以藏品为核心的知识图谱设计与应用

刘芳[1] 谢靖[2]
（1. 中国国家博物馆，北京 100006；2. 中国科学院文献情报中心，北京 100190）

摘要：在互联网时代，以藏品为核心的数字资源组织与应用成为博物馆智慧化建设的一项重要内容。本文通过对国家博物馆数据资源的研究与相关度分析，设计以藏品、多媒体、展览、项目、人员、机构、文献7类实体为核心的知识图谱，进而从知识抽取、知识融合、知识存储和知识应用4个方面设计了技术架构，重点分析知识图谱在知识抽取与知识融合中的关键技术问题，以及知识图谱在检索优化、智能推荐、可视化展示和智能问答领域的应用方式，以期对博物馆展览展示、文物保护、考古、修复和社教等业务的开展提供支撑。

关键词：博物馆；藏品；数字资源；知识图谱
中图分类号：G261
DOI：10.3772/j.issn.1673-2286.2020.06.002

博物馆近几年做了大量的数字化工作，如虚拟展示的制作，文物二维、三维影像的扫描，藏品管理系统的开发与升级改造等。但在数字资源的存储与管理方式上仍然沿袭了传统形式，藏品信息与其相关的多媒体资源、展览策划方案、研究文献等长期处于分离状态，没有实现有序的结构化管理，无法从根本上实现数据的有效整合并达到智能化与互联互通。而且随着数字资源体量的不断增大，在检索时由于数据关联性和语义性标注的缺陷以及人机交互技术应用的不足，导致检索的准确率较低，无法通过知识组织的方式为用户实现相关信息的关联与延展，严重影响了博物馆相关工作的顺利进行。知识图谱技术的应用，能够在一定程度上解决目前出现的问题。

知识图谱技术作为融合语义体系[1]和关联数据[2]特征的技术，最早由谷歌于2012年在其项目中提出。在通用领域已经形成以WordNet、DBpedia、YAGO、FreeBase等为代表的较为完整的体系规模。与通用领域相比，垂直领域的知识图谱更加注重概念之间的逻辑结构，因此专业性更强。如以藏品为核心的知识图谱创建，主要用于形象地展示文物的知识结构与知识脉络，从文物类型、历史进程、制造工艺等不同角度展示其相关关系及演变规律，为藏品展览、研究等各项工作的开展提供支撑。在国际上，有代表性的应用包括欧洲数字图书馆为实现全欧洲博物馆、画廊、图书馆和档案馆馆藏资源整合时创建的EDM（Europeana Data Model）模型[3]、国际文献工作委员会颁布的虚拟博物馆语义网络架构（CIDOC CRM）等[4]，具有较强的灵活性和扩展性，能够尽可能完善地保存资源数据的各种描述信息，同时能够在地理位置、时间序列、事件、主题内容、形状等上下文情境中应用相似度算法为分散、异构和跨领域的资源建立数据关联，在资源整合、展示、检索和相关推荐方面展现了良好的应用效果。在国内，2014年科学技术部科技支撑计划“文物数字化保护标准体系及关键标准研究与示范”项目，研究制定可移动文物（以书画与青铜器为例）和不可移动文物（以敦煌石窟寺为例）信息采集、加工、存储、传输、服务和交换等共87项标准，设计了多维度的文物分类主题一体化词表[5]。在2019年国家重点研发项目“文化遗产保护利用”重点专项中设立“基于知识图谱的文物知识组织和服务关键技术研发与示范”专项，针对文物数字化、考古报告、档案资料和文献、互联网数据制定适合文物领域知识的组织和表达模型[6]。本文旨在基于已有研究，围绕博物馆藏品生成的各类数据创建以藏品为核心的知识图谱，期望改进数字资源应用中出现的部

分问题与不足。

本文采用从数据环境到应用服务建设的设计思路，论述以藏品为核心的知识图谱设计流程。首先，明确抽取数据实体、属性和关联关系的范围，建立基于本体的资源描述框架。然后，在此基础之上完成知识的抽取、融合、存储、推理和表示等工作，为用户提供数据服务。

