

面向简单知识组织系统的术语注册管理平台构建研究^{*}

贾君枝, 薛秋红

(山西大学经济与管理学院, 太原 030006)

摘要: SKOS为知识组织系统提供了表达词表基本结构和内容的通用模型, 为实现数据关联化奠定基础。但是词表以SKOS描述后还需要对其进行集中存储和管理, 促进词表的发现、重用、管理、标准化和互操作。目前国外术语注册管理系统的技术已经非常成熟, 而我国尚缺乏比较成熟的受控词表注册平台。文章对词表注册过程进行分析, 结合我国GB/T18391.4—2009/ISO/IEC 11179-4:2004元数据注册标准, 构建了词表注册管理平台的系统架构, 设计跨职能的词表注册业务流程并分析词表状态变化, 最后概括出术语注册管理平台所需的五个基本功能, 并对其实现可利用的关键工具进行说明。

关键字: SKOS; 术语注册; 管理平台

中图分类号: G250

DOI: 10.3772/j.issn.1673-2286.2016.2.003

1 引言

简单知识组织系统(Simple Knowledge Organization System, SKOS)的出现为知识组织系统(包括叙词表、分类表、标题表、术语表、大众分类表和其他类似的受控词表等)提供了表达词表的基本结构和内容的通用模型^[1]。其将词表的概念、概念之间的关系和知识结构进行语义化描述, 采用机器可理解的方式表达出来, 以实现数据关联化, 形成庞大的关联数据网络, 扩展词表在其他领域的可用性, 提高数据的重用性和互操作性。目前使用SKOS模型的数据集已经超过39种^[2], 如农业叙词表、杜威十进制分类法、美国国会图书馆标题表等多种受控词表。但是, 词表以SKOS编码方式描述后还需要对其注册, 旨在对受控词表进行集中存储和控制, 实现词表的更新管理, 促进受控词表的发现、重用、管理、标准化和互操作, 以更好地提供词表服务^[3]。目前具有代表性的术语注册的系统有联合国粮农组织FAO^[4]的VEST Registry和Voc Bench、开放元数据注册(Open Metadata Registry, OMR)^[5]等。国内学者欧石燕对国外主流的词表注册系统展开

详细讨论后认为受控词表注册系统中功能较为强大的是OMR^[6], 不仅提供词表的注册、上传、编辑、维护, 还支持词表校验、格式转换的功能, 在OMR上注册的受控词表和数据集已经有385个, 注册组织和个人达到156个^[7]。目前国外术语注册系统的技术已经非常成熟, 而我国尚缺乏比较成熟的受控词表注册平台。

基于以上分析, 本文将在对国外OMR系统注册过程以及系统提供的功能详细分析的基础上, 结合我国GB/T 18391.4—2009/ISO/IEC 11179-4:2004^[8]标准, 构建词表注册平台的系统架构, 设计多职能的词表注册业务流程图, 分析词表记录项的状态变化, 最后对注册平台构建的相关功能和关键技术进行研究, 以期推动我国词表注册平台的开发进程。

2 词表内容注册流程分析

词表注册不仅有术语注册, 还有术语服务。术语服务是在术语注册的基础上提供的各类受控词表的成员术语、概念和关系的Web服务, 两者相辅相成, 术语注册是术语服务的前提和保障。

