

数字人文视野下专题历史文献资源 数字共享平台构建探究^{*} ——以西南联大历史文献为例

陈海玉¹ 王聪² 陈雨¹ 杨娴¹ 何茜林¹

(1. 云南大学历史与档案学院, 昆明 650091; 2. 广西民族大学相思湖学院, 南宁 530031)

摘要: 传统的历史文献专题研究多通过人工搜集海量的文献资源, 从中挖掘、分析、发现新的历史信息, 耗时耗力。随着数字人文研究的兴起, 历史文献研究的理念与方法得到新的突破。中华民族历史悠久, 形成的各类专题历史文献资源规模庞大、保管分散, 目前亟需构建具有统一数据管理模式的专题历史文献资源数字共享平台, 实现各专题历史文献资源可视化、语义信息关联化、内容浏览便捷化, 以满足数字时代历史研究的新需要。研究借鉴数字人文研究理念与技术方法, 阐述专题历史文献资源数字共享平台的建设目标、建设模式、建设框架及可行路径, 并以“数字联大”共享平台建设实践为例, 开展西南联大历史文献资源的数据采集处理、文献本体信息网络构建及文献数字平台知识服务, 实现专题历史文献资源从静态信息到动态知识的关联和可视化呈现, 满足平台用户深度学习和历史研究的需要。

关键词: 历史文献; 数字人文; 西南联大; 共享服务平台

中图分类号: G255 DOI: 10.3772/j.issn.1673-2286.2024.04.003

引文格式: 陈海玉, 王聪, 陈雨, 等. 数字人文视野下专题历史文献资源数字共享平台构建探究: 以西南联大历史文献为例[J]. 数字图书馆论坛, 2024, 20(4): 22-32.

近年来随着数字人文与历史文献研究的深度融合, 利用现代信息技术推进历史文献研究的潜能与优势彰显, 主要体现在以下几个方面。①历史文献研究迎来便利与创新。通过数据库建设实现分散、海量、碎片式历史资源的数字化集中保存与有机缀合, 使学者从海量庞杂的信息中解放出来, 促进了高效资源整合与检索利用, 益于学者的知识创新与智慧发现。②通过文本挖掘等数字技术“来重新定义、提出以及解答人文问题”^[1], 助力历史研究者和受众发挥主观想象力, 加强历史文本研究中的人文关怀。③推进历史文献研究的

纵深方向拓展。数字人文方法不仅可以有效提取文本要素并揭示各个要素之间的内在联系, 也可进一步分析内部要素与外部环境之间的有机联系。如利用地理信息系统 (Geographic Information System, GIS) 技术与虚拟现实 (Virtual Reality, VR) 技术可完成一些动态追踪与多要素综合研究, 实现历史文献内部关联, 辅助学者进行研究创新^[2]。可见, 数字人文为历史文献研究与历史资源服务利用提供了较清晰的坐标和更明确的发展方向。

大数据时代, 图书馆、博物馆、档案馆等文化机构

收稿日期: 2024-01-14

^{*}本研究得到国家自然科学基金项目“中共云南地方组织沿革档案资料整理及专题数据库建设 (1926—1950)” (编号: 22BTQ084)、国家自然科学基金重大项目“英国藏中印边境问题相关档案整理与研究” (编号: 23&ZD263)、教育部重大招标项目“元明清时期中国边疆治理文献整理与数据库建设研究” (编号: 21JZD042) 资助。

储藏的历史文献资源海量增长,其载体表征多样,分散异构,破坏了历史文献群的时空连续以及整体关联,造成信息内容碎片化、信息孤岛问题突出。文献资源的利用率普遍不高,影响了资源价值的发挥。此外,近年来各机构虽对馆藏实体文献进行数字化处理,开展了多模态资源整合加工,建立了历史文献资料库和多媒体库,促进了用户访问和获取信息资源,但其对文献资源的内容揭示不足,缺乏深层次描述,知识聚类、文本挖掘和语义关联还较少,特别在基于历史文献本身还原历史面貌、展现文献蕴含的完整信息方面还有很大发展空间。

数字时代,在历史文献挖掘利用方面,无论是传统实体资源的数字化转化,还是网络原生数字资源的保存开发,都应当强调知识共享的重要功能,并将其作为支持学术研究、开展教学育人和生成其他形式知识的重要信息来源。如“美国记忆”(American Memory)项目通过记录大量美国数字化历史文献的同名网站,将这些载有美国历史的数字记录作为“教育和终身学习的资源”供公众使用,履行“维持、保存知识和创造力的使命”^[3]。又如欧洲“时光机项目”(Time Machine Project),它不仅是世界文化遗产保护的重要成果,又是面向公众开放的互动性强、细节性高的历史教育工具^[4]。基于此,笔者认为探讨数字人文视野下历史文献资源数字共享平台的构建十分必要,通过数字人文共享平台能逐步实现各类研究模式、技术方法的聚合,消除传统历史文献研究中的资源数据获取障碍,并借助数字人文技术解决程序化研究工作和个人无法完成的研究任务,这对于历史文献研究问题的提出和问题的分析具有至关重要的作用。

1 相关研究与实践

(1) 历史文献数字整理方法。数字人文改变了历史文献整理格局,促进整理范式朝着数字整理转型,这为各类专题历史文献数据库建设和知识信息服务系统搭建提供支撑。刘石等^[5]提出基于大数据技术,用算法和模型来发现和组织知识,从海量数据中发现隐藏在传统文献中的知识、模型、关系、趋势与规律。梁继红^[6]探索构建了“文本阅读—元数据搭建—文本内数据化—数据分析和可视化”的历史文献数字化整理体系。邓君等^[7]以抗战老兵口述资料中心数据为数据源,采用

自顶向下方法构建抗战老兵口述历史档案资源知识图谱,即通过模式层构建、数据层组织和知识图谱可视化展示,实现口述历史档案资源多维知识发现。上海交通大学图书馆通过文献关联聚类、信息自动获取与空间可视化等数字化手段开展对33万件地方历史文献的数字化加工整理,在历史地理信息获取和标准化方面取得了显著成效^[8]。

