

科技知识组织体系素材管理系统的设计与实现*

□ 钱庆 胡铁军 李丹亚 李军莲 孙海霞 吴思竹 / 中国医学科学院医学信息研究所 北京 100020

摘要: 为满足我国面向外文文献的科技知识组织体系 (STKOS) 建设的迫切需求, 在全面分析现有国内外知识组织系统和知识组织系统管理工具的基础上, 提出并实现STKOS素材管理系统的整体框架和主体功能, 重点解决多来源、多领域、海量异构知识组织系统快速、有效存储、质量控制和多用户、多任务协同机制等关键问题。该系统基本实现了多来源、多格式、多类型知识组织系统的统一存储、加工与管理, 为科技知识组织体系的建设奠定基础。

关键词: 知识组织体系, 词表管理系统, 词表整合, 质量控制

DOI: 10.3772/j.issn.1673—2286.2012.12.001

1 引言

计算机网络和信息技术快速发展, 大量传统知识组织系统, 如分类表、主题词表、叙词表、术语表等被网络化, 并且新型的知识组织系统, 如本体、语义网等不断地被建立, 用以支持网络环境下对海量科技信息快速、智能的组织、分析以及检索利用。我国长期以来缺乏大规模、综合性科技知识组织工具, 难以满足外文科技文献的知识发现、知识关联和知识计算的要求。构建面向计算机应用的科技知识组织体系 (STKOS) 对于支撑我国海量外文科技文献信息的组织和利用, 实现国家科技文献信息战略资源的有效组织、深度揭示, 提升我国科技文献信息机构的知识服务能力等具有重要的理论和现实意义。然而, STKOS的建设需要大量来自不同科

技领域的、多类型、异构的知识组织系统作为基础素材, 需要构建有效的知识组织系统管理工具和辅助工具对这些多样化、异构的海量素材进行有效管理和使用。

本文在深入分析STKOS海量、多来源、多类型、异构知识组织系统 (包括术语表、叙词表、主题表、分类表、科技词典、规范科技文档等) 素材特点, 充分借鉴国内外已有知识组织工具和知识组织系统管理工具建设成果与经验的基础上, 顺应网络化知识组织系统发展趋势, 尝试对STKOS素材管理系统的构建进行了探索和实践, 开发了支撑STKOS建设的素材管理系统, 用以支持对多来源、多格式、多类型的异构知识组织系统的统一存储、管理和利用, 为面向外文科技文献的知识组织系统的集成和整合奠定了重要基础。

2 现有知识组织系统管理工具相关研究

现有研究主要针对单一词表或某一领域中的多词表展开, 主要有两种^[1]: 1) 集成应用词表管理软件, 属于某一特定词表的专门管理工具, 它不能脱离该词表而独立使用, 如国内的《军用主题词表》应用管理系统^[2]; 2) 独立软件包, 可用于不同词表的编制和管理, 多为商业软件, 目前数量超过50种, 其中以下三种在国外较为常用: a) 澳大利亚This to That Pty Ltd公司开发的Term Tree 2000^[3]; b) 美国Multisystems公司开发的MultiTes系列^[4]; c) 美国洛杉矶WebChoir公司开发的WebChoir产品套装^[5] (包括TermChoir、ViewChoir、LinkChoir和SeekChoir)。但这些工具主要服务于单一词表的创建和维

* 本文系国家“十二五”科技支撑计划课题“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) 研究成果之一。

护, 缺乏对海量多词表的统一管理、还原、互操作、集成和整合能力。

目前, 对多领域、多学科、多来源、异构的海量知识组织系统进行大规模整合和集成的, 基于计算机的知识组织系统管理工具的研究在国外及国内尚未有成熟开展, 但当前, 知识组织系统整合集成、共建共享的发展趋势日益明显。如: 联合国农粮组织和欧共体合作开发的Agrovoc, 是涵盖农业、林业、渔业、食品安全等相关学科领域的多语种叙词表^[6]; 美国国立医学图书馆设计和维护的一体化医学语言系统UMLS, 对医学领域多类型词表进行了映射、转换和整合^[7]; 中文叙词表本体共建共享系统OTCSS, 实现了我国现有的130多部中文叙词表(主题词表)的本体化升级和网络化共建共享^[8]。这些已有研究成果对本文研究具有重要的借鉴经验, 引发了笔者对STKOS素材管理系统构建的深入思考。

