DOI: 10.13530/j.cnki.jlis.160022

面向人文研究的国家数据基础设施建设*

刘 炜 谢 蓉 张 磊 张永娟

摘 要 人文科学研究正在向以数据为驱动的新型研究模式转型，形成所谓“数字人文”研究新范式。“数据科学”为这些变化提供了方法论基础，支撑研究活动的基础设施也需要经过改造或重新建立。新的人文研究平台以数据为基础，以方法为导向，提供统一的数据资源管理、大数据分析、可视化展示和智慧型服务。这些变化将带来更大范围和深度的数据资料的应用和协同研究，将使研究人员以过去难以想象的尺度和维度提出问题。数字人文基础设施建设可以是国家层面的，也可以是地区行业或组织机构层面的。图书馆、博物馆、档案馆等人类记忆和文化遗产机构近20年来馆藏内容大规模数字化，可以通过升级为数字人文研究平台，转型为数字人文基础设施的重要组成部分，继续在数字时代发挥其巨大的价值。图1。表3。参考文献13。

关键词 数字人文 基础设施 数字图书馆

分类号 G250

Towards a National Data Infrastructure for Digital Humanities

LIU Wei, XIE Rong, ZHANG Lei & ZHANG Yongjuan

ABSTRACT

Like natural sciences and social sciences research, the humanities research is undergoing a data-driven paradigm transition towards the formation of so-called “Digital Humanities”. With the new paradigms, research materials, methods and approaches, procedures, forms of collaboration and communication among scholars are experiencing dramatic changes. “Data science” provides a methodological basis for these changes. The infrastructure to support research activities also need to be transformed or re-established. The new humanities research platform is based on data and method-oriented, providing an integrated data resource management, big data analysis, visualization and smart services. These changes will promote data applications and collaborative research which enables researchers to raise questions in unimaginable scale and dimensions.

At present, there are three models for research libraries to launch digital humanities projects, namely Resources-Based Model, DH Lab Model and Scholar Publisher Model. Libraries, museums, archives and other cultural heritage institutions should upgrade their large-scale digital collections built in two decades (usually called digital library system) to become a digital humanities research platform. It is an important

* 本文系国家社会科学基金重大项目“面向大数据的数字图书馆移动视觉搜索机制及应用研究”(编号：15ZDB126)的研究成果之一。(This article is an outcome of the key project of “Exploring the Mobile Visual Search of Digital Library in Big Data Era: Mechanism and Application” (No. 15ZDB126) supported by National Social Science Foundation of China.)

通信作者：刘炜，Email：wliu@libnet.sh.cn，ORCID：0000-0003-2663-7539(Correspondence should be addressed to LIU Wei, Email: wliu@libnet.sh.cn, ORCID: 0000-0003-2663-7539)

part of the national data infrastructure for humanities research, which makes collections continue to play a great role in digital age.

Digital Humanities infrastructure can beat the national level, or at the regional/institutional level. At the national level the following are recommended: 1) Develop the DH research plan, and establish DH research fund; 2) Publish the humanities data construction directory and the project development guideline, and gradually foster a comprehensive or specific DH research center; 3) Establish registration systems for the DH big data platform; 4) Develop subject/domain basic data demonstration platform; 5) Develop instrumental datasets, and provide basic data services on request to the public for free; 6) Enacting open data sharing policies to encourage crowdsourcing and open access; 7) Educate data librarian and train data personnel to curate and analyze data; 8) Strengthen international exchanges and integrate into a comprehensive national and global data infrastructure. At the institutional level, all types of libraries and cultural memory institutions are able to contribute to infrastructure building while carrying out their own research and developmental projects. It requires the industry to build a certain coordination mechanisms and follow certain technical standards and protocols. 1 fig. 3 tabs. 13 refs.

KEY WORDS

Digital humanities. Infrastructure. Digital library.

