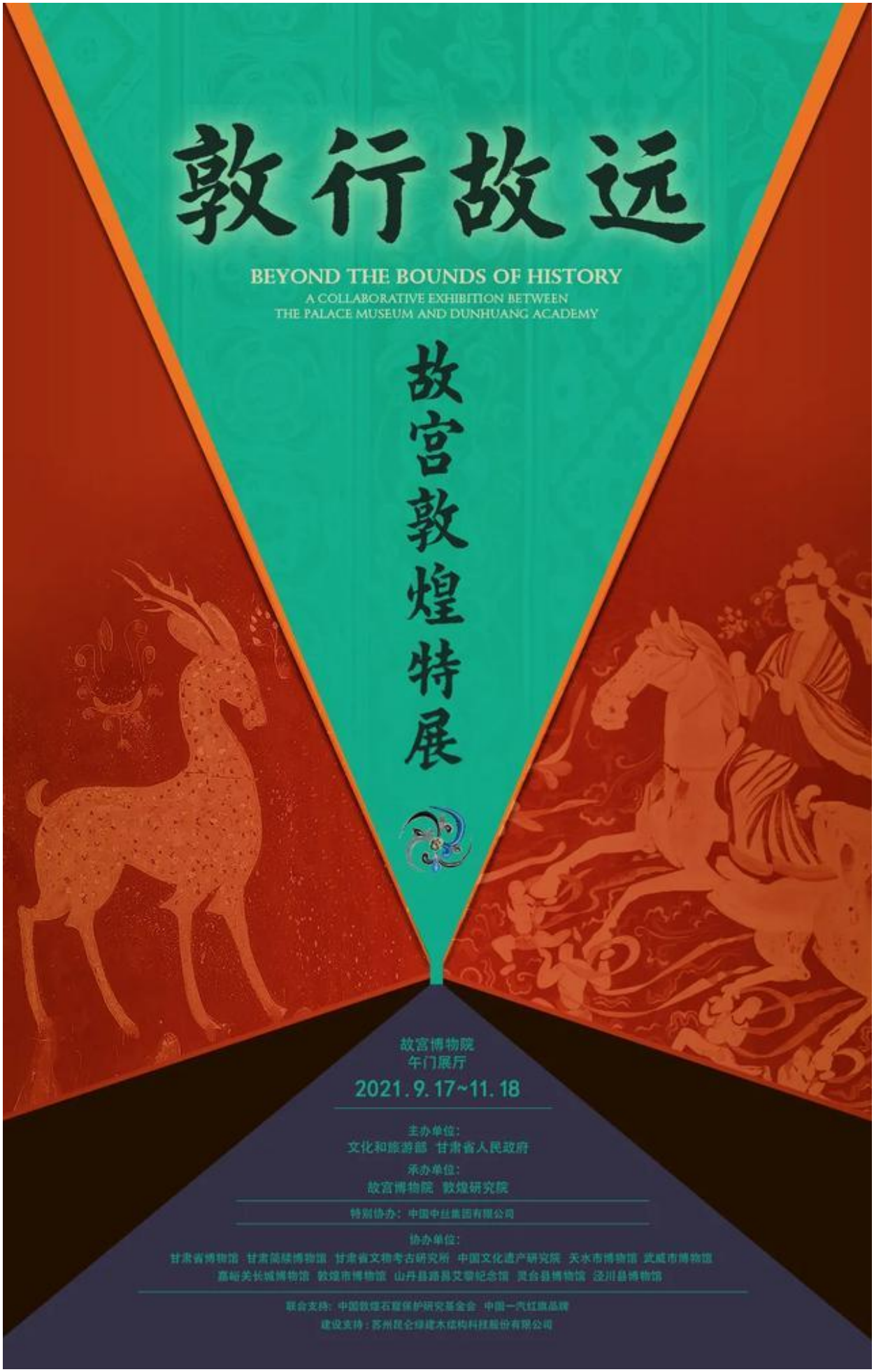
数字敦煌

敦煌壁画叙词表构建与关联数据发布

Thesaurus construction and associated data publishing of Dunhuang frescoes

开启数字敦煌之旅

近两年来，敦煌热席卷了各个领域，与敦煌文化相关的综艺影视节目、美妆产品、游戏联动等层出不穷。2021年9月，故宫开设了敦煌特展，吸引了无数对敦煌文化感兴趣的人。敦煌文化已然成为了具有代表性的中国传统文化之一。


