

敦煌壁画叙词表构建与关联数据发布*

王晓光 侯西龙 程航航 夏生平

摘 要 叙词表在知识组织中发挥着重要作用,然而敦煌壁画领域叙词表的缺失,限制了敦煌壁画研究及其价值的挖掘。通过建立一套规范、全面的敦煌壁画领域叙词表,可以为网络环境下敦煌壁画数字资源语义标注、信息检索和知识组织等提供智力和技术支撑。本文详细阐述敦煌壁画叙词表的构建与其关联数据发布过程,采用自顶向下与自底向上相结合的领域叙词表构建方法,提高叙词表建设效率的同时保证叙词表的高质量。在此基础上,研究实现敦煌壁画叙词表关联数据发布和可视化,与外部词表进行语义关联,并提供开放数据服务。敦煌壁画叙词表的构建和编制,可促进敦煌壁画领域数字人文应用的开发,也为我国文化遗产领域受控词表的建设提供了重要参考和借鉴。图8。表5。参考文献40。

关键词 敦煌壁画 叙词表 知识组织系统 关联数据 数字人文

分类号 G254.2

The Construction and Linked Data Publishing of Dunhuang Mural Thesaurus

WANG Xiaoguang, HOU Xilong, CHENG Hanghang & XIA Shengping

ABSTRACT

Thesauri play an important role in knowledge organization. Nevertheless, the absence of Dunhuang mural thesaurus sets limits on Dunhuang mural studies and its value exploration. The development of the thesaurus is a foundation work of information resource management of Dunhuang murals, which can not only provide a controlled and structured vocabulary for the standardized information resource organization of Dunhuang murals but also provide a clear knowledge system for Dunhuang murals researches. In order to speed up the construction of thesaurus and meet the needs of dynamic updating of knowledge organization, the development of thesaurus needs to adopt the machine assisted semi-automatic cooperative development strategy, and give full play to the advantages of machine and experts, and realizes the rapid development and maintenance of thesaurus through human-computer cooperation.

This paper presents the method and procedure of the constructing and linked data publishing of Dunhuang mural thesaurus. A combination of top-down and bottom-up approach was used to develop the thesaurus.

* 本文系国家自然科学基金创新研究群体项目“信息资源管理”(编号:71921002)和湖北省自然科学基金创新研究群体项目“大数据资源语义组织与管理应用研究”(编号:2019CFA025)的研究成果之一。(This article is an outcome of the project “Information Resource Management”(No. 71921002) supported by National Natural Science Foundation of China and the project “Semantic Organization and Management of Big Data Resources”(No. 2019CFA025) supported by Natural Science Foundation of Hubei Province of China.)

通信作者:王晓光,Email:wxguang@whu.edu.cn,ORCID:0000-0003-1284-7164 (Correspondence should be addressed to WANG Xiaoguang,Email:wxguang@whu.edu.cn,ORCID:0000-0003-1284-7164)

The top-down approach starts at the higher-level categories, and establishes a general framework for the thesaurus based on the objectives of the thesaurus. The basic structure of the thesaurus was developed by referring to the fundamental dictionary of Dunhuang studies and Art & Architecture Thesaurus (AAT). The thesaurus contains 5 facets including agents, objects, activities, time and physical attributes. Sub-categories were defined for each facet with total 25 second-level categories. The bottom-up approach builds up lower-level categories from the concepts extracted from the corpus with the help of natural language processing technology, and through the process of manually grouping and classification of the terms, definition of the semantic relations, adjustment and optimization of the structure. At present, Dunhuang mural thesaurus contains 4 276 terms. After that, this paper studied and implemented linked data publishing of the Dunhuang mural thesaurus. The thesaurus was semantically represented and described with SKOS model, and the terms were mapped to AAT concepts using the SPARQL query. Finally, about 21.2% of the terms was linked to AAT. In the meantime, a linked data service platform of Dunhuang mural thesaurus was produced, providing concepts browsing, term retrieval, visualization, open data service, etc.

The originality of this research lies in the compilation of a controlled vocabulary for the Dunhuang murals with standard specific semantic relations and authoritative multi-source scope notes. The thesaurus comprehensively covers mural protection and restoration, archaeology, iconography, human geography and other research perspectives. The construction of Dunhuang mural thesaurus lays an important foundation for the semantic annotation, information retrieval and knowledge organization of Dunhuang mural digital resource. It also has an important exemplary significance for the development of smart data resources in the field of cultural heritage in China. The future work contains further studies to improve the thesaurus's structure, categories, and terms. Systematic evaluation of the thesaurus remains to be studied. 8 figs. 5 tabs. 40 refs.

KEY WORDS

Dunhuang murals. Thesaurus. Knowledge organization system. Linked data. Digital humanities.

0 引言

叙词表是规范化组织和受控的词汇表,以特定的结构排列,显示词汇之间的关系,是知识组织最常用的工具,常应用于信息检索系统中^[1]。叙词表由特定领域内表达事物概念的术语及术语之间各种语义关系组成,具有词汇控制和概念关联的功能,其知识描述方式具有形式化、结构化的特点^[2]。叙词表的互操作性及其在关联数据中的应用,使其在知识组织与信息资源组织方面起着重要的补充作用^[3]。

在文化遗产及人文艺术领域,叙词表的建设已取得一些成果。人文艺术领域较为知名的叙词表有盖蒂词表(Getty Vocabularies)和荷兰

图像志分类系统(Iconclass);文化遗产领域有法国建筑与遗产部创建的与可移动文物叙词表^[4],欧洲遗产网建立的欧洲遗产多语言叙词表(HEREIN Thesaurus)^[5]。英国在建筑、遗址、纪念碑等文化遗产的保护上也制定了一系列词表^[6],英国维多利亚与艾伯特博物馆的“中国图像志索引典”(Chinese Iconography Thesaurus, CIT)项目创建了首个用于中国艺术图像志研究的叙词表^[7]。我国在叙词表编制方面也有丰富的经验,先后编制了逾百部叙词表^[8],其中以《汉语主题词表》和《中国分类主题词表》最有代表性,但在文化遗产与人文艺术细分领域几乎为空白,这在一定程度上影响了我国文化遗产信息资源的开发管理与传承传播。

敦煌壁画叙词表的编制是敦煌壁画相关信

息资源管理的基础性工作。构建敦煌壁画叙词表,不仅可以为敦煌壁画信息资源的规范组织提供受控术语,还能对敦煌壁画研究提供简略清晰的知识体系,是敦煌壁画相关的智慧数据资源建设的重要组成部分。本文围绕敦煌壁画叙词表的编制,系统阐述自顶向下与自底向上结合的叙词表设计与构建方法,在此基础上研究叙词表语义组织过程,利用语义网技术实现叙词表的关联数据发布,提供叙词表可视化和开放数据服务。本研究对我国文化遗产领域的智慧数据资源建设具有重要的示范意义。

