

从历史档案到历史大数据:基于威尼斯时光机十年路径的探索

祁天娇

摘 要 在全面数字转型的国家战略背景下,包括档案馆在内的历史文化遗产收藏机构正面临着建设历史大数据的机遇和挑战。将历史馆藏转化为历史大数据,改变历史档案原有封闭、单维、低价值密度和隐晦知识结构的状态,用新的网络化、开放化、关联化、结构化的状态连接更多类型的历史资源,以走出历史档案在当代的价值实现困境,是历史档案数字转型的可行方案。本文以历史档案如何转化为历史大数据为研究问题,以历史档案的价值实现为研究目标,以案例分析为研究方法,选择当前以历史档案为依托建设历史大数据最为典型的“威尼斯时光机”项目为分析对象,解读该项目关于历史大数据的理论成果,总结该项目十年间探索的数字化、数据化、语义化和知识化的实践路径,并将其转化为具有通用性、可复用价值的经验,以期为数智时代的历史文化资源保护、量化历史研究以及新文科背景下的交叉学科发展,贡献档案学的方法与力量。图1。参考文献27。

关键词 数字记忆 历史档案 历史大数据 数字转型 威尼斯时光机

分类号 G270

From Historical Archives to Historical Big Data:Exploration Based on the 10-year Path of Venice Time Machine

QI Tianjiao

ABSTRACT

In the background of social comprehensive digital transformation and construction of new liberal arts, the construction of historical big data is becoming the common mission of historical collection institutions. As the most important historical data source, historical archives have been closed, one-dimensional, of low value density and of obscure knowledge structure for a long time. How to transform historical archives into historical big data is the core problem to be solved in this paper.

Based on the value of historical archives implementation perspective, through analyzing on the path and key elements of historical archives to historical big data, this paper explores the way through which historical archives may walk out of the contemporary dilemma. The exploration may contribute to the history cultural heritage protection, quantitative research and new arts development in the background of cross domain.

With case analysis as the core method, this paper chooses the “Venice Time Machine” project as the analysis object, interprets the project’s exposition of historical big data, and summarizes the digitalization, datafication, semantic organization and knowledge transformation of historical archives explored by the project in the past ten years. This paper transforms the case into the experience of transforming historical

通信作者:祁天娇,Email:qtjjoy@163.com,ORCID:0000-0001-5595-4774 (Correspondence should be addressed to QI Tianjiao,Email:qtjjoy@163.com,ORCID:0000-0001-5595-4774)

archives into historical big data with universal and reusable value.

The author believes that the historical archives could provide a huge contribution to the construction of historical big data, including resources and methods. From historical archives to historical big data, the fundamental change lies in the new network and open characteristics of data structure, which promotes the upgrading of the technical competencies of data processing and the transformation of knowledge production paradigm. This transformation generally needs to go through four stages of digitalization, datafication, semantic organization and knowledge producing, pursuing the version conversion, syntax conversion, semantic conversion and pragmatic conversion of historical archives, in order to realize the recollection and remapping of historical narration in the digital era. 1 fig. 27 refs.

KEY WORDS

Digital memory. Historical archives. Historical big data. Digital transformation. Venice Time Machine.

0 引言

随着社会全面数字转型以及新技术的渗透性应用,历史遗产的传承体系也被快速积累的数据打断并重塑。数字时代历史资源的未来,正在超越资源数字化的技术问题,走向新的结构与逻辑。其中,历史档案作为最重要的历史资源之一,大规模的数字化已经开展多年,虽然在技术突破上仍然存在许多挑战,但从价值视角来看,如今面临的最重要问题是如何从数字载体中提取和阐释它们的信息结构^[1],尤其是要以符合当代数据思维的历史逻辑来重新理解这些信息及其结构,形成时代叙事。在这个问题引导下,历史大数据建设成为不可逃避的基础工作。历史大数据是指将形成于过去的历史记录,以当代叙事逻辑重新汇聚、组合,所形成的具有广泛、共享、结构化特征的数据集合。历史档案是具有原始记录性的历史记录,是最重要的历史大数据源。在不同历史时期,历史档案的汇聚都曾带来广泛的、共享的信息洪流,支持过对历史问题的寻解^[2]。历史档案转变为历史大数据具有天然的、经过时间验证的可行性。

但令人遗憾的是,在很长时间里,被束之高阁似乎是历史档案不得已接受的宿命。正如威尼斯国家档案馆所藏的长达 80 公里排架的历史档案一样,这些不间断记录着从中世纪到 20 世

纪威尼斯珍贵历史的大量手写档案,内容从官方行政文书到公众的出生证明、死亡证明、纳税证明等日常文书,再到城市规划设计、建筑图纸和地图,类型多样、内容全面,几乎覆盖了威尼斯上千年历史的方方面面。但因为这些历史档案大多采用拉丁语或威尼斯方言手写体,在数字化之前几乎没有被现代历史学家识读过。直到近十年,这些历史档案的珍贵面貌及其记录的宏大历史,才借助“威尼斯时光机”项目重现在世人面前。“威尼斯时光机”(The Venice Time Machine)项目于 2012 年由洛桑联邦理工学院(École Polytechnique Fédérale de Lausanne, EPFL)和威尼斯卡福斯卡里大学(Ca' Foscari University of Venice)主导发起,威尼斯国家档案馆作为最主要的历史档案资源提供者参与合作,另有 300 多名来自基础学科、工程学、计算机科学、建筑学、历史学和艺术学等不同领域的学者和学生参与其中,愿景是通过历史档案数字化重建威尼斯千余年历史的多维模型,并为科研与教育领域开放大型历史数据库。“威尼斯时光机”项目自 2015 年开始关注数字人文中的大数据问题,至 2017 年提出独有的“历史大数据”(Big Data of the Past)核心概念,在历史档案数字转型为历史大数据的问题上,探索出一条已被欧盟认可并广泛推广的路径。

本文选择“威尼斯时光机”作为案例分析对象,关注其在世界范围内具有示范意义的历史

大数据建设经验,以其项目团队公开发表的学术论文、会议报告、博客以及各类开放文本为研究资料,分析项目开展近十年来在历史大数据的理解、建设与应用方面的成果与路径,探讨该项目如何将威尼斯千年历史档案通过数字化、数据化、语义化、知识化等工程,转化为可支持历史重建与仿真的历史大数据,并以此启发我国历史档案的数字转型工作,为数字时代的历史文化遗产保护、量化历史研究以及新文科背景下的交叉学科发展,贡献档案视角的创新与力量。

