

国家科技图书文献中心科技词表的构建与应用

梁 冰 白海燕 王 莉 乔晓东
(中国科学技术信息研究所, 北京 100038)

摘要: 本文主要针对国家科技图书文献中心(NSTL)科技词表的构建特点及智能化文献服务应用方法开展研究。采用实例剖析的方法, 分析词表通用构建方法与常规应用模式, 在此基础上重点研究NSTL科技词表的构建特点和应用方式。NSTL文献服务平台从生产到服务的各个业务流中引入NSTL科技词表, 提升了国家科技文献信息战略资源的有效组织、深度揭示和知识关联, 提供智能检索服务。在新的互联网应用环境下, 词表规范不再是一种约束, 而是一种引导, 一种知识关系汇聚方式, 融入科技文献服务的方方面面, 产生新的应用方式。

关键词: 知识组织系统; NSTL科技词表; 词表应用; 词表构建; 科技文献服务; 文献检索

中图分类号: G250.76

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2016.02.001

Construction and Application of NSTL Scientific & Technological Knowledge Organization Systems

LIANG Bing, BAI Haiyan, WANG Li, QIAO Xiaodong

(Institute of Scientific and Technical of Information of China, Beijing 100038)

Abstract: Study the construction features of NSTL SKOS and develop intelligent document applications based on NSTL SKOS. Analyze usual construction method and application model of knowledge organization system using method of case analysis. Emphasize on studying of NSTL SKOS construction and application method. NSTL document service plat introduce STKOS to promote information organization effectively, describe deeply, knowledge linkage and intelligent document retrieval. Under network environment, KOS is not only a kind of specification, but also guidance, knowledge relationship aggregation model. NSTL STKOS has been integrated into all aspects of science and technology document service to create new applications.

Keywords: Knowledge Organization Systems (KOS), NSTL Sci-tech Knowledge Organization Systems (NSTL STKOS), vocabulary application, vocabulary construction, sci-tech document services, literature retrieval

1 引言

科技信息资源数量的激增及信息服务向知识

服务的转变, 要求我们尽快提升海量信息的自动处理能力, 实现从文献提供向科技对象实体的关联发现、科技热点监测、科技趋势分析等服务的

作者简介: 梁冰*(1974—), 男, 中国科学技术信息研究所高级工程师, 博士, 主要研究方向: 科技文献网络信息服务、数字图书馆技术研究等; 白海燕(1973—), 女, 中国科学技术信息研究所信息技术支持中心研究馆员, 中国科技情报学会信息技术专业委员会秘书, 主要研究方向: 信息组织、数字图书馆、关联数据、知识组织系统等; 王莉(1974—), 女, 中国科学技术信息研究所研究员, 主要研究方向: 科技文献服务体系研究、科技信息知识组织研究、文献标准制定等; 乔晓东(1965—), 男, 中国科学技术信息研究所研究员, 主要研究方向: 科技信息服务、科技政策研究、数字图书馆建设研究等。

基金项目: 国家“十二五”科技支撑计划项目“信息资源自动处理、智能检索与STKOS应用服务集成”(2011BAH10B05)。

收稿时间: 2015年4月29日。

转变,实现从传统检索向知识导航、自动主题聚类、内容相关揭示、自然语言检索和双语查询等的转变。但由于我国长期以来缺乏大规模综合性科技知识组织工具,难以满足外文文献的知识发现、知识关联和知识计算的要求^[1]。建设和应用综合性、集成化、工具化的科技超级词表,将更好地开发利用科技文献资源,加强和巩固国家科技文献信息战略保障水平,提升科技文献服务能力,实现从文献服务向知识服务的跨跃式发展提供基础性支撑。