敦行故远
BEYOND THE BOUNDS OF HISTORY
A COLLABORATIVE EXHIBITION BETWEEN
THE PALACE MUSEUM AND DUNHUANG ACADEMY
故宫敦煌特展
故宫博物院
午门展厅
2021.9.17~11.18
主办单位：
文化和旅游部 甘肃省人民政府
承办单位：
故宫博物院 敦煌研究院
特别协办：中国中丝集团有限公司
协办单位：
甘肃省博物馆 甘肃简牍博物馆 甘肃省文物考古研究所 中国文化遗产研究院 天水市博物馆 武威市博物馆
嘉峪关长城博物馆 敦煌市博物馆 山丹县焉支山纪念馆 灵台县博物馆 泾川县博物馆
联合支持：中国敦煌石窟保护研究基金会 中国一汽红旗品牌
建设支持：苏州昆仑绿建木结构科技股份有限公司

项目背景

敦煌壁画对敦煌历史、美术、乐舞、宗教等方面的研究具有重大的利用价值。然而，目前对敦煌壁画数字资源的标注与描述工作，由于缺少一套专门设计的词表，无法以一致性的控制词汇作为标准，也无法进一步对数字资源进行整合并展开语义互操作的工作，限制了敦煌壁画的研究与壁画价值的挖掘。

团队介绍

DH Center for Digital Humanities
Wuhan University

王晓光

武汉大学信息管理学院教授，博导，副院长，武汉大学大数据研究院常务副院长，武汉大学数字人文研究中心主任，国家级人才称号入选者。

侯西龙

武汉大学图书情报与档案管理博士后

项目实施过程

自顶向下的叙词表结构设计

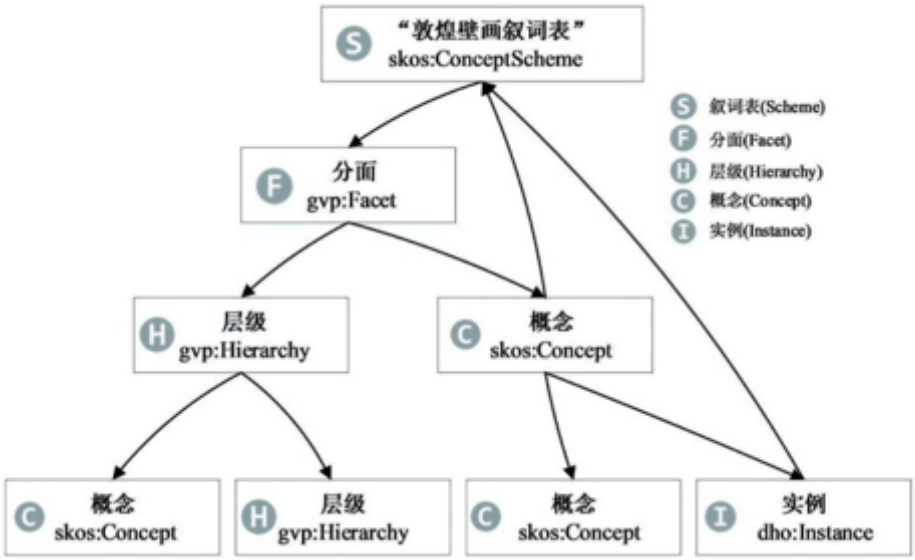


图1 敦煌壁画叙词表的层级逻辑关系

01

对敦煌壁画研究文献

进行内容分析

02

艺术与建筑叙词表

调研

03

参考叙词表

制作分层级结构

04

参考敦煌文献

设计叙词表

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,该团队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。

同时,**对国际知名的Getty“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。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。

参考AAT词表结构,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,构建多层级的分类结构;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,**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。**

