基于Semantic Turkey的主题词表及本体构建应用研究*

姚晓娜[1,2]　王思丽[1,2]　张旺强[1,2]
（1. 中国科学院西北生态环境资源研究院干旱区生态安全与可持续发展重点实验室，兰州 730000；
2. 甘肃省知识计算与决策智能重点实验室，兰州 730000）

摘要：主题词表及本体是语义化知识管理系统的基础数据支撑，对领域知识的语义化组织及知识图谱的构建具有重要意义。在建设公共危机案例知识集成平台的过程中，采用开源软件Semantic Turkey开发主题词表及本体构建功能，并在此基础上实现规范数据录入、词表导航、知识映射等功能，从而支持进一步的语义检索和知识推理。构建的主题词表及本体模型基于语义网标准与技术，具有良好的规范性和互操作性。开源软件Semantic Turkey提供了功能完备的应用程序编程接口，与完全自主开发相比，降低了开发成本，缩短了开发时间，为语义化知识管理系统的开发工作提供新思路和参考依据。

关键词：主题词表；本体；Semantic Turkey；SKOS；OWL
中图分类号：G250.7　DOI：10.3772/j.issn.1673-2286.2024.05.004
引文格式：姚晓娜，王思丽，张旺强. 基于Semantic Turkey的主题词表及本体构建应用研究[J]. 数字图书馆论坛，2024，20（5）：28-34.

在信息时代，知识表示、语义表达与推理等成为知识组织的重要研究方向，主题词表及本体模型构建也是语义化知识管理系统必不可少的功能。主题词表，又称叙词表，是概括某一学科或若干学科领域，并由语义相关的名词术语组成的规范化的动态词汇表[1]。主题词表可促进自动标引、实体识别等自然语言处理技术进行数据规范及约束，促进完整性、规范性数据检索的实现，是知识深度挖掘和本体模型构建的基础数据支撑[2]。本体是指共享概念体系的正式、明确的规范，用于定义知识领域内的概念和关系，并使不同系统和用户之间能够进行通信和知识共享[3]。本体用于描述和表示领域知识，具有良好的概念层次结构并支持逻辑推理，是语义表达与推理的重要工具。

公共危机案例知识集成平台是兰州大学应急管理研究中心主持建设的用于管理突发公共危机事件案例及相关知识的集成平台，该平台对碎片化的信息、情报、数据等知识进行规范化的形式描述，支持语义检索和知识推理。因此，在公共危机案例知识集成平台的建设中，不仅需要提供一种词汇控制工具，以指导标引者和用户使用一致的词进行标引和检索，还需要提供一个本体构建工具，使得用户可以对公共危机领域内的知识进行描述和表示，并在此基础上生成知识图谱，实现基于知识的检索等应用。本研究在对国内外相关开源软件调研的基础上，选择了Semantic Turkey以提供后端服务中的词表及本体服务，开发主题词表及本体构建功能，并在此基础上实现规范数据录入、词表导航、知识映射等功能。

收稿日期：2023-12-15
*本研究得到甘肃省自然科学基金项目“甘肃省医疗健康大数据资产管理模式与再利用机制研究”（编号：23JRRA581）资助。

1　相关开源软件调研

按照存储模型，主题词表及本体构建的相关开源软件可以分为三大类。

第一类采用关系数据库进行存储，需要结合具体的业务逻辑，进行相关表和字段的设计，相当于把主题词表的层级结构和本体的图形结构“扁平化”到关系数据表的二维行列结构中。相关的开源软件有Wikibase[4]、TemaTres[5]和iQvoc[6]。这一类软件虽然支持RDF三元组的管理，但由于采用关系数据库进行内部存储，缺乏更深层次的语义，无法支持进一步的逻辑推理，且查询效率比较低。

第二类采用属性图数据库（如Neo4j）进行存储。对主题词表而言，概念对应节点，关系对应节点之间的边；对于本体而言，类和实例对应节点，对象属性对应边，数据属性对应属性，需要增加属性用于区分类和实例。相关的开源软件有由北京大学研发的知识图谱自动化构建平台gBuilder[7]，其采用了自然语言处理、机器学习、人工智能、知识图谱以及图数据库等技术，支持基于结构化数据和非结构化数据的知识图谱构建，实现数据向知识的转化。gBuilder中的本体更像是图数据库中的模式，并不是严格意义上的本体，没有采用语义网标准和技术，不利于发布和共享，缺乏互操作性。此外，gBuilder也不支持主题词表的构建。这一类软件的优点是设置和使用简单快捷，甚至不需要提前定义数据模式，但由于缺乏标准化处理，难以与不同的数据库共享或交换数据。

