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*

李宝玲，李 珂，郭立鑫

摘 要：本文在数字人文视角下，对历史档案专题知识库相关概念及理论进行解析，以中福公司历史档案数据库为基础，通过引入本体思想，完善元数据分类及关联规则，提出历史档案专题知识库的构建原则、方法、功能和展示方式，为面向深度利用的历史档案知识服务做出了有益探索。

关键词：历史档案；档案专题；知识库；本体；数字人文；元数据；中福公司

Abstract: From the perspective of digital humanities, this paper analyzes the relevant concepts and theories of the historical archives thematic knowledge base, and based on the historical archives database of Zhongfu Company, puts forward the construction principles, methods, functions and display methods of the historical archives thematic knowledge base by introducing ontology, improving metadata classification and association rules, and making a beneficial exploration for the in-depth utilization of the historical archives knowledge service.

Keywords: Historical archives; Archival topics; Knowledge base; Noumenon; Digital humanities; Metadata; Zhongfu company

DOI:10.15950/j.cnki.1005-9458.2023.02.009

1 研究现状

本文以“档案、知识库”为关键词组合，利用中国知网和万方中文数据库进行检索，共有351篇国内研究文献，呈现逐年稳中上升趋势。从成果看，研究主要集中在档案知识库的构建模型、构建原则、构建方法、可视化设计等理论层面。徐拥军[1]通过分析文件管理（RM）系统、档案管理（AM）系统、知识管理（KM）系统与OA系统的关系，提出了档案知识管理系统构建的原则和策略；牛力[2]等提出了数字记忆视角下学术名人知识库建设的基本模式，对学术名人知识库的融合、建构与服务具体内涵进行阐释；张斌[3]等构建了基于档案知识库的档案知识服务模型，强调了档案知识检索和呈现能力。此外，还有一些学者基于本体开展了知识库构建理论研究。陆铭[4]基于本体构建了档案馆藏资源语义知识库模型； 孙振嘉[5]等参照 CIDOC-CRM概念模型，以五四运动为例构建了资源对象的本体模型。实践层面，青岛市档案馆历史档案知识库支持多种搜索模式和基于时间域进行知识浏览，中国历代人物传记资料库（CBDB）提供可视化查询、人名检索、地名查询、职官查询、亲属/人际关系查询、社会区分查询、两人社交网络查询 等多种检索，这些研究实践为档案知识库的构建提供了借鉴。

但就实践层面，在数字人文指导下的档案知识库研究和建设实证依然偏少，特别是相关档案知识库标准规范缺位，现有案例不同程度存在数据结构不统一、原始档案资源挖掘层次浅，知识展现用户体验差等问题，难以满足档案知识服务深度利用需求，亟待在后续研究解决。

2 历史档案专题知识库的相关概念

2.1 档案数据库与档案知识库。近年来，随着数字档案馆建设的全面推进，各省市级档案馆都建立起了覆盖馆藏的档案数据库，部分档案馆正在建立各种类型专题档案资源库，方便了档案规范化管理和检索利用。钱毅[6]认为档案数据库在不同阶段的发展中会出现许多称谓，如机读目录数据库、索引数据库，照片档案数据库、全文数据库、多媒体档案数据库等。知识库是由数据库概念发展变化而来，是一种以数据库为基础技术面向某一领域知识进行抽取和序化，通过一些技术手段对析出知识加以组织，与大量推理规则共同以特定存储方式贮存，为用户提供可视化的策略性知识服务系统。孔繁胜[7]认为知识库是合理组织陈述型知识和过程型知识的集合，不但包含了大量的简单事实，还包含了规则和推理。张斌认为档案知识库是一个档案知识系统，档案部门对原始的数字档案进行加工处理，从数字档案全文中提取出具有知识价值的内容，按照适当的知识表示和知识组织方法将其存储进知识库中。可以看出，与专题档案资源库、档案数据库不同，档案知识库不仅包含大量的数字档案资源，还包含资源之外的知识挖掘及推理规则，利用者可以通过档案知识库系统的人机交互界面，精准又迅速地找到自己感兴趣的档案知识。