自底向上的叙词表拓展优化

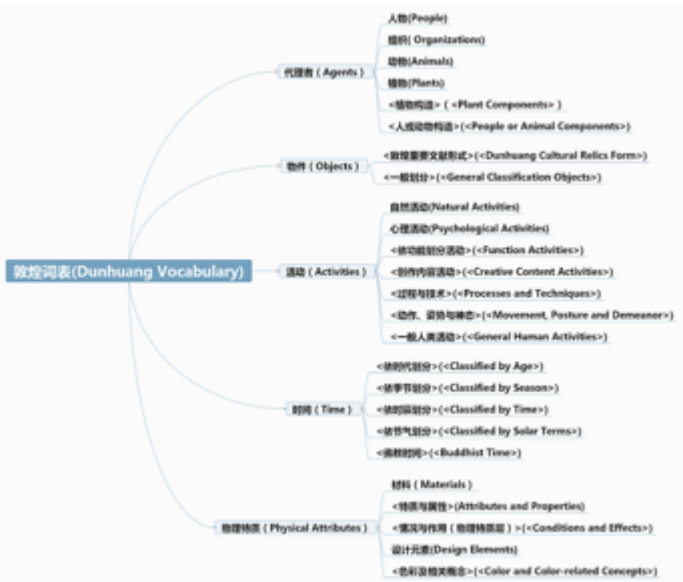


图2 敦煌壁画叙词表的顶层框架结构

01

领域概念与术语

收集

02

领域主题词
发现与归类

03

叙词表结构
调整与优化

04

叙词表
管理与维护

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。

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,该团队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。

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克夫模型,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,团队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,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。

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。通过不断地迭代,充分发挥人机协同编制的优势,实现叙词表的扩展与优化。

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,团队利用TemaTres开源词表管理系统对叙词表进行管理与维护。

项目成果

目前，敦煌壁画主题词表包括**代理者、物理特质、活动、时间、物件**五大分面，并设置25个二级类目，最深达十层；敦煌壁画主题词表共包含3896个词汇，其中与AAT（艺术与建筑叙词表）关联的主题词共430余个，数据库中三元组数量共27500余个。

项目构建了一个规范、全面的敦煌壁画领域主题词表，为敦煌壁画数字资源的深度语义标注、语义检索、知识组织、信息关联与共享等提供一套受控词表。其作为知识组织的框架体系和概念集，将提供自动标引、信息抽取、自动分类等信息加工自动化的支撑，也是智能化知识检索、知识挖掘、知识发现的基础工具。

在此基础上，在遵循W3C词表RDF（资源描述框架）发布的最佳实践，参考SKOS（简单知识组织系统 Simple Knowledge Organization System，SKOS）模型与Getty词表本体模型，**建立敦煌壁画主题词表本体模型以规范词表的语义转换**；然后，按照叙词表本体将敦煌词表进行语义转换，并对数据进行质量检验；通过SPARQL（为RDF开发的一种查询语言和数据获取协议）查询的方式与AAT进行**概念关联匹配**；最后，完成**主题词表关联数据集的存储与发布。**基于Apache Jena框架搭建敦煌壁画主题词表关联数据服务平台，提供主题词表关联数据浏览、词表可视化、主题词查询、SPARQL查询等功能以及关联开放数据服务。

成果使用

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。平台实现了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，查询结果显示符合查询条件的概念属于及其URL。

