知识图谱驱动的科研档案大数据管理系统构建研究*

雷洁[1] 赵瑞雪[1,2] 李思经[3] 鲜国建[1,2] 寇远涛[1,2]
(1. 中国农业科学院农业信息研究所，北京 100081；2. 农业农村部农业大数据重点实验室，北京 100081；
3. 中国农业科学院农业经济与发展研究所，北京 100081)

摘要：为实现档案资源的智能采集、语义化组织与档案大数据的统计分析，探索科研档案智能管理新路径，本文基于应用编辑工具Protégé、建模语言OWL建立的计算机可理解的科研档案知识图谱语义模型，依据档案大数据管理流程，针对多种档案类型的资源进行数据解析、抽取和结构化处理，并对各类实体进行识别、抽取与消歧，以及语义关系的计算、关联、校对、审核等。利用本体、知识图谱、机器学习等语义技术对科研档案核心知识资源进行组织和表示，实现档案智能化采集与归档、档案数据碎片化加工、档案数据自动抽取与智能识别、语义关联以及档案数据审核与发布等功能，推动科研档案智能管理。

关键词：科研档案本体；知识图谱；语义关联；碎片化加工；机器学习；Jena
中图分类号：G270.7
DOI：10.3772/j.issn.1673-2286.2020.02.003

当前我国正快速步入大数据驱动的智能时代，档案大数据是数据价值的富集地。大数据时代下，档案的归档对象、范围和数量远超传统模式，所有具有保存价值的文件、数据、视频、实物都可归档，并可利用数字化智能加工处理技术，将其转化成富含语义知识的“大”档案数据，这也给科研档案的数字化、知识化管理提出了新的挑战。近年来，知识图谱、机器学习等高新技术的快速发展和场景式应用，也为科研档案大数据管理提供了技术支持。如何在知识图谱驱动下构建新型的科研档案大数据管理系统，是值得研究与实践探索的重要课题。

随着社会发展和科技进步，科研手段发生了巨大变化，大规模、跨地域、跨机构的科研活动日益广泛。各种智能传感终端在各领域广泛应用，使得数据能够被快速精准获取，数据来源更加广泛、类型更加多样。在数据驱动环境下，科研档案资源作为重要战略资源，蕴含着组织、个人在科研、基建或生产、管理活动中的研究智慧。科研档案管理的信息化是科研信息化中重要的内容之一，有助于提高科研整体水平和效率，促进科技人才培养，加快成果转化，实现科研信息化跨越式发展。本文在科研信息化形成的平台、标准、规范与数据的基础上，改变传统以件或卷为单位进行档案管理的模式，对科研档案资源数据开展细粒度加工，应用编辑工具Protégé、建模语言OWL建立计算机可理解的科研档案知识图谱语义模型，在模式层将科研档案中的科研项目、科研成果、人员、机构等元素与档案中的相关文件建立关系，基于机器学习、自然语言处理等技术在数据层实现科研档案资源的语义关联，构建科研档案领域智能型大数据管理系统，支撑科技创新与发展。

*本研究得到公益性科研院所基本科研业务费项目“信息所档案数字化管理调研与设计”（编号：JBYW-AII-2017-28）、中国农业科学院科技创新工程项目重点任务“多源数据融合与知识组织增值”（编号：CAAS-ASTIP-2017-AII）及国家科技图书文献中心专项“文献元数据集成加工管理系统”（编号：2018XM34）资助。

1 相关研究

档案管理系统正从信息管理模式向知识管理模式转变，核心技术从数据库存储和XML元数据转向知识库和语义技术。目前国内在档案资源集成与整合方面的研究主要有：吕元智[1]设计了数字档案资源体系的语义互操作和转换框架；张卫东等[2]讨论了资源优先级选择—方法应用—服务集成的档案资源整合路径；梁孟华[3]通过调研中美数字档案资源跨媒体集成现状，基于语义技术构建面向用户的跨媒体知识集成服务平台。关于档案本体实例转换的研究如段荣婷等[4]提出档案著录信息本体标准化方法，并进行了实例转换；李婷[5]利用关联数据将档案文化资源转换为语义集成系统。关于档案语义技术应用的研究如陶水龙[6]利用语义技术分析了档案资源挖掘功能以及档案系统的建设方向。国外主要是对档案资源开展知识表示和数据挖掘应用层面的研究。在各个领域开展了语义技术的应用研究，如Lamharhar等[7]构建了知识库模式应用于电子政务，采用本体技术对数据和服务的描述进行了改进；Marcondes[8]利用关联数据技术对图书馆、档案馆、博物馆的数字资源进行整合；Saravana等[9]采用语义对描述文档进行解析，并构架云服务推荐系统；Bouyerbou等[10]通过设计GEO-MD综合地理本体，减少语义差距，提高自动分类的准确性。在本体模型构建和语义融合等方面，Llanes-Padrón等[11]研究了国际档案理事会关于档案著录标准中的概念模型RiC-CM，并提出一个OWL本体模型；Beniaminov[12]讨论了组织本体库和服务器开发本体及其应用的趋势；Yang等[13]针对知识库融合等问题对注释逻辑、规则结构进行修正以便对数据进行查询。