1 相关研究综述

1.1 叙词表构建方法研究

叙词表作为一种人工情报语言,一直都以专家手工构建为主。近年来,随着计算机技术的发展,半自动化或自动化编制叙词表开始兴起。目前,叙词表自动构建方法有基于自然语言处理、基于WEB挖掘和基于wiki挖掘的方法^[9]。其中,基于自然语言处理的方法将统计学方法和语言学方法结合起来,从语料库中自动提取概念并识别概念间的语义关系,核心工作涉及术语和主题词的选定,以及词间关系处理等多个方面。在术语提取和新术语发现方面,张昱提出了基于启发式规则的、多测度融合的新词发现方法^[10];Mohsen研究了阿拉伯文词表的自动构建方法,并对TF-IDF、PMI、LSA三种概念抽取方法的准确性进行了评价^[11];雷晓基于术语的分布特征进行领域新术语的识别,有效识别了领域的低频词^[12];曾文认为主题词表自动化构建应理解为一种辅助性质的自动化构建技术,中文主题词表的构建必须结合学科发展和汉语语言特点^[13];Yang和Wu采用自顶向下与自底而上相结合的方法构建了地震灾害领域分类表^[14];司莉对网络叙词表的现状进行调研,发现叙词表的编制向多语言化、分面化、可视化与互操作化的方向发展^[15];赵捷认为协同编制机制能够实现不同系统和不同用户对叙词表系统的互操作,是网络环境下叙词

表编制的发展趋势^[16]。

总的来说,叙词表的编制整体上呈现出构建方式网络化与半自动化、更新维护动态化、应用方式多样化的特点与趋势^[17],但是现阶段还远达不到自动化水平。为了加快叙词表的构建速度,满足知识组织动态更新的需求,叙词表的编制还需采用机器辅助的半自动协同编制策略,同时发挥机器和专家各自的优势,通过人机协同实现叙词表快速编制和维护。

1.2 叙词表关联数据发布

关联数据技术是轻量级的语义网实现方式,它使得海量异构数据资源的形式化、语义化、关联化及共享化成为可能,以其独特优势引起学界和业界多方关注。近年来,国内外对叙词表的语义化表示与关联化发布开展了众多研究与实践,越来越多的知识组织系统开始借助关联数据技术实现互通互联。

在叙词表的语义化描述与表示方面,W3C发布的简单知识组织系统(SKOS)标准提供一种轻量级的描述机制,能够较好地支撑叙词表的语义化表示^[18]。围绕我国《汉语主题词表》《中国分类主题词表》等综合性叙词表的SKOS表示问题,曾蕾^[19]、范炜^[20]、曾建勋^[21]、欧石燕^[22]等学者分别从不同角度进行了探索。曾新红研究扩展SKOS以实现对中国叙词表和分类法等传统受控词表的全描述^[23];鲜国建对《中国农业科学叙词表》采用SKOS和SKOS-XL语言实现语义化描述与关联数据发布^[24];石泽顺等对图情学科SKOS构建进行实证研究,并构建了图情学科叙词表检索系统^[25]。SKOS被广泛应用于分类法、叙词表、主题词表的语义化描述,成为实现知识组织系统语义化、关联化的首选途径。

在叙词表关联发布与服务方面,曾蕾从关联开放数据集生产者、词表生产者和词表使用者等多维度探讨了关联数据发布中词表的作用^[26],并提出语义网框架下的知识组织系统作为开放数据应该遵循的FAIR原则^[27];马费成从服务视角对知识组织系统研究进展进行系统

总结,认为词表与本体是知识组织系统提供服务的重要工具,可视化是提供服务的有效手段^[28];陶俊将词表语义化组织研究演进划分为“语义描述、本体转化、发布为关联数据”三个阶段^[29];贾君枝将知识组织系统的发展归纳为描述对象概念化、语义关系多样化、描述语言形式化、数据开放关联性四个特点^[30]。

综合国内外的研究可以发现,采用关联数据等技术在网上发布词表等知识组织工具,并与外部开放数据集建立关联关系,使其成为关联开放数据的一部分,是目前知识组织系统的发展趋势。国内在叙词表建设及其关联数据发布研究方面主要集中在叙词表语义化描述方法与模型和针对特定叙词表语义化转换实践两个方面。关联数据可视化作为关联数据探索性发现与分析的工具,是关联数据发布的重要环节,能帮助用户更好地理解其数据结构^[31],目前国内在叙词表语义化表示后的关联数据发布、词表间的关联匹配、质量控制以及可视化方面的研究仍不够深入^[32]。

2 敦煌壁画叙词表的构建

敦煌壁画叙词表是针对敦煌壁画领域编制

的,旨在为敦煌壁画相关信息资源进行语义标注、知识组织等提供一套规范的受控词表。敦煌壁画叙词表是敦煌学研究基础设施建设中的重要组成部分,在壁画内容细粒度语义知识标注、抽取、加工、挖掘和集成展示等方面具有重要意义,是敦煌壁画知识开放互联和传播共享的基础。

敦煌壁画叙词表的构建采用计算机辅助的自顶向下与自底向上相结合的方法。自顶向下与自底向上相结合的方法被认为是构建分类体系的最佳实践方法^[14];自顶向下的方法根据词表编制的目的从顶层宏观级别建立一个整体框架,提供对词表顶级类别的控制;自底向上的方法则从语料库中提取领域概念或术语,通过概念聚类与归类构建低级别的细分类别。具体来说,在叙词表结构设计时,首先由专家分析人文领域叙词表与敦煌学相关词典,进而设计叙词表结构,确定叙词表的初步框架^[33]。然后,从语料库中进行机器分词和新词发现,经过多人协作的主题词归类和领域专家的校对与审核,自底向上地扩充词表术语,同时不断调整优化词表结构,实现叙词表的迭代扩展与完善。具体的构建流程如图1所示。

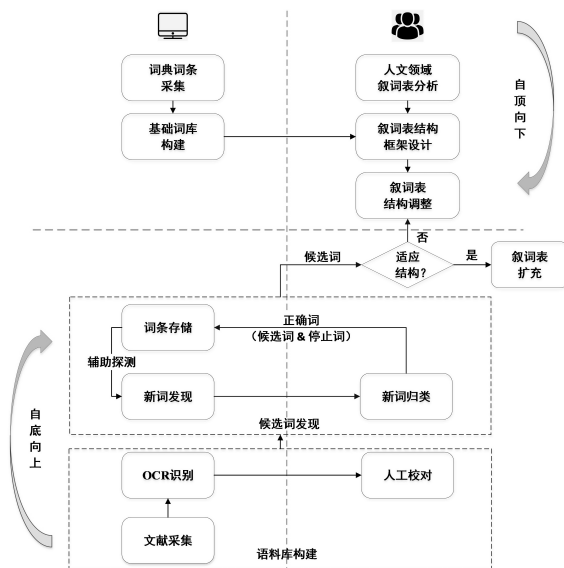


图1 敦煌壁画叙词表构建流程

2.1 自顶向下的叙词表结构设计

叙词表的结构直接决定了叙词表的功能与应用,为充分发挥叙词表的功能,必须为其设计合理的结构。为更好地理解壁画涉及的主题,我们对敦煌壁画研究文献进行了内容分析,进而识别了敦煌壁画涉及的主题及关系结构^[34]。同时,我们对国际知名的艺术与建筑叙词表(AAT)进行了调研,深入分析了 AAT 在领域通用性、组织结构、注释与元数据规范等方面的特征。