图3 关键词检索界面

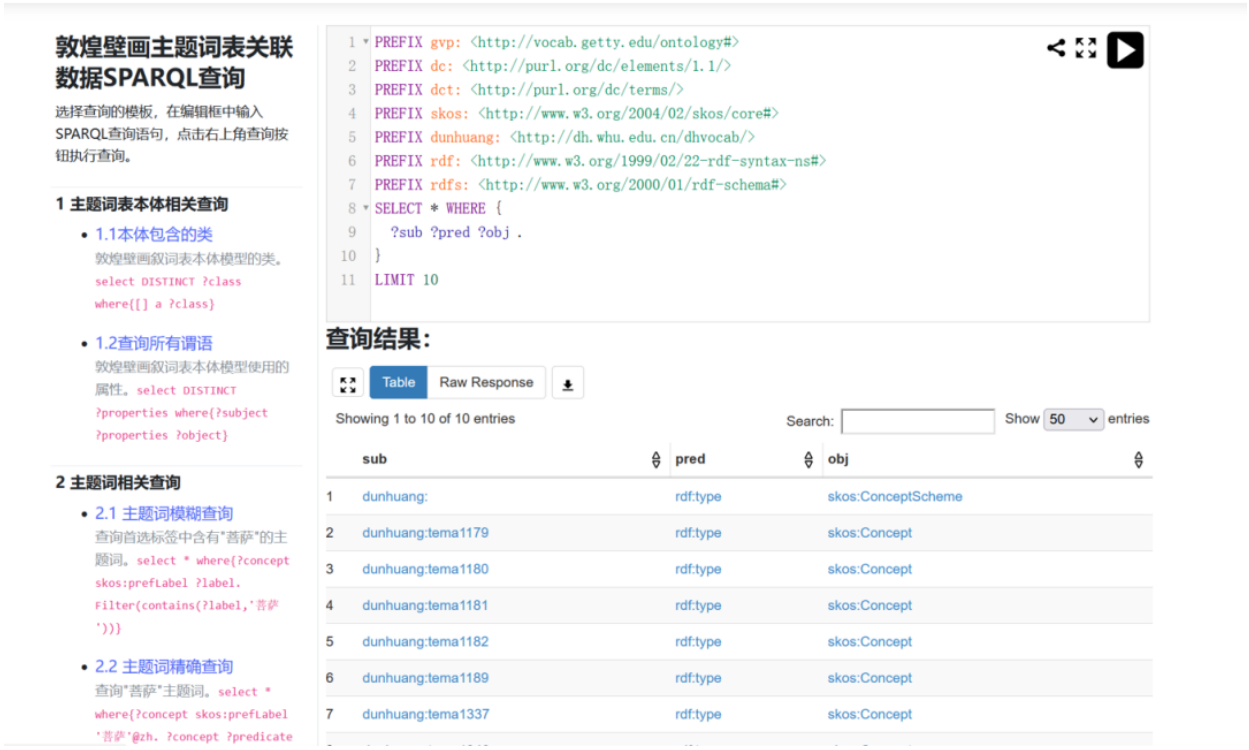
