农业科学叙词表的 SKOS 转化及其应用研究*

鲜国建　赵瑞雪　朱　亮　寇远涛
(中国农业科学院农业信息研究所　北京 100081)

【摘要】采用 W3C 推荐的标准 SKOS 将农业科学叙词表(CAT)中的叙词及词间语义关系进行描述,通过编写转化程序,将存放于数据库中的 CAT 自动批量地转化为 CAT/SKOS 格式。基于转化得到的 CAT/SKOS 以及开源全文检索引擎 Solr 和优化后的 Lucene - SKOS 插件,完成了百万级中外文农业科技期刊文摘的语义索引。同时,基于 CAT/SKOS 中的语义关系,实现一个可支持跨中英双语、可扩展和重构检索条件的智能检索原型系统,并对 CAT 在关联数据的构建与应用等方面进行展望。

【关键词】农业科学叙词表　SKOS　Lucene - SKOS　语义索引　语义检索
【分类号】G25

Conversion and Consumption of Chinese Agricultural Thesaurus as SKOS

Xian Guojian　Zhao Ruixue　Zhu Liang　Kou Yuantao
(Institute of Agricultural Information, Chinese Academy of Agricultural Sciences, Beijing 100081, China)

【Abstract】This paper represents the descriptor, non - descriptor and semantic relationships of Chinese Agricultural Thesaurus (CAT) as SKOS, and develops an application to convert CAT to CAT/SKOS automatically, and indexes millions of agricultural abstracts semantically based on CAT/SKOS, Solr and the improved Lucene - SKOS plugin. Then it realizes a semantic retrieval prototype system, which supports cross - language search (Chinese and English), and can also extend and reconstruct the query based on semantic relationship in CAT/SKOS. Finally, the authors make some prospects on the future applications of CAT/SKOS such as the publication as Linked Data.

【Keywords】Chinese agricultural thesaurus　SKOS　Lucene - SKOS　Semantic indexing　Semantic retrieval

1　引　言

在构建语义网络的过程中,尤其是从传统的文件网络(Web of Document)向具有结构化和富含语义的数据网络(Web of Data)演进过程中[1],传统知识组织系统(如叙词表、主题词表、分类法等)正在发挥越来越重要的作用。然而,在当前网络信息环境下,这些知识组织系统也需要与时俱进,为适应新的需求变化而不断推进其自身的发展和进化。

农业科学叙词表(Chinese Agricultural Thesaurus, CAT)作为一部大型、综合性农业叙词表,共收录了包括农业、林业、生物等领域在内的 6 万多个叙词、非叙词以及 13 万多条词间语义关系,为有效组织和利用我国的农业信息资源发挥了重要作用。近几年,针对 CAT 开展了大量的研究与实践,如将 CAT 中叙词及词间关系转换为资源描述框架(Resource Description Framework, RDF)、本体网络语言(Ontology Web Language, OWL)格式的轻量级本体[2,3],与FAO合作完成CAT与AGROVOC叙词表的映射[4]。然而,与互联网创始人Berners - Lee[5]提出的五

收稿日期: 2012 - 09 - 13
收修改稿日期: 2012 - 10 - 15
* 本文系国家"十二五"科技支撑计划项目"科技知识组织体系共享服务平台建设"(项目编号: 2011BAH10B03 - 3)的研究成果之一。

星级评价标准相比，目前的农业科学叙词表仅能获得两到三颗星，因为其还没有完全对外开放，也未提供便捷的获取和利用途径，所以非常有必要采用包括 RDF、SKOS、SPARQL 和关联数据等被推崇的开放标准和最佳实践，使农业科学叙词表变得更加开放、有用、可用和尽可能地被更多利用。

本文在简要介绍 SKOS 的基础上，将农业科学叙词表中的叙词、词间关系用 SKOS 提供的语言标签进行描述，并通过开发的转化程序，自动批量地将 CAT 向 CAT/SKOS 转化。基于 CAT/SKOS、开源的全文检索系统 Solr 及其插件 Lucene – SKOS，完成了对百万级中外文农业科技文摘数据的语义索引，最后实现简易的智能检索原型系统。

