数字图书馆

# 基于 SKOS 的叙词表到本体的转换研究

刘春艳 陈淑萍 伍玉成

（长春理工大学图书馆 长春 130022）

【摘要】 针对从叙词表到本体的转换中存在的没有统一转换标准的问题，引入 W3C 推荐的知识组织体系 SKOS 作为叙词表转换的标准；分析知识组织体系 SKOS 的特点；实际应用 SKOS Core 词汇表完成 UKAT 叙词表节选段落到本体的转换；并应用骨架法对该本体进行评价，指出该本体的特点及局限。

【关键词】 SKOS 本体构建 叙词表 【分类号】 G254

## The Transformation from Thesaurus to Ontology Based on SKOS

**Liu Chunyan Chen Shuping Wu Yucheng**

（*Library of Changchun University of Science and Technology, Changchun 130022, China*）

【**Abstract**】 According to the problem of no unified standard existed in transformation from thesaurus to Ontology, this paper introduces the knowledge organization system SKOS recommended by W3C as a standard in the transformation, analyses the characteristics of SKOS and transforms the chosen paragraph of UKAT thesauri to Ontology using SKOS core vocabulary, then evaluates the Ontology using skeletal methodology, and points out the characteristics and limitations of the Ontology.

【**Keywords**】 SKOS Ontology construction Thesaurus

本体是一种在语义层次上描述信息的建模工具，作为系统内部成员进行交流的语义基础，用于知识表示和管理、信息系统建模、信息集成、信息检索以及 Web 上异构信息的处理、语义 Web 等。本体自提出后得到了普遍的关注，各领域都在专家参与指导下构建本领域的领域本体。在领域本体的构建过程中存在一些问题：

（1）质量难以保证[1]。本体的建设需要花费大量的人力物力来搜集领域内各种概念术语及概念之间的关系，由于人为的因素，容易出现领域概念搜集不完全，概念关系定义不准确等情况，影响本体建设的质量；

（2）概念关系描述不一致[2]。目前，对于领域本体的建设没有一个统一的知识组织体系，本体中对于概念关系的描述使用不同的语言和标签，导致本体之间语义互通性降低，也降低了本体重用性。

鉴于以上原因，学者提出了利用传统的叙词表经过改造构建领域本体的方法。国外很多学术团体相继开始了利用现有的叙词表建立本体的尝试，已有多种叙词表被转换成本体系统：如阿姆斯特丹大学艺术和建筑叙词表（AAT）转换本体的项目[3]；世界农业信息中心信息及传播处把 AGROVOC 叙词表转换为农业本体的项目，各国农业叙词表包括中国的农业叙词表（CAT）通过映射到该系统中，实现对不同语言层面的农业词表建立相互联系[4]；加州环境资源评估系统（CERES）和国家生物信息基础工程（NBII）联合开发了一套基于 RDF 格式集成的有关环境的叙词表和叙词网络工具等[5]。目前，叙词表到本体的转换还没有一个统一的标准，在转换的过程中各项目使用的描述语言以及描述广度和深度不尽相同，给本体之间语义互操作和重用造成了困难。

收稿日期：2007－03－21
收修改稿日期：2007－03－30

## 1 叙词表转换本体使用的语言

为了适应语义 Web 的发展，更好地与本体的表示相融合，国际标准化组织负责信息和文献的委员会正在修订原有的叙词结构相关标准 ISO5964，将制定一个叙词的转换方式标准。在关于叙词表向本体转换的项目中，考