2.2 历史档案专题知识库的内涵。综上概念，本文所提历史档案专题知识库是指以特定历史档案资源为管理对象，在历史档案数据库的基础上，借鉴本体理论，完善元数据分类，构建语义规则，借助人工智能、数字人文等先进技术，按照一定知识体系进行整序和分析而组织起来的数据库系统。因此，历史档案专题知识库应集历史档案资源管理、后台知识管理、前台知识展示功能为一体，具有专题性、知识性、交互性、共享性、可扩展性等特点。实践中，应充分利用已有档案数据库资源基础，即把专题历史档案的数据化、有序化、叙事化和可视化工作作为研究重点；应自下而上，在构建历史档案资源元数据、分类标准、语义规则等工作基础上构建知识库；应先易后难，即以已有一定研究基础的专题历史档案作为切入点形成方法积累经验为其它专题提供参

考。

3 中福历史档案专题知识库的构建依据

3.1 理论支撑。首先是数字人文理论。数字人文起源于文学与语言学领域，是新型的跨学科研究领域，数字技术的进步及其在科学领域的普及应用促使它的产生与发展。[8]数字人文富有层次化的理论框架与技术体系不断发展，自然语言检索、知识图谱、VR/AR、可视化、AI等新兴技术应用，为历史档案资源深度开发利用带来了无限契机，推动了历史档案资源从“数字化”向“数据化”“知识化”方向转型发展。其次是本体理论。本体是对某一领域内概念类及其类之间关系的形式化表示。[9]本体一词原是哲学领域的一个名词，但当前已经广泛应用于知识工程、系统建模、信息处理、数字图书馆、自然语言理解、语义web等领域。本体通过定义类、属性等要素赋予数据语义关系，对相应知识集合实现细粒度的描述与归纳。[10]借助本体方法构建知识库系统可弥补档案数据库建设中重实体管理、轻知识服务的缺陷，有效地提高知识的可获取性、可互操作性、可共享性、可重用性和可维护性等，可以更好地进行历史档案资源的知识组织及相关研究。最后是知识服务理论。知识服务就是从各种显性和隐性信息资源中，根据对象的需求将知识提炼的过程，是依托资源建设为基础的高级阶段信息服务。[11]知识服务是基于知识管理的一种新的服务形态，知识管理是信息管理发展的新阶段，是知识发现、知识组织、知识利用的过程，它同信息管理不同，要求把信息与知识、信息与活动、信息与人连接起来，知识服务提供者针对不同类型用户多样化知识需求，围绕自身所拥有的显性知识与隐性知识，提供快速知识服务。

3.2 技术支撑。一是复用本体模型，CIDOC-CRM是一套应用于文化遗产的信息集成概念参考模型，支持图书馆、博物馆、档案馆等不同领域不同类型的专业研究，已有 81 个类、160 个属性。DC是成熟的描述数字文献的通用元数据标准框架，包含了DCMI术语和应用纲要，包含了15个核心元素集。FOAF是一种遵循W3C体系标准的资源描述框架（RDF）词表，用于描述人、人群、人的活动的特定属性及人与人、人与物间的社会网络关系。这3种成熟本体各有特点互为补充，档案专题知识库的构建在档案资源库的基础上，吸收成熟本体的部分元素，以此弥补原档案资源数据库的不足。二是自然语言处理技术。自然语言处理是档案知识库的核心应用技术，通过自然语言接口，用户在查询知识库内容时可以利用自然语言式的文本精确定义自己的知识需求；通过文档自动处理，使用NLP工具对词、短语和句子进行分析可以得出词、短语和句子之间的逻辑关系；通过知识自动获取，可以对档案资源库进行库数据挖掘进行知识抽取。尤其是自然语言处理的知识图谱构建应用，能根据不同逻辑实现知识的相互关联和图谱化输出。三是信息可视化技术。基于H5的虚拟现实、现实增强、地理信息系统等信息可视化技术在知识服务中逐渐兴起，使得知识库更具有人文关怀，它提供多重感官体验，支持交互式操作，增加服务对象的自由度。中福公司历史档案知识库除在线知识检索外，搭建虚拟展厅，用叙事方式和可视化技术，展示多个历史主题，令公众有穿越历史的真实体验。

