

国家社科基金资助项目成果
广东省哲学社会科学规划项目成果

中文知识组织系统

——语义描述、共建及共享服务

曾新红 等著



化学工业出版社

国家社科基金资助项目成果

广东省哲学社会科学规划项目成果

中文知识组织系统

——语义描述、共建及共享服务

曾新红 等著



化学工业出版社

· 北 京 ·

图书在版编目 (CIP) 数据

中文知识组织系统——语义描述、共建及共享服务/
曾新红等著. —北京: 化学工业出版社, 2016. 7

ISBN 978-7-122-27171-6

I. ①中… II. ①曾… III. ①数字图书馆-自动化系
统-研究 IV. ①G356.6

中国版本图书馆 CIP 数据核字 (2016) 第 115015 号

责任编辑: 张安超 段志兵

文字编辑: 颜克俭

责任校对: 王素芹

装帧设计: 关 飞

出版发行: 化学工业出版社 (北京市东城区青年湖南街 13 号 邮政编码 100011)

印 装: 大厂聚鑫印刷有限责任公司

787mm×1092mm 1/16 印张 23½ 彩插 4 字数 583 千字 2016 年 10 月北京第 1 版第 1 次印刷

购书咨询: 010-64518888 (传真: 010-64519686) 售后服务: 010-64518899

网 址: <http://www.cip.com.cn>

凡购买本书, 如有缺损质量问题, 本社销售中心负责调换。

定 价: 98.00 元

版权所有 违者必究

Z 族首词 1 § 族首词 2

图 4-7 叙词表文本分隔符定义 (见108页)

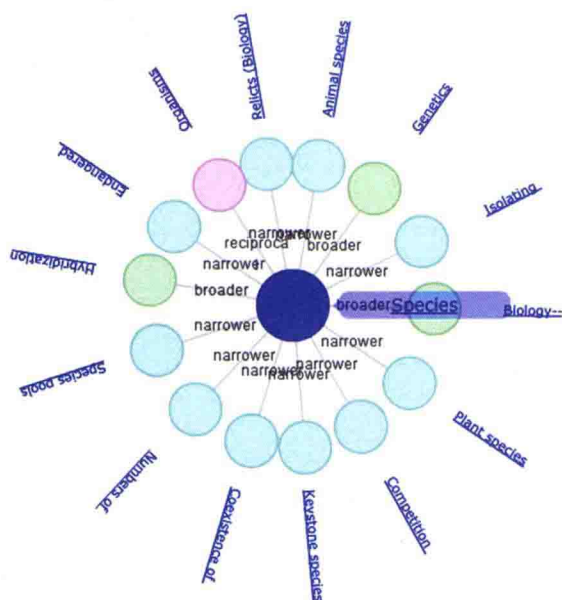


图 10-13 LCSH 的可视化效果 (见296页)

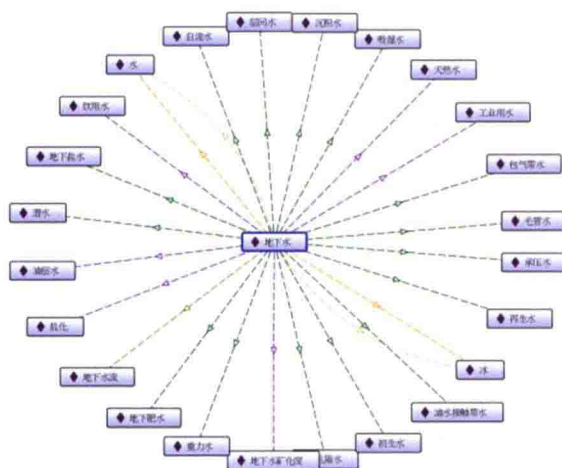


图 10-14 protégé 的 Radial 展示效果 (见297页)