第三类采用RDF图数据库进行存储，存储模型为RDF三元组，支持SPARQL查询，最适合存储主题词表及本体。相关的开源软件有Protégé[8]、WebProtégé[9]、Semantic Mediawiki[10]、SKOS Editor[11]和Vocbench[12]。RDF图数据库的缺点是在大型RDF图中搜索效率较低，这是深度搜索的复杂性导致的。

本研究从词表管理、本体管理、互操作性、团队协作以及应用程序编程接口（Application Programming Interface，API）等方面对相关开源软件进行对比分析（见表1），从中选择功能较为完备、满足研究应用需求的软件。

表1　相关开源软件比较分析

软件名称	词表管理	本体管理	互操作性	团队协作	API
Wikibase	部分支持	支持	较好	支持	无
TemaTres	支持	不支持	较好	支持	有
iQvoc	支持	不支持	较好	支持	无
gBuilder	不支持	支持	较差	不支持	无
Protégé	Protégé 4支持，最新版不支持	支持	较好	不支持	无
WebProtégé	不支持	支持	较好	支持	无
Semantic MediaWiki	支持	部分支持	较好	支持	无
SKOS Editor	支持	不支持	较好	支持	无
Vocbench	支持	支持	较好	支持	有

通过对比分析可以看到，同时完全支持主题词表和本体管理的开源软件并不多，目前只有Vocbench能够满足研究的应用需求。Vocbench是由意大利罗马第二大学ART研究小组开发的一个Web端应用程序，用于管理OWL本体、SKOS词表以及本体词典等RDF数据集，支持多语言和协作开发，最早被联合国粮农组织的农业信息管理标准小组用于对农业组织多语种叙词表Agrovoc的语义化转换。Vocbench在系统架构上是前后端分离的，前端采用Angular框架实现，后端采用开源平台Semantic Turkey。Semantic Turkey是一个用于知识管理和获取的开源平台，基于Java语言开发，采用了OSGI模块化架构，具有良好的可扩展性，提供了一系列API，用户可以基于API开发专用的工具或系统，定制特有的用户界面[13]。所提出的公共危机案例知识集成平台也是前后端分离的系统架构，前端采用Vue框架实现界面开发，需要在已有前端的基础上开发主题词表及本体构建功能，因此也选择将Semantic Turkey作为后端服务的开发方式。

2　主题词表及本体构建工具的系统架构

主题词表及本体构建工具是公共危机案例知识集

成平台的子系统，系统架构如图1所示，采用典型的三层架构，包括表示层、服务层和数据层。

数据层用于访问三元组数据，由一系列访问RDF数据的API构成，通过SPARQL语言进行检索。支持多种三元组存储方式，包括内存、本地文件系统以及外部RDF数据库等，并可以根据采用的技术进行微调，如进行逻辑推理时，将数据存放在内存中，并实时同步到大容量数据库中。