2　CAT 向 CAT/SKOS 的转化

2.1　SKOS 简介

简单知识组织系统(Simple Knowledge Organization System，SKOS) 是 W3C 在 2005 年制定的规范标准，是以资源描述框架(RDF) 为基础，为知识组织体系(包括叙词表、分类法、主题词表、术语表等) 提供了一套简单、灵活、可扩展的机器可理解的描述和转化机制，目的是为了资源的共享和重用。SKOS 由核心词汇(SKOS Core)、映射词汇(SKOS Mapping) 和扩展词汇(SKOS Extensions) 三部分组成。其中比较成熟的是 SKOS Core，已经形成了相应的语法标准和应用标准，而后两者目前还处于发展阶段。

在提出 SKOS 标准后，国内外的图书情报界已开展了一系列的知识组织系统的 SKOS 描述转化研究，如荷兰视听档案通用词汇表(GTAA)[6]、医学主题词表(MeSH)[7]、美国国会图书馆标题表(LCSH)[8]、FAO 多语种农业叙词表 AGROVOC[9] 等；2009 年，《杜威十进分类法》(简称 DDC) 以 SKOS 格式发布[10]，目前提供了前三级类目数据的开放下载；张士男等[11] 提出了《中国科学院图书馆图书分类法》中类目、类号、关系、类目注释等的 SKOS 转换；刘丽斌等[12] 建立了《中国分类主题词表》的 SKOS 描述自动转换方案。

2.2　CAT 的 SKOS 表达

笔者曾利用本体描述语言 OWL 对 CAT 进行了规范化描述和转化，但由于 OWL 语义描述很强，在常规语义网络环境下难以加以应用并发挥其功能，而具有轻量级语义描述能力和更广泛应用场景的 SKOS，能最大程度地兼容传统知识组织体系，实现其在网络环境下的转化和应用。因此，本文将基于 SKOS 的描述词汇，并借鉴国内外知识组织系统向 SKOS 转换的经验，实现农业科学叙词表 CAT 向 SKOS 的转换。

(1) 叙词向概念转化

在将 CAT 向 SKOS 转换时，每个叙词都将被转化为 SKOS 的一个概念。作为 SKOS 的概念，唯一标识符(URIs) 是必备要素，用以唯一标识 skos: Concepts 的实例。在关联数据的构建实践过程中，推崇应用 HTTP URLs 来标识资源。CAT 的叙词及非叙词都拥有一个稳定、唯一的编号(Term – Code)。因此，在将叙词转换为概念时，term – code 将作为 HTTP URL 模板“http://aii.caas.net.cn/cat/concept/{term – code}”的一部分，以确保可通过稳定的 URLs 来解析 CAT 中的各个概念。

(2) 标签属性的应用

SKOS 提供了 skos: prefLabel 和 skos: altLabel 等标签属性，以将优选和替换的自然语言标签与特定概念相关联。本文将 CAT 叙词的中文字符串和英文对译字符串分别映射为带有语言标记的 skos: prefLabel，而其“代”的非叙词则以同样带语言标记的 skos: altLabel 来表达。

(3) 语义关系的转化

农业科学叙词表中的语义关系主要包括“用、代、属、分、参”等类型。“用、代”已通过标签属性进行了表达，本文采用 skos: broader、skos: narrower 和 skos: related 来分别转化“属、分、参”三种关系。在描述 CAT 与 AGROVOC 这两个叙词表的映射成果时，则主要应用 skos: inScheme、skos: narrowMatch 和 skos: broaderMatch 等词汇。在转化后的 CAT/SKOS 中，所有概念资源之间都通过它们的唯一标识符 URIs 来建立各种语义关联关系。利用 SKOS 的数据描述模型，将 CAT 中叙词“大豆”知识片断及其与 AGROVOC 映射成果的描述如图 1 所示。

2.3　CAT 向 CAT/SKOS 的批量转化

笔者开发设计了自动批量转化程序，将 CAT 中 6 万多个叙词、非叙词，以及 13 万余条词间关系自动批量转化为 SKOS 格式的 RDF 文档，转化结果如图 2 所示。