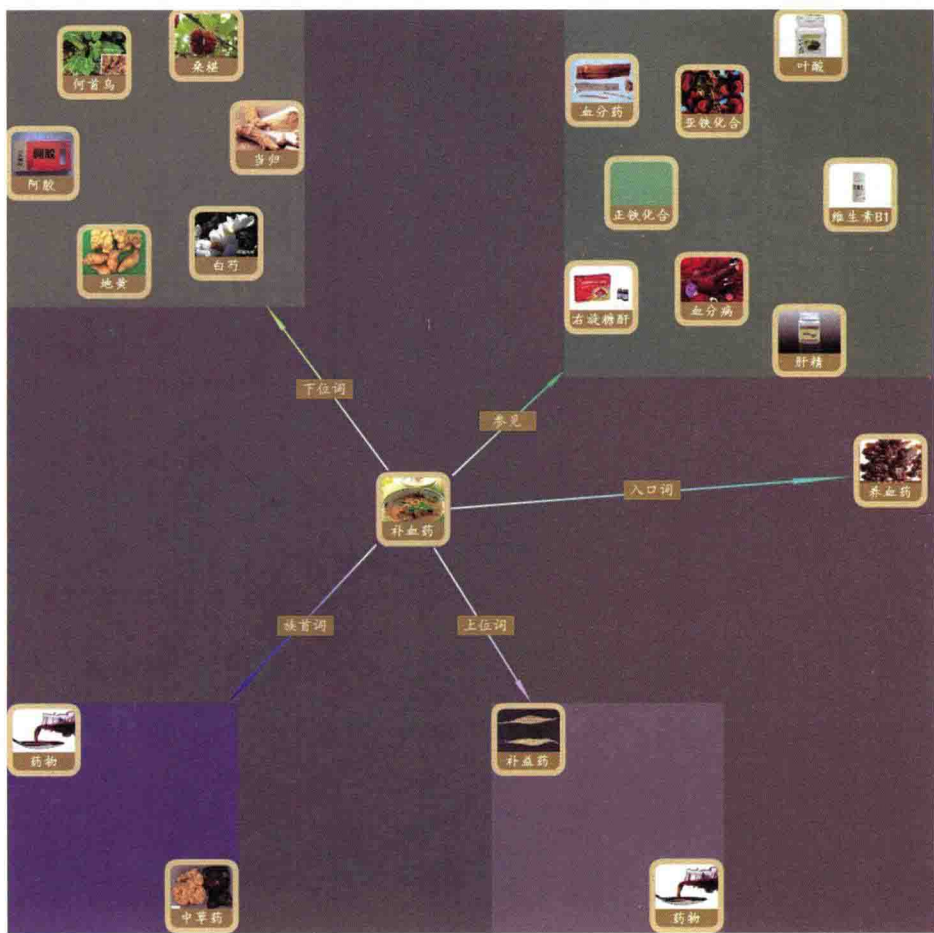


图 10-22 概念实例“补血药”可视化效果图(见304页)

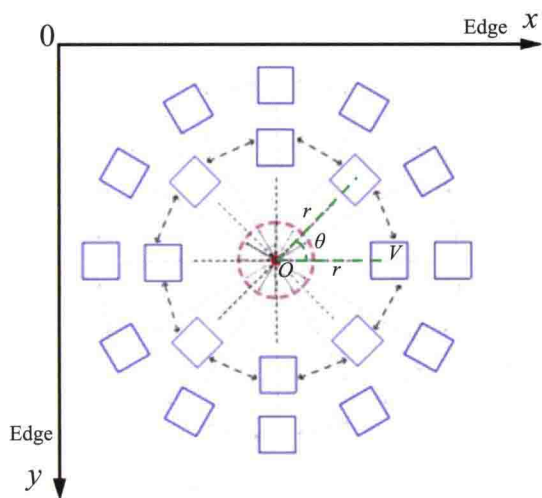


图 10-25
生成节点初始位置原理示意图 (见309页)

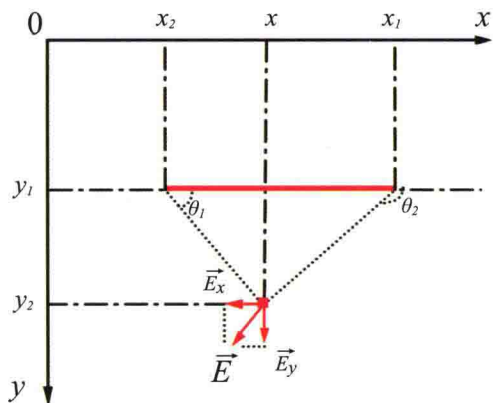


图 10-28
有限带电导体求场强示意图 (见310页)

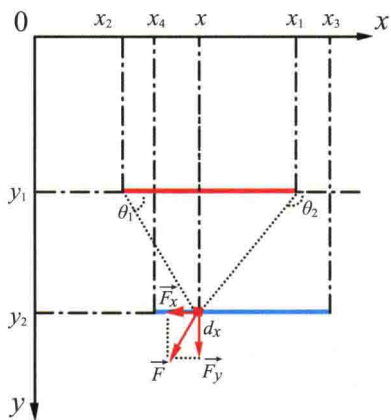


图 10-29

两边沿 x 轴平行但不共线 (见311页)