AAT 是多层级结构化的叙词表,其层级结构包括分面、层级、引导词和概念。具体分为代理者、物理材料、相关概念等八个分面,一个分

面可以看作是某种方式聚集的一组概念的集合;通过分面可以把抽象的概念组织成为具体的、实际的文物^[35]。每个分面下包含多个层级,层级下包含引导词和概念;层级与引导词主要用来创建分类层级体系,但都不可用于建立索引或编目^[36]。参考 AAT 词表结构,本文将敦煌壁画叙词表的层级结构分为分面、层级、概念、实例四个类型。其中分面是最高等级,直接在叙词表体系下;层级与概念在分面之下,层级主要用来构建词表层级结构,通过层级与概念的混合使用,构建多層级的分类结构;实例表示某个概念或层级包含的具体实例对象。敦煌壁画叙词表的层级结构与逻辑关系如图 2 所示。

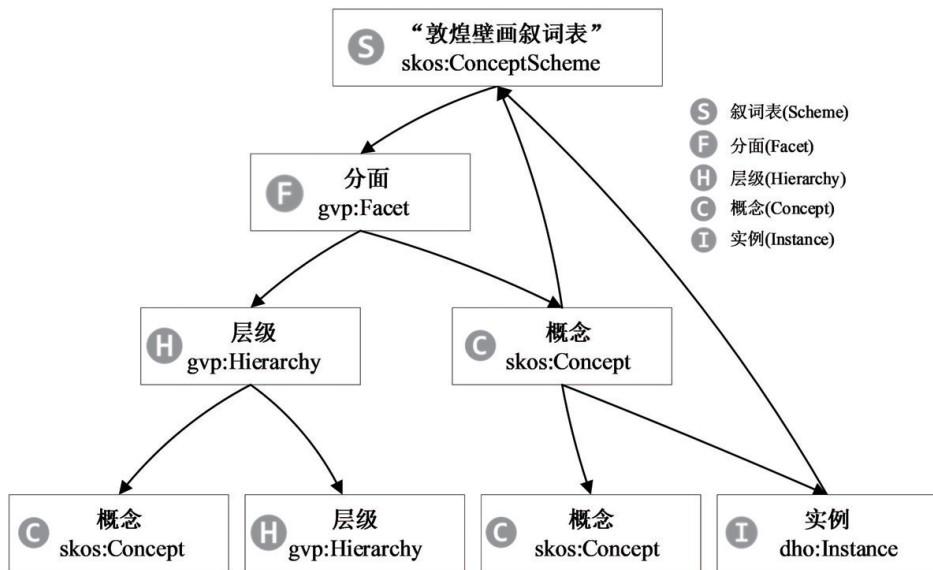


图 2 敦煌壁画叙词表的层级逻辑关系

敦煌壁画的内容主题丰富,涉及宗教、史地、美术、乐舞、民俗等诸多领域。敦煌壁画相关的研究也涉及考古、壁画保护与修复、图像志、人文地理等多个方面。为便于检索和反映敦煌壁画相关概念及其结构,我们在参考《敦煌学大辞典》《敦煌石窟内容总录》《敦煌人物志》等敦煌学基础文献后,设计了叙词表的基本结构。如图 3 所示,叙词表包括五大分面,分为代理者、物理特质、活动、时间、物件,每个分面下

设置二级类目,共设置 25 个二级类目。

2.2 自底向上的叙词表扩展优化

自底向上的叙词表扩展优化采用自然语言处理技术,从专业性语料库中自动提取领域候选词,经过人工的主题词归类、概念间关系的定义、结构优化以及质量校对等过程,最终实现人机协同的叙词表内容不断扩充及词表结构的优化。

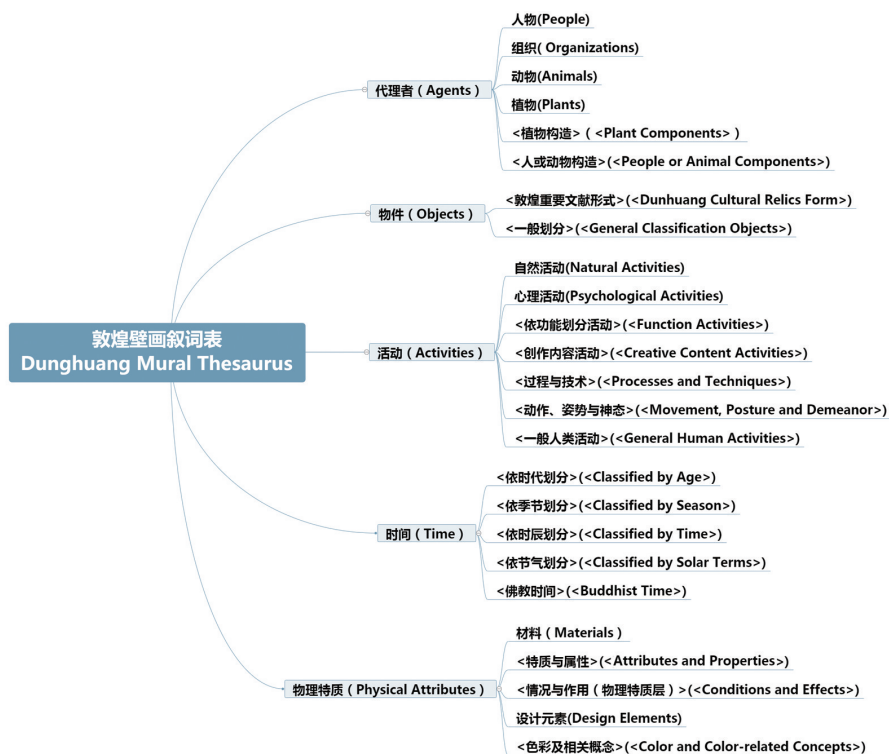


图3 敦煌壁画叙词表的顶层框架结构

2.2.1 领域概念与术语的收集

通过网络机器人采集与人工收集方式,我们收集了敦煌学基础辞典《敦煌学大辞典》及两本敦煌学中文权威期刊《敦煌研究》《敦煌学辑刊》自发刊以来与敦煌壁画相关的700余篇论文。利用OCR技术与人工校对相结合的方式对文献进行格式转换,进而构建了适用于机器学习的敦煌壁画初始语料库,保证了词汇的全面性和完整性。

2.2.2 领域主题词的发现与归类

采用词典分词和新词发现相结合的方式对敦煌学文献进行中文分词与新词发现。借助Jieba中文分词工具提取敦煌壁画领域候选词。Jieba中文分词是基于词频度统计的分词方法,其采用动态规划查找最大概率路径,找出基于词频的最大切分组合;对于未登录词,其采用基于汉字成词能力的隐马尔科夫模型,并使用了

Viterbi算法。在分词过程中通过动态调整词典,调节单个词语的词频,使其能(或不能)被切分出来。对《敦煌学大辞典》与研究论文进行分词处理,分别得到30854和122991个新词汇;之后对《敦煌学大辞典》的分词结果进行分类,将发现的新词分为相关词、停止词和错误词三类;经过统计计算,分词结果正确率达到72.13%,包括58.50%的相关词和13.63%的停止词。敦煌壁画相关研究论文的分词结果仍在进一步分类审核过程中。

通过对错误词的分析后发现,分词工具对古代国外地名、古代官职、朝代年号分类效果不佳,未能良好地适应文化艺术领域的部分特征。为此,本文对算法进行了基于词典和规则的优化,通过收集整理中国古代的朝代年号、古代地理名词、佛学规范库等丰富自定义词典,同时针对官职和数词构建语法规则匹配式正则表达