1 知识图谱建模

知识图谱是结构化的语义知识库，以符号形式描述物理世界中的概念及其相互关系。其基本组成单位是“实体-关系-实体”三元组，以及实体的相关属性值对。实体间通过关系相互联结，构成网状的知识结构[7]。在大数据环境中，基于知识图谱对基础业务数据进行归纳分析和统计推理，可便捷有效地掌握博物馆在业务、管理和研究方面最真实的情况，为今后的决策分析提供依据。由于博物馆的主流业务（如展览、保管、研究和交流等）均是以藏品作为核心，因此将藏品作为核心创建博物馆的知识图谱更为妥帖。围绕藏品创建多媒体、展览、机构、研究人员、项目和文献相关实体。7类实体分别概括如下。

（1）藏品。瓷器、牙骨角器等可移动文物。包含其基本要素、造型、功能要素、技术要素、文化背景等属性类型，如藏品名称、藏品总登记号、文物类别、文物级别等。

（2）多媒体。藏品的二维影像、三维影像和视频等。包括影像标号、编码方式、存储位置、影像级别（高清/浏览）、体积等。作为藏品的一个衍生实体，通常默认为拥有藏品的机构拥有影像。

（3）展览。举办的各类线上和线下展览。包括展览名称、展览时间、展览主题、观展人数等属性。

（4）机构。相关藏品保存和研究机构，如博物馆、美术馆、高校等。包括单位名称、单位类型、地理位置等属性。

（5）研究人员。藏品的修复、考古、研究和展览策划的相关人员。包含姓名、学历、年龄、性别、研究方向等属性。

（6）项目。针对藏品进行的保管、考古、修复、教育、文创等业务项目，以及各类国家级、省/部级、地市级、馆级的科研项目。

（7）文献。与藏品相关的各类著作。如期刊论文、著作、图录、导览词等出版物。

7类实体及其关联关系如图1所示。该知识图谱的创建主要用于解决针对藏品的展览、保管、研究工作的资源整合与关联分析。在实际应用中，实体间、实体与属性间存在大量的关联关系。可将业务需要作为出发点，有偏向性地选择知识图谱的组织内容和结构以保证模型的简洁性，可以有效降低模型的构建和维护成本。另外，知识的开放性赋予知识图谱强大的延展性，通用领域的知识图谱以及其他垂直领域的知识谱图能够在一定程度上实现对本领域数据的完善和补充。例如，应用图情领域的知识图谱对出版文献进行统计和分析，利用维基百科丰富文物的描述信息等。

为了实现上述目标，在进行概念模型设计和程序化实现时应优先选用已有的数据模型作为编码标准。CIDOC CRM对异构资源的统一存取具有较强的兼容性，出现至今已被应用于多个具有广泛影响力的项目，

