师范大学西北历史环境与经济社会发展研究院 GIS 实验室的潘威副研究员在报告——“‘数字历史黄河’平台应用个案”中,展示了其团队将河工史料,如涉及清代以来黄河流域的水利工程、河务财政、流域环境和社会等方面的治河类书、清代河务奏折、民国档案、共和国档案、民间文献、碑刻题刻等,与 GIS 基础地理信息数据(如三维微地貌的高精度复原、河道形态的高精度复原、灾害性降雨特征的重建、工程模拟与地形模拟等)相结合设计的三维动态数字历史黄河 GIS 系统平台,透过这个平台以及相关数据,他们不仅模拟出了一定时间序列下黄河水位淹没范围、水汛报告传递路径,而且还提出了“清前中期东河岁修河银制度的缺陷”等一些传统史学难以回答的史学新问题。

2.3 史料与历史文献数字整理

印本史料的数字化改善了传统史学研究中资料检索的方式,不过在数字人文大背景下,这已经不是新鲜事。本届论坛更多关注史料的大规模文本化与结构化,以及强调文本上下文语境与文本分析环境的构建。

(1)史料的大规模文本化与结构化。

全球网站数量已突破 5 亿、全国开放各类档案 10376 卷、CADAL 数字化 250 万册图书、GOOGLE 数字化 1500 册图书……这是目前的数据现状。浙江大学图书馆黄晨研究馆员在“大数据视野下的数字人文”的报告中,以学术论文中的“大数据”“数字人文”这两个关键词与关键词“图书馆”的强关联说明了大规模的结构化数据对数据分析和可视化的意义。黄晨研究馆员认为,目前的核心挑战之一是处理海量的非结构化数据,而要处理非结构化的数据,需要对数据进行清洗(包括同名处理、机构名目规

范、学科映射等)、可视化工具开发、数据使用等。南京大学艺术学院陈静副教授在“概念模型与中国现代广告社会”的报告中,提供了把散落在报刊中的图文广告从数字化到知识化的方法(如图 1 所示):



图 1 图文广告从数字化到知识化的方法

通过对 20 世纪的《汉口中西报》(1906—1937)、《越华报》(1931—1938)、《大公报》和《申报》中的黑白、卡通风格的图文广告进行数字化、元数据著录,可以为文化学者、历史学家等提供有用的广告数据,比如挖掘哪些国家的商品广告多、哪些商品广告多、哪种类型商品广告多、哪些公司广告多,理解商业文化生活是如何在 19 世纪末 20 世纪初进入中国并渗透入历史的。南京大学历史学院世界史学者王涛副教授提供了另一个类似的例子,通过将 56 卷《德意志人物志》中的 26500 个历史人物的数据结构化入库,把数据转化为知识,提出“谁是德国人”、“德意志人的平均寿命”、“不同职业人员的平均寿命分布”、“德意志是否是哲学家的国度”等问题,并很容易得到答案,这是传统史学研究不可想象的。

(2)史料文本语境和观察文本语境环境的构建。

现今数以亿计的历史文献文本数据往往让人文学者如大海捞针不知所措。台湾大学资讯系特聘教授项洁在报告——“数位人文与脉络分析:历史数据库的视野与取向”中认为,传统历史研究相关的史料数据库要求高求准率(即检索结果中有多少是使用者需要的)和高求全率(即使用者所需的文献有多少被检索到了),不过提供文本语境(textual context,也称“文件脉络”)和可分析操作与观察解释的环境(observation platform)作为大数据时代探索未知的重要内容尤应引起重视。其中的文本语境,包括图书馆行业所熟悉的基于元数据的上下文关系(metadata contexts,如文本中出现的时间、地点、出处或来源、作者、主题等)、基于统计的上下文关系(statistical contexts,即对检索结果、元数据进行量化统计和分析)和基于语义的上下文关系(semantic contexts,即针对文本的特性找出原始文本间特有的关联并展