(2) 历史文献数字项目实践。①借鉴数字人文理念对传统历史文献进行数字化加工和数据化处理,建设“以数字资源为基础、以数字人文技术为支撑、以内容关联和知识服务为特征的新型数据库”^[2]。其代表有国家图书馆建设的“革命文献与民国时期文献联合目录”^[9]、上海图书馆的“名人手稿档案关联开放数据集”^[10]、中国第一历史档案馆的“满文档案图像识别软件系统”^[11]、上海交通大学出版社的“中国地方文献数据库”^[12]。②将数字人文技术理念应用于历史文献资源的实践项目,其功能主要集中在以下维度:重视历史文献资源整合共享,如由国家图书馆构建的历史文化资源共享平台“中国记忆项目实验网站”^[13],该平台是多载体、多种类专题历史档案文献的集合;注重对历史文献资源进行细粒度治理,如“《古兰经》语料库”^[14]深入历史文献文本内容,形成独特的《古兰经》音译数据库和符合国际规范的手稿目录;开展对历史文献价值的挖掘,如“中国历代人物传记资料库”^[15]、“历代宝案脉络分析系统”^[16]等对史料中大量分散的人物文本数据进行结构化标引,形成大规模数据集合,并运用文本挖掘技术、GIS技术及关联数据技术等获取和收集文献内容信息和外部特征,在历史文献数据的基础上开展社交网络分析,以此全面呈现人物脉络;凸显可视化服务,如上海图书馆的“中国家谱知识服务平台”^[17]将家谱数字资源可视化,展示某一姓氏在某一地理空间范围内的分布情况,荷兰国家档案馆的“记忆宫殿”项目利用虚拟技术进行馆藏历史照片的3D转化,使公众了解照片档案的多样化内容^[18]。总体来看,目前实践项目以历史地理、古籍文献类专题项目居多,仅台湾大学数位人文研究中心和武汉大学数字人文研究中心设立了数字人文平台,提供了包含新闻资讯、出版物、数字人文项目、数字人文工具和会议信息在内的多种资源服务^[19]。

数字人文赋能历史文献研究与实践已展现出巨大潜力,但目前大部分数字实践成果囿于对历史文献资料的数字化处理,具有深度揭示分析功能、具备一定开放数据性质的数字人文数据库凤毛麟角^[20]。区别于以往

检索型的静态信息贮藏资源库, 本文从数据处理、信息组织、知识关联等方面探究动态知识关联分享型专题历史文献资源数字共享平台的开发与路径, 构建以人为本的知识聚合服务平台。

2 平台构建路径

当前数字人文领域研究兴起, 为各类历史文献资源的重新整理、组织、利用与问题研究提供了计算思维。专题历史文献资源数字共享平台应借助数字人文方法和技术, 建立健全各专题历史文献资源研究体系, 推进专题文献资源跨机构、跨区域、跨学科交流和研究, 并在现阶段数字化基础上, 探索深层利用和有序开放共享数字资源的新模式, 逐步完善服务功能, 实现历史文献资源在线开发利用一体化, 从而最大限度地提

升历史文献资源的服务效能、满足历史文献资源保护需求。

2.1 建设规划

2.1.1 明确建设目标

“专题历史文献资源数字共享平台”建设目标规划如图1所示。最终目标是在数字技术赋能下构建能够实现知识聚合查询的专题历史文献数字共享平台, 以用户个性化需求为导向, 创建可视化的历史文献数字应用场景。借助知识本体理论和数字工具, 通过构建专题历史文献资源领域本体, 进行数据组织和知识关联, 发现历史文献中蕴藏的隐性信息与新型问题, 在数字时代进一步讨论历史文献资源的知识价值和研究意义。

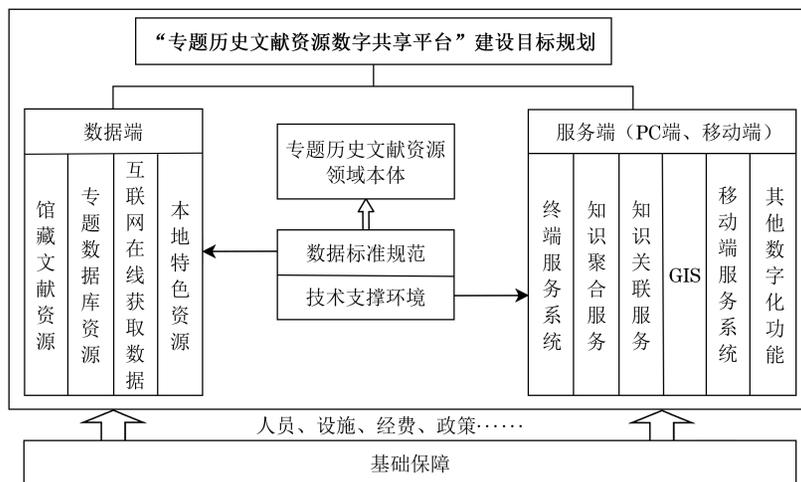


图1 “专题历史文献资源数字共享平台”建设目标规划

2.1.2 选择建设模式

各目标专题的历史文献资源数字共享平台建设模式需要根据所选择专题文献资源的管理现状来确定, 综合考虑历史文献资源的整理、保存、分布、管理等情况, 以及纸质文献数字化进展等。如调研目标历史文献资源主要收藏机构设置及相应文献管理状况, 分析保存主体是否包括各类型图书馆、档案馆、文博馆、纪念馆及民间社会组织与个人。另外, 需明晰历史文献来源类型, 如书籍、报刊、档案、手稿、日记、书信、图片、证书, 以及相关的音视频资料等。通过对目标专题历史文献资源的管理现状进行系统分析, 明确其