从国际上看,随着21世纪互联网的飞速发展,一些词表开始从单机电子版向网络环境移植,并逐渐被集成到信息检索系统中,如ERIC/HASSET/AAT/UMLS等^[2]。词表的发展呈现出网络化、编制模式半自动化与协作化、集成化与本体化、信息组织标准全面升级、专业领域词表发布为关联数据等特点^[3]。各类基于词表的应用急剧增多,对网络化,知识化的词表研究逐渐形成热点,重点集中在词汇自动丰富、不同词表间的互操作、基于叙词表构建本体以及通过词表形成领域知识化服务等方面。此外,互联网服务商开始建设类似百度词条、维基百科等以用户互动为基础的知識组织体系,以全新的模式对传统词表的建设和应用提出了挑战。

我国的词表建设也经历了相同阶段,代表性的成果有《中国分类主题词表》、《中国图书馆分类法(R类)与医学主题词表(MeSH)对应表》建设映射系统,并在中文医学文献主题标引中应用。北京大学、清华大学等单位研制了语义知识库、标注语料库,以实现计算机对文本内容的理解。一些学者也对传统知识组织体系在语义网环境下的发展和应用进行探索,如:用OWL表示《中国分类主题词表》,叙词表的RDF表示方法等。总体来看,我国的词表建设和应用规模较小,由于缺少大规模、综合性科技词表,尚无法支持我国科技文献信息机构进行知识组织和知识服务。

NSTL超级英文科技词表(STKOS: Scientific & Technological Knowledge Organization System, 以

下简称“NSTL科技词表”)由国家“十二五”科技支撑计划立项支持,借鉴国内外已有的知识组织系统建设成果与应用经验,结合当前互联网应用服务中的新特点,构建而成的面向计算机应用的超级英文科技词表,对于实现国家科技文献信息战略资源的有效组织、深度揭示和知识关联,文献智能检索服务提供基础性支撑,具有重要作用。

本文首先梳理和归纳了当前词表构建的技术模式与应用,然后介绍了NSTL词表的构建,并重点介绍了NSTL词表在智能检索系统中的应用方法,最后分析和介绍了NSTL科技词表的构建特点。本文的目的在于通过NSTL科技词表构建技术的分析研究与实际应用研究,提升和推广综合性、集成化科技知识组织系统在文献服务知识化转型中的应用。

2 词表构建模式及其应用

2.1 词表构建模式

(1) 词条的自动发现与更新

传统词表的词条发现和更新维护主要依靠人工实现。从科技领域词汇的发展变化来看,其更新速度远远跟不上实际领域发展的速度。因此,文献服务行业研究重点逐渐从词表的数字化加工处理转向词汇的自动发现与补充,不断丰富和更新词表的建设模式。不断从科技领域学术研究的自由文本中抽取词汇,挖掘词汇间关系,而这些关系的准确性与采用的挖掘算法密切相关,虽然无法比拟人工编制词表的严谨与精确,但是可以作为其补充,促进词表的快速更新^[4]。

(2) 词表间的互操作与集成化

不同的知识组织工具,如分类词表、主题词表、术语表等,具有不同的自然语言控制能力和语义化程度,通过词表互操作与集成,旨在发挥和集成各自的功能特点与作用。词表互操作包括两种方式:一是以统一标准进行数据整合,在其他系统中导入和使用;二是在两个不同词表之间建立映射关系,通过映射将不同类型、不同语种的词表集成整合到一起。在标准化方面,研究

热点集中在如何采用知识组织规范对叙词表进行描述。目前，通过词表间映射是解决互操作的最重要的方法。美国国立医学图书馆建设的UMLS (Unified Medical Language System) 就是非常典型且成功的案例，它不断整合国家医学系统的传统分类法、叙词表、标题表、术语词典等资源，通过映射、链接等对概念及概念关系进行自动抽取、集成归并，建立不同词表概念间的映射关系，形成了一部概念和语词覆盖面广、语义关系丰富的超大规模词表，广泛应用于医疗信息系统、病案系统、自然语言处理、文本自动标注、智能检索及搜索引擎领域^[5]。