1 研究与实践现状

我国学界对“威尼斯时光机”项目的关注,始于“十三五”期间数字档案馆建设与档案资源数字化的浪潮中,并在近几年数字人文的迅猛发展中对其有了更深的理解。早期关于该项目的案例研究,主要从“威尼斯时光机”的数字人文理念和历史档案图文识别技术角度开展^[3],强调数字人文理念对档案资源开发利用的影响^[4],并以此为目标改变历史档案的整理与开发思路^[5]。近两年在数字人文的新学科视角下,图情领域的学者发现并整理了“威尼斯时光机”在文化遗产数字化、智能化方面的贡献^[6]。“威尼斯时光机”是从历史档案出发走向宏观历史叙事的代表性数字人文研究,其项目规划已经超越资源数字化本身,而面向更庞大的城市与人文历史重建。研究者需要进一步突破现有学科框架,将历史档案作为历史数据源之一放在更宏大的历史叙事中,来理解“威尼斯时光机”为增强历史档案价值及与其他历史资源关联性所做出的巨大努力;尤其需要挖掘“威尼斯时光机”项目所代表的数字人文或数字文化遗产模式背后的核心创新概念与人文主义实现路径——历史大数据。

历史大数据是量化历史研究的产物,其本质是经过数据化的海量历史资料,主要表现为各种形式的历史数据库^[7]。20世纪90年代以

来,一种将大批历史材料结构化录入数据库,并依靠定量分析揭示其中隐含的史实,检验和发展历史认识与经验的新方法逐渐在国际学术界流行起来^[8]。例如,通过对美国国家谱学会保存的家谱人口档案进行数据库处理,美国医学领域实现了对家族乳腺癌基因遗传研究的重大突破^[9];通过对一个世纪以来美国教育、职业和收入的历史数据分析,美国经济学家分析了历史变革、技术革新、教育进步对收入分配结构的影响^[10]。受到这些重要研究成果及其方法的影响,史学界历史资料量化数据库的建设与研究快速增长。一方面,建设大规模史料数据库已经成为国内外史学研究的新共识。美国学者李中清团队花费20年时间建立基于八旗户口册和清代皇室族谱资料的“中国多代人口系列数据库”(China Multi-generational Panel Data Series, CMGPD)^[11];美国哈佛大学和中国北京大学合作建设的“中国历史人物传记资料库(CBDB)”,目前已收录52万余人的传记资料,其资料主要来源包括地方志、官方文书、明清档案等历史档案^[12];国内史学界也先后建设了“中国地方历史文献数据库”^[13]、“清代灾荒纪年暨信息集成数据库”^[14]等史料数据库。另一方面,随着大规模史料数据库的建设,学界逐渐认识到历史档案数字化与数据化是建库必需的基础性工作。实际上,国内外在2000年以后出现的不同规模的历史数据库,共有前提就是20世纪90年代开始的历史档案电子化与数据库建设工作。2011年,哈佛大学学者通过“谷歌图书计划”(Google Books Project)构建起大规模出版物数据库,并利用词典编纂、语法进展、语义分析、集体记忆等方面的研究,展示了如何通过信息资源数字化发现历史新知^[15],为信息资源管理视角下的量化历史研究提供了范例。近年来,在数字人文研究范式的影响下,很多文化遗产管理机构、历史资源典藏机构以及学术研究机构都在积极探索信息资源数字化背后巨大的文化发现潜力。Internet Archive、Europeana以及本文所研究的“威尼斯时光机”等都是国际典型的信

息资源管理视角下的历史大数据项目,国内的数字敦煌等项目也在积极探索量化历史研究的新路径。以上海图书馆团队为代表的图情领域学者已经有意识地对历史数据基础设施建设中的学科方法进行总结^[16],但遗憾的是目前档案学界还缺乏对这一领域的足够重视与充分参与。

本文立足档案视角,以历史大数据建设过程中的档案资源建设与转型为切入点,具体探讨历史档案与历史大数据的关系,以十年来“威尼斯时光机”的历史档案数据化实践为依托,探讨从历史档案到历史大数据的转型框架与关键点,旨在拓展历史档案的工作思维,为更多档案领域主导的历史大数据建设项目提供可复用的方案或有益参考。

2 历史大数据的档案视角解析

历史大数据是一个广义的学术概念,历史学、文化遗产保护、大数据等领域各有相关但又存在差异的学术阐释。本文立足档案视角,探讨历史大数据在广义且跨科学学术框架下的基本内涵与本质特征,并在此基础上探讨历史大数据与历史档案的关系,明确历史档案向历史大数据转型所需达成的目标。

2.1 历史大数据的内涵与特征

历史大数据发端于历史领域的量化研究,是一种以大数据方法重组历史记录所形成的、具有历史研究价值的数据集。肖恩等认为历史大数据为数字时代的历史学研究带来了新的“宏观视角”^[17]。从资源角度讲,历史大数据是一种全新的历史资料模态,是传统史料自我解构并重构所形成的、具备大数据特征的数据态史料,这种史料可更好地支持数字环境下公开、共享的历史研究与成果传播;从管理角度讲,历史大数据是一种全新的历史资料组织与管理手段,能够利用文本挖掘、主题建模、网络分析等数据管理手段,建立新的历史资料秩序,挖掘新

的历史资料关系;从方法角度讲,历史大数据是一种全新的历史研究方法,这种方法应用了大数据思维,但并非历史研究的最好方法或唯一方法,而是更适合解决一些微观研究方法无法解决的历史“远读”问题;从知识角度讲,历史大数据既是知识来源,也是新知识本身,正如《探索历史大数据》一书所说,历史大数据是一种“生成性”的方法,“不仅可以证实过去的故事,而且能够生成新故事、新观点”^[17]。本文从档案视角解析历史大数据,即主要从资源的角度探讨历史档案与历史大数据的联系与区别,并以此来指导历史档案的管理、方法和知识层面的转型。

历史大数据作为“大数据”的一种,也具备大数据的基本特征,包括容量大、种类多、速度快和价值密度低。但与其他记录社会生产生活领域的大数据不同,历史大数据主要用于记录、描述、分析、挖掘和研究历史,是历史领域的大数据,可作为一种重要的领域科研基础设施。历史大数据在其建设、发展与应用过程中,具有独特的领域大数据特征。