保存情况、数字化规模、数据库建设现状等, 从而选择适宜的平台建设模式。根据建设主体及建设目标, 现有历史文献资源数字共享平台的共建模式大致可概括为政府主导模式^[21]、文化机构合并模式^[22]、跨机构合作模式^[23]和共同项目驱动模式。基于各保存主体所藏文献资源实际情况及文献当代价值实现需求, 选择契合的文献资源共享平台建设模式, 明确各建设主体及建设目标, 将各主体优势进行系统整合, 通过发挥优势特长进行资源互补共享, 从而合作完成平台共建, 最终实现目标专题历史文献资源数字共享平台建设规划目标。

2.2 平台建设框架

专题历史文献资源数字共享平台建设框架如图2所示,该框架自下而上可分为资源准备阶段、资源重组阶段和资源利用阶段,对应的主要建设内容依次是信息存储与描述、数据加工与关联、知识管理与服务。参考计算机科学研究领域的细粒度模型^[24],对目标专题历史文献资源进行逐级分解重组,旨在将繁杂的专题历史文献资源构造成便于访问的细粒度知识关联网络。由此,提出的专题历史文献资源数字共享平台建设框架可简述为:将目标专题历史文献资源庞杂的模块信息分解为碎片化的离散数据,最终联结成逻辑性的关联知识,实现目标专题历史文献资源从静态信息到动态知识的流动发展,满足用户的深度学习需求。

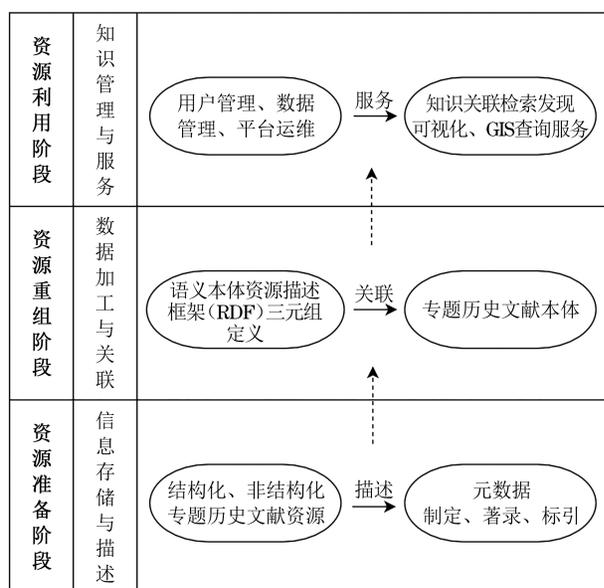


图2 “专题历史文献资源数字共享平台”建设框架

2.2.1 资源准备阶段: 信息存储与描述

专题历史文献资源数字共享平台建设的资源准备阶段主要实现专题历史文献资源的收集存储与内容描述。

(1) 专题历史文献信息存储。目标专题历史文献资源的收集存储是数字人文平台建设的基础工作,为平台建设提供各类文献资源。一方面,就各专题历史文献资源的来源格式而言,需收集存储结构化和非结构化信息资源数据。另一方面,就各专题历史文献资源的来

源地址而言,可以考虑将与目标专题历史文献资源相关的其他专题数据库设为主要资料收集源,以其他线上和线下的有关专题资料为补充资料来源。

(2) 专题历史文献信息描述。前期收集存储的目标专题历史文献资源表现类型多样且数据结构各异,又因来源地址的不同而有不同的基础数据属性描述。专题历史文献信息描述以构建元数据基础框架为主,旨在标注资源属性,统一文献内容信息的表达方式与资源结构标准,为下一阶段的资源重组提供数据支持。元数据设计是专题历史文献数据加工和知识关联的重要工作,也是文献资源细粒度描述的关键。需要考虑资源描述与知识表达的完整性与合理性,最大限度满足全部数据的描述需求,充分表示专题历史文献资源内容知识,以利于平台长期建设中的知识组织、知识关联和知识管理。

传统的文献资源属性著录集中于主题、时间、载体形式等外部特征,对资源内容信息的关注不足,主要用于文件管理与主题检索等粗粒度信息利用。而文献资源的语义知识本体模型以细粒度知识为描述对象,标识关联知识的概念和领域,能够揭示资源信息内容特征,进一步解释文献资源所蕴含的历史信息和隐含知识^[25]。知识本体亦是描述数据的数据,仍属于元数据概念范畴,但在描述对象、编码方式、标准规范等方面区别于传统元数据。

专题历史文献信息描述工作通过基础元数据描述资源外部特征,通过语义本体描述资源内部知识。下文描述的“数字联大”平台为例,其元数据基本架构如表1所示,囊括了西南联大历史文献资源内容所涉及的事件、组织、人物等不同本体概念类,各概念类中包含诸多子类,由不同子类指向同一对象实体而产生知识关联,进而构建出完整的西南联大历史文献信息描述体系。在资源描述和知识表示过程中,根据专题历史文献资源具体的外部特征和内部知识属性对元数据架构进行相应调整与扩展,若某一历史事件涉及人物众多、关系链条复杂,可相应在人物概念类下增加角色子类,以此厘清人物关系网络,此类资源信息描述过程中出现的相应概念扩展有利于梳理知识单元联系和细化资源知识粒度,提高目标专题历史文献资源知识关联和逻辑推理的可行性与高效性。

