

关联数据环境下我国海洋历史文献资源 知识组织研究^{*}

华林 冯安仪 谭雨琦
(云南大学历史与档案学院, 昆明 650091)

摘要: 海洋历史文献数量众多, 繁杂分散, 资源异构现象明显, 现有文献资源建设方式难以实现资源间的深层关联与有效发掘利用。以海洋历史文献资源为对象, 首先, 在分析文献资源建设现状的基础上, 总结出海洋历史文献资源建设中存在的理论认知、资源异构和知识关联等方面的问题。其次, 以文本文献、历史舆图、金石文献、口述资料和音像资料五大模块, 构建海洋历史文献资源体系, 结合文献特点设计元数据结构, 利用Protégé工具构建我国海洋历史文献资源本体模型。最后, 以实例可视化形式, 验证该本体模型的合理性, 揭示海洋历史文献中时间、地点、人物和资源要素与特定事件的关联关系, 为更好地开发利用海洋历史文献提供方案。

关键词: 海洋历史文献; 本体构建; 关联数据; 知识组织

中图分类号: G254 **DOI:** 10.3772/j.issn.1673-2286.2023.08.002

引文格式: 华林, 冯安仪, 谭雨琦. 关联数据环境下我国海洋历史文献资源知识组织研究[J]. 数字图书馆论坛, 2023 (8): 15-25.

中国正处于历史性发展的关键时期, “拥抱海洋, 发展海权”是我国从海洋大国迈向海洋强国的必由之路。习近平总书记在党的二十大报告中强调维护国家全能力和海洋权益, 坚定捍卫国家主权及安全利益, 加快推进海洋强国建设^[1], 充分体现了党和国家对海洋的高度重视。我国拥有大量的海洋历史文献, 这些海洋历史文献是指1949年以前, 各个历史时期的中央政府、地方机构和个人等在海洋经略、海疆保卫、维护权益、捕捞生产、航海贸易、文化建设、宗教信仰、生态保护等活动中形成的, 对研究海洋治理历史、构建海洋记忆、维护海洋权益有参考、凭证和史料价值的历史文书、涉海方志、海洋著述、地理舆图、音像材料和口述资料等。我国现存海洋历史文献数量众多、分布广泛, 但在资源整合、知识组织与关联共享等方面仍存在滞后问题。关联数据富含语义数据网络, 可有效描述数字资源

实体之间的相关关系。结合关联数据技术对我国海洋历史文献资源进行知识组织研究, 有利于促进多源异构的海洋历史文献资源建设完整化、规范化, 推动知识连接及资源的深度整合, 提升海洋历史文献资源组织水平, 为细化海洋历史文献资源组织的粒度、加深其开发程度提供有益借鉴。

1 研究综述

我国现存海洋历史文献内容涵盖经济、政治、文化、社会、安全、生态、海疆等方面, 在开展海洋治理、传承海洋记忆与维护国家领海主权方面发挥了重要作用。为更好地发掘利用这一珍贵历史文献, 学界对其知识组织问题开展相关研究。

在海洋历史文献知识组织层面, 已有针对南海文

收稿日期: 2023-07-13

^{*}本研究得到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。

献资源的相关研究,内容主要聚焦在知识抽取、知识图谱和知识发现方面。秦东^[2]基于证据链理论,构建南海历史文献关联数据支撑服务框架模型,对数据资料进行语义描述、本体构建和链接,生成并发布关联数据,实现对南海历史文献知识的有序重组,该研究对于完善南海历史文献证据链具有应用价值。王燕红等^[3]以《中国在南海的历史性权利及证据目录》及百度百科部分数据为主要数据来源,从知识发现视角对南海书证目录进行内容挖掘及分析,建立文献内容间的链接关系,构建南海书证目录知识图谱。司徒凌云等^[4]以南海疆文献资料为基础,结合证据链理论,提出基于多模态知识图谱设计南海疆维权证据链系统的技术体系,以实现多载体、多形式的南海疆文献知识关联、内容分析与维权文献的可视化。孙浩洋等^[5]基于知识元理论对民国南海文献进行知识元内容标引以及知识元描述规则与抽取规则的探索研究,以构造南海历史文献知识元模型,通过可视化呈现文献内容与其间关系,提高知识关联、挖掘及利用效率。以上研究主要面向南海历史文献,并以此为数据基础进行知识抽取、知识图谱和知识发现研究,但尚未产生涵盖海洋历史文献资源全领域的知识组织成果,未深入探讨海洋历史文献整体知识分类,难以整体反映和关联该领域的深层知识。

历史文献知识组织的研究集中在应用层面。①历史文献本体构建。孙振嘉等^[6]以五四运动历史档案资源为研究对象,利用Protégé工具构建本体模型并验证其合理性,深度揭示了相关历史档案资源要素的关联。李宝玲等^[7]利用数字人文技术和本体构建方法进行历史档案专题知识库构建研究。②历史文献知识图谱。陈海玉等^[8]通过对徽州文书进行实体概念抽取、异构资源知识融合,构建了民间历史文献知识图谱并以此进行深入研究。其他研究成果还有名人年谱资源知识图谱^[9]、中国近代史知识图谱^[10]、红色记忆知识图谱^[11]及民国报纸知识图谱^[12]等。③历史文献知识聚合及发现。严承希等^[13]结合交互式可视化和语义聚合技术,以清代档案文献为研究对象构建知识聚合模型。邓君等^[14]结合本体和知识图谱技术,设计了由数据资源层、知识组织层、知识关联层、知识发现层4个层次构成的口述历史档案资源知识发现模型。海洋历史文献属于历史文献的范畴,历史文献相关知识组织研究成果可从理论、技术与方法维度为本研究提供参考文献。鉴于此,上述相关历史文献本体模型与知识图谱构建方法以及知识可视化、知识关联等内容可供本研究参考,但还需结合海