3.3 资源支撑。历史档案专题知识库选择中福公司历史档案全宗为研究对象，具有四个特点：一是中福公司历史档案较其他全宗历史档案，内容更丰富、保存更完整，它形成于1897到1956年间的档案有4485卷，具有时间跨度大、形成主体多元、门类齐全、载体多样、领域宽泛，史料价值高的特点；二是中福公司历史档案依据《民国历史档案著录规则》结合中福公司档案特点，制定了《中福公司档案著录细则》，进行了数字化的整理和开发，形成了标准化目录数据库和全文数据库，析出了中福公司档案的主题和关键要素，为知识库构建打下良好的数据基础；三是中福公司历史档案在社会上成为研究热点，产生了一系列中福公司档案研究成果、文史资料、翻译作品、科研论文等知识产品，丰富了知识库的来源；四是中福公司档案内容丰富，涵盖了政治、经济、文化、工业、教育等方方面面，是河南近代工业发展的缩影，便于与馆藏其他档案进行知识关联。

4 中福历史档案专题知识库的元数据体系

4.1 元数据项的设计需求。梁继红[12]对走向文本的历史档案数字整理研究提出，历史档案数字整理包括了文本阅读的基底层，元数据搭建的桥梁层，文本内数据化的加工层，以及数据分析和可视化的知识发现层。可见，元数据是历史档案知识库构建的重要基础，元数据提供了知识的提取、聚类、关联，使得历史档案实体能够在数字空间中呈现多重脉络。前期中福公司历史档案实现了目录和内容层面的数字化，虽然按照《中福公司档案著录细则》进行了档案形式和内容元数据项著录，但是元数据是以资源管理为导向，是独立的、分散的，缺少语义关联，难以实现知识管理，不能完全满足专题知识库知识服务的需求。因此，需要在原有元数据基础上进行优化，构建能够体现语义特征的元数据方案。

4.2 元数据体系的设计原则。遵循三个原则：一是复用与自定义相结合原则。历史档案既有一般历史文化遗产档案的共性特点，也具有自身特殊性。因此，参考CIDOC-CRM、DC、FOAF模型对中福公司档案的核心元数据进行标准化描述和定义，选取通用核心指标直接复用，而其它特殊和次生指标通过专家协作进行增补和解释。二是有利于知识挖掘和关联原则。知识库建设大致分为自顶向下的基于本体和自底向上的基于人工智能两种表示方法，而历史档案知识库建设以挖掘隐性知识和深度利用为直接目的，应发挥两类方法各自优势，即在本体思想指导下进行元数据体系设计，并在此基础上通过人工智能技术支撑，进行实例抽取和知识关联，达到知识的深度挖掘。比如，针对中福公司历史档案特点，细化主题类目，规范定义每一个类目属性并辅助以同义词、近义词词表；在每一件文献著录主题词的基础上，增加所属一级主题类目、二级主题类目。三是突出历史档案专题特色原则。不同专题历史档案反映了不同历史阶段和专业领域，具有不同的档案类型和内容。在元数据体系设计上要考虑专题档案资源特点，体现出研究对象的特色。比如在对“事件”的界定上，既包括发生在这一历史时期的历史事件，也