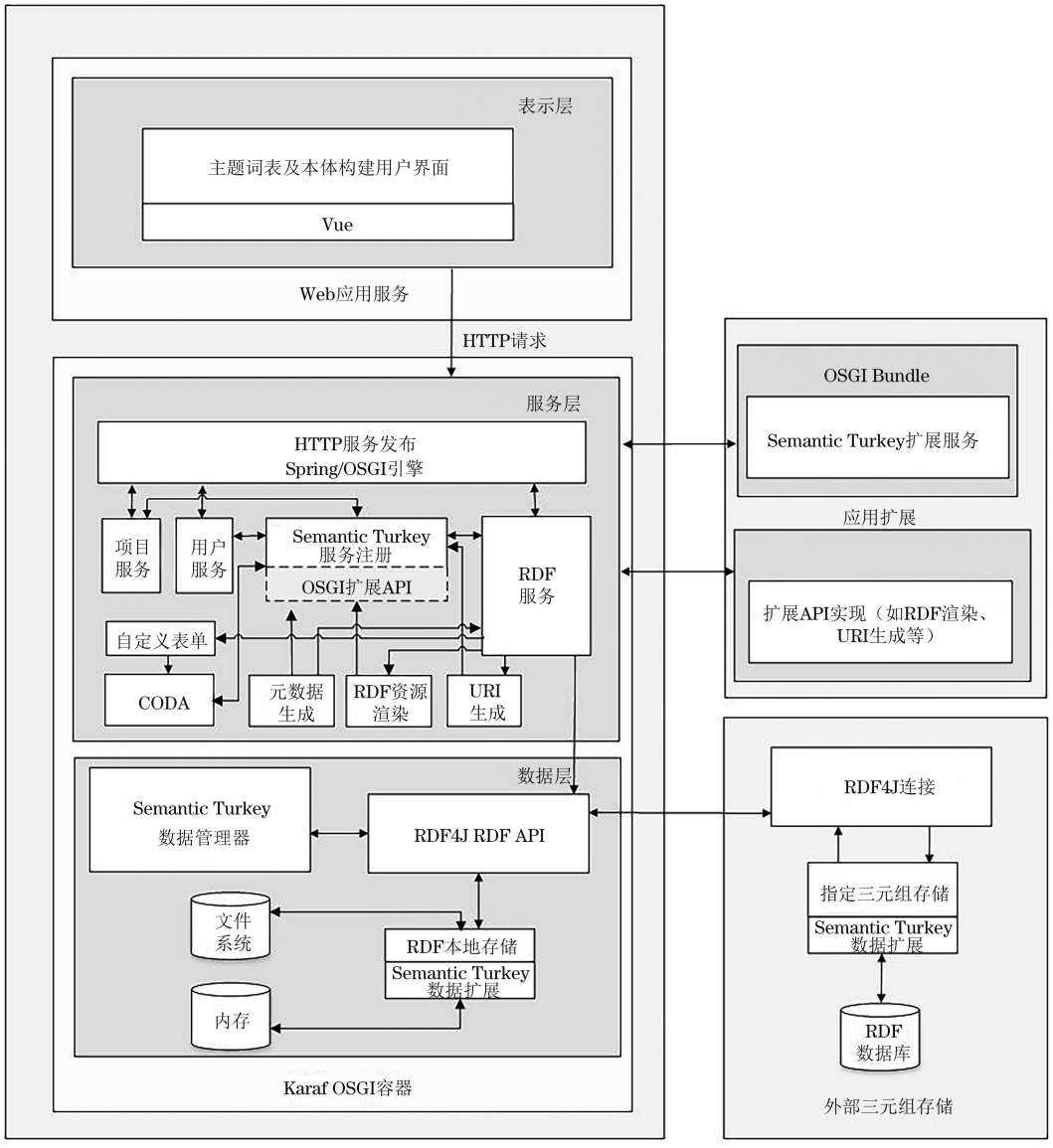

图1　主题词表及本体建构子系统的系统架构

服务层采用OSGI动态模块化框架，所有服务模块通过OSGI扩展API统一注册，并发布为HTTP服务。主要的服务包括项目服务、用户服务和RDF服务，Semantic Turkey的主题词表及本体构建都基于项目和用户进行，其提供一系列RDF服务，如元数据生成、RDF资源渲染及URI生成等核心功能，开发人员可以基于OSGI扩展API进行个性化定制。

表示层即Web应用服务，主要功能是页面展示及用户交互。本研究的开发工作集中在表示层，Semantic Turkey是公共危机案例知识集成平台的后端，集成平台的Web应用服务通过HTTP请求调用Semantic Turkey的API，前端采用Vue技术进行主题词表及本体构建用户界面的开发。