虑到所转换叙词表本身的特点,学者们尝试使用了很多描述语言,总结起来有以下几种:用 XML Schema 构建叙词标记语言;用 RDF Schema 表示叙词内容和关系;目前也有人探索用 DAML + OIL、OWL 语言表示叙词关系。比较以上几种本体描述语言,从 XML Schema、RDF Schema 到 DAML + OIL 以及 W3C 推荐的本体描述语言 OWL,语言的表述能力不断提高。XML Schema 语义描述能力太弱,不适合叙词表到本体的转换;OWL 语言虽具有很强的描述能力,但是描述起来过于复杂,成本过高;介于中间的资源描述框架语言 RDF Schema 在表达能力和逻辑严格性方面最适用于从叙词表到本体的转换,具有灵活性和扩展性,易于与其他资源组织体系结合应用,从这方面来看,适合叙词表到本体的转换;然而,RDF Schema 仅仅提供初级的语义关系表达,没有统一的标签支持更具体语义关系的描述,因此需要在 RDF 基础上建立支持表达更具体的语义关系的具有灵活扩展性的统一的知识组织体系标准。

## 2 知识组织体系 SKOS

2005 年,W3C 推荐了知识组织体系(Simple Knowledge Organization System,SKOS)[6]。SKOS 是资源描述框架语言 RDF(S)的应用,丰富扩展了 RDF(S)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[7]。SKOS 可以向横向、纵深扩展,可以和其他知识组织体系结合应用。SKOS 的目标是用机器可以理解的方式提供一个强有力的框架表达知识结构。在语义 Web 领域,SKOS 最早应用在欧洲语义 Web 高级发展(SWAD-E)计划中的一个子项目“分类和叙词的管理”(Catalogue and Thesaurus Management)。现在作为与 ISO 相关标准兼容的 RDF 词表模式草案,进一步发展多语种词表和词表之间的图示和导航工具[8]。

SKOS Core 词汇表提供了表达知识组织体系基础结构和内容的核心模型,该词汇表给出了一系列的标签作为叙词表转换的统一标签,应用该标签进行转换构建的本体具有很强的通用性,利于本体重用和互操作性。SKOS Core 词汇表由一系列 RDF 属性和类组成,这些属性和类涵盖了各种词表中使用到的标签,词表中的概念词条通过给出的这些标签进行标识,从而融入到 SKOS Core 的概念框架中,形成计算机可以理解的语义知识组织体系。

SKOS Core 词汇表中标签很多,例如,标识词表的类 ConceptScheme,标识词表中概念的类 Concept,上位概念属性标签 broader,首要主题词属性标签 prefLabel,停用主题词属性标签 altLabel 等。这些标签集合构造出知识组织体系框架,可以用来携带各种类型的概念词表。各种不同类型的概念词表依赖该框架可以相互连接,共同构成跨词表的知识组织体系。SKOS Core 也提供了连接概念和人们常用词语的框架,可以用于支持多个任务,例如 Web 文档的自动分类,多语种术语表的自动翻译等。

## 3 基于 SKOS 的叙词表到领域本体的转换

### 3.1 本体编辑工具

目前,很多研究机构及学者开发出了本体编辑工具,这些编辑工具能屏蔽当前众多的本体描述语言的不同之处,通过提供一个良好的用户图形界面,使本体的构建和对信息资源的本体标注变得更为简便。如美国斯坦福大学的 Protégé、德国卡尔斯鲁厄大学的 KAON、英国曼彻斯特大学和阿姆斯特丹公立大学的 OILED 等。笔者比较了几种本体编辑工具,觉得斯坦福大学的 Protégé 界面友好,并且可以通过插件扩展各种功能,对于本体的图示表示灵活,可以支持多种本体语言[9]。笔者选择该工具作为本体编辑工具,完成了基于 SKOS 的叙词表到领域本体的转换。

### 3.2 UKAT 叙词表分析

UKAT(UK Archival Thesaurus)是英国档案部门创建的主题词表,用于档案文件的标引和检索,在此基础上构建了国家档案网。它所收录的词汇来源于 UNESCO(联合国教科文组织)叙词表。UNESCO 是一个涵盖了教育、科学、文化、社会及人文科学、信息和通信、政治学、经济学、法律的高水平的术语词汇表。到 2004 年 8 月为止,UKAT 共收录词条 19 698 个,其中 6 356 来自于 UNESCO 词表。UKAT 提供了检索档案库和查阅档案目录时使用的受控词汇,让用户能够有效地从主题角度检索和利用国家档案网[10]。

