山西档案
*Shanxi Archives*
ISSN 1005-9652,CN 14-1162/G2

# 《山西档案》网络首发论文

题目：　拥有整体性记忆：档案领域数据本体管理论纲
作者：　赵生辉，胡莹
网络首发日期：　2020-09-30
引用格式：　赵生辉，胡莹．拥有整体性记忆：档案领域数据本体管理论纲[J/OL]．山西档案. https://kns.cnki.net/kcms/detail/14.1162.G2.20200929.2125.004.html

网络首发时间：2020-09-30 17:08:05
网络首发地址：https://kns.cnki.net/kcms/detail/14.1162.G2.20200929.2125.004.html

# 拥有整体性记忆：档案领域数据本体管理论纲*

赵生辉[1] 胡莹[2]

（1. 西藏民族大学管理学院 咸阳 71082；2. 云南大学历史与档案学院 昆明 650091）

**摘 要：**[目的/意义]应对外部技术环境和内部管理范式转型的双重挑战，探索具有整体性特征的档案信息资源管理新模式，是当前我国档案信息化研究的重大任务。[方法/过程]本文在回顾我国电子文件管理研究历程，分析“文档态电子文件”管理技术瓶颈的基础上，提出了体现整体性治理理念的“档案领域数据本体”概念，并对其术语内涵、生成机理、管理需求、组织机制、生命周期和技术支持等问题进行了框架性梳理。[结果/结论]档案领域数据本体（Archival Data Ontologies，ADO）是按照规范化的流程和方法从大量文档态电子文件内容和元数据中抽取并经过重组的，用来模拟和反映社会历史领域各类实体属性之间的语义关系及其运动变化过程的大规模关联数据集。档案领域数据本体管理要围绕“真实性、完整性和表征性”的保障需求，建立科学的机制、平台和流程，为社会记忆问答等智能化档案利用形式提供数据基础设施支撑。本文的研究结论对于人工智能时代档案信息服务智能化与可信社会记忆基础设施建设具有重要的理论参考价值。

**关键词：**电子文件管理；本体工程；档案数据化；整体性治理；

**中图分类号：**G270.7 **文献标识码：**A

## 1 研究背景

### 1.1 我国电子文件管理研究历程回顾

1988 年 1 月，《档案学通讯》杂志刊发了美国学者罗伯特·F·威廉斯的《电子文件管理——即将到来的文件管理革命》一文，标志着电子文件管理问题开始得到国内学者的关注[1]。此后十年间，陆续有学者探讨电子文件管理相关议题，安小米[2]、冯惠玲[3]、刘家真[4]、王健[5]等学者在 1998 年前后所做的研究工作，奠定了我国早期电子文件管理研究的学术基础。其中，以冯惠玲教授博士论文《拥有新记忆：电子文件管理研究》为基础形成的电子文件管理新思维系列论文，对于我国电子文件管理基础理论框架和档案记忆观的形成发展都产生了深远影响。“拥有新记忆”在当时的语境下是指拥有数字化形态的社会记忆，即档案载体在由纸质文献向电子文件迁移的过程中，由数字技术的新特性带来的档案管理和利用模式的诸多新变化。进入 21 世纪，我国的电子文件管理研究取得了长足进步，国内众多学者的共同参与，对电子文件技术特性、电子文件生命周期、电子文件“真实性、完整性和可靠性”保障、电子文件管理元数据设计、电子文件管理系统 ERMS 开发等问题进行了深入探索[6]，为信息时代的档案管理提供了诸多可供借鉴的智力成果。电子文件（electronic records）在档案学界是内涵非常丰富的

*【基金项目】本文系国家社科基金项目“多民族语言数字资源语义互联框架研究（项目号：19BTQ004）”的成果之一。

【作者简介】赵生辉（1977-），男，陕西宝鸡人，西藏民族大学管理学院教授，硕士生导师，研究方向：计算档案学、民族信息学、藏学数字人文；胡莹（1981-），女，浙江宁波人，云南大学历史与档案学院副教授，硕士生导师，研究方向：少数民族档案、历史文献学。

术语，电子公文、电子表格、数据库记录、数字图像、数字音频等形式都可能属于电子文件管理的范畴。尽管研究视角非常多元，对电子文件的理解也不局限于文档，但是毋庸置疑的是，无论是在理论研究还是应用实践方面，我国档案学界的绝大多数成果都是默认以“文档（documents）”为主要管理对象，元数据著录、鉴定和归档操作大多是针对文档进行的。为了研究方便，本文将以文档为主要管理对象，以管理元数据著录、维护和检索为主要手段，可以为用户提供数字化档案文档查找和阅读服务的电子文件管理模式称为“文档态电子文件管理”。