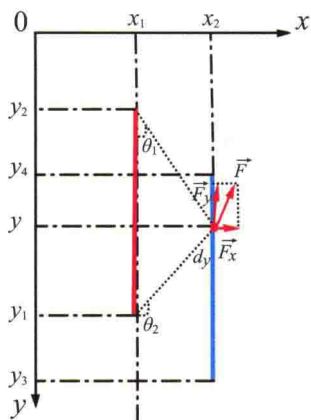


图 10-30

两边沿 y 轴平行但不共线 (见311页)

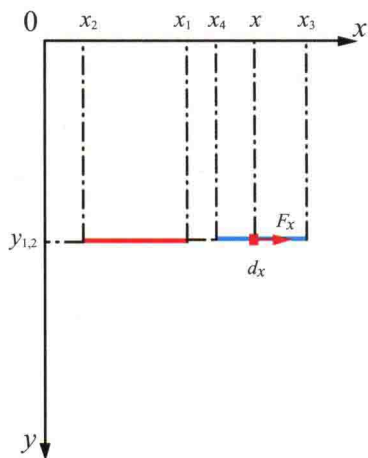


图 10-31

两边沿 x 轴共线 (见311页)

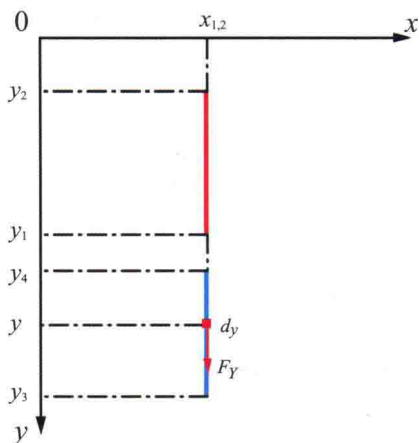


图 10-32

两边沿 y 轴共线 (见311页)

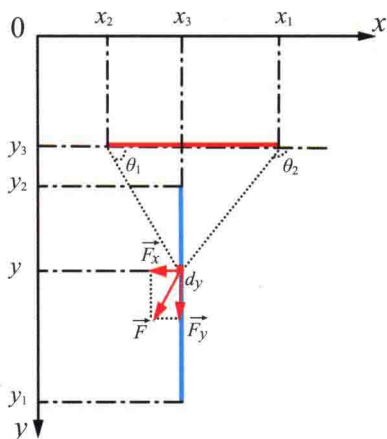


图 10-33

两边垂直, 红边平行于 x 轴 (见312页)

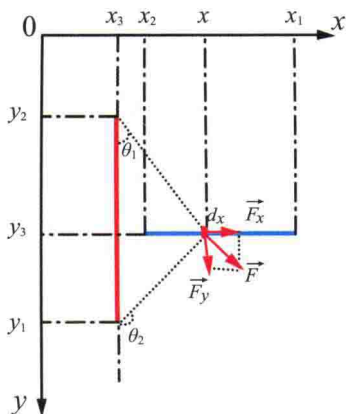


图 10-34

两边垂直, 红边平行于 y 轴 (见312页)

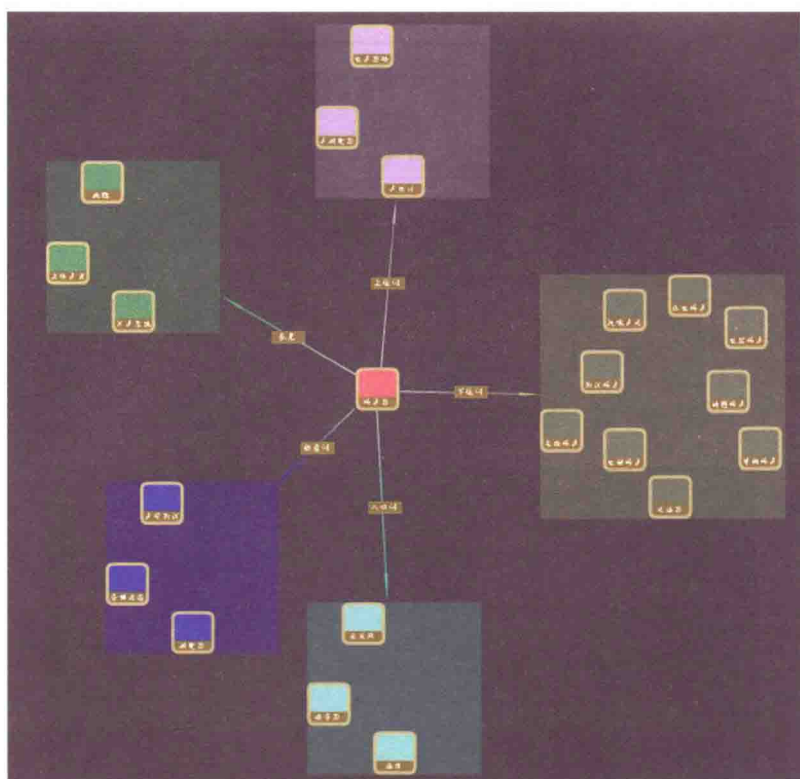


图 10-38 叙词“传声器”的布局 (见315页)