表1 “数字联大”平台元数据基本架构示例

外部特征		内部知识	
元数据	元数据扩展	一级类	二级类
题目 (Title)	系列 (Series)	时间 (Time)	时间点 (Point)
	环境 (Condition)		时间段 (Period)
日期 (Date)	归档日期 (Archive Date)	地点 (Location)	地理概念 (Geographical Concept)
	数字化日期 (Digitized Date)		行政区域 (Administrative Regions)
责任者 (Author)	档案保存单位 (Storage Unit)		军事区域 (Military Area)
	档案责任部门 (Archives Department)		政治区域 (Political Area)
来源 (Origin)	档案馆 (Archives)	人物 (People)	政治领域人物 (Political Figures)
	图书馆 (Library)		经济领域人物 (Economic Field Figures)
	博物馆 (Museum)		社会领域人物 (Social Figures)
	民间组织 (Civil Organization)		教育领域人物 (Persons in Education)
	个人 (Individual)		国际人物 (International Figures)
原始载体 (Original Carrier)	纸质载体 (Paper Carrier)	组织 (Organization)	经济组织 (Economic Organization)
	其他载体 (Other Carriers)		政治组织 (Political Organization)
关键词 (Key Words)			军事组织 (Military Organization)
著录者 (Describer)			国际组织 (International Organization)
体裁 (Genre)			社会群体 (Social Communities)
语种 (Language)	中文 (Chinese)		资源 (Resources)
	英语 (English)	通报宣言 (Notification Declaration)	
	日语 (Japanese)	报刊杂志 (Newspapers and Magazines)	
	其他语种 (Other Languages)	电报书信 (Telegram Letter)	
国家 (Country)	中国 (China)	事件 (Event)	政治事件 (Political Event)
	国外 (Abroad)		军事事件 (Military Event)
内容 (Content)	内容提要 (Summary)		社会活动 (Social Activity)
备注 (Remarks)			自然事件 (Natural Event)

2.2.2 资源重组阶段: 数据加工与关联

(1) 专题历史文献资源数据加工。数据加工是将专题历史文献资源由元数据描述转化为知识领域本体的桥梁。现有各大专题历史文献资源线上平台多展示所整理的相关文献的数字化图文信息,如广西师范大学出版社集团开发的“徽州文书数据库”可进行简单检索,展示所存徽州文书图片,为每幅图片添加题名、时间、区域等主题词,通过主题词进行简单归类。此类专题数据库采用不同的元数据体系对资源进行定义和描述,在对这些专题数据库资源进行整合重组时,需要解决元数据描述格式不同的问题。在现有数字化成果基础上,运用语义本体技术,以RDF三元组形式对各数据库资源进行定义和描述,即采用本体标准描述语言对需建设的结构化资源进行重新表达,从而完成数据加工。加工形成的专题历史文献资源RDF三元组描述框

架数据,可以通过资源语义关联和映射实现不同语义的RDF元数据关联与聚合。对于平台后期新增的专题历史文献资源,可借助文本分析、内容分析、分词、提取特征词条等数字技术进行关键词或属性的提取,通过元数据标准体系描述,并使用本体模型的标准描述语言进行标注,提取知识单元,形成规格统一的RDF数据存储格式。

(2) 专题历史文献资源数据关联。目标专题历史文献资源数据关联模式可根据平台建设目标、文献主题以及利用价值等来确定,如专题历史文献资源本体模型、文献资源知识图谱等。其中专题历史文献资源本体模型是数字共享平台的建设核心,是平台为用户提供知识关联、可视化检索、GIS服务的基础。通过为专题历史文献资源中的各概念类添加关系属性构建出资源本体模型,进而将资源实体数据导入本体模型,实现专题历史文献资源数据关联。基于结合传统元数据和语义本

体技术的专题历史文献资源数字共享平台,利用资源本体的语义描述能力表现专题历史文献资源中包含的诸多人物关系和历史事实,以此进行领域知识关联研究。基于逻辑推理的专题历史文献资源数字共享平台还可以根据用户需求进行高级知识关联,用户可关联查询历史事件前后一段时间的背景知识,以此获得更加丰富的信息,并产生更为有效的理解。基于专题历史文献资源基础数据,还可将文献资料中的历史建筑视为特殊资源实体进行知识描述与关联,如“威尼斯时光机”项目以威尼斯国家档案馆为主要资源对象,在同一地理空间展示不同年代的城市场景,利用数字化的手段再现威尼斯的历史与文化^[26]。基于此,专题历史文献资源本体除了构建人、事、物等的关系网络外,还需关联专题历史文献资源内容涉及的建筑物历史空间信息,借助GIS组件全方位展现目标专题文献历史空间层面的知识。

2.2.3 资源利用阶段: 知识管理与服务

(1) 专题历史文献知识管理。专题历史文献资源数字共享平台知识管理是链接数据关联与知识服务的中间模块,也是目标专题历史文献资源数字化建设参与主体最为关注的层面。专题历史文献知识管理需从历史文献本体构建出发,考虑平台知识服务的具体要求,为专题历史文献资源数字共享平台的智慧应用创造可供部署的相关技术支撑环境。基于数字共享平台的开放性原则,除了平台牵头建设机构和主要共建机构以外,其他专题历史文献资源保存主体或个人可以通过平台对外提供的统一接口,将自己创建的第三方应用链接到平台环境中,平台为用户提供良好的网络、资源和系统支持,以吸引更多专业用户参与专题历史文献资源数字共享平台的长期建设,不断更新创建个性化、专门化、趣味化的第三方专题应用模块,最终创建全民参与的数字化知识管理体系。

(2) 专题历史文献知识服务。专题历史文献知识服务是数字共享平台建设框架的顶层应用,也是平台智慧利用功能的效果展示层,为用户提供各类功能检索服务,包括知识关联检索发现、可视化和GIS查询服务。知识关联检索发现服务通过用户检索的关键词信息聚合知识,同时以此关键词为原点向外辐射知识点,遍历专题历史文献本体库,将各相关数据库资源、线上开放资源中与关键词有映射关系的知识单元聚合,并以

网状关联的可视化形式向用户展示结果。可视化服务泛指数字共享平台中以图像、表格等形式为用户直观提供资源信息的服务。GIS查询服务则基于专题历史文献资源内容中的地理空间信息,通过数据转换、GIS建模等技术处理,提供空间位置导航、辨识等应用服务。

3 “数字联大”平台建设实施方案

西南联大历史文献资源是我国历史文献遗产的重要组成部分,其蕴含的深厚历史文化内涵和革命精神有待深度开发与传承。本研究以西南联大历史文献资源为建设目标,借助数字人文理念与技术方法搭建“数字联大”平台,以此呈现专题历史文献资源数字共享平台的建设路径。

