

科技知识组织体系范畴协同构建系统设计*^{*}

鲜国建, 孙巍, 赵瑞雪, 寇远涛, 朱亮, 罗婷婷

(中国农业科学院农业信息研究所, 北京 100081)

摘要: 作为科技知识组织体系构建研究任务的一部分, 在调研分析国内外知识组织体系构建管理系统基础上, 文章开展了STKOS范畴体系协同构建系统的需求分析、功能设计和系统实现, 支持对多来源、多类型范畴素材的统一存储与管理应用, 为理、工、农、医等多领域专家协作开展STKOS范畴体系构建管理提供网络化、分布式协同工作系统。最后, 指出该系统还应从支持多套、多语种范畴体系构建、灵活版本控制机制、范畴类目自动映射、语义化开放接口和系统集成整合等方面加以丰富完善。

关键词: 科技知识组织体系; 范畴体系; 协同构建; 知识组织; 国家科技图书文献中心

中图分类号: G254

DOI: 10.3772/j.issn.1673—2286.2014.11.005

1 引言

在“信息泛滥、知识匮乏”背景下, 为实现对海量外文科技文献信息战略资源的知识组织、深度揭示和知识关联, 促进基于文献信息内容的知识发现、知识挖掘和知识计算应用示范, 整体提升科技文献信息机构的知识服务能力, 国家科技图书文献中心(NSTL)组织实施了国家科技支撑计划项目“面向外文科技文献信息知识组织体系建设和示范应用”(项目编号: 2011BAH10B00)。该项目将构建我国面向外文科技文献的科技知识组织体系(Scientific & Technological Knowledge Organization Systems, STKOS), 覆盖来源术语、基础词库、规范概念、范畴体系和本体网络等多类型、多层次内容, 以更系统全面地支持海量科技信息的揭示、组织、发现和应用^[1], 通过该项目国内又重新掀起了一股知识组织体系建设的热潮。

范畴(Category)作为人类理性思维的一种逻辑形式, 是反映事物本质属性和普遍联系的基本概念, 任何一门学科均由一系列范畴构成^[2]。在科技知识组织体系STKOS中, 范畴是概念的重要属性, 用以描述概念所适用学科或所归属的类, 对文献信息的主题聚类、分类组织及浏览具有重要意义。STKOS范畴体系将覆盖理、工、农、医各个领域, 将由众多不同机构、不同领域专家, 基于多种来源范畴素材, 共同完成范畴体系的协作共建。然而, 调研国内外现有知识组织体系构建管理工具发现, 目前还没有能完全适合网络环境下开展STKOS范畴体系协同构建的系统或工具。

因此, 本文在借鉴国内外知识组织体系构建工具建设经验的基础上, 开展了STKOS范畴体系协同构建与管理的需求分析、功能设计和系统实现, 用以支持对多来源、多类型来源范畴素材的统一存储和应用管理, 为理、工、农、医等领域专家开展STKOS范畴体系构建与管理提供网络化、分布式的协同工作系统, 并在

* 本研究得到国家“十二五”科技支撑计划项目课题“科技知识组织体系的协同工作系统和辅助工具开发”(编号: 2011BAH10B02)和“面向外文科技文献的超级科技词表和本体建设”(编号: 2011BAH10B01)资助。

范畴体系质量检测与应用测试等方面提供辅助支持。

2 国内外知识组织体系构建管理工具

当前, 国内外已经研制了支持叙词表、分类表、本体等多种类型的知识组织体系构建管理工具。由中国科学技术信息研究所研制的汉语主题词表网络编制平台, 实现了网络环境下多人分工协作的叙词表编制与管理^[3]。军事科学院研制的《军用主题词表》应用管理系统集成应用词表管理软件^[4], 属于某一特定词表的专门管理工具, 但不能脱离该词表而用。国外可用于不同词表的编制和管理工具, 常见的有MultiTes Pro系列产品^[5]、WebChoir产品套装和Term Tree 2000等^[6]。本体构建工具应用较广泛的有Protégé、KAON和WebOnto等^[7]。联合国粮农组织 (FAO) 开发了AGROVOC Concept Server Workbench, 通过提供基于网络的分布式协同工作平台^[8], 满足农业领域专家和信息管理专家对多语种术语和语义概念信息的管理和维护, 并支持向农业本体的逐步进化。