(1) 泛在性:指建设历史大数据所依据的广泛存在的数据源。一方面,传统史料广泛的“存量”数字化为历史大数据打下了最坚实的基础,这是包括档案馆在内的历史文化馆藏机构在过去二十年间开展的重要工作之一。《“十四五”全国档案事业发展规划》明确提出,“十四五”期间中央和国家机关传统载体档案数字化率要达到80%的整体性目标,这为历史档案参与历史大数据建设并作为历史大数据重要数据源做好了充分准备。另一方面,广泛存在的数据“增量”使得任何历史事件、活动或人物的信息都有可能被记录并保存下来,且现有的数据与方法已能够支持对这些信息的进一步管理和挖掘,而这些“增量”产生并留存的最主要方式就是建档或归档。《“十四五”全国档案事业发展规划》也明确提出,要统筹重大历史事件、重大活动、突发事件应对活动等档案专题库建设。这意味着历史数据建档归档将成为国家档案资源建设

的重要任务,也是档案领域参与历史大数据建设的重要形式。

(2)拓展性:指历史大数据在其发展过程中始终具有可拓展的数据结构。历史大数据在建设初期具有鲜明的主题性,并在该主题下根据数据源的数量和质量呈现一定的数据规模。但是历史大数据的数据结构是开放的、可拓展的,可以在主题约束下不断建立数据关系、扩大数据规模,甚至可以在相关主题之间建立联系,形成大规模数据网络。形象地说,历史大数据在建立的过程中,也将经历从单维数据表向关系型数据库再到知识库的转变。同样,历史档案在数字时代也在孕育着自己的“拓展性”,包括汇集层、文件层、实体层等不同层面的可拓展,即历史档案要进一步强化与其他类型历史数据之间的关联,加强一定汇集内文件与文件之间的关联,加强档案内容中实体与实体之间的关联。历史档案是否具有强拓展性,将决定历史档案能否最终汇入历史大数据之中。

(3)再生产性:指历史大数据的应用价值不仅在于本身,更在于可不断生成新的数据、新的知识和新的叙事。历史大数据的建设与发展过程,是一种从零散历史资源重组为新的历史数据产品的过程。在这一过程中,包括历史档案在内的大规模多源异构数据汇聚并融合,以一种完全不同于原始存在模式和状态的新形式被生产和呈现。也就是说,历史大数据的建设与发展是历史大数据的首次生产过程,它的应用可以有多种形式,例如可支持用户在数据库中对数据的直接检索和挖掘,可支持技术人员不依赖历史大数据存储库或平台而对数据集进行直接调用,可支持开发人员利用数据进行知识推理、形成可视化的数据产品或新知识等。历史大数据的最大价值,在于研究者可以利用它进行研究与创新,而这些灵感往往来自历史大数据的泛在性和拓展性。也就是说,历史大数据的应用是其第二次生产过程,历史大数据的泛在性和拓展性决定了它的再生产性。同样,历史档案的价值除了被长期保存所象征的证据

价值以外,还有巨大的、可支持知识与叙事再生产的信息价值与记忆价值。

2.2 历史大数据与历史档案的关系

正如上文所述,历史大数据源于历史研究领域,但如今已超越历史学范畴而成为数字时代支持交叉领域研究的重要基础设施。按照“威尼斯时光机”首席专家卡普兰教授的观点:“历史在社会转型和技术迭代的不同时期会呈现出不同的数据特征”^[2],即历史大数据之“大”是具有历史相对性的,在每个历史阶段,机构、组织和个人产生的具有重要价值的数据都会按照当时的逻辑被组织起来。在这些数据经年传承的同时,组织这些数据的逻辑与方法却在不断更迭。掌握不同的数据组织逻辑之间的映射方法,用当代的数据组织范式来重新整理不同历史时期的数据,是理解、掌握和应用历史大数据的核心要义。本文所探讨的历史大数据,是当代的历史大数据,是以当代叙事逻辑重新组织产生于过去的历史数据所形成的大数据集。历史档案正是产生于过去的历史数据,从历史档案到历史大数据不仅是资源状态的改变,更是从旧有数据逻辑向当代数据逻辑、从过往叙事规则向当代叙事规则的转变。这就意味着历史档案与历史大数据之间,至少具有资源和方法两个层面的关系。

(1)资源层面。历史大数据的资源规模与范围远超历史档案,还可能包括例如图书、文物以及其他历史文化遗产等在内的广泛的数据源,这些历史资源都具有转化为历史大数据的潜力。历史大数据之“大”就体现在多来源、多模态和多结构等方面,这也是历史大数据“泛在性”的表现。不同历史资源转化为历史大数据的路径和方法可能有所差异,但转化而来的数据结构应该是一致的或者是可相互映射的,转化所依据的数据标准应该是具有互操作性的。当历史档案转化为一定范围的历史大数据后,可以与更多来源的历史数据集形成关联,不断追求历史大数据的完整性,以全面重塑历史。

历史档案是不同历史阶段零散而广泛的历史数据中最具有长期保存价值的一部分。历史档案借助归档范围的确定,将历史复杂数据中最能够原始记录社会生产生活的那部分数据进行归档和长期保存,确保了从历史档案转化过来的历史数据集的真实性和准确性。在历史发展的长河中,由于保存和管理能力的限制,每个阶段的档案范围都小于当时社会生产的总数据,但毫无疑问这些档案是经过择选鉴定的具有更高长久保存价值的一部分,且与其他零散的社会数据不同,馆藏档案接受了科学系统的长期保存管理。因此在不同历史时期,档案都是记录、保存和寻找历史细节的首选数据集。例如,威尼斯国家档案馆(State Archives of Venice)自1815年成立以来,接收各类国家行政管理文件甚至部分来自意大利的文件,其广泛繁杂的档案接收范围,本身就确保了这些由历史档案转化而来的历史大数据的泛在性。

(2)方法层面。不同领域对历史大数据的建设、发展和应用可能贡献不同的方法论,就历史档案而言,因其极为稳定的组织逻辑与方法,更有利于解读产生于过去的历史数据,并在不同叙事逻辑之间建立桥梁。首先,历史档案的组织与管理具有长期的传统,能够在一定时期内以相对稳定的逻辑和方法开展历史档案的组织工作,遵循这样的逻辑和方法就能够检索与查询档案。例如,“来源原则”是最重要的历史档案管理原则之一,理解历史档案的来源及其在数字时代的丰富表现,就能够理解历史档案按来源分类、整理、索引的基本逻辑。早在古希腊和古罗马时期,就已经出现按来源、时间或职能分类的档案管理思想,与当代档案分类方法虽有细节差异,但底层逻辑具有一致性。依照这种逻辑,就可以帮助考古学家发现和整理历史档案的形成时间、主体与大致内容。其次,历史档案管理者长期积累完成的馆藏目录、主题词表、叙词表、分类方案等,都是历史档案组织的重要工具,也是转化为历史大数据索引、本体、关联图谱等自动化工具的重要基础。