3 系统整体框架设计

STKOS素材管理系统的主要目标是实现对多领域、多类型素材的整合管理, 为后继开展相关应用研究奠定基础。其传统的知识组织系统管理工具区别在于, 并非侧重于词表的构建, 而是侧重于对构建超级词表所需的多来源、多类型的知识组织系统进行有效的集成、整合、映射存储、浏览加工、编辑修改、质量控制、协同管理等。STKOS素材管理系统的建设需要紧密结合超级词表构建的需求, 结合待整合的海量素材的领域、格式、类型、内容等特点, 遵循统一的标准规范, 开展有效的系统建设工

作。因此, 本文首先对STKOS素材管理系统构建的关键问题进行深入分析, 然后在此基础上提出系统的整体框架和主要功能。

3.1 关键问题分析

(1) 海量多来源异构素材的入库

STKOS素材管理系统拟收录的科技知识组织系统有1000余部, 这些海量素材在领域、格式、逻辑结构、规模数量上存在较大差异。在类型上有术语表、分类表、主题词表、叙词表、本体、百科、同义词表等; 在格式上有word、txt、pdf、html、xml、数据库等; 在逻辑结构上, 不同编制单位采用不同的元数据体系结构; 在规模上, 少则几百条, 多则几百万条。海量素材的异构性, 导致素材管理系统收录和管理的困难。因此, 如何将这些形式、特点不同的素材进行有效规范、快速入库, 如何将元数据结构各异的素材有效存储并保留原结构特征, 整合为可用的有效资源, 成为STKOS素材管理系统构建要解决的关键问题之一。

(2) 异构素材的质量控制

质量控制能力是评价词表编制、维护和集成软件的重要指标之一^[9]。STKOS素材管理系统用以支撑超级科技词表的建设和应用, 其收录词表的质量至关重要, 直接影响后期词表建设。素材管理系统收录的知识组织系统格式各异、类型多样, 如何结合术语表、分类表、主题词表、叙词表本体等知识组织系统的结构差异, 提供系统、灵活的、有效的质量检测机制是STKOS素材管理系统构建要解决的又一关键问题。

(3) 多用户、多任务管理和系统协同机制

STKOS素材管理系统目标是支持多领域海量素材的整合和集成, 能够支持各领域、多单位异地用户同时进行素材的上传、质量检测、编辑修改、加工等操作, 因此, 如何保证多领域、多单位、多用户的协同管理与工作, 确保多任务并发控制机制的有效性是STKOS素材管理系统构建要解决的第三个关键问题。

3.2 系统功能框架

在对STKOS素材管理系统需要解决的关键问题进行深入分析的基础上, 本文提出STKOS素材管理系统构建的整体功能框架, 如图1所示, 主要包括7个功能模块: 素材登记、素材导入导出、元数据分析与管理、浏览与加工、统计报表、质量检测、系统管理。其中, 素材导入、导出、元数据分析与管理、质量检测和系统管理为重要的功能模块, 针对性地解决了前文提出的三个关键问题。

(1) 素材登记

支持对新到素材基本信息的登记和编辑, 包括名称、编制机构、类型、版本、学科、关系类型、版权信息、登记单位等信息的登记, 实现素材在词表层面的管理、描述和应用, 便于用户对素材概貌的了解。

(2) 素材导入、导出

现有知识组织系统格式各异、类型多样, 为保证异构素材的快速入库和统一存储与管理, 系统首先基于简单、易用、通用且能够充分保留原始语义信息的原则制定了统一的素材格式描述规范, 即尽可能统一采用以概念为描述中心的txt和xml格式, 图2是以概念为描述中心的规范txt格式示例。其次, 分类提供向导工具, 对于简单且规范的素材, 系统提供规范格式的素材入