0 引言:人文研究的范式转型

随着大数据时代的到来,数据正在取代文献成为所有科学探索的素材和媒介,这种“数据驱动型研究”被美国计算机科学家、图灵奖获得者蒂姆·格瑞总结为“科学研究的第四范式”[1]。

对于人文学科而言,以往基于对个别新颖材料的发现和个人灵感与洞察力的、个人书斋小作坊式的研究,逐渐被基于对某一主题领域资料的全面占有和各种数理统计,及数据挖掘、跨学科团队协作式的研究所取代,这种转变得益于数字技术在人文学科研究方面的全面应用,得益于数字化网络化带来的研究素材的大数据化、研究工具的软件化、研究方法的计算化、研究过程的众包化、研究结果的可视化、研究平台的协作化以及学术交流方式的社会化。这种以数字技术支撑的人文科学研究,就是刚刚兴起的数字人文研究。

每种创新在刚刚产生的时候总会同时带来两类人群:热情似火的拥趸者和嗤之以鼻的反对者。数字人文的目标是将现代信息技术融入传统的人文研究与教学过程中,从而改变人文研究成果的获取、标注、比较、取样、阐释乃至表现方式,实现人文研究范式的全面变革,其本质上属于一种方法论和研究范式上的创新。数字人文从内容到形式、从过程到方法都与过去如此不同,以至于传统人文领域的权威们纷纷惊呼:这哪里还是人文[2]!? 他们认为,人文研究的核心从来都是灵感和洞察,而不可能是技术与方法。但数字技术的确有这样的颠覆性,数字人文迄今为止的实践已经证明其无可置疑地站稳了脚跟。新问题、新角度、新方法正在使传统的人文学科焕发出青春,人文学科从此进入了一个崭新的时代。

数字人文实践先驱——意大利著名人文学者罗伯特·布撒神父(Roberto Busa)[3]认为数字人文(当时称为人文计算)能够实现:①将学者从繁杂琐碎的资料收集整理工作中解脱出来;②专注于提出问题和学术发现;③极大地提高研究效率,促进学科发展。布撒神父毕其一生,为编制数千万字的托马斯·阿奎那文集索引殚精竭虑。整个过程伴随了信息技术从大型

机到互联网的发展，布撒神父也被人尊为文本分析和数字人文的创始人之一。他的实践证明，数字人文研究不仅能给传统人文研究提供全新的研究方法、工具和平台，而且能够改变人们思考问题的方法及看待问题的角度和维度，从而实现人文研究范式的变革。

与传统人文研究相比（见表1），数字人文研究对基础设施的需求变化最明显的特征是基于数据的细粒度管理和基于众包的内容生产、管理与协作模式。众包模式以前是不存在的，而现在的基础设施不仅要支持众包产生内容，而且需要在平台层支持全过程众包，包括内容的生产、组织、加工，服务提供乃至社交协作，等等。

表1 人文研究的两种范式比较

项目	传统人文研究	数字人文研究
内容特征	基于文献，载体管理，粗粒度	基于数据，内容/语义管理，细粒度
方法特征	定性、逻辑、思辨、比较、思想实验等	定量、计算、模型；内容分析、时空分析、社会关系分析等
协作特征	讨论、分工、小规模团队	众包、大规模协同、社会交互
辅助设施	基金会、图书馆、研究机构	数字图书馆、云平台、数据分析和计算设施
成果形式	论文、图书出版	论文、博客、数据、知识地图机构库、知识库、工具软件
采用方法	田野调查、问卷、访谈等	本体化、语义标注、语义检索、文本分析、可视化

科学研究的范式转型在其提出者库恩看来是一种带有颠覆性的科学革命[4]，它不可能是一蹴而就的，而是经历一系列过程之后的总结。这种革命的最终实现通常取决于一种新的基础设施的形成，正是这种基础设施固化了新的研究范式，并为新型的研究提供强有力的支撑。回顾历史我们可以发现，现代科研体系的形成也是遵循了这个规律，自印刷革命促成教育和知识的普及之后，伴随工业革命经历了三百多年的演变，逐渐成熟和固化。这个体系主要是由政府、科研机构、大学、基金会、企业、社会组织等各类机构按照一定的运行规则构成，其中围绕知识的生产、交流、出版、组织、保存、利用的生态链最为重要。传统意义上，这一生态链是围绕以纸张为主要知识载体的形态来运行的，进入到数字时代，这一基础正遭遇颠覆，而新的基础设施尚在孕育之中，可以说我们正幸运地（也可以说不幸地）处于两个时代的中间过渡期。本文正是对这个形成过程中的基础设施进行初步探讨。

1 什么是数字人文基础设施

基础设施（Infrastructure）原本是指人们开展正常的社会经济活动所赖以维持的物质条件和必需的公共服务，例如水、电、煤、交通、医疗、教育，等等。由此概念引申出的信息基础设施（Information Infrastructure）和网络基础设施（Cyberinfrastructure）是指面向网络服务通用的基础支撑环境，包括计算资源、数据资源和服务资源，例如网络接入、带宽、存储、数据、设备设施等。而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（ResearchInfrastruture①），是

① 欧共体对RI的定义为：与社会、人文、自然科学等科学研究相关的各类设施、资源与服务，包括能有助于在各自领域取得顶级研究成果的各类条件。参见 http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=what.