基于资源和方法两个层面的关系,历史档案能够为历史大数据建设贡献两方面的核心价值:一是历史档案是具有最高永久保存价值和最原始记录特征的数据来源,能够确保历史大数据的可信性;二是历史档案具有长期稳定性和延续性的数据组织和叙事逻辑,能够确保不同历史时期数据的互理解与互操作,形成不断流的历史大数据。

2.3 从历史档案到历史大数据的本质变化

从历史档案到历史大数据的转变,其本质变化并不是记录内容的变化,而是数据结构的变化,表现在数据结构的网络化、开放化以及由此带来的数据处理的高技术难度和新知识范式的结构化^[2]。

(1)网络化。网络化是历史大数据的本质特征^[18],即数据与数据之间具有丰富的语义关系,当这些语义关联被建立起来时,历史大数据也就成为一种特殊的语义网络(Semantic Network)。网络化的前提是大规模的数据依照一定的结构被汇聚和关联,这就对计算和处理这些数据间相关性的大规模数据库提出了计算力的高要求。在转化为历史大数据之前,历史档案也天然地具有关联性,即每份档案与同一档案汇集内其他档案之间存在关联。这是因为档案所参与和记录的社会活动之间是相互关联的,甚至有学者认为档案之间的关联是档案的必要组成部分,当这种关联被破坏时,档案的真实性与内容可信性也会被破坏^[19]。但历史档案的这种关联性是垂直的、单维的、缺少交互的。一方面是因为这种关联存在于人工环境中,只能通过多级著录、树状分类体系这样的层级结构,或者依靠主题词表或叙词表中的参引、互引等单向标注来说明这种关联;另一方面是因为这种关联本质上是来自档案所记录的社会活动的关联,是一种强调档案与所记录活动因果关系的关联。而历史大数据关注的是数据语义层面的关联,这种关联除了来自档案所记录的社会活动,还可能来自档案内容所反映的社会知

识体系,是包括因果关系和相关关系在内的更广泛的知识联系。并且,这种关联凭借语义网技术,能够被扁平地、多维地、交互式地呈现,成为机器可理解、可操作的网络化结构。

(2) 开放化。开放化是历史大数据保持大规模和流动性的根本原因。历史大数据是开放的、动态的,数据集中会不断新增数据,这些数据可能来源于新的历史档案数字化、数据化成果,也可能来自不同类型历史资源的转型结果。更为重要的是,历史大数据的开放化是基于其网络化结构的,即并非所有数据都可以随意加入该数据集,只有具有语义关系的数据才被开放权限。历史大数据的开放化与网络化是相互促进的,开放接纳更多的关联数据,可以形成更广泛的数据网络,而更多的数据网络节点就可以识别更多相关数据纳入历史大数据。相较而言,历史档案是封闭的、界限明确的。呈现档案关联并不是人工环境下历史档案整理工作的首要任务,来源是档案汇集之间最明显的界限,一个历史档案汇集很难开放接纳不同来源主体生成的档案,不同类别之间的历史档案也很难融入同一个汇集。例如,在威尼斯国家档案馆的长期馆藏体系中,建筑图纸、地图和行政文书之间几乎没有交集,因为它们之间存在边界。历史档案的数据量是可统计的,例如多少全宗、多少排架、多少卷/份等,但历史大数据是趋于忽略统计的,或者以时间为单位统计一定时间段的数据流量,也就是说每一个历史大数据的理想状态不是一个数据集,而是一个数据流^[2]。

(3) 高技术难度。数据结构的变化,最直接的反映就是数据规模的急速扩大。历史大数据的规模已经超过传统手工环境所能承载的数据量,因此需要新的处理和管理策略以及密集型的存储和计算基础设施^[20]。对于档案机构来说,历史大数据的高技术难度是从历史档案的数字化开始的,并贯穿数字化之后的结构化与矢量化处理、语义关联与开放、知识构建与获取的每一个环节。例如,“威尼斯时光机”项目在威尼斯国家档案馆安装了专门服务器——“时

光机箱”(Time Machine Box),用于存储被数字化扫描的全部文档及其元数据,并可以通过国际图像交换标准 IFF 协议对外提供访问。该服务器不仅支持大规模历史档案数字化成果的存储,还支持威尼斯国家档案馆或者其他外部机构对服务器上的数字化图像应用各种图片分割或文本识别算法^[21]。此外,“威尼斯时光机”项目还开发了专门的手写文档识别系统,用于手写历史档案的文本识别。

(4) 知识结构化。琐碎的历史大数据对于科研人员或者公众来说,价值是低密度的,需要进一步将“流量”数据转化为结构化的知识系统,以聚合人物、地点、主题、事件等历史知识单元。聚合完成的知识系统能够以机器可读、可理解的形式化语言表达出来,以便深度的数据挖掘和交换。为此,“威尼斯时光机”采用一种基于语义网技术的历史知识编码方法,详细记录历史研究过程中的各种信息源,并将这些信息源与给定的编码链接起来,形成一种知识序列。不同历史研究中,使用不同信息源所产生的历史叙事结果,就形成了不同的知识序列,这些知识序列融合构建起一种能够表示多个历史事实的系统,共同记录可能存在的知识空间的形成过程^[22],从而实现多维历史叙事。广义来讲,这种流量聚集和结构化知识改造的过程,也是历史档案数据化(datafication)的题中之义^[23]。

3 “威尼斯时光机”的十年路径

威尼斯共和国作为欧洲最早的资本主义政权,在 13 世纪晚期就有了按时间线索对档案排序并装订成册的登记制度。在自 8 世纪传承至 20 世纪的稳定政治制度与行政体系支持下,威尼斯档案规模不断扩大并在 17 世纪达到顶峰。此外,受典型的层级制行政体系影响,威尼斯在很长一段时期内掌握着全欧数量最多、来源最广泛、内容最复杂的档案资源^[24]。据统计,目前威尼斯保存下来的各类档案文件总计 40 万件左

右,大部分档案的形成时间可以追溯到 1797 年威尼斯共和国的自治权结束之前。“威尼斯时光机”项目组深刻认识到如此大规模的历史档案中可能蕴藏巨大的价值,从 2015 年开始就面向“历史大数据”(Big Data of Past)启动了一系列子项目,对历史档案向历史大数据转型开展一系列实验。总的来讲,“威尼斯时光机”在 2012—2019 年的第一阶段,主要完成了历史档案的大规模数字化扫描和转录识别,在 2020 年之后的第二阶段将重点攻克历史档案的语义关联和新知识体系的建设问题。