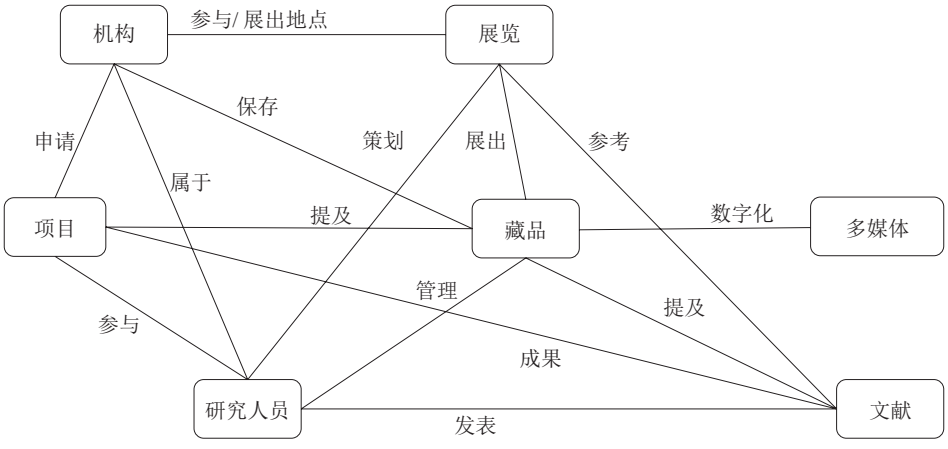

图1 知识图谱的概要模型

发展较为完善。例如，上海博物馆的“董其昌书画艺术大展”数字人文项目中应用该数据模型创建了明清文人书画本体[8]；天津大学张加万研究团队以CIDOC CRM作为数据模型与标准进行文物知识图谱的相关研究，通过与中国文物数据的匹配创建了CN-CRM，对CIDOC CRM进行修订和扩展[9]。知识图谱的结构化表达如图2所示。

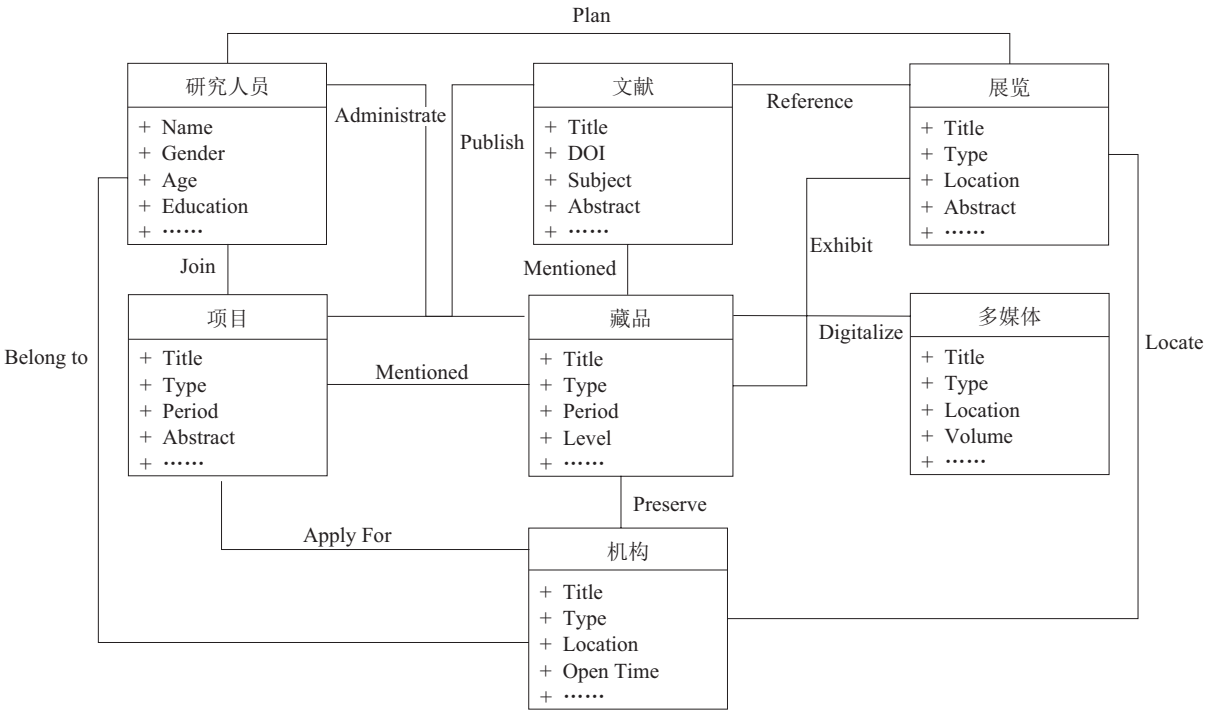

图2 知识图谱的结构化表达

2 知识图谱技术架构

以藏品为核心的知识图谱整体技术架构如图3所示，知识抽取、知识融合、知识存储和知识应用是技术架构的主要组成部分。其中，知识抽取用于完成藏品相关数据的收集工作；知识融合用于实现不同来源数据在同一知识图谱模型下的重新整合。这两部分作为知识图谱的核心组成部分将在下文“3关键技术问题分析”中做详细论述。