然而, 还没有一款系统工具, 能有效实现对多领域、多学科、多来源、异构的海量知识组织系统进行有效集成整合, 并适合网络环境下开展STKOS范畴体系协同构建。随着Web 2.0等技术的广泛应用, 知识组织体系在网络环境下的协同编制也将成为主流。因此, 还有必要专门研究建立与STKOS范畴体系创建管理相适应的协同机制, 满足不同领域、不同机构协同操作和不同层次用户的联机编制需求^[9]。

3 STKOS范畴协同构建系统设计

3.1 功能架构

STKOS范畴协同构建系统的建设目标是: 面向STKOS范畴体系内容建设的业务需求, 遵循统一的元数据标准规范, 实现对多领域、多来源分类表和范畴类素材的集中管理与对比分析, 为理、工、农、医等领域专家协作开展STKOS范畴体系构建与管理提供网络化、分布式的协同工作系统, 并在范畴体系质量检测与应用测试等方面提供辅助支持。

STKOS范畴协同构建系统主要由来源范畴体系管理、STKOS范畴体系构建与管理、范畴体系对比分析、质量检测、应用评估和系统管理等功能模块组成, 其

中STKOS范畴体系构建与管理是系统的核心功能, 如图1所示。



图1 STKOS范畴协同构建系统功能框架

3.2 核心功能

3.2.1 来源范畴管理

来源范畴是指从现有叙词表、分类表、主题表等知识组织体系中具有学科分类体系或等级结构中析取出来的知识内容。来源范畴管理是在对多来源的专用分类表、主题表范畴、叙词表范畴等进行规范统一存储和集中管理基础上, 将来源范畴表及其类目推荐到范畴构建过程中, 为构建STKOS范畴体系提供重要来源素材和参考依据。

该模块支持对来源范畴表在线登记管理, 可对范畴表的名称、创建者、语种、类目数量等核心元数据项进行编辑, 可以标准的XML文件导入来源范畴类目元数据, 提供按范畴来源表、类目名称、类目编号等进行查询, 提供字顺和等级结构两种浏览方式, 为范畴构建人员提供启发和参考。

3.2.2 STKOS范畴体系构建与管理

(1) 类目创建管理

范畴类目是构成范畴体系的核心要素, 类目创建是范畴体系建设的核心环节, 是建立系统、科学、完备范畴体系的必经步骤。类目创建有两种途径来实现: 其一是从零开始, 即直接添加新范畴类目, 填写类目英文名称、中文译名、范畴类定义、注释、备注等信息。其二是基于来源范畴表快速构建, 即在浏览来源范畴表 (如

DDC及其他专业范畴)过程中,直接将来源范畴的一个类目或类目分支拖动并添加到正在构建的STKOS范畴体系中。

其中,新增类目的唯一标识符将由系统按照范畴标识管理中的规则自动产生赋值。范畴类目的基本信息除可编辑外,还可进行逻辑删除或者物理删除。通过快速构建方式创建新类目的同时,系统还将自动生成STKOS范畴体系与该来源范畴“等同”类型的映射关系信息。

(2) 类目结构调整

范畴类目等级结构和同级类目之间上下相对位置信息,是除类目基本信息外,范畴体系中最重要结构描述信息。类目等级结构及位置调整是范畴体系构建会涉及的另一重要操作。系统需提供将范畴体系特定类目或类目下所有子类目,在范畴树型等级结构之间升降、同级类目上下排列位置之间进行拖动和释放操作。