洋历史文献自身特点、类别构成与属性关系设计具体研究方案。

综上所述,我国海洋历史文献资源知识组织研究有以下亟待解决的问题。①利用现有关联数据的理论及应用研究成果和本体论,深入挖掘海洋历史文献资源语义特征和关联关系,解决文献本体的描述、构建及知识揭示表层化问题。②借鉴学界在南海历史文献以及历史文献本体构建、知识图谱、知识聚合与知识发现方面的成果,解决海洋历史文献资源知识组织问题。为此,本研究借鉴历史文献知识组织相关研究成果,结合海洋历史文献产生领域独特、种类繁多、规模较大与多源异构等特点,从资源体系构建、知识本体建模、关联模型发布等维度,着力解决海洋历史文献资源知识组织问题。

2 实践工作分析

2.1 海洋历史文献资源建设现状

鉴于海洋历史文献的珍贵价值,档案馆、图书馆和高校等建设主体广泛开展文献资源建设开发工作,并取得显著成果。在汇编方面,主要有中国边疆研究资料文库成果《海疆文献初编 沿海形势及海防》《中国海疆文献续编 明代倭患》《中国海疆文献续编 台湾琉球港澳》,蝠池书院的《中国古代海岛文献地图史料汇编》《中国南海诸群岛文献汇编》《广东历代方志南海史料辑录》等。专题数据库建设成果有中国第二历史档案馆的民国时期海疆档案资源数据库^[15]、中国南海研究院的南海文献数据库^[16]、厦门大学图书馆的东南海疆研究数据库^[17]、中国海洋大学图书馆的海大文库^[18]、海南师范大学的南海及南海诸岛特色数据库^[19]、深圳市盐田区图书馆的海洋专题特色文献库^[20]等。这些建设成果为海洋历史文献资源整合、本体构建与知识聚合奠定了资源基础。

2.2 海洋历史文献资源建设滞后问题

2.2.1 海洋历史文献理论认知问题

学界对海洋历史文献进行了理论研究,如吴晓文等^[21]、李慧颖等^[22]探讨了海洋档案内涵外延问题,刘家真等^[23]、方泉等^[24]等对海洋历史档案的定义、类型与价

值开展了研究。整体而言,研究还停留在表层描述层面。①学界尚未从理论上对文献概念进行科学界定,主要以海洋历史活动内容(政治、经济、军事及外交等)划分文献类型,无法体现出文献载体、文种、形制等的多样性,这对文献知识本体的完整构建产生不利影响。②对文献文本的解读与研究不够深入,还存在对文献史学与现实价值认知不全的问题,这不利于后期对文献知识的挖掘与聚合利用。

2.2.2 海洋历史文献资源异构问题

①数据结构异构。由于保管主体、保管方式不同以及数据资源构建方式的差异,海洋历史文献资源格式不一,形成了Word、HTML、PDF等半结构化、非结构化数据,数据异构特征较为显著。②术语标准异构。由于各保管单位采用的文献管理标准规范不同,文献资源在概念、属性、实例及关系等的名称^[25]和代码表示方面具有差异。如在元数据标准方面,图书馆常用的元数据格式有MARC、Dublic Core (DC)等,博物馆常用的是CDWA、CIMI、VRA Core等,档案馆则主要使用EAD格式^[26]。机构间元数据结构及其格式的差异不仅影响机构间数据资源的互操作效率,也为进一步创建与发布关联数据带来了一定困难。③句法表示异构,即各机构的文献信息表示结构不同。随着资源描述框架(Resource Description Framework, RDF)、网络本体语言(Ontology Wed Language, OWL)的广泛应用,文献资源的表示句法也愈加丰富,所引发的文献资源句法表示异构问题将更加普遍。

2.2.3 海洋历史文献知识关联问题

一方面,海洋历史文献资源建设取得成效,但其知识组织缺乏深度。中国第二历史档案馆、中国南海研究院、厦门大学图书馆、中国海洋大学图书馆等纷纷建成不同专题的海洋文献数据库,数据库建设已成趋势。现有海洋文献数据库主要采用传统录入和检索方法建库,提供全文检索和文本检索功能,服务方式单一,缺少统一规范的元数据著录标准,难以实现对不同语义层面文献资源的细粒度组织,知识聚合和语义关联松散。另一方面,海洋历史文献资源内容表述单一,缺少深层次描述,无法揭示内在的本质属性与知识关系,缺乏对文献资源所含时空关系、事件-事件关系、事件-人

物关系、人物-人物关系等的完整描述,难以形成多维立体的知识体系和交叉关联的知识网络,这阻碍了海洋历史文献资源的知识性开发与价值实现。

3 研究思路与方法

3.1 关联数据环境下海洋历史文献资源知识组织思路

基于关联数据技术构建海洋历史文献本体模型,设计海洋历史文献的元数据结构,并以此为基础对已采集的文献资源进行语义化描述、知识本体建模,以解决目前海洋历史文献资源建设中存在的问题,借助关联数据技术,实现多源异构资源间的联系与共享。按研究过程将海洋历史文献资源知识组织分为收集清洗、知识描述、本体构建、知识关联及知识应用5个阶段。

(1) 收集清洗。一方面,将散存于档案馆、图书馆、海洋机构、高校等单位的海洋历史文献资源作为主要数据源,结合海洋历史文献自身分类特点,构建海洋历史文献资源体系。该体系分为文本文献、历史舆图、金石文献、口述资料和音像资料5个大类,并下设子类。以此体系为依据进行数据采集,实现完整化和体系化资源建设。另一方面,对异构海洋历史文献资源数据进行清洗,利用光学字符识别(Optical Character Recognition, OCR)、自然语言处理(Natural Language Processing, NLP)等技术,对非结构化与半结构化数据资源进行调整和完善,为后续本体构建及关联数据模型建设创造条件。

