面向农史领域的数字人文研究基础设施建设研究
——以方志物产知识库构建为引

徐晨飞[1,2] 包 平[1]

(1.南京农业大学 中华农业文明研究院,江苏 南京 210095;2.南通大学 经济与管理学院,江苏 南通 226019)

【摘 要】大数据时代的到来,为传统人文学科研究者带来了新的挑战和机遇,计算机技术与数据科学的介入为人文学科带来了"数字人文"研究的新理念与新范式,作为支撑数字人文研究活动的基础设施也正在全球范围开始建立与使用。农史研究具有较明显的跨学科特征,通过文献调研与分析,提出应建设面向农史领域的数字人文研究基础设施。以中华农业文明研究院特藏文献资源《方志物产》为例,从数字化、数据化、知识化、平台化这四个阶段阐述方志物产知识库构建的思路以及深度利用的场景,以期以此为标志开启面向农史领域的数字人文研究基础设施建设的新篇章。

【关键词】农业史;数字人文;方志物产;知识库

【中图分类号】S-09;K207 【文献标识码】A 【文章编号】1000—4459(2019)06—0040—12

Research on Infrastructure Construction of Digital Humanities Research for Agricultural History: Start with Construction of Knowledge Base of Local Chronicle: Produce

XU Chen-fei[1,2] BAO Ping[1]

(*1.Institution of Chinese Agricultural Civilization, Nanjing Agricultural University, Nanjing 210095;*
2. Economics and Management School of Nantong University, Nantong 226019)

Abstract: The arrival of the era of big data has brought new challenges and opportunities for researchers in the traditional humanities. The involvement of computer technology and data science has brought a new concept and new paradigm for the study of the "digital humanities" for the humanities. As a support for digital humanities research activities, the infrastructure is also beginning to be established and used globally. Agricultural history research has obvious interdisciplinary characteristics. Through literature research and analysis, this paper proposes that digital humanistic research infrastructure should be built for the field of agricultural history. Taking Local Chronicle: Produce, a special collection of literature resources of the Chinese Research Institute of Agricultural Civilization, as an example, this paper expounds the construction ideas and scenes of deeply utilization of the knowledge base from the four stages of digitization, datalization, knowledgeablization, and platformization, with that as a sign, in order to open a new chapter on digital humanities research infrastructure construction for the field of agricultural history.

Key words: agricultural history; digital humanities; Local Chronicle: Produce; knowledge base

[收稿日期] 2019-08-08

[基金项目] 国家社会科学基金重大项目"方志物产知识库构建及深度利用研究"(18ZDA327);教育部人文社会科学研究青年基金项目"基于语义的方志物产资料知识组织与知识聚合实证研究"(19YJC870027)

[作者简介] 徐晨飞(1981-),男,南京农业大学中华农业文明研究院博士生,讲师,研究方向为知识组织、数字人文;包平(1964-),男,南京农业大学中华农业文明研究院教授、博士生导师,研究方向为科技史信息组织。

自从2007年图灵奖获得者吉姆·格雷(James Gray)提出基于数据密集型计算的“科学研究的第四范式”[①]以来,相关领域特别是人文学科的研究进展令人兴奋。托马斯·库恩的“范式转换”理论认为,新范式的建立伴随着科学革命的产生,革命的结果是拥有新范式的新的科学共同体取代拥有旧范式的旧的科学共同体[②]。这十年间,数字人文作为一门正在兴起的新学科,其演进过程也印证了库恩的“范式转换”理论,历史学家、地理学家、文学家等人文学科研究者与数据科学家、计算机科学家、信息资源管理专家等正携手成为新的科学共同体,将大数据化的研究素材、软件化的辅助研究工具、计算化的研究方法、可视化的研究结果贯穿于人文科学研究过程,取得了诸多令人瞩目且颠覆传统研究思维的成果,也使得人文学科重新焕发了新的生命力。而伴随着数字人文研究的不断开展与深入,学界对研究基础设施的需求也迫在眉睫,目前全球在相关政府、机构的支持与规划下,各类数字人文研究基础设施建设正在如火如荼地进行中。习近平总书记在中共中央政治局就实施国家大数据战略进行第二次集体学习时就强调,要推动实施国家大数据战略,加快完善数字基础设施,推进数据资源整合和开放共享,保障数据安全,加快建设数字中国[③]。

在历史学界,乔·古尔迪(Jo Guldi)与大卫·阿米蒂奇(David Armitage)在《历史学宣言》中就呼吁当代历史学家要有宽阔的视野和考察大问题的雄心,而大数据时代的到来可在未来帮助史学家成为新时代的数据文献专家,并向公众交流其他学科的数据、方法和成果,并以自身的学术强项对此作出分析、比较和对照[④]。近半个世纪以来,数字人文研究的兴起给史学研究带来了前所未有的颠覆与争鸣,如量化史学研究方法之于经济史、教育史、宗教史等等[⑤]。农史研究作为一门相对独立的学科,只有百年的历史,其进展和积累对历史学、政治经济学及其他社会科学具有基础性科学的价值[⑥]。中华农业文明研究院作为国内农史研究重镇,是一个集科学研究、人才培养和信息收集与服务于一体的开放型学术机构,其特藏的《中国农业史资料》《方志综合》《方志物产》《方志分类资料》《二十五史水利史资料》《太湖地区农业史资料》《农业史简报数据》等资料被学界誉为“海内孤本”。本文基于《方志物产》资料的数字化、知识组织与知识挖掘等前期研究工作,提出方志物产知识库构建思路与方法,目的是将其建设成为面向农史领域的数字人文研究基础设施,以期在数字时代推进农史及其它专门史研究打开一个全新的篇章。

一、农史研究与数字人文研究基础设施概述

(一)农史研究概述

农史研究作为一种学科化的努力始于20世纪初期。在西方,美国、英国、德国、丹麦、荷兰、法国等均为农业史研究开展得较早的国家。“农业史”在西方分别有agrarianhistory,agriculturalhistory和rural - history等不同说法,笔者以这三个关键词为主题,在Web of Science TM核心合集中检索了农业史有关研究文献被Social Sciences Citation Index(SSCI)(1995–2019)收录的情况,共检索出论文4345篇。将检索结果导出由网络分析工具UCINET[⑦]生成高频关键词共现网络图谱,如图1所示。图中红色为高频关键

① Tony Hey, Stewart Tansley, Kristin Tolle, The Fourth Paradigm Data—Intensive Scientific Discovery: The Science Press, 2009, p.5.