(3) 词表的本体化

本体是以建立概念和概念之间关系为核心，提供对特定领域知识的描述，其构建较为复杂，并且在不同领域很难复用。叙词表由于其丰富的概念、词汇及词间的语义结构，为本体的构建提供基础条件，且已有较为成功的应用案例，如：联合国粮农组织将Agrovoc叙词表转换为农业本体；UMLS项目中基于传统知识组织体系关联扩展建设语义网络，建立一个语义网络知识源数据库^[6]。

(4) 词表建设的标准化与规范化

国外主要发达国家的标准化机构纷纷开展了知识组织体系统一的计算机化表示规范、集成映射标准和网络化服务技术接口的相关研究。值得重点关注的知识组织体系建设标准有：英国BS 8723 系列标准、美国国家标准 (Z39.19)、英国国家标准 (BS 8723)、IFLA多语种叙词表编制指南、W3C组织的SKOS草案等。

(5) 词表的Web服务和工具化

越来越多的词表在改造、扩建和集成完成后，其最终形式是工具化的Web服务，提供支持性软件工具与开放接口，供第三方用户下载、定制、集成、映射和调用。例如：UMLS提供了能够查询和获得UMLS的概念、语义类型、语义关系等的术语服务 (UMLS Terminology Services, UTS)，用于UMLS本地安装和剪裁定制的工具MetamorphoSys，UMLS概念的文本映射工具MetaMap等^[5]。

2.2 词表应用

(1) 信息组织

信息组织包括利用词表进行信息的深度提示、内容整合和语义关联。例如：联合国粮农组织发起的FOS项目，旨在通过创建、整合和利用本体，来加强渔业信息系统的信息整合和语义互操作能力。其中的OneFish是一个渔业项目的门户，采用等级主题树方式对信息进行组织，大约有超过1800个主题，主题含有简短的摘要、标识符、相关联的材料，如文件、网站、元数据。

(2) 自动标引

美国国立医学图书馆基于UMLS和MeSH开发的医学文本标引工具Medical Text Indexer (MIT)，应用于Medline/Pubmed的文献标引，经MIT自动标引系统处理后给出推荐MeSH主题词和副主题组配供标引员进行人工审核，这种人机相结合的标引工作模式极大提高了数据处理速度，减轻了标引员的工作负荷。

(3) 智能检索

基于MeSH、UMLS超级叙词表中同义关系、等级关系等简单语义的关系，PubMed实现了检索的主题扩展、副主题扩展、副主题限定、提问词自动转换、提问短语分析转换等智能化处理。Semantic Medline对PUBMED检索命中所返回的MEDLINE数据进行自动文摘和可视化展现。

(4) 资源整合

STERNA (Semantic Web-based Thematic European Reference Network Application) 项目与12个欧洲自然史和生物多样性机构合作，以鸟类和各种鸟类相关的信息为核心，将各种关于鸟、鸟类物种等多媒体资源，包括科学数据、文章、图片、视频和音频文件，汇集和显示在同一个信息空间。受控词表包括：术语表、分类表、叙词表、本体等，在STERNA系统中被称为“参考结构”，采用SKOS进行描述。

(5) 应用整合

MedWeaver系统以UMLS为基础，集成Dx-plain (诊断决策支持系统)、WebMedline (斯坦福大学开发的用于检索MEDLINE数据库医学文

献检索系统)和CliniWeb(WWW医学信息检索系统)3个系统的资源,并且将检索结果按内容进行有机组织后反馈给用户。

(6) 术语研究

据统计,UMLS最主要的应用方向包括术语研究、术语映射和创建本地术语,约占有应用的53%。Q.Li等开发了基于Protege 2000的一系列工具,可以用于转换形式化表达的或非形式化表达的知识,将从外部获取的知识进行整合集成,消除术语冗余和概念间的冲突,已实现了将远程的UMLS术语合并到Protege 2000知识库中。

3 NSTL科技词表构建及特点

3.1 体系框架

NSTL科技词表由来源词表仓库、基础词库、规范概念集和范畴表组成,总体框架如图1所示。其中,素材层是NSTL科技词表的来源,但并不是NSTL科技词表的组成部分。基础词库包含国际上重要知识组织体系中的术语,主要涉