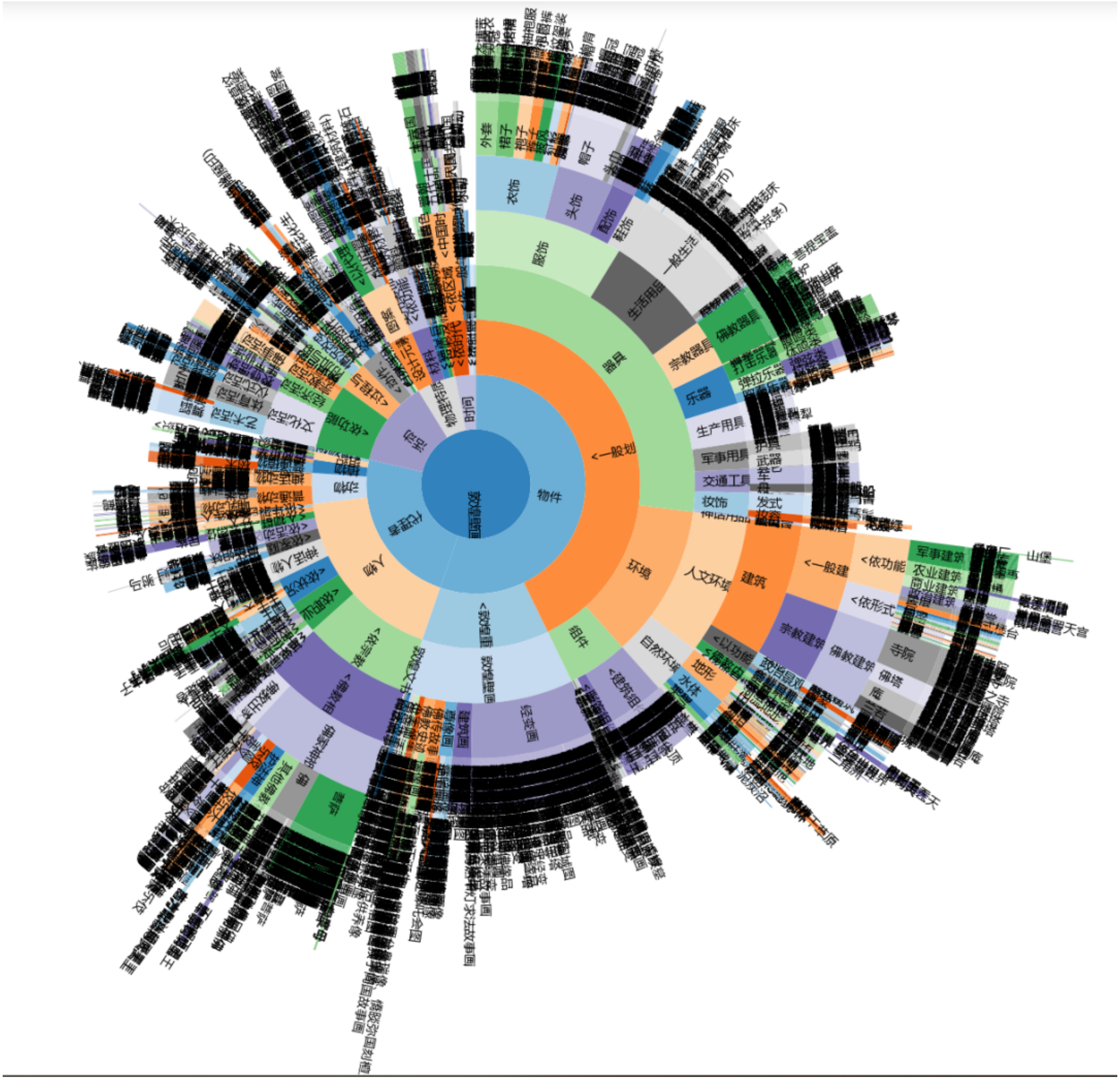