② [美]托马斯·库恩:《科学革命的结构》(第4版),金吾伦、胡新和译,北京大学出版社,2012年,第5–6页。

③ 习近平:《实施国家大数据战略加快建设数字中国》,http://www.xinhuanet.com/politics/2017–12/09/c_1122084706.htm。

④ [美]乔·古尔迪、[英]大卫·阿米蒂奇:《历史学宣言》,孙岳译,上海人民出版社,2017年,第138–139页。

⑤ 陈志武:《量化历史研究的过去与未来》,《清史研究》2016年第4期。

⑥ 王思明:《农史研究:回顾与展望》,《中国农史》2002年第4期。

⑦ UCINET官网地址:https://sites.google.com/site/ucinetsoftware/home。

词，分别为“Rural、Hiv、China、Agriculture、Risk Factors、Migration、Gender、Development、Depression”等，以这些高频关键词为中心，将与之连接的关键词进行聚类，从文献计量角度可以分析出，近二十多年来农史研究涉及的方向较为多元，如社会学、医学、环境科学、人类学、历史学、农学、政治学、地理科学等等，这与学界对“农业史”的学科定位分歧较多有一定关系。有人认为是历史学派生而来；也有认为是农学的分支，研究的是历史时期的农业，即古农学；还有不少研究经济史和农业经济史的学者认为它是经济学和农业经济学的一个分支；还有人认为它是科学史分化而来。经过百年的发展，越来越多的人认识到农史是一门介于自然科学与社会科学之间的一门交叉学科。它运用自然科学与社会科学相互交叉、农业科学与历史学相互结合的方法，探讨农业产生和发展的动因、动力、影响及规律[①]。正是由于农史研究的跨学科属性，其研究方法也越来越多元化，还包括需面对“数据洪流”的问题，这都对当今的农史工作者提出了更高的要求，然而从另一个角度，也为农史研究带来新的契机与活力。

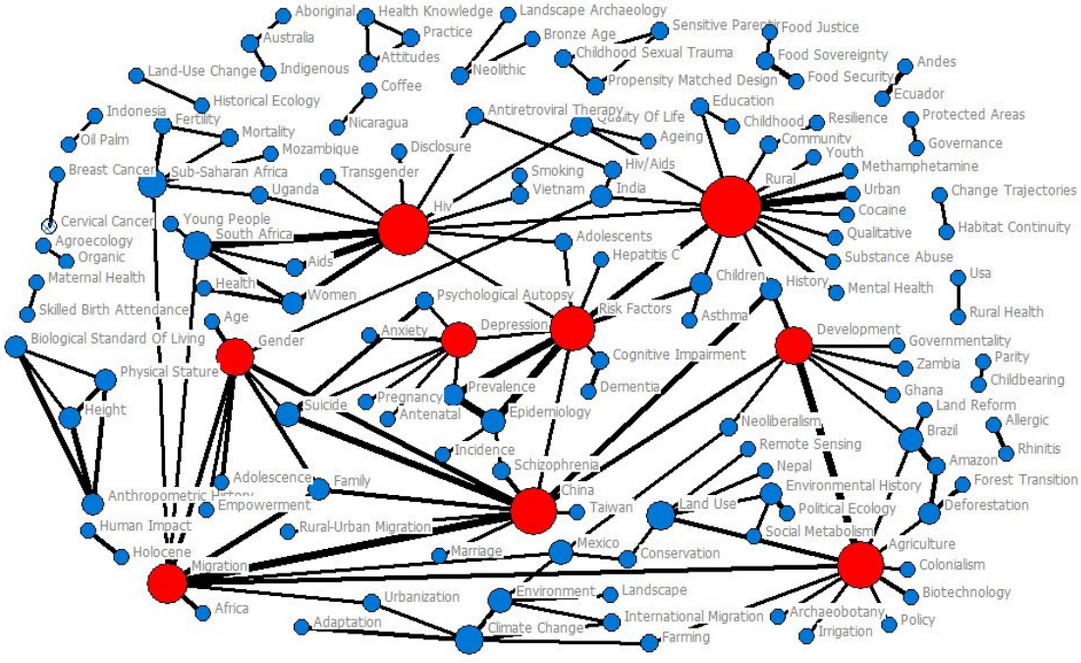

图1 SSCI(1995–2019)收录农业史相关研究文献高频关键词网络图谱

(二)数字人文研究基础设施

数字人文(Digital Humanities)可以追溯到1949年，当时基督教神父罗伯特·布萨(Roberto Busa)与IBM公司一起，将圣托马斯·阿奎那(St. Thomas Aquinas)的所有作品转为数字形式[②]。他本人被公认为数字人文实践先驱，他的创新性工作被认为是使用计算机进行量化分析的一种全新的人文研究方法，体现出的是方法论上的一种革新。之后，随着信息与网络技术的发展和数字化技术的广泛应用，使得数字人文的研究和应用领域更加宽泛。近十多年来，越来越多的专门“数字人文中心(系)”在北美、加拿大和欧洲成立，并开展了许多跨学科的数字人文项目，涉及历史、考古、艺术史、英语文学、建筑等多个学术领域。此外，在英美学术圈内也出现了专门的协会组织，这些协会将不同学校及地区的数字人文中心及学