3.1 大规模数字化扫描

“威尼斯时光机”项目的初衷,就是整合威尼斯大规模的历史档案,重建威尼斯城市的发展传记、政治动态、建筑变迁与社区变化,通过这些文书档案的相互参照、交织与关联,共同形成一个有关威尼斯千年历史的丰富叙事。要实现这个初衷,改变历史档案实体分散、内容离散的现状,开展历史档案的大规模数字化扫描是第一步。实际上,在“威尼斯时光机”项目之前,威尼斯国家档案馆自 2006 年起就在意大利文化遗产管理局的支持下启动了历史档案的数字化扫描工作,最初选择的扫描对象是只需单幅扫描、技术难度较低的城市地图。2012 年之后,“威尼斯时光机”项目为威尼斯国家档案馆创建了一个数字化扫描生产线,能够扫描更多复杂尺寸和形式的历史档案,包括地图、家谱、手稿、乐谱等^[25],并将其中可识别的文字直接转化为数字文本以供识读和标注。到 2017 年 10 月,“威尼斯时光机”项目已经完成时间跨度 200 年的 19 万份行政文书、72 万份照片档案、3 千本专著的大规模数字化扫描工作,共形成超过 200 万张数字化图像,同时由档案工作者手工对历史档案中所含的人名、地点、关键词等进行了 16 万次的抄录并标引^[21]。卡福斯卡里大学校长 Michele Bugliesi 为此曾说:“档案馆藏的数字化为研究和理解过去与当代文明的文化历史演变开辟了新的途径。”随后,“威尼斯时光机”项目又与工业界

合作,开发了历史档案大规模扫描机器人装置,可半自动化每小时扫描 1 000 页,大幅提高了数字化扫描的效率。2020 年起,项目继续突破数字化扫描的技术问题,开始使用激光雷达扫描和地面摄影测量技术对更丰富的历史资源进行 3D 扫描工作^[21],通过将二维历史档案与三维历史资源关联呈现,进一步构建威尼斯历史数字镜像空间。

3.2 自动化文本转录识别

历史档案的大规模数字化并不意味着这些档案就能够自动被人所识读和理解,进一步的文本识别和转录是揭示历史档案内容价值的关键一步。“威尼斯时光机”项目组认为,历史档案数字化的前景在于内容层面的可索引、可检索,因此必须完成手写文本的识别与转录工作。虽然在大规模扫描阶段,档案工作者已经手工完成了大量的转录识别工作,但手工建立的内容索引不具有可伸缩性,难以充分体现历史档案之间的关联。2016 年起,“威尼斯时光机”启动“档案文本的识别与丰富”(The Recognition and Enrichment of Archival Documents, READ)子项目,专门利用模式识别、图像分析、计算机视觉、自然语言处理等技术,解决历史档案的手写文本识别、关键字识别、布局分析、自动作者识别等问题,并于 2018 年对外公布了自动化手写文档识别系统和专门用于照片档案数字化副本识别的通用文档分割处理系统^[21]。2020 年之后,项目更多使用卷积神经网络和递归中立网络等新技术架构来处理手写文本的自动抄写和转录问题,相较此前的人工转录,效率大幅提高。

3.3 图谱式关联离散数据

历史档案的大规模数字化与自动文本处理技术,为历史数据的重组和挖掘提供了新的机会。通过大量的历史关键词标注,“威尼斯时光机”搜集了海量的人名、地名、交易信息、建筑名称等历史数据,这些数据零散分布于历史档案的内容中,且语义模糊、离散、缺少一致性表达

和关联性组织。但威尼斯大规模历史档案的真正价值,就在于内容语义之间的关联性。一方面,不同档案记录的内容中可能包含同样的关键词或实体,这就使得基于统一关键词或实体的交叉检索成为可能;另一方面,以这些共同出现的关键词或实体作为节点,就可以搭建起一个关联数据图谱,不同知识单元可以通过这些节点相互关联,知识的推理也成为可能。为了解决历史档案中离散数据的关联问题,“威尼斯时光机”主要应用了语义网技术,将位于历史档案文本不同位置的同一实体进行链接,识别实体含义并以形式化语言进行一致性表达,最终能够聚合单个实体的多来源信息,甚至构建起跨文本、跨档案汇集的实体图谱。而所有从大规模历史档案中提取出来的不同信息源的实体,都会被组织在这样一个图谱式的关联数据集中,真正具备了历史大数据的网络化特征。

3.4 模型化构建新知识体系

“威尼斯时光机”构建历史大数据的最终目的,是帮助研究者和公众找寻历史问题的答案,事实证明,有关千余年来威尼斯城市的公共卫生、健康、医疗、经济、测绘等历史问题,都可以在威尼斯历史大数据中找到答案或线索。这推动着历史大数据建设更向前一步,走向历史知识库的建设,并以虚拟空间模拟重建威尼斯历史。例如,“威尼斯时光机”以威尼斯地籍地图为基底制作了威尼斯城市 3D 模型,历史学家曾根据这份地图为 1740 年的威尼斯人口普查档案标注了地理背景,而这份人口普查档案记录了威尼斯所有拥有和出租房产的公民,再结合绘画档案中的建筑信息,不同信息源的离散数据在这份地图上被关联、汇聚起来。该 3D 模型能够以地理信息为新的数据组织逻辑,动态展示当时每座建筑中活跃的企业,构建起一个超越历史档案文本的新知识体系。与此类似,“威尼斯时光机”还利用历史档案中提取的大量人名,构建了威尼斯人物社会网络,形成了以人物关

系为基础逻辑的又一个知识体系。

4 从历史档案到历史大数据的转型框架

“威尼斯时光机”十年间所探索的历史档案向历史大数据的转型路径,大体可分为数字化、数据化、语义化、知识化四个阶段。对于大多数国内档案机构来说,大规模的馆藏历史档案数字化工作已经取得显著成果,但文本尤其是手写文本的转录识别是当前面临的棘手问题,语义化和知识化尚处在智慧档案馆建设的远期规划中。《“十四五”全国档案事业发展规划》已将“加快推进对重要档案数字化成果进行文字识别和语音识别”明确列为档案资源数字转型的重要任务。但遗憾的是,历史档案数字转型的讨论大多在一馆之内,档案馆之间、档案馆与其他历史文化馆藏机构之间、历史档案汇集与其他历史资料之间的巨大关系价值尚未被充分挖掘,历史档案作为传统社会留存下来的最具价值的大规模数据汇集,对当代数字社会历史融聚与连续性理解的重要作用尚未被充分认知。