(2) 知识描述。根据元数据标准规范具体描述资源的内容特征及语义结构,实现对数据的有效规范管理。参考元数据编码和传输标准(Metadata Encoding and Transmission Standard, METS)与RDF等有关规定,及国家档案局颁布的《文书类电子文件元数据方案》《照片类电子档案元数据方案》《录音录像类电子档案元数据方案》《口述历史电子档案元数据方案》等,结合海洋历史文献的特征复用并重新定义了VRA Core和DC元数据集中的元数据描述元素,设计海洋历史文献元数据规范,以加深海洋历史知识资源信息揭示程度,体现海洋历史知识间的概念逻辑关系,有效进行知识描述并实现知识标准化。同时,由表层到深层进行知识描述与概括,从形式、内容和整体层面多维构建海洋历史文献资源语义网络,缩小异构知识鸿沟,为知

识深度关联提供保障,实现组织层面的海洋历史知识融合。

(3) 本体构建。在对元数据进行规范描述的基础上,将得到的数据转换为富有语义的RDF元数据,对海洋历史文献资源属性及其属性值进行描述,实现信息资源的RDF化,并选用适当的工具设计海洋历史文献资源本体模型。具体通过明确本体领域、列举领域术语、定义类与属性、添加属性约束和本体实例化5个步骤进行本体构建,借助Protégé工具的OntoGraf功能实现本体可视化。

(4) 知识关联。经过实体消歧、共指消解、知识合并等过程,结合关联数据技术,将转化后的RDF元数据进行关联,将庞杂的海洋历史文献资源整合成相互联系的有机集合,采用双向长短期记忆网络(Bi-Directional Long-Short Term Memory, Bi-LSTM)和条件随机场(Conditional Random Field, CRF)技术抽取实

体外部关联关系,用规则模板和人工标注等方式抽取实体内在属性关系,并通过知识可视化的方式展现对应关联关系,以揭示海洋历史文献资源间的潜藏关系,实现知识网络化管理。可使用关系型数据库(如Oracle、DB2、MySQL)和非关系型数据库(如NoSQL、Neo4j、GraphDB)等可视化工具。

(5) 知识应用。在对海洋历史文献进行序化处理及本体构建的基础上,对其进行知识抽取,发布关联数据并提供文献资源的浏览与查询检索等功能。①利用知识抽取完成资源实体提取和关系提取,从多维度、细粒度的角度实现深层语义化链接。②将本体模型中的RDF数据导入三元组数据集,对重复、弱关联的资源数据进行深度消歧、序化处理。③将相互关联的数据链接成知识图谱,用以揭示知识的深层链接,实现资源的可视化检索,为用户提供数据关联检索、智能推荐匹配、知识发现、知识推理等服务。具体思路如图1所示。

