DOI:10.13530/j.cnki.jlis.160012

国家叙词库的体系结构与数据模型*

吴雯娜　鲍秀林

摘　要　国家叙词库是我国叙词表资源集成与服务系统,体系结构包括三层:数据获取与转换层、存储与语义集成层、服务与应用层。数据获取与转换层实现叙词表数据的采集、规范描述和格式转换;语义集成采用中心词表集成模式,顶层为分类表和本体,中间为概念层,底层为基础词库;服务方面,采用渐进服务思路,提供不同颗粒度和不同语义层次的数据服务。叙词表宏观结构方面,将叙词表及其各组成部分作为独立概念体系,通过描述概念体系间的关联,表达叙词表宏观结构;微观结构方面,将概念和词汇分别作为描述对象,词汇作为概念标签,在概念层面描述语义属性,在词汇层面描述非语义属性。基于 DC 元数据框架设计叙词表元数据方案,基于 SKOS+XL 设计概念和词汇描述的元数据方案。图 3。表 3。参考文献 19。

关键词　国家叙词库　叙词表结构　知识描述模型　元数据　语义集成系统　术语服务　SKOS+XL

分类号　G254

The Architecture and Data Model of National Thesauri Warehouse

WU Wenna & BAO Xiulin

ABSTRACT

It is the purpose of the National Thesauri Warehouse (NTW) project to integrate the thesauri constructed and published in China and provide services. Several integration projects at home and abroad were investigated, analyzed and compared. According to the situation of the Chinese thesauri, NTW adopted such integration mode: Chinese Thesaurus was chosen as the mapping center, while other thesauri would be mapped to it.

NTW has a three-layer architecture: data acquisition and conversion layer (DACL), storage and semantic integration layer (SSIL), service and application layer (SAL). DACL has three modules: thesaurus metadata registration, data import and verification, unified description and format conversion. SSIL has a classification and an ontology above, a concept library in the middle, and a base term library below. The top classification is used as categories of concepts according to their disciplines, subjects or topics. The top ontology is used as categories according to their essential attributes. The top classification and ontology gather concepts from different perspectives to facilitate multi-thesauri semantic integration and concepts navigation in one thesaurus or cross-thesauri. The concept library contains concepts and their semantic attributes from member thesauri. The base term library contains terms and their non-semantic attributes, not just the terms of member thesauri, also including other term sources, such as keywords, dictionary terms, etc. In serv-

* 本文系国家社会科学基金项目“国家叙词库构建方式与发展机制研究”(编号:13BTQ013)的研究成果之一。(This article is an outcome of the project “Research on Construction Mode and Development Mechanism of National Thesauri Warehouse” (No. 13BTQ013) supported by National Social Science Foundation of China.)

通信作者:吴雯娜,Email:wuwenna@istic.ac.cn, ORCID: 0000-0001-8017-3232(Correspondence should be addressed to WU Wenna, Email: wuwenna@istic.ac.cn, ORCID:0000-0001-8017-3232)

ices, NTW follows the principle of gradual improvement, by using its construction achievement, which have different granularity knowledge units and different semantic levels, to provide data services. The services include thesaurus metadata services, concept services, concept group services and thesaurus customization services. According to different users, these services show different forms: query, browse, edit, download services for human users, and third party invoking services for computer agents.

NTW data model was built based on the analysis of thesaurus' macrostructure and microstructure. Thesaurus' macrostructure is the conjunctive structure of concept schemes. Concept schemes in a thesaurus regularly include main tables, auxiliary tables, category tables and indexes. The macrostructure is represented by describing the relationship between concept schemes in a thesaurus, and between the versions of the thesaurus. Thesaurus' microstructure is a frame of knowledge description. In NTW, concepts and terms are separated. Terms are labels of concepts. Concepts and terms are both objects of description. Concepts are described by using semantic attributes, such as concept definitions, categories, relations between concepts, and so on. Terms are described by using non-semantic attributes including phonetic notations, non-semantic corresponding relations between terms, such as Chinese labels/English labels, full names/acronyms, wrong spellings/correct spellings, and so on. Concepts and terms are linked through the label attribute which attaches terms to their labeling concepts. Two special relationships in thesauri were discussed, including concept coordination relationship and guiding associative relationship. These two relationships both link concepts from informal concepts to formal concepts in the same thesaurus. To separate and describe the concepts, "core concepts" and "extended concepts" were put forward. Formal concepts are core concepts while informal concepts are extended concepts. Thesaurus metadata scheme was designed based on DC. Concept and term metadata schemes were designed based on SKOS+XL. 3 figs. 3 tabs. 19 refs.

KEY WORDS

The National Thesauri Warehouse. Thesaurus structure. Knowledge description models. Metadata. Semantic integration systems. Terminology services. SKOS+XL.

0 研究背景

叙词表是一种知识组织工具,主要用于文献信息的主题标引和检索。叙词表具有词汇控制和概念关联的功能,其知识描述方式具有形式化、结构化的特点。因此,叙词表知识体系具有机器理解的可能性。推动叙词表的应用,将加快信息资源组织和利用的语义化进程,提高文献信息加工处理的智能化水平。

我国编制出版的叙词表有 100 多部,目前,继续维护和对外提供服务的叙词表在 10 部左右,大部分叙词表还是印本形式,疏于管理,难以利用。因此,在国家层面集成各种叙词表,建设国家叙词库,对我国叙词表资源进行揭示,改变其存在状态,使之易于获取和利用,并在使用中完善和发展,成为很紧迫的任务。2013 年,笔者申请的社科基金项目"国家叙词库构建方式与发展机制研究"得以立项。

"国家叙词库"是我国叙词表资源的集成与服务系统,其主要任务是:①对我国各学科领域叙词表进行数字化保存和网络化服务,使之易于获取,易于利用;②逐步解决多词表的语义互操作问题,促进跨系统信息资源的共享;③为基于叙词表的信息组织提供各种语义层次的数据服务,为面向特定应用的叙词表构建提供语义资源和技术辅助。

1　叙词表集成系统的国内外研究现状

叙词表不同，其构建的知识结构、采用的描述方式、使用的语言、覆盖的专业等也不尽相同。叙词表品种越多，叙词表兼容的问题就越严重。早在20世纪90年代初，机械电子部机械科技情报研究所傅兰生、朱岩就撰文提出建立“国家叙词库”来解决叙词表兼容问题。当时提出的“国家叙词库”有以下特征：“词库中只容纳经过规范化的叙词，不容纳词间关系”[1]；“搞统一的范畴划分，并与《中国图书资料分类法》及世界上一些通用的分类法建立对照关系”[2]。近20年来，国内外出现了一些有代表性的研究与实践项目，下面介绍在集成模式上比较有代表性的几个项目。

