

中文叙词表本体共建共享系统研究¹⁾

曾新红¹ 明 仲² 蒋 颖³ 林伟明¹ 胡振宁¹ 张水英¹

(1. 深圳大学图书馆, 深圳 518060; 2. 深圳大学信息工程学院, 深圳 518060;
3. 中国社会科学院图书馆, 北京 100732)

摘要 本文阐述了中文叙词表本体(OntoThesaurus, 即基于中文叙词表建立的本体知识库)共建共享系统的设计思想和总体结构。描述了中文叙词表转换为 OWL 本体的扩展 TBox 定义, 叙词表文本的 ABox 实例自动转换, OntoThesaurus 的一致性检测机制; OntoThesaurus 在图书情报界及语义 Web 界的广泛共享应用前景, 在共享应用中采集标引员、领域专家和一般检索者知识实现本体共建和动态完善的完整过程。最后对我国叙词表编纂机构快速实现现有中文叙词表(主题词表)的网络化共建和共享服务提出了建议。

关键词 叙词表 本体 中文叙词表本体 本体构造 共建共享 动态完善 网络术语学服务 机辅标引 机辅检索 知识组织系统 语义 Web

Research on OntoThesaurus Co-construction and Sharing System (OTCSS)

Zeng Xinhong¹, Ming Zhong², Jiang Ying³, Lin Weiming¹, Hu Zhenning¹ and Zhang Shuiying¹

(1. Library of Shenzhen University, Shenzhen 518060; 2. College of Information Engineering of Shenzhen University, Shenzhen 518060;
3. Center for Documentation and Information, Chinese Academy of Social Sciences, Beijing 100732)

Abstract This paper presents the design idea and architecture of OTCSS (OntoThesaurus Co-construction and Sharing System), then delineates the TBox definition and ABox automatic conversion for translating Chinese thesaurus to OntoThesaurus in OWL, the consistency checking mechanism of OntoThesaurus, OntoThesaurus' promising applications both in Library community and Semantic Web, the process of realizing co-construction and dynamic updating for OntoThesaurus by collecting knowledge from indexers, domain experts and general users while they're using it. Finally the proposal that the thesauri owners could update existing Chinese thesauri to OntoThesauri and realize their co-construction and sharing services via Internet quickly is given.

Keywords thesaurus, ontology, OntoThesaurus, ontology construction, co-construction and sharing, dynamic updating, networked terminology service, computer-aided indexing, concept retrieval, KOS, semantic Web

1 前 言

我国现有 100 多部经人工精心编纂的综合性或专业性主题词表(叙词表), 几乎覆盖了所有学科领

域, 在图书情报界的标引和检索工作中得到广泛应用。但由于其面向专业人员、共享性差和更新速度慢等原因而未能成为网络中文术语学服务的主流, 有些甚至在图书情报界也渐被淡忘。它们面临着矛盾的处境: 理论上, 网络上的海量信息使得信息检索

收稿日期: 2007 年 7 月 30 日

作者简介: 曾新红, 女, 1968 年生, 毕业于北京大学, 研究馆员, 硕士生导师, 研究方向: 知识组织与知识管理, 数字图书馆相关技术。E-mail: zengxh@szu.edu.cn。明仲, 男, 1967 年生, 博士, 教授, 研究方向: 本体, 软件工程。蒋颖, 女, 1967 年生, 硕士, 研究馆员, 中国社会科学院图书馆副馆长, 研究方向: 文献计量学, 图书馆自动化。林伟明, 男, 1982 年生, 硕士, 研究方向: 计算机应用技术。胡振宁, 男, 1966 年生, 硕士, 副研究馆员, 研究方向: 图书馆自动化。张水英, 女, 1966 年生, 硕士, 副研究馆员, 研究方向: 主题标引。

1) 国家社科基金资助项目(05CTQ001)、国家自然科学基金资助项目(60673122)。

对叙词表的需求和机遇超过以往任何时候 ;另一方面 ,最终用户却认为其太复杂而拒绝使用。这些挑战使叙词表向着两个主要的方向发展 :一是寻找一种变通的方式使受控的叙词表可以被检索者更快捷、更容易、更直观地利用 ;二是系统间的协作意味着我们设计的词表更容易嵌入到诸如目录管理系统、搜索引擎和入口网站等下级应用系统中^[1]。笔者认为 ,形式化[实现 M2M(machine to machine)]、网络化[实现共建共享]是让传统的中文叙词表焕发新生的最佳途径。