及叙词表、术语表等富含语义关系的术语或概念,规模为500万条;规范概念集包含概念形成过程中涉及的词型规范、意义规范的同义词和准同义词,以及规范概念间的共现关系,收集科技概念规范名称80万条,覆盖理、工、农、医四大领域;范畴体系按照使用需求修改或重新编制的概念体系结构,主要用于规范概念的分类归并以及文献信息的宏观分类导航^[7-8]。

3.2 构建特点

(1) 素材采集、评价与遴选,强调广泛性与专业性

超级科技词表的素材包括两部分来源:一是各种词表、术语表;二是来自科技文献中作者的关键词和科技用户检索词。按照理、工、农、医领域采集国外叙词表、分类法、术语表等,并对其对比分析和评价。根据遴选标准从来源词表中遴选出骨干词表。

(2) 基础词库建设注重规范化和标准化

对已收集素材中的海量科技术语进行形式化

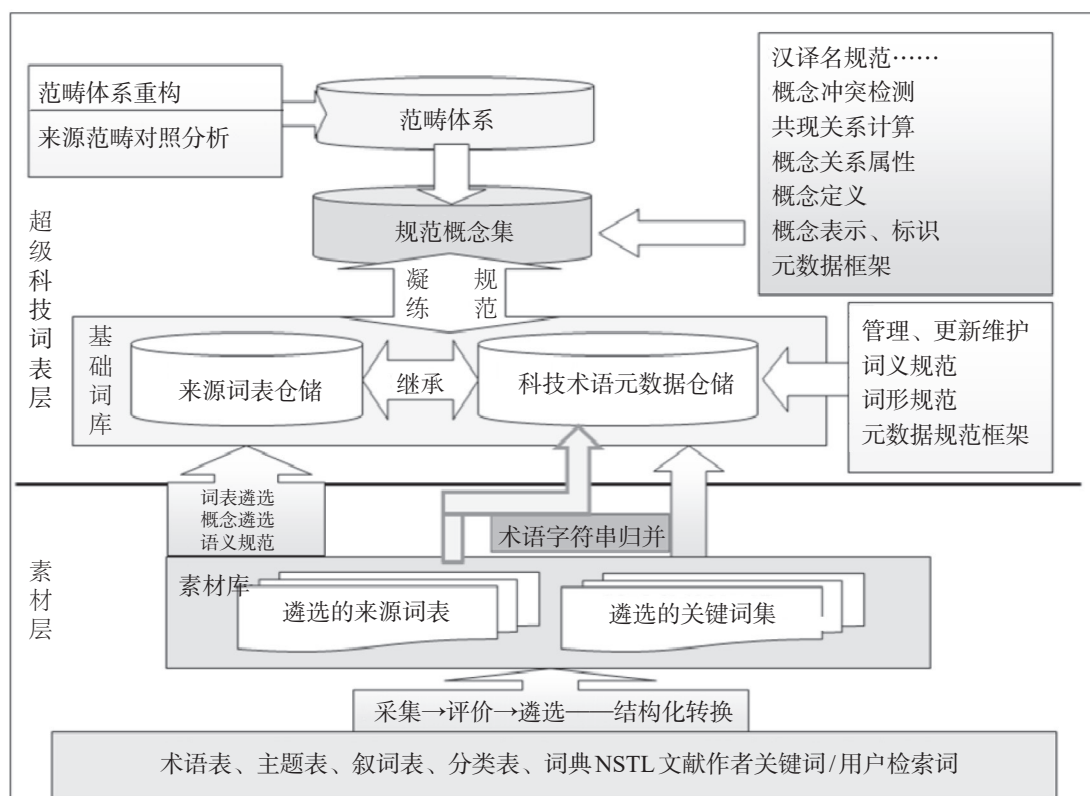


图1 NSTL词表的体系框架

汇总、整理、规范、去重、分类等处理，按照统一的基础词库元数据结构标准，建设基础词库。其中主要解决不同词表中异构数据的同构化表示（包括术语表达、属性的继承等），处理各词表间同形异义、异形同义和同形近义等问题。