### 1.2 文档态电子文件管理的技术瓶颈

2015 年之后，随着新兴信息技术的迅猛发展和应用，我国电子文件管理的外部技术环境和内部管理范式同时发生着非常显著的变化：一方面，国家《促进大数据发展行动纲要》和《新一代人工智能发展规划》的相继出台标志着社会信息化发展战略开始由“数字化”向“智能化”转型，必然对档案管理工作提出新的挑战；另一方面，按照“档案管理范式”理论，我国的档案管理工作经历了“档案史料管理”“档案实体管理”和“档案信息管理”三大范式之后，正在向着“档案知识管理”范式迁移，需要档案工作者从对档案本身的关注转向对档案内容的关注，通过对档案内容信息的深度整合，为用户提供更为便捷和智能的档案信息服务[7]。智能时代的到来，使文档态电子文件管理呈现出一些不能适应新需求的特征：第一，文档态电子文件管理的“碎片化”现象。由于篇幅的限制，每一份文档内容所描述的信息都是客观世界的一些侧面，关于外部世界同一实体对象的各类信息分散保存于不同的文档中。用户要想获得某一实体对象的全方位档案信息，必须查阅大量的档案才能完成。第二，文档态电子文件利用的“低效率”现象。通常情况下，用户查阅档案的目的并不是为了获取档案本身，而是希望找到蕴含在档案内容当中的部分关键性证据信息。为了找到少量证据信息，用户必须阅读大量与其核心需求并不相关的内容，在缺乏专业知识或注意力不集中的情况下，阅读中很容易遗漏关键信息。第三，文档态电子文件检索的“弱智能”现象。电子文件元数据只能为用户查找文档提供支持，由于系统知识库支持不足且不能对文档内容信息进行深度计算，因而无法完成具有智能化特征的复杂档案检索任务，更无法采用人机对话等智能化的档案服务模式。

### 1.3 整体性档案信息资源管理的现实需求

与此同时，全球范围内政府治理模式也在经历较为重大的变化，“整体性治理（Holistic Governance）”成为新公共管理、新公共服务之后的又一个受到理论和实践界广泛关注的热门领域。“整体性治理”是对新公共管理“碎片化”困境做出的战略回应，强调以公民需求为导向，以信息技术为手段，以整合与协同为机制，把治理体系当中的层级、职能、公私关系等碎片化的单元进行整合，实现由局部到整体，由“碎片化”到“一体化”的转变[8]。近年来我国政务云计算平台建设、政务大数据体系建设和“互联网 + 政务服务”都是“整体性治理”理念在电子政务领域的实践。要解决文档态电子文件管理面临的“碎片化、低效率、弱智能”问题，就必须从“整体性治理”视角出发，构建与云计算、大数据、人工智能等新兴信息技术相适应的电子文件管理新模式，使蕴含在海量文档态电子文件当中的内容信息能够实现深度融合，为用户提供“整体性、一站式、智能化”的数字社会记忆服务。基于上述背景，本文提出体现整体性治理理念的“档案领域数据本体”概念，分析其形成机理和管理需求并对其利用模式进行展望，研究结论对于人工智能时代档案管理工作的融合创新和人类数字化社会记忆的构建模式具有有重要的理论参考价值。

## 2 内涵解析

### 2.1 我国档案学界本体研究综述

本体（Ontology）的概念源于哲学领域，可以追溯到古希腊哲学家亚里士多德建立的世界事物分类体系。牛津（Oxford）英语词典将 Ontology 解释为“关于存在的科学或研究”，因而哲学当中的

ontology 与存在相关，是对世界上各类客观存在物及其关系的系统性描述，其复数形式 (ontologies) 可翻译为“本体论”或“存在论”[9]。基于对外部世界结构和规律的共同关注，在 20 世纪后期人工智能研究当中，本体成为连接哲学和信息系统科学的纽带和桥梁。1991 年，美国斯坦福大学知识系统实验室的汤姆·格鲁伯（Tom Gruber）首次将信息系统科学当中的“本体”解释为“构成相关领域词汇的基本术语和关系，以及利用这些术语和关系构成的规定该词汇外延的规则”[10]。此后，经过多位学者的补充和完善，计算机科学、信息系统和人工智能等学科对本体的理解基本趋于一致，即面向特定领域信息应用的明确的、详尽的、形式化的共享概念集，其最典型的应用就是各类基于受控词表建立的叙词本体。从公开文献来看，我国学者从 2007 年开始关注档案与本体之间的关系，大致分为三种类型：第一种，档案本体论研究，主要从哲学本体论视角研究档案现象，代表性学者主要有潘连根[11]、丁海斌[12]等；第二种，档案学领域本体研究，致力于构建档案学领域知识的本体模型，代表性学者主要有王应解、吕元智[13]等；第三种，基于本体的档案管理研究，主要通过档案领域本体词表的构建与应用来辅助档案文献的精准和智能检索，代表性学者主要有李海军[14]、段荣婷[15]等。需要说明的是，国内一些基于语义网（Semantic Web）的档案管理研究论文或专著题名并未出现本体，但是鉴于本体在语义网体系当中的核心地位，这类研究也应属于基于本体的档案管理研究范畴。在上述三种类型的研究当中，基于本体的档案管理研究在数量上占据主体地位。本文是在第三种研究视角基础上的拓展，即以本体为中心的档案管理，本体不仅是辅助检索的工具，也是档案管理的核心对象。