总体来说，目前国内研究成果多侧重从理论上论证基于语义技术开展档案管理的重要性，集中探讨了档案资源语义整合、语义模型框架构建以及档案系统关联数据转换构建等层面，实践性研究较少。国外主要是围绕特定领域的档案资源进行本体构建和系统应用研究。从国内外进展来看，很少有研究针对档案数据在进行数字化加工、档案目录著录及知识组织管理时，利用知识图谱等语义技术开展档案大数据管理系统性研究。因此，本文通过分析设计档案大数据系统的总体目标、功能框架和技术架构，旨在为新型科研档案管理系统的系统开发提供重要参考，实现科研档案大数据中各类知识单位的细粒度、规范化揭示和语义化组织关联，有力推动科研档案精细化、智能化和现代化治理与服务。

2 面向科研档案管理的知识图谱框架设计

为推动科研档案资源精细化管理，解决科研档案实际管理过程中存在的资源加工程度、关联程度低等问题，开展面向科研档案管理的知识图谱框架设计工作，梳理科研档案特点，结合知识图谱相关技术构建科研档案知识图谱，对档案知识进行抽取、融合与推理，唤醒科研档案这座“沉睡的宝藏”，为新型科研档案的智能管理模型设计提供思路。

2.1 需求分析与概念模型构建

科研档案知识图谱是为科研档案管理这一特定需求而构建的。明确目前科研档案管理工作中亟需解决的问题，梳理与分析管理需求，以指明科研档案知识图谱的构建方向。在大数据时代，数据已成为基础性战略资源，档案的传统形态使得档案价值难以挖掘。目前科研档案管理面临智能化采集、数据碎片化加工以及知识语义关联等需求，亟需新的管理思路和工具来满足，以发挥科研档案数据资源最大价值。在科研档案管理需求分析基础上，对知识图谱的功能模块、层次结构以及相互关系等开展设计与研究。基于已有的档案本体系统和标准，继承较为通用的模型，设计科研档案本体模型，定义核心类、对象属性及数据属性，丰富并规范实体及其语义关系。

本文面向科研机构档案管理智能化、精细化与关联化的需求，结合科研档案特点，开展科研档案管理的知识图谱模式层构建。明确将开展科研活动过程中产生的有价值资源作为科研档案的范围，分析机构、人员、档案、科研项目、科研成果等要素间的关系，梳理与规范构建所需的语言、工具，按照本体构建流程，继承、利用EAD、DCMI、VIVO、SWRC、Schema.org等现有较为通用的本体模型，参考CERIF、Nanopublication等模型框架，构建科研档案的知识图谱概念模型（见图1）。在此基础上，选取本体编辑工具Protégé，使用资源描述框架RDF和WEB本体表示语言OWL，构建一个计算机可理解的科研档案知识图谱模式，包含7个核心类、8个一级对象属性及29个数据属性。