3　主题词表及本体构建工具的开发

Semantic Turkey提供了功能完备的API，基本可满

足本研究的应用需求，因此主要通过调用已有API实现所需功能，并进行前端用户界面的开发。

3.1　主题词表管理功能

在主题词表管理方面，大部分工具以SKOS[14]为标准进行词表管理。SKOS是万维网联盟（World Wide Web Consortium，W3C）在2005年制定的规范标准，以RDF为基础，为知识组织体系（包括叙词表、分类法、主题词表、术语表等）提供了一套简单、灵活、可扩展且机器可理解的描述和转化机制，目的是实现资源的共享和重用[15]。基于SKOS开发主题词表管理功能，包括主题词表、概念体系、主题词、上位词、同义词以及相关词的创建与删除。

（1）主题词表创建与删除。一个主题词表对应Semantic Turkey中的一个SKOS项目，因此创建主题词表时需要调用Semantic Turkey API的createProject命令，指定项目模型为SKOS，词典模型为SKOS。删除主题词表时需要调用deleteProject命令。

（2）概念体系创建与删除。在SKOS模型中，主题词表对应为概念体系。SKOS概念体系可以看作是一个或多个SKOS概念的集合，对应一个独立的叙词表、分类法、主题词表等。一个SKOS项目包含一个或多个概念体系，用skos:ConceptScheme及其相应的属性描述，因此在创建SKOS项目之后还需要创建一个概念体系。创建概念体系的API命令为createConceptScheme，创建完成后还需要调用命令activeSchemes，激活该概念体系，才能进行概念（即主题词）的创建。删除概念体系时需要调用deleteConceptScheme命令。

（3）主题词的创建与删除。在SKOS模型中，主题词被视为SKOS概念，描述为实例skos:Concept。SKOS将概念的不同语言形式视为语言标签，采用skos:prefLabel描述概念的首选标签，不同的语言形式可以有不同的首选标签，但不同语言形式的首选标签最多只能有一个。对于主题词表而言，概念的首选标签就是主题词本身。创建主题词的命令为createConcept，如果创建的主题词不是顶层概念，需要指定上级概念（即上位词）的URI，参数名为broaderConcept。删除主题词的命令为deleteConcept。

（4）上位词的创建与删除。SKOS使用skos:broader表示两个SKOS概念之间的上位关系，对应地，使用skos:narrower表示两个概念之间的下位关系。Semantic Turkey并不支持直接在两个概念中创建上下位关系，而是将上下位关系隐含在概念的创建过程中。在某个主题词下创建子主题词时，创建完成后，该子主题词与上级主题词自动具有上下位关系。上下位关系可以是多对多的，一个主题词可以有多个上位的主题词，Semantic Turkey提供了添加上位词的API，命令为addBroaderConcept。同样提供了删除上位词的API，命令为removeBroaderConcept。当某个主题词具有某个上位主题词时，相应地，该主题词是其上位主题词的下位主题词，使用skos:narrower进行标记。

（5）同义词的创建与删除。在SKOS模型中，使用skos:closeMatch和skos:exactMatch标记两个相似的概念，这两个概念可以在信息检索应用程序之间交换使用[16]。Semantic Turkey提供了addValue命令，为概念添加任意属性。属性可以是数据属性，也可以是对象属性，当属性为对象属性时，即为两个概念增加映射关系。创建同义词时，需要调用addValue命令，指定属性为skos:closeMatch或skos:exactMatch。可以通过调用removeValue命令删除同义词。当作为同义词的概念并不能作为一个独立的概念存在于词表中时，可以采用skos:altLabel表示同义词。skos:altLabel在SKOS模型中用于描述概念的可选标签，与首选标签一样，可选标签可以有多种语言形式；但与首选标签不同，一个语言形式下可以有多个可选标签。创建可选标签的命令为addAltLabel，删除可选标签的命令为removeAltLabel。

（6）相关词的创建与删除。skos:related用来声明两个SKOS概念之间的相关链接，是对称属性。可以使用addValue命令添加相关词，传入的参数属性为skos:related，同样调用removeValue命令删除相关词。

3.2　本体管理功能

基于OWL开发本体管理功能，包括本体、类以及属性的创建与删除。

（1）本体的创建与删除。每一个本体对应Semantic Turkey中的一个OWL项目，因此创建本体时需要调用Semantic Turkey API的createProject命令，指定项目模型为OWL，词典模型为RDFS。删除本体时需要调用deleteProject命令。