### 2.2 档案领域数据本体的概念界定

由本体理论发展历程可知，信息科学视域早期研究当中的“本体”本质上就是领域可共享关联概念集，致力于通过规范信息系统的数据元素，解决自然语言“一词多义”“语义近似”“歧义”等问题对信息检索系统性能的影响，提高信息检索的精准程度。作为共享概念集的“本体”尽管也是对现实世界的一种模拟，但是这种模拟主要体现在对概念间语义关系的模拟，还没有在实体 (entity) 层面实现对应与关联，因而尚不具备作为社会记忆资源的属性。2012 年 5 月，美国谷歌公司推出知识图谱（Knowledge Graph）技术之后，本体的应用场景发生了显著的变化。知识图谱是一种用节点、属性和边的复杂组合来表示的特定领域语义知识的通用形式化描述框架，其核心原理是美国心理学家奎廉（Quillan M Ross）在 1966 年提出的语义网络（Semantic Network）模型。语义网络是一种通用语义关系连接的概念网络，由相互连接的节点和边构成，节点代表实体或事物，边代表节点与节点之间的语义关系[16]。在知识图谱技术架构当中，本体不仅可以描述概念与概念之间的关系，更为重要的是可以用来描述客观世界当中实体与属性、实体与实体之间的复杂关系。知识图谱中本体的另一个显著优势是通用资源标识符（Universal Resource Identifier,URI），它所有实体赋予互联网当中唯一的路径字符串作为识别符号，从而使客观世界的对象与信息世界的实体之间实现了一一对应，为跨系统的信息整合提供了统一的逻辑参照体系。此外，知识图谱技术框架下的本体与关系型数据库概念设计当中使用的实体关系模型（E-R 模型）完全对应，可以非常方便地将关系数据库当中生成的各类数据记录（record）整合到本体当中，形成关于特定领域各类实体关系的整体性数据图景。基于上述特性，知识图谱架构中的本体就具备了反映和描述人类社会生活关键信息的潜力，成为物理世界各类事物及其关系在信息世界当中的一种“投影（Projection）”。因此，知识图谱技术架构之下的“本体”概念与哲学意义上的“本体论”概念更为接近，本质上是运用语义网络模型对外部世界万事万物及其相关关系的一种结构化模拟，是人类对外部世界认知结果的一种整体性的形式表达，其原理如图 1 所示：

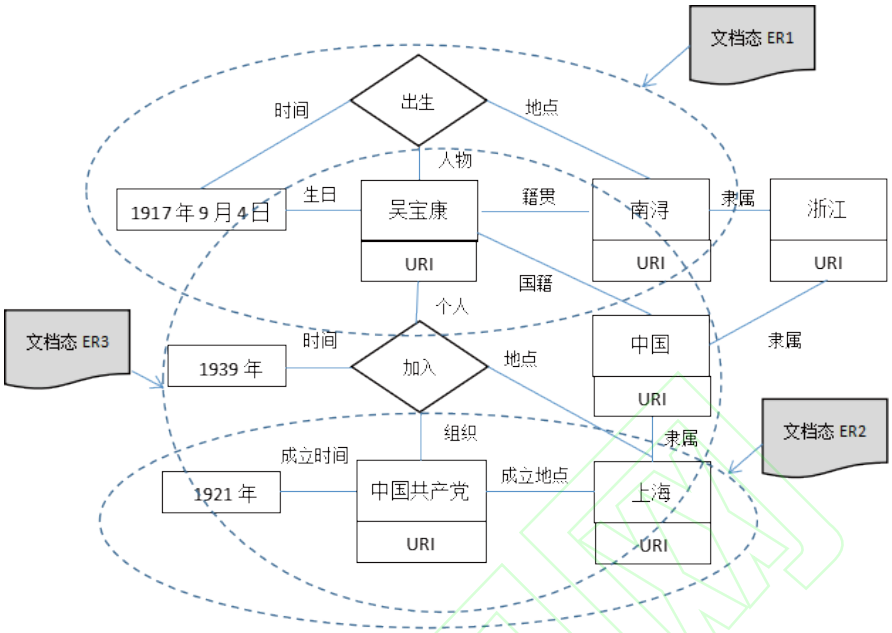


图 1 基于本体的历史领域结构化建模原理

图 1 以新中国档案事业奠基人中国人民大学吴宝康教授档案信息语义整合为例，展示了基于本体的历史领域结构化建模的原理。人类社会活动从来都是整体性的，由于机构和个人认知和记录社会系统的能力有限，在特定时间段所能产生的档案只能是对当时社会活动的某个阶段、某个侧面进行的碎片化的记录。例如，图 1 当中电子文件 1（ER1）主要记录了吴宝康教授的出生信息，电子文件 2(ER2) 记录了中国共产党成立的基本信息，电子文件 3(ER3) 记录了吴宝康教授加入中国共产党的相关信息。此外，部分档案文本当中实际上包含了大量的背景知识，如果不进行形式化表达必然会影响到档案检索的精度。例如，档案显示吴宝康教授的出生地是南浔，其隐含的南浔隶属于浙江这条数据就会在与身份相关的人物检索当中发挥重要作用。本体作为一种进行概念或数据组织的方法，最初是不具备原始记录性的。然而，在构成本体的所有数据都来源于具有原始记录型的文档，每一条数据背后都有可靠的电子或纸质文档作为原始依据，整个本体也就有了社会记录属性，本体当中所聚合的多个来源的数据也可以作为衍生性的聚合态证据使用。用户对本体数据的信任，根源还是在于对原始档案的信任。图 1 当中，各类电子档案当中所蕴含的内容信息最终都汇聚融合于本体，基于统一的逻辑体系实现了整合，构建起与真实历史领域相类似的一种大规模、立体化关联数据结构。由于信息科学视域当中的“本体”需要通过规范化的描述语言编程来实现，天然具有数字化属性。在知识图谱技术架构之下，本体通过从原始档案抽取数据而具有了社会记录属性，其内容信息可以作为证据使用。综上所述，本文将按照规范化的流程和方法从大量文档态电子文件内容和元数据中抽取并经过重组的，用来模拟和反映社会历史领域各类实体属性之间的语义关系及其运动变化过程的大规模关联数据集称为“档案领域数据本体（Archival Data Ontologies，ADO）”。