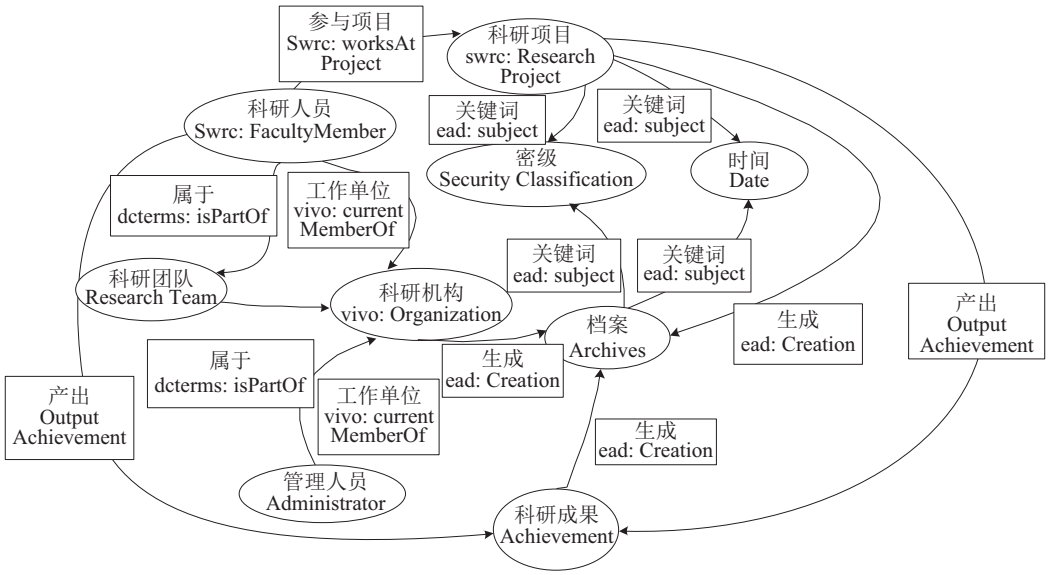

图1 科研档案知识图谱概念模型

知识图谱概念模型的构建是动态循环的。后续科研档案知识图谱数据层的知识抽取过程产生的实体（如具体的某项研究热点、关键技术等），经过综合分析后，将遴选高频词作为实体扩充到档案概念模型中，以进一步修订和完善模式层。在科研档案知识图谱模型构建过程中，采用专家咨询的方法，邀请知识构建领域专家2名、档案管理专家2名对科研档案知识图谱模式的结构合理性和可扩展性两方面开展质量评估。

2.2 知识抽取

在科研档案知识图谱数据层的构建中，结合知识图谱的概念模型，根据档案实体的分类情况，对档案知识的提取过程进行分析。档案知识抽取包含对实体的识别（实体和属性值）和对实体关系的抽取。对实体的识别需要从原始数据中辨别出实体，并划分到对应分类的实体集合中。关系的抽取是要分析两个实体之间是否存在关联以及分析实体之间关联关系的属性。根据档案数据来源和类型的不同，可以将数据划分为非结构化数据、半结构化数据以及结构化数据3个类别。根据不同类型的数据选取不同的模式识别技术、机器学习方法、知识图谱关键技术等进行实体的抽取。根据科研档案资源的数据结构和特点，在进行知识抽取时引入科技类相关字典，结合词性特征和实体组的概念，基于科研档案的特殊固定表达模式构建科研档案知识抽取规则，可提高科研档案知识抽取性能。目前常用的知识抽取模型有CRF模型、BiLSTM模型等。

2.3 知识融合

通过科研档案知识图谱模型的构建和知识抽取，定义了科研档案知识图谱的实体、关系、属性，实现从各类数据中获取实体、属性、关系的目标。由于科研档案数据源分散，在实际操作中，经知识抽取获得的数据会存在歧义等问题。且为增强科研档案知识图谱的可扩展性，丰富知识图谱语义关系，需要与外部知识库或知识图谱（如DBpedia等）进行融合。以上过程都要应用知识融合技术，将来自多个数据源的对同一实体或者概念的不同表达融合起来，提高科研档案数据内在逻辑性和表达能力，形成高质量的知识图谱。知识融合主要包含模式层和数据层，模式层主要包括概念、概念上下位、概念属性的统一，数据层主要是将不同来源的数据进行实体消歧和共指消解。根据科研档案知识图谱中的实体类型，构建特定词典进行语义层面的知识融合，能有效提高科研档案知识图谱知识融合的效果。目前常用的知识融合技术有实体链接、实体对齐、知识合并等。

科研档案知识是不断变化的，一次科研项目的完成不代表此项科研工作的结束。经过知识图谱模式层的构建，以及数据层的知识抽取和知识关联后，需要在已有高质量知识库的基础上，对知识之间隐含的关系进行深层次推理和挖掘，从而使科研档案知识图谱中

的信息不断更新与丰富。

3 基于知识图谱的科研档案管理系统构建

3.1 总体目标

在档案的实体管理阶段，业务流程主要包括收集、整理、鉴定、保管、统计、利用6个环节[14]。随着信息技术的快速发展和广泛应用，档案管理流程涵盖电子文件形成到电子文件保存和利用的全过程，突破了实态档案物理局限，在已有的业务流程基础上增加了检索和编研，发展为8个环节，可直接对档案数字信息进行管理、加工和利用。在大数据、语义技术的驱动下，档案的接收、归档环节可实现数据自动归集；档案管理环节主要功能是信息数据的一体化存储、语义化组织与知识化计算；档案利用环节也能实现智能分析、关联发现与预测功能，为社会创造更多价值。根据档案大数据管理流程，基于知识图谱的科研档案管理系统构建目标：①档案智能采集，基于系统智能化、自动化地开展工作，以实现档案数据的智能捕获、自动标引和智慧归档；②档案语义化组织，针对档案大数据异构且分散的特点，基于知识图谱等技术实现档案的语义化组织，通过对知识单元的抽取、组织并关联，构建数字档案资源本体库，实现数字档案资源之间知识关联、集成与共享；③档案大数据统计分析与利用，基于大数据、知识图谱等技术，完成档案整体信息的统计与分析，实现档案资源的进一步挖掘，绘制档案知识地图，实现档案资源的精细化管理，为领导决策提供更为精准的数据，为科研人员研究工作提供支撑。