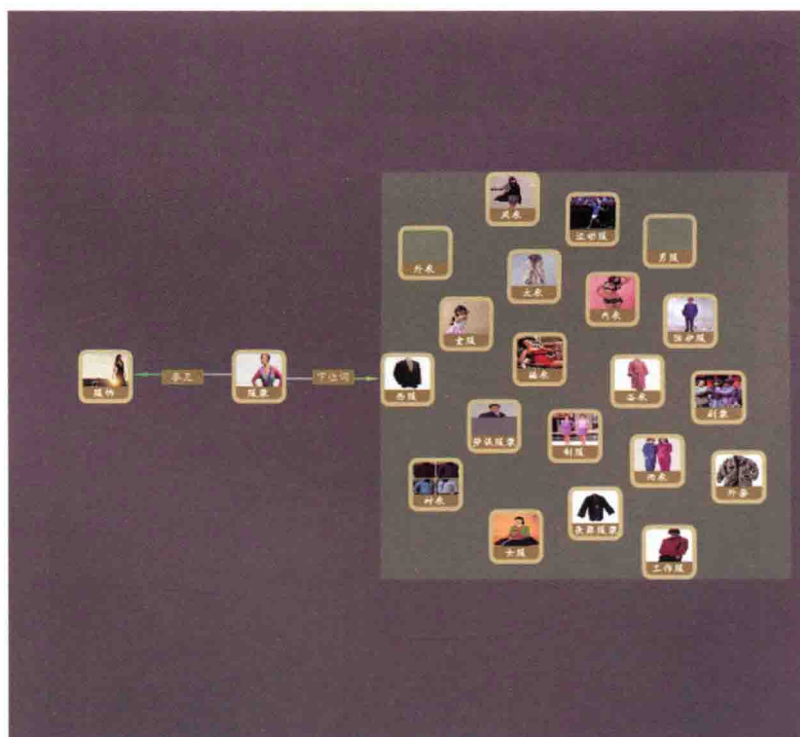


图 10-39 叙词“服装”的布局 (见315页)

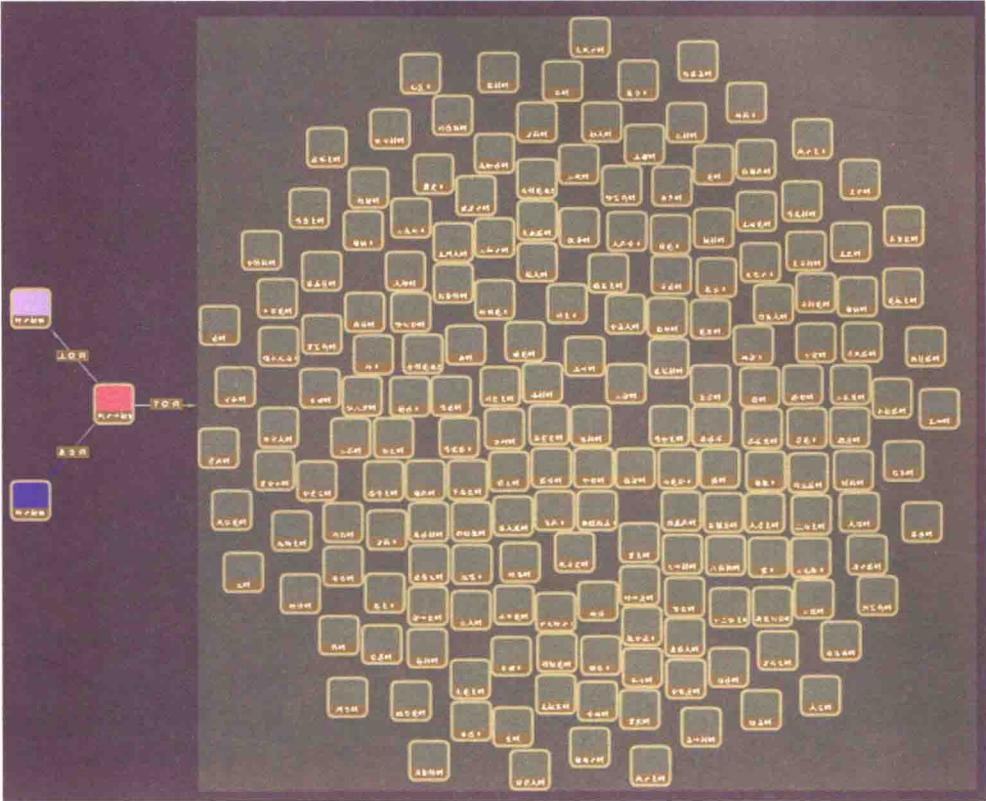


图 10-40 叙词“双子叶植物”的布局 (见316页)