包括中福公司机构变化、人事任免、重要会议、煤矿、安全事故等公司大事。

4.3 基于本体的元数据体系构建。历史档案专题知识库以“一站、两库、多专题”为基本结构，“一站”指历史档案知识服务网站，“两库”指专题资源库和专题知识库，“多专题”指不同的专题资源所对应的不同专题模块，不同专题知识库依照该专题档案的存量与整理情况具有相同模型和不同元数据项。这里我们以中福历史档案为例，引入本体思想，从历史档案资源的资源管理层和资源内容层分别进行分析，资源管理层的本体类目主要描述中福公司档案的形式特征和过程特征，资源内容层的本体类目设计旨在对中福公司档案内容进行多维度描述，为知识发现、挖掘和利用打下基础。

中福公司档案本体共160个类，其中包含11个一级类目，79个二级类，70个三级类，资源管理层面有“档案外形特征、数字化资源、档案类型”3个一级类目，资源内容层面有“人物、时间、主题、语种、事件、地点、责任者、文种”8个一级类目。其中，“主题”“事件”“地点”“时间”复用CIDOC-CRM，“人物”复用FOAF，“文献类型”复用DC，其余类目为自建而成。

在类目和层级关系上，可以看出该体系弥补了传统资源管理模式中以管理一级类为导向的不足，进而增加大量内容层面类目。其中，档案文献的外形特征包括该档案的档号、题名、目录号、页码等信息；“数字化资源”指对档案实体和相关资料实体进行数字化后形成的数字化副本；文献类型主要指文书、照片、音视频、人事、会计、科技、图书资料、实物等档案实体类型；“事件”主要指由行为主体设计和执行的具有一定预期目标的事情，主要指历史或公司的一些大事件及要事；“主题”主要描述中福公司历史档案的15个方面的特色内容，分别为矿案、教育、矿产、民窑、交通运输、矿警、抗战、工人运动、经营管理、行政管理、生产管理、外事、外贸、日记日志、医疗等；“语种”主要指档案文献的语言种类，包括中文、英文及其它语言；“地点”是中福公司或人物对象活动发生时所存在的位置空间，例如北京、河南、四川、湖北等主要地点；“时间”是指中福公司各类事件和行为发生所形成的阶段式时间范围；文种主要包括令、信函、电报、章程、票据、日记、报表、凭证、报告、呈等。

在对象属性上，中福历史档案本体中除了上下级关系的“包含”关系外，还涉及实体与内容间关系、事件情境关系、资源对象间关系、内容间关系、行为主体间关系、时间等诸多关系。比如，实体与内容间的记录关系，事件与人物、地点、时间之间的“谁参与事件”“事件发生时段”“事件发生地点”“子事件”“属于哪个主题”等关系。本文放弃通过描述对象属性进行语义关联的方法，采用人工智能语义关联方法，对上述人物、地点、时间、事件、责任者等概念进行元数据体系优化，实现知识的关联。

完成上述类目后，参考《中国档案主题词表》《民国档案分类主题词表》，按照5%进行等间隔抽样标注，人工标注档案近9000件，包含635个主题类属词、212个文种类属词等，再将提取到的元数据类属词表植入到自然语言处理语料词库，通过机器识别聚类，机器标注的关键词约17000个，包含16824个人名、953个地名和177个事件，以此构建中福公司档案的知识体系。

5 历史档案专题知识库的平台建设

5.1 建设框架。历史档案专题知识库以“一站、两库、多专题”为建设总体目标，借助本体元数据体系设计，通过智能化数据挖掘和抽取进行知识关联，最终以可视化形式流向利用者。建设框架划分为四个模块：专题资源层、技术融合层、知识组织层、展示应用层，如图1所示。