### 2.3 档案领域数据本体的术语辨析

（1）“档案领域数据本体“与”文档态电子文件”。“档案领域数据本体”是相对“文档态电子文件”而言的，其区别主要体现在以下方面：第一，实现形式不同。档案领域数据本体的实现形式是大规模关联数据集，文档态电子文件的实现形式是文档及其管理元数据；第二，组织形式不同。档案领域数据本体是模拟实现世界实体关系进行资源的整体性组织，内容的语义信息关系是连续的集合形态，能够反映对应社会领域的全貌；文档态电子文件是在社会实践当中直接生成的，语义信息在文档内部连续，文档与文档之间割裂，呈现出“碎片化”现象；第三，社会功能不同。档案领域数据本体的主要功能是供计算机进行智能计算，文档态电子文件的主要功能是供与其相关的个人直接阅读。“档案领域数据本体”

与“文档态电子文件”同时又密不可分：一方面，文档态电子文件是档案领域数据本体的内容来源，档案领域数据本体是由大量文档态电子文件进行知识抽取之后获得的数据集汇聚融合而成；另一方面，档案领域数据本体可以为领域档案管理系统提供公共数据基础设施，在文档态电子文件生成过程中，可以将文档模板与从本体检索的数据相结合自动生成新的文档，因而档案领域数据本体又可能成为文档态电子文件的内容来源。此外，档案领域数据本体和文档态电子文件可以功能互补，当用户需求超出了本体数据集的范畴，或者用户对于不满足于仅看到数据而是希望看到原始档案时，系统可以切换到文档态档案信息服务模式，供用户自行阅读和查找。在文档态电子文件管理模式之下，档案领域数据本体还可以充当与档案主题词表类似的功能，通过对文档态电子文件管理的关联标引，提升文档态电子文件检索的精准程度和智能化程度。因此，“档案领域数据本体”并不是要取代“文档态电子文件”，而是构建两者互补共存的多元档案信息生态。

（2）“档案领域数据本体”与“数据态档案”。“档案数据化”是当前我国档案学研究的热门领域，其理论基础主要是档案管理的“三态两化”学说。“三态两化”学说是由我国学者钱毅在 2018 年提出的一种档案管理对象演化理论，其核心观点是档案管理对象空间与技术环境的变迁相适应，总体上沿着“模拟态”“数字态”和“数据态”的路径演进，链接“三态”的是两大档案加工任务，即“档案数字化”和“档案数据化”，前者实现模拟态档案向数字态档案的迁移，后者实现数字态档案向数据态档案的迁移[17]。“数据态档案”即以结构化数据集形式存在的档案资源，目前在实践中主要有三种类型：第一，各类业务部门当中，支撑办公自动化系统运行的业务数据库；第二，文档态电子文件管理系统当中用来描述文档背景、结构和内容信息的元数据集；第三，由政府大数据管理部门或类似职能机构负责建立的旨在实现跨部门业务数据整合的政务大数据集。“档案领域数据本体”则是在档案管理部门主导下，从构建数字化社会记忆的需求出发，通过逆向数据抽取的方法所构建起的大规模、整合态关联数据集，属于“数据态档案”的新形式。因此，“数据态档案”的内涵较为丰富，“档案领域数据本体”属于“数据态档案”的一种新类型。

（3）“档案领域数据本体”与“政务大数据集”。“档案领域数据本体”与政府大数据部门构建的政务大数据集覆盖范围类似，但是两者建设思路和建设方法各不相同。政务大数据集是直接实各业务部门数据库的整合，建设目的是为了驱动跨部门的协同政务运转，对实时数据的关注超过通常会超过对历史数据的关注；“档案领域数据本体”则是由档案行政部门推动建设，从构建区域数字化社会记忆的需求出发进行结构设计，以为用户提供智能化档案信息服务为目标，通常情况下需要以文档态电子文件中抽取的事实型数据为基础进行创建。由于两者覆盖范围的重合性，实践当中也可以将“档案领域数据本体”作为“政务大数据集”的辅助和备份资源来使用。同时，为了提高档案领域数据本体的建模效率，政府大数据管理部门和档案管理部门也可通过协作共赢机制，通过交换获得各自需要的数据集。

（4）“档案领域数据本体”与“数字孪生模型”。数字孪生（digital twins）是近年来随着全球范围内智慧城市建设的兴起而受到广泛关注的研究领域。数字孪生是指利用现代信息技术，对物理世界当中的实体对象进行多维度综合仿真，从而在虚拟空间构建起与实体对象类似的虚拟对象，以此来实现对物理世界各类实体的感知、管理和控制[18]。“档案领域数据本体”和“数字孪生模型”都是整合形态的数字信息资源集合体，但是两者在应用场合和设计思路上还是有所不同：第一，“档案领域数据本体”的建模对象是社会历史领域，即社会组织或个人社会实践；“数字孪生模型”主要建模对象是自然科学领域，例如车辆、航空器、作业设备等。第二，“档案领域数据本体”的建模目的是反映社会活动的历史进程，并不一定要求数据状态与外部世界实体状态保持同步；“数字孪生模型”的建模目的保持系统数据状态与外部世界基本同步，因而现有数据是其关注的重点，对各实体数据的演变过程关注相对较少。在科技档案领域，“档案领域数据本体”与“数字孪生模型”可以合二为一，基于建筑信息模型（Building Information Model,BIM）的城建档案信息管理就是围绕数字孪生模型进行管理，从设计、施工、验收到运行维护全生命周期的信息都可以存入模型，档案利用不再需要翻阅二维图纸，而是直接