本文选择该词表进行本体转换,一是考虑该词表涵盖的范围较宽,是综合性的叙词表,从中可以选择大家比较熟悉的类目;二是等级排列的词表结构清晰,容易进行等级转换;三是词表中每一个词条都有 URI 标识定位,可以直接通过浏览器定位到相关词条。

以下节选了 UKAT 中信息与通信(Information and Communication)子表中的计算机应用(Computer Applications)子类下的各级类目作为 SKOS 表达的词条,在节选段落中以计算机应用作为一级类目,则节选词表可划分三级类目,如图 1 所示。

### 3.3 UKAT 叙词表到本体的转换

一个本体是关于一个领域概念的集合,概念的含义

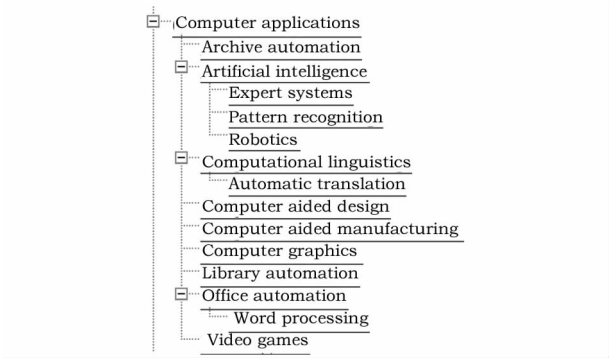


图 1 UKAT 节选段落图

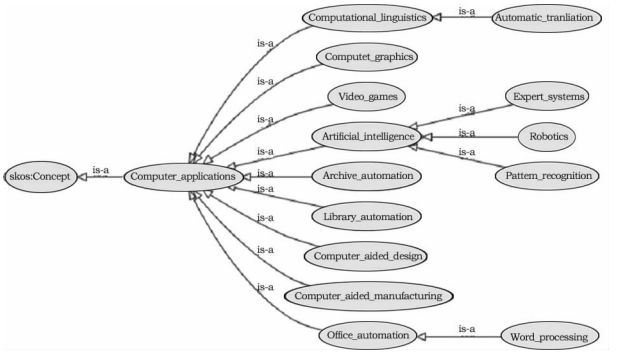


图 2 本体类目体系图

通过概念之间的关系来体现。而在一个具体的领域内，一个概念往往具有其特有的属性，即确定了一个概念，就决定了它所具有的属性；反之，由一个概念所具有的属性，也可以确定一个概念。一个本体包括一套关于某一领域概念的规范而清晰的描述类，描述了有关概念的各种特征的属性，还包括属性插件的限制条件，在构建完成的本体中还可以添加一系列与某个类相关的实例，这些实例共同构成了知识库系统[11]。

本体的构建过程具体实现步骤如下：

(1)引入名称空间 Namespace。名称空间是本体中使用一系列词条的前提，必须准确说明正在使用的特定的词汇表。一个本体标准的初始模块是包含在 rdf:RDF 标签中的一系列名称空间的声明。这些声明用以准确解释文档中的标识符，从而使本体的其他部分具有可读性。由于本例中所有词条都来自于 SKOS 前缀名称空间，因此在名称空间中添加 SKOS 前缀并给出对应的名称空间 URI。

(2)构建本体类 Class。类也叫做概念(Concepts)，表示对象的集合，它是本体最基本也是最重要的一个建模元语，是构建本体的基础。基于 SKOS 的词表本体构建中，类的构建主要用到 SKOS 中表示词表和概念的两个类 skos: ConceptScheme、skos:Concept。其中，skos:ConceptScheme 类为所选词表类，对于 skos:ConceptScheme 类给出相应的描述性元数据信息，说明所选词表的名称(dc:title)、节选的段落描述(dc:description)等信息。skos:Concept 类为词表中的概念类。词表节选段落中的所有概念都是 skos:Concept 的下位类。然后根据词表节选段落中概念的等级树的关系构建相应类目体系。以直观图的形式显示的类目等级关系如图 2 所示。

