

关联数据在知识库中应用的研究综述^{*}

尚渡新¹ 袁润² 夏翠娟³ 刘桂锋²

(1. 江苏大学科技信息研究所, 镇江 212013; 2. 江苏大学图书馆, 镇江 212013; 3. 上海图书馆, 上海 200031)

摘要: 近年来关联数据在知识库中得到广泛应用。本文对关联数据在知识库中的研究现状进行系统分析和梳理, 并总结关联数据在知识库的知识源、可视化和语义化3个方面的重点应用, 探讨关联数据应如何更好地促进知识库的建设以满足用户的知识需求, 进而指出关联数据在知识库的诸多领域皆有研究, 但相关项目未能充分发挥关联数据的特点, 存在过度依赖叙词表、忽视学科交叉等问题, 后续研究需要在构建理念、知识集关联、知识领域融合和构建目标方面增加关注度。

关键词: 关联数据; 知识库; 知识组织; 知识管理

中图分类号: G254 **DOI:** 10.3772/j.issn.1673-2286.2022.03.003

引文格式: 尚渡新, 袁润, 夏翠娟, 等. 关联数据在知识库中应用的研究综述[J]. 数字图书馆论坛, 2022 (3): 22-31.

1 关联数据与知识库的关系

知识库是运用知识组织理念和信息化技术等对各类知识进行有效组织、形式化表述和规范化利用的一整套知识系统, 能够以机器可解释的方式存储、分析和重用这些知识, 并富含知识集间的相互关联^[1]。其中知识集是知识库中一类特定知识的集合, 多个知识集共同构成知识库的知识源^[2]。知识库作为学术交流的基础设施, 提供有关某一专题或事物的知识内容, 同时随着研究的深入, 知识库的体量和类型也在不断丰富, 极大方便了研究人员对知识的获取和利用^[3]。进入语义网时代后, 传统知识库在一定程度上对内缺乏语义资源整合、对外缺少知识集关联, 在面对用户的知识需求升级时显得难以为继^[4], 使得大量的宝贵知识陷入了“信息孤岛”的困境。因此, 语义层面的知识组织成为提升知识库价值的关键。

关联数据是一种通过可链接的方式来发布、分享和组织Web中各类资源并利用Web在不同资源之间创建语义关联的数据原则, 也是语义网研究的主题之一。

作为一种语义网的实现方式, 关联数据所拥有的数据源动态扩展、知识元关联和知识语义化检索等特征^[5], 为改变传统知识库构建过程中面临的非动态更新、可视化功能薄弱和语义匮乏等局面提供了新思路。国内外相关学者利用关联数据在知识库领域开展了诸多研究: 康奈尔大学通过本体模型和关联数据技术构建了生命科学领域的知识库VIVO, 并提供语义检索和关联扩展功能^[6]; 麻省理工大学、W3C和惠普实验室也利用关联数据技术共同承建了名为SIMILE的数字图书馆项目, 有效地实现了知识库的数字教育资产、图像知识集与维基百科之间的语义互操作^[7]。

关联数据自2008年引入国内后虽经历了长期的技术积累, 但在知识库领域的应用集中于图书馆、档案馆和博物馆这些图书情报学科所扎根的实践阵地, 同时这些应用中有一部分仅借助DSpace框架与外部知识集进行关联实现联邦检索^[8], 并未利用关联数据在知识库的语义层面进行广泛的知识组织进而服务于用户, 更多是理论研究、局部研究性的实践。这就造成关联数据的研究成果没有很好地解决现实中知识库遇到的知识

^{*} 本研究得到江苏省研究生科研与实践创新计划“基于本体自增长模型的农业机械知识服务平台研建”(编号: KYCX21_3319)资助。

组织问题,或者说没有大幅度提升用户的知识服务体验,也就未能形成实践产生问题、问题催生理论、理论作用实践的良性循环。因此,探究关联数据的研究应如何助力于知识库的发展,以应对语义网环境下用户对知识库更高的服务需求,从而提升国内知识库的实用价值和使用率,这便成为一个值得讨论的问题。本研究以CSSCI来源期刊中的“图书馆、情报与文献学”学科类别下的21种期刊作为中文文献源;将Web of Science Core Collection中期刊学科类别为“Information Science Library Science”的文献作为外文文献源,搜集以关联数据在知识库中应用为主题的文献资料。限定时间2012—2021年,共获取相关中文文献75篇,外文文献87篇。以此为基础综述关联数据在知识库建设中所发挥的作用,针对研究现状分析关联数据在知识库中研究的不足,并对今后的相关研究和应用作出建议。

2 关联数据在知识库中应用的研究现状

2.1 研究概述

通过阅读和分析相关文献,近十年来关联数据在知识库中的研究整体呈现平稳上升的趋势。国外较早将关联数据引入知识库的建设中,如Baker^[9]在其文章中提出运用RDF(Resource Description Framework)和都柏林核心集来确保知识库中的数据满足质量和一致性的要求,进而重点关注了关联数据技术在知识库进行内外部数据源互操作时,对数据的一致性所发挥的约束作用,并分析了关联数据具备将一致性约束建立在创建数据的过程中的优势属性,以及可以满足将知识库的内部知识与遵循其他标准的知识源进行链接或合并的技术要求,为关联数据在知识库的早期应用提供了借鉴。

同时,受到传统数字资产管理工具DSpace、EPrints等研究理念的影响^[10],国内外的早期研究主要凭借关联数据的突出优势集中对知识库内部知识集的组织 and 外部知识集的关联等多源知识的关联化进行研究,如徐华^[11]论述了在DSpace基础上运用关联数据技术实现知识库间的资源整合的优势,将关联数据作为基础描述框架实现了机构知识库间合作模式的构建。