（1）HILT

HILT（High-Level Thesaurus，高级叙词表）项目[3]的词表集成模式是以DDC（Dewey Decimal Classification，杜威十进分类法）作为映射中心，将LCSH（Library of Congress Subject Headings，美国国会图书馆标题表）、UNESCO（United Nations Educational，Scientific and Cultural Organization，联合国教科文组织）叙词表、AAI（Art & Architecture Thesaurus，艺术与建筑叙词表）等15种受控词表的词汇对应到DDC类目，通过这种模式实现词表间的语义互操作。该项目之所以在众多的词表中选择DDC作为映射中心，主要基于以下考虑：①DDC是综合性类表，覆盖了全部的学科领域；②DDC的类目标识系统可以消除自然语言的歧义问题；③DDC被译成了30多种语言，有着广泛的应用基础[4]。

（2）OCLC词表映射项目

1994年，OCLC（Online Computer Library Center，联机计算机图书馆中心）开始在DDC和LCSH之间进行映射，后逐步扩展了词表范围，将包括ERIC（Education Resources Information Center，美国教育资源信息中心）叙词表、LCC（Library of Congress Classification，美国国会图书馆分类法）、MeSH（Medical Subject Headings，医学标题表）在内的其他6种叙词表映射到DDC或/和LCSH，包括LCC/DDC、MESH/DDC、ERIC/LCSH、MESH/LCSH等9个映射任务[5-6]，DDC和LCSH成为事实上的映射中心。OCLC之所以选择LCSH作为映射中心，主要考虑以下几个方面：①LCSH是应用最为广泛的词表；②LCSH被翻译成多种语言，被很多国家采用；③LCSH是英语语种最大的综合性词表[5]。

（3）KoMoHe项目

KoMoHe（Kompetenzzentrum Modellbildung und Heterogenitätsbehandlung）是德国最大型的词表映射项目。该项目集成了社会科学领域的25种受控词表，包括英、德、俄三种语言[7-8]，采用的是词表之间相互映射的语义集成模式。根据实际映射情况，THESOZ（Thesaurus Sozialwissenschaften，社会科学词表）是对内和对外映射最多的词表，所以在整个映射网络中处在最中心的位置。较次之，SWD（Schlagwortnormdatei，标准规范文档）、PSYNDEX（Psyndex Terms，心理学词表）也起了映射转换中心的作用。

（4）UMLS

UMLS（Unified Medical Language System，一体化医学语言系统）包括三个知识源（Knowledge Sources），即超级词表（Metathesaurus）、语义网络（Semantic Network）、专家词典及词典工具（Specialist Lexicon and Lexical Tools）[9-10]。超级词表是在来源词表之上构建的概念层：不同来源表的词汇依据等同关系聚集形成超级词表概念，即超级词表的一个概念可能对应不同来源表的概念及其各种词汇表达形式。UMLS 2015AA版超级词表整合了178部词表[9]。语义网络可视作顶层本体，为超级词表概念提供统一的范畴体系。专家词典及词典处理工具为跨词表词汇等同关系的发现提供支持。相比于之前介绍的项目，UMLS已基本形成自有的概念体系，说是“基本形成”，是因为超级词表概念只继承来源表概念的关系，而不对关系集合进行甄别梳理。

（5）NSTL英文超级科技词表项目

“英文超级科技词表”（简称“英表”）是我

国“十二五”国家科技支撑计划项目“面向外文科技文献信息的知识组织体系建设与应用示范”的主要研制内容之一。该项目由国家科技图书文献中心牵头,多家情报机构分工协作完成,是我国迄今为止规模最大的词表语义集成项目。

“英表”集成了英文语种的叙词表一百多部,覆盖自然科学和工程技术各学科领域。“英表”的语义集成模式基本是参照 UMLS,从逻辑结构上看,自下而上包含四层:素材层、基础词库、概念层、范畴体系。素材层包括相关专业的叙词表、分类表、专业词典、术语集等。经过遴选的“素材”(主要是叙词表)以整表为单位进入基础词库。基础词库中不同来源的叙词表词汇进行同义词归并后形成概念层的“概念”,“概念”继承各同义词在来源表中的原有关系,对关系不做梳理。在概念层之上设置范畴体系,范畴体系是概念层概念的分类框架,概念被归入相应类目,按学科特征得以聚集,并借助范畴体系的等级结构由一般到具体层层展开①。

“英表”范畴体系与 UMLS 不同。UMLS 的“语义网络”是一种形式分类体系,其等级结构遵循严格的属种关系语义限定,每种语义类型有明确的定义。“语义网络”有利于揭示类中概念横向的关联关系,在一定程度上能实现知识推理,可视作顶层本体。“英表”顶层的范畴体系是基于 DDC 构建,是典型的主题/学科分类思想,等级结构不是严格的属种关系,类目语义的界定和形式化表达也不够。这些特征使“英表”基本不具备知识推理的可能性,但这个范畴体系能起到按学科或主题聚拢概念的作用,有利于实现文献信息处理的分类主题一体化。

从语义集成模式看, HILT 采用卫星表向中心表映射的集成模式,这种集成模式较易控制,词表集成成果的服务和应用都可以基于中心表去扩展。但这种模式对中心表的选择有较高要求,中心表在学科范围上,在概念的深度和广度上要能够覆盖卫星表,中心表的知识结构和描述方式与卫星表要有较好的兼容性。KoMoHe 项目采用了多表相互映射的集成模式,这种集成模式复杂度较高,映射的完整性较难控制,映射中心的形成具有较大的随机性,在集成成果的服务和应用方面缺少可依托的主干。OCLC 词表映射则兼具以上两种集成模式的特征。UMLS 和“英表”采用在多来源表之上构建概念层的集成方式,傅兰生等在 20 世纪 90 年代提出的“国家叙词库”也是这样的集成思路。这种集成模式可以形成一个统一的概念集,有利于集成成果的服务和应用,但由于不具备自有关系网络,而是继承来源表关系,因此知识体系中最重要的等级结构必然变得混乱纠缠。虽然可追溯和利用来源表的关系进行弥补,但在来源表众多、关系网络复杂的情况下,又会给用户带来较大的使用负担。

2 国家叙词库的集成模式与体系结构

通过对国内外知识组织体系集成模式的调研和对比分析,并结合我国叙词表编制的实际情况,“国家叙词库”采用《汉语主题词表》(以下简称《汉表》)作为词表映射的中心表。这主要是基于以下两个方面的考虑。