① 王思明：《农史研究：回顾与展望》，《中国农史》2002年第4期。

② [美]克莉丝汀·L.伯格曼：《大数据、小数据、无数据——网络世界的数据学术》，孟小峰、张祎、赵尔平译，机械工业出版社，2017年，第130页。

者整合起来，形成了一个庞大的国际学术社区，定期组织会议及各种学术活动[①]。

基础设施（Infrastructure）原意是指为社会生产和居民生活提供公共服务的物质工程设施，用于保证国家或地区社会经济活动正常进行的公共服务系统，包括交通、邮电、供水供电、商业服务、医疗卫生、环境绿化、文化教育等等[②]。由此引申的概念有信息基础设施（Information Infrastructure）、网络基础设施（Cyberinfrastructure）、数字基础设施（Digital Infrastructure）、研究基础设施（Research Infrastructure）以及科研数据基础设施（Research DataInfrastructure）等。根据2003年美国国家科学基金会（National Science Foundation）的一篇报告，网络基础设施被计算机科学家Dan Atkins等人定义为支持大规模数字对象的存储、共享、分析的大型基础设施，并且他们认为“若基础设施的建设是为了工业经济，那网络基础设施建设则是为了知识经济”[③]。

相较于美国，欧洲在研究基础设施（RIs）建设方面走在了全球前列。研究基础设施被描述为“科研团体为开展研究以及创新培育而使用的工具、资源与服务集合”[④]，例如欧洲网格计算基础设施Europe - an Grid Infrastructure（https://www.egi.eu/）、学术交流基础设施OpenAIRE（https://www.openaire.eu/）、提供虚拟研究环境的数据基础设施D4Science（https://www.d4science.org/）等。GRDI2020（Global Research Data Infrastructures 2020）项目专家组将科研数据基础设施定义为一个以数字化科研数据为中心，包含服务与工具的管理型网络环境[⑤]。从概念的范围来看，“科研数据基础设施”属于“研究基础设施”，两者皆从属于“网络基础设施”（或“数字基础设施”），“数字人文研究基础设施”应从属于“研究基础设施”，即支持人文学者在数字环境下开展科研活动的必须具备的基础设施[⑥]，包括与主题相关的数字化文献资源、数据、软件工具、硬件（云存储），系统平台等对象，并支持人文科学研究数据分享与重用，促进科研成果在线出版、全球人文学科合作，加速科研创新的生态系统。

近些年，全球尤其是欧洲数字人文研究基础设施数量急剧增长，大多数研究基础设施都聚焦于人文学科的特定领域，比如面向考古学的ARIADNE（http://www.ariadne-infrastructure.eu/）、研究大屠杀历史的EHRI（https://www.ehri-project.eu/）、面向历史研究的Cendari（http://www.cendari.eu/）、面向语言学研究的CLARIN（https://www.clarin.eu/）、面向艺术与人文学科的DARIAH（https://www.dariah.eu/）以及面向文化遗产研究的IPERION（http://www.iperionch.eu/）等等。这些数字人文研究基础设施为相关学科领域学者提供了支撑跨学科研究的资源、工具、数据管理与检索的通用解决方案。

从目前全球数字人文研究的发展阶段来看，虽历经几十载，除欧盟成立了专门机构来落实数字人文研究基础设施以外，其他基于国家层面的广义数字人文研究基础设施还未能建成。一般是政府表明支持态度，由各类财团、基金会以及一些官方或非官方组织，在各自学科、领域进行相关主题的狭义数字人文研究基础设施建设。例如美国的数字人文研究基础设施建设思路就与欧洲截然相反，其建设并非

① 陈静：《历史与争论——英美“数字人文”发展综述》，《文化研究》2013年第4期。

② 刘炜、谢蓉、张磊：《面向人文研究的国家数据基础设施建设》，《中国图书馆报》2016年第5期。

③ D. Atkins, Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Founda - tionblue—ribbon advisory panel on cyberinfrastructure, 2003.

④ L. Candela, D. Castelli, P. Pagano, Virtual research environments: an overview and a research agenda: Data Science Journal, 2013, pp.75—81.

⑤ F. Karagiannis, D. Keramida, Y. Ioannidis, et al, Technological and Organisational Aspects of Global Research Data In - frastructures Towards Year 2020: Data Science Journal, 2013, pp.1-5.

⑥ Alessia Bardi, Luca Frosini, Building a Federation of Digital Humanities Infrastructures. https://ercim—news.ercim.eu/en111/special/building-a-federation-of-digital-humanities-infrastructures.

由政府的科技政策制定者与管理者来主导，而是由各学科领域的数字人文研究学者来推动[①]。这种自下而上的建设方式也催生出大量不同学科领域、不同专业方向的优秀成果，尽管这些成果目前可能还存在技术标准化、资源整合、版权等诸多问题。或许在未来，可期更多的组织机构携手联合，由国家层面制定并出台相关标准框架，真正形成体系完整、标准统一、数据共享、跨学科领域的综合数字人文研究基础设施。

（三）国内外农史领域数字人文研究基础设施建设现状

目前国内外与农史相关的数据基础设施建设还是以数字化资源存储项目居多，严格意义上来说，大多数还处于数字人文基础设施的初级阶段，相关平台还缺乏支持诸如文本挖掘、时空分析、社会网络分析等数字人文研究常用方法的工具与服务模块。美国农业部（USDA）下属的国家农业图书馆（National Agricultural Library，NAL）开发了多项农业史数字人文项目，如“Growing a Nation: The Story of American Agriculture”项目（https://growinganation.org/），是以剧本故事的形式来讲述美国的农业史，采用了农史编年体、视频、教师授课计划等多种多媒体形式来展现；“Homestead Act”项目（https://www.nal.usda.gov/homestead-act），是对林肯当年颁布与实施“宅地法”的相关历史文献资源进行了数字化；“Organic Roots Digital Collection”项目（https://naldc.nal.usda.gov/organic_roots/）收集了合成有机物被广泛应用之前的出版的农业历史期刊全文，主要是1942年之前的农业技术与有机农业信息。