目前《中国分类主题词表》第 2 版)等少数词表已实现了电子化 ,但由于价格、知识产权保护、传统观念等因素 ,应用面仍相当窄 ,有必要进一步实现通过 Internet 进行的网络化共建共享。国内外图书馆界在叙词表的形式化表示方面也进行了一些研究 ,一般采用基于 XML 语法的表示语言(如 RDF/RDFS ,DAML + OIL 等)将叙词表的原有结构表示出来^[2] ,由于 OWL 成为 W3C 推荐标准的时间较短 ,目前还鲜见采用 OWL 来表示叙词表的研究。

而在信息技术领域 ,构建和维护实用本体是一项艰巨的任务 ,当前网络上可利用的实用本体(尤其是中文本体)的数量和质量都远不能满足用户的需要。近年来 ,利用已有的知识组织系统(辞典、分类表、分类法、叙词表等)快速构造本体 ,取代从零开始构造本体的方法 ,已逐渐成为业界的共识。例如在生物医学领域 ,国际上已有多个项目利用已有的医学叙词表等知识组织系统来构建生物医学本体(如 UMLS^[3] ,NCF^[4]等)。国内也有一些试验性的研究利用现有词表构造本体^[2]。但目前的研究一般都采用从知识组织系统抽取需要的内容去构建本体 ,很少考虑兼顾这些知识组织系统原有的结构和服务功能^[2~7]。

本系统旨在为中文叙词表的升级、共享和动态完善提供一种富有生命力的、可同时满足人的需求和 M2M 需求的解决方案 ,将叙词表的网络化发展与本体的构建合二为一。本系统将中文叙词表转换为 OWL 本体 ,在本体中保留叙词表的原有结构并可进行子关系扩展 ,再借助本体技术、使用者的集体智慧和知识集成对其进行一致性推理检测、深化扩展和动态更新完善 ,让传统的中文叙词表不仅能更好地为图书情报界服务 ,还能在网络信息环境中发挥出更大的作用 ,成为语义 Web 上中文本体和中文术语学服务的主力。

2 系统总体结构

本系统先将已有的中文叙词表转换为 OWL 本体(初始 OntoThesaurus) ,再运用本体技术对其进行一致性推理检测/修改完善 ,通过网络发布实现全方位共享(网络术语学服务/共享格式版本和传统叙词表格式版本下载服务/嵌入式机辅标引和机辅检索) ,在共享服务中及时采集使用者(标引员/领域专家/一般检索者)的修订意见 ,发送到知识集成和修订中心 ,通过统计分析 ,为本体修订专家提供修订依据 ,辅以本体推理技术 ,半自动实现本体知识库的共建和动态完善 ,并定期发布更新版本供共享使用。系统总体结构如图 1。

3 中文叙词表转换为 OWL 本体

采用基于 XML 的语言来表示词表以实现其网络服务和 M2M 功能已成为国际共识。目前较为对口的可参考标准是 W3C 于 2005 年 11 月发布的基于 RDF 的 SKOS(Simple Knowledge Organisation Systems)工作草案 ,专用于支持在语义 Web 框架内将知识组织系统(如叙词表、分类法等)转换为网络上可应用的 RDF 格式文档。但 SKOS 本身仍处于初级发展阶段(working draft) ,结构较为简单 ,还需要完善和扩展^[8]。笔者认为 ,我们有必要在参考国外已有研究成果的基础上积极尝试其他适合我国国情的、可能的实现方式 ,以获得更多的实践经验 ,而不是消极地等待一个还未成熟的方案成熟之后再移植过来。国际共识本身需要建立在大量的实践基础之上。