(1)《汉表》与其他叙词表有较好的兼容性

1975 年,我国启动了“748”工程配套项目《汉表》的编制工作。在《汉表》成功编制的基础上逐渐形成了系列叙词表相关的国家标准,包括 GB 3860—83 文献主题标引规则、GB 13190—91 汉语叙词表编制规则、GB/T 15417—94 多语种叙词表编制规则。我国编制的叙词表绝大部分出现在《汉表》之后,基本是在《汉表》大的知识框架下进行学科专业的细化补充和调整,知识体系与《汉表》有较好的承接关系,在词表宏

① 关于“英表”集成模式的阐述引用了该项目课题一——“面向外文科技文献信息的超级科技词表和本体建设”任务书中课题技术方案的相关内容。

观和微观结构的表现形式上也与《汉表》基本一致。《汉表》是一部大型的综合性叙词表，包含社会科学、自然科学和工程技术方方面面，是学科覆盖最为完整的一部词表，其收录的专业词汇数量也最为庞大，因此，在概念体系的广度和深度上能较好地覆盖现有的专业叙词表。

(2)《汉表》有较好的维护和应用基础

我国叙词表的维护和利用总体不太好，大部分叙词表没有修订过，知识体系有待更新，原有的编制单位很多已不具备叙词表编制维护的能力。这些叙词表可以作为《汉表》的延伸和补充，却很难再独立进行服务和应用。相对而言，《汉表》维护较好，第一版《汉表》出版于 1980 年，收词 10.9 万条。1991 年修订出版自然科学增订本，包含自然科学和工程技术部分，收词 8.1 万条。2014 年，《汉表》(工程技术卷)出版，为满足文献信息处理自动化的需要，大量增补词汇，收词 36 万条[11]。2015 年启动了自然科学卷的修订工作。

基于以上原因，“国家叙词库”采用以《汉表》作为映射中心，其他词表向《汉表》进行映射的语义集成模式。其体系结构分为三层：数据获取与转换层、存储与语义集成层、服务与应用层(见图 1)。

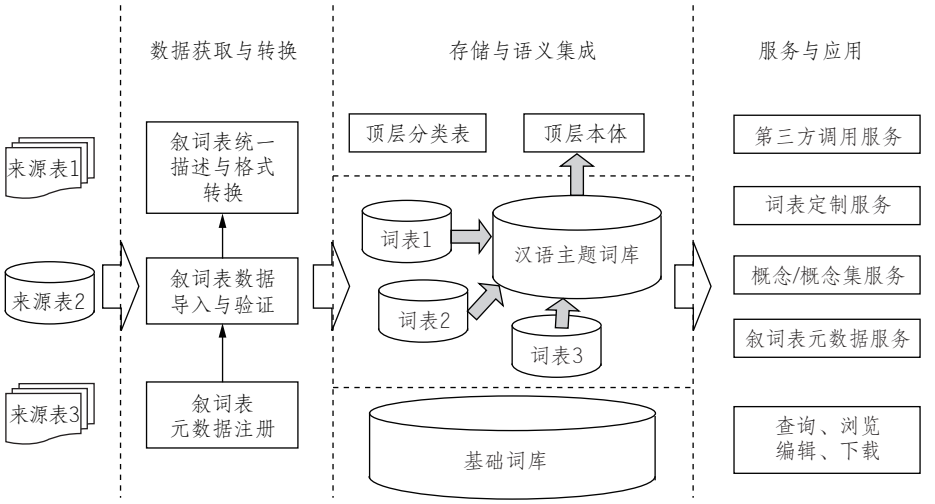

图 1　国家叙词库体系结构

2.1　数据获取与转换

我国的叙词表大多受《汉表》影响，在结构和表现形式上具有较多共同点，有利于对叙词表进行统一规范的描述，并采用相同的格式进行存储。但由于学科领域不同、编制单位不同、年代跨度大，叙词表在宏观结构和微观结构的表现形式上也不完全一致。国家叙词库需要对我国叙词表结构与表现形式上的共性和特性进行调研和分析，解决叙词表知识结构的统一描述问题。

“数据获取与转换层”的主要功能是采集叙词表数据，对之进行规范描述和格式转换。具体包括三个模块：叙词表元数据注册、叙词表数据导入与验证、叙词表数据统一描述与格式转换。

叙词表元数据注册是对“表”的元数据信息的著录。叙词表元数据包括叙词表名称、编制单位、出版时间、学科归属、版权信息等。叙词表元数据注册可汇总我国各阶段编制的叙词表基本信息，帮助用户了解和定位可利用的叙词表资源。叙词表数据导入与验证模块是将已注册的叙词表中的词汇、关系等数据导入系统，并对数据质量进行检查和控制。印刷本词表要先进行数字化，已经数字化的词表由于来源不同，

格式也可能不同,系统需具备支持多种格式的数据导入功能。早期编制的词表受限于当时的技术条件,词间关系或多或少有一些逻辑问题,比如存在关系冲突、冗余、循环等[12]。验证模块需具备逻辑问题查找和自动处理的功能,在保证逻辑正确的同时尽可能降低词汇信息的损失。经验证的叙词表数据即可按照统一的元数据方案进行描述,并以统一的格式进行存储。

2.2 存储与语义集成

2.2.1 顶层分类表与本体

叙词表一般都自带一个范畴表/分类表用于词汇归类。词表不同,所带的范畴表/分类表一般也不同。对于多表集成系统来说,顶层分类表为各来源词表的概念提供了统一的导航体系,有利于实现多词表概念的语义集成。国家叙词库语义集成系统从学科和本体的角度在顶层设立了两个类表:顶层分类表和顶层本体。顶层分类表对各来源词表的概念按学科、专业或主题进行归类;顶层本体对各来源词表的概念按事物本质属性进行归类。顶层分类表和本体可从不同角度对概念进行汇聚,辅助实现多词表概念的语义集成,也可进行单表或跨表的概念导航。

顶层分类表须采用能覆盖各学科领域的综合性分类表,要有一定的权威性和影响力,有广泛的认知和应用,以便和其他信息系统更好地兼容。新版《汉表》(2014 年)的分类体系以《中国图书资料分类法》(第 4 版)为基础,参考了《中国图书馆分类法》(第 5 版)进行编制,对两部分类法既有继承又有创新,既能满足文献信息分类的需要,又能更好地适应词汇/概念分类的需要。采用新版《汉表》的分类表作为国家叙词库语义集成系统的顶层分类表,是现阶段最好的选择。

叙词表概念体系的结构与本体有着相似之处,但叙词表没有对概念的属性空间进行明确定义,因此,叙词表概念间横向的相关关系是模糊的。引入顶层本体有利于叙词表往本体方向演进。将叙词表概念归入顶层本体相应的“类”中,可汇聚具有共同属性的概念,概念被归入相应“类”后,可继承“类”的属性空间,使相关关系进一步明晰成为可能。顶层本体的构建需要充分调研各来源词表词族构建的情况,分析各学科领域词族分布的特点,确定分面的构成情况,进一步确定基本的语义类型和等级框架。“类”的属性空间也需要结合学科特征和未来应用的需求,采用逐步深入和开放发展的思路进行设计。