美国康奈尔大学的“Core Historical Literature of Agriculture”项目（https://digital.library.cornell.edu/collections/chla）是一个收录了自19世纪早期至20世纪末出版的，涵盖农业经济学、农业工程学、动物科学、植物保护学、食品科学、人类营养学、农村社会学以及土壤学等专业领域的各类重要文献，数字化后支持全文检索；康奈尔大学还与美国农业部下属的美国农业统计局联合开发了“USDA Census of Agriculture Historical Archive”项目（http://agcensus.mannlib.cornell.edu/AgCensus/homepage.do），它对美国农场、牧场以及农民档案进行了详细的统计，档案资料涉及土地利用、土地所有权、经营者的情况、生产实践、收支情况等等方面。美国国会图书馆建设的“Historical Agricultural News”数字人文项目（http://agnews.net/）可对美国历史上的农业机构、农业技术以及生产实践活动等数字化报纸资源进行检索，这些历史农业数据还可以支撑诸如经济实践、移民活动、语言文字的应用、媒体的影响等其他领域的研究。

科罗拉多州立大学的“Colorado Agriculture and Rural Life”项目（https://lib2.colostate.edu/research/agbib/）对科罗拉多州历史上重要的农业与农村文献资料进行了整理，主要包括与农业历史相关的水资源、教育、矿产、旅游、娱乐产业等主题，文献类型主要有图书、期刊、学位论文、地图、图片、档案等。联合国粮农组织（FAO）根据各国农业科研和生产发展的需要，于1975年建立的题录型数据库ARGIS（http://agris.fao.org/agris-search/index.do），收录了FAO编辑出版的全部出版物和180多个参加国和地区提供的农业文献信息，特别是第三世界国家农业、林业及相关学科的应用研究方面的文献，1979年以后部分数据提供了文摘。

在亚洲，日本农林水产省建制的“Agriknowledge”知识库（https://agriknowledge.affrc.go.jp/）提供了大量日本农业科学与技术相关的信息资源，如论文、研究课题、研究成果、认定品种等，此外还提供明治时代至今百余年的农具检索，为其平台特色之一。

国内农史领域的数字人文基础设施建设项目目前还比较稀少，南京农业大学中华农业文明研究院相关学者之前在此领域做了一些基础性工作：例如在数据库建设方面，研制开发的中国农业遗产信息平台包含农史论文题录数据库、农业古籍目录数据库、中国农业遗产选集图文库、民国资料图文库、方志资料图文库、农业典籍善本图文库、农业古籍全文数据库及农史论文全文数据库等若干数据库，初步实现

① W. Kaltenbrunner, Digital Infrastructure for the Humanities in Europe and the US: Governing Scholarship through Coordinated Tool Development: Computer Supported Cooperative Work, 2017, Vol.26 No.3, pp.1-34.

了各类资源的数字化，在一定程度上促进了资源共享[①]；曹玲研究了古籍数字化整理方法与过程并列举了《齐民要术》知识库的构建实践[②]；王雅戈对民国时期农业文献数据库建设展开研究[③]。在文本挖掘与知识组织方面，黄建年研究并设计出农业古籍自动断句标点的算法，并实现了农业古籍断句标点的原型系统[④]；常娥对古籍自动编纂、自动校勘相关智能处理技术展开了研究[⑤]；何琳构建了古农书的本体，提高农史信息资源语义检索的效果[⑥]；唐惠燕利用GIS技术对1949–2011江苏水稻种植进行了时空变迁的实证研究[⑦]。

其实在历史学领域，国内外已经有诸多较为成功的数字人文基础设施项目，在此围绕中国历史研究举例一二。例如，哈佛大学费正清研究中心与北京大学中国古代史研究中心、台湾中研院史语所联合建设的“中国历代人物传记资料库（CBDB）”项目（https://projects.iq.harvard.edu/chinesecbdb），其负责人包弼德教授也多次在各种场合提出要建设服务于中国历史研究的网络基础设施，提出可通过API分享和文档分享来聚合网络上不同的数字资源，也可避免基础数据建设的重复劳动。台湾大学数位人文研究中心的“台湾历史数位图书馆（THDL）”（http://thdl.ntu.edu.tw/index.html）也是以提供数字人文研究基础设施为目的来建设的。在THDL中，不但有提供全文检索、元数据检索功能的全文数据库（淡新档案、明清台湾行政档案、古契书），还提供了可服务人文研究的各类软件工具集，如中西历日期对照查询、清代官职表、度量衡单位换算系统、THDL前后缀词分析工具等等，其设计理念已经超越了普通的数据库存储系统，而是可以帮助研究人员发现新问题的有效平台。

此外，还有上海交通大学研发的中国地方历史文献数据库（http://dfwx.datahistory.cn/pc/）、复旦大学的中国历史地理信息系统（CHGIS）（http://yugong.fudan.edu.cn/）、台湾中研院开发的中华文明时空基础架构（CCTS）（http://ccts.ascc.net/）等项目，均可作为相关主题研究的数字人文研究基础设施。从以上案例可窥探出，目前国内外数字人文项目大多数还是以服务特定领域与主题的人文研究为主。在欧洲以外的地区，国家层面主导的数字人文基础设施建设还存在诸多困难与问题，但是考虑到研究基础设施建设势在必行，因此以机构为主导的研究基础设施项目若在设计之初即着重思考数据的交互、资源的共享、工具的适用、用户的合作等标准化问题，就有可能在未来与国家级研究数据基础设施进行对接并成为其重要组成部分。

二、面向农史领域的数字人文研究基础设施建设

数字人文研究基础设施的建设应始终围绕人文学者的学术研究需求来展开，若要对人文学者的研究需求进行分析，则首先应深刻理解人文研究的活动过程，尤其是在e–Research[⑧]大时代背景下的虚拟研究环境（Virtual Research Environment，VREs）[⑨]之中的学术活动过程。美国数字人文研究学者John Unsworth早在2000年一次研讨会中就提出“学术基本体”（Scholarly Primitives，也有国内学者翻译为“学

① 曹玲、常娥、薛春香：《农史研究的新工具——中国农业遗产信息平台的设计与构建》，《中国农史》2006年第1期。

② 曹玲：《农业古籍数字化整理研究》，南京农业大学博士学位论文，2006年。

③ 王雅戈：《民国农业文献数字化整理及信息组织研究》，南京农业大学博士学位论文，2007年。

④ 黄建年：《农业古籍的计算机断句标点与分词标引研究》，南京农业大学博士学位论文，2009年。

⑤ 常娥：《古籍智能处理技术研究》，南京农业大学博士学位论文，2007年。

⑥ 何琳：《古农学本体的半自动构建及检索研究》，南京农业大学博士学位论文，2007年。

⑦ 唐惠燕：《基于GIS江苏种植结构演变研究(1949–2011)》，南京农业大学博士学位论文，2014年。

⑧ T. Anderson, H. Kanuka, E–research: Methods, strategies, and issues: Boston: Allyn and Bacon, 2003.