OWL 是 W3C 在知识表示语言方面的最新推荐标准 ,它面向 WEB 相对于 XML、RDF 和 RDF Schema 拥有更多的机制来表达语义 ,而又与它们兼容。OWL 能够被用来清晰地表达词表中的词汇含义以及这些词汇之间的关系 ,并具备良好的扩展性。一经推出即得到国际生物医学界的积极响应 ,率先将其应用于生物医学本体的构建 ,目前已积累了一定的实践经验^[4]。本系统采用 OWL 来表示中文叙词表 ,是为了能够利用 OWL 丰富的描述机制和良好推理能力来实现中文叙词表本体的一致性检测、语义关系扩展和未来多个词表本体的映射、集成。采用 OWL 来表示中文叙词表 ,可以直接利用支持 OWL 的工具来实现词表的一致性推理和进行应用系统开

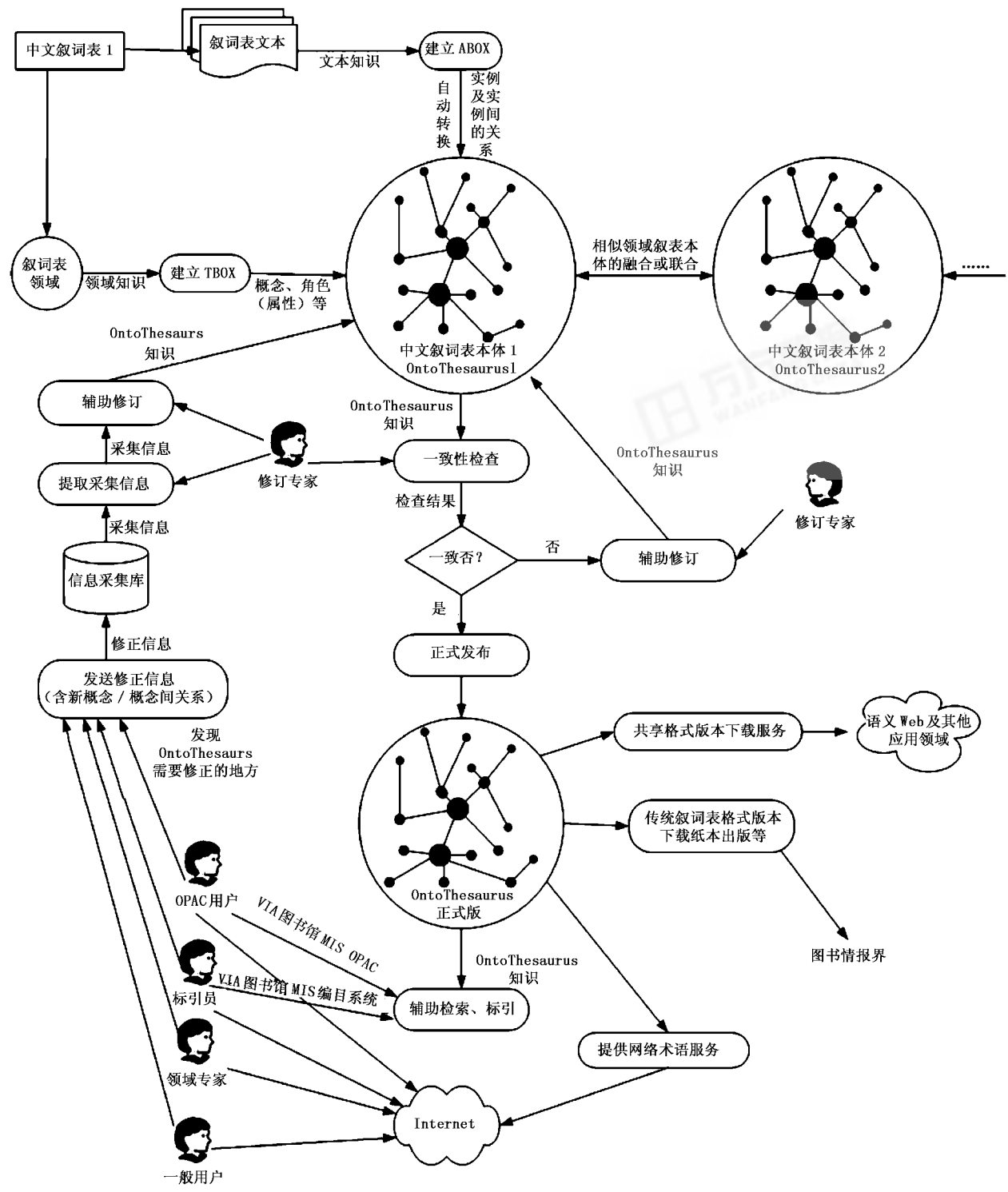


图 1 系统总体结构

发。目前国际上对于 OWL 工具的开发支持力度很强 ,OWL 格式的中文叙词表将具有更好的发展应用前景。

SKOS 的维护机构也在广泛征求使用案例以获得需求来改进 SKOS。从 W3C 于 2007 年 5 月刚刚发

布的工作草案 SKOS Use Case and Requirements^[9]中可以看出 ,使用 OWL 来扩展特殊概念类型、用 OWL 为词表进行编码等在实践案例中具有强烈的需求 ,实现 SKOS 与 OWL-DL 的兼容也被列入了下一版本 SKOS 的候选需求(R-CompatibilityWithOWL-DL)。如

果将来 SKOS 成为了 W3C 的正式推荐标准并实现了与 OWL-DL 的兼容,本系统也很容易将与 SKOS 对应的原语转换成 SKOS 形式并加入必要的 SKOS 元素,提供 SKOS 共享格式版本的下载。

3.1 转换规则

本体的构建、集成和演化很大程度依赖定义良好的语义和强大的推理工具,描述逻辑提供了上述两个方面,是理想的本体构建基础。笔者在现有国内外研究成果的基础上,根据中文叙词表的特点、描述逻辑及 OWL 语法,制定了中文叙词表转换为 OWL 本体的转换/扩展规则(本体知识库的 TBox 部分)其中的具体含义请参见参考文献 [10]。

对于可能就概念(和术语)的精确定义达到共识以及需要定义逻辑规则来处理关系和可能推出新知识的领域,可扩展子关系来定义明确的关系属性。参考 ANSI/NISO Z39.19 - 2005^[11],笔者对等级关系(上位/下位)和相关关系进行了扩展,定义了它们的子关系属性及相应的标识。上下位关系分别扩展出三种子关系:类属关系(the generic relationship),实例