“数字联大”平台以资源数据采集处理为基础、以资源本体模型构建为核心、以专题知识服务为目标。首先需完成西南联大历史文献数据采集与处理,进而统一专题资源元数据标准体系并构建资源本体模型,确定知识间关联情况,其次组织“数字联大”平台知识聚合服务模块,实现专题知识关联检索发现与空间信息的实时访问。

3.1 数据化资源采集与处理

3.1.1 资源采集

“数字联大”平台建设中的资源采集范围不能局限于现有的各公共机构馆藏资源和西南联大历史文献专题数据库,应包括各类学术文献资源库,如中国知网、万方、维普等。另外,互联网开放性资源也是资源采集的重要领域:①保存有西南联大历史文献资源的各个网站,如云南师范大学西南联大博物馆网站;②关于西南联大历史的新闻报道,如新华网2020年1月21日发表的《习近平考察西南联大旧址》;③关于西南联大历史的各类影视作品,如描绘抗战时期西南联大师生学习生活的电影《无问西东》、以西南联大学子口述为依据制作的电影《九零后》;④艺术团体和个人创作的西南联大历史文艺作品,如云南师范大学以西南联大历史背景创作的歌曲《西南联大组歌》,包括《大师们》《联大花开》《天降大任》等曲目;⑤民间收藏的有关西南联大历史的各类文献资源。

平台建设中的资源采集需要遵循不拘一格的采集

原则,广泛收集西南联大相关文献资源,以完整还原西南联大历史全貌,为用户提供真实、详尽的知识聚合服务。

3.1.2 资源处理

“数字联大”平台资源处理流程如图3所示。在统一数据标准规范下,对现有的数字化成果和采集到的其他非结构化或结构化资源进行结构统一、文本识别、内容分析、自然语言处理、语义本体描述、数据存储等。通过对结构各异的西南联大历史文献资源进行统一规范处理,为资源本体模型构建和知识关联做好基础数据准备。“数字联大”平台建设方案文献资源处理工作重点在于语义本体描述,这是数字人文本体论思想的具象体现。RDF是专门用于描述Web资源的数据模型,采用“资源-属性-属性值”三元组描述形式来表征事实,以统一资源标识符(Uniform Resource Identifiers, URI)为资源数据标识机制基础,能够描述任何实体对象,可囊括西南联大历史文献资源中所有结构化与非结构化资源数据。在RDF数据描述中采用URI来标识与引用西南联大历史文献资源中需要被

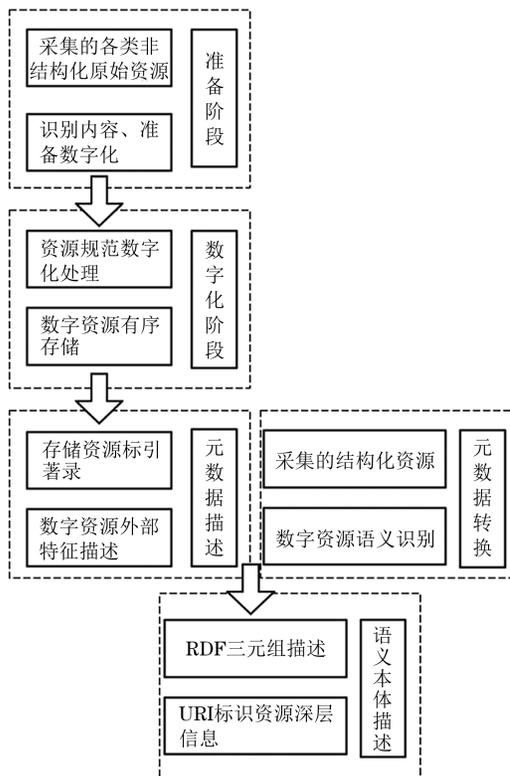


图3 “数字联大”平台资源处理流程

标识的内容,包括某个网页、某张图片、某本图书、某份文件以及具体内容语义中的概念对象等,从而涵盖西南联大历史文献资源的全部知识内容。

3.2 网络化资源本体模型构建

“数字联大”平台所构建的西南联大历史文献本体旨在描述西南联大从1938年2月19日南迁到1946年7月31日停止办学这段历史时期涉及的实体概念、术语、属性、关系等信息。其中包括西南联大历史文献资源内容涉及的人物、时间、地点、组织、事件等,以及政治、文化、经济、贸易、军事等各行业领域知识。通过构建本体,能够实现西南联大历史文献资源共享共建。同时,“数字联大”平台中基于逻辑描述的实体概念所建立的相应关联属性为线上开放资源知识框架的后期扩展打好基础。依据西南联大历史文献资源保存数据的事实特征,围绕平台后期应用的可视化关联检索目标,在借鉴本体建设骨架法(Skeletal Methodology)^[27]的基础上,设计“数字联大”资源本体构建步骤,在实际处理和操作过程中根据目标专题历史文献资源的具体内容进行适当调整。

(1) 确定资源类别与领域术语。西南联大历史文献资源来源多样,虽已对其实施统一规则的数据处理,但也需明确其来源类型,遵从历史文献的生命周期管理理念。西南联大历史文献资源涵盖文本、图片、音频、视频、地图册、实物等保存形式,内容涉及教育、战争、政治、人文、外交等历史活动。为实现较为清晰详尽的资源类型划分,施行自顶向下分级划分,在西南联大历史文献一级类目下按照资源类型划分二级类目和三级类目,如表2所示。

表2 “数字联大”平台历史文献资源类别划分体系

一级类目	二级类目	三级类目
西南联大历史文献	文本文献	西南联大历史档案
		西南联大报刊资料
		西南联大网络资源
	音像资源	西南联大图片
		西南联大音频
		西南联大视频
	地图资料	西南联大历史地图
		西南联大实时地图
	实物资源	西南联大建筑
		西南联大现存物品