图4 SPARQL查询语句检索界面

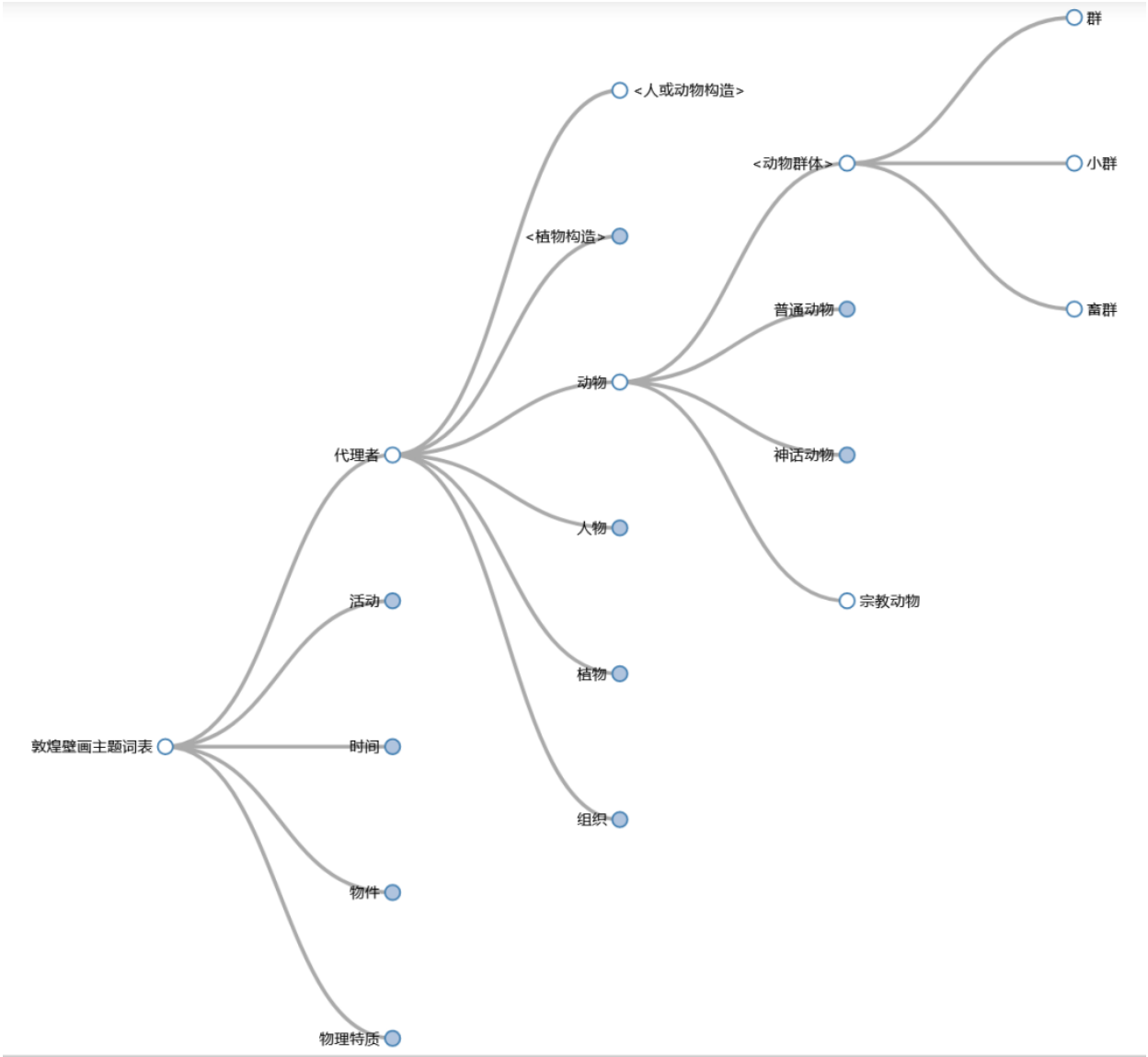
叙词表可视化

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**旭日图、圆堆图及树状图**等多种叙词表可视化。



物件
器具
服饰
衣饰
头饰
配饰
鞋饰
外套
帽子
一般生活
生活用品
宗教器具
佛教器具
乐器
生产用具
军事用具
交通工具
妆饰
发式
武器
环境
人文环境
自然环境
建筑
军事建筑
农业建筑
宗教建筑
佛教建筑
寺院
佛塔
地形
水体
组件
时间
人物
经变画
山峦


敦煌壁画主题词表
代理者
<人或动物构造>
<植物构造>
动物
人物
植物
组织
<动物群体>
群
小群
畜群
普通动物
神话动物
宗教动物
活动
时间
物件
物理特质