关系(the instance relationship)和整体-部分关系(the whole-part relationship)。相关关系的子关系属性则拥有更大的自由度,可以根据具体的应用环境从列表中进行选择,也可以在共享应用的过程中根据使用者(如领域专家)的要求进行扩展,从而细化概念间的关系粒度,使来自叙词表的粗粒度本体逐渐演变成细粒度本体,满足特定领域的逻辑推理需求。

中文叙词表转换为 OWL 本体的类定义及其说明参见参考文献 [10] 中的表 1(取消 PTerm 类)扩展后的属性定义见本文表 1(限于篇幅,与参考文献 [10] 中表 2 重复部分略)。

表 1 中带“*”号的是扩展的子关系属性。相关关系的子关系属性对应的叙词表标识是由笔者自定义的(粗体英文标识,在相应的标准中尚未明确定义)。在用户界面中,这些扩展关系的详细解释将随鼠标指针指向相应关系名(label)而出现(悬停),以便于用户理解和判断。由于篇幅有限,本表略去了 18 种相关关系子关系的定义(这些子关系定义将会在 05CTQ001 课题的研究报告中详细列出)。

表 1 属性定义

Domain	Property	Range	属性特征	label	叙词表对应标识	
	ObjectProperty				英	中
Concept	Broader	Concept	具有传递性。与 Narrower 互逆。	上位词	BT	属 S
	* BroaderGeneric		Broader 的子属性,表示类属关系(generic relationship)。具有传递性。与 NarrowerGeneric 互逆	类属关系上位词(扩展)	BTG	
	* BroaderInstance		Broader 的子属性,表示实例关系(instance relationship)。具有传递性。与 NarrowerInstance 互逆	实例关系上位词(扩展)	BTI	
	* BroaderPart		Broader 的子属性,表示整体-部分关系(whole-part relationship)。具有传递性。与 NarrowerPart 互逆	整体/部分上位词(扩展)	BTP	
Concept	Narrower	Concept	具有传递性。与 Broader 互逆	下位词	NT	分 F
	* NarrowerGeneric		Narrower 的子属性,表示类属关系(generic relationship)。具有传递性。与 BroaderGeneric 互逆	类属关系下位词(扩展)	NTG	
	* NarrowerInstance		Narrower 的子属性,表示实例关系(instance relationship)。具有传递性。与 BroaderInstance 互逆	实例关系下位词(扩展)	NTI	
	* NarrowerPart		Narrower 的子属性,表示整体-部分关系(whole-part relationship)。具有传递性。与 BroaderPart 互逆	整体/部分下位词(扩展)	NTP	