类目结构调整除了直接在树型结构中进行自由拖动,当范畴类目较多或层级较深时,也可提供新的窗口供用户选择目标类目。包括前面的类目创建管理等所有操作均需用户确认后执行,并即时刷新调整后的等级结构,让用户所见即所得。

(3) 范畴体系映射

异构网络信息环境和用户“一站式”的信息检索与知识发现需求,正在推动着知识组织体系互操作的研究实践。通过人工或者自动建立叙词概念和范畴类目的映射关系,是实现知识组织体系间互操作的重要方式。中国科学院国家科学图书馆的宋文研究员^[10]、南京农业大学的侯汉清教授^[11]等分别在叙词表、分类法映射方面已做了深入研究。

STKOS范畴体系映射,需要支持两种映射模式:其一是建立STKOS范畴体系与不同来源范畴表类目的各类型映射关系,映射类型包括等价关系、包含关系、包含于关系和相关关系等^[12]。系统将记录来源和目标范畴表、范畴类目唯一标识符、映射类型、操作用户和操作时间等信息。其二是在STKOS范畴体系类目之间建立内部映射关联,映射类型有用(Use)、代(Used For)和参见(See Also)三种,系统将记录范畴类目唯一标识符、映射类型等信息。两类映射结果都需要保存,并提供浏览和删除管理功能。

(4) 范畴标识管理

在数字信息环境中,唯一标识符承担着唯一且永

久地确认数字对象,并将它们与其他数字对象或服务系统连接起来的功能^[13]。在STKOS范畴体系中,共有两类范畴类目标识符:唯一标识符(DOI)和范畴编号(Code)。其中,范畴唯一标识符是唯一且永久的,主要用在建立类目等级关系和各类映射时对类目的标识引用;范畴编号唯一但不稳定,不同版本范畴编号可发生局部更改,主要用在对各类信息资源的分类标引处理。

范畴类目唯一标识符的生成规则是一位领域属性大写字母加六位数字组成。因此,在添加一级类目时,需要指定理(L)、工(G)、农(N)、医(Y)、通用(T)和综合(Z)等其中一种领域属性。一级类目及所有子类目的唯一标识符将由系统自动生成,子类目将自动继承一级类目的领域属性。

范畴编号生成规则主要遵循:两位数字表示一个大类;以数字的顺序反映大类的序列;各级类目编码均采用两位数字表示;每三个等级之间用圆点符号分隔。在创建一级类目时需要人工指定两位数字的范畴编号,如00。系统提供范畴编号自动批量生成功能,子类目将自动继承父类目的范畴编号,同级范畴类目编号由与其所排列的上下位置相对应,可按整表或局部类目(及所有子类目)自动生成,在版本已经稳定且需发布时才执行范畴编号生成操作。

(5) 类目查询浏览

为用户提供友好、便捷、易操作的交互界面,是用户愿意使用系统的重要因素。STKOS范畴构建系统需为用户提供多入口快速定位和多视图浏览多层级的类目。在范畴构建管理、类目映射和类目审核等操作界面,将提供按范畴类目英文名称、中文译名、唯一标识符、范畴编号等进行快速查询,检索模式包括精确检索、模糊检索、布尔逻辑检索等,可对查询结果进行快速定位,并在范畴树形结构上进行命中结果展开并高亮显示。

提供按类目字顺和类目等级结构体系两种浏览导航功能,为范畴等级结构及类目的构建管理提供操作入口。点击某个类目时,可以选项卡方式查看该类目的基本信息、下位类目、映射关系和审核状态等信息。

(6) 其他功能

数据导入导出是系统开放性和交互性的重要指标。类目导入导出功能是支持STKOS范畴构建系统与其他系统,以遵循一定描述规范的XML格式文件进行数据交换的主要方式。在操作时,可设置要导入导出的

类目等级范围,除范畴类目基本信息外,还可选择要导入导出的语义关系,包括STKOS内部类目上下位关系和映射关系,以及与来源范畴表的映射关系等。

类目审核发布是保证范畴体系质量的重要方式。系统将提供对范畴类目基本信息、类目间映射关系及与来源范畴类目映射关系的审核功能,对是否已经审核及是否通过审核的类目区别显示。在所有类目都通过审核的前提下,可以将STKOS范畴体系进行整体发布。