分类法共享服务系统 (CLSS)
Classification Scheme Sharing Service System

进入CLSS Linked Data 设为首页 收藏此页
欢迎您! 2015年12月09日 星期三

检索 浏览 主表: A B C D E F G H I J K N O P Q R S T U V X Z 附表: 1 2 3 4 5 6 7 8 说明 设置

└ B152 二十一世纪哲学

└ B17 马克思主义哲学的传播和发展

└ B2 中国哲学

└ B20 唯物主义与唯心主义

└ B201 唯物主义

└ B202 唯心主义

└ B21/26 中国各代哲学

└ B21 古代哲学

└ B22 先秦哲学 (~前220年)

└ B221 诸子前哲学

└ B222 儒家

└ B222.1 四书

└ B222.2 孔子(孔丘, 公元前551~前479年)

└ B222.3 孔子弟子

└ B222.4 子思(孔伋, 公元前483~前402年)

└ B222.5 孟子(孟轲, 公元前390~前305年)

B222.2 孔子(孔丘, 公元前551~前479年)

获取 细分 分类等级 对应主题 CNKOS SKOS 查OPAC

显示或隐藏关联数据 B222.2

uri http://nkos.lib.szu.edu.cn/CCT_CLC_V4.0#B222.2

skos:notation B222.2

rdf:type <http://www.w3.org/2004/02/skos/core#Concept>

skos:prefLabel 孔子(孔丘, 公元前551~前479年)

skos:closeMatch 孔丘(前551-前479) / 哲学

skos:broader B222

图 11-7

分类类目“B222.2”近似映射主题“孔丘(前551-前479) / 哲学”(见339页)



图 11-8

点击主题“孔丘(前551-前479)”后展示CCT2_OTCSS Linked Data 服务相应款目（见340页）



图 11-9

点击主题款目中的中图法分类号映射到CLC4_CLSS Linked Data 服务相应分类款目（见340页）

前言

网络信息的极度海量和无序已使越来越多的人在思考网络信息资源的有效组织和高效检索，人们重新把目光投向了传统的知识组织系统——以叙词表和分类法为代表的情报检索语言。网络知识组织系统 (Networked Knowledge Organization System, NKOS)，就是在这样一个背景下产生和发展起来的。

NKOS 有两种类型：一种是来自信息科学界 (Information Science，即国内所称的图书馆学情报学界或图书情报界) 的传统知识组织系统的延伸和发展，如分类法、叙词表、主题标题表、规范档等的网络应用；另一种则是在网络环境中产生和发展起来的语义工具，如本体 (Ontology) 和语义网络 (Semantic Network，如 WordNet——词汇数据库) 等。时至今日，网上可获取和应用的中文 NKOS 仍然十分匮乏。

NKOS 框架下的 KOS 是一种广义的知识组织系统概念，包括所有的 (广义的) 受控词表，大致可分为：从最简单的线性结构的、提供多义性 (ambiguity) 控制的各种术语列表，到具有等级关系控制的、树状结构的分类法和范畴表 (有些含少量横向相关关系，如美国国会图书馆标题表 LCSH)，再到基于关系的 (含纵向等级关系和横向相关关系)、网状结构的知识组织系统类型 (如叙词表、领域本体)。叙词表和本体同属于最后这种结构最复杂、控制程度最高的类型。我们有理由相信：只要解决了最高端类型的知识组织系统的形式化表示和网络应用问题，其他低端 KOS 的问题只是最高端类型的简化，自可迎刃而解。

笔者对中文叙词表本体的研究始于 2004 年。一次偶然的机会，笔者了解到“本体”的概念与叙词表是如此相似，通过查找资料发现了刚刚发布为 W3C 推荐标准的 OWL 以及国外在词表转换方面的先驱研究。凭着多年编目工作中积累下来的对《中国分类主题词表》(CCT) 的了解，笔者研究了将 CCT 转换为 OWL 的可能性。随后，在国家社科基金青年项目“基于本体和知识集成实现中文叙词表的升级、共享和动态完善” (编号：05CTQ001) 的支持下，将这一设想进一步细化为形式化表示的中文叙词表本体 (OntoThesaurus)，并组织课题组成员开发实现了其支持系统“中文叙词表本体共建共享系统” (OTCSS)。

