

《中国分类主题词表》的 SKOS 描述自动转换研究*

刘丽斌 张寿华 濮德敏 任瑞娟 米 佳

摘 要 SKOS 为知识组织系统(叙词表、分类法、主题词表、术语表等)提供了一套语义 Web 环境下简单灵活、机器可理解的描述和转换机制,使叙词表之间、叙词表与本体之间的共享成为可能。利用 Java 语言编写内容提取程序,将《中国分类主题词表》电子版中的主题词、词间语义关系、分类号、注释等提取,并分别存储在关系数据库中,然后通过 SQL 语言从数据库中提取相关字段,用 Java 语言写入 SKOS 描述的对应标签中,形成自动转换的 SKOS 描述代码。但对《中国分类主题词表》的附表部分如何在语义环境中实现 SKOS 描述及自动转换仍需进一步研究。图 2。表 2。参考文献 14。

关键词 SKOS 中国分类主题词表 自动转换

分类号 G254.24

ABSTRACT SKOS is an area of work developing specifications and standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web. Based on the programming language and database query language, a feasible scheme that automatically transforms *Classified Chinese Thesaurus* (CCT) in electronic edition into the form described in SKOS is put forward, and the specific practice is carried out. 2 figs. 2 tabs. 14 refs.

KEY WORDS SKOS. *Classified Chinese thesaurus*. Automatic transformation.

CLASS NUMBER G254.24

1 SKOS 简介

语义网资源共享的基础是领域本体的存在。传统叙词表等信息标引工具具备了领域本体的许多特点,且由领域专家多年有序组织而成,如何将叙词表成功地转换成领域本体并应用于语义网具有深远的现实意义。国内外许多学术团体相继进行了利用叙词表转换构建领域本体的尝试,其中最常用的描述语言是 RDF 和 OWL。RDF 是较为理想的资源描述语言,但它仅能表达浅层次的语义关系,无法支持更精确的语义关系描述。本体语言 OWL 在 RDF 的基础上扩展了语义描述功能,能够定义更加复杂

的概念结构,并具备了一定的知识推理能力。然而,由于过于苛求逻辑上的精确,构建基于 OWL 的领域本体是一项代价昂贵、复杂度相当高的工程,所以需要一种简单可行的描述方案来表达词表的结构和内容^[1-2]。

SKOS (Simple Knowledge Organization System) 是 W3C 语义网最佳实践及部署工作小组公布的推荐标准,目前尚处于工作草案阶段^[3]。SKOS 为知识组织系统(叙词表、分类法、主题词表、术语表等)提供了一套语义 Web 环境下简单灵活的描述和转换机制,目的是为了资源的共享和重用。SKOS 建立在 RDF 基础上,扩展了 RDF 的描述能力,提供了表达各种受控词表结构和内容的通用框架,将概念模式及语义关系

* 本文系国家自然科学基金资助项目“基于分布式数据库构建分布式 Ontology 的方案设计与实现”(项目编号:07BTQ024)的研究成果之一。

表达为机器可理解的方式^[4]。通过 SKOS 描述语言,受控词表将转换为与 RDF、OWL 兼容的概念模型,词表中的词汇对应转换为 SKOS 模型中的具体概念,并实现语义网中与其它 RDF 数据的合并与融合,从而实现真正意义上的资源共享。SKOS 语言是一种比 OWL 本体描述语言更简单,但又非常容易扩展的知识结构描述语言,通过 SKOS 知识组织方案,可以充分利用传统知识组织系统的现有成果,并实现与语义 Web 的

良好结合^[5]。

SKOS 词表是由一系列 RDF 定义的类和属性组成,这些类和属性涵盖了描述各种传统知识组织系统结构中使用到的标签。词表中的概念词条通过这些标签进行标识,从而融入到 SKOS 的概念框架中,形成机器可以理解的语义知识组织体系。SKOS 词表中的标签及其说明见表 1。

表 1 SKOS 词表的标签及其说明^[6-7]