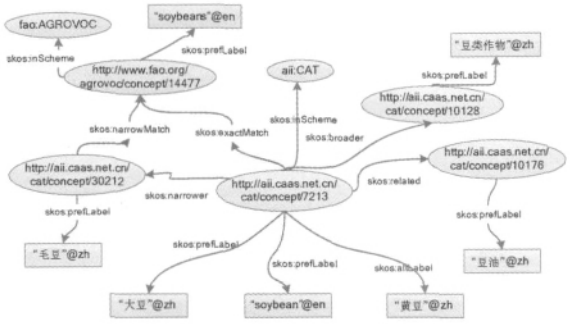

图1 农业科学叙词表的SKOS描述

图2 农业科学叙词表向SKOS的批量转化结果

3 基于CAT/SKOS的语义索引及智能检索

通过对农业科学叙词表基于SKOS开展转化研究,形成了CAT的SKOS版(CAT/SKOS)。在CAT/SKOS这一RDF文档中,已通过规范的语义标签建立了丰富的内外部语义关联关系,已是真正意义上的关联数据,这将为基于CAT/SKOS开展新型知识服务应用奠定良好的语义资源基础。本文将基于CAT/SKOS、开源软件Solr和农业科技期刊文摘数据库来开展语义索引和智能检索等方面的研究和实践。

3.1 全文检索引擎Solr和Lucene-SKOS插件

Solr是在全文索引工具Lucene基础上进行了封装和功能扩展,它是一个高性能的、可独立运行的企业级全文搜索引擎服务器。Solr在2007年正式成为Apache的子项目后,先后共发布了多个版本,最新版为2012年7月发布的Solr 3.6.11[13]。目前,Solr能为多种数据格式提供全文索引、检索、分面浏览、高亮显示、富文档处理、分布式检索、索引复制和空间数据检索等强大功能。

此外,Solr还提供了可灵活扩展的插件体系架构,用户可根据实际需要进行自定义扩展和配置。Lucene-SKOS就是基于该体系而实现的语义索引和检索插件。实质上,Lucene-SKOS是为Apache Lucene和Solr定制的一个分析器模块,它以存在于SKOS中的概念及其语义关系为基础,为建立Lucene文档索引及执行查询条件提供术语扩展。该插件目前支持两种形式的术语语义扩展:基于唯一标识符URI的扩展和基于字符串标签的扩展[14]。

3.2 基于CAT/SKOS构建语义索引

在测试Lucene-SKOS过程中,笔者发现该插件还存在一些不足。在启动Solr后,该插件首先会将指定的整个SKOS文档基于Jena进行解析,并将解析结果(如唯一标识符URI、概念、属性标签和语义关系等)在内存中建立对应的Lucene索引。然而,一旦SKOS文档较大(比如20MB),该插件就可能因内存不足而停止工作。在分析源码过程中还发现,该插件在进行语义扩展时,只考虑了"用、代、属、分"关系,没有将"参"纳入扩展范围。因此,笔者对该插件进行如下改进和完善:

(1)新增了可以加载磁盘中现成索引文件(事先建立SKOS的Lucnce索引,没有大小限制)的初始化函数,代码片断如下所示:

```
public SKOSEngineImpl( String indexFilePath) throws IOException {
    this . indexDir = new SimpleFSDirectory ( new File ( indexFile-
        Path) ) ;
    this. searcher = new IndexSearcher( indexDir) ;
    System. out. println( "SimpleFSDirectory created") ; }
```

(2)对源代码中的SKOSFilterFactory. java文件也作了相应修改,包括添加成员变量indexFilePath,重写inform(ResourceLoader loader)方法,确保该插件能支持指定SKOS文件和指定其索引文件两种模式。由于转化得到的CAT/SKOS文档较大,为避免解析该文档导致内存溢出的问题,笔者基于存放于关系型数据库的CAT构建了符合该插件规范的Lucene索引文件。

国家农业图书馆经过多年的数字化建设,已建成了专业领域集中、元数据著录完整规范、学术价值较高的中外文农业科技期刊文摘数据库。笔者利用Solr和改进后的Lucene-SKOS插件,并基于CAT/SKOS索引文件,对400多万条文摘数据建立了语义索引(基于字符串标签的术语语义扩展模式)。为避免过多的语义扩展,目前只在关键词和题名这两个字段建立了语义索引。若在文摘或全文层次建立语义索引,不但会导