^{*} 本研究得到国家社会科学基金重点项目“基于关联数据的中文名称规范档语义描述及数据聚合研究”(编号: 15ATQ004)资助。

现将词表注册流程分为六步, 先进行词表的机构注册, 再对词表的元数据及其内容进行注册, 如图1所示。

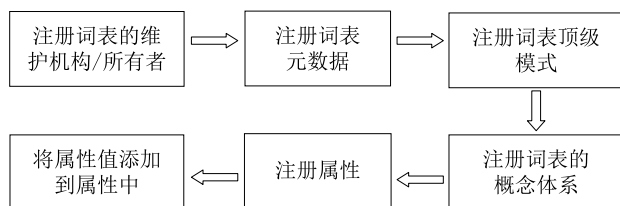


图1 词表内容注册流程

(1) 注册词表的维护机构/所有者

需要填写词表所属机构的基本信息, 比如机构名、国际代码分配符、组织标识符、文档语言、国家、通信地址、电话、E-mail、网址等机构属性与联系方式。

(2) 注册词表元数据

词表元数据即词表的基本信息, 包括词表所属机构、词表名称、词表URL、词表说明、词表开发团体、词表状态、基础定义域、令牌。其中词表URL即词表的地址, 该URL可以将整个词表加载进来; 词汇表的基础定义域为词汇表及其术语的脚本URIs; 令牌即词表的一个独特的表示, 可以用作RDF容器类; 如某词表的URL: <http://cct.nlc.gov.cn/Classification>, 基础定义域为<http://cct.nlc.gov.cn>, 令牌为Classification。

(3) 注册词表的顶级模式

记录词表整体的通用属性, 以便于对词表进行参考引用。包含以下信息: 词表的标签, 作为词表的全名, 将会应用在词表列表中; 命名空间的名称, 用来区分具体的模式; 命名空间URI, 等同于词表的基础定义域+令牌; 说明文档URL, 如果该模式有附注的说明文档, 可以通过此URL添加进来; 说明文档注释, 指对说明文档的解释; 语言, 标签名称、命名空间等使用的语言。

(4) 注册词表的概念体系

SKOS表示的受控词表中, 定义了概念体系、概念集合、顶层概念及概念, 分别用skos:ConceptScheme/skos:inScheme、skos:Collection (人名表、地名表)、skos:hasTopConcept、skos:Concept表示。此步骤需要明确所使用的SKOS概念词汇, 并添加所选词表的概念体系, 需确定概念体系对应的URI, 该概念是否是顶级概念或者在哪个顶级概念之下、概念术语的状态 (未完成/发布等)、发布的语言等。

(5) 注册属性

受控词表中, 词间的语义关系表示概念间的属

性关系, 通常采用SKOS词汇skos:prefLabel/skos:altLabel、skos:broader/skos:narrower、skos:related表示等同、等级及相关关系。另外概念的注释、标记符号分别用skos:definition、skos:example、skos:historyNote、skos:notation表示。需要将SKOS词汇注册到平台的属性中, 提交内容包含属性的标签 (人类可读的属性名称)、SKOS词汇 (计算机可读的SKOS元素)、URI、与此属性相关描述。

(6) 将属性值添加到属性中

选择概念对象, 为此概念添加已经注册好的SKOS属性, 将具体的属性值赋予属性中。填写的内容有属性类别、属性值、语言、状态。

3 术语注册平台系统架构设计

考虑到SOA架构的简单、精确定义的接口、粗粒度性、松耦合性、位置透明性、协议无关性等优势, 术语注册平台采用SOA架构, 共分为四层, 如图2所示。第一层是数据存储层, 包括词表元数据、词表内容数据、用户权限数据、用户评论等; 第二层是组件层, 因基于SOA架构的松耦合性, 基于SKOS/RDF数据的接口, 对该格式的数据进行转换从而存储到关系型数据库中, D2R/Drupal可以实现将关系型数据库转换为RDF数据^[9]; 第三层为服务层, 本层构建了与词表的四大基础应用相关的服务: 内容校验、词表管理、用户管理、词表发布; 第四层是应用层, 在应用层中, 提交组织的主业务是提交词表, 并管理自己的词表, 审核专家的主业务是审核词表, 注册机构主要负责词表管理及用户管理。总体而言, 就是将SKOS描述的词表经内容校验、格式转换后存储到关系型数据库中, 并对其管理, 符合一定的标准后, 便由注册机构使用D2R或Drupal将关系型数据库发布为关联数据。以D2R为例, 主要由三部分构成: D2RQ Mapping的主要功能是定义将关系型数据转换成RDF格式的Mapping规则, D2RQ Engine 功能是使用一个可定制的D2RQ Mapping文件将关系型数据库中的数据映射成虚拟的RDF格式。该文件的作用是在访问关系型数据时将RDF数据的查询语言SPARQL转换为RDB数据的查询语言SQL (结构化查询语言), 并将SQL查询结果转换为RDF三元组或者SPARQL查询结果; D2R Server是一个HTTP Server, 它的主要功能提供对RDF数据的查询访问接口, 以供上层的RDF浏览器、SPARQL查询客户端以及传统的HTML浏览器调用。而