通过将 OntoThesaurus 与 W3C 推荐标准 SKOS (简单知识组织系统) 进行深入比较，笔者意识到完全依赖已有的国际标准来全面实现中文知识组织系统的语义描述和网络共享是不可行的。中文 KOS 有其自身的发展历程，形成了有别于西方文字 KOS 的一些特点，这就决定了中文 KOS 的形式化表示不可能完全依赖国外组织制定的规范，而必须自行研究制定出一些符合自身特点而又与国际相关标准兼容的规范，先解决国内的需要，再进一步寻求国际支持。随后，在国家数字图书馆工程“知识组织标准规范”项目 (编号：D009)、广东省哲学社会科学“十一五”规划项目 (编号：GD10CTS02) 和国家社科基金项目“中文知识组织系统形式化语义描述标准体系研究” (编号：12BTQ045) 支持下，笔者研究建立了“中文知

识组织系统形式化语义描述标准体系”，并组织课题组成员开发实现了更完善的支持系统。

“中文知识组织系统形式化语义描述标准体系”研究的总体目标，是基于国内知识组织系统的特点和发展需要，参考国际上网络知识组织系统的现有成果和发展趋势，提出中文知识组织系统的形式化语义描述标准体系，为国内相关团体和个人构建各类网络知识组织系统提供一系列标准规范选择或参考。进一步，为该标准体系提供配套的技术服务，进而推动国内知识组织系统的网络化发展和应用。

“中文知识组织系统形式化语义描述标准体系”由以下三份标准文本构成：通用 CNKOS 语义描述规范；高受控词表的 OntoThesaurus 描述规范；中文领域本体的 OWL 描述规范。

“通用 CNKOS 语义描述规范”中包含了中文知识组织系统的纯 SKOS 描述规范以及 CNKOS 扩展版描述规范。其前身名为“受控表语义描述规范”，得到了国家图书馆的资助，已被国家图书馆采纳，为其知识组织标准规范的组成部分。后续研究将其提升为通用 CNKOS 语义描述规范，并制定了 CNKOS 的 RDF Schema，以符合语义 Web 规范的方式为机器提供类和属性定义的共享服务。其中，纯 SKOS 描述规范为所有的中文知识组织系统提供具体的 SKOS 语义描述方法，使它们能够参与 SKOS 级别的共享应用，而 CNKOS 扩展版则可以用来实现传统中文受控词表（如中文主题词表、分类法等）的全描述，使它们在保证不损失原有丰富语义的前提下提供网络共享服务。

“高受控词表的 OntoThesaurus 描述规范”是在中文叙词表本体（OntoThesaurus）研究基础上进行的升级，笔者进一步将 OntoThesaurus 提升为符合语义 Web 规范的描述语言形式，拥有独立的命名域、缩写和严格的形式化定义，制定了 OntoThesaurus 的 RDF Schema，以便更好地融入语义 Web 应用环境。并根据 CNKOS 研究过程中发现的新问题，对少许类和属性进行了调整。该规范为高受控的中文叙词表（主题词表）、专类分类表（Taxonomy）、规范档等提供了严格的语义描述规范，使它们能够最大限度地保留已有的词表控制成果，可实现严格的一致性检测，并有利于向细粒度本体演化。

“中文领域本体的 OWL 描述规范”是在全面梳理 W3C 推荐标准 OWL1 和 OWL2 的标准文本的基础上，针对国内中文领域本体研究中所涉及的描述需求，编制的一套易于入门和操作的参考性描述规范，由一份“中文领域本体的 OWL 应用指南”及本课题组翻译的 5 份 OWL1 和 OWL2 标准文本中译文组成，方便中文领域本体的建模者参考和使用 OWL 的标准定义。这 5 份中译文文本同时发布在 W3C 官方网站 W3C Translations 上，供全球需要者参考使用。

纯 SKOS 和 OWL 格式描述的 KOS，可以利用业界现有的 SKOS 和 OWL 支持工具进行处理和提供服务。我们 NKOS 研究室团队为自定义的 OntoThesaurus 和 CNKOS 开发了支持系统，即基于 OntoThesaurus 的叙词表类型（Thesaurus-like）中文知识组织系统支持系统 OTCSS，以及基于 CNKOS 的分类法类型（Classification-like）中文知识组织系统支持系统 CLSS。这两个支持系统分别面向叙词表类型和分类法类型（即为概念指派了唯一标识代码的 KOS，如分类法、范畴表、标准代码表等），提供集成、可重用和面向语义 Web 的统一解决方案，可以快速实现国内外现有中文主题词表或分类法的 Web 版、Linked Data 服务、Web Service 服务功能，并可快速移植到同类型的结构更简单的 KOS 上。这些解决方案对于国外同类型知识组织系统实现语义 Web 化共享服务也具有参考价值。本书相应的章节将详细介绍这些支持系统的功能设计和实现技术。