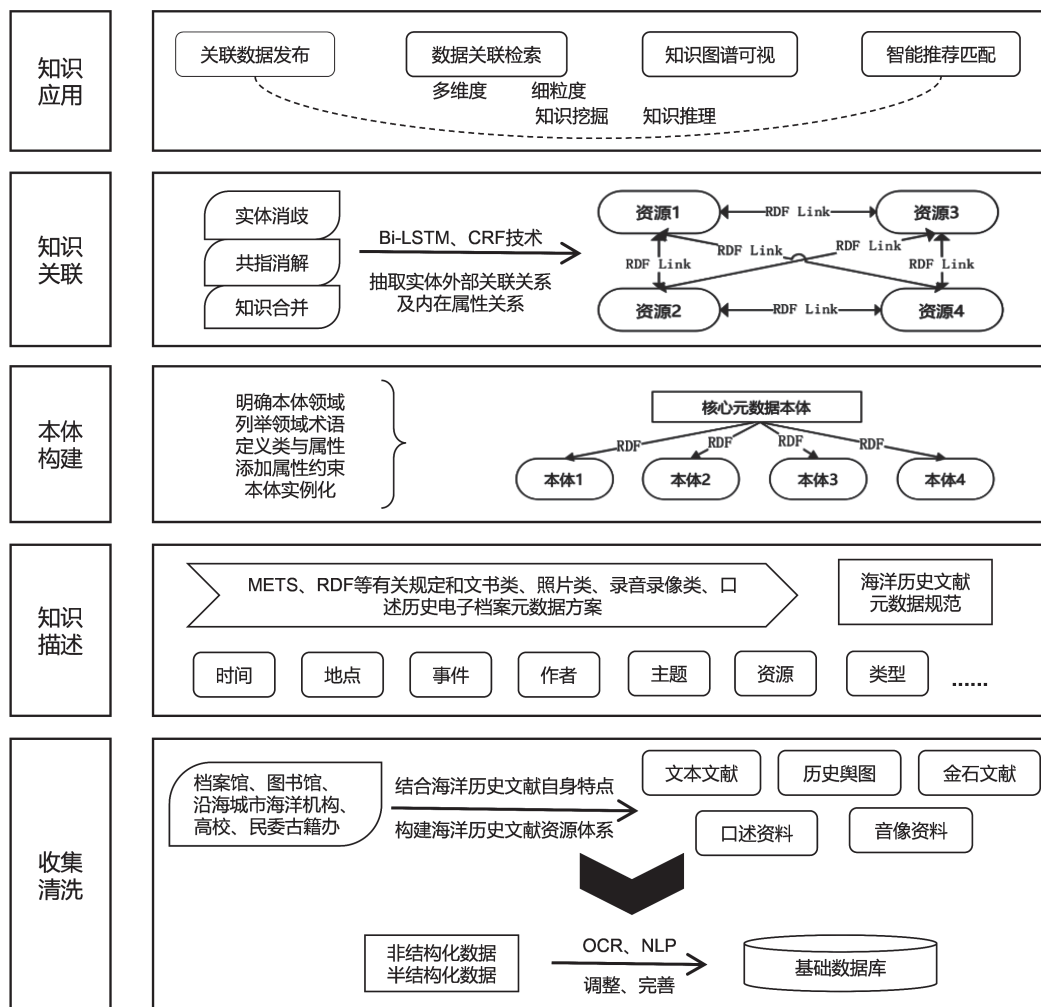


图1 海洋历史文献资源知识组织思路

3.2 关联数据环境下海洋历史文献资源知识组织方法

3.2.1 海洋历史文献资源体系构建

为明确区分海洋历史文献的构成模块,依据文献遗存状况,借鉴文献学相关分类方法,将海洋历史文献资源分为文本文献、历史舆图、金石文献、口述资料和音像资料5个大类:文本文献下设政府海洋公文、沿海地方志书、海防海战文献、海上交通文献、海洋邦交文献、海洋科技文献、海洋文化文献7个子类;历史舆图包括陆海古图、航海古图、海疆地图与海防地图4个子类;金石文献包括石刻、碑刻和铭文;口述资料包括人物历史、事件口述、海洋诗赋、海洋传说;音像资料包括海洋历史照片、历史音频、历史视频等。利用各档案馆、图书馆等机构收藏的档案、图书等文献资料如《海疆文献初编 沿海形势及海防》《中国海疆文献续编》和南海文献数据库、东南海疆研究数据库等收录内容,参照《海洋档案管理规定》《国家海洋专项档案管理办法》《海洋管理机关档案业务规范》以及各级地方志等权威资料,构建文献资源体系,以确保文献资源本体构建过程中核心概念的准确性。海洋历史文献以非结构化和半结构化数据为主,包含档案、方志、著作、金石与音像等内容,利用OCR将其转化成文本数据,将文本内容视为字符串并采用KMP(Knuth-Morris-Pratt)算法实现抽取,以二维表的形式存储于计算机中,为后续元数据结构设计与本体建模构建统一的数据格式。

3.2.2 海洋历史文献元数据结构设计

元数据是描述数据的数据,用于描述某类资源或数据的属性并实现对数据资源的检索及管理^[27]。目前,国内外常用的文献资源元数据标准有《编码档案著录》(Encoded Archival Description, EAD)、《国际档案著录规则(总则)》[ISAD(G)]、DC、《档案编码背景信息著录规范》(Encoded Archival Context, EAC)、《文本编码倡议》(Text Encoding Initiative, TEI)等。参考国家档案局发布的行业标准与方案及DC、EAD、ISAD(G)等元数据通用标准,抽取海洋历史文献资源元数据元素,进行海洋历史文献的语义特征描述(见表1)。

表1 海洋历史文献资源元数据元素

元素名称	元数据标识	元素名称	元数据标识
事件	event	类型	type
地点	place	地区	region
日期	date	来源	source
作者	author	类型	type
出版方	publisher	朝代	dynasty
主题	theme	海域	sea area
资源	resource	关键人物	personappearance
描述	description	字体	font
活动	activity	印章	seal
作品	work	载体形式	carrier

在关联数据的环境下,还需要以HTTP URI模式进行URI标识^[28]。URI的基本结构为“基地址+实体类型名称+实体编号”。以http://hylswx.com为海洋历史文献资源管理网址,以文本文献(Textual Literature)、历史舆图(Historical Map)、金石文献(Jinshi Literature)、口述资料(Oral Information)和音像资料(Audiovisual Materials)为实体类型。文本文献类的序号1数据所对应的URI就可以表示为http://hylswx.com/Architectural sites/000000000001。

3.2.3 海洋历史文献资源知识本体建模

海洋历史文献资源分布广泛,种类丰富,涵盖文本、图片、画册、音视频和实物等形式,各类文献间存在一定的差异性,通过进行海洋历史文献资源本体建模可建立各概念间的关系,形成个体与个体间的关联关系。结合海洋历史文献的特点,利用操作便捷的Protégé工具构建海洋历史文献资源本体模型,参考斯坦福大学的七步法(明确本体的领域和范畴、调研现有本体的可复用性、列举领域的术语和概念、定义类及类的层级结构、本体实例化、本体检测^[29]),将构建过程分为明确本体领域、列举领域术语、定义类与属性、添加属性约束和本体实例化5个阶段。

(1)明确本体领域。构建知识本体模型时首先应明确其领域,再划分应用范围,并基于此构建本体间的交叉关系。海洋历史文献遗存丰富,形成于海疆保卫、海洋外交、维护海权、航海贸易等活动中,涵盖文本、图片、画册、音频、视频、实物等形式。为此,可在海洋历史文献一级类目之下,按文献性质划分文本文献、历史舆图、金石文献、口述资料和音像资料等二级类目,

其下再设立三级类目，具体类目如表2所示。

表2 海洋历史文献资源类目

一级类目	二级类目	三级类目
海洋历史文献	文本文献	政府海洋公文
		沿海地方志书
		海防海战文献
		海上交通文献
		海洋邦交文献
		海洋科技文献
		海洋文化文献
	历史舆图	陆海古图
		航海古图
		海疆地图
		海防地图
	金石文献	石刻
		碑刻
		铭文
	口述资料	人物历史
		事件口述
		海洋诗赋
		海洋传说
	音像资料	海洋历史照片
		海洋历史音频
		海洋历史视频

(2) 列举领域术语。专业领域术语为类和属性的添加与设置提供支撑，罗列核心术语与关键词也为本体构建奠定基础。在分析及凝练领域内核心术语的基础上，选择尽可能多的渠道和方法，以全面获取与选择海洋历史文献领域内的关键术语。首先，依据《中国分类主题词表》等现有词表对历史文献领域的描述，选取适合海洋历史文献的相关词汇；其次，参考现有文献网站及数据库使用的术语；最后，研究海洋历史文献相关著作、论文，并从中获取领域术语，海洋历史文献资源领域部分术语如表3所示。

表3 海洋历史文献资源领域部分术语

类	术语列举
人物	姓名、性别、年龄、籍贯、民族、配偶、子女、出生地、死亡地
地点	国家、省、市、县、乡、镇
时间	朝代、年、月、日
资源	名称、作者、发表时间、主题、类型、载体形式、装订方式
类型	文书、图片、画册、实物、音频、视频
主题	海洋经略管理、海疆保卫、海洋外交、维护海权、捕捞生产、航海贸易