(3)构建本体属性 Property。属性可以被用来说明类的共同特征以及某些对象的专有特征。一个属性是一个二元关系。可以通过指定属性的 domain 和 range 以及定义子属性来约束一个属性更准确的表达特定范围对象之间的关系。SKOS Core 提供的标签绝大部分以属性的形式出现，本例中涉及到的属性主要有：skos:prefLabel、skos:scopeNote、skos:inScheme、skos:broader 、skos:narrower、skos:related、skos:altLabel。这些属性标签分别对应词表中的首选标签、范围注释、所属词表、上、下位概念、相关词条、替换词条等各种等级、相关关系。考虑到本文所做的仅仅是 UKAT 中节选的段落，不可能把整个 UKAT 中的相关词条都以类的等级形式列出，因此，本例中把 skos:related 标签作为注释属性。其中 skos:broader 、skos:narrower 属性涉及到的关系可以通过类的关系全部列出，因此设计为对象属性。

(4)为类添加属性。每一类对应着一些属性，用于对该类的描述和限定，如相关属性、首选标签、范围注释等。根据词表中给出的关系来添加这些属性，并给出对应的值。如 Computer_applications 类，它具有首选标签 skos:prefLabel 属性，其属性值为 Computer applications；它还具有 skos:scopeNote 属性，其属性值为 Use more specific descriptor。

对于构建完成的本体，可以通过图的形式更清晰地看出它的结构和关系，如图 3 所示。在图 3 中，每个方框代表一类的各种描述属性，包括相关类目、首选标签、范围注释等；箭头表明了各类之间的上下位等级关系，整体的结构清楚地展示出来。

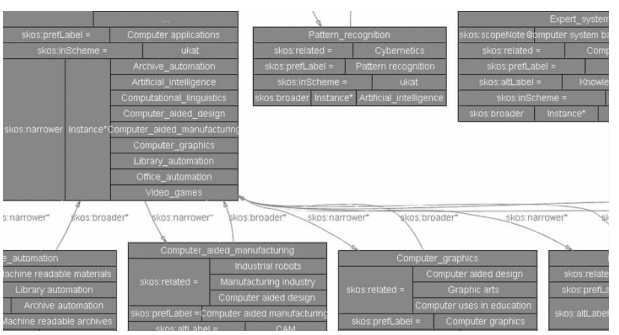


图 3 本体类目属性图

(5)为本体添加实例。完成本体的构建后，将简单的实例添加到本体中，这几个实例为单主题文献，可以直接添加到对应类下。对于添加的实例，通过属性 skos:broader、skos:narrower 标签给出彼此间的关系。完成后的类与实例对应如图 4 所示。

图中实例之间的等级关系可以明确地反映出来，如果该本体进一步扩展成完整的词表本体，那么其中实例

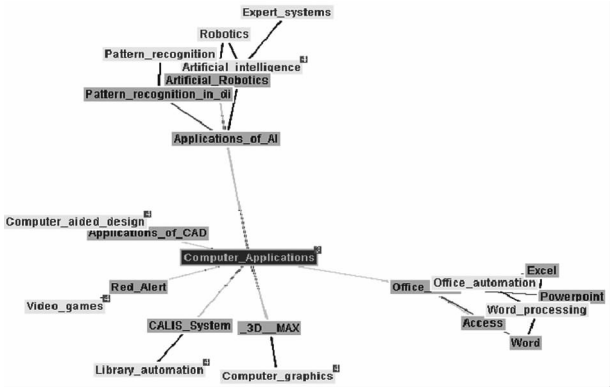


图 4 类与实例对应图

的关系不只限于等级关系,还包含相关的关系,实例之间的语义关系网络建立起来了。在此基础上如果对某一实例进行检索,除了可以得到这一实例外,还可以得到与其相关的其他实例,实现了基于语义的检索的扩展和延伸。