此外,对西南联大历史文献资源涉及各个实体,需规范表达术语,即标签化实体术语描述。参考《中国分类主题词表》《全国革命历史档案数据采集标准》等现有词表中的各领域术语,并借鉴权威词典、经典著作中的专业事实描述,确定西南联大历史文献资源领域本体术语体系,表3为部分术语展示。

表3 “数字联大”平台历史文献资源领域术语(部分)

实体分类	术语列举
人物	姓名、字号、性别、年龄、籍贯、民族、出生日期、死亡日期
时间	年份、月、日
地点	国家、省份、市、区县、乡镇
资源	著作、作者、发表时间、发表刊物、主题、文体、载体
主题	教育、军事、文学、政治、哲学、数学
组织	师范学院、理学院、法商学院、体育部、常委会
院校	北京大学、清华大学、南开大学、云南大学
事件	讲演活动、冲突事件、主题报告、会议活动
类型	文本、图片、地图、画册、实物、音频、视频

(2) 细化概念关系与属性描述。为实现西南联大历史文献资源的知识关联发现功能,还需对各资源类

别中的各个术语建立联系,旨在将资源内容涉及的所有历史活动信息有机关联起来,组建专题知识网络。通过统一的术语来描述具体的人、事、物相关属性,具有相同属性的人、事、物便会自动建立起联系,这是知识网络建立的原始逻辑。如西南联大教授闻一多先生,本体建设中需要将“闻一多”看作一个实体对象,归为人类,他的属性有字号、性别、年龄、籍贯、出生日期等,其中籍贯又涉及地点类,出生日期涉及时间类,如此“闻一多”实体便逐步丰富起来。另外,对描述实体对象的各个属性需要制定相应的约束条件,旨在维持本体构建过程中的表述一致性,防止出现事实描述混乱和语义矛盾。为此对实体描述属性设置定义域(Domain)和值域(Range):一是表达实体与实体间关系的对象属性,Domain表示属性描述的事实对象,Range表示属性描述指向的实体类;二是显示实体数据描述类型的数据属性,Domain是事实的实体类,Range则是数据类型。西南联大历史文献资源本体建设中的实体属性及其约束条件如表4所示。

表4 “数字联大”平台历史文献资源本体建设中的实体属性及约束条件示例

对象属性			数据属性		
分类	Domain	Range	分类	Domain	Range
出生日期	Person	Time	姓名	Name	xsd:string
出生地	Person	Place	性别	Gender	xsd:string
死亡日期	Person	Time	民族	Nation	xsd:string
死亡地	Person	Place	国家	Place	xsd:string
任职于	Person	Department	时间	Time	xsd:string
著作	Person	Work	籍贯	Place	xsd:string
属于	Thing	Thing	省份	Place	xsd:string
就读于	Person	University	市	Place	xsd:string
包括	Thing	Thing	刊物	Journal	xsd:string
搬迁至	University	Place	主题	Theme	xsd:string
发表于	Work	Time	文体	Style	xsd:string
刊登在	Work	Journal	载体	Type	xsd:string
发生地点	Thing	Place	文学	Theme	xsd:string
发生时间	Thing	Time	教育	Theme	xsd:string

(3) 资源本体实例化。在规范西南联大历史文献资源本体建设涉及的实体、关系、属性等数据体系后,利用Protégé本体构建工具对系统准备数据进行整理建设,构建西南联大历史文献资源本体实例,如图4所示。本体构建完成之后,需利用数字插件工具对本体表达逻辑进行事实验证,检验实体关系与属性表达的一

致性,确保资源本体模型对历史事实表示无误与清晰。将构建完成的西南联大历史文献本体资源数据上传数据库后台保存,可用于其他文献功能应用开发如西南联大历史文献知识图谱的构建。构建完成的西南联大历史文献资源本体能够将诸多离散的文献知识单元统一归档,通过关系节点互相关联,按照计算机标准化表达