知识存储将规范化的数据存储在数据库中，并按需建立缓存和索引。目前常用的知识图谱存储有关系数据库存储和图数据库存储两种形式。两种方式在数据存储和运行性能上各有利弊。其中，RDF关系数据库延续了以往关系数据库基于表的存储和索引方法，具备了较为成熟的操作说明。图数据库更符合人的思维方式，索引方法也更加灵活，但成熟度略低于RDF图[10]。Neo4j、Virtuoso、OrientDB等图数据库在主流应用中均有良好的表现。考虑到Neo4j用户生态较为成熟，能够以高度自由且规范的方式管理和存储数据，将其作为知识图谱的存储数据库。

知识应用基于知识图谱实现精准检索、相关推荐、可视化展示和智能问答功能。将在下文“4以藏品为核心的知识图谱应用”中做详细论述。

3 关键技术问题分析

3.1 知识抽取

以藏品为核心的知识抽取，需获取的数据可以分为结构化、半结构化和非结构化3种类型。其中，结构化数据是优先抽取的数据类型，获取来源包括馆内信息系统和第三方数据库，如藏品管理系统、影像管理系统、展览档案管理系统、考古发掘系统、科研项目管理系统、人力资源管理系统、中国知网数据库等。这类数据的结构通常较为清晰，通过映射或者D2R工具等方式即可实现从原系统数据到知识图谱数据库的转换，

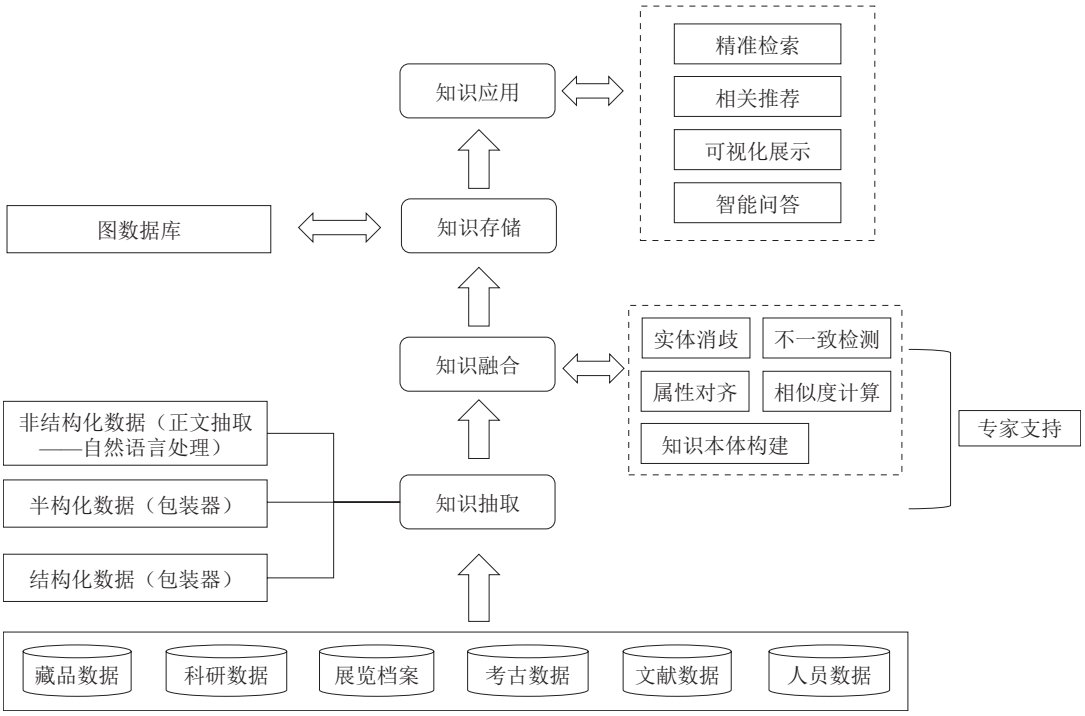

图3 知识图谱的技术架构

仅需完成数据清洗和预处理工作。包括数据的取值归一、标准化处理、无效数据剔除、回补和修订等，实现实体、属性和关系的有效抽取和归并[11]。一般情况下，出于吞吐量以及系统压力的考虑，通常采用批量采集方式实现各系统数据库到知识图谱数据中心的数据同步，以数据容量限制和时间阈值限制相配合的方式作为新批量数据的采集条件。