Drupal除了可实现关联数据的发布,在可视化方面表现突出:支持RDF数据可视化,以及与数字地图、D3.js的结合使用。

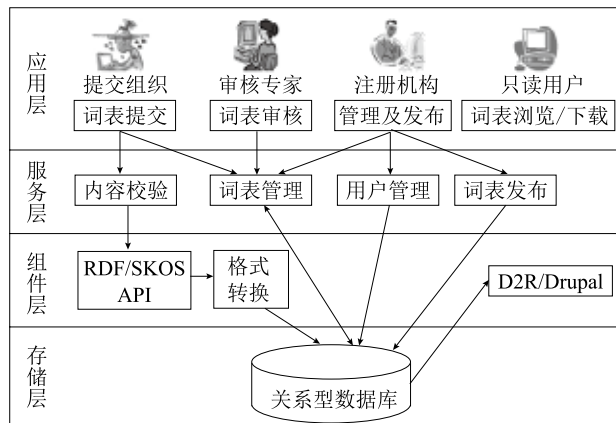


图2 注册平台系统架构设计图

4 业务流程分析

4.1 多职能业务流程

词表注册系统需要不同角色的人员协调管理,基于GB/T 18391系列标准和技术报告,将使用注册平台的

用户角色分为四种:提交组织、审核专家、注册机构以及只读用户,业务流程见图3。

(1) 提交组织,即词表的所有人。提交组织首先要注册,经过注册机构对其身份审核后,便可登录提交词表元数据,即词表的基本信息,提交完成后便生成词表元数据表和未完成词表;点击未完成词表便可进行词表内容的提交,可以是完整词表上传给系统,经系统进行词表校验后生成候选词表,也可以是单个概念的提交,也会记录到候选词表中。

(2) 审核专家,即业内相关领域的权威人士,经注册机构对其身份审核后可登录查看词表列表,点击某个词表便可以对该词表中处于候选状态的词进行审核,专家审核后词表的状态转变为合格词表、失效词表或者被替代词表。需要注意的是提交组织和审核专家登录平台后对相关词表的操作都会记录下来,包括生成时间、修改人员等,生成历史记录,便于后期对词表来源进行追踪。

(3) 注册机构,其拥有注册平台的最高权限,需要对提交组织和注册机构的用户信息进行身份核实,查看用户操作记录,对恶意用户进行剔除,编辑词表列表,删除虚假词表和恶意信息;查看词表内容,把专家审核通过后生成的词表(合格词表、失效词表或者被替代词