2.2.2 概念库与基础词库

国家叙词库的网络化存在与发展决定其描述须符合语义网规范。国家叙词库采用基于 SKOS+XL 的知识描述模型(参见 4.1),将叙词表概念和词汇进行区分,词汇作为概念“标签”,在概念层面描述语义属性,在词汇层面描述非语义属性。国家叙词库中的各成员词表逻辑上是独立的,物理上则统一进行存储。概念及其语义属性的集合形成概念库,词汇及其非语义属性则统一存放于基础词库。概念库中的概念通过标签属性链接到基础词库中的词汇。

基础词库为各类词汇及其属性的集合,不仅包括国家叙词库成员表中的词汇,也包括其他术语来源的词汇,比如文献关键词、用户检索词、专业词典词汇、全国科学技术名词审定委员会发布的规范名词、术语标准中的术语等。随着国家叙词库建设的推进和深入,在逐步实现叙词表概念语义集成的同时,将基础词库中除叙词表之外的高频词、规范词、新词等逐步纳入叙词表概念体系中,建立起概念和词汇间更多的关联,这将有效地提升叙词表应用性能,进而改进文献信息资源组织和利用的效果。

2.3 服务与应用

网络环境中的知识组织系统有以下几个应用方向[13]:①作为信息组织的框架体系和概念集;②作为信息加工自动化的支撑,以其结构化的概念层级和语义关系,提供自动标引、自动分类、信息抽取、自动文摘、自动索引等功能;③作

为智能信息检索的后台控制；④作为知识挖掘、知识发现、学科发展分析预测的工具。

叙词表应用有很大的想象空间，但总体上我国叙词表应用还处于研究阶段，成熟的工具或产品比较少见。现实的情况是很多信息系统没有可用的叙词表来进行词汇控制，更谈不上利用叙词表丰富的语义信息。出现这种情况，主要原因是受限于词表数据难于获取和利用。因此，为基于叙词表的应用提供数据支持仍然是国家叙词库服务设计中首先要解决的问题。

范炜提出了包括三个层次的术语服务架构[14]：第一层是基于词表整体，提供各种词表的元数据描述和使用统计等相关描述；第二层是跨词表的基于术语层次的概念语义揭示与检索服务；第三层是在前两层的基础上，以概念、术语、语义层次的规范结构化机制对外向第三方应用提供服务。借鉴范炜提出的渐进服务思路，国家叙词库可以利用不同语义层次的内容建设成果，针对不同的叙词表应用和建设的需求，提供不同颗粒度的知识单元的数据服务，包括叙词表元数据服务、概念服务、概念集服务，以及词表定制服务。这些服务又可依据服务对象不同而表现为面向用户（人）的查询、浏览、编辑、下载服务，和面向机器的第三方调用服务。

2.3.1　叙词表元数据服务

叙词表元数据服务是基于叙词表注册信息的查询、浏览、下载服务。叙词表元数据服务可展示我国叙词表编制的总体情况，以及各部叙词表学科专业、编制和结构方面的特征，方便用户依据自身需求对叙词表进行选择。叙词表元数据服务是促进叙词表资源被发现和利用的最为基础的数据服务。

2.3.2　概念和概念集服务

概念和概念集是叙词表知识体系中不同粒度的知识单元。概念是叙词表知识体系最小的描述单元。概念集为多个概念按照一定的逻辑组合而成的知识单元，比如一个概念及其关联（上位、下位、相关）概念的组合，或一个词族中以等级结构关联的多层级概念的组合，或以映射方式相互关联的多表概念的组合，或共属同一个类目或本体类中的多个概念的组合等。概念服务还可依据所提供的概念描述信息的完整程度分成若干层次，比如，只提供词汇级别的中英对照，或词汇等同关系，或概念的等级、相关关系，或附加概念跨表映射关系等。基于概念或概念集的服务方式需依据叙词表应用中对知识单元的需求特征来设计和逐步实现。

2.3.3　词表定制服务

我国编制的叙词表数量有100余部，基本能覆盖所有的学科领域，但却不能满足所有的应用需求。其问题主要表现在以下方面：①很多叙词表比较陈旧，大部分叙词表完成于20世纪90年代，之后疏于维护，很多概念没有及时增补；②传统叙词表主要满足手工标引和检索的需要，人工语言痕迹较重，收录词汇与文献实际用词匹配度不高，很多较为专指的概念需要人工进行组配，难以满足计算机自动处理文献信息的需要；③传统叙词表一般覆盖一个或多个学科专业，学科广度和知识深度往往难以两全，很难满足专题信息系统文献信息组织揭示的需要。另外，叙词表作为一种形式化的知识体系，除了应用于文献信息组织，也可用于知识挖掘、知识发现等方面，但现有的叙词表很难完全支撑这些应用需求。因此，基于国家叙词库集成系统，提供词表定制服务，为新表构建提供数据和工具支撑是国家叙词库一个重要的应用方面。

基于国家叙词库的词表构建包括以下几个环节：①专题词汇集确定，②叙词表知识体系裁剪，③多表融合。数据层面，利用国家叙词库分类表确定词汇抽取范围，基础词库提供词表词汇，并根据用户需求补充文献关键词，以及其他来源的词汇。专题词汇集的确定还需要跟用户供词、专题文献抽词相结合。基础词库中不同来源的词汇属性信息可用于词汇评价，比如词汇出处可用于权威性判断，词汇在文献中的词频及学科分布情况可用于词汇重要性或专业相关性评价等。专题词表定制过程实质是多表知

识体系的裁剪与融合过程。国家叙词库概念层可为新表构建提供关系属性集合。来源于相同或不同词表的知识片段需按照一定的规则进行续接,由于词表知识体系存在重复、交叉、异构的特征,多表融合后必然存在概念语义模糊、关系缠结的现象,需要有相应规则和工具对这些问题进行处理。

3 叙词表知识体系的宏观结构及其描述

3.1 叙词表宏观结构

叙词表宏观结构指叙词表中多概念体系的组成结构以及该表不同版本之间的关联关系。叙词表一般包括主表、附表、范畴表、索引表。主表是叙词表主体部分,包含若干以词汇表达的概念及其完整的属性信息。附表一般为人名表、地名表、设备型号表、产品名称表等。范畴表/分类表是叙词表的概念/词汇分类体系,其基本构成单元为范畴/类目。主表中概念的范畴号/分类号即是对范畴表/分类表中的范畴/类目的引用。索引表是手工检索年代为方便查找叙词表概念而编制的从不同角度对概念进行定位的辅助表,类似于信息系统中的导航体系。索引表一般不会提供更多主表不包含的信息,因此,去除索引表通常情况下不会影响数据的完整性。