基于此,本文尝试探讨历史档案在历史大数据建设过程中的重要价值,并基于“威尼斯时光机”经验,为记录中华文明悠久历史的大规模珍贵历史档案,探索走出架阁库、走进知识海洋的可行路径。

4.1 总体框架

正如卡普兰教授所说,历史档案向历史大数据的转变,真正完成了两个使命:回忆(recollection)和重置(remapping)^[2],分别是以当代的数据思维与逻辑理解历史档案和重建历史数据。如图 1 所示,为完成这两大使命,历史档案需要依托计算手段,经历数字化、数据化、语义化、知识化四大阶段,分别实现版本转换、语法转换、语义转换和语用转换四大变化,需要载体控制、表达规范、关系模型和翻译机制四大工具。

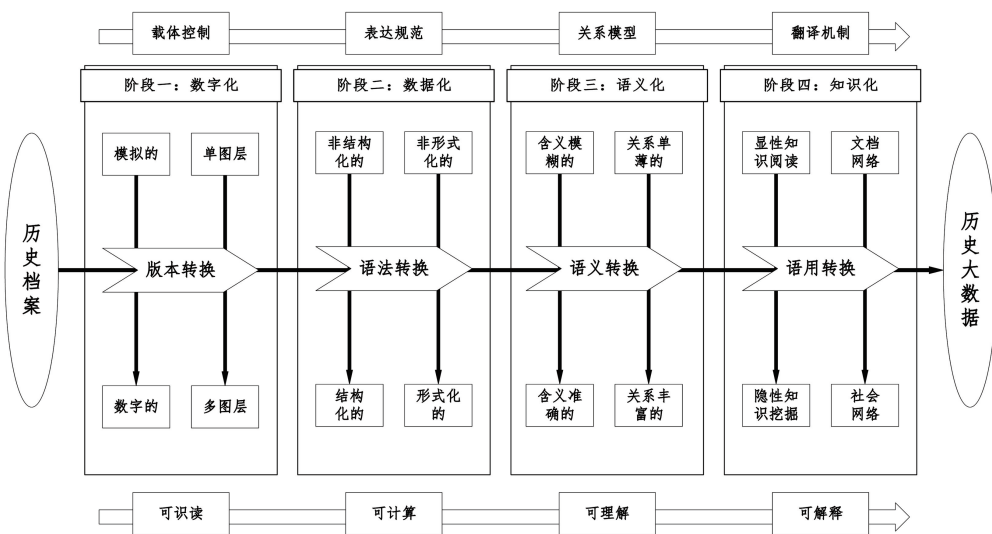


图1 从历史档案到历史大数据的转型框架

(1) 数字化阶段

大规模数字化阶段主要解决历史档案版本状态不适于机器识读的问题。这一阶段的核心任务是将历史档案从模拟状态批量转化为数字状态,即从纸质或实体版本转化为具有一定格式的数字版本。在这一过程中,载体控制即选择什么样的存储体来更好地完成数字版本历史档案的长期保存是一个关键性问题,而历史档案的内容挖掘与增值并非本阶段的核心任务。但这并不意味着数字化阶段不需开展内容层面的工作,将模拟态档案从只包含档案内容文本层的单一图层页面,转换成封装有文本识别数据、元数据、标注数据和标签数据的多图层页面,可为未来的数据处理和分析工作奠定良好基础。总的来讲,数字化阶段对历史档案的要求是可识读,即从面向人工识读转为面向机器识读,实现历史档案生命周期从模拟环境向数字环境的延伸。

(2) 数据化阶段

在大规模数字化基础之上,数据化阶段主要解决历史档案数据结构不适于机器计算的问题。这一阶段的核心任务是将历史档案的语法结构从自然语言转换成机器语言。一方面将历

史档案从非结构化数据转化成结构化数据,突破基于传统叙事逻辑的历史档案数据组织模式,重新对数据结构进行定义;另一方面在被明确定义的数据结构基础上,将历史档案的文本识别数据、元数据、标注数据和标签数据等重新用形式化语言表达,使其具有计算能力。在这一阶段,历史档案将突破长期保存的载体和版本限制,借助自然语言处理、人工智能等手段,以更灵活的数据形态追求更强大的计算能力,以实现未来更多的内容分析与挖掘。总的来讲,数据化阶段对历史档案的要求是可计算,即从面向人工处理转为面向机器处理,进一步增强历史档案对计算环境的适应能力。

(3) 语义化阶段

在具备计算能力的数据库上,语义化阶段主要解决历史档案内容表达不适于机器理解的问题。这一阶段的核心任务是将历史档案的语义组织从隐式组织转换成形式化组织。一方面,要明确定义历史档案数据中包含的概念与实体的含义,建立相同概念、相同实体的准确映射关系;另一方面,要多维度建立历史档案数据中包含的概念与实体的关系,建立不同概念、不同实体的合理相关关系。在准确含义与丰富关

系的基础上,建立历史档案内容、背景与结构中可能存在的概念体系,并将其形式化地表示为本体或图谱。在这一阶段,历史档案将借助语义网络等手段重新理解、组织并建构新的语义体系,“回忆”数据在历史档案形成时的含义与关系,并在当代逻辑下“重置”这种含义与关系。总的来讲,语义化阶段对历史档案的要求是可理解,即从面向人工理解转为面向机器理解,实现历史档案在数据状态下的内容增值。

(4) 知识化阶段

在明确的语义基础上,知识化主要解决历史档案语用情境不适于当代叙事的问题。这一阶段,对于历史档案内容的处理已告一段落,将更多依靠历史档案的背景信息,在特定的时空背景与叙事逻辑下,理解蕴藏在历史档案文本内容中的更宏大的历史叙事。也就是说,知识化除了对历史档案文本内容记录的显性知识进行挖掘外,还会对历史档案形成时空中存在的人、地、时、事、物等隐性历史知识进行挖掘,以充分理解并解释历史档案的形成在历史情境中所发挥的真正作用。简单地只分析某一历史档案汇集,或者只分析历史档案,都不足以充分挖掘这样的隐性知识,并与其他历史资源建立联系。只有突破文档网络,在一定时空序列中建立多源异构历史数据之间的更广泛的网络,才能够实现从历史档案向历史大数据转型的最终一步,而这一步将决定历史档案能否真正参与到当代数据逻辑指导下的社会网络建构中。总的来讲,知识化阶段对历史档案的要求是可解释,即以当代叙事逻辑解释历史档案形成时的叙事逻辑,以当代数据组织逻辑解释历史档案过往被整理时的组织逻辑,实现历史档案在不同时空之间的价值转换。