随着信息需求的提升,有学者开始聚焦于知识库中的知识消费,关联数据的应用也转向了知识库的可视化服务和知识库的语义化操作。如Li等^[2]通过知识服务阶

段的关联数据可视化建立事件关系网络,为人文研究中的历史事件表述提供一个通用的描述规则,这种可视化的知识服务方式将有助于历史学者研究经济和微观社会等客观存在对历史事件的影响。

此外,为进一步了解关联数据在知识库中的研究热点,本研究采用词频分析、主题词共现结合文献阅读和案例分析的方法归纳出近年来关联数据在知识库中应用的3个主要趋势:以数字人文、知识融合为代表的主题词词频的提升,显示了关联数据知识库建设中融合多源知识的优势;可视化、知识服务等主题的重要性愈发凸显,反映了关联数据可视化在知识库服务中的重要作用;文献计量结果中语义检索、语义关系的研究比重上升也表达了语义化将成为知识库在应用关联数据时不可或缺的一部分。

2.2 关联数据在知识库的典型应用

随着关联数据相关理论和技术的成熟,国内外诸多研究机构利用关联数据技术以知识库的形式组织并表达其领域知识。在生物医学领域,上海曙光医院联合华东理工大学利用关联数据发布了中文症状知识库,包含症状、科室、疾病、药物和检验,以及症状与上述相关实体的关系^[12];斯坦福生物医学信息研究中心(Stanford Center for Biomedical Informatics Research)通过关联数据技术检索和聚合了多个开放数据源、网络服务和生物医学本体中埃博拉病毒的知识,并开发了以埃博拉病毒为中心的知识库(Ebola-KB),同时转换为RDF,链接到Bio2RDF数据集,并通过SPARQL端点提供服务^[13]。在文艺领域,意大利国家研究委员会(Italy National Research Council)信息科学与技术研究所(Institute of Information Science and Technologies)采用关联数据将文学与地理相结合,构建了中世纪和文艺复兴时期人文主义的拉丁地理文学知识库^[14]。在图书情报学领域,周晨^[15]收集了各成员馆的开放书目数据,基于关联数据构建了联合目录知识库,实现联合目录的开放共享。关联数据在不同领域的知识库中的重点应用大致集中在机构知识库、领域知识库、基础知识库和网络知识库,如表1所示。

纵观国内外已发布的关联数据知识库,上海图书馆开放数据平台依托上海图书馆丰富的馆藏资源,利用关联数据技术持续整合相关知识,并不断以关联数据的方式向互联网发布上海图书馆数字人文项目所涉

表1 关联数据在重点知识库中应用的类别

知识库类别	机构知识库	领域知识库	基础知识库	网络知识库
知识类别	收录某一机构的研究成果，以该机构的期刊论文、会议论文、学位论文、研究报告为主	某一主题知识的系统化收录	科学研究中横断性知识的规范化收录	以通识性知识和用户贡献知识为主
服务目标	无限期地保存机构所属的出版物，提高引用率，从而快速分发研究方面的研究成果	通过联合目录实现知识的聚合，实现知识发现和知识进化	满足用户在相关学习、交流和研究活动中对基础知识易于扩展、便于重用和共享的需求	以结构化、语义化的知识公开供用户取阅
关联数据的重点应用	规范研究成果的外部特征，实现书目控制	进行细粒度描述，对知识资源进行重新组织，利用全网域的互联网知识实现知识域的关联和组织	实现内容分类系统与关联数据集建立连接，使其知识库能够呈现从关联数据云自动更新内容	通过多源数据的整合实现“去中心化”，并对结构化的知识赋予语义关系
典型案例	SIMILE	盛宣怀档案知识库、Ebola- KB	上海图书馆人名规范库	Dbpedia

及的基础知识库（人、地、时、事、物）、文献知识库（家谱、手稿档案、古籍等），以促进数据的开放获取、共享和重用。其中上海图书馆开放数据平台上线的人名规范库^[16]以丰富的功能、便捷的调用方式、规范翔实的记录成为关联数据在知识库中应用的典型案例。

上海图书馆根据人名数据的内部特征和业务需求首先构建了人名规范库本体（shlnames）用以描述和规范所创建的人名数据，作为关联数据构建人名规范库属性和关系的基础，如图1所示，同时复用了foaf, bibframe等词表的部分属性。其次，对已有的人名数据进行数

据清洗，提取实体及其关系，如《中国近现代人物名号大辞典》《中国近现代名人辞典》《上图古籍数据库》《上海年华》《国图规范档》等。而后，根据“主-谓-宾”的三元组数据模型对已有实体进行描述并根据关联数据规范赋予URI，同时与外部数据库构建关联，如与上海图书馆馆藏唱片知识库、上海文化总库等进行关联。最后，将RDF序列化编码后储存于Virtuoso数据库中，通过关联数据发布原则并结合历史人文大数据平台以可视化的形式展示数据，利用Web平台实现知识发现和利用。