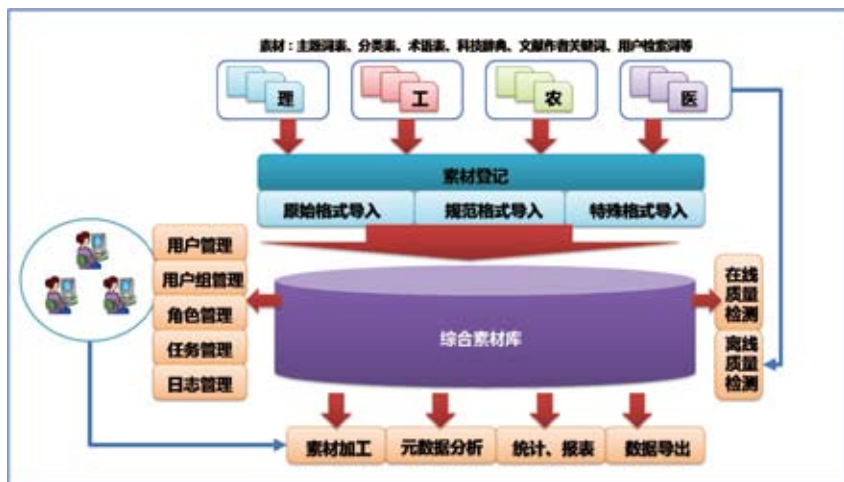


图1 STKOS素材管理系统整体框架

库向导工具；对于结构较为复杂的非xml格式素材，系统支持个性化处理。素材入库向导工具能够自动解析需导入素材的元数据结构 and 识别元数据标签，并实现素材元数据标签和系统指定的基本元数据标签（如表1所示）间的映射。对于基本数据标签不能涵盖原有素材的元数据标签，系统可辅助用户进行自定义和扩展，并对异常标签进行警示。

此外，为便于素材数据共享与交换，系统提供规范输出接口，支持规范格式（txt、xml）素材数据输出与调用。

(3) 元数据分析与管理

知识组织系统的入库要保证其原有结构和知识单元的正确性，元数据映射的准确性至关重要。为减轻用户负担，快速、准确实现海量异构素材的结构转换，系统提供元数据辅助分析和功能，由计算机辅助用户进行元数据结构的分析与映射，用户只需对结果进行审核，并根据需要对素材元数据显示模式进行编辑和调整即可。

(4) 质量控制

STKOS素材管理系统的目标是在素材词表充分还原的基础上，实现各种素材词表的统一存储和管理，为后续词表统一建设、浏览、互

表1 STKOS素材管理系统定义的基本元数据标签

序号	元数据标签	说明
1	ID	概念编号
2	PT	首选词、正式词
3	BT	上位词、广义词
4	NT	下位词、狭义词
5	RT	相关词
6	UF	非首选词、同义词
7	UFA	组合词
8	DEF	定义
9	CN(Tree Number)	树编号
10	Category	分类
11	USE	描述该非首选词（非描述元）所用的首选词（描述元）
12	USA	描述词条由两个描述元组成

操作和集成奠定基础，因此保证入库素材知识单元的正确性，词表质量的可靠性至关重要。STKOS素材管理系统拟收录的科技知识组织系统类型，包括术语表、分类表、叙词表、本体等，不同类型的知识组织系统的质量检测侧重不同。在对当前知识组织系统质量检查机制分析基础上，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素材管理系统建设的全过程。

1) 形式错误检测。包括对知识组织系统中知识单元的重复性、形式错误、编码错误等问题的检测。具体包括：

- 词形规范性检测，如词形拼写错误；
- 重复记录检测，如一个叙词及其属性出现了多次；

```

biotic community
USE biotope(5211)

biotope
SN A biotope is a homogeneous area of similar ecology which constitutes the habitat
of a specific animal and vegetable community.
MT 5211 natural environment
UF biocenosis
UF biotic community
UF ecological niche
BT1 ecosystem
BT2 physical environment

bipolarisation
SN Situation in certain countries where the political scene is dominated by the
struggle for power between two political movements. Not to be confused with two-
party system.
MT 0431 politics and public safety
UF bipolarization
BT1 politics
RT two-party system

bipolarization
USE bipolarisation(0431)

```

图2 规范txt格式示例

- 异常字符(乱码)检测,如不符合UTF-8编码的字符;

- 异常属性名称检测,如因换行符引起属性值作为叙词属性名的异常问题;

- 属性值为空检测,此项可能是数据项缺失引起的。

2) 一致性错误检测。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^[1,10]。任何一个词表编辑和维护软件都应该确保通过一致性检测,避免词表术语和结构的非法特征。具体包括:

- 一致性检测,主要包括对检测词条和关系唯一性的控制,避免同一词条或词间关系重复建立;

- 关系互逆和完整性检测,如根据词A的“代”、“属”、“参”词关系属性自动在其相应的对照词B的“用”、“分”和“反参”关系属性中检测A是否在其相应的值域;