⑨ L. Candela, Virtual research environments: GRDI2020 Scientific Report, 2011.

术原语”[①])的概念,认为具有共同特征的学术活动是超越学科与时代的,具体包括:探索(Discovering)、注释(Annotating)、比对(Comparing)、咨询(Referring)、取样(Sampling)、阐释(Illustrating)、表达(Representing)等七个方面[②]。C.L. Palmer等学者定义了在虚拟网上研究环境中的五个核心学术基本体:搜寻(Searching),收集(Collecting)、阅读(Reading)、写作(Writing)与协作(Collaborating),其中每一个学术基本体中又细分为若干个,总计16个二级学术基本体,比如“合作”中又有协同(Coordinating)、联网(Networking)、咨询(Consulting)等[③]。Tobias Blanke与Sheila Anderson等学者基于数字人文研究基础设施的使用角度,通过对人文研究学者的深度访谈调查[④],总结出五个核心学术基本体:探索(Discovering)、收集(Collecting)、比对(Comparing)、发布(Delivering)和协作(Collaborating),以及多个细粒度的二级学术基本体[⑤]。以上学者提出的“学术基本体”研究,可以看成是数字人文研究“方法共同体(Methodological Commons)”[⑥]的概念化及具体阐释。虽然在虚拟研究环境中,数字人文研究方法存在一定的共性,但是也要深刻意识到不同人文学科之间的研究对象、研究方法、研究过程的差异性。

中国农史研究有近百年历史,从工作的主要内容来看主要分为两大阶段:一是20世纪初到20世纪80年代中期,工作重心为农史研究基本资料的收集与整理,在这期间基本上中国最重要的古农书均已被梳理一遍,这也为现今的数字人文研究基础设施建设奠定了数据基础;二是20世纪80年代后期,完成资料收集与整理阶段性任务后,逐步向农业科技史和农业经济史研究,研究方法也更具多元化趋势。张波对农史学科的研究方法体系进行了详细划分,他提出基本研究方法包括传统的文献研究方法、考古学与民族学研究方法以及新兴的科学研究方法如比较农史研究、计量农史研究、系统农史研究等[⑦];王思明认为传统农史研究主要采用历史学、文献学、版本目录学和古文字学等研究方法,现代的研究开始大量借鉴其他学科的研究方法,如经济学、社会学、人类学、民族性、计算机科学、统计学、考古学、农学等,特别是吸收了欧美及日本等国的研究经验与视角,开始注重比较研究方法的运用(时间、空间及时空的综合比较)、计量学与统计方法的应用等[⑧]。

综合来看,农史学科具有历史学、生物学、环境科学、土壤学、经济学等多学科的特征,其跨学科属性也决定了它与广义的历史学研究还存在一定的差异性,近些年许多优秀农史研究成果中定量分析的比重明显提升,领域学者也开始注重将前期整理的农业古籍资料开始数字化,并运用计算机信息技术诸如文本分析、内容挖掘、地理空间分析、社会网络分析等对资料进行处理。如南京农业大学科学技术史(农业史)博士点在2003年就开辟了“科技史信息组织”方向,以侯汉清为首的研究团队取得了令人瞩目的研究成果,出版了“中国文化典籍计算机整理与开发技术研究系列”丛书。可以说,相较于其他人文学

① 刘炜、叶鹰:《数字人文的技术体系与理论结构探讨》,《中国图书馆学报》2017年第5期。

② J. Unsworth, Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this, Symposium on Humanities Computing: Formal Methods, Experimental Practice, London: King's College, 2000, Vol. 13, pp.5–00.

③ C. L. Palmer, L. C. Teffeau, C. M. Pirmann,Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development, 2009.

④ Anderson S, Blanke T, Dunn S, Methodological commons: arts and humanities e–Science fundamentals, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 2010, Vol. 368, pp.3779–3796.

⑤ T. Blanke, M. Hedges, Scholarly primitives: Building institutional infrastructure for humanities e–Science: Future GenerationComputer Systems, 2013, Vol. 29 No.2, pp. 654–661.

⑥ W. McCarty, H. Short, Mapping the field, Report of ALLC meeting held in Pisa, 2002.

⑦ 张波:《试论农史学科层次结构和理论方法体系》,《中国农史》1992年第2期。

⑧ 王思明:《农史研究:回顾与展望》,《中国农史》2002年第4期。

科，我国农史学者是较早意识到采用数字人文研究的方法来拓展研究领域和内容的，尽管在当时，“数字人文”的概念还未在国内落地与普及。而这些领域学者的研究活动也是具有一定的共同性，也就是上文提到的“学术基本体”，笔者将Tobias Blanke提出的学术基本体与农史领域已有相关研究成果中的研究情景以及刘炜、叶鹰提出的数字人文技术体系[①]进行映射，结果如表1所示。