从三维建筑模型查找所需的关键性数据。

# 3 管理框架

## 3.1 档案领域数据本体的生成机理

档案领域数据本体是在档案数字化工作的基础上，针对文档态电子文件管理的不足而演化出的一种新的档案信息服务模式，其本质是是对大量文档态电子文件语义内容的整体性融合建模，如图 2 所示。

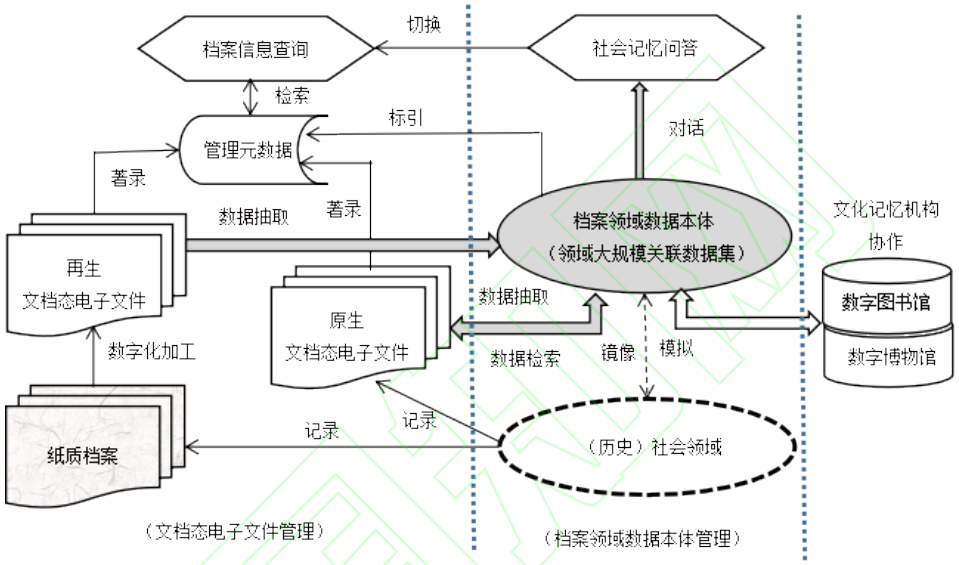


图 2 档案领域数据本体的生成机理

图 2 描述了我国档案信息化建设由“文档态电子文件管理”向“档案领域数据本体管理”演化的过程。机构或个人在社会实践活动中有意识或无意识地产生了各种类型的文件，其内容是对当时所处的社会环境和参与社会活动的各类主体相关信息的客观反映，本质上是大量分别描述整体性社会侧面的“信息碎片”。在文档态电子文件管理过程中，首先是通过数字化加工将纸质文档转换为电子文档并对相关信息进行元数据著录。随着信息系统的普及，社会生活中持续产生大量原生态的电子文档，其元数据结构在生成之初就已形成。无论是再生电子文件还是原生电子文件，最终目的都是通过元数据的集成，供用户进行档案信息检索。因此，“文档态电子文件管理”主要解决的是档案的分散保存和用户的集中利用之间的矛盾，可以让用户在同一地点访问到存储在不同地理位置的档案信息资源。然而，查找到档案之后，用户必读逐一阅读才能确认是否包含要查找的关键性证据信息。在“档案领域数据本体管理”模式之下，蕴含在文档态电子文件当中的内容信息经过抽取操作被描述为规范化格式的数据，并添加到档案领域数据本体当中。如果某一实体已经创建，则后续数据抽取的相关数据直接连接到之前的实体标识符当中。经过多轮数据抽取和添加操作，原本被分散保存于不同文档当中的信息又重新聚合起来，越来越接近实体和属性在历史社会领域当中最初的状态。与此同时，如果图书馆、博物馆等公共文化记忆机构收藏有与建模主题相关的藏品，为了本体数据集的完整性，档案管理部门可以通过建立数字化协作机制，将图书、报告、文物所蕴藏的历史信息也录入本体。当汇聚和融合的数据量足够大的时候，档案领域数据本体就相当于建立起了历史社会领域的“镜像模型”，通过对档案领域数据本体的访问就能获知社会历史领域各类实体属性关系的全貌。

“档案领域数据本体”并不是凭空出现的，它是建立在档案数字化和文档态电子文件管理工作的基础上，旨在实现文档内容信息深度融合的整体性档案信息资源，是智能社会人类数字化社会记忆的主要表征形式，也是支撑智能社会运行的数据基础设施。有了档案领域数据本体所提供的领域知识库，计算

机就可以在自然语言处理技术的支持下，通过人机智能对话方式与用户进行交互。用户只需要用自然语言向计算机提出社会记忆领域的问题，计算机就可以直接向其反馈问题的答案，不再需要翻阅大量案卷进行查找。例如，用户需要查证新中国第一个档案学本科专业的创办时间，使用自然语言问句提问之后，计算机将通过语音或文字直接反馈该问题的答案“1952 年 10 月”。这种以人机问答为特征的交互模式与人际交流模式类似，用户在不需要掌握任何检索技术的情况下就可以轻松完成档案查阅。此外，问答系统可以大幅度减少用户用于查找和阅读档案文献所耗费的时间和精力，甚至可以完成一些需要通过分析和推理才能找到答案的复杂性查档任务。例如，用户向计算机提问“新中国档案学开创人吴宝康教授曾经担任过国内哪些高校的兼职教授？”，在人机对话模式下计算机将直接反馈答案“南京大学、苏州大学、上海大学、空军政治学院”[19]。上述模式当中，问答系统之所以可以回答用户的问题，得益于服务系统后台对领域知识的结构化存储，而“档案领域数据本体”正是这种可以实现领域知识大规模、结构化存储的基础性资源。