指在数字环境下为开展人文研究而必须具备的基本条件，包括全球范围内与研究主题相关的所有文献、数据、相关软件工具、学术交流和出版的公用设施及相关服务等。

具体来看，数字人文涉及所有通常被人们称为人文学科的领域，如语言学、文学、历史、哲学、宗教、法律、艺术，以及一些边界模糊的社会科学，如人类学、地区研究、传播学、文化研究等，这些学科的研究素材十分相似，大都是文献形式，但在研究方法和过程上可能会有不同，这也会反映在数字化之后的软件工具和平台上。数字人文的基础设施可以分为三个层次，其核心是由文献资源及其服务机构组成，提供了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，数据科学提供了它们共同的方法论基础；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（详见图1）。这些因素相互作用，形成一个可自我运行并持续发展的有机整体。它的主要功能就是为人文研究提供全面深入的资源和完善的存取手段，以支持现代研究过程（主要是基于数据的研究）得以顺利进行。

图1　数字人文基础设施的主要构成

在基础设施提供的众多职能中，知识的长期保存是其中非常重要的基础功能之一。伴随人类文明始终的图书馆、档案馆等人类记忆机构，是人类社会不断进步和可持续发展的重要保障。纸张这种载体经历数百年依然能够为人们方便地提供关于过去的记录，迄今仍旧是现代人文研究的主要依据。由于数字信息具有依附性和易逝性等特性，其永久保存问题显得非常突出。人们清晰地记得，CD-ROM 光盘这种介质曾经号称有 50—100 年的寿命，但这项技术还没有存在 50 年，人们已经不再依赖它了，它的实际寿命比缩微介质还要短很多。介质的特性决定着数据的持久性，但也只是数据永久保存的众多因素之一，其他因素还包括数据格式、组

织方式、对软件和系统的依赖性，等等。许多20年前的数字图书馆项目，即使已经完成，但没有多少增量内容和维护工作，流传至今依旧能访问的也已是凤毛麟角。这是一个非常严峻的问题，也是数字人文基础设施建设首先要解决的问题，但至今还未引起人们的足够重视。

数字人文研究并非只需要“老”的素材就够了，大量鲜活的“正在发生的历史”越来越成为主流，这也为基础设施建设提出许多现实的挑战：如何在传统出版行业走向没落、新的数字出版生态尚未形成的困难时期，利用社会性网络和开放存取的信息作为数字人文的信息来源，纳入到数字人文平台中去，也是当下数字人文基础设施迫切需要解决的问题。

2 数字人文基础设施的建设现状

数字人文的前身——人文计算可以追溯到二十世纪四十年代罗伯特·布撒神父的工作，其发展几乎贯穿了整个计算机发展史，但第一部真正以数字人文作标题的研究成果 *A Companion to Digital Humanities* 出版于2004年11月[5]。2006年，美国国家人文研究基金（NEH）启动了首个以“数字人文”冠名的项目——数字人文先导计划（Digital Humanity Initiative），如同数字图书馆先导计划一样，成为数字人文兴起的一个标志性事件。紧接着数字人文研究的基础设施建设被提上了议事日程，欧美均于2006年发布了报告，分别是“美国学习型社会委员会”（the American Council of Learned Societies）的《我们的文化共同体》（*Our Cultural Commonwealth*）[6]和“欧洲研究基础设施战略论坛”（the European Strategy Forum on Research Infrastructures）的《欧洲研究基础设施建设路线图》（*the European Roadmap for Research Infrastructures*）[7]，正式宣告启动数字时代科学研究基础设施建设的发展蓝图。

从这两个报告可以看出美国和欧洲在推进科研事业方面完全不同的做法。美国的报告只是一份数字人文基础设施建设的宣言，表明态度，宣示这项事业的重要意义，但具体如何做，还是市场经济的做法，交给各基金会和有关机构自行制定方案；而欧洲则是将数字人文基础设施作为整个欧盟研究设施和e-Infrastructure及e-Science的一部分，成立了各类专门的机构，设计了庞大的运行体系并落实了巨额预算（每年至少3亿欧元），进行统一的规划、管理和推进，该报告每两年更新一版。