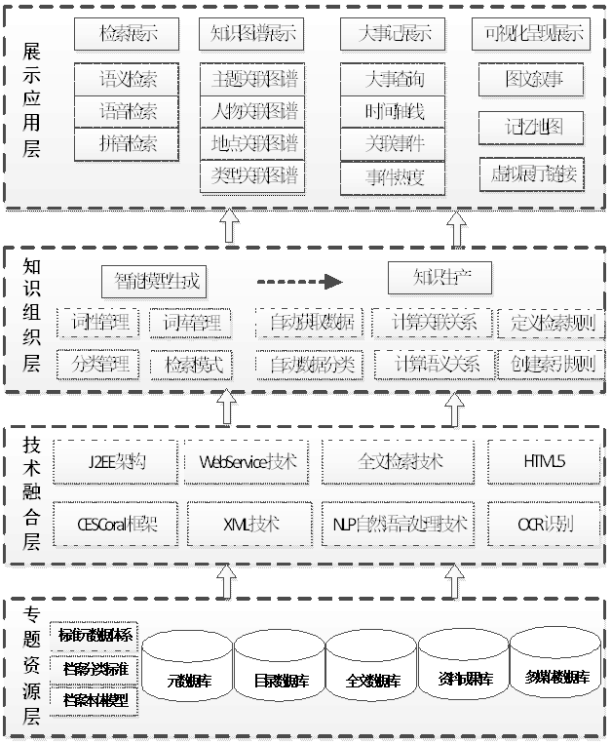

图1 历史档案专题知识库平台建设框架

专题资源层由五个数据库组成提供原始数据。元数据库按照基于本体的元数据体系方案，形成结构化的专题元数据库；目录数据库由本馆和外部征集整理产生而成的电子目录，包括181768条基础目录数据；全文数据库由双层PDF挂接而成，这部分以文书和图片档案为主，包含180845条全文数据。资料成果数据库由收集整理并实施数字化的相关研究成果组成，包括著作、论文等；多媒体数据库包括了中福历史档案相关的展览、纪录片、文献片、采访报道等数字资源，以音视频为主。技术融合层，通过选型先进的J2EE架构、CES-Coral框架、WebService技术、XML技术、OCR技术、多媒体技术、全文检索技术、NLP技术、HTML5技术等，为知识组织和展示应用提供强有力的技术支撑。知识组织层首先通过NLP技术自动定义不同种类词性，通过对126万个词汇自动识别，形成了395832个有效词组或短语，而后按照元数据方案，抽取地名、人名、同义词等应用词汇或短语39880个，形成不同类型词库；其次，自动计算应用词语或短语的权重、频次、距离及关联关系和语义关系，形成检索和索引规则；最后进行知识提取，即根据词汇模型进行数

据分类，实现专题档案和资料数据的自动获取。展示应用层按照主题分类、知识图谱、虚拟展厅、图文叙事、时间轴大事记、人物介绍、在线交互等形式进行可视化呈现。

5.2 系统功能。知识库应用平台建设中引入知识工程方法，探索历史档案资源从“卷”“件”深入到内容层面的知识化处理，系统功能上体现后台数据处理的智能化和前台利用的人文化。

后台模块，系统管理包含人员、权限、日志、访问等管理功能，专题数据管理包含档案资源数据接收、基础词库维护、数据挖掘、利用审核等功能，知识库管理提供知识入库审核、知识关联、语义推理、知识生成、知识维护等功能，专题发布提供知识离线数据包生产功能，包括大事记、图文叙事、知识图谱等。为保证档案资源安全，后台部署在局域网，中间通过单项离线摆渡传输方式更新知识包，既保证知识利用的广泛性，又确保系统平台的安全可靠。