3.2 功能框架

为实现构建知识图谱驱动的科研档案管理系统的总体目标，在进行系统功能框架设计时应涵盖科研档案立项、研究、结项、成果申报、成果转化与应用的全生命周期。传统的档案管理系统包括“收”“管”“存”“用”等基础管理功能。知识图谱驱动的科研档案大数据管理系统是面向细粒度和聚合度知识单元的数字档案资源开发，在档案管理全生命周期中更加强调提高档案资源的关联性、层次性和完整性。核心功能包括档案智能化采集与归档、档案数据碎片化加工、档案数据自动抽取与智能识别、档案资源语义增强与关联、档案数据审核与发布。

（1）档案智能化采集与归档。数据采集是档案管理工作的基础。利用智能化技术收集不同数据源产生的信息数据，并提取潜在可用的信息，为知识图谱驱动的档案数据管理提供基础条件。从文档载体形式来看，大多数科研机构采集的档案资源主要包含纸质资源和电子资源两种类型。对于历史存量的纸质档案资源，在数字化之后，需要进行扫描、图像处理、数据挂接等工作才能归档。在数字化扫描过程中，OCR识别是关键技术，识别档案扫描后的文字、图像特征等效果直接决定了后续文件检测效果。基于大数据与深度学习方式相结合的通用文字识别OCR接口可得到较好的印刷体文本图像OCR识别效果[15]。在大数据环境下档案数量激增，各类业务系统复杂且数据关联性强，并呈现出异构分散的特点。档案管理系统基于文档一体化管理的需求，与办公自动化、科研等系统实现数据共享及互联互通，应用智能化技术，对海量档案信息资源进行搜索与智能检测，分析并过滤冗余无效的信息，以满足用户多方位的查档需求。

（2）档案数据碎片化加工。传统的档案管理以卷和件为基础，档案的组织方式只能解决物理及逻辑结构上的异构，语义异构的问题依然存在。为提高档案资源的关联度，满足知识图谱驱动的档案系统精确检索与分类、统计的需求，需要对结构化、半结构化和非结构化的档案数据资源进行碎片细粒度加工。基于《档案分类标引规则》（GB/T 15418—2009），利用大量人工标引的档案分类号实例构成实例库并开展相似度计算，根据计算结果进行档案自动分类。档案的结构分为著录信息和正文信息两大部分。利用人工智能技术、自然语言处理技术，基于科研档案领域知识，通过著录信息（如分类号、保管期限、责任者、题名、页码范围等条目）的元数据智能提取，针对OA系统、科研系统等业务系统中异构元数据需要进行映射、转换与互操作。对于档案正文信息部分，为满足语义知识图谱的构建，依据如时间、地点、事件的模板框架，智能识别句法及章节层次，自动抽取关键数据与图表信息，完成对档案内容的细粒度加工。

（3）档案数据自动抽取与智能识别。面对海量的数字档案，需利用知识图谱等技术来抽取与识别大量的科研知识单元数据，完成语义关联关系的构建，实现数字档案资源之间知识关联、集成与共享。基于前期研

究，笔者构建了科研档案知识图谱模型，且利用Protégé构建了一个计算机可理解与计算的科研档案本体，支持以科研档案为中心的知识单元的集成、关联和融合，主要包括7个核心类、8个一级对象属性及29个数据属性。对档案资源知识图谱数据层构建主要是获取档案资源中实体、关系、属性等RDF三元组。科研数字档案资源的抽取包含档案资源实体抽取、关系抽取和属性抽取。基于深度学习模型等对档案系统中的数据进一步训练和集成，以实现档案知识图谱中各类实体和语义关系的填充，并在数据层实现图谱数据的自动构建。通过对科研档案领域的数据进行实体、关系、属性抽取得到的实验数据，还需要与科研项目、人事信息等进行知识融合，才能进一步完善。利用语义理解、档案知识库中名词解释等进行智能校对，再由人工审核入库，将错误率降到最低。

（4）档案资源语义增强与关联。档案管理长久以来“重藏轻用”的意识限制了档案资源中真正有价值信息的流动和传播。语义技术能促进档案数据的关联与利用。现有的科研档案数据比较复杂，包括元数据记录、知识组织系统和业务系统相关资源等，而且数据之间明显存在异质性。将原有的传统知识组织系统通过语义转换成关联数据格式，才能最大限度地实现档案领域知识的关联[16]。将档案数据转换成关联数据，需要针对本地源数据实行语义转换，发布RDF序列化格式，并构建语义链接。档案资源本体构建包含需求分析、本体设计、数据清洗与RDF数据生成、语义标注与发布等过程。