半结构化数据是知识抽取过程需要解决的重点和难点问题，主要包括研究文献、展览档案、藏品图录等。这部分数字资源占据比重较大，包含大量对文物造型、功能、技术和文化背景要素的描述信息，具有重要研究意义。需要在目录改编的基础上，以文物作为基本单元进行碎片化处理，将其中的文物图片、纹饰拓片、线图以及描述性文字分别进行拆分和标注。在工作初始阶段，为保证知识抽取的正确性，通常借助人工完成文档的碎片化和数据标注工作，然后编写封装器实现对目标数据源的针对性解析。同时，可借助自然语言处理技术对上述结构化文档进一步细化和数据挖掘分析，实现对藏品数据的分类、聚类、相似性分析、关键词自动标引，同时实现基于规则的实体关系发现。经确认后，形成XML格式结构化文档存储[12]。以“守望家园——陕西宝鸡群众保护文物成果”展览为例，九年“卫”鼎的可抽取信息如图4所示。

3.2 知识融合

文物定名要求“观其名而知其貌”，需按照时代、特征、通称顺序排列，如“明成化斗彩高足瓷杯”等。在实际应用中，因为理解和描述上存在的个性化差异，导致不同的藏品名称会指向同一件实物。在中国国家博物馆的展览策划时，为了对藏品名称的描述更加准确，有时会进行部分修改，出现展出的藏品名称与保管系统中的藏品名称不一致的现象。在同类藏品较多的情况下，可能出现指代错误。因此，依据上下文背景和属性分析设计更高效的消歧方法，挖掘正确的实体并连接到知识图谱中，将分散的藏品相关实体以最低的代价建立关系，变成可以利用的知识，是知识图谱构建的关键。本文设计的知识图谱中，含有的指代类型较为丰富，如实体的名称存在、别称、缩写、同义词、曾用名等。通用领域的处理工具如NLTK（Natural Language Toolkit）和Stanford NER等，在未经过训练时，难以达到满意的处理效果。在知识融合过程中，可通过词典、叙词表建设对实体、属性和关联关系进行相似度计算，达到消除歧义及正确识别的目的。对于计算机无法识