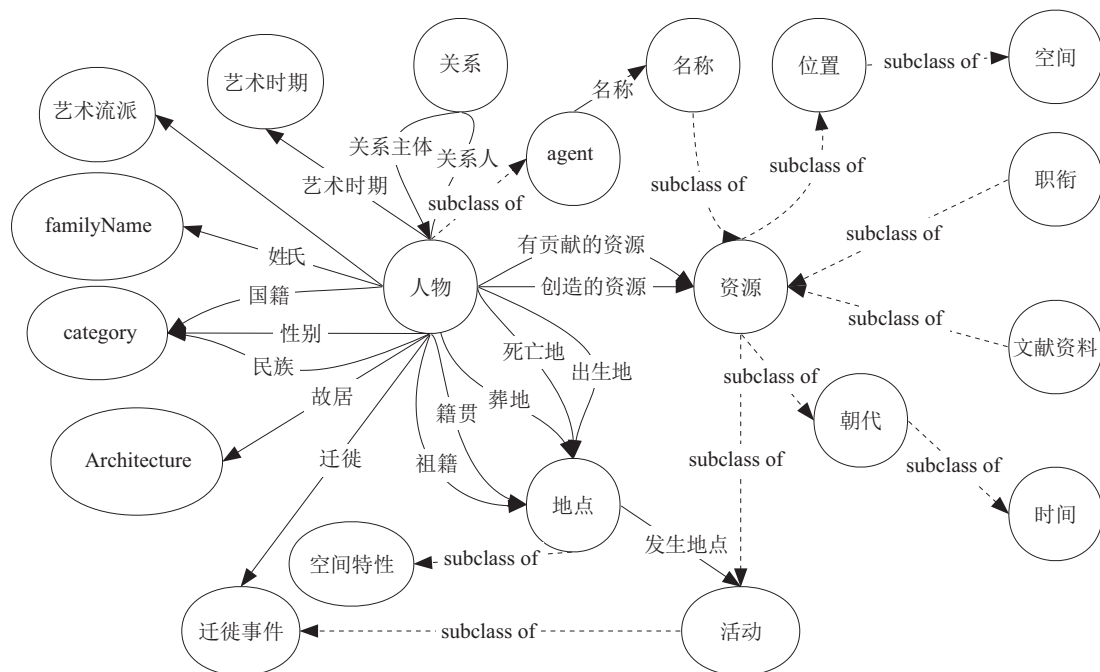


图1 上海图书馆人名规范库本体

从上述文献计量所呈现的结果结合人名规范中关联数据的发布流程, 我们可以看出关联数据在知识库中的实践及应用重点集中在多源知识的关联化、知识服务的可视化和知识库中知识的语义化描述, 这些主题往往是关联数据应用于知识库的主要优势, 同样也是研究重点。

2.3 知识库中多源知识关联化研究

在以往的研究中, 知识库所包含的知识往往是知识密集型资源, 如学术期刊、图书和专利文本等。随着信息技术的快速发展与广泛应用, 更多的信息生产环节得到关注, 知识库中的知识类型也逐渐多样化, 如产业信息、文化知识和科研数据等都已成为知识库的收集对象。但这些知识通常囿于系统隔阂而被禁锢于少数群体或某些区域, 对共享和重用该类型知识造成了壁垒和障碍。关联数据的引入可以将多源知识和物质载体相关联, 如后现代主义和梵高的《向日葵》之间的知识融合实践^[17], 该方案不仅提高了知识的完整度和系统性, 也为异构知识集之间的传递和利用提供了依附。在关联数据理念刚出现时TimBerners-Lee^[18]便制定了关联数据内容描述的4项基本原则: 使用URI来标识事物; 使用HTTPURI可以让用户访问到这些标识; 当有人访问到标识时, 提供有用的信息; 尽可能提供关联的URI, 以使人们可以发现更多的事物。

同时, 我国学者也在4项基本原则上针对关联数据的发布做了进一步的总结和规范^[19]: 用RDF数据模型描述要发布的数据资源, 为其生成HTTPURI (通常是一个CoolURI), 并生成资源的RDF描述文档; 在数据之间建立RDF链接。

综上所述, 无论是内容描述还是资源发布都有统一的格式要求, 这就为多样化的知识集融合提供了统一的范式。其中RDF是关键, 关联数据采用RDF描述信息资源, 可利用三元组 (实体、属性和属性值) 描述知识资源之间的相互关系, 同时借助于URI对资源的定位, 可以实现多源数据的互操作, 极大地促进了知识间的跨系统融合。

(1) 过程蕴含知识。事物发展过程中的每个环节或节点都会产生大量的信息, 同时又会有大量的因素对过程产生影响, 这些信息和因素通过关联数据的结合会产生有价值的知识, 是知识库所要收集和组织的对象。以往对过程的研究多集中于生产过程, 侧重于实体

生产中的价值创造与价值获取^[20], 如今国内外学者对诸如创新的过程、知识创造的过程加以关注。这些过程虽无实体, 但却涉及众多实体之间信息的传递和共享, 这些信息的传递和共享有利于组织的持续创新和发展。例如, Zhang等^[21]借助关联数据技术, 对供应链中各节点企业的知识进行搜索、组织和分析, 以实现不同地区、不同行业企业之间的信息共享、流程协调, 并建立了一个完整的知识库。关联数据将知识库中知识集链接的同时也进行了网状的知识序化, 使得各领域间的知识得到了规范化表达, 形成知识网络。比如, 一个地区的发展过程中会产生诸多知识, 对某一个地区构建知识库便可利用关联数据对多源知识进行组织。徐晨飞等^[22]以《方志物产》云南卷中的知识为组织对象构建关联数据知识库, 利用关联数据框架首先对分散的多类知识元素构建本体, 使得该领域的知识得到时间序列上的规范化表达, 然后对源数据进行实体抽取并转换为RDF格式, 并进行实体关联和发布, 实现知识的网状序化。与此同时, 该知识库在时间线上与上海图书馆的“中国历史纪年表”中对应的历史条目实体和属性相关联, 不仅丰富时间维度上的知识, 也对时间线做了标准化描述; 在利用方面与华东师范大学数字方志库相关联, 可以更好地建立跨地域的知识连接, 充实方志机构库。