（5）档案数据审核与发布。科研档案是科研单位极富价值的资产，科研档案资源价值的鉴定与审核是一项专业性工作。基于知识图谱驱动的科研档案系统可以辅助档案管理人员进行数据质量控制与逻辑校验，完成档案鉴定与审核。通过对各类科研档案知识对象、对象属性、数据属性进行统计分析与智能挖掘，形成支撑档案精细化管理的图谱，通过人工弱干预审核校对，进一步优化与完善管理流程。建立图谱数据的语义索引与多类接口发布机制，开放与共享科研成果，为机构的科研创新与决策分析提供支撑。

3.3 技术架构

本文所设计的知识图谱驱动的科研档案大数据管理系统参照国家档案局颁布的档案系统设计与实现功能要求，吸收了知识图谱语义技术的特色，对现有档案信息系统进行改造升级。该系统架构自下而上划分为档案数据层、知识加工层、语义关联层、智能管理层4个部分（见图2）。底层数据（包含各类业务系统数据、纸质档案数字化数据等）基于语义中间件Jena搭建数据模型，以适用于语义知识库系统。语义模型设计上采用Protégé为本体构建工具，语法设计以XML为基础，采用RDF/RDFS，语义设计以OWL为基石。采用API对档案知识本体中类、属性及实例进行解析。在技术上，充分借鉴一些主流的技术标准（如Web Services、Agent等），实现知识图谱驱动的科研档案大数据管理系统的智能采集与归档、加工、自动化抽取与智能识别、语义关联与知识发现等功能，为用户提供智能化档案知识管理服务。

3.3.1 档案数据层

档案数据层主要是为知识图谱驱动的科研档案管理系统提供数据与规则。数字档案资源一部分来自办公自动化系统与科研、人事、财务系统等业务类的数据库，还有一部分来自历史存量纸质档案的回溯数字化加工资源。目前档案数据主要形式有文本档案、音视频档案、档案元数据、XML档案等，储存在数据库中的是结构化数字资源，而日常办公使用Word、PPT、Excel、PDF等数据是非结构化数字档案资源。纸质档案数字化获取的文件是TIFF图片格式，通过OCR识别转换为TXT文本格式。数据资源分散多元且类型各异，半结构化和非结构化档案没有进行统一定义，不利于计算机的处理。因此，在进行智能采集时，要设计并预定义档案数据收集规则，基于OA系统以及科研系统等数据制定相应的元数据方案。依据各档案数据源的接口和地址，利用API等方式对数据进行采集，使用ETL引擎等技术进行数据解析、抽取和结构化处理[17]，通过对海量档案资源的数据清洗、过滤、除噪，以及分词、词性标注等预处理操作，嵌入数据封装、加盖电子签章等技术，实现数据归档智能化、自动化。在档案资源采集和归档过程中增加人工审核模块，提高档案归档准确率。