（3）以概念为核心的规范概念集建设

NSTL词表的核心是概念，通过对基础词库中科技术语进行词形规范、词义规范，并按照概念遴选标准遴选概念，确定概念的规范名称。以概念为核心，以继承来源骨干叙词表的原有关系为基础，通过概念与原有叙词表来源术语的语义关系，以及概念间的共现关系进行关联，形成以概念为核心的概念集合。通过概念定义、概念范畴、概念间的关系、概念的规范汉译名以及概念被使用的信息等多种方式加以完善和扩充。

（4）范畴体系建设体现语义映射

范畴是概念的重要属性，用来说明概念所适用的学科或所归属的类。对于文献信息的主题聚类、分类组织及浏览具有重要意义。同时，范畴体系的建设也是科技文献信息通用本体建设的基础，有利于控制通用本体的维度和颗粒度，便于建立通用本体与超级科技词表概念的映射关系，有利于解决因学科交叉、表达产生的维（粒）度不同、冲突和重叠等方面问题。

4 NSTL科技词表在文献服务平台的应用

NSTL文献服务平台在其从生产到服务的各个业务流中引入NSTL科技词表，其目标在于实现国家科技文献信息战略资源的有效组织、深度揭示和知识关联，提供智能检索服务。

4.1 NSTL科技词表的应用

NSTL科技词表在文献服务中的应用主要体现在以下几个方面：一是能够支持对科技文献信息在概念层面的自动标引，一定程度上支持自然语言的检索，支持一定意义上的双语检索，支持基于概念层面的智能检索，包括扩检和缩检，支持基于关联词表的主题聚类等。二是通过建立受控优选词与自由词之间的映射，实现信息资源的自动标引；对用户检索而言，用于检索词的

扩展，发挥引导作用，而不是早期的检索约束。

三是NSTL科技词表中的范畴类目具有较强的结构性和丰富的语义关系，可作为科技领域浏览的知识框架，提供科技信息的分类导航^[9]。可见，NSTL科技词表主要解决了文献的概念化揭示、知识化组织和基于知识内容的内在关联构建，从而实现具有一定程度的智能化服务。

（1）自动标引

自动标引包括概念标引、学科范畴标引以及关键词标引三部分^[10]。其中，概念标引以STKOS基础术语为标注词典，对文献标题、摘要等文本信息进行处理，综合考虑位置、词长等因素改进TF-IDF算法，制定权重阈值，控制术语的标注数量，降低标引结果中的噪音。学科范畴标引同样依据STKOS词表，标引概念对应的范畴，并制定范畴合并与遴选的规则，标引范畴号和范畴类目。关键词标引主要针对原始科技文献元数据中关键词缺失问题，从未登录词和最终标引的概念中进行遴选，在标引的同时完成关键词与词典规范术语的映射。

（2）扩展检索

扩展检索是词表在搜索引擎中非常典型的应用。考虑到用户在检索过程中面临的问题往往是信息量过大，在实际应用中智能检索系统中采用了扩展提示方式，而不是由系统自动对检索词进行扩展检索。检索扩展提示包括输入词、同义词、语义关系3个层面。首先，在用户输入关键词时，系统采用简单字符匹配方式查找由历史检索词、收录文献关键词、STKOS词表组成的词汇集合，实时给出输入提示，提升界面友好度。同义词提示和语义关系提示分两个区域在检索结果列表页面展示，扩展是由用户的点击行为触发。其中，同义词提示是利用STKOS词表中概念和基础术语的映射关系实现的，只出现存在等同和替代关系的词汇，即词形不同词义相同；语义关系则与之相反，提示的是那些同形异义的词汇，并且以树形方式给出提示词的范畴及邻近的上位词，支持用户点击树形结构查看更详细的语义关系。

(3) 引导式检索

集成STKOS词表数据,允许用户在学科领域中直接浏览词表内容,选定规范概念/术语,用选定词汇进行检索,有效提高检索效率。也可在用户检索过程中,通过同义词和反义词、规范概念和概念范畴等多角度提示用户进行引导式检索。