（2）类的创建与删除。OWL模型中主要的概念是类和属性，顶层类为owl:Thing，有且仅有一个。创建类的命令为createClass。为了符合中文用户的使用习惯，还需要给类设置中文名称，设置中文名称的API与SKOS添加属性的API一样，为addValue API，此时传入的参数属性为rdfs:label。rdfs:label与skos:altLabel类似，可以有多种语言形式，且一个语言形式下可以有多个可选标签。删除类的命令为removeClass。

（3）属性的创建与删除。OWL模型的项目中默认有很多属性，如rdfs:label、rdfs:subClassOf、owl:sameAs等，可以直接使用。用户也可以创建自己的属性，创建属性的API命令为createProperty，需要指定属性的类型。在OWL模型中有5个属性类型，属性（property）、注解属性（annotationProperty）、数据属性（datatypeProperty）、对象属性（objectProperty）、本体属性（ontologyProperty）。删除属性的命令为deleteProperty，本研究目前实现了数据属性和对象属性的创建与删除功能。

4 主题词表及本体构建应用

4.1 主题词表构建及应用

基于上述实现的主题词表管理功能构建公共危机事件类型、行政区划、动物疫病分类以及传染病分类等主题词表。以公共危机事件类型主题词表为例，该词表依据国家标准《突发事件分类及编码》[17]划分为3个层级：第一级包括自然灾害、事故灾难、社会安全事件、公共卫生事件4个大类；第二级是第一级事件类的细分，包括地震灾害、火灾事故、网络安全事件等46个亚类；第三级是更加具体的事件类别，包括泥石流、洪水、台风等253个细类。构建时按照层级关系依次添加主题词，主题词表构建界面分左右两个区域，左侧为主题词的层级导航，右侧显示当前主题词的中英文描述、上位词、同义词及相关词。点击树形列表左上方的按钮，可以添加主题词或下位词。在获取树形结构的主题词表时，先获取所有顶层的主题词，再查询每个主题词的下位主题词。可使用Semantic Turkey提供的getTopConcepts API命令获取所有的顶层主题词，使用getNarrowerConcepts API命令获取某个主题词的所有下位主题词。

构建完成的主题词表主要用于公共危机案例知识集成平台的规范数据录入和前台浏览检索。如在自然灾害案例的规范数据录入界面，事件类型和事件子类字段关联了事件类型词表，用户只能在词表范围内选择事件类型及事件子类，从而实现数据的规范录入。

在前台浏览检索时，左侧导航提供了事件类型的树形主题词表，并显示了各个主题词下的案例统计数量，方便用户按照规范数据进行浏览和检索。

4.2 本体构建及应用

目前在公共危机领域已构建的本体模型并不多，且大多为面向整个公共危机领域的综合性本体模型，难以有效服务于具体场景下的知识查询和推理。在实际应用中，针对某一类具体的危机事件，需要加入危机事件的具体特征信息，如洪水灾害，需要有洪水水位信息、降雨量信息、地质条件、受灾人口密度等与该类事件密切相关的信息。因此在分析处理具体的危机事件时，需要在通用或公共本体的基础上，增加新的概念和关系，构建面向具体危机事件的本体模型。

公共危机案例知识集成平台把类的创建和属性的创建分为两个界面。类的创建界面左侧为类的层级导航，右侧显示当前类的URI、中文名称、父类及注释等。点击树形列表左上方的按钮，可以添加类或子类。

属性的创建界面左侧为属性的列表导航，右侧显示当前属性的URI、中文名称、定义域、值域及注释等。点击树形列表左上方的按钮，可以添加对象属性或数据属性。

本体的整体结构可以展示为节点关系图。Semantic Turkey提供了getGraphModel API命令以获取本体项目的节点关系图，图中的节点代表类，连线代表类之间的关系，包括父子类关系及数据属性。

参考李文娟[18]提出的公共危机事件案例表示框架构建公共危机案例基础本体。基于基础本体，选择泥石流为一类具体突发事件，进行泥石流本体模型的构建。本研究以公共危机案例知识集成平台上已有的122个泥石流灾害事件案例为语料，人工抽取相关概念及概念间的关系，并在领域专家的指导下，进行概念及关系的合并及泛化，最终构建完成的泥石流本体模型包括63个类（见表2）、21个对象属性以及2个数据属性。