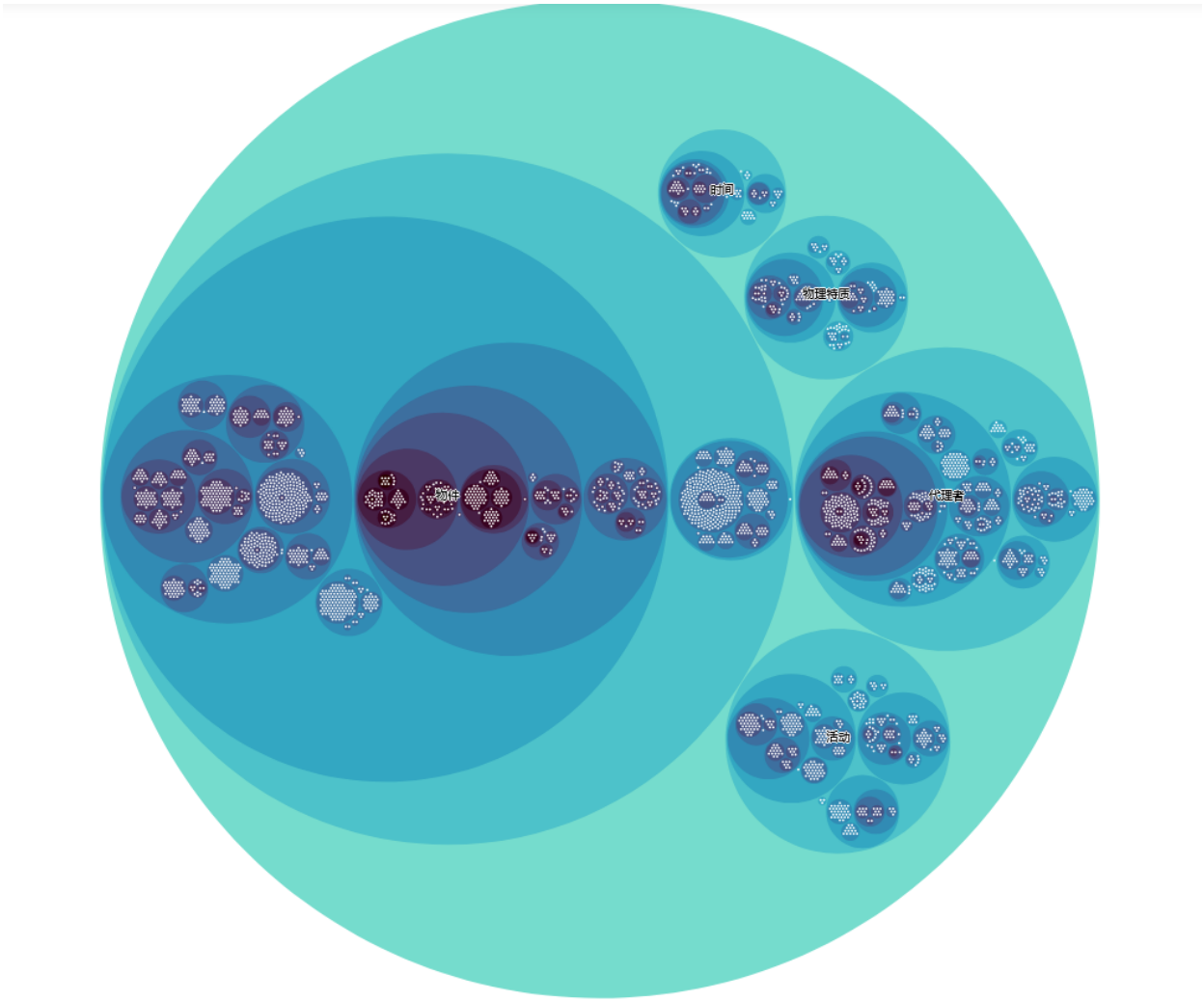


图5 旭日图、树状图和圆锥图

项目贡献

满足实际需求

过往由于缺少一套专门围绕敦煌壁画而设计的词表，导致对敦煌壁画数字资源对象描述时无法以一致性的控制词汇作为标准，也无法进一步对数字资源整合与语义互操作的工作，限制了敦煌壁画的研究与壁画价值的挖掘”的问题。而词表的出现有效解决了这些问题。

引领领域新实践

敦化壁画主题词表关联数据促进了人文学者的研究与敦煌壁画数字人文应用的开发，也为我国文化遗产领域相关受控词表与主题词表的构建提供参考借鉴。

填补学术空白

在世界范围内的文化遗产及人文艺术领域，叙词表的建设已经取得了一定的成果。我国在叙词表编写上也有着丰富的经验，但在文化遗产与人文艺术细分领域几乎为空白。敦煌壁画叙词表的诞生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播传承。

项目特色

敦煌壁画叙词表关联数据

左右滑动查看更多

人机协同机制构建

采用自上而下的词表结构设计以及自下而上的词表优化过程，利用自然语言处理与人工相结合的协同机制进行分词与校对。

敦煌壁画叙词表关联数据

左右滑动查看更多

关联数据发布

遵循关联数据的基本原则以及词表RDF发布的最佳实践方法，建立敦煌壁画主题词表概念关系模型，采用SKOS/RDF对主题词进行语义化描述，提供关联数据服务；同时与Getty AAT词表进行关联。

敦煌壁画叙词表关联数据

左右滑动查看更多

高质量规范词表

在多位敦煌学专家的协助下，迭代优化词表结构与词汇分类，词汇内容注释来源于敦煌学、佛教等多方面中英文权威词典；并与国际人文与艺术领域词表进行关联。

项目评价

1.**开创性**：对于学术领域空白的填补。

2.**实用性**：可供敦煌相关领域研究的学者使用，可获得更为深入与全面系统的资料。

3.**未来可开发**：目前敦煌壁画叙词表关联数据还是一个学术专业领域可用的系统，属于研究工具，还未开发出接近普通民众的产品。未来可以借助于此，深入挖掘敦煌壁画的内容，让其带着深厚的文化底蕴、历史文化价值，以更加生动、亲切的形式走进民众。

DUNHUANG

编辑 | 郭滢辰 唐雨菲

排版 | 郭滢辰 唐雨菲

公众号账号：rucdh2019

网址：http：//dh.ruc.edu.cn

邮箱：rucdh@ruc.edu.cn

中心简介

中国人民大学数字人文研究中心集人民大学多学科优势，秉持融合文理、协同创新之理念，开展数字人文理论研究、实践探索、人才培养和学术交流。