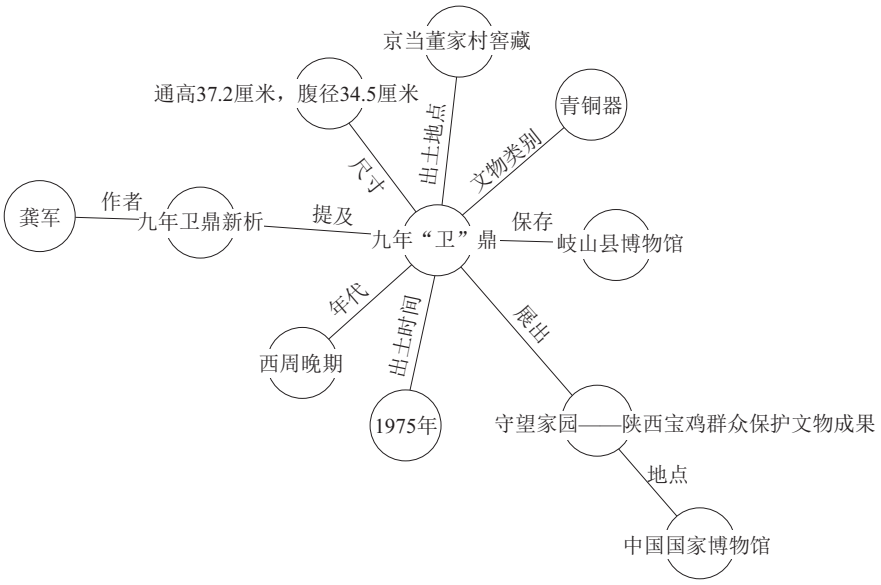

图4 九年“卫”鼎数据抽取示例

别的实体需要专家给予决策意见，并在识别完成后，依据国际通用标准为其标注唯一标识符。

近些年，出现了不依靠本体模型，通过“自训练”的方式，主动寻找“关系”实现大规模的实体扩展与融合，但是在垂直领域的准确度有待提高[13]。为保证准确性，可采用半监督学习实现三元组数据的抽取工作。首先，在专家的指导下创建正确的关系模式，并以手工方式创建“关系”集合；然后，以“关系”作为触发词，在数据中手工标注样本[14]；最后，通过相似度匹配找到新的样本扩充训练集，并将符合条件的实体加入知识图谱中不断扩充。例如，“藏品-保存-机构”可抽取出“九年‘卫’鼎-保存-岐山县博物馆”。这种构建方式，保证了知识图谱的准确性，但是需要人工更新关系集合，以确保更加丰富的信息能够添加到知识图谱中。

4 以藏品为核心的知识图谱应用

知识图谱的应用重点不在于堆积数据，而是基于对大量数据的整理和分析，制定可执行的方案解决博物馆业务中出现的问题，然后投入到大规模的应用中产生效益。数字资源、知识图谱和数据技术是实现应用服务的3个重要组成部分，为优化数字资源应用服务提供了支撑。数字资源是应用服务的内容主体，数字资源结构的规范性和内容的丰富性，决定了应用服务所能提供的知识范围。知识图谱是实现应用服务的知识组织和整合工具，知识图谱的准确性和完整性，决定了应用服务所能提供知识的深度及知识的可获取性。数据技术，是应用服务的技术支撑环境，先进的数据处理技术（如分布式计算、数据推理技术、人工智能等）决定了应用服务的交互性和智能化程度。本节重点介绍知识图谱在精准检索、资源推荐、可视化数据展示和智能问答4个方面的使用价值。

4.1 检索优化及资源推荐

首先，知识图谱采用图结构建模和记录藏品相关实体间的关联关系。在用户使用关键词进行检索时，可不拘泥于词语的字面含义，通过实体数据匹配准确检索到用户查找的信息，并能包容自然语言的多样性，有效实现更加精准的对象检索。例如，中国国家博物馆馆藏文物“司母戊鼎”，经考证后被更名为“后母戊鼎”，需将“司母戊鼎”作为曾用名存储到词典中，在检索时两者应指向同一件文物。

其次，博物馆涉及大量多媒体资源的检索问题。多媒体资源的语义标注对提高检索结果的准确度具有重要意义。有赖于专家经验提取影像的颜色、纹理、轮廓、形状、领域概念等用户感兴趣的特征进行标注，并基于这些信息实现对实体分类、属性和关系的描述。

最后，以藏品为核心的知识图谱将存在关联关系的多媒体、展览、文献、机构和研究者等实体数据整合

成网状结构，极大地丰富了数据的上下文环境。在检索时，能实现与检索词存在关联关系的所有实体的多维度检索，并根据用户对初次检索结果的偏好逐渐调整并缩小检索范围。以中国国家博物馆新入藏的流失文物“虎鎣”为例，“青铜器”“水壶”“老虎”“周代”“祭祀”“海外流失”“拍卖”“圆明园”是从器型、铭文、功能、纹饰、历史背景等不同角度对藏品具有代表性的描述。选择1~2个描述信息能够快速缩小检索范围，提高检索效率。这是传统的藏品管理系统（只涵盖藏品基本描述信息）不能达到的效果。当用户有清晰的检索目标，但是由于记忆的缺失，无法回忆出该信息的任何关键字作为检索条件时，可将与查找实体存在关联关系的属性和其他实体作为补充描述。在搜索某件文物时能将其同一时期、同一质地、同一类型或同一发掘地出土的其他文物推荐给用户。如同为殷墟出土的“后母戊鼎”和“司母辛鼎”，在形制、纹饰和铭文风格上具备较强的相似度，可作为对方的相关推荐进行展示。