## 4 本体的评价

Mike Ushold 和 Micheal Gruninger 提出的本体构建方法——骨架法[12]认为,构建本体的流程中,对本体的评价标准是清晰性、一致性、完善性、可扩展性。清晰性是本体中的术语应被无歧义的定义;一致性指的是术语之间关系逻辑上应一致;完整性指本体中的概念及关系应是完整的,应包括该领域内所有概念和全部的关系,这一点是理想化的,在实际本体的构建中是无法达到的,随着领域不断的发展,本体需不断完善;可扩展性指本体应用能够扩展,横向扩展该领域新出现的概念,纵向扩展描述的深度。

对于本文建立的本体,依据以上的评价标准定性评价后,可以认为基本上满足清晰性、一致性、完善性和可扩展性的要求。从清晰性来看,本体中的术语取自专业的叙词表中经过规范控制的正式叙词,不存在语义上的歧异性;从一致性来看,本体的等级组织体系严格依据叙词表中树状结构建立,其它各种关系的建立也依据叙词表"用、代、属、分、参"的语义关联建立,因此在逻辑结构的划分上符合逻辑上的一致性;从完善性来看,本体中概念术语取自于领域概念术语的集合——叙词表,基本包含本领域的大多数概念和术语,随着领域不断地发展,新的概念不断出现,这些概念不断地充实到本体中;从可扩展性来看,该本体基于描述语言 RDF(S),具有良好的描述和推理能力,具有良好的扩展性。例如,需要丰富叙词表中给出的语义关系时可以扩展使用 OWL 提供的丰富的描述语言;需要进行复杂的逻辑的推理时,可以扩展使用 OWL 的限制约束关系。

从以上本体的构建过程可以看出,SKOS 基于 RDF 表达词表的结构和内容,相对于其他用于精确推理的本体描述语言来说逻辑上比较简单,不需要更多的约束和限制,词表本体的构建重点在于类、属性、关系的区分和交叉关系的把握。本文构建的本体只是初次 SKOS 应用的尝试,真正实现 SKOS 表示词表的完整本体构建还需要完成更多的工作。

**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. 图书情报工作,2006,50(6):23 - 27

2 刘春艳. 语义 Web 环境下基于 SKOS 的叙词表到本体的转换研究:[学位论文]. 吉林:吉林大学,2006

3 The Getty. http://www. getty. edu/research/conducting_research/vocabularies/aat/index. html(Accessed Mar. 28,2006)

4 Agriculture Ontology Service / Concept Server (AOS/CS). http://www. fao. org/aims/aos. jsp(Accessed Mar. 25,2006)

5 The CERES/NBII Thesaurus Partnership Project. http://ceres. ca. gov/thesaurus/(Accessed Mar. 28,2006)

6 Alistair Miles , Dan Brickley . SKOS Core Vocabulary Specification. http://www. w3. org/TR/2005/ WD - swbp - skos - core - spec - 20050510/(Accessed Apr. 5,2006)

7 Stair Miles, Brian Matthews, Dave Beckett, et al. SKOS: A Language to Describe Simple Knowledge Structures for the Web. http://iealliance. org/proceedings/xtech05/papers/03 - 04 - 01/(Accessed Apr. 8,2006)

8 Dave Reynolds, Steve Cayzer, et al. SWAD - Europe Deliverable 12. 1. 1: Semantic Web Applications - Analysis and Selection. http://www. w3. org/2001/sw/Europe/reports/chosen_demos _rationale _report/hp - applications - selection. html(Accessed Apr. 4,2006)

9 Protégé. http://protege. stanford. edu/(Accessed Mar. 20,2005)

10 UKAT. http://www. UKAT. org. uk/(Accessed Mar. 10,2006)

11 Brian Matthews, Alistair Miles, Michael Wilson. Modelling Thesauri for the Semantic Web. http://www. w3c. rl. ac. uk/SWAD/papers/thesaurus/swdbpaper. html(Accessed May. 10,2006)

12 Mike Ushold, Michael Gruninger. Ontology: Principles, Methods and Applications. Knowledge Engineering Review, 1996, 11(2):93 - 126


(作者 E - mail:tsglcy@126. com)