表1　　农史领域数字人文研究情景、学术基本体与技术体系映射

研究情景	学术基本体	技术体系
基于XML技术对《齐民要术》进行文本数字化，构建知识库	收集	数字化技术，数据管理技术
收集近现代农史研究和著作30000余篇（册），进行数字化加工和整理	收集	数字化技术
从网络数据源中筛选并导入引文数据，构建国际和国内的杂交水稻育种研究引文编年图	探索，收集	数据分析技术，可视化技术
基于已有元数据标准设计农具藏品元数据方案，实现藏品信息的详细而准确的著录	比对，探索	数据管理技术
构建基本词典群包括人名、地名、书名、职官名、物产名等，用于农业古籍计算机自动分词标引研究	比对	机器学习技术
采用基于关联规则挖掘和基于自然语言处理两种方法相结合的方式，从古农学研究论文中获取领域概念的属性关系	比对	机器学习技术，数据管理技术
建立GIS农业种植信息专题数据库，按照江苏地级市和县域归属为基本单元进行数据收集和整理规范，形成与农作物种植相关的数据集	收集，比对	数据管理技术，数据分析技术
构建并发布面向概念检索的农史信息门户	发布	数据管理技术
多位学者对于《方志物产》研究素材的多角度、持续性研究	协作，发布	数据管理技术，数据分析技术

将领域学者的学术研究活动进行归纳得到学术基本体，而与之相关的资源、工具、服务等，均为研究基础设施建设需涉及的方面。刘炜等学者提出数字人文研究基础设施框架应分为三个层次，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，这一层是数字人文研究活动的主体；外层是数字人文成果发布、与社会交互、产生社会影响的界面层，由门户或平台形式呈现[②]。对于农史领域的数字人文研究，此框架同样适用，其核心层文献资源大致包括史书、古农书、地方志类编物产资料、文人文集游记、农史研究文献、自然科学研究数据、农业经济数据等等。

纵观农史领域前期数字人文研究，大多数为个人的特定选题研究（以学位论文为主），其问题在于多数文献资源的数字化、数据化乃至知识化的过程存在不可通约性，如元数据标准设计缺乏评价、数据库构建缺乏规划、相关本体的不可复用、软件工具非开源等，这些也导致前期的研究数据无法进一步为其他研究者所用，与其他各类数据源的数据无法融合以及软件工具的功能扩展性较差等一系列问题。基于此，面向农史领域的数字人文研究基础设施建设势在必行。

三、方志物产知识库构建及深度利用研究

农史领域数字人文研究基础设施建设需以文献资源为核心，资源的独特性与唯一性是研究基础设施建设必要性的前提，也是区别其他以机构为导向的研究基础设施的标志。在农史领域，古籍方志中记载的物产资料是重要的研究史料，是领域学者进行相关研究不可忽视的重要文献资料。在本节中，笔者以中华农业文明研究院特藏文献《方志物产》资料为核心资源，结合前期相关研究成果以及未来研究工

① 刘炜、叶鹰：《数字人文的技术体系与理论结构探讨》，《中国图书馆学报》2017年第5期。

② 刘炜、谢蓉、张磊：《面向人文研究的国家数据基础设施建设》，《中国图书馆报》2016年第5期。

作计划谈一谈面向农史领域的数字人文研究基础设施建设构想。

(一)《方志物产》简介

方志是历史研究的必需文献,从清代开始,已经形成了一门独立的学问。方志以志为主体,有述、记、志、传、图、表、录等,在历时性的维度下对特定区域的建置沿革、分野、疆域、城池、山川、坊郭镇市、土产、风俗、户口、学校、军卫、郡县廨舍、寺观、祠庙、桥梁、古迹、宦迹、人物、仙释、杂志、诗文进行描述和记载[①]。其中的"物产"几乎一直是方志必载项目,简称方志物产。方志物产记载一地的动植物资源(*部分方志物产也包含货物,如矿物资源*),方志物产是方志中记载农业最多、最集中的部分,这在以农立国的中国有着更为重要的地位,传统中国是农业社会,无论是研究古代史还是近代史都或与农业发生联系。

1924年,主政金陵大学农业图书研究部的万国鼎先生,开始计划辑录古书中有关农业的资料"片段的农学记载",汇编为《先农集成》,开始了方志的搜集工作,后由于战争中止;1949年,中国农业遗产研究室成立伊始就开始重启方志的查抄工作,依旧由万国鼎先生负责,其工作团队足迹遍布40多个大中城市和100多个文史单位,到1958年查抄方志工作基本完成,1959年整理,1960年初编成《方志物产》449册、《方志综合》111册、《方志分类》120册,共680巨册3600余万字,成为今天中华农业文明研究院的镇院之宝[②],其中以《方志物产》价值为最大。

概言之,《方志物产》是上个世纪建国前后,大批有识之士在万国鼎先生的策划和组织下集一代人心血精心搜集、挑选和抄写装订起来的大型方志类文献汇编,具有唯一性和不可替代的丰富性,海内外未见同类型的其它文献可与之媲美。

(二)面向农史领域的数字人文研究基础设施建设规划——以方志物产知识库构建及深度利用为例

本文提出面向农史领域的数字人文研究基础设施建设可先以方志物产知识库构建为首要工程,基于该知识库可对方志物产资料展开知识发现、知识考证以及深度利用研究。方志物产知识库构建步骤具体可分为四个环节:数字化、数据化、知识化及平台化。打一个比喻,"数字化"的工作是将活牛进行屠宰与清洗;"数据化"的工作是将牛进行肢解,并将各部位按照用途进行初步加工;"知识化"的工作是将初步加工的部位按照食谱与其他食材按严格比例进行烹煮并得到最终的食物,如一块"菲力"牛排(*牛之里脊肉*);"平台化"的工作就是要解决用什么样的餐具、配合什么样的美酒或是在什么样的就餐环境去消费这一块牛排,让食客得到更完美的体验。数据科学家、领域专家这些专业的"厨师"将贯穿在基础设施的建设过程之中。方志物产知识库构建框架如图2所示。

(1)*数字化——方志物产资料数字化整理与加工*

对手抄孤本《方志物产》进行数字化是其得以保存和利用的重要手段之一。在此基础上,还需以国内外各种方志目录为线索,对相关资料进行二次辑录、整理与查漏补缺,形成更为完整、全面的方志物产资料,这也是数字人文研究基础设施建设的前期基础性工作。