(3) 定义类与属性。在参考模型CIDOC CRM的基础上，针对海洋历史文献资源的特征，结合自上而下和自下而上的方法对领域内术语进行组织分类，形成知识模块，通过建立知识模块间的关联搭建类之间的层级关系。选择人物、地点、时间、类型、主题、资源、事件、作品、关系等作为海洋历史文献本体的核心类细化单元。人物类可细化为个体、团队和组织，地点类可细化为省、市、县、乡、镇等，事件类可细化为军事事件、经济事件、商业事件、文化事件等，时间类可细化为年、月、日、时、分等，资源类包括文书、图片、画册、实物、音频、视频等。可利用Protégé工具生成具体的海洋历史文献本体图。

(4) 添加属性约束。在定义类的基础上，需添加对象属性 (Object Property) 和数据属性 (Data Property)。对象属性用于描述类与类间的关系，其定义域和值域都是类；数据属性描述对象的数据类型，其定义域是类，值域则为数据类型^[30]。以海洋历史文献资源为对象，设人物 (Person)、地点 (Place)、名字 (Name)、时间 (Time)、类型 (Type)、主题 (Theme)、资源 (Resource)、事件 (Event)、作品 (Work) 及关系 (Relationship) 10个类，下设包括 (contains)、属于 (belongs_to)、关于 (is_about) 等31个对象属性和姓名 (Name)、国家 (Country)、省 (Province)、市 (City) 等31个数据属性。根据数据类型分别设置其定义域及值域，在工作区添加属性，并对对象属性和数据属性进行语义约束，类及属性的对应关系见表4。

(5) 本体实例化。根据海洋历史文献知识元素及领域内术语，在确定海洋历史文献本体模型涵括的类、添加属性之后，可制作海洋历史文献本体类可视化图，如图2所示。知识本体构建完成后，需利用Protégé工具的HermiT 1.4.3插件对其进行推理验证，其目的是检验本体的一致性，并检测类间的属性关系。若存在逻辑不合理、属性关系不符等问题，错误部分会标红显示。在推理验证完成后，可利用OWLviz插件查看本体前后变化情况。

3.2.4 海洋历史文献资源知识组织实例

添加实例是进行本体修正的重要环节，同时也是验证海洋历史文献知识组织方案适用性的基础。以郑和海洋历史文献为例，通过查阅现有文献记载及网络数据，提取郑和知识本体实例，并构建实例间的关联关

表4 对象属性约束与数据属性约束

对象属性			数据属性		
属 性	定义域	值 域	属 性	定义域	值 域
event_name	Event	Name	Name	Name	xsd:string
event_place	Event	Place	Country	Place	xsd:string
is_about	Event	Person	Province	Place	xsd:string
has_time-span	Event	Time	City	Place	xsd:string
date_of_event_initiation	Event	Time	Native_Place	Person	xsd:string
date_of_event_completion	Event	Time	Working_Place	Person	xsd:string
participate_in	Person	Event	Official_Experience	Person	xsd:string
is_from	Person	Place	Event_Experience	Person	xsd:string
is_relative_to	Person	Person	Created_Work	Person	xsd:string
has_position	Person	Post	Dynasty	Time	xsd:string
take_office	Person	Place	Year	Time	xsd:integer
has_experience	Person	Event	Document	Type	xsd:string
is_attendant_to	Person	Person	Picture	Type	xsd:string
is_member_of	Person	Group	Material_Object	Type	xsd:string
was_born_in	Person	Time	Audio	Type	xsd:string
died_in	Person	Time	Video	Type	xsd:string
has_name	Resource、Work	Name	Marine_Strategic_Management	Theme	xsd:string
has_subtitle	Resource、Work	Subtitle	Maritime_Defense	Theme	xsd:string
was_created_by	Resource、Work	Creator	Ocean_Diplomacy	Theme	xsd:string
was_assigned_by	Resource、Work	Identifier	Maintain_Sea_Power	Theme	xsd:string
has_type_of	Resource、Work	Type	Fishing_Productione	Theme	xsd:string
has_format	Resource、Work	Format	Maritime_Trade	Theme	xsd:string
has_theme_of	Resource、Work	Theme	Title	Resource、Work	xsd:string
has_language	Resource、Work	Language	Theme	Resource、Work	xsd:string
is_classified	Resource、Work	Format	Creator	Resource、Work	xsd:string
is_stored_in	Resource、Work	Place	Type	Resource、Work	xsd:string
is_bound_as	Resource、Work	Pattern	Publisher	Resource、Work	xsd:string
is_originated_from	Activity	Event	Character_Name	Event	xsd:string
involves	Event	Activity	Origin	Event	xsd:string
belongs_to	Activity	Event	Process	Event	xsd:string
moved_to	Place	Place	Ending	Event	xsd:string