(2) 文化孕育知识。文化资源作为人类社会精神文明成果的外化存在, 孕育出了丰富的表现内容和载体。通过调研发现, 在初期的知识库建设中更多收集的是各种文化载体的数字化资源^[23], 其中所蕴含的知识关联并未得到充分的揭示。如今, 语义网的发展为文化的传承和发扬提供了新的契机。例如: 传统文化遗产的知识本体可将文化与知识融为一体^[24], 实现了文化因子的有序化和文化关联的系统化收集; 上海图书馆的知识库借助数字人文实践^[19], 从理论和实践两个层面论证图书馆的书目控制、规范控制、文献循证方法和关联数据、知识图谱、大数据等技术结合起来的前景和优势, 为知识库建设提供了参考。可以看出, 将虚拟的文化元素作为知识库的知识单元既为人文学者的研究提供了有力帮助, 也为文化服务发展提供了强劲推力。如LINKED CANVAS^[25]在Synaptica开放注释语义索引系统 (Open Annotation Semantic Indexing System, OASIS) 的基础上, 利用关联数据平台和国际图像互操作框架 (International Image Interoperability Framework, IIIF) 共同开发艺术品知识库, 该知识库连接全球范围内的学校、社区等机构, 使得本地图像可以突破储存的

限制, 获取到全球范围内与图像相关的非物质文化遗产数据和注释, 实现知识集间的跨平台融合, 帮助人们从视觉、概念和背景上更全面地探索艺术品。

(3) 数据积蓄知识。大数据技术从数据中挖掘知识, 而知识库则是在数据有序化的基础上叠加现有知识, 催生出新的知识。未来科技领域的知识治理能够运用关联数据联合各类知识库, 对接数据和知识的节点, 推进数据开放互联和知识融通。与此同时, 建设公共性、开放性、持续性的“数据基础设施”已经提到议事日程^[26]。例如, 夏翠娟^[27]利用文献调研、数据建模、比较研究、实验研究等方法, 调研和比较了CBDB项目、复旦大学历史地理研究中心、台湾地区“中研院”GIS研究中心等, 在熟悉中国历史地理数据库的建设及利用情况的基础上, 探索在图书馆的数字人文项目建设中引入现代历史地理学的研究成果, 利用知识组织和规范控制方法丰富现有知识库, 实现历史地理数据在图书馆知识库开放应用的目标, 拉近历史地理学与人文社会科学研究者之间的距离, 达成数据即知识的转化; 在联机图书馆中心(OCLC)的数据科学项目^[28]和LVK(Library Knowledge Vault)项目中, 研究人员利用关联数据技术对三元组数据进行更细粒度的加工, 使得图书馆知识库中的数据从关联形态向知识形态转变, 便于知识库和互联网搜索引擎和Web应用程序的调用。

2.4 知识库服务的可视化研究

知识库的可获得性和易用性是决定信息用户是否利用该系统的最重要因素, 正如著名的穆尔斯定律所言: “一个情报检索系统如果对用户来说, 取得情报比不取得情报更伤脑筋和麻烦的话, 用户就会倾向于不使用该系统”。而影响知识库易用性的关键因素是知识的表示方式和知识结构, 其中后者又是前者实现的前提和基础。关联数据作为一种新兴的数据表现方式, 具有统一的数据结构(RDF)和交换格式(N3、Turtle), 同时关联数据基于领域本体构建知识库时又赋予了知识规范化的网状关系, 可视化方法又能够勾勒出知识之间的脉络和趋势, 有利于实现视觉增强分析。关联数据的这些独特之处为知识集交互和可视化呈现提供了便利和优势, 诸多学者利用该优势对多领域的知识进行了可视化表达, 其中尤以刘炜、夏翠娟、陈涛等学者为代表的上海图书馆研究团体所承建的知识库最为显著和系统。该团队在2015年系统研究了关联数据可视化的

多种方式, 包括利用R语言对SPARQL端点的统计分析可视化、通过JSON数据格式的可视化等^[29]; 并于2016年将可视化技术应用于家谱知识库的开发实践, 在该项目中不仅实现了本体的Class结构可视化, 而且结合GIS信息利用SIMILE Timemap在地图上实现知识的检索和发现^[30]; 2017年, 在复用本体的基础上, 又将同一地区不同时代的GIS信息和历史信息转化为关联数据实现知识的关联, 并对关联数据实现了多图层的叠加展示^[27]; 2019年, 在实践成熟的基础上, 该团队提出了关联数据可视化的系统框架, 并发布基于关联数据的中国历代人物传记资料库(CBDBLD), 该知识库以可视化的形式、丰富的图运算功能展现了历史人物之间丰富的亲属及社会关系, 形成特有的社会关系网络^[31]。知识库的用户在利用知识时受到知识本身可获得性和易用性的影响, 知识服务可视化技术将知识库内抽象的知识具象为人类易于理解的符号, 提高了复杂知识的可获得性和易用性, 从而极大地促进了知识的消费、传播和再创造。可视化知识服务也逐步成为知识库的研究趋势, 得到众多学者的深入研究。

国际语义网会议在2015年增设了本体和关联数据的可视化与交互研讨会, 目的是讨论关联数据实践中如何利用可视化技术方便快捷地协助信息需求者获取知识资源及构建知识脉络。不仅仅是讨论, 知识库中知识集可视化研究在各领域的实践环节已经取得丰硕成果。在前沿知识发现工作中, 知识库的知识集可视化可以及时发现并准确识别前沿知识, 如Wu等^[32]将LDA可视化模型和知识库结合起来, 使用MeSH(Medical Subject Headings)对生物医学概念进行标注, 对主题、演化和知识结构进行了可视化, 在癌症治疗领域的前沿热点问题发现上发挥了重要作用。在数字人文的研究中, 知识库的可视化可以建立事件关系网络, 为人文研究中的历史事件表述提供一个通用的描述规则, 这种可视化的知识服务方式将有助于历史学者研究经济和微观社会等客观存在对历史事件的影响^[2]。

关联数据、知识库和可视化三者的结合, 显现出数据结构和新兴技术的优势。上述案例也说明关联数据不仅可以满足知识库的新需求, 还能以直观、便捷的方式挖掘出“冰山”表面下的知识, 利于知识的发现。然而, 通过梳理文献发现, 关联数据集发布后仅采用SPARQL作为数据利用方式的项目占据多数, 可视化技术的嵌入和服务方式仍待推广。