从此数字人文基础设施建设进入了一个加速发展的时期，大型建设项目持续推进。如2008年启动、2014年完成的Bamboo项目，紧接着的“数字研究工具指南”（DiRT）项目、十九世纪学术在线（NINES）项目、联结十八世纪（Eighteenth Connect）项目、中世纪学术联盟电子档案（MESA）项目等。欧洲这类比较有名的基础设施建设项目有：欧洲研究基础设施联盟（ERIC）所属的艺术人文数字研究基础设施（DARIAH），该项目从2014年开始正式进入实际运营阶段；还有专门领域的欧洲大屠杀研究（EHRI）、欧洲数字档案协作研究（CENDARI）、考古数据库（ARIADNE）等。这些基础设施建设与早期的数字图书馆项目，如美国的Hathitrust、DPLA和欧洲的Europeana研究平台也开始进行整合[8]。

各类基金会成为数字人文基础设施建设的重要推动力量，同时它们自身也是基础设施的一部分。过去资助信息高速公路建设和数字图书馆的官方、非官方组织都成了数字人文研究的金主，除了前述的美国国家人文基金会（HEH）和欧盟研究基础设施联盟（ERIC）之外，诸如日本科学技术振兴机构（JST）、德国研究基金会（DFG）、英国信息系统联合委员会（JISC）、澳大利亚联邦政府创新、产业、科学与研究部（DIISR）、梅隆基金会（Mellon Foundation）、麦克阿瑟基金会（MacArthur Foundation）、图书馆与信息资源委员会（CLIR）、美国博物馆和图书馆服务协会（ARL）等都纷纷加入进来，资助不同主题领域不同方向的大量研究项目。其中既有新型的研究方法和成果示范，也有大

量的数据加工、资源组织、平台开发、开放存取和学术交流等基础性建设，正在形成体系完整、面向未来的数据驱动型人文学科研究的基础设施。

当前许多优秀的数字人文研究成果案例，已开始受益于这些基础设施。谷歌图书、万维网存档等项目扫描的大量图书，目前已成为 Huthitrust 和 DPLA 的重要馆藏，其中有大约 800 万种均经过全文识别，有 500 万种有较为完整规范的元数据信息，可用于文本分析和挖掘，这些已经在语言学、历史文化等学科领域发挥巨大作用。Erez Aiden 利用这个系统开发了一个词频分析工具[9]——N 元词频词组查看器，且已经在历史文化、人类语言、社会名望和群体记忆等方面取得了一批重要发现，成为大数据人文研究的典范，他本人还因此获得了美国优秀青年总统奖。国内也有这方面的例子，如金观涛、刘青峰夫妇的研究课题“观念史研究：中国现代重要政治术语的形成”，以十年之功，建立起一个庞大的“中国近现代思想史专业数据库”（1830—1930），汇集了大约 1.2 亿字的文本，成为他们开展研究、著书立说的利器。

除了文本分析之外，时空分析、社会关系分析也是数字人文研究通常采用的方法，成为许多数字人文平台提供的基本功能。斯坦福大学开发的启蒙运动知识界通信地图（Mapping the Republic of Letters）系统，是描述文艺复兴时期知识分子之间相互通信情况的一个数据库，可以提供通常的文本和字段查询，而且将往来通信的时间和点的信息在地图时空坐标上形象地展示出来，通过数万封信的流向，结合内容，可以了解这些启蒙思想家相互影响和思想演变的情况。斯坦福大学开发的一些数据处理工具（如 Palladio）和中间成果也能为多个类似的项目提供服务。以往的研究往往只能分析少数几个地区和学者，而有了基础设施平台，则可以从整个宏观的时空范围和群体角度进行提问和观察。这也是建设基础设施平台的意义所在。

3 作为数字人文基础设施的图书馆

图书馆作为人类文字记忆的保留地，从古希腊晚期集图书馆、大学和研究机构三位于一体的“雅典学园”伊始，就一直是从事知识与理性探索的天堂。印刷革命之后进入现代社会，图书馆之于传统人文研究，就像实验室对于自然科学、田野调查对于社会科学一样，仍然是举足轻重的。目前正全面进入数字革命时期，图书馆在其承担的四大职能中，除了文化保存，其他三大职能都已不具有主导作用。但欧美国家在人文研究基础设施中仍然无一例外地将图书馆、博物馆、档案馆作为重要组成部分，甚至出现一个新词“遗产科学（Heritage Science）①”，用来描述对于文化遗产进行研究的学科领域。图书馆如何支撑数字人文研究，不仅关系到其收藏的海量内容能否为基于数据的人文研究提供不竭的素材和丰富的工具，而且关系到其自身在数字时代的生存和发展。为数字人文研究提供全方位的支持，是许多研究型图书馆未来最重要的生存方式之一。