示彼此之间的关系)。一个典型的例子,是他们通过扫描复制、建立元数据、全文输入、全文标记(如时间标准化、地理位置定位、大致分类等)“台湾明清档案”(含 46721 件明清奏折、题本、谕旨、照会、笔记、方志、起居注、内阁大库等档案资料)、台湾“古契约文书”(含 40454 件 20 世纪前台湾民间的古契约文书)和“淡新档案”(含 14000 余件清代台湾北部淡水县与新竹厅地方行政档案)而建设的“台湾历史数位图书馆(THDL)”系统^[8],这个系统可以很方便地基于元数据、基于统计和基于语义分析上下文关系,让人文学者有鸟瞰史料的能力和更广阔的诠释空间。在如何对文献或文本上下文关系进行关联和揭示方面,加州大学伯克利分校东亚图书馆的林海青先生以数据化宋代官修《崇文总目》和南宋私藏《直斋书录解題》目录为例提供了一个案例。林海青先生认为数字环境下去“辨章学术,考镜源流”,将各类历史文献或著作按照科学、系统、辨证的原则进行分类,将各种学术派别和流派的作品、论著进行梳理分门别类两个工作极为重要,一是对历史文献进行描述,二是对历史文献的关系进行揭示,并认为采用 RDF 资源描述框架的关联数据(linked data)是将文献数据转化成机器可读可理解格式的一种方法,是中国古典目录数据化的有效方式之一,并在实验中成功实现。

2.4 史学研究数字环境与网络基础设施

数字人文视野下的历史学者如何开展研究,在需求的推动下已经开始有一些共识。这种共识推动人文数字环境和网络基础设施建设,未来必将成为数字人文领域的新锚点。

(1)可信的数据基础设施建设。

从“量化史学”的角度,中国人民大学清史研究所的林展博士和香港科技大学的康文林教授的报告都认为大规模、高质量、公认且统一的结构化史料数据库非常关键,不仅可以减少重复性劳动,而且各领域学者也不用花时间去争论数据来源是否可靠,而是把精力和争论都集中在所研究的问题上。目前无论是陈志武团队主持的“中国利率数据库”“历史时期吏治数据库”“清代府级治理信息数据库”项目,还是前述李康团队所主持的“中国多代人口数据库”“精英教育”和“缙绅录”历史大数据项目,都是此目标。上海图书馆刘炜研究馆员在“以数据重现历史——数字人文的技术方法体系”的报告中认为,将

数字图书馆(digital library)变为数据图书馆(data library),并免费公开数据,是面向人文研究的国家数据基础设施非常重要的一部分。

(2)网络基础设施建设。

所谓“网络基础设施(cyber infrastructure)”,维基百科(Wikipedia)解释为一种跨机构、跨地区,且支持数据采集、存储、管理、整合、挖掘、可视化和计算处理的研究环境。哈佛大学费正清研究中心徐力恒博士的报告——“服务中国史研究的网络基础设施——CBDB 的探索和建议”认为,所构建的这种网络基础设施,可以帮史学研究者采取互通的方法(如 API 数据接口)打开并利用各种研究资料,促进群体或个人研究。上海图书馆刘炜研究馆员认为数字人文的基础设施是一种支持人文科研活动的基础设施,可以分为三个层次(如图 2 所示):其核心是由文献资源及其服务机构组成,提供基本研究素材的保障;中间层由基金会、资源库、机构仓储、计算设施、系统平台、工具软件、领域专家和数据科学家等构成,是数字人文研究活动的主体;外围是数字人文成果发布、与社会交互、产生社会影响界面层,以门户或平台形式呈现。



图2 数字人文网络基础设施

(3)数字人文技术体系。

上海图书馆刘炜研究馆员在报告中建构了一个支持人文学者发现、收集、比较、发布、协作的技术体系框架(如图 3 所示)。他以上海图书馆为例,规划上海图书馆未来数字人文技术体系的主要工作将集中在“基于资源的书目控制”“领域资源对象化、数据化”“人地时事基础数据的模型化本体化”“提供可视化、各类处理及分析挖掘工具”“服务系统平台化、社会化、互动化”“数据增值:语义出版、虚拟现实”等方面,并会从上海图书馆馆藏家谱、盛宣怀档案、手稿着手,逐步建成“上海记忆”和中国近代历史人文数