### 3.2 档案领域数据本体管理的功能需求

档案领域数据本体是由大量文档态电子文件的内容和背景信息汇聚而成，其管理需求与文档态电子文件管理既有相似之处，同时因为结构和形式的差异性在部分方面有所不同，其核心功能需求主要有以下方面：

（1）档案领域数据本体的真实性保障需求。真实性（Authenticity）是电子文件管理的核心目标，也是电子文件可以被作为证据使用的主要原因。文档态电子文件的真实性是指文件内容、逻辑结构和背景信息经过传输、迁移等处理后依然保持不变，与形成时的原始状态一致[20]。文档态电子文件的真实性保障旨在通过各类措施确保核心内容的原始性，对电子文件内容信息的关注相对较少。档案领域数据本体需要从文档态电子文件当中进行数据抽取，因而其关注的重点是电子文件的内容，默认为其原始性已经实现。因此，档案领域数据本体的真实性是建立在文档态电子文件真实性的基础之上的，其重点是确保从文档中抽取的数据集能够最大限度体现文档作者的原意。档案领域数据本体真实性保障的困难来源于多个方面：第一，由于自然语言和结构化数据生成原理的差异性，通常情况下只有部分关键信息可以抽取成为结构化数据，由文档态电子文件向档案领域数据本体转化的过程中，必然存在“语义损耗”现象。第二，对文档语义信息的解读能力决定着所抽取出的结构化数据集的质量，通常情况下计算机只能完成一些事实性信息的自动抽取，人工抽取效率低且受到抽取人本身知识结构和偏好的影响，要求从事档案数据抽取的人员必须具备相应领域的知识积累。第三，同一实体对象在不同的文档当中的描述可能是不一致的，而这些文档本身都具有真实性，此时就需要通过可靠的机制从中评审和认定，尽最大可能保障录入本体的数据是符合事实的。

（2）档案领域数据本体的完整性保障需求。完整性（Integrity）是指电子文件作为证据使用时所要求具备的各类要素都得到妥善保存。文档态电子文件的的完整性通常包含三个方面的要求：一是内容、结构和背景信息没有缺损；二是与某项社会实践活动相联系的各类电子文件数量齐全；三是电子文件相关辅助软件程序及其参数说明等数量齐全。档案领域数据本体的完整性是相对建模的社会历史领域而言的，是指本体数据集应当覆盖领域绝大多数的社会机构、个人及社会活动，以最大化还原社会历史原貌。档案领域数据本体完整性保障的挑战主要有以下方面：第一，档案领域数据本体是从大量文档态电子文件当中抽取的数据，因而数据集当中所包括的机构、个人和社会活动实际上是由文档态电子文件所决定的，数据范围本身就在档案鉴定环节经历了筛选；第二，由于社会领域的连续性，本体建模的边界是不清晰的，数据完整性的判定缺乏清晰和统一的标准，例如以某级综合档案馆的馆藏为依据进行建模时，与当地党政机关相关的上级机构、下级机构和有业务往来的不相隶属的机构的相关信息也需要进行数据抽取。为了提高档案领域数据本体数据的完整程度，档案机构必要时可以通过与图书馆、博物馆等公共文化机构的合作丰富本体中的数据内容。当然，与档案数据抽取一样，从图书或者文物资源当中向社会历史领域本体抽取数据时，也必须做到查之有据，每一条数据都必须有相应的具有可信度的图书或文物

作为证据支撑。

（3）档案领域数据本体的表征性保障需求。表征性（Representativeness）即相似性，是本体与社会历史领域结构和规律的相似程度，表征度越高说明两者越相似。档案领域数据本体是以大规模关联数据集方式对社会历史领域的一种结构化模拟，本体与历史之间能够实现良好拟合是建模的最终目标。尽管档案领域数据本体不需要像“数字孪生模型”一样与现实系统做到实时对应，但是数据集内部结构也要符合实现世界当中各类实体之间的互动关系，各类模块之间逻辑上保持一致，从中检索和分析出的结果符合社会活动演进的客观规律。档案领域数据本体表征度保障面临的挑战主要来自于语义网络模型的复杂性，当本体涉及的实体数量增加到一定的量级之后，每新增一个实体都意味着与与原有实体之间存在复杂关联，建模出错的可能性就会增大。此外，实体与实体之间存在众多隐含的语义关系，人工建模时通常难以完整识别。 要提高档案领域数据本体的表征度，一是要录入尽可能多的实体，二是尽可能要对实体之间的关系进行完整性描述。为此，在技术上可以借助本体建模工具的自动语义推理功能，对本体逻辑一致性和完整性进行判别，以发现人工构建的本体模型在实体或关系方面的逻辑冲突，同时对潜在的语义关系进行自动化补全。

### 3.3 档案领域数据本体管理的组织机制

档案领域数据本体是体现“整体性治理”理念的关联数据集，其管理需要在文档态电子文件管理现有机制的基础上，通过全局性、系统性的规划，建立相应的组织机制来推动落实。为了保障档案领域数据本体的真实性、完整性和表征性，其管理机制设计需要重点考虑以下方面。