3.3.2 知识加工层

知识加工层主要是对各类实体的识别、抽取与

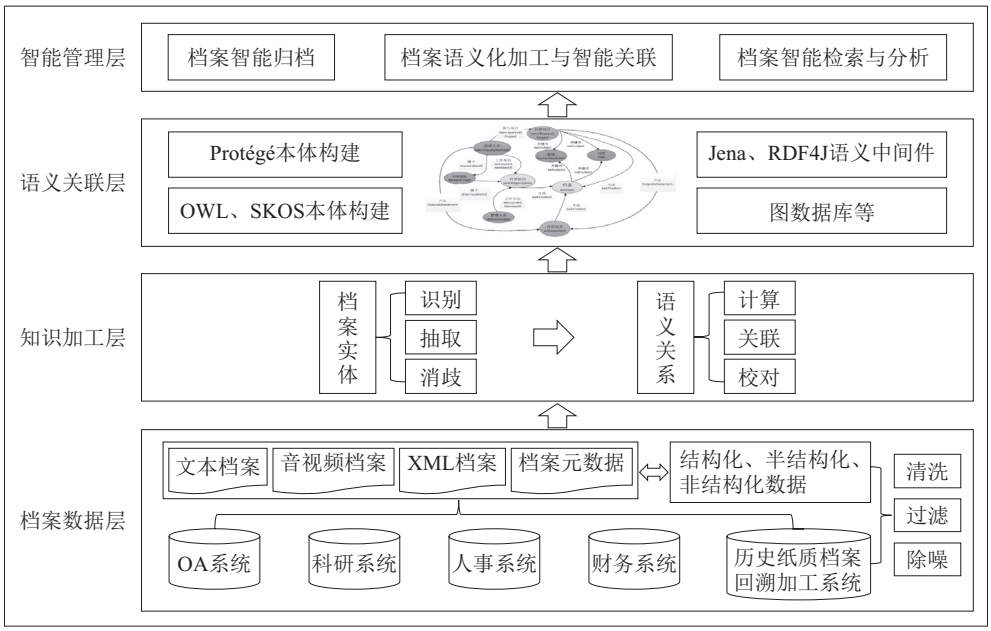

图2 技术架构

消歧，以及语义关系的计算、关联、校对、审核等，为档案知识图谱实例数据自动构建提供支持。科研档案知识图谱实质是揭示科研档案实体关系的复杂网络。为构建科研档案知识图谱本体模型，需从科研档案元数据中提取结构化数据，生成档案（archives）、人员（person）、科研成果（achievement）、密级（security classification）、时间（date）等不同类型的实体，并获取相关实体的属性值。

不同来源的档案数据中责任者可能存在名称相同、机构缩写或别名相同等歧义问题，需要对抽取的档案实体进行对齐和消歧。对不同来源的实体进行筛选，通过邮箱、姓名、所属部门、资助编号、项目名称、年份等信息分别判断责任者、项目实体等，消歧与归并相同实体，添加到知识图谱中，利用科研档案规范库对项目名称、研究方向、机构、密级等关键词进行规范化处理。

利用自然语言处理技术OpenNLP或LTP（Language Technology Platform），从非结构化档案文本中提取如项目任务书、合同书、课题实施方案等实体或概念，发现实体之间的语义关系。通过建立实体间的链接关系（如机构与档案、科研人员与科研成果的隶属关系、科研人员间的合作关系、科研成果间的引用关系等），增加知识图谱中边的密度。目前各类业务系统主要用关系型数据库存储数据，由于关系型数据库不适用于大量数据高效读写，且扩展性较低，使用Virtuoso、Neo4j等图数据库来存储关联数据，易于扩展与查询。三元组是知识图谱的一种通用表示方式，为充分表达与揭示各类数据之间的关系，利用R2RML映射框架将关系型数据库中的科研数据转化为RDF三元组，识别实体、关系及二者间的匹配关系并进行转换、生成和管理，以支撑基于知识图谱的科研档案管理系统语义组织与管理等功能的实现。

3.3.3 语义关联层

利用《中国档案主题词表》及档案领域相关知识与语料，通过本体、知识图谱、机器学习等语义技术对科研档案核心知识资源进行组织和表示，推进档案语义化组织与档案知识资源发现的智能化。基于前期知识加工层对数据的转化，使用自然语言处理和机器学习方法对档案资源中的实体进行概念提取，类及其等级体系的确定，类的对象属性及数据属性的确定，以及本体评价等过程，选取BERT和LSTM等多类深度学习模型[18]，对档案系统中的数据进一步训练和集成，以实现档案知识图谱中各类实体和语义关系的填充，并在数据层实现图谱数据的自动构建。关系获取是知识抽取过程中的关键一步，主要方法有基于规则和基于机器学习两种。由于科研档案资源数据量较大，可利用机器学习方法，基于依存句法分析规则及工具完成档案资源关联关系的抽取。属性抽取的主要任务是获取科研档案文本中实体的属性以及相应的属性值。抽取的属性可能是档案文件

中某个属性名，也可能是档案实体属性三元组，因此可采用基于特征词的词性序列匹配进行抽取。

采用Protégé为本体构建工具自上而下构建科研档案知识本体，综合应用OWL、SKOS等建模语言与Jena、RDF4J等语义中间件进行科研档案知识融合、知识计算、图挖掘和图计算，基于Jena框架自带的OWL本体语言操作接口进行实例添加和实例转换，并可结合开放的关联数据集进行资源内容的语义化增强。可视化作为知识图谱最直接的体现，提供直观的档案数据分析和研究。知识图谱中的“图”更多地是体现在知识组织方面，而不单单是可视化图表中的“图”。根据科研档案知识图谱中已有的知识，使用Jena的推理引擎进行推理，推断出新的科研档案知识关联关系，实现科研档案数据智能更新与管理。运用文本相似度中的余弦相似性等方式判定档案内容具体的相似程度，同时采用实体消岐和知识合并方法对重合或冗余的抽取结果进行档案知识融合。

分析已有档案资源与档案知识组织系统，并结合SKOS、OWL、RDFS、FOAF等已有词表设计档案资源本体抽象模型，与DBpedia、Freebase、基金项目等外部资源关联，创建链接路径，增强用户对档案信息及其背景的理解。从档案文献的元数据中提取人、地、时、事等实体，赋予HTTP URIs，利用统一资源标识符建立相互之间的关联，实现语义互操作与语义知识检索。