在本体构建完成后，可以基于本体实现知识映射功能，将结构化的数据映射为本体中的实体及关系，从而形成知识图谱。Semantic Turkey集成了Sheet2RDF

表2　泥石流本体中的类

一级类	二级类	三级类
事件特征	事件编号	-
	事件标题	-
	事件类型	-
事件时间	开始时间	-
	结束时间	-
	持续时间	-
发生地点	国家、省、市、县、乡、村、地理位置	-
事件结果	受灾人群	-
	基础设施	建筑、交通、电力系统、通信系统、其他设施
灾害对象	-	-
灾害级别	-	-
灾害损失	人员伤亡	死亡人数、受伤人数、重伤人数、轻伤人数、失踪人数
	经济损失	-
	其他损失	-
	损失情况	-
事件成因	地震因素	-
	地形因素	地形分布、植被覆盖、岩性
	降雨	事前降雨量、事后降雨量、持续降雨量
	人为活动	-
环境因素	-	-
应急管理	应急组织	政府部门、非政府组织
	应急物资	通信设备、工程设备、防护用品、应急资金、器材工具
	应急响应	-
	应急文件	-
损坏设备	-	-
原生灾害	-	-
次生灾害	-	-

工具包，支持将关系数据库或Excel中的数据转化为RDF三元组。Sheet2RDF采用了功能强大的映射规范语言PEARL，具有半自动的知识映射功能，用户可以在界面上进行映射配置，系统自动生成映射代码，用户也可以直接修改映射代码，进行更加细致的转化映射[19]。Sheet2RDF使用UIMA框架来解析Excel表格，主要处理标题行，将各个列名与本体中的属性相对应。默认表格中的每一行对应一个实体描述，列对应描述中的三元组。每一行对应的实体为subject，列对应的属性是predicate，当前行中列的取值视为object。

公共危机案例知识集成平台的知识映射功能支持从Excel文件或关系数据库中导入数据，完成数据加载后，用户需要添加实体映射、属性映射及关系映射，完成映射设置后，提交并保存到知识图谱项目中。

映射完成的知识图谱可视化效果如图2所示。可以看到，关于“2010年7月24日河南栾川泥石流”的结构化数据被映射为知识图谱中的节点和边，从而支持进一步的语义检索或知识推理。

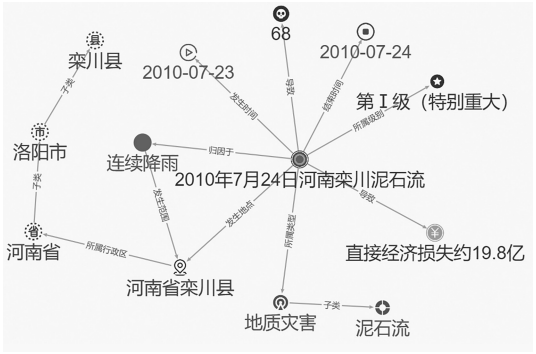

图2　知识图谱可视化效果

5　结语

主题词表可以实现对信息资源科学有序的组织与管理，为实现信息资源的有效检索、准确交换和全面共享提供技术标准。本体是领域知识共享、集成和重用的基础，本体的应用使得网络中的信息有了明确的语义描述，可以让计算机系统理解信息的含义，提高自动化和智能化处理能力。本研究基于开源软件Semantic Turkey完成了对公共危机案例知识集成平台中主题词表及本体构建功能的开发，其中主题词表的构建基于SKOS模型，本体的构建基于OWL模型，均采用W3C推荐的语义网技术标准，从而使得构建出的主题词表和本体具有良好的规范性和互操作性。在此基础上，本研究进行规范数据录入、词表导航、知识映射等功能的开发，支持进一步的语义检索或知识推理，从而实现了对公共危机事件案例及相关知识的语义化管理。开源软件Semantic Turkey提供了功能完备的API，与完全自主开发相比，基于开源软件进行二次开发可以大大缩短开发时间，降低开发成本，但同样也存在异常问题不好定位、优化改进难度较大的问题。因此，在后续的工作中，还需要进一步加深与Semantic Turkey的集成程度，利用已有的OSGI动态模块机制，尽量在不改变源码的基础上，扩展自定义的服务与模块，从而实现更多的功能应用。

参考文献

[1] 陈青云，曹建飞，陈荣祯. 从叙词表到知识图谱的构建研究与实践[J]. 农业图书情报，2019，31（1）：44-53.