标 签	说 明	标 签	说 明
skos:Concept	概念	skos:broader	上位关系(属)
skos:ConceptScheme	概念表	skos:broaderTransitive	上位传递关系
skos:inScheme	某个概念在某概念表中	skos:narrower	下位关系(分)
skos:hasTopConcept	叙词表有某个族首词	skos:narrowerTransitive	下位传递关系
skos:topConceptOf	概念表中的顶层概念	skos:related	相关关系(参)
skos:altLabel	非首选语词标签	skos:semanticRelation	语义关系
skos:hiddenLabel	隐藏标签	skos:Collection	概念集合(等级关系)
skos:prefLabel	首选语词标签	skos:OrderedCollection	有序的概念集合
skos:notation	分类号	skos:member	成员
skos:changeNote	变更注释	skos:memberList	成员列表
skos:definition	概念的定义	skos:broadMatch	不同词表的概念上位关系
skos:editorialNote	编者注释	skos:closeMatch	不同词表的概念相似关系
skos:example	实例	skos:exactMatch	不同词表的概念等同关系
skos:historyNote	历史注释	skos:mappingRelation	不同词表的概念映射关系
skos:note	通用注释	skos:narrowMatch	不同词表的概念下位关系
skos:scopeNote	范围注释	skos:relatedMatch	不同词表的概念相关关系

2 《中国分类主题词表》及其网络化

随着计算机技术和网络技术的普及与发展,传统分类法、主题词表等知识组织系统的网络化及其应用研究已成为创新图书馆知识组织和知识服务研究的重要课题。

《中国分类主题词表》(以下简称《中分表》)是我国目前规模最大的分类主题一体化标引工具,共收录分类法类目 52992 个、主题词 110837 条、主题词串 59738 条、入口词 35690 条,包括哲学、社会科学和自然科学、工程技术等各领域的学科和主题概念^[8]。该词表目前广泛应用于全国各类型图书馆和信息机构的文献标引工作。2005 年 9 月国家图书馆修订并研制出版

《中国分类主题词表》(第二版)视窗电子版。虽然《中分表》电子版在机助标引和互检索文献方面开发了应用接口技术,可实现单机或局域网内服务,但由于描述形式的非结构化、价格、知识产权保护、传统观念等因素,应用面仍相对狭窄,还不能完全满足网络共享、交互发展和网络服务的需求。因此,《中分表》需从语义网知识组织系统的发展特点出发,从面向单一用户到面向多用户、多语种、多元化的服务来设计确定它的数据表现形式和语言描述形式。

SKOS 是基于传统分类法、主题词表等组织、检索网络信息,并应用网络语义表示的推荐标准,为各种叙词表提供了一套语义 Web 环境下简单灵活的描述和转换机制。SKOS 非常适用于《中分表》的网络化改造。如果将《中分表》

转换成 SKOS 的描述形式,将方便其在网络上的共享和重用,并可运用到网络信息检索的各个环节和领域。

目前,LCSH(美国国会图书馆标题表)^[9]、NALAT(美国农业图书馆叙词表)、MeSH(美国医学主题词表)等词表已经启用 SKOS 表示其数据元素,开展术语网络服务^[10-11]。美国肯特州立大学图书情报学院教授曾蕾博士、中国科学院国家科学图书馆的范炜和国家图书馆的卜书庆等人也在合作研究《中分表》的 SKOS 化^[12-13]。

3 《中国分类主题词表》电子版自动转换为 SKOS 描述的构想

《中分表》涉及的概念范围广、数量多,如果人工手动用 SKOS 语言描述该词表,工作量之大可想而知,因此考虑用计算机程序实现自动转换。

《中分表》电子版的浏览表视图是由 54 个 HTML 文件组成。页面文档只明确表达了族、参、代三种关系,属、分关系隐含在词族里,用关系是代关系的反向。HTML 文档虽然不包含内容的结构信息,即文档成分及其相互关系的信息,但是 HTML 语言还描述了文档的格式。分析其页面源代码,从中找出格式标签嵌套的规律,可以通过计算机程序设计语言将主题词、词间语义关系、分类号、注释等提取出来。SKOS 是简单的结构化语言,将提取的内容与 SKOS 词表的标签一一对应,写入相应标签里,最终形成 SKOS 形式的 RDF 文档。

4 《中国分类主题词表》电子版自动转换为 SKOS 描述的具体实现方案

首先要用 Java 语言编写一段内容提取的代码(图 1),根据所处理的 HTML 文件中的各种字符与标签,判断出每一行文本的类型,即分辨出叙词、分类号、关系、相关叙词、注释等内容。考虑到《中分表》页面文档中的叙词间属、分关系不明显,需要从词族里分解,并且为了方便以后叙词表数据的提取和查询,先将从页面文档中