续表 1

Domain	Property	Range	属性特征	label	叙词表对应标识	
	ObjectProperty				英	中
Concept	TopConcept	Concept		族首词	TT	族 Z
Concept	Related	Concept	参见	RT	参 C	
	* Cause _ Effect		Related 的子属性 ,表示原因/结果(Cause/Effect)关系	原因/结果相关词 (扩展)	RTCE	
	* Effect _ Cause		Related 的子属性 ,Cause _ Effect 的逆属性	结果/原因相关词 (扩展)	RTEC	
	* Process _ Agent		Related 的子属性 ,表示处理/工具(Process/Agent)关系	处理/工具相关词 (扩展)	RTPAg	
	* Agent _ Process		Related 的子属性 ,Process _ Agent 的逆属性	工具/处理相关词 (扩展)	RTAgP	
						
	DatatypeProperty					
Concept	CLCCode	&rdfs ; literal	除非特别注明 ,默认为此分类法类号。CNMARC 对应字段 690	中图法分类号	CLC	
	LCCASCode		中国科学院图书馆图书分类法类号。CNMARC 对应字段 692	科图法分类号	LCCAS	
	UDCCode		国际十进制图书分类法类号。CNMARC 对应字段 675	国际十进分类号	UDC	
	DDCCode		杜威十进图书分类法类号。CNMARC 对应字段 676	杜威十进分类号	DDC	
	LCCCode		美国国会图书馆图书分类法类号。CNMARC 对应字段 680	LC 分类号	LCC	
Concept	PinYin	& rdfs ; literal	Cardinality = 1 (即只能出现一次)	汉语拼音	PY	
Concept	EngCounterpart	&rdfs ; literal		英译名		E
Concept	ScopeNote	&rdfs ; literal		范畴注释	SN	注 :

在实际转换中 ,不同的中文叙词表可以根据其自身的特点选用其中的某些定义。分类号属性可以根据需要进行扩展。确定后的 TBox 可利用 Protégé 或其他支持 OWL 格式的本体编辑工具生成 OWL 文件。图 2 为《中国分类主题词表》的 OWL 文件 TBox 部分的片段。

3.2 各种格式中文叙词表的处理及转换(ABOX 自动转换)

我国现有的叙词表大部分以印刷形式存在 ,而

电子版词表的主要形式有 :文本格式(纯文本、word 等)、数据库格式、MARC 格式和 RDF 格式等。在实践中我们发现 ,采用预定的带特殊字符分隔符的纯文本格式作为转换中介格式对大多数词表来说是一种比较理想的方法。现行的文字编辑/数据库软件大都支持到纯文本格式的转换 ,可先利用原编辑/数据库软件系统的功能将各种格式的词表转换为本系统指定的纯文本格式(仅有印刷形式的词表直接输入/扫描转换为该格式) ,然后用统一的转换程序将

```
.....
<owl:Class rdf:about="#Concept">
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >概念</rdfs:comment>
</owl:Class>
<owl:Class rdf:ID="PersonConcept">
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >人物概念</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Concept"/>
  </rdfs:subClassOf>
</owl:Class>
.....
<owl:TransitiveProperty rdf:about="#Broader">
  <rdfs:domain rdf:resource="#Concept"/>
  <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >上位词</rdfs:label>
  <owl:inverseOf rdf:resource="#Narrower"/>
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
  <rdfs:range rdf:resource="#Concept"/>
</owl:TransitiveProperty>
.....
```