图书馆行业从二十世纪九十年代中期开始建设数字图书馆，不仅积累了大量的数字化资源，而且一直在探索作为互联网基础设施的一部分，即如何从资源描述、知识组织、移动应用、智慧服务等各方面，真正解决海量、多媒体、异构信息的互操作和开放存取问题，力求为读者提供体验良好、无所不在的知识服务。为此几乎所有新产生的信息技术，例如网格化、云计算、大数据、语义网等，都会第一时间在数字图书馆中找到应用。数字人文的兴起让数字图书馆的努力没有白费，找到了新的发展方向：从文

① 参见 https://en.wikipedia.org/wiki/Heritage_science，始见于 2006 年，指文化遗产类机构的数字人文研究学科。斯普林格专门有《遗产科学》（*Heritage Science*）杂志。

献到知识，从数字化到数据化，从全文化到模型化，从文本检索到语义检索，从门户到平台，从搜索到发现，从统计分析到关系推理，从分面展示到大数据可视化，从专家编目到众包加工……，图书馆真正从保存知识载体的库房变成提供内容服务的知识中介。

表 2 数字图书馆与数字人文应用的功能比较

数字图书馆	数字人文
资源数字化	资源内容数据化、实体化、对象化
资源管理	数据看护
资源门户	服务平台
搜索发现	提供环境工具方法
资源整合	挖掘与可视化

一项对图书馆与数字人文关系的调查表明[10]，绝大多数受访馆员认为数字人文的素材和成果应当保存在图书馆，过半数馆员认为图书馆员可提供数字人文项目开展初期的咨询服务，41%的受访图书馆已经开始提供事实上的数字人文服务，21%的图书馆设置了相应的服务岗位（如数字人文馆员），甚至有 17%的图书馆已经建立了数字学术中心，提供包括数字人文在内的服务。图书馆在数字人文研究中的角色定位是：数字人文机构库（54%），协助制订保存计划（50%），强化元数据功能（46%），促进和支持机构间的协作（43%），在数字化初期提供咨询意见（37%），提供项目的可持续支持（33%），帮助申请项目经费（26%），促进机构间的共同投入（22%），建立数字人文中心（19%），打包现有服务成为虚拟的“数字人文中心”（15%）；相关的服务类型主要有：项目初期的咨询顾问（51%）、数字保存（47%）、项目管理（31%）、研究支持（24%）、宣传推广（24%）、高性能计算保障（8%），等等。

从目前国外研究型图书馆参与数字人文项目的方式来看，主要有三种模式。

（1）资源中心模式。图书馆继续承担资源保障者角色，提供数字学术研究（Digital Scholarship）支撑，只是此时资源的形式是海量数据。图书馆通过深度标引，采用知识本体、关联数据等方式组织并开放资源，提供平台化服务，并提供能够嵌入个人知识环境的工具，一些高校图书馆参与建设 MOOC 资源中心也可归入此类。这类图书馆的学科馆员也应具备基本的数据操控和分析能力。这可以看成是数字图书馆服务的自然延伸，许多正在从事数字人文服务的著名大学和研究型图书馆都开始朝这个方向发展。

（2）数字人文实验室（Lab）模式。图书馆转型为提供基本研究条件的数字人文研究中心，例如提供各类媒体资源的加工以及工具平台（Media Lab），提供编程人力、数据分析挖掘（如提供 SPSS 分析工具）或可视化的团队，依靠机构知识库、GIS 平台和基本的工具软件平台支撑项目研发，并不断积累，共享给更多的研究人员使用等。这是一种深度参与，图书馆员与具体的数字人文研究团队呈现一种“你中有我、我中有你”的状态，甚至可以起到培养人才和孵化成果的作用。斯坦福大学图书馆的数字人文中心就属于此类，也有一些大学图书馆采用与专业团队合作的方式朝这个方向发展。

（3）科研成果发布出版者模式。承担类似传统知识交流模式中出版社的角色。开放存取运动、科学 2.0、Citizen Science 等一系列新生事物正在带来与过去完全不同的社会知识交流模式，直接的变化是科研成果的出版由原来的向用户收费，转为作为国家科研经费投入的一部分，由新知识的生产者来承担，而使用者零成本，这样更有利于成果的传播和利用。传统上图书馆作为机构库的建设和维护者，很自然地向产业链上游扩展，数字人文研究成果直接提交给图书馆来运营和发布，这也是顺理成章的事情。此外，该模式的平台还可以支持语义出版、数据分析、API 输出、众包优化等多种崭新功能。