提取的各个字段存储到关系数据库中。关系数据库中包含三个表:第一个表存储所有族首词,第二个表存储所有的词族,第三个表存储用、代关系和相关关系的词条。词族关系表中按照上下位分为一级、二级、三级……的形式存储。最后通过 SQL 语言从数据库中提取相关字段,用 Java 语言写入 SKOS 描述的对应标签中。具体实现方案流程图见图 2。

```
.....
for(int i=0;i<str.length();i++) {
    char ch = str.charAt(i);
    if(Character.isDigit(ch)){
        d = true;
    }else if(Character.isUpperCase(ch)){
        if(i == 1 && (Ue || Le)){
            notr = true;
        }
        Ue = true;
        count++;
    }else if(Character.isLowerCase(ch)){
        .....
    }else if(ch == "|" || ch == "=" || ch == "-" ){
        count++;
    }
}
if(count == str.length()){ return 2; }
if(d && ! sp && href){ return 0; }
if(Ue && ! Le && href && ! sp){ return 0; }
if(str.getBytes().length != str.length()){
    if(Ue && notr == false) return 1;
    else return 3;
}
return 2;
.....
```

图 1 文本判别程序

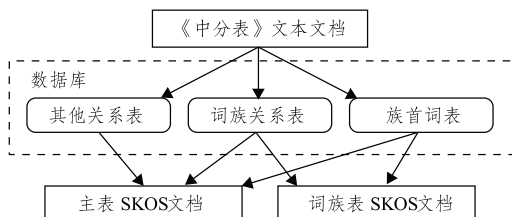


图 2 自动转换方案流程

通过对《中分表》F 类(经济)、Q 类(生物科学)、X 类(环境科学、安全科学)三个类目的词条进行自动转换试验,程序生成的描述代码完全符合 SKOS 语法要求。《中分表》主表以及词族表的节选及其相应的自动转换后的 SKOS 描述代码见表 2。

表 2 SKOS 描述代码

自然环境 Natural environments X21 D 自然地理环境 S 环境 F 沉积环境 大气环境 地貌环境 Z 环境 C 经济环境	< rdf:RDF xmlns:rdf = "http://www.w3g.org/1999/02/22 - rdf - syntax - ns#" xmlns:rdfs = "http://www.w3g.org/2000/01/rdf - schema#" xmlns:skos = "http://www.w3g.org/2004/02/skos/core#" > < skos:Concept rdf:about = "http://lib.hbu.cn/thesaurus/concepts#自然环境" > < skos:note > Natural disasters </skos:note > < skos:notation > X21 </skos:notation > < skos:prefLabel xml:lang = "zh" > 自然环境 </skos:prefLabel > < skos:altLabel xml:lang = "zh" > 自然地理环境 </skos:altLabel > < skos:broader rdf:resource = "http://lib.hbu.cn/thesaurus/concepts#环境" /> < skos:narrower rdf:resource = "http://lib.hbu.cn/thesaurus/concepts#沉积环境" /> < skos:narrower rdf:resource = "http://lib.hbu.cn/thesaurus/concepts#大气环境" /> < skos:narrower rdf:resource = "http://lib.hbu.cn/thesaurus/concepts#地貌环境" /> < skos:related rdf:resource = "http://lib.hbu.cn/thesaurus/concepts#经济环境" /> </skos:Concept > </rdf:RDF >
生物工程 • 蛋白质工程 • 发酵工程 • 基因工程 • • 植物基因工程 • • 转基因技术 • 酶工程 • 农业生物工程 • 细胞工程 • • 染色体工程	< rdf:RDF xmlns:rdf = "http://www.w3g.org/1999/02/22 - rdf - syntax - ns#" xmlns:rdfs = "http://www.w3g.org/2000/01/rdf - schema#" xmlns:skos = "http://www.w3g.org/2004/02/skos/core#" > < skos:Concept rdf:about = "http://lib.hbu.cn/thesaurus/concepts#生物工程" > < skos:narrower > < skos:Collection > < skos:member rdf:resource = "http://lib.hbu.cn/thesaurus/concepts#蛋白质工程" /> < skos:member rdf:resource = "http://lib.hbu.cn/thesaurus/concepts#发酵工程" /> < skos:member rdf:resource = "http://lib.hbu.cn/thesaurus/concepts#基因工程" /> < skos:member > < skos:Collection > < skos:member rdf:resource = "http://lib.hbu.cn/thesaurus/concepts#植物基因工程" /> < skos:member rdf:resource = "http://lib.hbu.cn/thesaurus/concepts#转基因技术" /> </skos:Collection > </skos:member > < skos:member rdf:resource = "http://lib.hbu.cn/thesaurus/concepts#酶工程" /> < skos:member rdf:resource = "http://lib.hbu.cn/thesaurus/concepts#农业生物工程" /> < skos:member rdf:resource = "http://lib.hbu.cn/thesaurus/concepts#细胞工程" /> < skos:member > < skos:Collection > < skos:member rdf:resource = "http://lib.hbu.cn/thesaurus/concepts#染色体工程" /> </skos:Collection > </skos:member > </skos:Collection > </skos:narrower > </skos:Concept > </rdf:RDF >