- 孤儿概念检测,检测没有任何关系存在的孤立概念;

- 概念及其参项在同一属分链上检测,检测出现参照关系和属分关系的逻辑错误;

- 上下位关系闭合(死循环)检测,检测词在上下位等级传递关系中,出现越级或循环关系。

(5) 浏览与加工

STKOS素材管理系统支持对入库素材的浏览、检索和加工。素材检索主要支持单表和跨表检索,检索模式包括精确检索、模糊检索、布尔逻辑检索等。素材浏览包括两种方式,按词表的概念等级关系(树状结构导航)和字顺方式。同时,素材浏览中兼具素材内容加工功能:包括概念等级关系、等同关系的建立,概念词条的维护(正式词、非正式词的新建、编辑、删除、锁定、解锁等),以便用户在素材浏览中发现

素材中存在问题时及时进行修正。

(6) 统计与报表

系统提供统计分析和报表功能,目标是帮助用户实时了解系统内素材概况。

1) 统计分析,支持从不同维度(如学科、规模、类型等)、不同粒度(如统计对象可以是以整个素材为单位,也可以是素材中的定义、优选词等某个字段或字段值)对素材库中的词(概念)数量与分布、关系类型与数量进行统计,并生成相应的报告。

2) 报表,主要通过可定制的报表模板,供用户定制报表内容和格式,进行报表下载。报表内容包括素材基本登记信息、元数据信息、加工数量统计信息等,系统提供多领域综合素材报表或单个领域多表或单表的素材报表的下载。

(7) 系统管理

STKOS素材管理系统的系统管理功能主要实现对素材加工用户、用户角色和加工任务的协同管理,由用户管理、用户组管理、角色管理等重要模块组成。

STKOS素材管理系统整合多领域素材,涉及多领域、多家单位、多学科的人员,为实现不同类型人员的协同、灵活管理,系统将用户管理对象分为用户组 and 用户两种。用户组的设立实现了以学科领域为单位的用户管理,其可灵活组织、管理并指定组内特定用户的角色和权限。系统通过用户角色模块为各个功能模块提供角色定义和用户访问权限的管理功能,目前系统内共有五大类角色系统管理员、单位管理员、数据加工员、审核员、浏览用户。不同角色具有相应的系统操作权限。通过权限的控制可以便于各领域及各层次用户的素材导入、调

整、管理等操作。

4 系统物理实现

STKOS素材管理系统的实现采用应用服务层-数据访问层-数据层三层B/S结构(如图3所示)。该结构提高了客户端应用程序的通用性、独立性,通过应用服务层中集成多种功能组件,能够提高系统开发、使用和升级效率。

(1) 应用服务层

作为人机交互的界面,负责处理用户的输入、输出以及大批量的事务处理,它将各种复杂业务分解为多项功能单一的服务,每一项服务通过相对独立的组件来实现。

(2) 应用程序层

包括系统管理程序模块和系统自身应用程序模块。系统管理模块和应用功能模块在逻辑上相互独立,有较为明显的系统边界,保证今后相关系统开发时的功能调用。各个应用功能模块设计划分时,也以界面、逻辑、实现分层次分隔的方式实现,通过对数据库的共享来实现逻辑上的集成。

(3) 数据库层

主要涉及两类数据:一类是系统数据,用于存储系统的用户帐户、工作任务安排、用户加工操作日志等系统管理模块产生的数据;另一类是素材内容数据,采用关系型数据库保存导入的各种素材词表。数据层与应用层通过JDBC连接。

系统整体开发环境采用JAVA开发平台,以Java Development Kit(JDK)为开发工具。系统建立和运行于64位数据处理环境之下,使数据库进程和应用程序进程可以直接访问2G以上的物理内存,提高系统的整体性能,适于大规模数据

处理。系统服务器操作系统采用 Windows Server 2008 (64位), 数据库管理系统是Oracle 11 (64位), JAVA应用服务器为WebLogic (64位), 用于部署和管理大型分布式Web应用和数据库应用。

目前, STKOS素材管理系统已经完成开发, 实现了用户素材登记、数据上传与下载、多维度统计分析(学科、加工状态、素材类型

等)与报表、检索浏览与编辑、质量检测、系统管理等核心功能, 并在国家科技支撑计划项目“面向外科技文献信息的知识组织体系建设与应用示范”中得到了应用, 为科技知识组织体系建设提供稳定和可靠的海量素材管理平台, 为后期的基础词库建立、海量外科技文献信息的组织和利用、知识的深度发现和揭示提供了重要基础支撑。