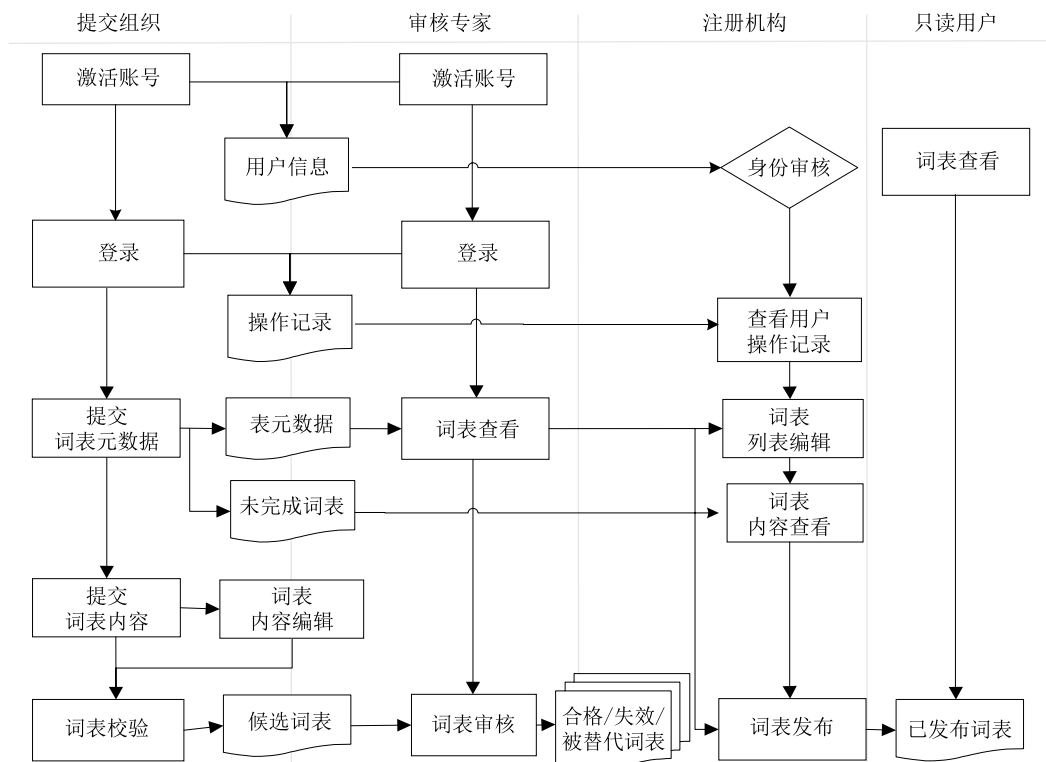


图3 术语注册业务流程图

表)连同词表元数据发布出去,形成已发布词表。

(4) 只读用户,通常可以查看已经发布的词表。

4.2 词表状态管理

从图3中可以看出系统设定了根据词表的管理级别来追踪管理项从一种状态转变为另一种状态的进程,将词表的状态分为未完成、候选、合格、失效和被替代五种。

(1) 未完成状态。提交者在正常工作的过程中,需要定义词表的概念类、属性及其属性值,以完成词表元数据及内容的编辑,需确保定义的准确有效。未完成状态到候选状态的转变由提交者来决定,当提交者认为所附加的属性足够完整时,便可将词表或者概念的状态更改为候选。

(2) 候选状态。提交者确认必选的术语属性已经完善,便进入候选状态。候选状态术语由业内相关领域的专家来评审,审查术语属性是否正确,是否符合相关标准。如果提交的元数据属性不符合上述标准,专家需要反馈给提交组织相关信息或者帮助指南。若确认术语属性符合管理项的质量要求,包括标识符的唯一性和定义的准确性,专家就将该管理项升级到合格状态。

(3) 合格状态。处于合格状态的管理项意味着主管组织已经确认元数据必选属性完整且概念属性的质量是符合要求的。处于合格状态的术语项也可能进入到被替代和失效状态。

(4) 被替代或失效状态。专家需要定期对处于合格状态的术语进行审核,判断其是否应该进入被替代或失效状态,同样处于被替代状态的术语项也可能会进入失效状态,专家做出这两个判定时需要给出简短说明。

(5) 发布状态。经专家审核通过的词表由注册机构检查其完整后连同词表元数据一起发布出去,此时词表的状态改变为已发布。

5 基本功能分析

术语注册管理平台所需的基本功能有用户管理、词表上传、词表下载、词表发布和词表维护。

(1) 用户管理功能

包括用户个人资料管理、用户权限管理、用户操作记录管理和用户意见管理。其中,个人资料管理需存储四种角色的个人资料;权限管理是指不同角色的用户权

限不同,由注册机构从安全级别的角度对资源和对象进行划分,对不同级别的资源进行访问控制,并将这种权限赋予角色中;操作记录管理是指不同角色的用户对词表进行修改时,需要对其操作进行记录,以便对词表的来源以及状态的变化情况进行追踪;意见管理是指不同角色的用户都可以发表对平台、术语改进方面的意见,参与到术语注册管理的流程中。

(2) 词表上传功能

提交组织除了提交单个术语,还可以进行整个词表的上传,默认支持SKOS/RDF序列化格式的词表文档,上传的词表最终存储在关系型数据库中,涉及SKOS数据如何存储到关系型数据库中,便于对其管理。在解析数据时,任何依据特定本体的RDF工具,都可加载SKOS本体,并可创建SKOS概念,然后在属性中填入适当的元数据。

(3) 词表下载功能

以词表列表的方式展示所有词表,提供某个词表整体下载,支持词表中单个概念下载、可视化图形下载,同时提供XML/SKOS/JSON三种格式的数据下载方式。

(4) 词表发布功能

注册机构通过使用关联数据发布工具将关系型数据库内容发布出去,允许数据消费者包括原始数据提供者,以不同的数据格式访问、聚合中央存储库的数据,广泛支持基于数据的下游服务,所有完整的数据来源对下游用户是完全透明的。