式。由于敦煌学领域专业词汇较为偏僻,传统的中文训练语料库中很难涉及此类词汇。未来我们将优化算法,从而达到更精确的分词效果。

由于敦煌壁画叙词表建设处于冷启动阶段,缺乏专业的训练数据集,难以实现基于机器学习的词汇自动分类与扩充,对提取的候选词仍需要通过人工方式进行归类。项目召集具有相关背景知识的标引员,经过分类培训后,周期性地分配标引与归类任务,对候选词库中的术语进行归类。然后由领域专家对新增术语进行审核,保留合格词汇,反馈不合格词汇。同时,我们对叙词表前三层级的词汇进行词族词性分析,“代理者”“物件”分面下一般多为名词,“活动”分面中动词占绝大多数,“物理特质”分面包含众多形容词,“时间”分面的术语多为时间副词与名词。项目使用斯坦福大学自然语言处理工具进行词性标注,辅助标引员在候选词归类时进行参考,从而加快候选词归类速度。

2.2.3 叙词表结构调整与优化

在对候选词进行归类与审核时,如果词表结构不能适应新的词汇,则需要考虑叙词表结构的调整,以使其更加科学合理。在叙词表宏观结构框架下,根据主题词的成组与归类情况,确定更细级别的类目。在细化类目时,充分考虑了敦煌壁画领域的特殊性。比如,在“代理者”分面中区分敦煌壁画出现的“佛家神祇”和“世俗人物”,在“时间”分面中增加“佛教时间”特有层级,“活动”分面设置“动作、姿态与神态”层级来描述壁画描绘的人物或动物的姿势、动作及神态,“物理特质”分面则包含壁画特有的材料、病害、状况等物理属性以及壁画绘制的图案、装饰等设计元素,在“物件”分面除了突出敦煌重要文献形式外,在一般物件层级侧重记录佛教基本概念。通过不断的迭代,充分发挥人机协同编制的优势,实现叙词表的扩展与优化。

2.3 叙词表管理与维护

为提高敦煌壁画叙词表协同编辑、术语管理、词表结构与词表发布等方面的科学管理,我

们利用 TemaTres 开源词表管理系统对叙词表进行管理与维护。TemaTres 能够提供词表中词汇及其关系的编辑与存储,支持多种词表间关系的构建,具有全文检索、等级浏览等功能;其在词表结构定义、一致性控制、数据输出、词表显示、多语言支持及词表互操作等方面表现比较突出^[37]。通过借助 TemaTres 系统在线协同编辑的方式,对提取的主题词进行范围注释、定义注释与图像注释,并细化主题词之间的语义关系。为确保词汇注释来源的专业性与权威性,依据《佛学大辞典》《佛学常见辞汇》《汉语大词典》《敦煌石窟知识辞典》《牛津英语大词典》《朗文英文词典》等中英文权威词典对术语进行定义。在图像注释上,创新性结合国际图像互操作框架(International Image Interoperability Framework, IIIF)实现词汇与数字图像的关联,用以解释与描述相关术语;采用图像注释不仅丰富了词表注释形式,亦为壁画及其元素提供了视觉特征上的理解,便于用户理解与学术研究。此外,还邀请不同领域专家对叙词表质量进行周期性审核与校对,以实现叙词表的不断更新和维护。

当前已发布的敦煌壁画叙词表的规模如表1所示。未来我们将会对更多的语料资源进行加工,丰富机器学习的语料库,以发现更多的敦煌壁画相关词汇,从而扩充与完善叙词表;同时,进一步明确领域实例词的定义与范围,扩展实例词的数量。本文采用机器辅助的半自动协同编制策略,构建了一个规范、全面的敦煌壁画领域叙词表。叙词表包含的丰富概念及其相互

表1 敦煌壁画叙词表的统计数据

项 目	数量
词汇总量	4 276
分面数量	5
层级术语数量	83
主题词数量	3 199
实例词数量	989
候选词	18 050

之间的层级与语义关系,不仅呈现了敦煌壁画蕴含的知识网络,也为数字人文环境下敦煌壁画的知识组织、信息检索与知识图谱构建等提供了基础性的智力支持。

3 敦煌壁画叙词表关联数据的发布

关联数据技术能够在不同数据集之间建立关联,形成一个结构化且富含语义信息的数据

网络,从而实现资源的真正共享。关联数据为知识组织提供了语义化的技术、方法、工具和实现路径,国内外众多具有影响力的知识组织系统陆续发布为关联数据。将敦煌壁画叙词表以关联数据的形式进行语义发布,能够为世界范围内的信息资源共享奠定重要的语义基础。敦煌壁画叙词表关联数据发布遵循 W3C 词表 RDF 发布的最佳实践方法^[38],具体发布流程如图 4 所示。

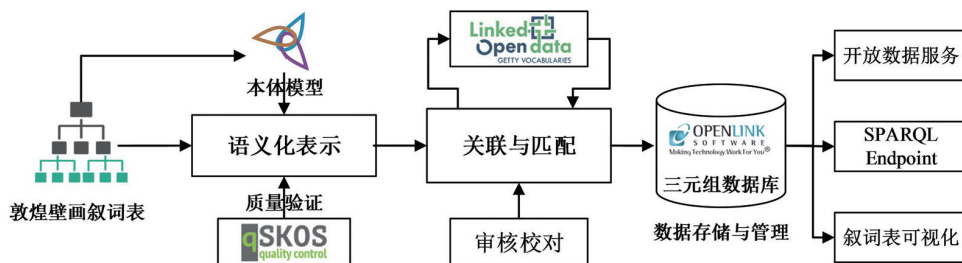


图 4 敦煌壁画叙词表关联数据发布流程

首先,在参考 SKOS 模型与盖蒂词表本体模型的基础上,建立敦煌壁画叙词表本体模型以规范叙词表的语义化描述;其次,按照叙词表本体模型对叙词表进行语义化转换,并对转换后的数据质量进行检验;然后,通过 SPARQL 查询的方式与 AAT 进行概念关联匹配;最后,使用三元组数据库进行叙词表关联数据的存储与管理,完成叙词表关联数据发布,提供对外开放服务。在叙词表关联数据集的基础上,本文使用 Apache Jena 框架开发了敦煌壁画叙词表关联数据服务平台^①,平台提供基于 Web 的用户界面,并提供面向机器可读与可理解的开放数据接口。

3.1 叙词表的本体模型

本体是关联数据集的语义结构主体,所有要素按照本体的结构进行规范化表示与组织。

在敦煌壁画叙词表逻辑结构的基础上,本文定义了敦煌壁画叙词表本体以规范叙词表的语义化描述与表示。依据本体构建的原则与方法,复用盖蒂词表的 GVP 本体^②、SKOS 以及 DCMI 元数据标准,本文定义的敦煌壁画叙词表本体模型如图 5 所示,并实现了叙词表本体模型的发布^③。

根据上文对敦煌壁画叙词表层级逻辑结构的设计,叙词表本体模型共定义 5 个核心类,其中分面和层级类复用 GVP 本体,概念类使用 SKOS 中的 Concept 类进行表示,自定义实例类作为 skos:Concept 的子类。本体模型的对象属性包含等级关系、等同关系和相关关系等,具体体现为术语间的上位词、下位词、同义词、相关术语等。本体模型的数据属性复用 DC 元数据标准描述概念的注释信息来源、创建与修改时间等信息,具体属性如表 2 所示。