据中心,在开放度上做到支持开放建设(众包)、开放存取、公众参与,支持关联数据和语义查询,支持文本分析、图像标注、时序空间分析、社会关系分析等基本分析功能和支持基本的可视化展示,通用数据(人地时事)提供标准的公共数据服务。

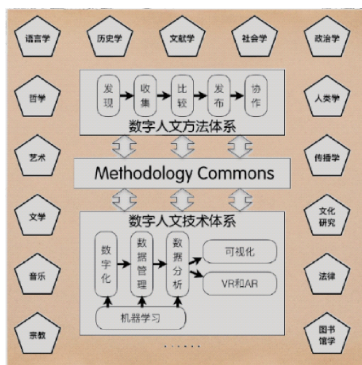


图3 数字人文技术体系

3 结语

学科服务是高校图书馆支持教学科研的一项基础服务。不过,在数字人文概念下,学科服务有了新内涵,需要图书馆在文献资源建设与服务方面做一些改变。以本次会议的“支持史学研究”为例,图书馆可以在以下方面进行新的布局谋篇:

一是将文献的数字对象化思路转变为数据化(即数字文本)思路。以古籍为例,尽管古文献的数字影印在一定时期还能满足史学研究者的部分需求,但从长远看,这种不彻底的数字化实际上囿史学研究于原地,无法开展基本的文本分析与关联。

二是将非结构化数字文本进行大规模结构化。结构化的文本数据会赋予史学研究新的内涵,也是将传统史学研究范畴边界向外延展的重要基础支

撑。从事数字史学和量化史学研究的学者已经严重感觉到此方面的不足。

三是史学研究工具的推介与培训。利用新的数字工具可以提升传统研究的效率,有目的开发、推介、培训史学研究数字工具会非常有帮助。

四是数字人文基础设施的构建。这种宏大的基础设施不会仅限于某个单一的学科,将是跨学科的,包括可信任的基础数据设施、可互操作的网络基础设施和技术工具。

参考文献

- 1 China Multi-Generational Panel Datasets (CMGPD) [EB/OL]. [2017-06-16]. <http://www.shss.ust.hk/lee-campbell-group/china-multi-generational-panel-databases-cmgpd/>.
- 2 China University Student Dataset (CUSD) Project [EB/OL]. [2017-06-16]. <http://www.shss.ust.hk/lee-campbell-group/china-university-student-dataset-cusd-project/>.
- 3 China Government Employee Database-Qing (CGED-Q) [EB/OL]. [2017-06-16]. <http://www.shss.ust.hk/lee-campbell-group/china-government-employee-database-qing-cged-q/>.
- 4 AboutCESTA [EB/OL]. [2017-06-16]. <https://cesta.stanford.edu/about/about-us>.
- 5 China Historical GIS [EB/OL]. [2017-06-16]. <http://www.fas.harvard.edu/~chgis/>.
- 6 曹树基. 中国人口史(第五卷,清时期)[M]. 复旦大学出版社, 2001.
- 7 葛剑雄. 中国人口史(第六卷,1910-1953年)[M]. 复旦大学出版社, 2005.
- 8 台湾历史数字图书馆 [EB/OL]. [2017-06-16]. <http://thdl.ntu.edu.tw/index.html>.

作者单位:北京大学图书馆,北京,100871

收稿日期:2017年6月18日

Interaction and Coexistence: Digital Humanities and Historical Research

—Conference Review for “The 2nd Peking University Digital Humanities Forum”

Zhu Benjun Nie Hua

Abstract: The 2nd Peking University Digital Humanities Forum, jointly organized by Peking University library and Institute of Humanities and Social Sciences of Peking University, presented various topics such as “digital history and qualitative history”, “GIS and historical research”, “historical books and literature collation”, “digital environments and cyber infrastructure for historical research” and discussed possibilities and approaches of practicing DH in university libraries.

Keywords: Digital Humanities; Conference Review; Peking University Digital Humanities Forum