4.2 转型关键

如图1所示,为了顺利完成从历史档案到历史大数据转型的数字化、数据化、语义化和知识化,每个阶段都需要一些关键的方法与工具,例如包括载体控制在内的长期保存方法,包括

表达规范在内的数据结构转换方法,包括关系模型在内的语义组织工具,包括翻译体制在内的价值实现工具等。关于历史档案数字化的长期保存问题,我国档案领域已探讨多年,本文不再赘述,下面主要对后三个转型关键问题进行阐释。

(1) 表达规范

不同历史时期的档案有着不同的物理载体和内容结构,但因同一时期历史档案的数据表示都遵循一定的规范,后代能够识读和理解这些数据。例如,清代皇家档案中,不同级别的文书写作体例必须遵循一定的规范,掌握这种规范不仅有利于理解历史档案的内容,还有利于将历史档案的旧体例映射、转化为适应当代读者思维的新体例。这种旧规范与新规范之间的转化,就意味着历史数据被以新的组织方式阐释出新的关系,即历史被重新解读或再现的过程。例如,原本分散在墓志铭、家谱等不同载体上每一人物的信息,被聚集、重新书写成人物传记的形式,这就是使用某种规则再现或再生产历史档案的过程,也是“规范表示”的过程^[26]。历史档案管理过程中所创造的各种用于检索或查询的工具,例如索引、目录、词表、谱系图等,都是规范表示的产物。然而在漫长的人工管理与阅读环境中,这种规范表示的结果都是人可阅读的,随着机器成为新的阅读通道,历史档案规范表示的结果也将趋于机器可读。

在“威尼斯时光机”项目中,地图以及地图上的地理信息,是标注和建立历史数据之间关系的重要基底。项目将访问、输入、输出、缩放、旋转、分层等操作都内化为地图的使用规则,使原本无法进行这些操作的纸质地图被重新规范表示,转化为数字地图,而数字地图会进一步聚集更多来自其他载体的数据,例如在地图上标注的人物信息、建筑信息、贸易事件信息等,这些信息都以地理信息为标准被规范化重组。由此可见,在历史档案的数据化过程中,数据表示规范的制定至关重要,即各类零散、分散的历史档案数据,应以何种规范、在何种空间中被重

组,以聚合成新的统一的系统。实际上,在历史档案传统管理工作中的编研环节,长久以来都在遵循先制定数据表示规范再重组历史档案信息的习惯,即常见的编研体例,如大事记、组织机构沿革等,都是一种数据表示规范,只不过这种规范是依赖人工阅读习惯而形成的,往往是单一维度的。在历史档案数据化时代,数据表示规范将更加多样、面向人工和机器共同阅读、协同处理,能够支持更加多维度、跳转式、网络化的历史档案信息重组和表达。

(2) 关系模型

经过规范化表示后,历史档案就有了建构模型的基础,而历史档案的语义化也需要遵循一定的模型开展。这里的模型主要处理历史档案的各种关系,在不同的历史领域或主题下,模型具有不同的形式,遵循不同的逻辑,可推算出不同的新知识。

最常见的是时间模型,可用于推算历史档案中的时间节点、大事顺序与因果关系等。例如,将气候气象数据、地理数据、历法数据等统一在时间模型中,可以帮助考证某些历史事件发生的时间、顺序与触发关系。其次是基因模型,可用于推算历史档案中的人物、事件、实物等实体之间是否具有“基因”关系,以及这种产生自历史档案所记录的历史事件本身的基因,为历史档案中记录的人物、事件、实物等实体赋予相关、相等与不相关关系等。其中,继承关系是基因模型中最能推断出来的关系,例如家谱中的人物谱系等。历史档案主题词表、索引等工具某种程度上都是一种基因模型。再者是事件模型,可用于推算出历史档案所记录的各类活动或现象的过程,例如手工艺生产、商业贸易、居民迁徙等。在不同关系模型中,为概念和实体命名都是非常基础且重要的事情。在很长一段时间里,寻找高效且适应性强的命名方法来描述看似无限的生物多样性,是大数据方法在自然领域应用的基石。档案领域的著录、标注等描述体系也都是重要的概念和实体命名系统,这些系统本身就是一种逻辑体系,可以通过

形式化语言转化成最适合机器描述和机器计算的模型。

(3) 翻译机制

历史大数据是用当代互联网环境下的大数据思维来重组历史信息,可能汇聚不同历史节点上的历史数据,但每个历史节点上的档案都只具有有限的“时间视野”,在缺少翻译机制的情况下,这种有限的时间视野使得超越时空局限的跨时代、延续性思考更加困难。当代的逻辑和思维有时甚至会误读、错解历史信息,而使得知识的推断更加困难^[27]。为了扩大这种“时间的视野”,一方面需要聚集更多的数据,由点到线地汇聚更多的时间点以形成时间流、数据流,并且保持这种时间流、数据流的开放性,以吸收和容纳更多的时间与数据;另一方面就要设计并形成有效的翻译机制,即使用符合当代逻辑的规范去“重新记录过去”。这种翻译机制并非简单的语言或文字上的翻译,而是叙事逻辑的翻译。

5 总结

历史档案转化为历史大数据的巨大潜力已经在“威尼斯时光机”项目中充分显露,当前越来越多的国家和组织意识到历史大数据建设与开发的重要性。2020年,“威尼斯时光机”走向更高阶段,升级为“时光机:欧洲大陆通史”(Time Machine: A Common History for the Continent)项目,其宗旨就是提取和利用历史大数据,或者说“释放历史大数据”。作为“威尼斯时光机”的第二阶段,“欧洲时光机”将发展为一个超大型的城市网络,项目将更多使用数字化和人工智能技术,来充分挖掘欧洲丰富的文化遗产,将大量碎片数据转化为高价值的产业知识,从本质上讲是通过大规模计算和数字基础设施建设来模拟欧洲整个社会、文化与地理的演变。“威尼斯时光机”的勃勃雄心,是建立在从历史档案向历史大数据转型基础上的。

我国拥有世界上规模最大、数量最多、记录

时间最长的历史档案资源,其转型为历史大数据的潜力是巨大的。威尼斯国家档案馆在漫长时间里经历的历史档案价值展示与实现的困境,如今我国诸多档案馆也正在经历;“威尼斯时光机”项目所提供的从数字化到数据化、再到语义化、知识化的转型路径,对于我国历史档案

馆藏机构是具有可复用价值的。但我们还需要在数据表达的规范、知识生产的模型以及叙事翻译的体制上,找寻适合中国故事的模板。我们期待历史档案走向历史大数据,以新的数据思维爆发新的力量。