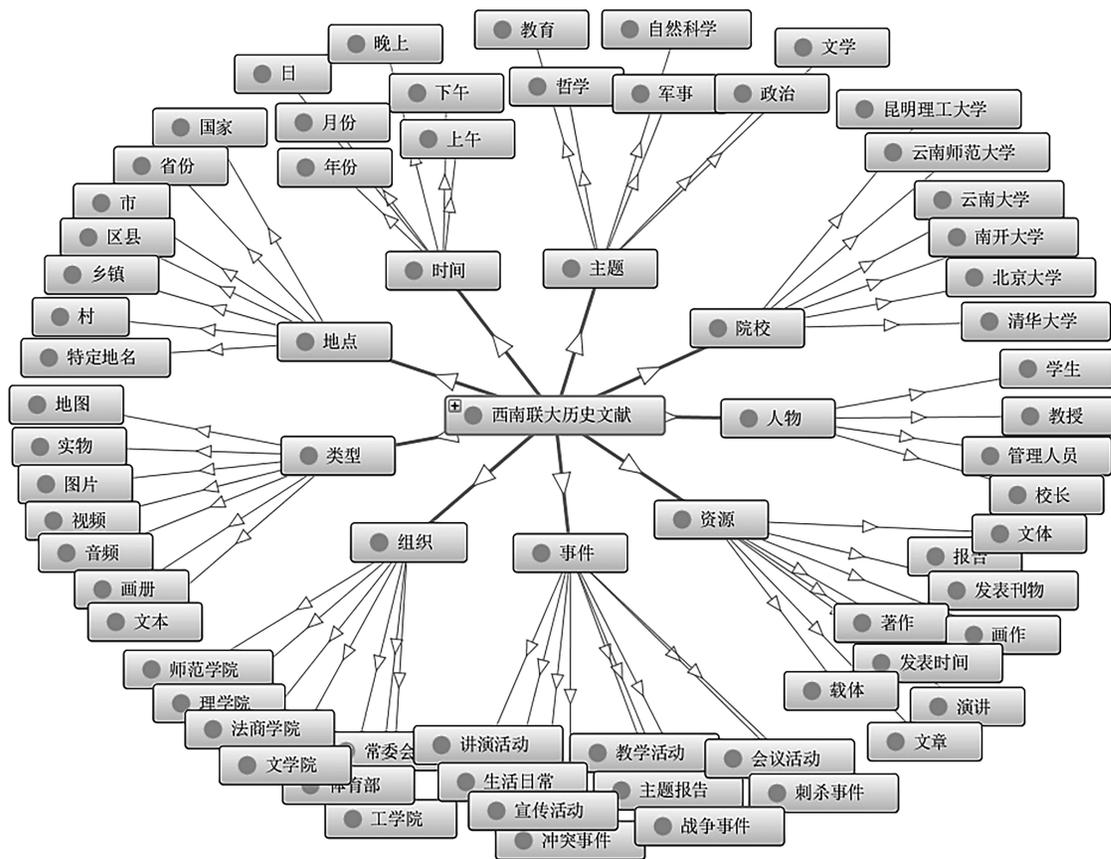


图4 “数字联大”平台历史文献资源本体示例

模式形成同一范式的西南联大历史文献结构化文档，既可以实现文献内容关联检索也为其他特色功能的开发提供基础数据。

3.3 数字化平台知识聚合服务

3.3.1 知识关联检索发现

以昆明市档案馆馆藏“闻一多惨案”有关的档案历史文献资源为实例，将“闻一多惨案”作为西南联大历史文献本体中的一个实例对象进行构建和展示。“数字联大”平台为用户提供历史文献内容的知识关联检索发现、可视化、GIS查询等服务，本例的“闻一多惨案”实体知识服务也以平台总规划为实现纲领，以知识聚合展示形式为用户提供查询服务。根据本体构建确定的核心概念，设置“人物”“组织”“资源”等主要栏目以供用户快速访问。以平台“事件”栏目为例，用户在“数字联大”平台以“闻一多惨案”为检索关键词进行

查询，得到查询结果，点击界面上教育事业人物下属的“闻一多”按钮，平台即可展示人物简介、相关图片，以及和闻一多有关的文献资源信息。本体结构图显示，闻一多为中国民主同盟的主要负责人之一，且为西南联大文史系国文教授，与李公朴互为战友，涉及“人物”“组织”等核心概念。类似于此，通过点击平台其余应用栏目同样能够获取相应知识服务。同时，“数字联大”平台本体结构图可通过可视化的形式向用户展示西南联大文献资源知识组织和管理过程，用户可根据本体结构图中的知识单元或核心概念进行再次检索，从而满足一站式、语义化、关联性的文献资源利用需求，从细粒度层面实现西南联大历史文献资源知识聚合服务。

3.3.2 空间信息实时访问

“数字联大”平台对西南联大历史文献资源中的历史地理空间信息进行采集分析，根据文献资源内容特征和平台功能需求，通过GIS技术进行地理信息元数据

转换, 形成机器可以理解的计算机语言, 生成西南联大历史文献资源历史地理信息图层, 结合二维地图平台提供的底图实现GIS查询服务。GIS查询服务将西南联大历史文献资源中重要的元数据描述信息加入数字地图, 如文献资源保存主体地址与历史事件发生的地名、人物籍贯、人物居住地等信息, 以可视化形式在地图中展示信息, 帮助用户系统了解西南联大历史文献资源分布情况等位置信息。同时借助线上地图导航功能, 实现西南联大历史文献资源实时导航。当用户在平台中点击具有地理位置信息的文献资源内容时, 平台即可提供GIS查询服务, 根据用户检索的地理位置名称进行相应的数据导航分析, 如《闻一多被害侦查报告书》中“西仓坡”的地理位置概念。用户点击界面中的“西仓坡”后, 界面左侧显示导航功能, 平台根据用户实时地理位置给出最佳导航路径或关联到第三方专业地图平台, 为用户提供空间位置参考信息。

4 结语

数字人文背景下, 智能化聚合、挖掘和利用历史文献资源成为研究热点。为了更好地促进历史文献资源知识聚类, 将各类历史文献信息元素进行有效关联, 形成完备的历史知识体系, 需搭建具有统一文献数据管理规范和多类数字化功能服务的历史文献共享平台。提出的专题历史文献资源数字共享平台旨在为用户提供一系列专题资源、聚类数据、数字工具和便捷服务, 促使用户积极学习并鼓励研究者进行深度的研究与讨论。在此基础上, 也可对专题历史文献资源进行“故事化”组织开发, 围绕文献内容涉及的历史事件、历史人物、地理信息等梳理历史发展脉络^[28], 增强历史文献资源的可读性、趣味性, 激发专题历史文献的文化活性, 扩大平台的用户规模, 让专业领域研究者有所得、大众读者有所获, 体现寓教于乐的教育学习理念。