本书将两个国家社科基金 05CTQ001 和 12BTQ045 的结题研究报告以及这两份研究报告间隔期间发表的相关学术论文或研究报告进行了整合。笔者汇集多年的研究积累和实践经

验,试图在图书情报、语义 Web、信息技术等相关各界之间架起桥梁,吸取各学科的成果精华,结合各界对中文知识组织系统的需求,为 NKOS 的研究者、构建者和服务者提供广泛深入的背景知识和行之有效的解决方案。由于中文叙词表本体及其共建共享系统 OTCSS 的研究是整个研究的起点和基础,因此将其置于本书的第一篇(升级后的“高受控词表的 OntoThesaurus 描述规范”作为第二篇的第八章出现),分为以下五章:第一章 概论;第二章 中文叙词表的本体化升级及形式化语义描述;第三章 OTCSS 的功能机制;第四章 OTCSS 的设计与实现;第五章 系统与应用。第二篇为“中文知识组织系统形式化语义描述标准体系”,分为以下四章:第六章 知识组织系统语义描述的研究基础;第七章 通用 CNKOS 语义描述规范;第八章 高受控词表的 OntoThesaurus 描述规范;第九章 中文领域本体的 OWL 描述规范简介。第三篇为“中文知识组织系统形式化语义描述标准体系之支持系统”,分为以下两章:第十章 叙词表类型的中文知识组织系统之支持系统:OTCSS,第十一章 分类法类型的中文知识组织系统之支持系统:CLSS。

此项研究持续了十余年,期间参与者众多,具体分工如下。

曾新红:研究命题和研究思路确定,语义描述标准规范及转换方案研究,相关学术论文、结题研究报告及本书撰写,系统需求制定,系统测试,整体研究工作的主持和推进,所有研究论文和 OWL 标准文本中译文的修订和审校,NKOS 研究室网站的整体设计和内容提供。

林伟明:OTCSS 系统 Jena 版的具体实现,研究论文相关内容的起草,系统说明书撰写,Lucene 版方案的提出,OWL 标准文本中译文的发布。

黄华军:OTCSS 系统 Lucene 版的具体实现,CLSS 系统的具体实现,系统各项后续功能的整合,相关研究论文的起草,系统说明书撰写。

蔡庆河:可视化系统的具体实现,相关研究论文的起草,3 份 OWL2 标准文本的初译。

张水英:研究资料翻译。

陈若韵:分类法细分规则研究。

唐铨:NKOS 研究室网站的具体实现,可视化概念图片自动搜索的初始研究。

田金凤:叙词定义自动构建研究。

方圆:入口词自动构建研究。

曾汉龙:可视化系统的初始研究。

徐乐、吴鹏:一份 OWL 标准文本的初译。

此外,在本研究刚刚开始的最困难阶段,明仲、蒋颖、胡振宁等老师在人员设备、立项申请等方面提供了大力支持,使本研究得以开展和延续;敦煌研究院的李鸿恩老师提供了《敦煌检索词表》Word 版原始数据,为我们研究 OTCSS 系统贡献了宝贵的素材;在承担国家数字图书馆“知识组织标准”子项目期间,国家图书馆词表组的卜书庆老师及其团队提供了《中图法》B 大类的 MARC 数据、国图 SKOS 方案及其说明,并与北京大学的王军老师、南京大学的欧石燕老师一起,对 CNKOS 语义描述规范的研究提出了宝贵的意见,使 CNKOS 规范的研究得以进行和不断优化;本研究是在国内外已有相关研究的基础上进行的,这些参考文献的作者,为我们的研究打开了思路;还有在历次课题评审、成果鉴定、成果评奖、论文审理中无私帮助过我们的匿名专家们,他们的支持和鼓励是我们不断向前的强劲动力。在此,著者一并对他们表示衷心的感谢!