参考文献

- [1] Kaplan F. A map for big data research in digital humanities[J]. *Frontiers in Digital Humanities*, 2015(2): 1-7.
- [2] Kaplan F, di Lenardo I. Big data of the past[J]. *Frontiers in Digital Humanities*, 2017(4): 1-12.
- [3] 陈婧. 威尼斯的时光机[J]. *IT 经理世界*, 2016(18): 76-80. (Chen J. Venice time machine[J]. *CEOIO China*, 2016(18): 76-80.)
- [4] 董聪颖. 穿梭千年: 数字人文对档案信息资源开发利用的影响[J]. *档案管理*, 2018(2): 11-14. (Dong C Y. Through the millennium: digital humanities' influences on development and utilization of archival information resources[J]. *Archives Management*, 2018(2): 11-14.)
- [5] 杨茜茜. 数字人文视野下的历史档案资源整理与开发路径探析——兼论档案管理中的历史主义与逻辑主义思想[J]. *档案学通讯*, 2019(2): 17-22. (Yang Q Q. The historicism and logicism in archival management: paths of arranging and exploring historical archival resources in perspective of digital humanities[J]. *Archives Science Bulletin*, 2019(2): 17-22.)
- [6] 王晓光, 梁梦丽, 侯西龙, 等. 文化遗产智能计算的肇始与趋势——欧洲时光机案例分析[J]. *中国图书馆学报*, 2022, 48(1): 62-76. (Wang X G, Liang M L, Hou X L, et al. A preliminary exploration on intelligent computing of culture heritage: a case study of European Time Machine project[J]. *Journal of Library Science in China*, 2022, 48(1): 62-76.)
- [7] 陈志武. 开启“历史大数据时代”[EB/OL]. (2018-09-06)[2021-07-11]. <https://www.huxiu.com/article/261358.html>. (Chen Z W. Open the “historical big data era”[EB/OL]. (2018-09-06)[2021-07-11]. <https://www.huxiu.com/article/261358.html>.)
- [8] 梁晨, 董浩, 李中清. 量化数据库与历史研究[M]//付海晏, 徐剑. 大数据与中国历史研究: 第1辑. 北京: 社会科学文献出版社, 2017: 34. (Liang C, Dong H, Li Z Q. Quantitative database and historical research[M]//Fu H Y, Xu J. Big data and the study of Chinese history: Vol. 1. Beijing: Social Sciences Academic Press, 2017: 34.)
- [9] Davies K, White M. Breakthrough: the race to find the breast cancer gene[J]. *Nature Medicine*, 1996(2): 241-242.
- [10] Goldin C, Katz L F. The race between education and technology[M]. Cambridge: Belknap Press, 2008.
- [11] China multi-generational panel datasets series[EB/OL]. [2022-01-19]. <https://www.icpsr.umich.edu/web/ICPSR/series/265>.
- [12] 中国历代人物传记资料库(CBDB)[EB/OL]. [2021-07-21]. <https://projects.iq.harvard.edu/chinese-cb-db>. (China biographical database project(CBDB)[EB/OL]. [2021-07-21]. <https://projects.iq.harvard.edu/cbdb>.)
- [13] 中国地方历史文献数据库[EB/OL]. [2021-07-21]. <http://ndfwx.datahistory.cn>. (Database of Chinese lo-

- cal historical documents[EB/OL]. [2021-07-21]. <http://ndfwx.datahistory.cn>.)
- [14] 国家社科基金重大项目“清代灾荒纪念暨信息集成数据库建设”开题报告会举行[EB/OL]. (2014-03-13)[2021-07-21]. <https://news.ruc.edu.cn/archives/75504>. (The opening report of the major project of the National Social Science Fund “disaster memorial and information integrated database construction in Qing Dynasty” was held[EB/OL]. (2014-03-13)[2021-07-21]. <https://news.ruc.edu.cn/archives/75504>.)
- [15] Michel J B, Shen Y K, Aiden A P, et al. Quantitative analysis of culture using millions of digital books[J]. Science, 2011(331):176-182.
- [16] 夏翠娟. 面向人文研究的“数据基础设施”建设——试论图书馆学对数字人文的方法论贡献[J]. 中国图书馆学报, 2021, 47(3):24-37. (Xia C J. The construction of “data infrastructure” for humanities research: the methodological contribution of library science to digital humanities[J]. Journal of Library Science in China, 2021, 47(3):24-37.)
- [17] 肖恩·格雷厄姆, 伊恩·米利根, 斯科特·魏因加特梁. 探索历史大数据: 历史学家的宏观视角[M]. 梁君英, 刘益光, 黄星源, 译. 浙江: 浙江大学出版社, 2019: iv. (Graham S, Milligan I, Weingart S. Exploring big historical data: the historian's macroscope[M]. London: Imperial College Press, 2015: iv.)
- [18] Loader B D, Dutton W H. A decade in Internet time: the dynamics of the Internet and society[J]. Information, Communication and Society, 2012(15):609-615.
- [19] Duranti L. The archival bond[J]. Archives and Museum Informatics, 1997(11):213-218.
- [20] Dayalan M. MapReduce: simplified data processing on large clusters[J]. Commun. ACM, 2008, 51(1):107-113.
- [21] Chronology[EB/OL]. [2021-08-25]. <https://www.epfl.ch/research/domains/venice-time-machine/chronology>.
- [22] Colavizza G, Ehrmann M, Rochat Y. A method for record linkage with sparse historical data[C]// Digital Humanities Conference 2016. Krakow, Poland, 2016.
- [23] Mayer-Schonberger V, Kenneth C. Big data: a revolution that will transform how we live, work, and think[J]. American Journal of Epidemiology, 2013, 179(9):1143-1144.
- [24] 王岑曦. 来源原则历史源流新探[D]. 南京: 南京大学, 2018. (Wang C X. A new exploration to formation of principle of provenance[D]. Nanjing: Nanjing University, 2018.)
- [25] Abbott A. Venice get a time machine[J]. Nature, 2017(6):341-344.
- [26] Kaplan F. How books will become machines[J]. Lire Demain: Des Manuscrits Antiques à L'ère Digitale, 2012: 27-44.
- [27] Brand S. The clock of the long now: time and responsibility[M]. New York: Basic Books, 1999.

祁天娇 中国人民大学信息资源管理学院师资博士后, 讲师; 中国人民大学数字人文研究中心、电子文件管理研究中心、档案事业发展研究中心研究员。

(收稿日期: 2022-02-07)