统计分析是为管理人员全面了解、准确分析范畴构建工作整体进展提供数字参考。需统计各等级、各领域范畴类目数量分布,也可按加工状态统计,统计时可动态选择所属领域、类目等级、起始日期等参数。统计结果以图表等形式输出。

3.2.3 范畴体系对比分析

范畴体系对比分析是指对范畴体系学科覆盖的完整性、体系结构的科学性和实用性等方面进行对比分析和综合评价。建立范畴体系对比分析辅助功能,可为构建STKOS范畴体系确定可重点参考的主干范畴体系,并明确主干范畴的缺陷及调整方向提供参考依据。

系统支持两两范畴表之间的对比分析,即来源范畴与STKOS范畴及其他来源范畴之间,基于类目名称相似度计算,提供类目等级分布、类目相同或相似等对比分析。可动态设置要对比的目标和来源范畴体系、对比分析的类目等级、类目名称相似度阈值等参数,分析结果可以保存到数据库,供用户随时查看分析结果,也可以重新计算并覆盖分析结果。

3.2.4 范畴质量检测

范畴质量检测控制是指在范畴构建过程中,遵循相关规范和标准,对用户构建的范畴体系进行规范性测试,保证通过本系统构建的范畴体系达到质量要求。本系统主要侧重于建立两类重要质量检测机制:形式错误检测和一致性错误检测,并能够自动生成检测报告供用户查看和分析,且将质量检测贯穿于范畴体系构建全过程。其中:

形式错误检测相对简单易操作,重点包括对范畴类目的重词形规范性检测,如类目拼写错误;重复类目

检测,如同等级中同一类目出现多次;异常字符(乱码)检测,如不符合UTF-8编码的字符;必备项检测(属性值为空检测)等。在STKOS范畴体系建设中,理、工、农、医等各领域的范畴类目交叉重复的问题较为突出,系统将提供各层级类目英文名称和中文译名的查重分析,为领域间的范畴合并提供参考。

关系结构和术语的一致性控制是词表管理系统最重要的基本功能^[14],在范畴体系构建系统中,类目结构等级及映射关系一致性检测也极其重要。主要包括对检测类目和关系唯一性的控制,避免同一类目或类目间关系重复建立;检测没有任何关系存在的孤立类目;上下位关系闭合(死循环)检测,检测类目在上下位等级传递关系中,出现越级或循环关系。

3.2.5 范畴应用评估

范畴体系主要应用于信息资源的分类标引、主题聚类和导航浏览等场景。范畴体系类目的设置是否科学、类目层级分布是否合理,都需要在实际应用中进行测试、检验和评估,并根据反馈进行调整完善。根据范畴体系内容建设需要,建立范畴体系应用评估辅助功能,提供类目层级设置、各层级类目分布、类目下概念归类情况进行统计分析。其中,类目下概念归类情况将与STKOS超级科技词表整体建设工作联合开展。

3.2.6 系统管理

系统管理是STKOS范畴协同构建系统正常运行的基础保障,由于STKOS范畴协同构建系统整合多领域的范畴素材,涉及多领域、多家机构、多学科的人员,为实现不同类型人员的分布、协同和精细化管理,系统将提供对范畴构建用户、用户角色、系统权限、加工任务、版本控制、加工日志等进行协同管理。

3.3 元数据标准规范

元数据标准是描述某类资源对象所有规则的集合。面向STKOS范畴体系构建与管理的实际需求,建立来源范畴表、来源范畴类目、STKOS范畴类目以及范畴映射等多类型元数据标准规范,便于范畴体系及其类目的规范描述、发现、管理和交换。表1是描述STKOS范畴体系类目的核心元数据项:

表1 STKOS范畴体系类目核心元数据

字段名	字段说明	数据类型	长度	必备	备注
doi	类目唯一标识符	varchar	20	是	基于规则在创建类目时自动生成
pdoi	父类目唯一标识符	varchar	20	是	父范畴类目, 一级类目的父范畴类目唯一标识符为“N”
prefLabel	优选名称	varchar	200	是	STKOS类目英文名称
altLabel	非优选名称	varchar	200		STKOS类目中文译名
classLevel	类目等级	int	11	是	类目在范畴树型结构所处层级
seqPosition	类目位置	int	11	是	同级类目前后(上下)排列位置
code	类目编号	varchar	100	否	范畴类目范畴号, 系统按规则自动生成
parentCode	父类目编号	varchar	100	否	父范畴类目范畴号, 一级类目的父范畴类目编号设为“N”
remark	备注	varchar	500	否	对范畴类目的备注说明
definition	定义	varchar	500	否	对范畴类目的定义
note	附注	varchar	500	否	对范畴类目的一般注释
status	加工状态	int	1	是	0新建待审核、1通过审核、2未通过审核、3已发布
checkNote	审核备注	varhcar	500	否	在审核类目时, 进行相关信息的备注说明

在表1中, 类目唯一标识符在新建范畴类目时, 按范畴标识管理中的规则自动生成且不再改变。为形成范畴树型结构, 需将上级类目唯一标识符写入父类目唯一标识符中, 一级类目的父标识符和范畴编号都将统一设为“N”。类目等级是为了便于快速组建树型结构和按层级统计分析, 而类目位置主要用于同级类目的排列显示和后期排版打印。

4 STKOS范畴协同构建系统实现

STKOS范畴体系内容构建是一项知识密集型任务, 需要整合多领域来源范畴素材, 将有多个领域、多家机构和多学科专家参与其中。为避免众多单机版或者C/S架构知识组织体系构建管理工具的缺陷, STKOS范畴协同构建系统必将以互联网为基础, 为各类用户提供便捷的访问操作入口, 以更低计算机技术要求随时随地参与范畴体系内容协同构建管理。

STKOS范畴体系协同构建系统将采用B/S的系统架构, 用户可直接通过网页浏览器就能完成范畴体系内容的构建和管理功能。在系统开发过程中, 采用了用户表示层(UI)、业务逻辑层(BLL)和数据访问层(DAL)三层分离的结构, 确保系统有良好的灵活性和可扩展性。系统开发语言以Java为主(JDK 1.6以上), 数据库选择MySQL 5.0以上。

在人机交互界面方面, 使用了富客户端开发技术(ExtJS 4), 以及Ajax异步更新等先进技术手段, 为用户提供简洁、美观、便捷、人性化、类Windows风格的操作界面和功能菜单, 以树节点拖动、下拉列表选择为主, 人工填写为辅, 最大程度提高范畴体系内容建设的效率, 如图2所示。



图2 STKOS范畴协同构建系统操作界面

目前, STKOS范畴体系协同构建系统各项核心功能已开发完成, 经过多轮用户测试和优化完善, 该系统已正式运行, 并为科技支撑计划项目课题“面向外科技文献的超级科技词表和本体建设”STKOS范畴体系研究构建工作组, 以及理、工、农、医等领域专家提供了协同工作系统支持, 先后完成了38个一级类目、319个二级类目、1441个三级类目, 最深层级为9级, 共计11408个类目的协同创建与管理维护。

5 结语

随着新一轮知识组织体系建设热潮的到来,对各类型、各层次的专业化知识组织体系构建支撑平台的需求也日益强烈。STKOS范畴体系协同构建系统的研究与建设,是在网络环境下对多领域科技知识组织协同构建与集成管理的一种新的探索和尝试,初步实现了对不同来源范畴体系的统一存储和管理,满足了跨机构、跨领域、多用户协同构建范畴体系的需要。然而,要使得该系统更具应用价值,还有诸多方面需要后续改进和加强,例如:同时支持多套、多语种范畴体系构建、灵活版本控制机制、质量监测与控制、范畴类目自动映射、语义化开放接口^[15]、范畴向SKOS本体自动转换^[16]等。同时,也迫切需要与其他相关知识组织建设系统的对接集成和联合测试,最终实现对术语、概念、范畴体系和本体在内的各类知识组织体系的一体化协同构建与集成管理。