2.5 知识库的语义化研究

知识库中语义化操作的研究经历了从同义词典映射到RDF扩展和关联,虽然研究的目的大同小异,但研究目标随着网络环境的变迁时异事殊。传统知识库的语义化研究往往作为优化检索的渠道具有很强的依附性^[33],而在当前语义网的理念下,知识库进行语义化研究更多的是一种赋能,是后续基于语义对知识进行深度利用和探索的前提。语义化在知识库中的研究也逐渐由幕后走向台前,变得不可或缺。所谓语义化,是指把知识客体中的知识因子和语义关系表示出来,以便人们识别和理解知识^[34]。关联数据作为一种语义网的实现方式,应用于知识库的语义研究中,这不仅迎合了知识库语义化的发展趋势,也进一步推进了信息资源向语义Web的融合。与此相关的研究也一直在持续。

国内,国家图书馆在2009—2021年完成关联数据服务系统的构建,成功发布《中国分类主题词表》语义版,对元数据进行URI标识规范^[35],使得知识库的语义化关联有章可循。在具体实践中,学者利用IIIF和关联数据技术,通过对印章资源的语义揭示和知识推理构建“印章知识中心”,实现异构知识库和不同应用程序中知识的语义关联^[36]。国外,关联数据对知识库中的文本和多媒体等资源的语义化标注更为关注^[37]。Alamri^[38]通过构造语义关联数据本体,实现路径数据库和疫情知识库的语义解释,以应对COVID-19引发的室内导航问题并用于减少室内空间的人口密度,以确保社交距离和防止疫情传播;Hu等^[39]利用关联数据技术对跨领域的信息进行语义注释并推断隐含的知识,在建筑节能评价的领域可以识别低效的建筑操作,提高能源效率。由此可见,关联数据在语义化层面可发挥重要作用,这一方面源于传统词表语义化的陆续开展,另一方面得益于关联数据基于语义本体构建的先天条件,两者共同形成关联数据语义化的发展动力。

语义互操作也是知识库长久以来的努力方向,而语义互操作的理想前提则是计算机充分掌握和理解知识的语义。因此,揭示知识中的隐含语义便成了构建知识库的重要一环,相关的研究也是踵事增华,如Boteram^[40]在研究更大范围内的知识库语义互操作时,也强调了关联数据支持语义互操作的特性对于有效管理这些信息系统的重要性。该研究提出“元知识组织系统”具备在概念层次的语义互操作,这就要求将所有相关的子系统整合为一个全面的国际知识组织系统,提

供统一的语义互操作范式也将成为跨知识库间互联的前提和基础。王思丽等^[4]以中国科学院机构知识库平台(CASIR)为例,研究了利用关联数据实现机构知识库意义扩展的关键技术,并实现了针对可扩展实体的语义标注。

3 关联数据在知识库中应用的不足与思考

通过对文献的调研梳理和分析,我们能看到近年来关联数据作为一种新兴技术受到知识库领域的广泛重视。但是,关联数据在知识库中的研究现状也体现出了一些不足。

首先,目前的研究中存在过度依赖传统词表实现知识库关联数据转化的现象。传统知识组织理念在灵活性、知识的交叉与融合、语义描述和体现知识间关联关系方面存在不足,如等级结构或聚类揭示概念间的关系往往采取最主要的、单一的特点进行标注,同时基于传统的分类表、主题词、关键词等知识组织体系的知识组织颗粒度较粗,导致被掩盖的知识特点多于被揭示和显现的知识特点,以这样的方式无法揭示知识内部隐藏的大量有价值的数据以及知识元素之间的关系,只能提供一维、线性的组织模式,语义化程度低,造成诸多潜在特征信息被淹没在传统的知识组织理念中,显然不适应于网络信息资源组织与构建的目标需求。

其次,诸多学者在利用关联数据时对知识库进行“方块状”开发,即在数据源选取时便建立明晰的知识边界,并未调研已有的关联数据集是否可复用,造成重复劳动。例如,在农作物的关联数据构建中的实体抽取研究中^[41-42],两项研究存在明显的交集,但是由于较早发文的作者并未将关联数据处于开放获取状态,不仅使得后续研究难以承前启后,而且数据构建完成后的封闭存储也有悖于关联数据的初衷。这样的方式可能不利于关联数据的良性发展,即在空间上造成不同知识关联数据集无法进行互操作,难以在更大范围内实现关联;在时间上造成先后顺序的研究者不具备复用和引用已有研究成果的研究思路和技术条件。

最后,从知识融合角度来看,当前研究多侧重于组织独立领域内的知识,而忽视了对交叉知识的吸纳。知识库在应用关联数据时未兼顾知识间的联系,使得单独领域的知识库变成了“知识孤岛”,虽然降低了知识

组织难度,却限制了本领域研究的广度。

关联数据的理念源于计算机的实践问题,目的是将不同系统的数据进行关联。当图情学科对该项技术进行研究和利用时,难免会出现由学科理念不同而造成的实践偏差。因此,我们在积极吸纳其他领域的研究方法时,应该主动从图情领域的实践问题出发,内化该方法后,使之发展方向迎合图情领域的发展趋势。为此,本文在指出上述不足的基础上,结合关联数据的特点,对相关研究作出以下4点思考。

3.1 构建理念:网状的,而非树形的

知识库的构建理念和思想经历了从线性思维到树形思维,再到网状思维的转变,在发展过程中深受数据库开发理论基础的引导和知识组织研究实践的影响。以树形组织方式为代表的传统数据库开发方式由于历史渊源,成就斐然,时至今日,在知识库领域仍占据主导,甚至对新兴知识组织方式影响颇深。