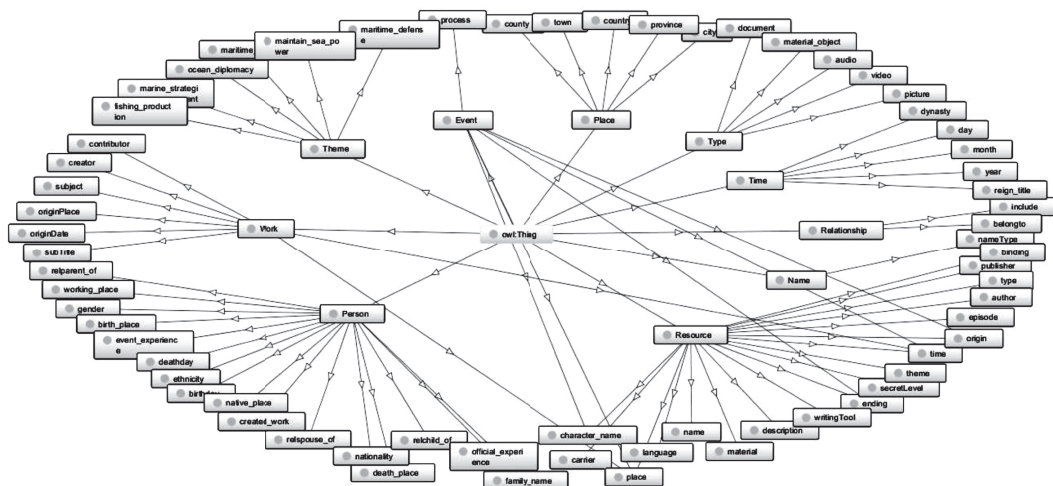


图2 海洋历史文献本体类可视化图

系。文本文献实例有《郑和家谱》《郑和家谱考释》《影印原本郑和家谱校注》《郑氏家谱首序》《赛氏总族牒》《赛典赤家谱》《马氏家乘》，金石文献实例有《故马公墓志铭》《郑和铜钟》《娄东天妃宫通番事迹记》《天妃灵应之记碑》《古里碑》《柯枝镇国之山碑》《锡兰布施碑》，历史舆图实例有《郑和航海图》。现以郑和海洋历史文献为例对人物本体和文献本体进行实例化：将数据分别添至对应类中，设置实例，并给实例添加相应数据属性。

在对多源异构的郑和海洋历史文献进行知识抽取的过程中，初步形成的数据集仍存在数据冗余或表达歧义等问题，因此需要利用知识融合技术补充与整合信息数据，解决数据冗余问题。海洋历史文献中存在

对同一实体的不同表述问题，例如郑和父亲的名字在《故马公墓志铭》中的表述为“公字哈只，姓马氏”，而《郑和家谱》记载“赛典赤后裔，其父米里金”。对此种情况，应统一同一实体的表述方式，可采用实体分组（Mention Pair）的方法，依照句段标注顺序，将两个实体描述为组并依序配对，比较对象间的共指关系，从而将相同指代项合并为一个实体标注对象^[8]。对上述“同人异名”问题，将统一以“姓+名”标注，在规范标注的同时体现出人物的家族谱系关系。

具体以《故马公墓志铭》为例进行分析，依据海洋历史文献本体模型中的实体及关联关系，抽取原文中郑和之父的姓名、祖籍、亲缘关系、生卒时间等信息的实例（见表5）。

表5 《故马公墓志铭》知识提取结果

《故马公墓志铭》内容	知识提取结果
公字哈只，姓马氏。世为云南昆阳州人。祖拜颜，妣马氏。父哈只，母温氏。公生而魁岸奇伟，风裁凛凛可畏，不肯枉己附人。人有过，辄面斥无隐。性尤好善，遇贫困及鳏寡无依者，恒保护赈给，未尝有倦容，以故乡党靡不称公为长者。娶温氏，有妇德。子男二人，长文铭，次和，女四人。和自幼有材志，事今天子，赐姓郑，为内官监太监。公勤明敏，谦恭谨密，不避劳勤，缙绅咸称誉焉。呜呼！观其子而公之积累于平日，与义方之训可见矣。公生于甲申年十二月初九日，卒于洪武壬戌七月初三日，享年三十九岁。长子文铭奉柩安厝于宝山乡和代村之原，礼也。铭曰：身处乎边陲而服礼仪之习，分安乎民庶而存惠泽之施，宜其馀庆深长，而有子光显于当时也！永乐三年端阳日资善大夫、礼部尚书兼左春坊大学士李至刚撰。马氏第二子太监郑和，奉命于永乐九年十一月二十二日到于祖宗坟茔，祭扫追荐。至闰十二月吉日回还记耳。	姓名：马哈只 祖籍：云南昆阳州 父亲：马拜颜 妻子：温氏 出生时间：甲申年十二月初九日（1344年） 死亡时间：洪武壬戌七月初三日（1382年） 葬于：昆阳和代村 长相：魁岸奇伟 性格：乐善好施，谦恭谨密 子女情况：两子，长子马文铭，次子郑和；四女，姓名不详 撰写人姓名：李至刚 撰写时间：永乐三年端阳日 撰写人职务：礼部尚书兼左春坊大学士 现存地点：云南晋宁县昆阳郑和公园

在Protégé工具中可设置映射（Functional）、逆映射（Inverse Functional）、传递性（Transitive）、对称（Symmetric）、非对称（Asymmetric）、自反（Reflexive）、非自反（Irreflexive）等关系，并据此进行知识组织与推理评估，即通过识别分析基础类与属性间的关联，设置对应关系并对属性进行赋值，推理出新的关联并修

复错误关联，进一步提升知识关联的准确度。将所提取的信息，包括郑和之父马哈只的生年（1344年）、出生地（云南昆明昆阳和代村）、卒年（1382年）、身份（郑和之父）、妻子（温氏）、子女（长子马文铭，次子郑和，另有四女）等信息导入Protégé工具，以生成关联图谱，较为清晰地表示出马哈只的人物概况与人物关系，见图3。