参考文献

- [1] 孙坦,刘峥.面向外科技文献信息的知识组织体系建设思路[J].图书与情报,2013(1):2-7.
- [2] 彭彦琴.中国心理学思想史范畴体系的重建[J].心理学探新,2001(1):9-13.
- [3] 贺德方.国内外知识组织体系的研究进展及应对策略[J].情报学报,2010,29(6):963-972.
- [4] 赵建华,赵建国,杨肖宁,等.《军用主题词表》应用管理系统[J].情报理论与实践,1999,22(5):348-352.
- [5] MultiTes-Thesaurus and Taxonomy Authoring Tools [EB/OL]. (2014-01-25) [2014-10-25]. <http://www.multites.com>.
- [6] HEDDEN H. Comparative evaluation of thesaurus creation software [J]. The Indexer, 2008, 26(2): 50-59.
- [7] 李景.本体理论在文献检索系统中的应用研究[M].北京:北京图书馆出版社,2005.
- [8] YONGYUTH P, THAMVIJIT D, SUKSANGSRI T, et al. The AGROVOC Concept Server Workbench: A collaborative tool for managing multilingual knowledge [C]// IAALD World Congress, Tokyo (Japan), 24-27 August 2008.
- [9] 赵捷,曾建勋,吴雯娜.网络环境下叙词表协同编制系统的构建[J].图书情报工作,2011,55(22):6-10.
- [10] 陈辰,宋文.叙词表映射研究综述[J].图书情报工作,2014,56(12):113-119.
- [11] 戴剑波,侯汉清.文献分类法自动映射系统的构建:以《中国图书馆分类法》与《杜威十进分类法》为例[J].情报学报,2006,25(5):594-599.
- [12] 戴剑波,侯汉清.图书分类法映射系统设计原理:以《中国图书馆分类法》与《杜威十进分类法》为例[J].情报学报,2005,24(3):299-303.
- [13] 张晓林.数字对象的唯一标识符技术[J].现代图书情报技术,2001,(3):8-11,14.
- [14] 杜慧平.国外计算机辅助编制叙词表软件评价[J].现代图书情报技术,2009(9):17-21.
- [15] 杨锐,汤怡洁,刘毅,等.知识服务环境下语义化开放接口应用研究[J].图书情报工作,2014,58(4):99-104.
- [16] 曾新红.传统受控词表的SKOS描述规范及应用指南[J].数字图书馆论坛,2014,(4):9-19.

作者简介

鲜国建,男,1982年生,博士,研究方向:知识组织、信息系统开发,E-mail: xianguojian@caas.cn。

The Design and Realization of Collaborative Construction System of Category of Science and Technology Knowledge Organization System

XIAN GuoJian, SUN Wei, ZHAO RuiXue, KOU YuanTao, ZHU Liang, LUO TingTing
(Agricultural Information Institution, Chinese Academy of Agricultural Sciences, Beijing 100081, China)

Abstract: As one of the tasks of building the Scientific & Technological Knowledge Organization Systems (STKOS), this paper firstly makes some survey and analysis of existing knowledge organization system managing tools, then focuses on the requirements analysis, main function modules design and realization of the collaborative construction system of category of STKOS, to support the storage and application of different sources and types of category related materials, and also to provide a Web platform for groups of domain experts to create and manage category system of STKOS collaboratively. Finally, we think more functions are needed in the future, such as supporting managing more categories at the same time, working with multi-languages and flexible version control, automatic mapping of classes of categories, open semantic API and system integration.

Keywords: STKOS; Category; Collaborative construction; Knowledge organization; NSTL

(收稿日期: 2014-11-20)