5 结语

传统叙词表的存在为发展语义网中的本体提供了基础。如果不借助当前受控词表的成

果,语义网的发展将会变得缓慢。SKOS 一方面可以作为受控词表在语义 Web 中的描述语言,避免了重新构建本体的繁重工作,又有效地利用了传统知识组织系统的已有成果。另一方面,SKOS 虽然不是领域知识本体,但它提供了

一种对现有受控词表进行转换的通用模型,相当于模型的中间层,也可以在此基础上建立资源与概念的关系,这比基于受控词表直接构建领域本体要简单、容易得多^[7]。

本文以程序设计语言、数据库查询语言为基础,提出了将《中国分类主题词表》电子版自动转换为 SKOS 描述形式的构想,并进行了具体的实践。利用自动转换替代人工转换,不仅提高了构建本体的效率,而且为其他受控词表的自动转换提供了参考。但是对于《中国分类主题词表》的附表部分,在语义网环境下如何实现 SKOS 描述以及自动转换尚需进一步研究。

参考文献:

- [1] 刘春艳,曹锦丹,李佳军. 语义 Web 环境下知识组织体系 SKOS 应用研究[J]. 图书情报工作, 2006,50(6):23-27.
- [2] 贾君枝. 简单知识组织系统与汉语主题词表[J]. 中国图书馆学报,2008(1):75-78,84.
- [3] Baker T, Schreiber G, Swick R. W3C semantic web deployment working group[EB/OL]. [2009-04-14]. <http://www.w3.org/2006/07/SWD/>.
- [4] Introduction to SKOS - SKOS simple knowledge organization system [EB/OL]. [2009-04-14]. <http://www.w3.org/2004/02/skos/intro>.
- [5] 王茜,陶兰,王弼佐. 语义 Web 中基于 SKOS 的知识组织模型[J]. 计算机工程与设计,2007, 28(6):1441-1443.
- [6] Miles A, Bechhofer S. SKOS simple knowledge organization system reference [EB/OL]. [2009-04-14]. <http://www.w3.org/TR/2009/CR-skos-reference-20090317/>.
- [7] 曾蕾,范炜,史国良. SKOS 词汇表[EB/OL]. [2009-04-14]. http://www.metadataetc.org/wiki/index.php5?title=Skos_zh.
- [8] 卜书庆,贺玲勇.《中国分类主题词表》电子版研制概述[J]. 国家图书馆学刊,2006(2):10-14.
- [9] LCSH. INFO [EB/OL]. [2009-04-14]. <http://lcsch.info/>.
- [10] 卜书庆. 知识组织系统的发展与《中图法》修订[J]. 图书馆论坛,2008,28(6):162-165,192.
- [11] 刘炜. 数图研究笔记:SKOS 版的 LCSH [EB/OL]. [2009-04-14]. <http://www.kevenlw.name/archives/655>.
- [12] SKOS Data Model 中文注释[EB/OL]. [2009-04-14]. <http://www.metadataetc.org/wiki/index.php5?title=SKOS>.
- [13] 曾蕾. 中分法 SKOS 化的讨论和试验[EB/OL]. [2009-04-14]. <http://cnlib20.ning.com/profiles/blogs/509559:BlogPost:17459>.
- [14] 国家图书馆《中国图书馆分类法》编辑委员会. 中国分类主题词表(第二版)电子版[CP/CD]. 北京:北京图书馆出版社,2005.

刘丽斌 河北大学管理学院硕士研究生,河北大学图书馆馆员。通讯地址:河北保定。邮编 071002。

张寿华 河北大学数学与计算机学院讲师,硕士。通讯地址同上。

濮德敏 河北大学图书馆副研究馆员,硕士。通讯地址同上。

任瑞娟 河北大学图书馆研究馆员,硕士生导师。通讯地址同上。

米佳 河北大学图书馆馆员。通讯地址同上。

(收稿日期:2009-04-15)