(4) 关系发现

关系发现是一种交互性检索应用。以文献元数据和STKOS词表数据为基础构建知识关联网络,采用可视化技术,以检索结果集为对象,展现文献资源、学科领域、人物、科研活动四大类型实体及实体之间的关系,支持渐进式探索。关系发现应用中可包含3个功能:一是以检索得到的文献集合为基础,采用图的形式展示不同节点之间的关系,支持用户点击任意节点进行浏览,可以跟随节点的链接扩展到更大的节点空间,突破检索结果集合的限制;二是专注于一篇文献,观察它在整个节点空间中与其他节点之间存在的关系,支持渐进式的探索;三是在一次或多次检索的结果列表中收集关注的实例(如作者、论文、机构等信息),通过实例分析功能在整个节点空间中观察它们之间的关联关系^[9]。

(5) 跨语言检索

STKOS词表中的条目拥有中英文两种表述,利用这一对照关系部分实现跨语言检索功能。

4.2 问题与局限

从实际应用来看,NSTL科技词表作为综合性词表,一是体现在多科学技术领域的集成,二是体现在主题词表、分类词表、叙词表、词典等多种知识组织工具的融合。因此,NSTL科技词表应用于海量、全学科科技文献资源的组织与检索,在提高序化程度和组织深度方面,显示了突出的优势,对于交叉学科,也有很好的适应性。但对于具体学科,与专业词表相比,其深度和专业程度还有所欠缺,有待通过底层词表深度映射等技术方法加以改进。

5 结语

信息服务日益发达的今天,用户对科技信

息需求日益多样化,这迫使科技信息服务系统建设者不断寻找新的建设方法和模式,词表被日益隐藏化、后台化。但无论如何演变,词表作为一种有效的知识规范工具并没有被放弃,它对从主题领域深度分析、同义词聚合到概念关系识别等依然具有重要作用,其发展呈现出数据格式多元化、多类表集成以及从术语模式向引导概念模式的本体化转变等发展趋势,其基本功能及形态也不断发生着深刻变化。在新的互联网应用环境下,词表规范不再是一种约束,而是一种引导,一种知识关系汇聚方式,将融入科技文献服务的方方面面,从而焕发新的生机。

NSTL科技词表的出现,解决了我国长期以来缺少全领域、集成式科技知识组织工具这一文献服务转型的长期瓶颈问题与根本桎梏;对于提升海量文献的自动处理水平,实现从基于传统检索的信息服务向基于知识导航、自动主题聚类、内容相关揭示、自然语言检索和双语查询等智能化服务转变具有重要意义。NSTL词表本身的构建技术和应用模式,也成为国内在科技外文文献深度揭示与序化、知识化关联和智能化服务方面的最佳实践与典型案例。

同时,我们也应该看到,NSTL科技词表还面临着更新和持久构建的问题,以应对不断出现的新知识、新概念;在解决了全领域科技文献有效处理的同时,还应强化提升领域和深度服务。

参考文献

- [1] HILL Linda, BUCHEL Olha, JANÉE Greg, 等. 在数字图书馆结构中融入知识组织系统[J]. 现代图书情报技术, 2004, (1): 4-8.
- [2] 王军. 词表的自动丰富: 从元数据中提取关键词及其定位[J]. 中文信息学报, 2005(6): 36-43.
- [3] 张忠秋. 近年来我国分类表与主题词表研究与实践进展综述[J]. 图书馆工作与研究, 2013(2): 102-107.
- [4] 李育嫦. 传统知识组织系统的重构及其在网络环境下的应用[J]. 情报杂志, 2011, 30(7): 114-118.
- [5] 白海燕, 梁冰, 王莉. UMLS 及其在智能检索中的应用[J]. 现代图书情报技术, 2012, 28(4): 1-9.
- [6] 王军, 张丽. 网络知识组织系统的研究现状和发展趋