① <http://dh.whu.edu.cn/dhvocab/home>

② <http://vocab.getty.edu/ontology>

③ <http://dh.whu.edu.cn/dhvocab/ontology#>

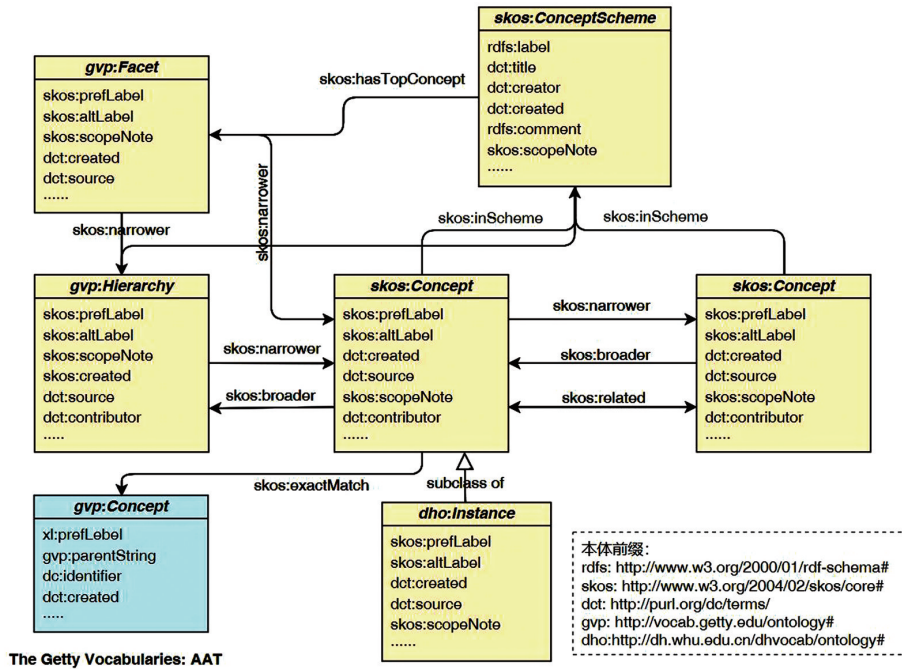


图 5 敦煌壁画叙词表本体模型

表 2 敦煌壁画叙词表本体模型的属性列表

属性名称	定义域	值域	注释
skos:hasTopConcept	skos:ConceptScheme	skos:Concept	顶级概念属性;本文用于表示叙词表的分面。
skos:broader	skos:Concept	skos:Concept	直接上位概念
skos:narrower	skos:Concept	skos:Concept	直接下位概念
skos:related	skos:Concept	skos:Concept	相关术语
skos:exactMatch	skos:Concept	skos:Concept	准确匹配;本文主要用于与外部词表的关联。
skos:inScheme	skos:Concept gvp:Facet gvp:Hierarchy	skos:ConceptScheme	某个概念在该叙词表中
skos:prefLabel	rdfs:Resource	rdfs:Literal	首选语词标签;rdfs:label的子属性。
skos:altLabel	rdfs:Resource	rdfs:Literal	非首选语词标签;rdfs:label的子属性。
skos:scopeNote	rdfs:Resource	rdfs:Literal	范围注释;通用注释 skos:note 的子属性。
dct:created	rdfs:Resource	rdfs:Literal	主题词创建时间
dct:contributor	rdfs:Resource	rdfs:Literal	主题词贡献者
dct:source	rdfs:Resource	rdfs:Literal	数据来源,记录主题词或注释信息的来源。

3.2 叙词表语义化表示

在本体模型的指导下,本文采用 SKOS 模型进行叙词表的语义描述和表示。遵循关联数据的四原则,将每个主题词表示为相应的资源类型,为叙词表中的每个主题词定义唯一且可网络解析的 HTTP URI。主题词 URI 采用“http://dh.whu.edu.cn/dhvocab/{编号}”的规则,如“菩萨”的唯一标识符为“http://dh.whu.edu.cn/dhvocab/tema239”。主题词之间的层级

关系、等同关系及相关关系则通过 SKOS 的对象属性建立,同时借助 D2RQ、OpenRefine、RDFizer 等转换工具实现叙词表的自动语义化转换。如表 3 所示,以主题词“菩萨”为例,对其进行 SKOS/RDF 序列化描述。首先,对其进行编号并赋予唯一的 HTTP URI,其中文名称使用“skos:prefLabel@zh”标签描述,并描述其范围注释、创建时间等属性;然后对其上下位词、等同概念等语义关系进行描述。

表 3 主题词“菩萨”的 SKOS 描述示例

<pre> <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" xmlns:skos="http://www.w3.org/2004/02/skos/core#" xml:base="http://dh.whu.edu.cn/dhvocab/" > <rdf:Description rdf:about="tema239"> <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/> <skos:prefLabel xml:lang="zh">菩萨</skos:prefLabel> <skos:altLabel xml:lang="zh">菩提萨埵</skos:altLabel> <skos:scopeNote xml:lang="zh">梵文音译全称为“菩提萨埵”,意为“觉有情”,即“上求菩提(觉悟)、下化有情(众生)”的人。佛教指达到自觉(自身得到解脱)、觉他(使众生达到解脱)两项修行果位者……—汉语大词典 </skos:scopeNote> <skos:broader rdf:resource="tema204"/> <skos:narrower rdf:resource="tema2388"/> <skos:narrower rdf:resource="tema2389"/> <skos:exactMatch rdf:resource="http://vocab.getty.edu/aat/300264360"/> <skos:inScheme rdf:resource="http://dh.whu.edu.cn/dhvocab/" /> </rdf:Description> </rdf:RDF> </pre>	<pre> <命名空间信息> 资源类型:skos:Concept 首选标签:菩萨 非首选标签:菩提萨埵 范围注释(中文):意为“觉有情”,即“上求菩提(觉悟)、下化有情(众生)”的人。佛教指达到自觉(自身得到解脱)、觉他(使众生达到解脱)两项修行果位者……—汉语大词典。 上位词:佛家神祇 下位词:普贤菩萨 月光菩萨 等同概念:aat:300264360 所属词表:敦煌壁画叙词表 </pre>
---	---

使用 SKOS 进行叙词表语义描述和转换过程中,往往会出现文档、语法、字符与编码等方面的问题,影响叙词表的关联数据发布与使用。本文采用 qSKOS 工具^①对敦煌壁画叙词表语义化转换的 SKOS 数据进行正确性和一致性检测;qSKOS 是 W3C 推荐的 SKOS 文档质量检测工

具,支持对 SKOS 中的标签和文档、结构、关联数据,SKOS 表达四个方面 29 个具体问题的一致性检验^[25]。本文采用本地命令行的形式对 SKOS 文档进行问题检验,根据生成的检测报告对 SKOS 文档进行错误纠正和优化,为叙词表关联数据的发布及其使用提供了质量保证。

① <https://github.com/cmader/qskos>

3.3 与 AAT 的关联匹配

在不同 RDF 数据集之间建立关联是关联数据标准的基本要求。虽然敦煌壁画叙词表收录的多为该领域的专业词汇,但仍有一部分通用性概念术语;为丰富与增强叙词表的关联性、互操作性与开放性,我们将其与 AAT 中的词条进行关联。采取以概念为单位的对应分析方法,即以该概念所代表的意义为主要考量,从 AAT 中判断出与其精确等同的概念词汇。使用 W3C 推荐使用的 `skos:exactMatch` 属性连接两个词表间语义相同的概念,表示彼此可以在所属的词表体系中进行互换。