致索引文件急剧增大，也会导致过多的语义扩展，影响检索结果。

3.3　基于 CAT/SKOS 的智能检索原型系统

基于建立的农业科技文摘语义索引文件和开源前端展示交互脚本库 Ajax – Solr[15]，实现了基于 CAT/SKOS 进行语义扩展的智能检索原型系统。除 Solr 提供的全文检索、检索结果高亮显示、分面浏览与检索、相似文献检索等强大功能外，该系统还支持对关键词字段进行中英双语检索，以及基于农业科学叙词表中丰富的词间关系，进行检索条件的语义扩展和重构，如图 3 和图 4 所示：

图 3　基于 CAT/SKOS 的智能检索原型系统
（只检索中文）

图 4　基于 CAT/SKOS 的智能检索原型系统
（中英双语检索）

可以看出，在基于 CAT/SKOS 实现语义索引基础上，即使用户只输入中文检索词“大豆”，在命中结果中包含与“大豆”对应的“黄豆”、“青豆”、“Soybean”等语义密切相关的其他中英文词的信息也被检索出来，因此，简单的智能检索功能基本实现。通过点击页面左边的分页检索智能导航区域的超链接，则可以直接检索与“大豆”存在语义关系的其他词的信息，实现了简单的智能导航功能。

4　结　语

本文利用 SKOS 对农业科学叙词表中的叙词及词间关系进行了描述，并将其批量地转化为具有关联数据特性的 CAT/SKOS。同时，基于开源软件建立了农业科技文摘数据的语义索引和智能检索原型系统，实现将叙词表嵌入检索系统，提高了检索效率。近年来，关联数据作为一种被推荐的最佳实践，广泛应用于语义网，通过使用 URIs 和 RDF 发布、分享、连接各类数据、信息和知识[16]。以下工作将进一步开展：把 CAT/SKOS 发布为公开的关联数据，提供 SPARQL 查询终端、RDF 片断/文档下载等服务，并与 AGROVOC、LOD 云图中的其他开放关联数据进行语义映射和广泛互联；在 CAT 的应用方面，继续完善智能检索原型系统的各项功能，将“参”关系也纳入语义扩展范围，支持语义关系扩展类型任意组合的自定义设置，并且以可视化方式将语义扩展和查询条件重构的结果和链接提供给用户。

参考文献：

[1] The Rise of the Data Web [EB/OL]. [2012 – 06 – 18]. http://www.dataspora.com/2009/08/the – rise – of – the – data – web/.

[2] 常春. Ontology 在农业信息管理中的构建和转化 [D]. 北京：中国农业科学院研究生院，2004.（Chang Chun. Construction and Conversion of Ontology in Agricultural Information Management [D]. Beijing: Graduate School of Chinese Academy of Agricultural Sciences, 2004.）

[3] 鲜国建，孟宪学，常春. 农业科学叙词表的 OWL 表示研究 [J]. *中国农业科学*，2007，42（S2）：91 – 95.（Xian Guojian, Meng Xianxue, Chang Chun. Study on the Presentation of China Agricultural Thesaurus in OWL [J]. *Scientia Agricultura Sinica*, 2007, 42（S2）: 91 – 95.）

[4] Liang A, Sini M, Chang C, et al. The Mapping Schema from Chinese Agricultural Thesaurus to AGROVOC [EB/OL]. [2012 – 06 – 18]. ftp://ftp.fao.org/docrep/fao/008/af241e/af241e00.pdf.

[5] Berners – Lee T. Linked Data – Design Issues [EB/OL]. [2012 – 06 – 20]. http://www.w3.org/DesignIssues/LinkedData.html.

[6] Resulting RDF: gtaa/GTAAinstancesSKOSv7.rdf [EB/OL]. [2012 – 10 – 03]. http://thesauri.cs.vu.nl/eswc06/gtaa/GTAAinstancesSKOSv7.rdf.

[7] MeSHToSKOS [EB/OL]. [2012 – 10 – 09]. http://code.

google. com/p/hive - mrc/wiki/MeshToSKOS.