4.2 可视化数据展示与智能问答

在智慧博物馆建设要求的不断驱动下，博物馆的藏品展示与教育方式已不仅满足于面向公众的静态说明，要求以生动的方式引导观众基于文物介绍去思考和获得衍生的知识。通过制造工艺、功能、造型和文化背景等多个角度了解文物，并对其存在的历史环境有全面的认识。基于知识图谱的藏品数据可视化展示能够在有限的空间环境中，为藏品提供丰富的上下文背景并基于时间、地理位置、文物类型等维度展示文物的相关性。同时通过触碰的方式不断基于某个知识点对图谱进行延展和钻取，增强与用户之间的交互性。例如，基于知识谱图展示青铜器上铭文字体的演变规律、青铜器制造工艺的改进方式、从青铜器族徽的变化解释家族迁徙和融合的规律等，具有重要的实践意义。这种展示方式，在上海博物馆“董其昌书画艺术大展”数字人文项目的在线展示中，取得了良好的社会效益。

智能问答是知识图谱的一种重要应用形式。智能问答首先需要应用自然语言处理技术对用户的问题进行分词处理从中找到核心实体和关系，然后与知识图谱中的实体和关系进行映射和匹配，并将查找到的知识点反馈给用户。由于用户对实体和关系的描述往往较为模糊，且可能跨越了多个三元组结构，因此系统需要具备较强的相似度计算能力和对本体规则结构推理能力。例如，“在国家博物馆与后母戊鼎同时期的展品有哪些？”需要定义规则“（?x right held by 中国国家博物馆），（? x belongs to 商后期）->（?x belongs to same period as 后母戊鼎）”。对于复杂或模糊的问题，可先将与描述问题相关的结果全部推送给用户，然后通过一问一答的方式不断明确用户提交的问题。智能问答系统可应用在大厅的讲解机器人上，便于快速解决用户需求，提高用户参观体验的趣味性，同时可节约人力成本。

5 结语

以云计算、物联网、移动通信、大数据和人工智能为代表的新技术，不但改变了人类的思维观念、价值取向和生活方式，同时也驱动着智慧博物馆建设成为历史趋势。知识图谱技术作为大数据管理和应用的一种有效手段，能够对智慧博物馆在资源组织、学术研究和业务管理等方面的发展起到积极的促进作用。本文在总结分析相关项目成果和研究资料的基础上，从理论和技术角度论述了以藏品为核心的知识图谱设计和应用模式，可行性仍需在实际应用中进行验证、改进和完善，并从以下方面进一步落实相关工作。

（1）知识图谱的终极目标是将非结构化、无显示关联的粗糙数据逐步提炼为结构化、高度关联的知识体系。核心为实体识别与关系建立，其质量能够影响图谱的描述能力与准确性。前期通常需要专家依据词表手工建立基本的数据模型，实现概念、实体、关系和规则的抽取，以及实体的消歧、指代消解和文本理解等，再依靠自动化知识抽取和图谱补全进一步提升图谱的质量。

（2）词表包含大量的概念词汇，以及词汇之间的等同、等级和相关关系，对知识体系的建立具有重要的指导意义，应做好文物的词表分类和建设工作。网络词表和非逻辑标签也是词表的重要获取来源。

（3）知识图谱是一个开放、动态的模型，应注重词表的训练与更新以提高应用准确率。从技术维度上看，知识图谱是一个新型的信息系统基础设施，集成了自然语言处理、机器学习、知识标识、数据库和多媒体等多个技术领域的技术与思维方式[15]。这些技术的发展为知识图谱的创建提供了基础和发展条件，同时也促进知识图谱建设的不断完善。

博物馆的数字化建设现状，与科技、金融、医疗等

领域相比，存在滞后性的同时也为基于知识图谱的大数据平台建设提供了契机。可以在原生态环境中搭建统一的数据模型和技术架构，降低研究、设计、管理、应用和维护的难度，为博物馆的研究、展览、修复、考古和社会教育等各项工作提供有力支撑。

参考文献

[1] Ontology [EB/OL]. [2020-05-15]. https://ont.io/.