人文资料的整理和保存是图书馆与生俱来的工作，数字人文“开放、合作、连接、多元、实践”的核心价值[11]与图书馆的职业理念是完全一致的。尽管目前图书馆在数字人文研究中的贡献度还很低，但目前几乎所有的数字人文项目中都有图书馆员的身影，只是图书馆员做的工作还比较初级，例如开展数据管理、资源监护、数字保存、发现与推介等服务，还没有体现更强的学术性、专业性和技术性。

此外，图书馆有一个很重要，但往往又被人忽视的特点，就是它的中立性。图书馆的中立性可以使它参与纠正或避免“数字学术”可能带来的以下四个方面弊端。①数字鸿沟：占有和获取信息能力的不均衡将在很大程度上影响研究能力；不同年龄的研究人员在研究方法使用上也会存在代沟，这也会影响到成果的价值。②技术伪装：有意识地通过技术选择性应用，从而倾向于达成研究结果，或误导结果。③知识遮蔽：由于数字化、可获得性、描述错误等造成一部分知识永远得不到利用而“失踪”，这很可能会造成结论的片面性。④认知异化：长期使用某些“数据库”或分析工具带来认知上的偏见，或内容与工具的本末倒置。

去除上述四方面的弊端是图书馆行业自存在起就树立的职业理念：消除信息鸿沟、信息无障碍流动、知识平等、客观中立。这些理念使图书馆的服务更为客观和可信，这在网络时代尤其难得，而这应该是图书馆作为一种制度设计最能体现其价值的地方。

4 如何建设数字人文基础设施

据 CenterNet 网站统计，目前加入“国际数字人文研究中心网络”的成员有 190 家之多①，系统、平台、软件工具、语料库等渐成规模，数字技术已基本成为人文研究的基础手段。近 20 年来，保存人类历史记忆的图书馆、博物馆、档案馆等机构已经将大量的馆藏资源进行了数字化。国外目前已进入大规模数字人文平台的建设阶段，大都是通过“文本化”和“数据化”，对其内容进行描述、组织和揭示，结合社会化众包方式，建立更多的语义关联，并提供各类分析处理和可视化工具。这就是数字人文平台建设的基本要求，也是数字人文基础设施建设的主要内容。

我国数字人文研究大致起步于 2011 年，武汉大学成立国内第一个数字人文研究中心是标志性事件[12]。2012 年以来陆续有一些数字人文研究的文章和论文见诸于报刊，有许多研究机构虽然没有以“数字人文”冠名，但做的是同样的事情，如古籍全文本数据库、历史地图 GIS 系统等。2014 年，上海图书馆学会召开了“数字人文与语义技术”研讨会，2016 年 5 月，北京大学图书馆召开了“首届数字人文论坛”，这些让人明显感受到国内数字人文领域的研究在持续升温。

与国外轰轰烈烈的现状相比，我们起步晚了近十年，目前仍然处于单兵作战、“小米加步枪”的阶段。虽然国内的数字图书馆建设起步不晚，至今也取得了丰硕成果，然而决策层还没有充分意识到数字人文的发展趋势，还不知道应该采取怎样的对策，差距有进一步拉大的趋势，甚至还赶不上我国的台湾地区。一个最明显的例子就是，在近年大规模的古籍整理和“善本再造”项目中，基本没有考虑借鉴数字人文的思路和方法来建设，没有充分利用当前信息技术的巨大优势，还是以印刷出版为成果指向，缺乏“互联网思维”，无法真正使古代文献为普通研究人员甚至大众所用。

传统的中文文献进行文本化的过程本来就比西文数据更为复杂和漫长，这导致我们基础语料库存在规模不大、质量不高、描述信息不够、循证信息缺乏以及重复建设等问题。另外，进行时空分析所需的历史地图数字平台和进行

① 参见 http://dhcenternet.org/centers.