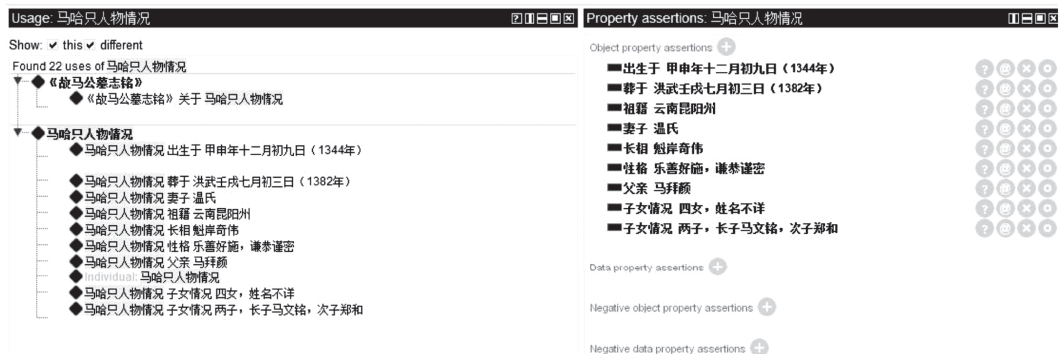


图3 《故马公墓志铭》实例添加

对郑和人物本身、文本文献、金石文献、历史舆图、郑和文献汇编添加实例后, 郑和海洋历史文献本体基

本构建完成, 选用Protégé工具的OntoGraf功能对其进行可视化展示(见图4)。

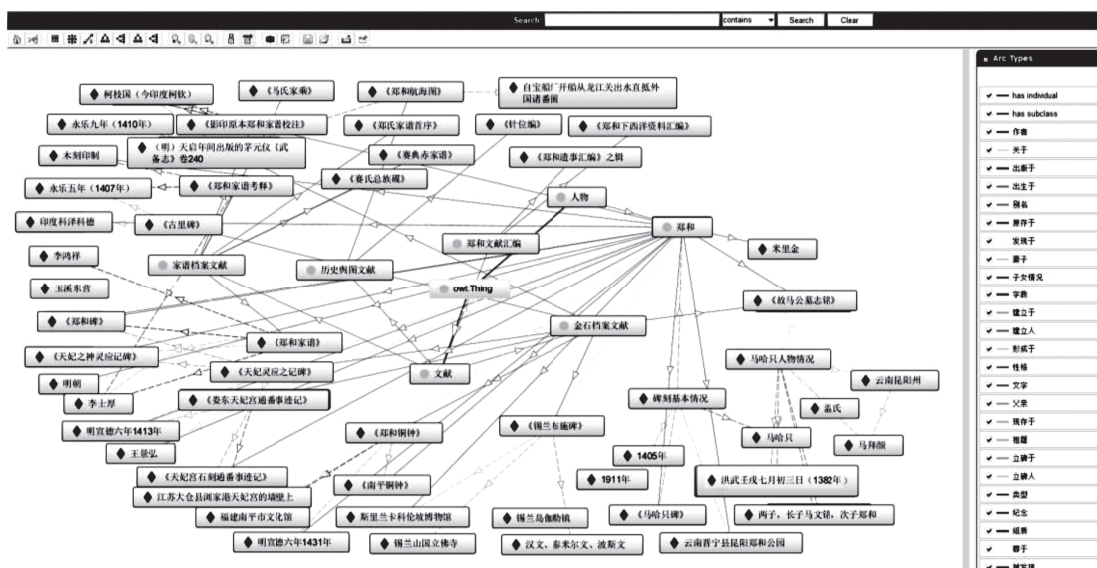


图4 郑和海洋历史文献本体可视化展示

以郑和海洋历史文献为例, 基于文献知识提取、分析与组织整理、本体模型构建和实体关系建立, 开展海洋历史文献的知识组织研究, 将文献涉及的人、事、物进行语义化聚类, 揭示了时间要素、地点要素、人物要素、信息资源要素与特定事件的关联关系, 并将整体组织结果以可视化方式呈现, 实现了对文献资源的知识描述及知识组织。研究表明, 通过运用数字技术方法与工具对海洋历史文献资源进行知识组织和可视化展示, 可更好地整理、挖掘与关联海洋历史文献资源中碎片化、繁杂化的隐藏信息。例如, 借助文献可视化成果可发现人物间的隐含关系, 也能清楚反映与某一地点相关的人物或事件, 为深刻揭示海洋历史文献资源间关系, 更好地发掘利用这一文献遗产提供有力支撑。

3.2.5 海洋历史文献资源关联模型发布

使用D2RQ平台搭建海洋历史文献关联数据模型, 即把相关数据转换为RDF格式, 并通过Web将其以关联数据的形式发布, 从而提高用户检索数据的便捷性。D2RQ平台主要包括D2R Server、D2RQ Engine及D2RQ Mapping语言, 其主要作用是提供对RDF数据的查询访问接口, 以供上层的RDF浏览器、HTML浏览器和SPARQL查询客户端调用, 且支持Oracle、MySQL、

PostgreSQL及Microsoft SQL Server等数据库。关联数据发布的重心是通过D2RQ Mapping语言(d2rq:ClassMaps和d2rq:PropertyBridges)将关系型数据库中的数据库表(Table)转化为关联数据中的类, 并建立关联关系。D2R Server的作用则是通过mapping工具生成的ttl文件发布关联数据。将信息资源作为关联数据链接到Web网络中, 根据语义链接机制挖掘资源实体间的关联关系, 并提供访问、浏览与查询功能, 以有效实现海洋历史文献资源的关联及聚合。

4 结语

现阶段海洋历史文献资源建设中还存在研究认知层次较浅、知识关联较弱、数据粒度较粗等问题。结合关联数据技术, 对文献资源进行知识收集、抽取、组织、关联, 利用Protégé工具构建海洋历史文献本体, 并对其进行知识分析与组织, 可实现文献数据间的RDF语义互联, 深层次揭示资源间关联关系, 促进海洋历史知识服务的创新与增值。本研究的应用价值体现在以下几个方面。首先, 可结合关联数据技术进行知识关联, 深入挖掘海洋历史知识内容及知识间潜藏关系, 实现海洋知识多维互联, 更好地把握海洋历史活动的内在关联及发展脉络。其次, 依托知识本体及知识图谱等工

具,可实现海洋历史人物及事件的关联查询,并通过关系节点(Cypher语法中的深度运算符)查询人、事、物及其相关信息,可极大地提高海洋历史知识实体检索效率。最后,当前海洋历史文献资源开发仍以文献汇编和数据库构建等方式为主,难以有效实现知识关联应用,诸多知识性开发方式亟待创新。本研究结合海洋历史文献自身特征进行知识本体构建及知识组织研究,有助于细化文献资源粒度,为实现资源语义组织、语义关联、可视化及多维知识性开发提供新的思路与方法。海洋历史文献纷繁芜杂,其知识组织仍存在简繁体转换不精准、知识语义理解不精准、实体识别与知识元抽取效率较低等问题,这些问题还有待进一步探讨、分析与解决。