[2] Linked Data [EB/OL]. [2020-05-15]. http://linkeddata.org/.

[3] EUROPEANA [EB/OL]. [2020-05-15]. https://pro.europeana.eu/.

[4] Definition of CIDOC Conceptual reference model. Version 6.2.2. ICOM/CIDOC Documentation Standards Group [EB/OL]. [2020-05-15]. http://icom.museum/resources/publications-database/publication/definition-of-the-cidoc-conceptual-reference-model/.

[5] 文物数字化保护元数据标准规范征求意见稿发布 [EB/OL]. [2020-05-15]. https://www.lib.pku.edu.cn/portal/cn/news/0000001494.

[6] 关于对“重大自然灾害监测预警与防范”重点专项（文化遗产保护利用专题任务）2019年度项目申报指南征求意见的通知 [EB/OL]. [2020-05-15]. https://www.sohu.com/a/302689451_120029064.

[7] 刘峤，李杨，段宏，等. 知识图谱构建技术综述 [J]. 计算机研究与发展，2016，53（3）：582-600.

[8] 童茵，张彬. 董其昌数字人文项目的探索与实践 [J]. 中国博物馆，2018（4）：114-118.

[9] 杨伟强. 文物知识图谱的构建与应用 [D]. 天津：天津大学，2018.

[10] 王鑫，邹磊，王朝坤，等. 知识图谱数据管理研究综述 [J]. 软件学报，2019，30（7）：2139-2174.

[11] 阿里巴巴数据技术及产品部. 大数据之路——阿里巴巴大数据实现 [M]. 北京：电子工业出版社，2017：7，8-15.

[12] 王昊奋，胡芳槐. CCKS-2017-行业知识图谱Tutorial [EB/OL]. [2020-05-15]. http://www.cipsc.org.cn/ssatt2018/.

[13] HU W，CHEN J，QU Y. A Self-training Approach for resolving object coreference on the semantic Web [C] //International Conference on World Wide Web. ACM，2011：87-96.

[14] 张娜. 文物知识图谱构建关键技术研究与应用 [D]. 杭州：浙江大学，2019.

[15] 漆桂林，高桓，吴天星. 知识图谱研究进展 [J]. 情报工程，2017（1）：4-25.

作者简介

刘芳，女，1985年生，硕士，馆员，研究方向：博物馆数字化建设、数据组织、数据存储与服务，E-mail：liufang@chnmuseum.cn。

谢靖，男，1983年生，硕士，副研究馆员，研究方向：大数据计算、数据组织、知识图谱、用户画像。

Design and Application of Knowledge Graph with Collection as the Core

LIU Fang[1] XIE Jing[2]

(1. National Museum of China, Beijing 100006, China; 2. National Science Library, Chinese Academy of Science, Beijing 100190, China)

Abstract: In the Internet era, the organization and application of digital resources with collections as the core has become an important part of the intelligent construction of museums. Based on the research and correlation analysis of the data resources of the National Museum, this paper designs a knowledge graph with the collection, multimedia resources, exhibitions, projects, people, institutions and documents. Then, the technical framework is designed from knowledge extraction, knowledge fusion, knowledge storage and knowledge application. Focused on the analysis of the key technical problems of knowledge extraction in knowledge extraction and knowledge fusion. Then, it analyzes the application of knowledge graph in the fields of retrieval optimization, intelligent recommendation, visual display and intelligent Q & A in order to provide support for museum exhibition, collection protection, archaeology, restoration and social education.

Keywords: Museum; Collection; Digital Resource; Knowledge Graph

（收稿日期：2020-03-25）