（1）档案领域数据本体管理的责任机制。档案领域数据本体的建模对象是特定区域或特定层级的社会组织和个人一定历史时期的社会活动，区域范围或行政层级的设定可以参照我国档案行政机关现有的层级管理体系。考虑到档案领域数据本体本身的特殊性，较为科学的建设路径是由国家档案行政机关总体负责，建立全国统一的“国家社会记忆大数据平台”，各地区、各层级档案行政机关基于同一基础设施，按照各自的权限负责相关模块的建设并访问相应的数据子集。按照“一对多”模式进行档案领域数据本体管理的优势在于每个机构都是在其他机构工作的基础上进行的，可以保证相同实体在本体当中表征的唯一性。

（2）档案领域数据本体管理的合作机制。与文档态电子文件管理关注档案本身有所不同的是，档案领域数据本体管理必须关注档案的内容，并实现基于本体实现档案内容信息的深度融合。因此，本体态档案管理必须建立在档案管理专家、领域专家和技术专家三方协同的基础上。此外，社会记忆资源并不局限于档案部门管理的电子文件，图书馆、博物馆、文化馆等机构的部分藏品也可以视为社会记忆资源，因而档案领域数据本体管理也需要打破机构的限制，通过与其它公共文化服务机构的协作丰富数据集的内容。

（3）档案领域数据本体管理的评议机制。档案领域数据本体管理涉及的数据融合机制面临非常多的挑战，例如工作人员对文本的不同理解，不同档案对同一事件的不同记录，对同一事件的不同语义表征形式等问题都可能造成数据抽取结果的非唯一性，需要由具有广泛代表性的专家组按照合理的程序进行评议，以便确定录入本体的数据内容和数据形式。此外，档案领域数据本体的数据质量也需要由专家组进行评议，只有真实性、完整性、表征度评价达标以后，档案领域数据本体才能作为基础设施为用户提供服务。

（4）档案领域数据本体管理的认证机制。档案领域数据本体的利用的目标模式是人机问答，直接为用户找到所需的关键性数据而不是提供相关文档。在档案领域数据本体的真实性可以通过文档向数据传递，系统向用户反馈的集合态的数据经过档案管理机关签章之后就可以作为整体性证据使用。由档案机构出具的本体数据集检索结果清单具有法力效力，必要时才需要向文档态电子文件甚至纸质档案回溯。

### 3.4 档案领域数据本体管理的生命周期

档案领域数据本体管理既是对文档态电子文件管理的衔接，又是对文档态电子文件的融合创新，其

管理生命周期可以划分为以下流程环节:(1)档案领域数据本体的前端设计。主要涉及基础建模平台构建、社会记忆领域顶层本体框架、数据模板体系设计、属性类型设计等任务；（2）档案领域数据本体的数据抽取。主要是通过人工和自动相结合的方式，从文档态电子文件当中抽取语义数据；通过对外协作方式，从图书馆、博物馆等机构获得部分文献，并进行数据抽取；（3）档案领域数据本体的数据评议。由专家委员会对数据抽取质量进行评价，讨论并提出数据歧义、数据冲突等问题的解决方案；（4）档案领域数据本体的数据融合。按照本体数据模板创建数据实例，将抽取的各类数据录入到本体当中，并对实体关系进行整体性描述；（5）档案领域数据本体的语义推理。按照语义推理规则，对本体进行补全，设计完成支持人机对话的语义推理模块；（6）档案领域数据本体的性能测试。对完成建模的本体数据集进行总体测试，对其与社会历史领域真实轨迹的拟合度进行评价，测试通过才能具有对外提供服务的资格；（7）档案领域数据本体的智能利用。基于问答系统、人机对话等方式，向社会组织和大众提供可信社会记忆智能问答服务；（8）档案领域数据本体的持续进化。通过服务了解用户需求，对档案领域数据本体的性能进行优化和提升。

### 3.5 档案领域数据本体管理的技术支持

（1）“国家社会记忆大数据平台”建设工程。鉴于全国范围内还没有以档案内容整合为目标的信息化建设项目，档案领域数据本体管理可以通过“国家社会记忆大数据平台”建设工程为驱动力量，采用云计算架构，为各地各层级档案部门提供基础设施服务，以保证本体数据集当中的各类实体关系的逻辑一致性。

（2）“自上而下”的本体工程技术路线。本体工程的支撑技术主要有 UNICODE、URI、XML、RDF、RDF Schema、OWL、SPARQL 等，可以按照“自上而下”的思路进行数据建模：首先设计完成领域顶层本体框架（Top Ontology Frame,TOF），再根据顶层本体框架定义类（class）和类层次（class hierarchy），完成数据模板（RDF Schema, RDFS）的设计，最后为数据模板添加大量数据实例（Data Instance）。

（3）多语言本体数据集关联技术。国家社会记忆大数据工程涉及多种语言文字，除了国家通用的汉语和汉字之外，还有数十种少数民族语言文字。在 OWL Protégé 本体建模平台当中，多语言数据集的语义关联主要通过设置等价实体或实体别名等方式实现，确保同一实体对象的不同文字表述拥有唯一的 URI。