参考文献

- [1] 习近平:高举中国特色社会主义伟大旗帜 为全面建设社会主义现代化国家而团结奋斗:在中国共产党第二十次全国代表大会上的报告[EB/OL]. [2022-11-05]. <http://jhsjk.people.cn/article/32551583>.
- [2] 秦东.南海维权历史文献的关联数据研究[D].南京:南京大学,2016.
- [3] 王燕红,司徒凌云,杨海平,等.基于知识图谱的书证目录知识发现研究:以南海书证目录为例[J].情报杂志,2022,41(3):173-180.
- [4] 司徒凌云,石进,杨海平,等.基于多模态知识图谱的南海疆维权证据链系统构建[J].情报杂志,2021,40(12):23-29,44.
- [5] 孙浩洋,沈固朝.民国南海文献知识元内容抽取规则研究[J].情报杂志,2022,41(12):132-139.
- [6] 孙振嘉,汪泽,邓君.数字人文视域下历史档案知识组织研究:以五四运动为例[J].兰台世界,2021(12):32-38.
- [7] 李宝玲,李珂,郭立鑫.面向深度利用的历史档案专题知识库构建研究:以中福公司档案为例[J].档案管理,2023(2):52-55.
- [8] 陈海玉,王聪,陈雨,等.民间历史文献知识图谱构建:以徽州文书为例[J].图书馆论坛,2022,42(11):141-150.
- [9] 徐彤阳,黄映思.名人年谱资源的知识图谱构建:以徐朔方《晚明曲家年谱》为例[J].数字图书馆论坛,2022(12):29-36.
- [10] 曾桢,赵浩宇.基于文献的中国近代史知识图谱构建与实证研究[J].数字图书馆论坛,2022(4):35-42.
- [11] 郭嘉欣.基于多源异构数据挖掘的“红色记忆”知识图谱构建[J].知识管理论坛,2020,5(1):59-68.
- [12] 陈玖瑜.数字人文视阈下民国报纸知识图谱构建研究[D].长春:吉林大学,2020.
- [13] 严承希,李少建,胡恒.面向知识服务的清代档案文献知识聚合[J].档案学通讯,2023(3):54-62.
- [14] 邓君,王阮.数字人文视域下口述历史档案资源知识发现模型构建[J].档案学研究,2022(1):110-116.
- [15] 中华人民共和国国家档案局.这里的档案很特别[EB/OL]. [2023-05-03]. <https://www.saac.gov.cn/daj/yaow/201002/9a1871f-bc29045c997018466ecb26f46.shtml>.
- [16] 中国南海研究院.文献资料[EB/OL]. [2023-04-08]. <http://www.nanhai.org.cn/bookdata.html>.
- [17] 厦门大学图书馆.东南海疆研究数据库[EB/OL]. [2023-05-08]. <http://210.34.4.13:8080/asia>.
- [18] 中国海洋大学图书馆.海大文库[EB/OL]. [2023-04-26]. <http://library.ouc.edu.cn/Home/Detail/2>.
- [19] 李敏,刘应平.“南海及南海诸岛”特色数据库建设研究[J].图书馆论坛,2015,35(5):89-93,99.
- [20] 许丰满.发掘海洋文献资源 做好图书馆特色服务:以深圳市盐田区图书馆为例[J].图书馆界,2015(4):73-75,94.
- [21] 吴晓文,孙杰,杨帆.大数据背景下海洋档案数据挖掘研究[J].办公室业务,2017(19):186-187.
- [22] 李慧颖,林凤君.色彩管理在海洋档案仿真复制中的作用[J].海洋信息,2010(1):11-12.
- [23] 刘家真,薛惠芬.国家海洋档案信息管理与国家海洋数字档案馆[J].海洋信息,2014(1):37-40,45.
- [24] 方泉,薛惠芬.近代海洋档案略谈(上)[J].海洋信息,2014(4):6-12.
- [25] 宋文.知识组织体系语义互操作研究[J].图书馆论坛,2012,32(6):117-121.
- [26] 刘爱琴,刘洁,尚珊.基于知识集成的LAM数字资源整合模型设计[J].图书馆学研究,2018(4):73-79.
- [27] 梅海燕.元数据的研究进展[J].现代图书情报技术,2002(4):17-19,53.
- [28] 许鑫,张悦悦.非遗数字资源的元数据规范与应用研究[J].图书情报工作,2014,58(21):13-20,34.
- [29] 贾琼,王萍.基于关联数据的历史档案资源聚合研究[J].图书情报工作,2021,65(10):105-112.
- [30] 李永卉,周树斌,周宇婷,等.基于图数据库Neo4j的宋代镇江诗词知识图谱构建研究[J].大学图书馆学报,2021,39(2):52-61.

作者简介

华林, 男, 博士, 教授, 研究方向: 历史档案、民族档案。

冯安仪, 女, 硕士研究生, 通信作者, 研究方向: 海洋历史文献、知识管理, E-mail: 1418662104@qq.com。

谭雨琦, 女, 硕士研究生, 研究方向: 海洋历史文献。

Knowledge Organization of Marine Historical Literature Resources in China Under the Context of Linked Data

HUA Lin FENG AnYi TAN YuQi

(School of History and Archives, Yunnan University, Kunming 650091, P. R. China)

Abstract: There are a large number of marine historical documents, which are complex and scattered. The phenomenon of resource heterogeneity is obvious, and the existing methods of document resource construction are difficult to achieve deep correlation and effective exploration and utilization of resources. This article takes marine historical literature resources as the research object. Firstly, based on the analysis of the current situation of literature resource construction, it summarizes the problems of theoretical cognition, resource heterogeneity, and knowledge correlation in the construction of marine historical literature resources. Secondly, a framework system for marine historical literature resources is constructed using five modules: text literature, historical maps, epigraphy literature, oral materials, and audio-visual materials. Based on the characteristics of the literature, metadata structure is designed, and the Protégé tool is used to construct an ontology model for Chinese marine historical literature resources. Finally, in the form of instance visualization, the rationality of the ontology model is verified, which reveals the correlation between time, location, people, and resource elements in marine historical literature and specific events, and provides a solution for better development and utilization of marine historical literature.

Keywords: Marine Historical Literature; Ontology Construction; Linked Data; Knowledge Organization

(责任编辑: 王玮)