通过关键词查询的方式,编写 SPARQL 查询语言对 AAT 词表进行检索,具体查询语句如表 4 所示。由于 AAT 使用的是中文繁体,将术语简繁转换后搜索 AAT 的首选标签和非首选标签中包含关键词的词条,检索结果经过人工判断后存储在数据库中。本文通过编写程序代码自动实现对 AAT 的 SPARQL 查询及查询结果的三元组关联。最终,在构建的敦煌壁画叙词表关联数据集中,五个分面共包含术语 3 853 个,其中与 AAT 关联的术语共 816 个,三元组数量共 27 500 余个,具体数据统计结果如表 5 所示。

表 4 查询 AAT 的 SPARQL 语句

```
prefix skos: <http://www.w3.org/2004/02/skos/core#>
prefix gvp: <http://vocab.getty.edu/ontology#>
select ?subj ?label ?title where {
    ?subj skos:prefLabel | skos:altLabel "keyword"@zh-Hant.
    ?subj gvp:parentString ?title.
    ?subj skos:prefLabel ?label. FILTER (lang(?label) = 'en').
}
```

表 5 敦煌壁画叙词表关联数据集统计结果

敦煌壁画叙词表分面	分面包含的术语数量	与 AAT 关联的术语数量	比例 (%)
代理者分面	911	144	15.8
时间分面	131	58	44.3
活动分面	494	128	25.9
物件分面	2 075	384	18.5
物理特质分面	242	102	42.1
总计	3 853	816	21.2

叙词表间的关联仅映射了等同关系,叙词表中概念间同义、近义、相关等语义关系的数量较少。今后还需要在领域专家与知识组织专家的帮助下,经过人工加工与校对的过程,完善叙词表的映射与审核工作。

4 敦煌壁画叙词表可视化与开放数据服务

4.1 叙词表访问与检索

敦煌壁画叙词表关联数据服务平台面向用

户通过 Web 交互界面,提供概念解析、概念浏览、主题导航、智能检索和术语服务等关联数据常规服务。在叙词表检索服务方面,普通用户可以设置检索的范围和条件,通过关键词进行全文模糊检索或精准检索;专业用户则可以编写 SPARQL 查询语句进行叙词表高级检索。如图 6 所示,平台基于 YASGUI 库^①实现了 SPARQL 语句的自动补全和基本语法检测功能,用户在输入框中编写 SPARQL 查询语句,查询

^① <https://doc.yasgui.org>

结果显示符合查询条件的概念术语及其 URI。

DH 敦煌壁画主题词表关联数据 Search

敦煌壁画主题词表关联数据SPARQL查询

选择查询的模板，在编辑框中输入 SPARQL 查询语句，点击右上角查询按钮执行查询。

1 主题词表本体相关查询

- 1.1 本体包含的类
敦煌壁画叙词表本体模型的类。
`select DISTINCT ?class
where{[] a ?class}`
- 1.2 查询所有谓词
敦煌壁画叙词表本体模型使用的属性。
`select DISTINCT ?
properties where{?subject ?
properties ?object}`

2 主题词相关查询

- 2.1 主题词模糊查询
查询首选项邻域中含有“菩萨”的主题词。
`select * where{?concept
skos:prefLabel ?label.
filter(contains(?label,“菩`

```
1 PREFIX gvp: <http://vocab.getty.edu/ontology#>
2 PREFIX dc: <http://purl.org/dc/elements/1.1/>
3 PREFIX ddt: <http://purl.org/dc/terms/>
4 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
5 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
6 PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
7 PREFIX dunhuang: <http://dh.wbu.edu.cn/divocab/>
8 select distinct ?Concept where {[] a ?Concept}
```

查询结果:

Table Raw Response

Showing 1 to 50 of 97 entries

concept	label
dunhuang:tema2432	*严土菩萨
dunhuang:tema2424	*快乐菩萨
dunhuang:tema2419	*供养菩萨
dunhuang:tema2568	*维摩菩萨

图6 SPARQL 查询页面

此外,平台面向计算应用程序提供 RDF\XML、JSON、JSON-LD、N-Triples 和 Turtle 不同的数据序列化格式,并提供 SPARQL Endpoint 开放数据获取端口,为应用程序开发与机器理解提供所需的数据接口和可处理的数据。

4.2 叙词表可视化

关联数据的序列化着眼于机器处理,难以

供人们直观地理解与有效识别概念间的语义关系。因此,叙词表关联数据的可视化十分必要。通过叙词表可视化,可以降低叙词表的认知难度,实现从专业化叙词表到适用于大众用户利用的过渡。为了更直观展示敦煌壁画叙词表的结构与内容,在叙词表关联数据发布的基础上,平台提供知识图谱、旭日图、圆锥图及树状图等多种形式的叙词表可视化,如图7所示。

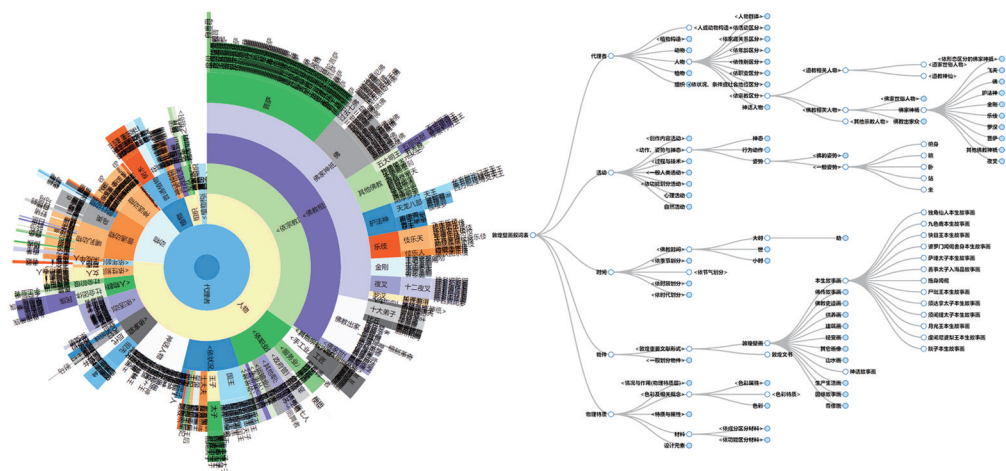


图7 敦煌壁画叙词表旭日图与树状图可视化

此外,为便于用户直观浏览叙词表的语义关系,平台基于 LodLive^① 插件以及 Ontodia 开源库^[39] 对叙词表关联数据以知识图谱的形式进行可视化与浏览,如图 8 所示。图中每个节点代表

一个概念术语,概念术语之间的关系用方向箭头连接,包括与 AAT 关联概念的属性。通过配置多个 SPARQL Endpoint,实现不同数据集中关联数据的展示,可以发现更多资源信息^[40]。