《方志物产》原始文本的地域范围几乎涵盖国内所有行政区划,时间跨度从宋代至民国,内容体系包括目录、序言、正文和结语。首先,需由数据科学家确定重新整理与辑录后的方志物产资料数字化的整体框架,针对原始《方志物产》体例进行编码设计,编制历史时间索引、来源志书索引、行政区域索引等,选取人工录入和机器扫描相结合的策略实现方志物产资料的数字化;其次,根据方志物产资料的字词分布特征,对于文本内容的繁简呈现、汉字编码集的确定以及生僻词造字方法的选择等相关汉字录入问题给出行之有效的解决方案;再次,针对方志物产资料的文献资源内外部特征,借鉴国内外多种元数据标准,例如都柏林核心元数据、国家图书馆地方志描述元数据等,设计方志物产描述元数据;最后,对于录入的方志物产电子文本,依据行文格式及相关内容设置数据库字段,同时结合机器扫描的图像及其相应

① 仓修良:《方志学通论》,华东师范大学出版社,2013年。

② 万国鼎:《中国农业史整理研究计划草案》,载王思明、陈少华主编:《万国鼎文集》。

处理，完成涵盖序言（序）、检索样例说明（叙例）、来源方志名称拼音检字、行政区域拼音检字、年代和正文以及手抄孤本原貌（图像）的基本素材库的构建。文献资源数字化是整个基础设施建设的基石，其资源数字化的质量决定了后续基础设施建设的成败。

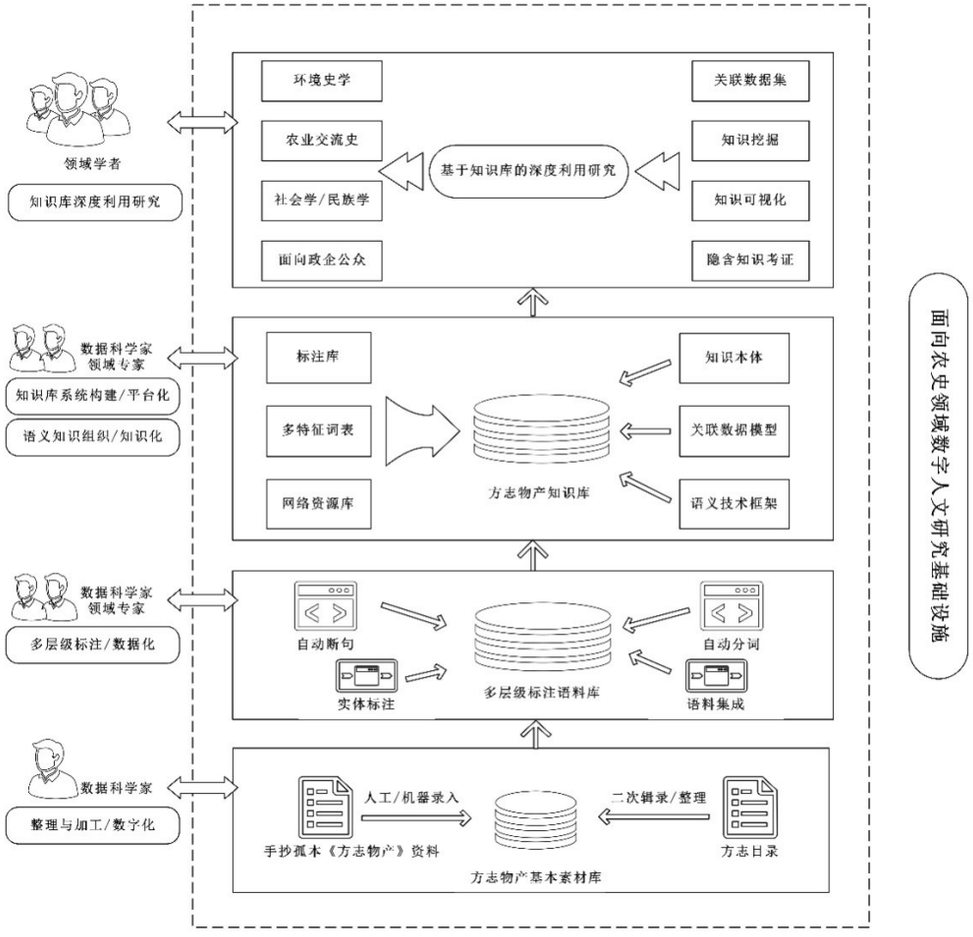

图2 方志物产知识库构建框架

（2）*数据化——方志物产资料多层级标注*

在这一环节将要实现数字化文本到数据化语料库的转换。首先，在方志物产基本素材库的基础之上，由领域专家研究并归纳各时期不同地域方志物产资料的知识书写的差异性，得到方志物产多特征词汇底表；再由数据科学家针对词汇的显性与隐性语义知识，通过人工标注、机器学习的策略完成方志物产资料数字化文本的分词、词性、命名实体和断句的自动标注，同时构建基于方志物产资料的自动分词、自动词性标注、自动命名实体识别和方志自动断句模型，通过不断修正模型提升标注数据集质量，实现对方志物产资料中蕴含的“人、时、地、物、事”等命名实体的一体化标注，最终得到一个多层级标注的结构化的方志物产语料库。一些有丰富软件开发或工具使用经验的数字人文研究者可直接利用语料库中的数据集来展开他们预设的各类主题研究。

（3）*知识化——方志物产资料语义知识组织*

对方志物产资料文本仅仅进行浅层次的数字化与数据化，并不能满足领域学者对古籍方志进行文本挖掘、语义分析和知识发现的需求。在这一个阶段仍然需要领域专家与数据科学家通力合作，先需对