我们在利用关联数据构建知识库时,可以保留传统组织方式的成果,但应摆脱刻板的组织方式,也不必进行非黑即白的分类。同时,在进行关联数据应用时容易忽略知识本身特点和传统数据库一维线性化的缺点,将更多的实践研究建立在信息组织的技术创新之上,这意味着虽然运用了新兴的知识组织技术,其应用理念却保留了十足的传统机制。“新瓶装旧酒”的方式虽简化了关联数据的发布流程,却很难最大化发挥关联数据的价值。因此,我们在今后的研究中应从两方面入手:首先在关联数据建模阶段对领域词表进行充分调研和专家论证,若复用已有的领域词表应对领域知识进行概念细分,在传统知识组织系统转化为关联数据时要进行非唯一性转化,一对一的映射仍具有单一线性的缺点;其次,依据属性和特征进行网络化组织,根据用户的实际知识需求抽取、确定知识元的实体概念类,并进一步确定知识元的URI命名、属性和知识元间关系。由于当前计算机技术的迅速发展,储存空间和计算能力已不再是瓶颈,良好的组织效果更应是我们追求的目标。

3.2 知识集:关联的,而非孤立的

关联数据最重要的价值在于“关联”,它支持结构化数据及非结构化数据的任意链接,利用关联数据技

术实现知识库之间的关联构建成为关联数据在知识库中发挥优势的重要体现。在研究中,我们应当将所有的知识域视为未拓之地,每个知识集的关联数据化开拓都应连点成线,化线为面,目的是利用关联数据实现知识库与外部知识域的知识关联。因此,关联数据的研究若要形成良性循环,一方面,业内学者在关联数据创建初期应积极考虑数据复用性问题,数据完成发布后对数据的开源共享性说明可借鉴国外的标注方式(论文指引性标注等);另一方面,在进行领域知识的构建时应充分借鉴和汲取已有的数据集,建立完成的数据集也应遵循相关共享协议进行知识共享,如业界已有的针对开放数据的协议ODC(Open Data Commons)非常适用于关联数据中的公开数据,形成开源的关联数据集,推动中文关联开放数据云的发展。

3.3 知识领域:融合的,而非拼接的

随着社会经济和科技研究的发展,领域知识交叉现象肩背相望,屡见不鲜。同时,也带来了交叉学科研究的兴起,越来越多的研究在朝着交叉融合的方向发展,学科或知识边界正在变得模糊。关联数据的“关联”特点在研究交叉知识时具有得天独厚的优势,其提供的数据接口能够使大量的数据或数据集之间相互关联,创建者可以任意调用不同知识库间的知识进行再创造。此举不仅可以促进交叉知识的共同发展,也方便对交叉学科研究进行动态的追踪和管理。研究人员在知识构建中应以本学科的核心知识为基础,广泛吸纳相关领域知识,为创造新的学科生长点提供积极作用。因此,在利用关联数据建设知识库的过程中,应主动吸纳具有交叉关系的知识,促进不同学科之间知识的交流碰撞、彼此借鉴,进而实现单一学科知识为其他学科研究提供研究成果和方法论贡献的同时又为本学科的研究开拓新领域、新方向。

3.4 构建目的:面向服务的,而非面向技术的

关联数据利用RDF描述数据、URI标注数据及HTTP传输数据等信息技术,再结合知识库的应用场景自然会衍生出诸多技术问题,但这些问题经过各个领域学者的长期研究和技术积淀,已有众多的理想解决方案和规范发布,如利用Drupal的关联数据发布、关联数据转化的W3C规范、知识的检索技术等一系列技术

已相对成熟。随着这些技术的发展, 关联数据的生长曲线也应转向服务和应用, 尤其是在知识库的研究视域下, 我们利用关联数据进行知识库建设是为了服务, 过多的关注技术问题将会造成“耕了别人家的地, 而荒了自家的田”。有学者也指出: 在图书馆和信息学(LIS)的硕士和博士课程中添加关联数据课程是为了未来的信息管理人员可以了解关联数据服务理念, 而不是精通关联数据技术^[43]。在今后的研究中, 领域内的学者应尝试转变研究思路, 将“为了发布关联数据而研究关联数据”的理念转变到“为了利用关联数据服务而进行关联数据研究”。为了应对这一问题, 相关学者在研究中可以积极借助计算机领域的研究成果, 如OWA算子算法^[44]、语义权值^[45]等支持自然语言检索、语义检索的关联数据应用层实现知识真正面向对象的服务。

4 结语

关联数据如何帮助知识库提升知识价值和利用率, 不仅取决于关联数据的发展方向与知识库的需求趋势之间的匹配度, 更取决于研究人员和从业人员能否以关联数据为桥梁, 不仅实现知识集间的关联, 更实现理论与实践间的关联, 即把“形而上”的技术探索 and 学术研究转变为“形而下”的可用性实践, 下沉到“寻常百姓家”, 写在“大地上”, 同时最大化利用关联数据的特性开发知识库的应用产品。知识库作为图情学界紧贴理论的实践产品, 应紧密把握信息用户的需求方向, 以关联数据为例, 相关学者应以需求为导向发展关联数据在知识库中的应用, 并针对关联数据应用的规范化问题, 制定统一和科学的引导性技术标准、政策, 在内化关联数据理念的基础上, 将研究成果外化于其他领域, 以谋逐步增强图情学科的学术影响力。