参考文献

- [1] 成一农. 抛弃人性的历史学没有存在价值: “大数据”“数字人文”以及历史地理信息系统在历史研究中的价值[J]. 清华大学学报(哲学社会科学版), 2021, 36(1): 181-190.
- [2] 霍艳芳, 何思源. 数字人文视域下城市记忆资源整合与开发路径研究[J]. 档案学研究, 2018(1): 29-34.
- [3] 刘燕权, 韩志萍. 美国记忆: 美国历史资源数字图书馆[J]. 数字图书馆论坛, 2009(7): 66-70.
- [4] 周耀林, 刘晗. 数字记忆建构: 缘起、理论与方法[J]. 山东社会科学, 2020(8): 50-59.
- [5] 刘石, 李飞跃. 大数据技术与传统文献学的现代转型[J]. 中国社会科学, 2021(2): 63-81.
- [6] 梁继红. 走向文本的历史档案数字整理: 历史追溯与时代转型(下)[J]. 档案学通讯, 2022(1): 60-66.
- [7] 邓君, 王阮. 口述历史档案资源知识图谱与多维知识发现研究[J]. 图书情报工作, 2022, 66(7): 3-16.
- [8] 汤萌, 李芳, 王昕. 历史地理信息的获取与标准化: 以上海交通大学馆藏地方历史文献为例[J]. 图书馆杂志, 2018, 37(1): 50-57.
- [9] 国家图书馆. 革命文献与民国时期文献联合目录[EB/OL]. [2024-03-28]. <http://pcpt.nlc.cn/mgwx/historyDoc/connection.action>.
- [10] 上海图书馆. 上海图书馆开放数据平台[EB/OL]. [2024-03-28]. <https://data.library.sh.cn/index>.
- [11] 中国第一历史档案馆. 满文档案图像识别软件的开发与应用项目[EB/OL]. [2024-03-28]. https://fhac.com.cn/research/subject_research/detail/5577.html.
- [12] 上海交通大学出版社. 中国地方历史文献数据库[EB/OL]. [2024-03-28]. <http://ndfwx.datahistory.cn>.
- [13] 国家图书馆. 中国记忆项目实验网站[EB/OL]. [2024-03-28]. <https://www.nlc.cn/web/ziyuanfuwu/zhuantiziyuan/cmptest/index.shtml>.
- [14] Language Research Group University of Leeds. The Quranic Arabic Corpus[EB/OL]. [2024-03-28]. <https://corpus.quran.com>.
- [15] Harvard University, Academia Sinica, Peking University. China Biographical Database[EB/OL]. [2024-03-28]. <https://projects.iq.harvard.edu/cbdb>.
- [16] 台湾大学数位人文研究中心, 台湾大学图书馆. 历代宝案脉络分析系统[EB/OL]. [2024-03-28]. <http://lidaibaoan.digital.ntu.edu.tw>.
- [17] 上海图书馆. 中国家谱知识服务平台[EB/OL]. [2024-03-28]. <https://jiapu.library.sh.cn/#/>.
- [18] 王萍, 黄小宇. 基于数字人文的历史档案文献开发实践解析: 对98个国内外项目的实证研究[J]. 档案与建设, 2023(3): 55-60.
- [19] 程静. 国内外数字人文服务平台建设现状及思考[J]. 图书馆学研究, 2020(2): 37-42.
- [20] 金玲娟. 我国图书馆数字人文服务现状、障碍与对策研究[J]. 图

- 图书馆工作与研究, 2018 (9): 15-20.
- [21] 王萍, 黄新平. 基于关联开放数据的数字文化资源语义融合方法研究: 欧洲数字图书馆案例分析[J]. 图书情报工作, 2016, 60 (12): 29-37.
- [22] 余波, 姚明, 刘孟. 加拿大国家图书档案馆发展模式研究[J]. 兰台世界, 2014 (14): 16-17.
- [23] 张亚宏. 《改造社区: IMLS2018—2022战略规划》解读及思考[J]. 图书馆工作与研究, 2019 (1): 94-99.
- [24] 陈姝睿, 梁子然, 饶洋辉. 细粒度语义知识图谱增强的中文OOV词嵌入学习[J]. 计算机科学, 2023, 50 (3): 72-82.
- [25] 夏翠娟. 文化记忆资源的知识融通: 从异构资源元数据应用纲要到一体化本体设计[J]. 图书情报知识, 2021 (1): 53-65.
- [26] 翟姗姗, 张纯, 许鑫. 文化遗产数字化长期保存策略研究: 以“威尼斯时光机”项目为例[J]. 图书情报工作, 2019, 63 (11): 140-148.
- [27] 段宇锋, 黄思思. 本体构建方法研究[J]. 情报杂志, 2015, 34 (11): 139-144.
- [28] 李晓珊. 历史文献资源创新知识服务研究: 以新加坡国家图书馆委员会Curiosity平台为例[J]. 数字图书馆论坛, 2023, 19 (1): 34-42.

作者简介

陈海玉, 女, 博士, 教授, 研究方向: 数字人文、文献资源建设、公共文化服务, E-mail: 2374188301@qq.com。
王聪, 男, 硕士, 助理馆员, 研究方向: 数字人文、知识组织。
陈雨, 女, 硕士, 研究方向: 知识组织、文献资源建设、公共文化服务。
杨娟, 女, 硕士, 研究方向: 文献资源建设。
何茜林, 女, 硕士, 研究方向: 文献资源建设。

Construction of Digital Sharing Platform of Thematic Historical Document Resources from the Perspective of Digital Humanities:
Taking Historical Document of Southwest Associated University as an Example

CHEN HaiYu¹ WANG Cong² CHEN Yu¹ YANG Xian¹ HE XiLin¹

(1. School of History and Archives, Yunnan University, Kunming 650091, P. R. China;
2. Xiangsihu College of Guangxi Minzu University, Nanning 530031, P. R. China)

Abstract: Traditional historical document thematic research often excavates, analyzes, and discovers new historical information through manual collection of massive document resources, which is time-consuming and labor-intensive. With the rise of digital humanities research, the concept and technical methods of digital humanities have brought new perspectives to the research of historical documents. The Chinese nation has a long history, and the thematic historical document resources formed are large in scale and scattered in storage. At present, it is urgent to build a digital sharing platform of thematic historical document resources with a unified data management model to realize the visualization of each thematic historical document resource, the correlation of semantic information, and the convenience of content browsing, so as to meet the new needs of historical research in the digital age. Based on the concept and technical methods of digital humanities research, this paper expounds the construction goal, construction mode, construction framework, and feasible path of the digital sharing platform of thematic historical document resources. Taking the construction practice of the sharing platform of Digital Associated University as an example, the data collection and processing of historical document resources of Southwest Associated University, the construction of document ontology information network, and the knowledge service of the digital platform are carried out, which realize the correlation and visual presentation of thematic historical document resources from static information to dynamic knowledge, and meet the needs of deep learning and historical research of platform users.

Keywords: Historical Document; Digital Humanities; Southwest Associated University; Shared Service Platform

(责任编辑: 王玮)