社会关系分析所需的基础数据库都还没有建立起来。究其原因，除了僵化迟钝的社科人文科研体制是最大原因之外，我们在数字人文基础设施上的落后，无法给相应的研究提供必要的支持，也是一个重要原因。

基础设施可以是国家层面的，也可以是地区行业或组织机构层面的。为尽快建成这样的基础设施，在国家宏观层面建议开展以下工作：①制订数字人文发展规划，设立数字人文发展基金；②发布基础人文数据建设目录和项目开发指南，逐步培育综合性和专门领域的数字人文研究中心；③建立注册登记系统，为建立数字人文大数据平台做好准备；④按学科或领域开展基础数据平台的研发示范；⑤研发工具型的数据集，如历史、地理、人物、机构、事件、名称概念词表等语料库，按需求分阶段提供公益型基本数据的网络服务；⑥制订相应的数据开放共享政策，鼓励众包建设，开放共享；⑦开展数据科学专业教育和人员培训，培养数据管理与分析方面的紧缺人才；⑧加强国际交流，并融入国家综合性数据基础设施和全球数据基础设施中。

在机构层面，数字人文应用系统的开发可能因选题和规模的不同有很大差异，但总体上大都遵循表 3 所列出的步骤。因为现在还没有数字人文大数据仓储（谷歌图书文本除外），数据的完整性、可获得性、开放程度以及技术条件等常常得不到满足，所以大多数字人文项目都不得不从数据收集、清洗、加工、组织和整理开始做起，并且缺乏基本的数据整合、分析与可视化工具。正因为如此，各类图书馆和文化记忆机构都可以在进行自身项目研发时，同时为基础设施建设添砖加瓦。这需要在业界形成一定的协调共建机制，并遵循一定的技术标准规范和协议。

表 3　数字人文应用开发的一般过程

步骤	工作内容	说　明
1	基于领域资源的书目控制	掌握某一领域或主题所有已发表的文献和成果情况，有条件拥有该主题领域的主要内容。通常可通过建立联合目录数据库等方式进行书目控制。
2	领域资源的对象化、数据化	根据需求设计，将内容中所包含的各类实体通过单独标引或描述的方式提取出来，进行结构化处理，并赋予规范控制。
3	基础数据（人地时事）的模型化、本体化	建立领域应用模型（通常以资源模型为基础）和领域本体，设计应用场景，将通用型、有共享需求的数据开发成单独的服务，提供数据开放接口，允许外部整合调用。
4	工具开发（处理、分析挖掘、可视化）	将必要的数据处理、整合、分析、挖掘和可视化功能开发成工具包。
5	建立服务平台（社会化）	提供支持各类设备（包括移动设备）的自适应平台化服务，支持机器调用，支持用户贡献内容（UGC），以及系统数据和功能的众包优化。
6	开发增值服务（VR、AR、MR、语义出版等）	提供面向企业二次开发的增值服务。包括各类个性化的可视化呈现、模板报表输出、移动视觉搜索、虚拟现实、增强现实、混合现实和语义出版等。

上海图书馆在深化数字图书馆资源库建设、使之能直接支撑数字人文研究的过程中，采取了一些行之有效的办法。首先在满足每个特色资源库建设目标的同时，将一些可以共享的资源单独设计成独立的数据服务。例如在开发盛宣怀档案库时，同时设计了名人规范档服务系统，将来可以开放为人名规范档来使用；其次非常注重采用已有的标准规范，例如家谱知识

库系统借鉴了美国国会图书馆正在研制的 BIBFRAME 数据模型，该模型也是对图书馆界 RDA 模型的一种应用，在格式中也大量复用了业界成熟本体中的许多元素；第三是有意识地将一些能够提供同行借鉴的做法设计成规范的标准，例如规范数据的发布命名规则，以及家谱的谱系图模式（Schema）等。这样，不仅每个项目的数据能够共享出来，而且数据格式和规范也能供人借鉴，将来一些通用的工具软件和平台也能开放出来。上海图书馆通过开展一次开放数据应用竞赛证实了这个想法的可操作性，在一个多月的时间里，几十支参赛队伍就利用家谱知识库中的关联数据 API 和 SPARQL 接口，开发出了非常具有创意的原型系统和移动应用[13]。如果从事数字人文开发和服务的机构都这样做，一个共建共享的基础设施就有望更快建成。

上海图书馆根据自身馆藏情况，计划以“上海年华”和“特色资源”为主题开发两类面向不同用户对象的数字人文平台。“上海年华”以晚清民国期间（从 1843 年上海开埠前后至 1949 年新中国成立）的期刊、报纸、手稿、尺牍、照片、名人档案、笔记、地图、地方文献等资料为主，按照其中的“人地时事”重构目前基于文献的知识系统，通过文献中丰富的细节多维度地呈现上海的文化定型和城市演变。“特色资源”则希望通过对古籍、家谱、碑帖等特藏，在目前“数字图书馆”系统的基础上，广泛采用关联数据技术所赋予的规范控制功能，重建基于内容的揭示系统，推出一批数字人文应用原型。