参考文献

- [1] 李聪, 王晓光. “知识库”概念的扩散与内涵演化[J]. 图书情报知识, 2012(4): 81-87.
- [2] LI Z, LI Q, ZHANG R, et al. Knowledge structures in a knowledgebase[J]. Expert Systems, 2016, 33(6): 581-591.
- [3] 赵继海. 机构知识库: 数字图书馆发展的新领域[J]. 中国图书馆学报, 2006(2): 33-36, 50.
- [4] 王思丽, 祝忠明. 利用关联数据实现机构知识库的语义扩展研究[J]. 现代图书情报技术, 2011(11): 17-23.
- [5] 刘炜. 关联数据: 概念、技术及应用展望[J]. 大学图书馆学报, 2011, 29(2): 5-12.
- [6] MCCUE J, CHIANG K, LOWE B, et al. Vivo: connecting people, creating a virtual life sciences community[J]. D-Lib Magazine, 2007, 13(7): 7-8.
- [7] BUTLER M H, GILBERT J, SEABORNE A, et al. Data conversion, extraction and record linkage using XML and RDF tools in project simile[EB/OL]. [2022-02-01]. <https://www.hpl.hp.com/techreports/2004/HPL-2004-147.pdf>.
- [8] 陈美华, 刘文云, 刘昊, 等. 国内外机构知识库建设研究[J]. 情报理论与实践, 2015, 38(9): 55-59, 63.
- [9] BAKER T. Libraries, languages of description, and linked data: a Dublin core perspective[J]. Library Hi Tech, 2012, 30(1): 116-133.
- [10] 朱丽雪. 基于Dspace的机构知识库构建[D]. 天津: 天津师范大学, 2010.
- [11] 徐华. 基于Dspace和关联数据的机构知识库合作服务模式构建研究[J]. 情报科学, 2012, 30(7): 1021-1024.
- [12] RUAN T, WANG M, SUN J, et al. An automatic approach for constructing a knowledge base of symptoms in Chinese[C]// 2016 IEEE International Conference on Bioinformatics and Biomedicine IEEE, 2016: 1657-1662.
- [13] KAMDAR M, DUMONTIER M. An Ebola virus-centered knowledge base[J/OL]. Database, 2015, 1-11 [2022-02-02]. DOI:10.1093/database/bav049.
- [14] BARTALESI V, METILLI D, PRATELLI N, et al. Towards a knowledge base of medieval and renaissance geographical Latin works: the IMAGO ontology[J]. Digital Scholarship in the Humanities, 2022, 37(1): 34-50.
- [15] 周晨. 基于关联数据的数字图书馆联合目录知识库构建研究[J]. 新世纪图书馆, 2018(7): 61-64.
- [16] 夏翠娟, 贺晨芝, 刘倩倩, 等. 数字人文环境下历史文献资源共建共享模式新探[J]. 图书与情报, 2021(1): 53-61.
- [17] 陈涛, 刘炜, 孙逊, 等. IIIF与AI作用下的文化遗产应用研究新模式[J]. 中国图书馆学报, 2020, 47(2): 67-78.
- [18] BERNERS-LEE T. Linked data[EB/OL]. [2022-02-02]. <http://www3.org/DesignIssues/LinkedData.html>.
- [19] 夏翠娟. 面向人文研究的“数据基础设施”建设——试论图书馆学对数字人文的方法论贡献[J]. 中国图书馆学报, 2020(3): 24-37.
- [20] 张华, 顾新, 王涛. 基于过程管理视角的开放式创新关系治理研究[J]. 科技进步与对策, 2020, 37(11): 1-8.

- [21] ZHANG Q Q, LIU S F, TU Q. Knowledge organization of node enterprises technological innovation under supply chain environment [J/OL]. Complex & Intelligent Systems, 2021: 1-15 [2022-02-02]. DOI:10.1007/s40747-021-00388-9.
- [22] 徐晨飞, 包平, 张惠敏, 等. 基于关联数据的方志物产史料语义化知识组织研究 [J]. 大学图书馆学报, 2020, 38 (6): 78-88.
- [23] 侯西龙, 谈国新, 庄文杰, 等. 基于关联数据的非物质文化遗产知识管理研究 [J]. 中国图书馆学报, 2019, 45 (2): 88-108.
- [24] 李永卉, 周树斌, 黄柏芝. 地方诗词资源的关联组织研究——以苏轼镇江诗词为例 [J]. 数字图书馆论坛, 2020 (2): 28-36.
- [25] LINKED CANVAS. engaging people, art and ideas [EB/OL]. [2022-02-02]. https://www.synaptica.com/wp-content/uploads/2015/03/Linked_Canvas_Factsheet.pdf.
- [26] 王敬, 王彦兵. 国外科研数据基础设施研究及实践的调研与分析 [J]. 情报资料工作, 2016 (6): 99-104.
- [27] 夏翠娟. 中国历史地理数据在图书馆数字人文项目中的开放应用研究 [J]. 中国图书馆学报, 2017, 43 (2): 40-53.
- [28] 徐玉梅. 从OCLC的数据科学项目看图书馆数据的形态演变 [J]. 图书情报工作, 2017, 61 (3): 125-130.
- [29] 陈涛, 夏翠娟, 刘炜, 等. 关联数据的可视化技术研究与实现 [J]. 图书情报工作, 2015, 59 (17): 113-119.
- [30] 夏翠娟, 刘炜, 陈涛, 等. 家谱关联数据服务平台的开发实践 [J]. 中国图书馆学报, 2016, 42 (3): 27-38.
- [31] 陈涛, 刘炜, 单蓉蓉, 等. 知识图谱在数字人文中的应用研究 [J]. 中国图书馆学报, 2019, 45 (6): 34-49.
- [32] WU Q, KUANG Y, HONG Q, et al. Frontier knowledge discovery and visualization in cancer field based on kosand lda [J]. Scientometrics, 2019, 118 (3): 979-1010.
- [33] BRÄSCHER M. Semantic relations in knowledge organization systems [J]. Knowledge Organization, 2014, 41 (2): 175-180.
- [34] 马费成, 宋恩梅. 信息管理学基础 [M]. 2版. 北京: 科学出版社, 2011: 229.
- [35] 李一秀. 图书馆语义化组织现状及对策研究——以国家图书馆实施情况为例 [J]. 新世纪图书馆, 2021 (1): 57-61, 74.
- [36] 张永娟, 刘炜, 于建荣, 等. 基于IIIF和语义知识图谱的印章资源整合与知识发现研究 [J]. 图书情报工作, 2020, 64 (7): 127-135.
- [37] SAUMURE K, SHIRI A. Knowledge organization trends in library and information studies: a preliminary comparison of the pre-and post-web eras [J]. Journal of information science, 2008, 34 (5): 651-666.
- [38] ALAMRI A. Semantic-linked data ontologies for Indoor Navigation System in response to COVID-19 [J]. International Journal of Geo-Information, 2021, 10 (9): 607.
- [39] HU S, WANG J, HOARE C, et al. Building energy performance assessment using linked data and cross-domain semantic reasoning [J]. Automation in Construction, 2021, 124: 103580.
- [40] BOTERAM F. "Content architecture": Semantic interoperability in an international comprehensive knowledge organisation system [EB/OL]. [2022-02-02]. <https://www.emerald.com/insight/content/doi/10.1108/00012531011074654/full/html>.
- [41] 任妮, 孙艺伟, 鲍彤, 等. 农业领域本体构建方法研究——以番茄病虫害为例 [J]. 情报探索, 2021 (4): 51-57.
- [42] 赵明, 杜亚茹, 社会芳, 等. 植物领域知识图谱构建中本体非分类关系提取方法 [J]. 农业机械学报, 2016, 47 (9): 278-284.
- [43] ALI I, WARRAICH N F. Linked data initiatives in libraries and information centres: a systematic review [J]. The Electronic Library, 2018, 36 (5): 925-937.
- [44] ZADEH P D H, REFORMAT M Z. Fuzzy semantic similarity in linked data using the OWA operator [C] //2012 Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS). IEEE, 2012: 1-6.
- [45] 贾丽梅, 郑志蕴, 李钝, 等. 基于动态权值的关联数据语义相似度算法研究 [J]. 计算机科学, 2014, 41 (8): 263-266.