叙词表各子表中概念的属性特征不完全相同,比如主表中的概念和范畴表中的范畴一般具有不同的属性信息和描述需求。因此,按照概念原本的聚合特征,即叙词表的宏观结构来划分概念体系是比较简单和易于接受的处理方法。具体做法是:将叙词表及其各组成部分分别作为不同的概念体系,通过描述概念体系之间的关联关系来表达叙词表宏观结构(见图2)。叙词表作为母表,主表、附表、范畴表作为子表,母表与子表之间为包含(dct:hasPart)与被包含(dct:isPartOf)的关系。主表中的概念需引用范畴表中的范畴,主表和范畴表的关系为引用(dct:references)与被引用(dct:referencedBy)的关系。对于有多个版本的叙词表,各版本分别作为不同的"概念体系"。新旧版本之间为替代(dct:replaces)和被替代(dct:isReplacedBy)关系。

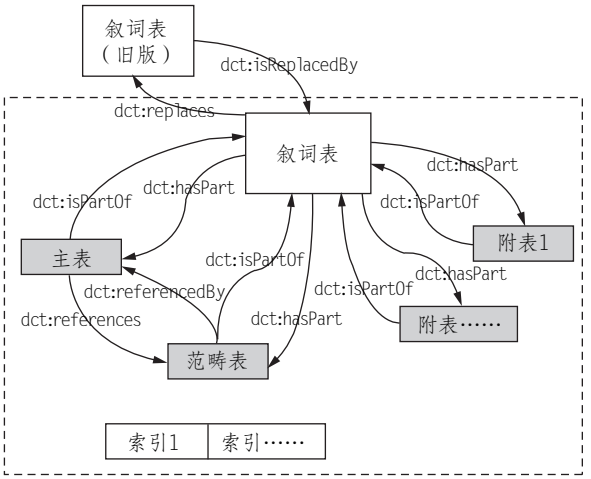

图2 叙词表宏观结构

3.2 叙词表元数据

叙词表元数据是叙词表"表"层面的描述框架。叙词表描述在很大程度上类似于文献的描述,可基于都柏林核心元数据框架[15]设计国家叙词库叙词表元数据方案(见表1)。

表 1　叙词表元数据方案

元　素					数据类型说明
表 ID			dc:identifier		URI
资源类型			rdf:type		skos:ConceptScheme（概念体系类型进一步细分为叙词表、主表、范畴表/分类表、附表等。）
表名称		正式名称	dc:title		
		其他名称	dct:alternative		
简介			dc:description		
专业			dc:subject		CLC（《中图法》类号）
语种			dc:language		
关联	外部关联	下一版本	dc:relation	dct:replaces	表 ID
		上一版本		dct:isReplacedBy	
	内部关联	母表		dct:isPartOf	
		子表		dct:hasPart	
		引用表		dct:references	
		被引用表		dct:isReferencedBy	
创建者		主要创建者	dc:creator		
		其他创建者	dct:contributor		
出版单位			dc:publisher		
版权		版权说明	dc:rights		
		版权单位	dct:rightsHolder		
出版时间			dct:issued		
统计		概念总量	totalConceptAmount		
		顶层概念量	topConceptAmount		
		非优选词量	altLableAmount		
		关系总量	relationAmount		
		层级	levels		
		……	……		

国家叙词库中的成员词表及其各组成部分均作为独立的概念体系进行描述，资源类型（rdf：type）为“概念体系”（skos：ConceptScheme），概念体系可依据需求再细分为叙词表、主表、范畴表、附表等。这些不同级别的概念体系均采用 URI 作为唯一标识。虽然主表、范畴表、附表可作为独立的概念体系进行描述，但一般情况下，叙词表各构成部分不具有独

立的“表名称”,具体著录时可继承叙词表的名称并加以扩展,比如“汉语主题词表—范畴表”。叙词表所属或所覆盖的学科领域或专业,用 dc:subject 描述,取值引用国家叙词库顶层分类表类目。

叙词表宏观结构通过概念体系之间的“关联”关系来描述。“关联”分为“内部关联”和“外部关联”。叙词表及其内部各组成部分之间的关联为内部关联,版本之间的关联为外部关联。内部关联包括 dct:hasPart(包含)、dct:isPartOf(被包含)、dct:references(引用)、dct:isReferencedBy(被引用);外部关联包括 dct:replaces(替代)、dct:isReplacedBy(被替代)。

叙词表统计元数据一般情况下可包括概念总量(totalConceptAmount)、顶层概念量(topConceptAmount)、非优选词量(altLableAmount)、关系总量(relationAmount)、层级(levels)。概念总量反映概念体系的规模;顶层概念量和层级可在一定程度上反映概念体系中词族构成情况和知识体系表现的深度;非优选词量和概念总量能间接反映概念体系的入口率,这是评价叙词表应用性能的主要指标之一;关系总量能反映概念体系中概念关联的疏密程度,也是评价叙词表应用性能的主要指标之一。以上统计可根据叙词表管理的需求进一步细化。

4 叙词表微观结构描述

叙词表微观结构是知识描述的框架结构。为了与语义网标准对接,叙词表微观结构的描述在早期较多采用 OWL(Web Ontology Language)进行格式层面的叙词表本体化转换。为了在语义网框架下更为简明地表示知识组织系统,2009 年,W3C 推出了 SKOS(Simple Knowledge Organization System)[16]。SKOS 是基于语义网规范对知识组织工具进行建模和描述的框架。我国在这方面已有较多的研究和探索[17-19]。为了更便捷地实现叙词表网络化,国家叙词库采用了基于 SKOS+XL 进行局部扩展的方案来实现叙词表微观结构的描述。

4.1 叙词表中的概念与词汇

传统叙词表以词汇作为描述对象,通过词间关系揭示概念之间的语义关联。自然语言词汇存在“一词多义”(多义词)和“多词一义”(同义词)的现象。为了解决多词一义问题,传统叙词表在多个同义词中选定一个作为优选词,其他词汇作为非优选词,并通过“等同关系”指引到优选词。为了区分多义词,传统叙词表一般在词后添加限定成分对词义进行说明,词汇与其后的限定成分共同构成一个新的“词”,从而消解词汇歧义问题。通过以上方式,叙词表实现了对词汇的控制。

词汇为词义和词形的复合体。同一词形可能表达多个词义,同一个词义可能对应多个词形。传统叙词表将“词汇”作为知识体系的构成单元来描述,必然要面对词形、词义纠缠不清的问题。对多义词不加区分,势必造成语义关系的混乱。而为了区分歧义词,又需要改变词形。添加限定语的方式虽然有助于明晰语义,但这时的“词汇”却失去了与文献信息用词和用户语言用词的天然关联。