3.3.4 智能管理层

在智能管理层进行语义展示，按照科研档案的细粒度管理模式，形成档案智能归档模块、档案语义化加工与智能关联模块、档案智能检索与分析模块，提升档案知识管理效率。利用档案语义化特性，对档案知识库中数据进行深层次地分析、挖掘，构建科研档案知识全景图，为科研档案的知识管理与领导决策提供支撑。

4 实证研究

本实证研究的数据来源于某研究所2013—2019年125项国家级科研项目生成的科研档案（项目具体分类见表1）。档案的数据源是研究所纸质档案数字化回溯系统、办公自动化OA系统、科研人事财务一体化ARP系统，包含综合文书类、科研课题类、财务类及人事类档案资源。其他数据源如CN-DBpedia、维基百科，以及DOI、ISSN、ISBN、ORCID等唯一标识符用于科研档案知识图谱构建中的实体去重及实体对齐等过程。研究所科研管理系统原型的构建主要包括科研档案知识图谱数据的采集、加工、关联、存储，以及应用系统的开发。

表1 科研项目类型分布情况表

科研项目类型	数量/项
国家重大科学研究计划（863计划）	2
国家重点研发计划	19
国家科技支撑计划	2
国家自然科学基金	21
国家社会科学基金	11
中国博士后科学基金	2
其他专项	68

在智能采集归档平台，科研档案系统的数据接入致力于实现统一数据接入和实时接入。利用Web Service对来自纸质档案数字化回溯系统、办公自动化系统、科研人事财务一体化ARP系统及DBpedia等外部资源的数据进行调用，使用统一自动转换和封装成JSON格式的API接口抽取数据。在碎片化加工平台，利用数据解析机实现数据的细粒度标引。在自动识别与语义关联平台，基于科研档案知识图谱构建框架，利用JIEBA中文分词工具、BIO标注方式，基于BiLSTM-CRF模型和BiGRU-Attention模型进行实体识别、语义关系抽取及属性抽取。在语义增强与关联平台，采用实体对齐技术与CN-DBpedia关联实现知识融合。目前系统在知识抽取与知识融合的数据准确度方面还需进一步完善。利用Virtuoso数据库进行存储管理，设置安全技术保障，提供防篡改技术（如入侵检测、数字签名等）保障电子文件安全性。在语义发布层基于SpringData搭建，实现对科研档案知识图谱的检索和应用服务。图3为“农产品质量安全采集作业场景下的语音识别鲁棒性研究”这一科研档案中科研项目、科研成果、人员、机构，以及档案内相关文档之间的知识图谱展示。

5 小结

大数据背景下，我国科研信息化进程迎来巨大的机遇和挑战。为进一步实现研究数据共享复用，推动成果转化，促使科研管理实现信息化，知识图谱驱动的科研档案大数据管理系统作为科研信息化的一部分，旨在为科研档案的智能化和精细化管理提供思路。相对

图3 研究所科研档案知识图谱展示效果图

于传统的档案系统，本文利用知识图谱开展科研档案大数据管理系统构建，依据档案大数据管理流程，针对多种档案类型的资源进行数据解析、抽取和结构化处理，并对各类实体进行识别、抽取与消歧，语义关系的计算、关联、校对、审核等。利用本体、知识图谱、机器学习等语义技术对科研档案核心知识资源进行组织和表示，开展科研档案本体建模、实例转换和可视化呈现。实现档案智能化采集与归档、档案数据碎片化加工、档案数据自动抽取与智能识别、语义关联与知识检索、档案数据审核与发布等功能，形成科研机构知识全景图，推动科研档案智能管理。未来将进一步细化科研档案大数据管理系统设计方案并开展系统研发，充分挖掘与利用科研机构档案资源的知识宝库，建立知识图谱驱动的新型档案大数据智能管理系统，为档案研究创新与发展提供支撑。

参考文献

[1] 吕元智. 面向资源架构的数字档案资源跨媒体整合研究 [J]. 档案学研究，2016（4）：91-96.

[2] 张卫东，左娜，陆璐. 数字时代的档案资源整合：路径与方法 [J]. 档案学通讯，2018（5）：46-50.

[3] 梁孟华. 面向用户的数字档案资源跨媒体知识集成服务研究 [J]. 档案学研究，2016（6）：49-54.

[4] 段荣婷，马寅源，李真. 档案著录本体标准化构建研究 [J]. 档案学研究，2018（2）：63-71.

[5] 李婷. 基于关联数据的档案文化资源整合探索 [J]. 教育现代化，2018（39）：157-158.