[8] Summers E, Isaac A, Redding C, et al. LCSH, SKOS and Linked Data [C]. In: *Proceedings of the International Conference on Dublin Core and Metadata Applications*, Berlin, Germany. 2008: 25 - 33.

[9] Morshed A, Keizer J, Johannsen G, et al. From AGROVOC OWL Model Towards AGROVOC SKOS Model [EB/OL]. [2012 - 06 - 20]. http://www.fao.org/docrep/012/al300e/al300e00.pdf.

[10] Dewey Decimal Classification/Linked Data [EB/OL]. [2012 - 07 - 03]. http://dewey.info.

[11] 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. (Zhang Shinan, Song Wen. Description Scheme of LASC in SKOS [J]. *New Technology of Library and Information Service*, 2010(6): 7 - 11.)

[12] 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*,2009,35(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, 35(6): 56 - 60.)

[13] Apache Solr [EB/OL]. [2012 - 07 - 20]. http://lucene.apache.org/solr/.

[14] Lucene - SKOS/SKOS Support for Apache Lucene and Solr [EB/OL]. [2012 - 07 - 20]. https://github.com/behas/lucene - SKOS.

[15] Ajax - Solr [EB/OL]. [2012 - 07 - 22]. http://evolvingweb.github.com/ajax - solr/.

[16] Linked Data - Connect Distributed Data Across the Web [EB/OL]. [2012 - 06 - 18]. http://linkeddata.org/.

(作者 E - mail: xgj@mail.caas.net.cn)

《现代图书情报技术》特邀专栏组稿

《现代图书情报技术》是中国科学院主管、中国科学院国家科学图书馆主办的计算机信息管理技术方面的学术性刊物。刊物拥有清晰的定位,即以跟踪技术的研究、应用、交流为主体,服务于广大信息技术人员。

本刊从 2004 年起开设不定期栏目——《特邀专栏》,每一期专栏集中发表关于某个特定方面的技术研发与应用的研究型文章,汇集科研成果、聚焦研究前沿。

1 《特邀专栏》操作办法及流程

(1) 本栏目特邀国内外知名专家、学者、教授担任专栏主编,专栏的设立一般由期刊的策划编辑和特邀专栏主编沟通,根据国内外图书情报技术学科的发展需要提出选题。

(2) 选题一旦确定后,由特邀专栏主编承担稿件的组织,审核并撰写前言。一期特邀专栏一般为 4 - 6 篇文章为宜。稿件组织过程中,策划编辑将与特邀专栏主编进行定期的沟通,及时掌握稿件的撰写情况,并对稿件的撰写提出适当的建议和意见。

(3) 稿件经特邀专栏主编审核通过,提交给编辑部。后期由策划编辑负责与作者的联系沟通及安排出版等事宜。

(4) 专栏的选题一旦确定后,将确定基本时间表。一般的操作周期为 3 - 5 个月。以正式确定特邀专栏题目为起始点,在 1 个月内确定约请论文的作者和题目,3 个月内确定初稿,5 个月内确定采用稿。

2 《特邀专栏》稿件内容要求

(1) 深入反映本专栏选题方向的前沿研究成果或重大应用成果,侧重理论研究、技术分析、系统论证或设计等,注意理论与实践相结合。

(2) 特邀专栏稿件应该主要是原始性和原创性研究论文,也可以有一篇综述性论文,但综述性论文必须可靠地覆盖该方向的原始核心文献。

(3) 文章按照严谨的学术文章体例写作,即明确扼要地界定研究问题,简要说明研究方法,系统精炼地描述国际国内发展状况,进而详细地描述作者自身研究工作的技术线路及研究结果。

(4) 特邀专栏的一系列文章应注意覆盖专栏选题所涉及的各个研究方向和多个研究单位,充分覆盖可能存在的多种观点和技术线路。

(5) 充分承认前人/别人的工作,充分引证所参考引用的文献(尤其是本研究工作中的原始核心文献和国内最先出现的研究文献),严格遵守著录规范。

3 《特邀专栏》稿件格式要求

(1) 论文版式请参照本刊网站"下载专区"中"论文模板"。

(2) 多个作者时,请注明通信作者,并注明各个作者的单位。

(3) 每篇稿件以 6 - 8 千字为宜(按篇幅字数计算,包括图、表)。