受著者的学识限制,本书中难免会有不当之处,敬请读者批评指正。

曾新红

2016 年 1 月于深圳

目 录

第一篇

中文叙词表本体及其共建共享系统 OTCSS / 1

引言	2
----------	---

第一章 概 论 / 3

1.1 知识组织系统的发展和应用	4
1.1.1 叙词表的发展历程及背景知识介绍	4
1.1.2 网络信息社会对知识组织系统的需求	9
1.1.3 信息科学界(图书情报界)的应对研究	12
1.1.4 来自知识组织系统其他相关各界的启示	14
1.1.5 叙词表与本体的比较研究	15
1.1.6 叙词表与本体的构建方法研究	17
1.2 中文叙词表本体共建共享系统(OTCSS)的构建	19
1.2.1 解决的主要问题	19
1.2.2 总体结构	20

第二章 中文叙词表的本体化升级及形式化语义描述 / 22

2.1 现有中文叙词表按照统一方法升级为本体的可行性	22
2.2 中文叙词表本体(OntoThesaurus)的构建	23
2.2.1 OWL 简介	23
2.2.2 TBox 设计及转换	26
2.2.3 ABox 转换	31
2.3 OntoThesaurus 与 SKOS 的深层次比较	32
2.3.1 SKOS 和 OntoThesaurus 简介	32
2.3.2 在目标层次上的比较	32
2.3.3 二者的类与属性的对应关系	33
2.3.4 二者的类与属性的具体比较	36
2.3.5 对 SKOS 的一点使用建议	38

2.3.6 对建立中文 KOS 形式化表示标准体系的建议	38
------------------------------------	----

第三章 OTCSS 的功能机制 / 40

3.1 中文叙词表本体的共享应用机制	40
3.1.1 网络术语学服务的概念	40
3.1.2 中文叙词表本体的网络术语学服务	41
3.1.3 中文叙词表本体的应用程序接口	48
3.2 中文叙词表本体的共建和动态完善机制	56
3.2.1 中文叙词表本体用户知识采集元数据	57
3.2.2 用户知识提取机制研究	60
3.2.3 用户界面和修订专家界面功能设计	60
3.3 中文叙词表本体的一致性检测机制	89
3.3.1 OntoThesaurus 中存在的一致性问题的形式化描述	89
3.3.2 OntoThesaurus 一致性问题的形式化描述	90
3.3.3 OntoThesaurus 一致性检测机制的应用效果	92
3.3.4 可移植性	103

第四章 OTCSS 的设计与实现 / 104

4.1 OTCSS 的需求分析	104
4.2 OTCSS 系统体系结构设计	107
4.3 OTCSS 的整体实现	107
4.4 OntoThesaurus 的构建实现	108
4.5 OntoThesaurus 的检索及获取实现	114
4.5.1 OntoThesaurus 的检索问题	114
4.5.2 面向本体的检索技术	114
4.5.3 SPARQL 检索语言	114
4.5.4 OntoThesaurus 网络检索的实现思路	115
4.5.5 OntoThesaurus 的检索实现方法	116
4.5.6 OntoThesaurus 检索效果测试	117
4.5.7 检索结果中叙词款目信息的获取	118
4.6 OntoThesaurus 一致性检测机制的实现	118
4.6.1 自定义规则	118
4.6.2 运用自定义规则检查一致性	119
4.6.3 OntoThesaurus 一致性检测机制的具体实现	121
4.7 超大型中文叙词表本体的检索与推理	121
4.7.1 研究超大型方案的必要性	121
4.7.2 国内外研究及技术背景	121
4.7.3 超大型中文叙词表本体的检索和推理解决方案	122
4.7.4 超大型 OntoThesaurus 检索和推理方案的设计与实现	126
4.7.5 实验与分析	129
4.7.6 方案的应用	130

4.8 中文叙词表本体的自动构建 131

第五章 系统与应用 / 132

5.1 示范系统 132

 5.1.1 敦煌学检索词表 132

 5.1.2 社会科学检索词表 135

 5.1.3 中国分类主题词表 135

5.2 系统应用可行性方案 137

5.3 结论 138

第一篇附录 139

一、主要缩写和术语表 139

二、现有中文叙词表目录 144

第一篇参考文献 155

第二篇

中文知识组织系统形式化语义描述标准体系 / 161

引言 162

第六章 知识组织系统语义描述的研究基础 / 163

6.1 叙词表（主题标题表）SKOS 描述相关研究 164

 6.1.1 AGROVOC 164

 6.1.2 LCSH (Library of Congress Subject Headings) 165

 6.1.3 对中文知识组织系统的 SKOS 转换研究 168

 6.1.4 部分参考资料要点 170

 6.1.5 结论 172

6.2 中文叙词表本体 (OntoThesaurus) 173

6.3 分类法的语义描述研究及中国传统分类法相关理论 173

 6.3.1 分类法的语义描述研究 173

 6.3.2 分类法语义描述研究及中国传统分类法相关理论参考资料要点 175

附录 语义 Web 建模语言基本知识 180

 1. RDF——语义 Web 的基础 181

 2. RDFS 182

 3. RDFS-Plus 182

 4. RDFS-Plus 的“在野”使用：SKOS 和 FOAF 183

 5. OWL 184

 6. OWL 的“在野”使用：NCI 185

参考文献 186