[6] 陶水龙. 海量档案数字资源智能管理及挖掘分析方法研究 [J]. 档案学研究，2017（6）：75-79.

[7] LAMHARHAR H，CHIADMI D，BENHLIMA L. Ontology Based Knowledge Representation for E-Government Domain [C] //. Proceedings of the 17th International Conference on Information Integration and Web-based Applications & ServicesDecember 2015 Article. Iiwas，At Brussels，Belgium，2015：51.

[8] MARCONDES C H. Interoperability between digital collections in archives，libraries and museums：potentialities of linked open data technologies [J]. Perspect.ciênc.inf，2016（21）：2.

[9] SARAVANA B B，KARTHIKEYAN N K，RAJ K R. Fuzzy service conceptual ontology system for cloud service recommendation [J]. Computers & Electrical Engineering，2018（69）：435-446.

[10] BOUYERBOU H，BECHKOUM K，LEPAGE R. Geographic ontology for major disasters：Methodology and implementation [J]. International Journal of Disaster Risk Reduction，2019（34）：232-242.

[11] LLANES-PADRÓN D，PASTOR-SÁNCHEZ J. Records incontexts：the road of archives to semantic interoperability [J]. Procedia Social and Behavioral Sciences，2017，51（4）：387-405.

[12] BENIAMINOV E M. Ontology libraries on the web：Status and prospects [J]. Automatic Documentation and Mathematical Linguistics，2018（52）：117-120.

[13] YANG F, YANG Z, COHEN W W. Differentiable learning of logical rules for knowledge base reasoning [C] //31st Conference on Neural Information Processing Systems (NIPS), Long Beach, 2017 (2): 11.

[14] 贾玲, 吴建华, 陆江. 大数据视野下档案业务流程再造方法研究 [J]. 档案学研究, 2018 (6): 90-94.

[15] 周一枫. 基于图像识别的敏感文件检测技术研究 [D]. 杭州: 浙江理工大学, 2018 (12): 2-10.

[16] 郭学敏, SHAW R. 基于关联数据的档案语义转换实践分析 [J]. 档案学通讯, 2019 (5): 50-55.

[17] 王颖, 钱力, 谢靖, 等. 科技大数据知识图谱构建模型与方法研究 [J]. 数据分析与知识发现, 2019 (1): 15-25.

[18] 李超. 视频侦查的知识图谱构建研究 [D]. 北京: 中国人民公安大学, 2019 (6): 1-25.

作者简介

雷洁, 女, 1988年生, 硕士, 助理研究员, 研究方向: 信息资源管理、知识组织。
赵瑞雪, 女, 1968年生, 博士, 研究员, 研究方向: 信息管理与信息系统、信息资源管理、知识组织及数字图书馆。
李思经, 女, 1962年生, 博士, 研究员, 通信作者, 研究方向: 科技管理与科技政策, E-mail: lisijing@caas.cn。
鲜国建, 男, 1982年生, 博士, 副研究馆员, 研究方向: 知识组织、关联数据、语义出版、信息管理与信息系统。
寇远涛, 男, 1982年生, 博士, 副研究馆员, 研究方向: 数字图书馆理论与技术、信息管理与信息系统。

Construction of Knowledge Graph for Scientific Research Archives Big Data Management System

LEI Jie[1] ZHAO RuiXue[1,2] LI SiJing[3] XIAN GuoJian[1,2] KOU YuanTao[1,2]
(1. Agricultural Information Institute of CAAS, Beijing 100081, China; 2. Key Laboratory of Agricultural Big Data, Ministry of Agriculture, Beijing 100081, China; 3. Institute of Agricultural Economics and Development, CAAS, Beijing 100081, China)

Abstract: In order to realize the intelligent collection of archive resources, the semantic organization and statistical analysis of archive big data, and explore a new path for the intelligent management of scientific research archives, this paper is based on the computer-understandable scientific archive knowledge map semantic model established by the editing tool Protégé and modeling language OWL. According to the archive big data management process, data analysis, extraction and structured processing are performed for multiple archive types of resources, and various entities are identified, extracted and disambiguated, and the semantic relationship calculation, association, proofreading, auditing, etc. Use ontology, knowledge map, machine learning and other semantic technologies to organize and represent the core knowledge resources of scientific research archives, to achieve intelligent collection and archiving of archives, fragmentation and processing of archive data, automatic extraction and intelligent identification of archive data, semantic association and knowledge retrieval, And file data review and release functions to promote intelligent management of scientific research files.

Keywords: Scientific Research Archives Ontology; Knowledge Graph; Semantic Association; Fragmentation; Machine Learning; Jena

（收稿日期：2020-02-03）