SKOS 数据模型将词汇的词义和词形进行分离,词义部分即为“概念”,词形部分作为概念的“文字表达形式”,即“标签”,在概念之间构建等级关系和相关关系。每个概念都有自己的词汇标签,一个概念可以有多个标签,这样就解决多词一义的控制问题。一个标签(非优选标签)可以对应多个概念,但不影响概念语义关系的表达,由此可解决一词多义的问题。显然,分离了词汇的词形和词义,叙词表知识体系的描述逻辑更加清晰。

但 SKOS 还不能完全满足中文叙词表描述的需求,比如词汇汉语拼音的描述,同一概念的多个标签之间潜在关系的描述等。SKOS+XL 模型能更好地解决词汇层面的描述问题:将概念和词汇均作为描述对象,在概念层面主要描述与语义相关的属性,包括概念的定义、范畴、概

念间的关系(包括映射关系)；词汇层面主要描述与语义无关的属性，包括词汇的汉语拼音，词汇间的非语义对应关系，比如中文与英文译称的对应关系，全称与缩略语的对应关系，错误拼写与正确拼写的对应关系等；词汇作为概念的标签，将两个描述对象“概念”和“词汇”关联起来(见图 3)。

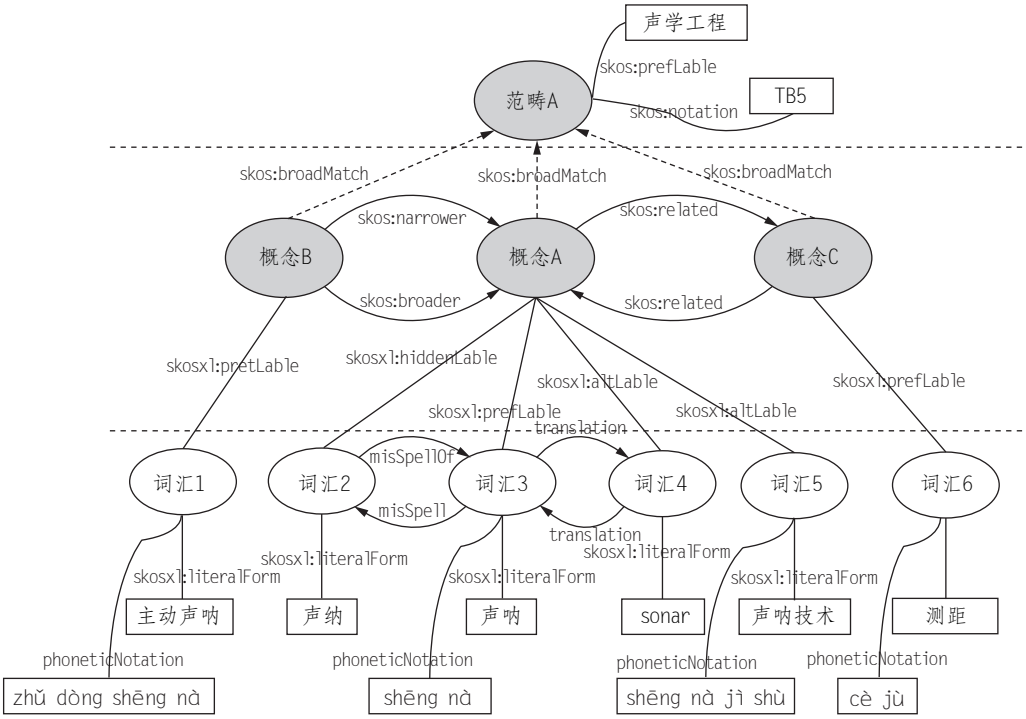

图 3　基于 SKOS+XL 的叙词表微观结构描述模型

图 3 对概念“声呐”进行了描述。“声呐”的词义部分抽象为“概念 A”，“概念 A”有四个词汇表达形式，包括三个中文词“声呐”“声纳”“声呐技术”，以及一个英文词“sonar”。在词汇层面，“声纳”是“声呐”的错误拼写，两个词间有对应关系；“sonar”是“声呐”的英文，两个词间也有对应关系；除了隐藏标签“声纳”外，其他词汇都有汉语拼音。在概念层面，“概念 A”的下位概念是“概念 B”(标签为“主动声呐”)，相关概念为概念 C(标签为“测距”)。

4.2　核心概念和扩展概念

我国编制的叙词表中，有约 10%采用组配方式表达概念，即多个概念按一定的逻辑组合表达成一个更为专指的概念。概念组配一般可分为概念交叉组配和概念限定组配。示例 1 和示例 2 为概念交叉组配。示例 1 中“天平”是各类天平的集合，“电子装置”是各类电子装置的集合，“电子天平”则为前述两个集合的交集。概念交叉组配生成的概念是参与组配的多个概念的下位概念。即“电子天平”既是“天平”的下位概念，又是“电子装置”的下位概念。

示例 1

Dianzi tianping

** 电子天平　BFO.FFO.X

** Electronic balances

→天平　BFO.FFO

+ 电子装置　KY

注:摘自《标准文献主题词表》(1988 年)

示例 2

合成润滑剂

synthetic lubricants

Y 合成材料

+ 润滑剂

注:摘自《核科学技术叙词表》2006 年

示例 3 和示例 4 都为概念限定组配。示例 3 中“聚合物溶液粘度”由“聚合物溶液”和“液体粘度”组配表达。“聚合物溶液粘度”是“液体粘度”的下位概念,但与“聚合物溶液”不构成上下位关系。“聚合物溶液”对“液体粘度”起限定作用。

示例 3

jù hé wù róng yè nián dù

聚合物溶液粘度

viscosity of polymer solutions

[6620]

Y 聚合物溶液

液体粘度

注:摘自《物理学汉语主题词表》(1994 年)

示例 4

道教艺术

CS36

Y 道教 + 宗教艺术

注:摘自《社会科学检索词表》(1996 年)

在叙词表中,还有一类概念是通过参见的方式指引到其他更为具体的概念,这类概念往往比较宽泛,见示例 5。“化学药品”为泛指概念,直接用于检索文献价值不大,通过参见的方式将其指引到更为具体的概念,“螯合剂”“去垢剂”“染料”……

示例 5

化学药品

chemicals

(见专指化合物或化合物的各类,例如“CARCINOGENS”“DETERGENTS”“PLASTICIZERS”和“ORGANIC COMPOUNDS”。)

J 螯合剂

H 去垢剂

H 染料

H 石化产品

H 添加剂

H 无机化合物

H 显影剂

H 有机化合物

H 指示剂

注:摘自《核科学技术叙词表》(2006 年)