近十余年数字人文领域的进展为人文科学研究带来了范式变革，也为数字图书馆的未来发展指明了一条可行的道路。然而目前业界尚未普遍认识到这一点，有能力进行尝试的机构更是少之又少，我们可能正在错过一个绝好的发展机遇。因此希望当前的一些成功实践能够起到一定的示范作用，吸引更多的图书馆加入到数字人文基础设施建设的队伍中来。唯有如此，图书馆才能在数字时代继续发挥人类知识的基础架构的作用。

参考文献

[1] Hey H，Tansley S，Tolle K. 科学研究的第四范式：数据密集型科学发现 [M]. 潘教峰，张晓林，译. 北京：科学出版社，2012.（Hey H，Tansley S，Tolle K. The fourth paradigm：data-intensive scientific discovery [M]. Pan Jiaofeng，Zhang Xiaolin，trans. Beijing：Science Press.）

[2] Hall G. There are no digital humanities [EB/OL]. [2016-07-25]. http://dhdebates.gc.cuny.edu/debates/text/21.

[3] RobertoBusa S J，The Invention of the Machine-Generated Concordance. Faculty publications，classics and religious Studies Department [EB/OL]. [2016-07-25]. http://digitalcommons.unl.edu/classicsfacpub/70.

[4] 托马斯·库恩. 科学革命的结构 [M]. 金吾伦，胡新和，译. 北京：北京大学出版社，2003.（Kuhn T S. The structure of scientific revolutions [M]. Jin Wulun，Hu Xinhe，trans. Beijing：Peking University Press，2003.）

[5] Schreibran S，Siemens R，Unsworth J. A companion to digital humanities [EB/OL]. [2016-07-25]. http://www.digitalhumanities.org/companion.

[6] Welshons M. Our cultural commonwealth：the report of the American Council of Learned Societies Commission on cyberinfrastructure for the Humanities and Social Sciences [EB/OL]. [2016-07-25]. http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf.

[7] ESFRI. The European roadmap for research infrastructures [EB/OL]. [2016-07-25].https://ec.europa.eu/research/infrastructures/pdf/esfri/esfri_roadmap/roadmap_2006/esfri_roadmap_2006_en.pdf.

[8] Dunning A. Drafting priorities for Europeana research [EB/OL]. [2016-07-25].http://pro.europeana.eu/blog-posts/drafting-priorities-for-europeana-research.

[9] 埃雷兹·艾登，让-巴蒂斯特·米歇尔.可视化未来数据透视下的人文大趋势[M].王彤彤，沈华伟，程学旗，译.杭州：浙江人民出版社，2015.（Aiden E，Michel J B. Uncharted：big data as a lens on human culture [M]. Wang Tongtong，Shen Huawei，Cheng Xueqi，trans. Hangzhou：Zhejiang People's Publishing House，2015.）

[10] Stewart V，Patricia H. Special report：digital humanities in libraries [EB/OL]. [2016-07-25].http://americanlibrariesmagazine.org/2016/01/04/special-report-digital-humanities-libraries.

[11] Spiro L. "This is why we fight"：defining the values of the digital humanities [EB/OL]. [2016-07-25].http://kompetenzzentrum.uni-trier.de/files/9113/9696/2499/Values_in_the_DH_Lisa_Spiro.pdf.

[12] 张肖雯，刘金波. 首个数字人文研究中心落户武汉[N].中国社会科学报，2011-05-05(002).（Zhang Xiaowen，Liu Jinbo. The first study center of digital humanities established in Wuhan [N]. Chinese Social Sciences Today，2011-05-05(002).）

[13] 上海图书馆.上海图书馆举办2016开放数据应用开发竞赛展及分享会[EB/OL]. [2016-07-25].http://www.nlc.gov.cn/newtsgj/yjdt/2016n/8y_11626/201608/t20160819_128768.htm.（Shanghai Library. Shanghai Library held exhibition on open data application development contest 2016 and sharing meeting [EB/OL]. [2016-07-25].http://www.nlc.gov.cn/newtsgj/yjdt/2016n/8y_11626/201608/t20160819_128768.htm.）

刘 炜 上海图书馆副馆长，研究员。上海 200031。
谢 蓉 上海对外经贸大学图书馆信息咨询部主任，副研究馆员。上海 201620。
张 磊 上海图书馆系统网络中心开发部主任。上海 200031。
张永娟 中国科学院上海生命科学信息中心。上海 200031。

（收稿日期：2016-08-31）