网络环境下方志物产资料语义知识组织的深度序化机制及实现路径进行探究；然后在此基础上构建方志物产领域知识本体，利用本体对相关资源（包括获取的网络资源）进行语义标注，建立词汇间的语义关系，如上位词（属关系）、主题词（用关系）、其下位词（分关系）、参见词（代关系）等，并存储对象类、属性以及对象之间的语义关系，作为后续方志物产知识库构建提供深度的语义层面知识。对方志物产资料展开语义知识组织，满足书目控制和规范控制、数据重用与共享等需求，是数字人文研究基础设施建设的重要环节。

（4）平台化——方志物产知识库平台构建

“平台化”是数字人文研究基础设施的“门户”建设也是核心部分，即采用关联数据的一整套技术、方法和流程，实现为领域用户提供各种知识服务的知识库系统平台。首先，基于方志物产知识本体设计关联数据模型，尽可能复用互联网已有成熟词表，对方志物产元数据进行数据清洗，提取概念实体并赋予HTTPURI，基于方志物产知识本体定义的类和属性来描述实体及实体间的关系，采用机器可读的RDF序列化格式进行编码与存储。接下来，使用关联数据四原则发布方志物产关联数据，运用SPARQL查询语言与语义技术开发框架存取和操作数据，同时运用可视化技术构建知识库呈现平台展现数据，提供数据开发接口供其他程序调用，采用关联数据开放与消费的方式实现知识聚合应用。最后，基于知识库为领域学者与公众用户提供面向数字学术与公众史学的各项知识服务的软件应用模块，实现诸如研究数据存储、知识检索与可视化、众包编辑、专题服务以及社交化应用等平台功能。

（5）数字人文研究基础设施的作用——方志物产知识库深度利用研究

方志物产知识库的建成将会是面向农史领域的数字人文研究基础设施的重要组成部分，但基础设施应是一种生态系统，即需有领域专家以及其他用户的参与，不断提出问题与需求，通过迭代在知识库中解决问题才是一套成熟的基础设施应实现的“落地”功能。

方志物产知识库的深度利用应首先建立在知识发现与考证基础之上，即通过人工甄别与机器比对相结合的方式，对提取的物产数据集进行考证，借助社会网络分析、地理信息系统等技术方法对知识库中的隐含知识进行挖掘与分析，如明至民国时期特定植物在全国范围内的分布情况、不同地区物产的丰富程度以及物产随时间变迁的消长情况、特定物产在时空框架下的变迁路线、物产与相关人物的关系等。

接下来再由领域专家对获得的隐含知识采用历史文献学的方法进行多轮专业考证，如物产的同名异物、同物异名，不同版本的志书与引书比对，特定物产的消长变化与变迁路线等等问题；在此基础上，领域学者可运用农学、动物学、植物学、生态学、历史地理学、农业经济学、社会学、民族学和人类学等学科理论与研究方法，研究特定物产与人类社会之间的复杂互动的整体关系，例如以下三个主题研究可按此路径展开：

一是基于环境史对动植物资源的数量和种类的分布及增减加以探源分析。二是中外农业交流路径上物产的时空变迁问题。进一步梳理一带一路上的外来作物的传播时间、路径及其经济价值。三是基于社会学及民族学的视野，立足于物产本身，梳理与该物产有关的社会、经济、文化，从而具而微地展现当地生活文化，解剖当地区域文化、民风民俗的形成与演变，增强文化自信。

同时除面向领域学者以外，还应兼顾政企与公众需求，围绕方志物产资料开发与利用模式展开研究，如促进科普知识传播、扩大旅游资源开发及提升农业遗产保护等。

四、总结与展望

我国历史悠久，文化遗产丰富，古代典籍文献中有许多农业科学技术方面的珍贵资料，可以帮助今

人考证农产品与农业技术的历史起源、辨别有关农业动植物和器物的名实异同，以及为当前农业生产和科学研究提供启示与借鉴。在农史研究中，古农书与方志向来是农业历史文献的主体，万国鼎先生曾明确指出：辑录古籍上有关农业的资料，方志最为大宗。

时至今日，各种古农书与方志的搜求、编目、校勘、注释、今译、辑佚、典藏、影印等工作已颇具规模且成果累累，但是在数字化、知识库建置等环节相关研究工作还刚刚起步，未成气候。数字人文研究基础设施是一种支持人文科研活动的通用基础架构，是在数字环境下为开展人文研究而必须具备的基本条件，可以是国家层面的，也可以是地区行业或组织机构层面的。研究基础设施的建设对于农史乃至历史学研究均具有深远的意义，有利于学科中各个项目数据资源的共享与关联、通用型软件工具与应用开放接口（APIs）的互操作以及人员协作模式的平台化与制度化。

目前，对于学界而言比较紧迫的任务是制定一些可持续发展的机制来构建并改进相关研究基础设施，正如之前在上海哈佛中心举行的“中国历史研究的网络基础设施国际研讨会”就已汇聚国内外诸多领域专家共商此事。

本文提出以中华农业文明研究院的特藏文献资源《方志物产》为例，通过数字化、数据化、知识化、平台化等步骤构建方志物产知识库，以此拉开面向农史领域的数字人文研究基础设施建设的序幕。可以预期的是，该基础设施的建成将不仅有助于农史领域内数字人文研究的深入开展，同时也可为未来行业乃至国家层面的研究基础设施建设添砖加瓦，从而推动具有中国风格的数字人文研究体系的形成。

[参　考　文　献]

[1] 王思明 . 农史研究：回顾与展望[J] . 中国农史，2002，(4) .
[2] 张　波 . 试论农史学科层次结构和理论方法体系[J] . 中国农史，1992，(2) .
[3] 刘　炜，谢　蓉，张　磊，等 . 面向人文研究的国家数据基础设施建设[J] . 中国图书馆学报，2016，(5) .
[4] 刘　炜，叶　鹰 . 数字人文的技术体系与理论结构探讨[J] . 中国图书馆学报，2017，(5) .
[5] 包　平，李昕升，卢　勇 . 方志物产史料的价值、利用与展望——以《方志物产》为中心[J] . 中国农史，2018，(3) .

杂志社版权页声明