[2] 词表管理系统[EB/OL]. [2023-12-13]. http://www.ninemax.com/wap/cp-cbgl.html.

[3] GRUBER T R. A translation approach to portable ontology specifications[J]. Knowledge Acquisition，1993，5（2）：199-220.

[4] Wikibase[EB/OL]. [2023-12-13]. https://wikiba.se.

[5] TemaTres[EB/OL]. [2023-12-13]. https://vocabularyserver.com/web.

[6] iQvoc[EB/OL]. [2023-12-13]. https://iqvoc.net.

[7] 北京大学知识图谱自动化构建平台gBuilder[EB/OL]. [2023-12-13]. http://www.openkg.cn/tool/gbuilder.

[8] Protégé[EB/OL]. [2023-12-13]. https://protege.stanford.edu.

[9] WebProtégé[EB/OL]. [2023-12-13]. https://webprotege.stanford.edu.

[10] Semantic MediaWiki[EB/OL]. [2023-12-13]. https://www.semantic-mediawiki.org.

[11] CONWAY M，KHOJOYAN A，FANA F，et al. Developing a web-based SKOS editor[J]. Journal of Biomedical Semantics，2016，7：5.

[12] Vocbench[EB/OL]. [2023-12-13]. https://vocbench.uniroma2.it.

[13] Semantic Turkey[EB/OL]. [2023-12-13]. https://semanticturkey.uniroma2.it.

[14] SKOS Simple Knowledge Organization System[EB/OL]. [2023-12-13]. https://www.w3.org/2004/02/skos.

[15] 鲜国建，赵瑞雪，朱亮，等. 农业科学叙词表的SKOS转化及其应用研究[J]. 现代图书情报技术，2012（10）：16-20.

[16] SANCHEZ-ALONSO S，GARCIA-BARRIOCANAL E. Making use of upper ontologies to foster interoperability between SKOS concept schemes[J]. Online Information Review，2006，30（3）：263-277.

[17] 国家质量监督检验检疫总局，中国国家标准化管理委员会. 突发事件分类与编码：GB/T 35561—2017[S]. 北京：中国标准出版社，2018.

[18] 李文娟. 基于本体的公共危机事件案例表示研究[D]. 兰州：兰州大学，2013.

[19] STELLATO A，FIORELLI M，TURBATI A，et al. VocBench 3：a collaborative semantic web editor for ontologies，thesauri and lexicons[J]. Semantic Web，2020，11（5）：855-881.

作者简介

姚晓娜，女，硕士，馆员，研究方向：知识管理系统建设，E-mail：yaoxn@llas.ac.cn。
王思丽，女，博士，馆员，研究方向：知识计算技术研发。
张旺强，男，硕士，馆员，研究方向：知识管理系统建设。

Application Research of Thesaurus and Ontology Construction Based on Semantic Turkey

YAO XiaoNa[1,2] WANG SiLi[1,2] ZHANG WangQiang[1,2]

(1. Key Laboratory of Ecological Safety and Sustainable Development in Arid Lands, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, P. R. China; 2. Key Laboratory of Knowledge Computing and Intelligent Decision, Gansu Province, Lanzhou 730000, P. R. China)

Abstract: The thesaurus and ontology are the fundamental data support for semantic knowledge management systems, which are of great significance for the semantic organization of domain knowledge and the construction of knowledge graphs. In the process of building a public crisis case knowledge integration platform, this study uses the open-source software Semantic Turkey to develop the functions of thesaurus and ontology construction and realizes functions such as standardized data input, vocabulary navigation, and knowledge mapping to support further semantic retrieval and knowledge reasoning. The thesaurus and ontology model are built based on semantic web standard and technology and have good standardization and interoperability. Semantic Turkey provides a fully functional API. Compared with fully self-developed mode, the study reduces cost and shortens time, providing new ideas and references for the development of semantic knowledge management systems.

Keywords: Thesaurus; Ontology; Semantic Turkey; SKOS; OWL

（责任编辑：王玮）