传统叙词表将组配生成的专指概念,以及需要指引到其他具体概念的过于宽泛的概念都处理成“非优选词”,因为传统叙词表中词汇和概念是一体的,“非优选词”可理解为“非优选概念”,这样的处理无可厚非。需要指出的是,此处“一对多”的“非优选词”实质上是叙词表中一类特殊的、不欲正式收录的概念。

在 SKOS 数据模型中,“非优选词”是概念的“标签”,是对应概念的一种文字表达形式。显然,传统叙词表中的这类“非优选词”并不能作为其指向概念的标签。为解决这一矛盾,“国家叙词库”提出了“核心概念”和“扩展概念”用以区分叙词表中的优选词和上述“一对多”的非优选词所对应的概念。传统叙词表中组配生成的专指概念和需指向具体概念的过于泛指的这类“非优选词”都为“扩展概念”,而优选词对应的概念则为“核心概念”。

扩展概念同样有自己的词汇标签,以及词汇层面的描述属性,但扩展概念之间没有关联关系。扩展概念与核心概念之间“一对多”关系的形成有多种情况,在具体处理时需要区分不同情况将关系类型转换为等级或相关关系,比如将概念交叉组配转换成扩展概念与核心概念间的上下位(等级)关系。

5 概念与词汇描述的元数据方案

5.1 概念描述的元数据方案

国家叙词库的“概念”包括叙词表主表和附表中的概念,也包括范畴表/分类表中的范畴/

类目。主表和附表中的概念又可分为核心概念和扩展概念。不同类型的概念，其属性空间不完全等同，但都可在 SKOS 框架下进行描述。概念描述的核心元数据方案如表 2 所示。

表 2　概念描述的核心元数据方案

元素				说明
概念 ID		dc：Identifier		URI
资源类型		rdf：type		skos：Concept，可进一步扩展为普通概念和范畴/类目。普通概念可进一步扩展为核心概念和扩展概念。
来源		skos：inScheme		取值表 ID。所归属的概念体系。
代码		skos：notation		范畴/类目的代码，即范畴的范畴号，或类目的分类号，只用于范畴/类目型概念的描述。
标签	优选词	skosxl：prefLable		取值词汇 ID
	非优选词	skosxl：altLable		取值词汇 ID。英文译称作为概念的非优选词。
	隐藏标签	skosxl：hiddenLable		取值词汇 ID
注释	定义	skos：note	skos：definition	
	范围注释		skos：scopeNote	
语义关系	属	skos：semanticRelation	skos：broader	
	分		skos：narrower	
	参		skos：related	
映射	约等映射	skos：mappingRelation	skos：closeMatch	取值概念 ID 或范畴 ID。概念所归属的范畴也用映射关系描述。
	等同映射		skos：exactMatch	
	上位映射		skos：broadMatch	
	下位映射		skos：narrowMatch	
	相关映射		skos：relatedMatch	

概念是网络环境中一种独立的可溯源的资源，采用 URI 作为唯一标识。概念所属的概念体系用 skos：inScheme 表示。概念的文字表达，即概念的标签通过三组属性组合来实现，概念优选标签（优选词）用属性组合 skosxl：prefLable，skosxl：literaForm，概念非优选标签（非优选词）用属性组合 skosxl：altLable，skosxl：literaForm，概念隐藏标签用属性组合 skosxl：hiddenLabel，skosxl：literalForm。隐藏标签一般用于词汇的错误拼写，现存的叙词表极少出现将词汇的错误拼写作为非优选词的情况，因为错误拼写一般情况下是不允许出现在出版物中的。但从实际使用的角度，对于一些“常用”的错误拼写，适当收入叙词表中有利于文献查全。

概念与概念之间的链接关系分为两种情况。①同一概念体系中概念的链接：概念可通过语义关系 skos：semanticRelation 及其子属性“属”skos：broader、“分”skos：narrower、“参”skos：related 来实现与其他概念的链接。②跨体系的概念链接：可通过映射 skos：mappingRelation 及其五个子属性来链接跨体系的概念。跨体系包括同一部叙词表的不同概念体系之间，如主表

与范畴表之间,以及不同叙词表之间,比如国家叙词库成员表与《汉表》之间,以及成员表与顶层本体之间等多种情况。

属性 skos:notation 用于范畴/类目型概念的描述,表示范畴的范畴号,或者类目的分类号。一个普通概念所归属的范畴需要通过概念与范畴的映射关系,比如 skos:broadMatch 来表示。

5.2 词汇描述的元数据方案

词汇是概念的标签(skosxl:lable),在国家叙词库中,词汇本身也作为一种资源实体进行扩展描述。SKOS-XL 数据模型提供了 skosxl:literalForm、skosxl:labelRelation 两种属性来扩展对词汇的描述。skosxl:literalForm 用于词汇的文字表达,每个词汇有且仅有一个文字表达,即词形。skosxl:labelRelation 用来表达词汇间的关系。传统叙词表中的词间关系指词汇间的语义关系,此处的词间关系指词汇间的非语义关系(参见 4.1)。SKOS-XL 数据模型只提供 skosxl:labelRelation 来表示词间关系,需依据具体描述需求对该属性进行扩展,比如词汇的正确拼写与错误拼写之间的对应关系(misSpellOf, misSpell),全称与简称的对应关系(abbreviationOf, abbreviation)等。由于 SKOS+XL 模型中没有相应的属性用于描述词汇的注音,需要在 SKOS 框架下对 skos:note 进行扩展,定义其子属性 phoneticNotation 用于词汇汉语拼音的描述。词汇描述元数据方案见表 3。

表 3 词汇描述的核心元数据方案

元素				说明
词 ID		dc:Identifier		URI
资源类型		rdf:type		skosxl:lable
词形		skosxl:literalForm		有且仅有一个
注音		phoneticNotation		词汇的汉语拼音
关系	正确拼写	skosxl:labelRelation	misSpellOf	词汇间的非语义关系,取值词汇 ID。
	错误拼写		misSpell	
	全称		abbreviationOf	
	简称		abbreviation	
	译称		translation	

6 结语

叙词表知识体系除了应用于文献的主题标引和智能检索,也可作为知识挖掘、知识服务的语义工具。我国有着丰富的叙词表资源,让这些叙词表能够被发现和利用,是“国家叙词库”的建设目标。

国家叙词库是一个叙词表集成系统。我国的叙词表大部分遵循国家标准编制,在叙词表结构和描述上有共同之处,但部分词表也因为学科和知识体系表达的特殊需要而具有一些不同的特征。如果对所有词表都完整描述,势必要对元数据进行扩展,但自定义的元数据项太多,对数据的理解和交换是不利的。因此,国家叙词库的知识描述需要在共性和特性间进行平衡,在尽量不损失原始信息的情况下,保证元数据的通用性。