5 结语

在传统知识组织系统网络化、新的知识组织系统不断涌现、多知识组织系统共建共享快速发展的趋势下, STKOS素材管理系统建设, 是在网络环境下对多领域科技知识组织系统整合和集成管理的一种新的探索和尝试, 初步实现了对不同来源、不同类型的异构海量素材的统一存储和管理, 实现了多领域用户、多任务的协同工作。然而, 这只是面向外科技文献的科技知识组织体系建设任务的基础, 下一步笔者等将继续探索知识组织系统的计算机化表示规范、集成映射标准, 探讨海量异构知识组织系统素材在内容层面的关联、集成机制, 以更好地支持基于科技文献的知识发现、语义关联、语义分析和语义计算等大规模语义应用。

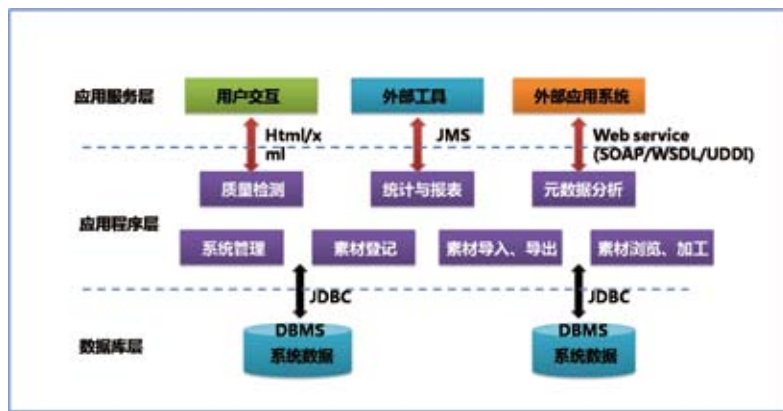


图3 STKOS素材管理系统实现的三层架构

参考文献

- [1] 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.
- [2] 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(5):348-352.
- [3] Term Tree [EB/OL]. [2012-04-15]. <http://www.termtree.com.au/html/termtree.html>.
- [4] MulTies Pro [EB/OL]. [2012-04-15]. <http://www.multites.com>.
- [5] WebChoir [EB/OL]. [2012-04-15]. <http://www.webchoir.com>.
- [6] Agrovoc [EB/OL]. [2012-04-15]. <http://aims.fao.org/website/AGROVOC-Thesaurus/sub>.
- [7] UMLS [EB/OL]. [2012-04-15]. <http://www.nlm.nih.gov/research/umls/>.
- [8] 曾新红, 明仲, 蒋颖, 等. 中文叙词表本体共建共享系统研究[J]. 情报学报, 2008, 27(3):386-394.
- [9] GANZMANN J. Criteria for the evaluation of thesaurus software [EB/OL]. [2012-04-15]. <http://www.willpowerinfo.co.uk/ganzmann.html>.
- [10] 白海燕. 传统知识组织工具管理软件的发展与评价[J]. 现代图书情报技术, 2008(9):47-52.
- [11] 曾新红, 林伟明. 中文叙词表本体一致性检测机制研究与实现[J]. 现代图书情报技术, 2008(5):1-9.

作者简介

钱庆 (1970-), 中国医学科学院医学信息研究所研究员。研究方向: 数字图书馆、知识组织。E-mail: qianq@imicams.ac.cn

The Design and Realization of the Management System of Science and Technology Knowledge Organization System

Qian Qing, Hu Tiejun, Li Danya, Li Junlian, Sun Haixia, Wu Sizhu / Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, 100020

Abstract: To meet the urgent needs of Chinese science and technology knowledge organization system construction for foreign papers, based on the analysis of the current status of the dominating knowledge organization tools and management system, this paper presents the design and realization of the material management system of System of Science and Technology Knowledge Organization System (STKOS), and discusses the key problems of accurate and fast storage, quality control for mass heterogeneous vocabularies from different fields, and multi-users and multi-tasks collaboration mechanisms. The system implements basically the unified storage, processing and management of knowledge organization system with multiple sources, various formats and complicated types, which lays a foundation for STKOS construction.

Keywords: Knowledge organization system, Vocabulary management system, Vocabulary integration, Quality control

(收稿日期: 2012-08-01)