前台模块，历史档案专题知识库提供多维度知识服务：主题分类形式，专题知识库系统根据预设的主题分类提供知识查询功能，实现专题档案资源高级检索和主动推送；大事记形式，把与主题相关的知识按日期进行组织排序，展现历史档案涉及的大事要事；图文并茂形式，对图片类历史档案进行标注，挖掘和解读图片档案背后的故事；知识图谱形式，把与主题相关的人物、事件、地点等要素进行逻辑关联，在整个馆藏数据资源库中进行语义分析和逻辑关联匹配相近档案，以图谱组织排序方式展现，并在知识之间标注关联关系；众筹翻译形式，利用众筹方式，借助社会力量，对历史档案中大量英文档案进行在线中文翻译，让利用者更易读懂档案原文，实现档案与用户互动交流；人物介绍形式，借助档案及资料，对中福公司档案涉及的主要历史人物，按时间顺序对其生平进行串联，使用户能够了解主要人物的主要经历和社会活动。虚拟现实形式，对历史档案部分特色场景虚拟化，达到重现历史的逼真效果。同时，前台档案全文展示自动调用通用浏览器，并通过流加载的方式实现边下载边查看的功能，提高用户知识服务体验。其中，知识图谱作为知识库的核心，按照历史档案本体中的类目，在整个馆藏数据资源库中进行关联和语义分析，匹配相似档案，实现知识关联。为面向最广大用户提供最广泛的知识服务，前台部署在互联网，采用统一用户认证机制接入。

6 价值与不足

面向深度利用的中福公司档案知识库建设把资源整合、知识建构、多维呈现作为重点，相较一般专题知识库单一把时间、事件、人物、地点、物件等要素独立建库，知识结构上更综合、更丰富，能有效突出历史档案的知识性和专题性，是对数字技术与人文研究有机融合的有益探索。价值有三：一是资源整合上，采取文本、照片、音视频多类型数字历史资源的采集方式，多元整合汇集馆内外相关的数字资源，实现档案、资料、研究成果等资源间的相互补充与引证，利用数字技术完成历史数字资源的汇总聚合。二是知识构建上，以现有数字档案馆资源库为基础，通过本体构建和数字人文技术，在面向深度利用的数字记忆建构观下，将中福档案及资料中的时间、人物、事件、地点等历史记忆要素转换为类目，形成基于本体的规范化元数据方案，再通过人工智能技术实例化类间关系对知识进行关联，将碎片化记忆转换为叙事型记忆，从而形成完整的历史知识形态；三是呈现展示上，引入大事记、图文并茂、时空地图、知识图谱、虚拟展厅等方式，通过H5多维呈现，有效提升档案文化传播能力，激活历史档案社会价值。

但在知识库构建过程中，也面临着理论和实践研究不够深入，历史档案资源数据化任务艰巨，特别是建设实证依然偏少，没有成熟的市场产品，相关标准规范仍然缺位，人工智能技术快速迭代等问题，影响了历史档案专题知识库的建设质量，有待后续继续完善。

*本文系国家档案局科技项目“面向深度利用的历史档案资源专题知识库构建技术与方法研究”（编号：2021-X-30）阶段性研究成果。

参考文献：

[1]徐拥军.“档案知识管理”系统构建的原则与策略[J].档案学通讯,2009(02):58-62.

[2]牛力,刘慧琳,高晨翔.数字记忆视角下的学术名人知识库研究[J/OL].情报理论与实践.

[3]张斌,郝琦,魏扣.基于档案知识库的档案知识服务研究[J].档案学通讯.

[4]陆铭.基于本体的档案馆藏资源语义知识库构建研究[D].吉林大学,2019.

[5]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)

[6]钱毅.档案数据库建设中存在的问题及解决思路[J],档案学通讯,2006(04)

[7]孔繁胜.知识库系统原理[M].杭州:浙江大学出版社,2000:10.

[8]王晓光.“数字人文”的产生、发展与前沿[M].武汉:武汉大学出版社,2010:5-8.

[9]杨建林.基于本体的文本信息检索研究[J].情报理论与实践,2006(05):598-601.

[10]沈立力,朱蓓琳,姜鹏.基于本体的民国文学专题数据库知识组织研究.

[11]张文静,刘婕,徐永全.知识组织和知识服务的基本理论和基本方法[J].商情,2013(31)

[12]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(01)

(作者单位：河南省档案馆 李宝玲，副馆长；李珂，处长，副研究馆员；郭立鑫，科员，馆员 来稿日期：2022-12-20)