作者简介

尚渡新, 男, 1996年生, 硕士研究生, 研究方向: 知识组织与应用, E-mail: shang-dx@qq.com。
袁润, 男, 1965年生, 硕士, 教授, 研究方向: 知识组织与服务、数字图书馆。
夏翠娟, 女, 1979年生, 硕士, 高级工程师, 研究方向: 知识组织、关联数据、数字人文。
刘桂锋, 男, 1980年生, 博士, 研究馆员, 研究方向: 知识组织与应用、科研数据管理。

Review on the Application of Linked Data in Knowledge Base

SHANG DuXin¹ YUAN Run² XIA CuiJuan³ LIU GuiFeng²

(1. Institute of Science and Technology Information, Jiangsu University, Zhenjiang 212013, P. R. China;

2. Library of Jiangsu University, Zhenjiang 212013, P. R. China; 3. Shanghai Library, Shanghai 200031, P. R. China)

Abstract: In recent years, linked data has been widely used in knowledge base. This paper systematically analyzes and combs the research status of linked data in knowledge base, summarizes the key applications of linked data in the knowledge source, visualization and semantics of knowledge base, and discusses how the linked data can promote the construction of knowledge base to meet the user's knowledge demand. At last, it is pointed out that linked data has been studied in many fields of knowledge base, but related projects fail to give full play to the characteristics of linked data, and there are problems such as over-reliance on thesaurus and neglect of interdisciplinary, etc. Subsequent research needs to pay more attention to concept construction, knowledge set association, knowledge domain fusion and goal construction.

Keywords: Linked Data; Knowledge Base; Knowledge Organization; Knowledge Management

(收稿日期: 2022-03-04)

■ 征 文 通 知 ■

2022年中国科学技术信息研究所博士后科研工作站 设站20周年学术&主题征文通知

2002年中国科学技术信息研究所获批成为国内首家具有独立招收资格的“图书情报与档案管理”学科博士后科研工作站。2022年, 20周年之际, 为回顾和总结图情档学科博士后科研工作站的发展历程和建设成就, 中信所拟汇聚学界同仁就图情档学科发展、人才队伍培养、支撑科技决策实践等内容开展征文探讨。

本次学术&主题征文均组织专家评选, 设立一、二、三等奖若干, 颁发荣誉证书并给予一定奖励。获奖稿件将推荐到《中国软科学》《情报学报》《情报工程》《中国科技资源导刊》《数字图书馆论坛》《全球科技经济瞭望》《高技术通讯》等期刊遴选发表。获奖作品将集结成册, 收录到设站20周年纪念册。

一、学术征文选题

- (1) 新时代图情档学科发展研究。
- (2) 图情档学科博士后“学—研—产”实践创新研究。
- (3) 图情档学科设站单位高水平人才培养特色研究。

二、主题征文选题

- (1) 表达对工作站设站20周年的真情实感等。
- (2) 回忆或记录在中信所从事博士后科研工作的心路历程和经历故事等。
- (3) 介绍作为博士后合作导师的指导经验方法与心得体悟等。
- (4) 回顾设站20年的发展历程, 以见证者、亲历者的身份记述工作站的发展变化, 讲述20年来各个发展时期的重大事件、重要人物、发展成果等。

三、征文要求

(1) 投稿文章必须是未公开发表的原创性研究成果。学术征文要求论据充分、数据可靠、结构严谨, 字数在6000字以上, 格式参考《数字图书馆论坛》投稿模板。主题征文体裁不限, 字数在2000字以上(诗歌除外)。

(2) 投稿请发送至huayf@istic.ac.cn(邮件主题注明“设站20周年学术征文/设站20周年主题征文”, 论文以“文章名称—作者姓名”的方式命名, 稿件内附个人简介, 包括作者姓名、联系方式、工作单位、职务等基本信息)。

(3) 投稿截止日期为2022年6月30日。

联系人: 花老师

联系方式: 010-58882046