众多词表的语义集成将是一项巨大的工程。之前已有一些大型的词表集成项目取得成功,但就目前看,集成成果的应用情况并不是很理想。所以,国家叙词库须采用渐进建设

的思路，并利用不同时期的建设成果提供不同语义层次的服务。首先是词表数据的统一存储和描述，使叙词表数据可发现、可获取、可利用。之后根据具体的应用需求，比如在词表定制服务的过程中逐步实现叙词表概念的语义集成。

另外，国家叙词库本质上是一个资源共建共享问题，我国100多部叙词表分属不同单位，有些词表版权不清晰，或者找不到维护单位，词表资源的共享模式也无太多经验可借鉴，因此，探索我国叙词表资源的开放共享和持续发展机制是国家叙词库发展需要面对的另一个重要问题。

参考文献

[1] 傅兰生. 我国叙词表兼容两大方案的分析——兼论国家级叙词兼容词库的建立[J]. 情报学报, 1991, 10(4): 257-264.(Fu Lansheng. Analysis of two schemes of Chinese thesauri compatibility and discussion about establishment of national compatible thesauri warehouse[J]. Journal of the China Society for Scientific and Technical Information, 1991, 10(4): 257-264.)

[2] 朱岩. 对建立国家叙词库的认识和思考[J]. 科技情报工作, 1991(2): 15-17.(Zhu Yan. Cognition and consideration about national thesauri warehouse[J]. Scientific and Technical Information Service, 1991(2): 15-17.)

[3] High-Level Thesaurus [EB/OL]. [2015-03-10]. http://www.webarchive.org.uk/wayback/archive/20110602225735/http://hilt.cdlr.strath.ac.uk.

[4] Nicholson D, Joseph A, McCulloch E. HILT IV: research report[EB/OL]. [2015-03-10]. http://www.webarchive.org.uk/wayback/archive/20110602225755/http://hilt.cdlr.strath.ac.uk/hilt4/documents/ResearchReport-May2009.pdf.

[5] Vizine-Goetz D, Hickey C, Houghton A, et al. Vocabulary mapping for terminology services[EB/OL]. [2015-03-10]. https://journals.tdl.org/jodi/index.php/jodi/article/view/114/113.

[6] 司莉, 徐丽晓, 吴钢, 等. OCLC术语服务研究:背景、进展与启示[J]. 中国图书馆学报, 2007, 33(1): 58-61.(Si Li, Xu Lixiao, Wu Gang, et al. OCLC terminologies service: background, development and inspirations[J]. Journal of Library Science in China, 2007, 33(1): 58-61.)

[7] Mayr D, Petras V. Building a terminology network for search: the KoMoHe project[EB/OL]. [2015-03-20]. http://arxiv.org/ftp/arxiv/papers/0808/0808.0518.pdf.

[8] Mayr P, Petras V. 交叉语词索引:术语映射及其对信息检索的影响[EB/OL]. 刘华梅,译. [2015-03-20]. http://www.nlc.gov.cn/newtsgj/iflaygt/gjtlzwyyzx/tlzlby/201011/P020101130497872306658.pdf.(Mayr P, Petras V. Building a terminology network for search: the KoMoHe project[EB/OL]. Liu Huamei, trans. [2015-03-20]. http://www.nlc.gov.cn/newtsgj/iflaygt/gjtlzwyyzx/tlzlby/201011/P020101130497872306658.pdf.)

[9] U.S. National Library of Medicine. Unified medical language system[EB/OL]. [2015-03-25]. https://www.nlm.nih.gov/research/umls.

[10] 李丹亚,胡铁军,李亚子,等. UMLS多词表整合机制研究[J]. 数字图书馆论坛,2012(4):28-36.(Li Danya, Hu Tiejun, Li Yazi, et al. Integration mechanism of multi-vocabulary in UMLS[J]. Digita Library Forum, 2012(4):28-36.)

[11] 中国科学技术信息研究所. 汉语主题词表(工程技术卷)[M]. 北京:科学技术文献出版社,2014.(Institute of Scientific and Technical Information of China. Chinese thesaurus (engineering and technology)[M]. Beijing: Scientific and Technical Documentation Press, 2014.)

[12] 吴雯娜, 王星. 叙词表融合方法研究[J]. 中国图书馆学报, 2012(4): 67-75.(Wu Wenna, Wang Xing. A method for thesauri merging[J]. Journal of Library Science in China, 2012(4): 67-75.)

[13] 薛春香. 网络环境中知识组织系统构建与应用研究[M]. 南京: 东南大学出版社, 2009.(Xue Chunxiang. Research on the construction and application of networked knowledge organization systems[M]. Nanjing: Southeast University Press, 2009.)

[14] 范炜. 受控词表的术语服务研究[J]. 图书情报工作, 2012,56(14): 34-39,97.(Fan Wei. Terminology service for controlled vocabularies[J]. Library and Information Service, 2012, 56(14): 34-39, 97.)

[15] Dublin Core Metadata Initiative. DCMI metadata terms[EB/OL]. [2014-12-17]. http://www.dublincore.org/documents/2008/01/14/dcmi-terms.

[16] World Wide Web Consortium. SKOS Simple Knowledge Organization System[EB/OL]. [2014-12-17]. http://www.w3.org/2004/02/skos.

[17] 刘丽斌, 张寿华, 濮德敏,等.《中国分类主题词表》的 SKOS 描述自动转换研究[J]. 中国图书馆学报, 2009(6): 56-60.(Liu Libin, Zhang Shouhua, Pu Demin, et al. Automatic transformation of *Classified Chinese Thesaurus* (CCT) description with SKOS[J]. Journal of Library Science in China, 2009(6): 56-60.)

[18] 段荣婷. 基于简约知识组织系统的主题词表语义网络化研究——以《中国档案主题词表》为例[J]. 中国图书馆学报, 2011, 37(3): 54-65.(Duan Rongting. The semantic networking of thesaurus based on the Simple Knowledge Organization System:a case study of the *Chinese Archival Thesaurus*[J]. Journal of Library Science in China, 2011, 37(3): 54-65.)

[19] 曾新红. 中文知识组织系统形式化语义描述标准体系研究(一)——扩展 SKOS 实现传统受控词表全描述[J]. 中国图书馆学报, 2012(3): 57-68.(Zeng Xinhong. Formal semantic description standard family for Chinese KOS (I): realizing full description for Chinese controlled vocabularies by expanding SKOS[J]. Journal of Library Science in China, 2012(3): 57-68.)

吴雯娜 中国科学技术信息研究所副研究馆员。北京 100038。

鲍秀林 中国科学技术信息研究所馆员。北京 100038。

(收稿日期:2014-12-18;修回日期:2016-01-06)