## 4 研究结论

我国档案信息资源管理的技术环境和内部范式正在经历非常显著的变化，以“档案知识管理”范式为理论基础，探索具有整体性特征的档案信息资源管理新模式，是当前应对档案智能化利用挑战的迫切需要。本体（Ontology）的概念源于哲学领域的存在论，是对世界上各类客观存在物及其关系的系统性描述，在信息科学领域当中用来描述面向特定领域信息应用的明确的、详尽的、形式化的共享概念集。档案领域数据本体（Archival Data Ontologies，ADO）是按照规范化的流程和方法从大量文档态电子文件内容和元数据中抽取并经过重组的，用来模拟和反映社会历史领域各类实体属性之间的语义关系及其运动变化过程的大规模关联数据集。档案领域数据本体的管理围绕“真实性、完整性和表征性”三大核心功能需求进行，围绕社会历史领域本体数据集的建模、抽取、维护和利用，立需要建立合理的责任机制、合作机制、评议机制和认证机制，实现电子文件管理的重点由“文档态电子文件”向“档案领域数据本体”的迁移。档案领域数据本体管理的基于“国家社会记忆大数据平台”，按照“自上而下”的思路进行本体构建和多语言数据的语义关联，其生命周期可以划分为前端设计、数据抽取、数据评议、数据融合、语义推理、性能测试、智能利用和持续进化等环节。档案领域数据本体是智能社会人类数字化记忆的一种整体性表征形态，其形成、运行和管理的基本规律还有待进一步探索。

## 注释与参考文献

[1] 罗伯特·F·威廉斯．电子文件管理——即将到来的文件管理革命 [J]．档案学通讯，1988(1)：100-103.

[2] 安小米．档案现代化管理面临的挑战——电子文件管理问题 [J]．档案，1997(12)：14-15.

[3] 冯惠玲．认识电子文件——《拥有新记忆：电子文件管理研究》摘要之一 [J]．档案学通讯，1998(1)：5.

[4] 刘家真．传统文件与电子文件形成比较——电子文件与传统文件比较系列论文之一 [J]．四川档案，1998(4)：15-18.

[5] 王健，张宁．敲响电子文件管理的警钟 [J]．档案学通讯，1998(4)：3.

[6] 冯惠玲主编．政府电子文件管理 [M]．北京：中国人民大学出版社，2004(5)：1-4.

[7] 原宜青，丁敬达．论档案知识管理范式的形成与发展 [J]．档案管理，2020(2)：23-26.

[8] 邵岩．运用整体性治理理念推进服务型政府建设 [J]．中国党政干部论坛，2019(9)：71-73.

[9] 冯志勇，李文杰，李晓红．本体论工程及其应用 [M]．北京：清华大学出版社，2007(5)：1-2.

[10] 黄映辉，李冠宇．Ontology 的 Gruber 定义：中文语境理解 [J]．计算机工程与设计，2008(4)：25-30.

[11] 潘连根．要重视档案学基础理论——文件、档案本体的研究 [J]．浙江档案，2007(4)：8-10.

[12] 丁海斌．档案学本体论——兼谈档案学的基本原则 [J]．档案学通讯，2015(6)：14-19.

[13] 王应解，吕元智，聂璐．档案学领域本体构建初探 [J]．档案学通讯，2015(6)：19-25.

[14] 李海军．档案管理信息化之本体论方法讨论 [J]．山西档案，2007(6)：16-18.

[15] 段荣婷．《中国档案主题词表》语义网络化应用研究 [J]．档案学研究，2010(6)：66-70.

[16] 赵军主编．知识图谱 [M]．北京：高等教育出版社，2019(4)：3.

[17] 钱毅．技术变迁环境下档案管理对象空间演化初探 [J]．档案学通讯，2018(2)：10-14.

[18] 徐辉．基于“数字孪生”的智慧城市建设发展思路 [J]．人民论坛，2020(5)：6.

[19] 范学贵．铁血柔情吴宝康——纪年吴宝康诞辰 100 周年 [J]．档案与建设，2017(8)：58-60.

[20] 冯惠玲主编．政府电子文件管理 [M]．北京：中国人民大学出版社，2004(5)：27.

## Having Holistic Memory:An Outline of Archival Data Ontologies Management

ZHAO Sheng-hui[1],HU Ying[2]

(1.Management School of Xizang Minzu University,Xianyang 712082；
2.School of History and Archival Science of Yunnan University,Kunming 650091)

**Abstract:**[Purpose/Significance]It is an significant task of current archival information researching to explore a new mode of archival information resources management with holistic characteristics in order to cope with the dual challenges of external technical environment and internal management paradigm transformation in China.[Method/Process] Based on the review of the electronic records management researching in China as well as the analysis of the current service dilemma of electronic records in document,the conception o f

archival data ontologies with idea of holistic governance is put forward in this paper,then the framework of it's conception content,management requirements,organization system,life-cycle model and technical support are discussed respectively.[Result/Conclusion]It is suggested in this paper that,the archival data ontologies (ADO) is a kind of large-scale linked data set which are extracted from electronic records in documents,and is used to simulate and reflect the semantic relationship between various kinds of entities,properties as well as their movement and changing of social historical domain.The management of archival data ontologies should focus on the guarantee of it's authenticity, integrity and representativeness,establish scientific mechanism,platform and process, and establish a data infrastructure for intelligent serveries of archives such as social memory Question & Answering.The conclusion of this paper is theoretically significant to intelligent services of archives as well as the establishment of trusted social memory infrastructure in artificial intelligence era.

**Keywords:**Electronic Records Management;Ontology Engineer;Archival Data Process;Holistic Governance