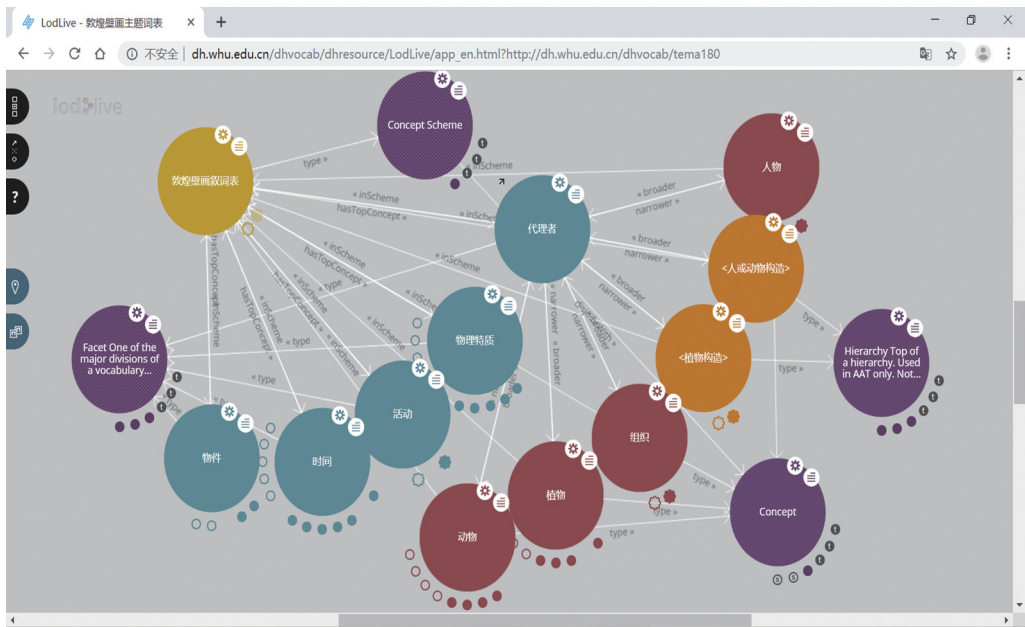


图 8 敦煌壁画叙词表语义关系可视化

5 总结与展望

围绕敦煌壁画叙词表的建设和研究,本文深入探讨了叙词表的构建方法及其关联数据发布的过程。自顶向下与自底向上相结合的机器辅助词表构建方法,丰富了领域叙词表编制理论与方法,在提高叙词表建设效率的同时,保证了叙词表的质量。遵循关联数据与词表语义发布最佳实践,实现了敦煌壁画叙词表的关联数据发布,并与艺术与建筑词表进行语义关联,增强了敦煌壁画叙词表的可发现、可获取、可共享与可重用性。敦煌壁画叙词表的建设,将加快敦煌壁画信息资源组织和利用的语义化进程,

提高信息加工处理的智能化水平。

敦煌壁画叙词表是针对敦煌壁画研究领域而编制的具有规范细分语义关系、权威多源注释内容的一套受控词表。叙词表全面涵盖了壁画保护与修复、考古、图像志、人文地理等壁画相关研究视角。敦煌壁画叙词表作为重要的语义工具,其应用将促进敦煌壁画细粒度描述、多源异构数据关联、领域知识图谱构建等数字人文应用的开发。敦煌壁画叙词表的编制是敦煌学研究基础设施建设中的重要组成部分,亦对我国文化遗产领域智慧数据资源建设具有重要示范意义。同时,高质量领域叙词表的构建也面临着众多挑战,比如,如何精准提取领域主题词及主题词间丰富的语义关系,叙词表分类体

① <http://en.lodlive.it>

系的科学性和适用性等问题,有待进一步的深入研究与评估。未来将丰富敦煌壁画叙词表的词汇,优化并评估叙词表质量;同时,深入开展敦煌壁画叙词表的应用实践。

参考文献

- [1] Currás E. Ontologies, taxonomies and thesauri in systems science and systematics[M]. Chandos Publishing, 2010: 131-133.
- [2] 吴雯娜, 鲍秀林. 国家叙词库的体系结构与数据模型[J]. 中国图书馆学报, 2016, 42(2): 81-96. (Wu Wenna, Bao Xiulin. The architecture and data model of national thesauri warehouse[J]. Journal of Library Science in China, 2016, 42(2): 81-96.)
- [3] Clarke D, Stella G. The information retrieval thesaurus[J]. Knowledge Organization, 2019, 46(6): 439-459.
- [4] Département du Systèmed' Information de l'Architecture et du Patrimoine. Thesaurus for the designation of architecture and structural works [EB/OL]. [2019-06-08]. <http://data.culture.fr/thesaurus/page/ark:/6717/T96>.
- [5] Council of Europe. Thesaurus [EB/OL]. [2019-06-08]. <https://www.coe.int/en/web/herein-system/thesaurus>.
- [6] FISH. FISH vocabularies [EB/OL]. [2019-06-08]. <http://www.heritage-standards.org.uk/fish-vocabularies/>.
- [7] Victoria and Albert Museum. Chinese Iconography Thesaurus (CIT) [EB/OL]. [2019-06-08]. <https://www.vam.ac.uk/research/projects/chinese-iconography-thesaurus-cit>.
- [8] 吴雯娜. 我国叙词表的编制历史与发展模式[J]. 情报理论与实践, 2018, 41(6): 39-44. (Wu Wenna. Construction history and development modes of Chinese thesauri[J]. Information Studies: Theory & Application, 2018, 41(6): 39-44.)
- [9] 常春. 面向叙词表构建的知识组织生态系统研究[J]. 图书情报工作, 2016, 60(15): 101-107. (Chang Chun. Research on knowledge organization ecosystem for the construction of thesauri[J]. Library and Information Service, 2016, 60(15): 101-107.)
- [10] 张昱, 于薇. 档案领域词表自动化辅助构建及知识组织应用探析[J]. 数字图书馆论坛, 2018(6): 67-72. (Zhang Yu, Yu Wei. The construction and knowledge organization application of archives thesaurus[J]. Digital Library Forum, 2018(6): 67-72.)
- [11] Mohsen G, Al-Ayyoub M, Hmeidi I, et al. On the automatic construction of an Arabic thesaurus[C]//2018 9th International Conference on Information and Communication Systems (ICICS). IEEE, 2018: 243-247.
- [12] 雷晓, 常春, 刘伟. 面向叙词表更新的新术语分布特征研究[J]. 图书情报工作, 2019, 63(20): 121-128. (Lei Xiao, Chang Chun, Liu Wei. Research on the distribution characteristics of new terminology for the update of the thesaurus[J]. Library and Information Service, 2019, 63(20): 121-128.)
- [13] 曾文. 网络化数字化时代主题词表自动构建技术的探索与实践[J]. 国家图书馆学刊, 2012(4): 78-82. (Zeng Wen. The exploration and practice of thesaurus automatic construction technology under network and digital age[J]. Journal of the National Library of China, 2012(4): 78-82.)
- [14] Yang L, Wu Y. Creating a taxonomy of earthquake disaster response and recovery for online earthquake information