(5) 词表维护功能

词表维护包括词表状态管理和词表综合管理。其中词表状态管理是指根据词表的状态去管理词表。词表综合管理是根据词表概念的生成事件日志、跟踪错误记录以及信息反馈系统获取诊断和帮助支持,编辑已经注册的词表元数据,更新词表文档的版本,对同一词表的不同版本进行控制,扩展支持词表间的自动映射与集成。

6 结论与展望

术语注册管理可以实现对词表的集中存储和控制,促进受控词表的发现、重用、管理、标准化和互操作,以更好地提供词表服务。本文对该平台的系统架构、注册流程、状态管理以及基本功能做了具体分析,但是对词表存储与管理方面国内的技术尚不成熟,国

外有许多可利用的工具,如iQvoc用于管理词汇表的开源工具,具有SKOS导入与导出功能^[10],此外SKOS是以RDF为模型,所以还可以利用RDF-aware应用程序开发工具与库来自己构建SKOS编辑系统。如何将这些工具嵌入到术语注册管理的平台中,将是下一步研究的关键问题。

参考文献

- [1] SKOS Simple Knowledge Organization System [EB/OL]. [2016-01-01]. <http://www.w3.org/2004/02/skos/>.
- [2] SKOS/Datasets - Semantic Web Standards [EB/OL]. [2016-01-03]. <http://www.w3.org/2001/sw/wiki/SKOS/Datasets>.
- [3] 欧石燕.国外术语注册与术语服务综述[J].中国图书馆学报,2014(5):110-126.
- [4] Liang A C, Lauser B, Sini M, et al. From AGROVOC to the Agricultural Ontology Service/Concept Server: An OWL Model for Creating Ontologies in the Agricultural Domain [C]// Proceedings of the International Conference on Dublin Core and Metadata Applications, 2006.
- [5] Hillmann D, Sutton S A, Phipps J, et al. A metadata registry from vocabularies up: the NSDL registry project [C]// Proceedings of the International Conference on Dublin Core and Metadata Applications, 2006: 65-75.
- [6] 欧石燕.基于SOA架构的术语注册和服务系统设计与应用[J].中国图书馆学报,2011(5):13-25.
- [7] The Registry! [EB/OL]. [2016-01-03]. <http://metadataregistry.org/>.
- [8] GB/T 30524-2014科技平台元数据注册与管理国家标准(GB)[EB/OL]. [2016-01-06]. <http://www.csres.com/detail/240771.html>.
- [9] 唐艳春.D2R在图书馆书目数据关联服务中的应用分析[J].图书情报工作,2014(14):132-138.
- [10] 使用W3C SKOS标准提高分类管理效率[EB/OL]. [2016-01-10]. <http://www.ibm.com/developerworks/cn/xml/x-skostaxonomy/>.

作者简介

贾君枝,女,1972年生,博士,教授,研究方向:信息组织和信息检索,E-mail: junzhij@163.com。
薛秋红,女,硕士研究生。

Research on the Construction of the Terminology Registry and Management Platform Facing SKOS

JIA JunZhi, XUE QiuHong

(School of Economics and Management of Shanxi University, Taiyuan 030006, China)

Abstract: SKOS provides the knowledge organization systems a common model of the expression of the basic structure and content of the vocabularies, which lays the foundation for the realization of data association. However, after encoding vocabulary to SKOS description, it needs to be centrally stored and managed, which promotes the discovery, reuse, management, standardization and interoperability of the vocabularies. At present, the technology of foreign terminology registry and management system has been very mature, while China is still lack of relatively mature controlled vocabularies registry platform. It analyzes how vocabularies registration, combined with China GB/T 18391.4-2009/ISO/IEC 11179-4:2004 metadata registry standards. This paper constructs the system architecture of the vocabularies registry and management platform, designs the business processes of the cross-functional vocabularies registered, and analyzes the changing state of vocabularies. Finally, this paper summarizes the five basic functions required for the terminology registry and management platform and describes tools that can be used.

Keywords: SKOS; Terminology Registration; Management Platform

(收稿日期: 2016-01-18)