(下转第12页)

政策,规划其发展;二是农业科研工作者,这类人群对于元数据标准的理解相对透彻,结合某一深层次的科研范围,其需求也更加具有针对性、专业性;三是数量最多的广大的农业基层工作者,他们需要元数据标准来执行具体工作,其需求更为基础务实。

4.3 规整元数据标准的应用

在元数据标准的应用中有一项重要的环节即著录环节,这个过程既基础又关键,它严重影响着数据的质量,因为在未来标准建设道路上,需严格规整著录环节。首先,应将著录环节制定相关原则以及著录标准来指导相关人员开展此项工作,不能因个人意愿来决定完成质量,更不能对难以著录的数据选择避开;其次,应对元数据著录标准的工作人员提出要求,需对元数据以及元数据标准有深度的理解,尤其是对每一个元素项都要有所提炼;最后,应参照类似农业行业标准构建流程的平台或者管理系统,建立一个既实用又简练的著录平台(软件系统),用于管理、组织元数据,提供某一类别中所有的元素描述,提供获得信息的入口,容易开发,方便应用,同时专家还可以通过此平台(或者软件系统)第一时间对元数据进行评审与核查。综上所述,利用最便捷的手段规整元数据标准的应用,提高数据质量,扩充应用领域和深度,为我国农业领域元数据标准的发展发挥作用。

5 结语

对已制定的农业科技信息核心元数据标准,政府及管理部门应组织专家论证,进一步修改和发布,进行大范围的推广、试点,以促进该标准

的应用。鼓励各级政府部门、科研院校大力应用农业元数据标准,在此基础上建立专门的应用方案,加速农业科技信息元数据标准体系的建设。将农业信息元数据标准体系的研究纳入国家级科研项目,为后续研究的顺利进行有所保障,早日制定完整的农业领域各类元数据标准,提升我国农业元数据标准的管理水平以及应用范围。

参考文献

- [1] Metadata for managing scientific research data [EB/OL]. [2013-12-20]. http://www.niso.org/news/events/2012/dcmi/scientific_data/.
- [2] 朱虹.农产品流通信息元数据标准化研究[J].中国标准化,2014(7):54-59.
- [3] 崔运鹏.农业科技信息核心元数据标准框架研究及其著录信息管理系统[J].中国农业科学,2007,40(4):685-692.
- [4] MICHAEL J, DAVID M. Universal meta data models [M]. U.S.A.: Wiley Publishing, Inc., 2004.
- [5] 冯项云.国外常用元数据标准比较研究[J].大学图书馆学报,2001(4):15-21.
- [6] 吴显义.我国元数据研究现状分析[J].情报科学,2003,22(1):55-58.
- [7] 钱平,苏晓鹭,崔运鹏.农业科技信息核心元数据标准的研究[J].农业网络信息,2006(2):18-21.
- [8] 毕强.元数据标准及其互操作研究[J].情报理论与实践,2007,30(5):666-670.
- [9] YOU S. Evaluative metadata in educational digital libraries: how user evaluative metadata in the process of document selection[J]. TCDL Bulletin, 2010, 4(2): 1-11.
- [10] 马蕾.数字文献元数据标准比较分析[J].情报理论与实践,2003,26(1):72-74.
- [11] 孟宪学.农业科学数据核心元数据[M].北京:中国农业科学技术出版社,2005.

(上接第6页)

- 势[J].中国图书馆学报,2008,(1):65-69.
- [7] 张士男,宋文.《科图法》SKOS描述方案设计[J].现代图书情报技术,2010(6):7-11.
 - [8] 王刘安,常春.《英文超级科技词表》编制中概念优选词的选择方法研究[J].图书情报工作,2014(13):

107-112.

- [9] 王莉,梁冰,白海燕.以数据空间理念建立关系发现应用: NSTL智能检索平台的实践[J].数字图书馆论坛,2014(6):48-54.
- [10] 章成志.自动标引研究的回顾与展望[J].现代图书情报技术,2007(11):33-39.