- management[J]. Knowledge Organization, 2019, 46(2): 77-89.
- [15] 司莉, 柴源, 周李梅, 等. 国外网络叙词表的现状调查及发展趋势[J]. 图书馆杂志, 2011, 30(7): 22-26. (Si Li, Chai Yuan, Zhou Limei, et al. The status and trends of online thesauri abroad[J]. Library Journal, 2011, 30(7): 22-26.)
- [16] 赵捷, 曾建勋, 吴雯娜. 网络环境下叙词表协同编制系统的构建[J]. 图书情报工作, 2011, 55(22): 6-10. (Zhao Jie, Zeng Jianxun, Wu Wenna. Thesaurus collaborative construction system in network environment[J]. Library and Information Service, 2011, 55(22): 6-10.)
- [17] 宋培彦, 陈白雪, 王星. 语义网环境下叙词表构建方法研究[J]. 情报科学, 2018, 36(2): 14-17, 42. (Song Peiyan, Chen Baixue, Wang Xing. Study of the thesaurus construction method in semantic web environment[J]. Information Science, 2018, 36(2): 14-17, 42.)
- [18] Martínez-González M M, Alvite-Díez M L. Thesauri and semantic web: discussion of the evolution of thesauri toward their integration with the semantic web[J]. IEEE Access, 2019, 7: 153151-153170.
- [19] Panzer M, Zeng M L. Modeling classification systems in SKOS: some challenges and best-practice recommendations[C]//Proceedings of the 2009 International Conference on Dublin Core & Metadata Applications, 2009.
- [20] 范伟, 邹庆. 《中国分类主题词表》的术语网络服务探索——以主题词规范数据为例[J]. 图书情报工作, 2012, 56(14): 40-46. (Fan Wei, Zou Qing. An exploratory study on terminology web service of Chinese Classified Thesaurus: taking subject authority data as an example[J]. Library and Information Service, 2012, 56(14): 40-46.)
- [21] 曾建勋, 常春, 吴雯娜, 等. 网络环境下新型《汉语主题词表》的构建[J]. 中国图书馆学报, 2011, 37(4): 43-49. (Zeng Jianxun, Chang Chun, Wu Wenna, et al. The construction of Chinese Thesaurus in web environment[J]. Journal of Library Science in China, 2011, 37(4): 43-49.)
- [22] 欧石燕. 中文叙词表的语义化[J]. 图书情报工作, 2015, 59(16): 110-118. (Ou Shiyan, Semantic representation of Chinese thesauri[J]. Library and Information Service, 2015, 59(16): 110-118.)
- [23] 曾新红. 中文知识组织系统形式化语义描述标准体系研究(一)——扩展 SKOS 实现传统受控词表全描述[J]. 中国图书馆学报, 2012, 38(3): 57-68. (Zeng Xinhong. Formal semantic description standard family for Chinese KOS (I): realizing full description for Chinese controlled vocabularies by expanding SKOS[J]. Journal of Library Science in China, 2012, 38(3): 57-68.)
- [24] 鲜国建, 赵瑞雪, 寇远涛, 等. 农业科学叙词表关联数据构建研究与实践[J]. 现代图书情报技术, 2013(11): 8-14. (Xian Guojian, Zhao Ruixue, Kou Yuntao, et al. Study and practice on converting and publishing Chinese agricultural thesaurus as linked open data[J]. New Technology of Library and Information Service, 2013(11): 8-14.)
- [25] 石泽顺, 肖明. 基于网络叙词表的图情学科 SKOS 构建与可视化研究[J]. 情报学报, 2018(3): 274-284. (Shi Zeshun, Xiao Ming. Study of the construction and visualization of SKOS in the library and information science field based on online thesaurus[J]. Journal of the China Society for Scientific and Technical Information, 2018(3): 274-284.)
- [26] Zeng M L, Mayr P. Knowledge Organization Systems (KOS) in the semantic web: a multi-dimensional review[J]. International Journal on Digital Libraries, 2019, 20(3): 209-230.
- [27] Zeng M L, Clunis J. FAIR + FIT: guiding principles and functional metrics for Linked Open Data (LOD) KOS products[J]. Journal of Data and Information Science, 2020, 5(1): 93-118.

- [28] 马费成,姜愿,赵一鸣. 服务视角下的知识组织系统研究新进展[J]. 情报杂志,2015(7):152,165-172. (Ma Feicheng, Jiang Yuan, Zhao Yiming. Research progress on knowledge organization system: a perspective of services[J]. Journal of Intelligence, 2015(7):152,165-172.)
- [29] 陶俊. 词表语义组织研究的演进(1998—2018)[J]. 图书情报工作,2018,62(21):140-148. (Tao Jun. A survey for research on semantic organization of vocabularies (1998-2018)[J]. Library and Information Service, 2018,62(21):140-148.)
- [30] 贾君枝. 面向数据网络的信息组织演变发展[J]. 中国图书馆学报,2019,45(5):51-60. (Jia Junzhi. Development of information organization for the data web[J]. Journal of Library Science in China, 2019,45(5):51-60.)
- [31] Dadzie A S, Pietriga E. Visualisation of linked data-reprise[J]. Semantic Web, 2017,8(1):1-21.
- [32] 韩燕,何琳,彭爱东. 近10年我国传统知识组织系统关联化研究述评[J]. 图书情报工作,2017(24):135-143. (Han Yan, He Lin, Peng Aidong. Review on the research of open link of knowledge organization systems in China in recent ten years[J]. Library and Information Service, 2017(24):135-143.)
- [33] Wang X, Cheng H, Li H, et al. Chinese Dunhuang Mural vocabulary construction based on human-machine cooperation[C]//Proceedings of the International Conference of Digital Humanities. Utrecht, Netherlands, 2019.
- [34] Wang X, Song N, Zhang L, et al. Understanding subjects contained in Dunhuang Mural images for deep semantic annotation[J]. Journal of Documentation, 2018,74(2):333-353.
- [35] 张俊娥. 艺术和建筑词表 AAT 的中文文化研究[J]. 数字图书馆论坛,2015(7):36-43. (Zhang June. Analysis of applying AAT to digital museum of China [J]. Digital Library Forum, 2015(7):36-43.)
- [36] The Getty Research Institute. About the AAT [EB/OL]. [2019-10-12]. <https://www.getty.edu/research/tools/vocabularies/aat/about.html>.
- [37] 白海燕,姜波. 基于开源软件构建数字图书馆的知识组织体系[J]. 现代图书情报技术,2009(4):7-13. (Bai Haiyan, Jiang Bo. To build knowledge organization system of digital library based on open source software [J]. New Technology of Library and Information Service, 2009(4):7-13.)
- [38] W3C. Best practice recipes for publishing RDF vocabularies [EB/OL]. [2019-10-12]. <https://www.w3.org/TR/swbp-vocab-pub>.
- [39] Mourmoutsev D, Pavlov D, Emelyanov Y, et al. The simple ,web-based tool for visualization and sharing of semantic data and ontologies[C]//International Semantic Web Conference, 2015.
- [40] 陈涛,刘炜,朱庆华. 中文百科概念术语服务平台 SinoPedia 的构建研究[J]. 中国图书馆学报,2018,44(4):4-18. (Chen Tao, Liu Wei, Zhu Qinghua. SinoPedia: an unified Chinese terminology service platform based on linked data[J]. Journal of Library Science in China, 2018,44(4):4-18.)

王晓光 武汉大学信息管理学院教授,博士生导师,教育部人文社科重点研究基地武汉大学信息资源研究中心副主任。湖北 武汉 430072。

侯西龙 武汉大学信息资源研究中心博士后。湖北 武汉 430072。

程航航 武汉大学信息管理学院硕士研究生。湖北 武汉 430072。

夏生平 敦煌研究院副研究员。甘肃 敦煌 736200。

(收稿日期:2019-12-10;修回日期:2020-02-19)