图 2 《中国分类主题词表》的 OWL 文件 TBox 部分的片段

其转换为 OWL 文件的 ABOX 部分(即实例部分)。此法成本低、效率高,应用效果良好。若存在 MARC 格式或 RDF 等共享格式的电子数据则转换起来更加容易。图 3 为根据《中国分类主题词表》(一版)的带分隔符纯文本电子版(1996)转换的 OWL 文件 ABox 部分的片段。

4 中文叙词表本体的一致性检测机制

我国现有的中文叙词表一般都是人工编制的(近年来少量词表借助计算机进行简单的一致性判断),难免存在结构和逻辑上的错误。有必要在发布共享之前借助本体工具,对转换后的初始本体进行全面而严格的一致性检测,并根据检测结果对其逻辑和层次结构进行完善。在中文叙词表本体的共建

和动态完善过程中也有必要在新版本正式发布前进行一致性检测,以使错误率降至最低。结合中文叙词表和本体的特点、编制规范和描述逻辑,本系统建立了针对 OntoThesaurus 的一致性检测机制,制定了 9 条自定义规则,采用 Jena 对 OntoThesaurus 进行基于自定义规则的推理,可检测出值域不一致、语义不一致、约束冲突以及信息缺失、逻辑错误等问题,并提供网络界面,供修订专家根据推理结果进行相应的处理。限于篇幅,此问题将另文专述,详见《中文叙词表本体一致性检测机制的研究与实现》一文^①。

5 中文叙词表本体的共享应用

本系统的产品 OntoThesaurus 在图书馆界和语义 Web 界都具有广泛的应用前景,可应用于检索目的

```
.....
<Concept rdf:ID="知识">
  <CLCCCode rdf:datatype="http://www.w3.org/2001/XMLSchema#string">[B023]</CLCCCode>
  <CLCCCode rdf:datatype="http://www.w3.org/2001/XMLSchema#string">G302</CLCCCode>
  <Related rdf:resource="#基本知识"/>
</Concept>
<CompoundConcept rdf:ID="知识-传播">
  <CLCCCode rdf:datatype="http://www.w3.org/2001/XMLSchema#string">G2</CLCCCode>
</CompoundConcept>
<Concept rdf:ID="知识表达">
  <CLCCCode rdf:datatype="http://www.w3.org/2001/XMLSchema#string">TP18</CLCCCode>
  <TopConcept rdf:resource="#知识工程"/>
  <Related rdf:resource="#产生式系统"/>
  <Related rdf:resource="#环境知识"/>
  <Related rdf:resource="#人工智能"/>
</Concept>
<Concept rdf:ID="知识产权">
  <CLCCCode rdf:datatype="http://www.w3.org/2001/XMLSchema#string">[D912.1]⑦</CLCCCode>
  <CLCCCode rdf:datatype="http://www.w3.org/2001/XMLSchema#string">D913⑦</CLCCCode>
  <CLCCCode rdf:datatype="http://www.w3.org/2001/XMLSchema#string">D(9)5⑦</CLCCCode>
</Concept>
.....
```

图 3 《中国分类主题词表》的 OWL 文件 ABox 部分的片段

① 文中提到的其他论文也是本课题的系列成果论文,敬请关注。

(如实现多方位的扩展检索),分类与标引目的、语言学目的、人工智能目的(本体推理)等。本系统目前将重点放在前两种目的的应用开发上,提供网络术语学服务、内嵌式机辅标引/机辅检索服务(Web Service接口)。此外,本系统还可以提供多种共享格式版本的下载服务,以便为更广泛的应用提供机器可理解的术语学工具。

5.1 网络术语服务(Terminology Service)

广义的网络术语服务可以是M2M的,也可以是交互式的、面向用户的服务,可应用于检索(retrieval)过程的所有阶段。这种服务通过网络(如经由术语服务注册中心)展示和应用词表(包括受控和非受控),其目的包括:检索、浏览、发现、翻译、映射、语义推理、主题标引和分类、收割、警告(如提示新的或修改过的术语状态)等。这种服务可以作为最终用户的直接元素使用,也可以根据应用环境支持场景之后的服务^[3]。

本系统为OntoThesaurus开发的网络术语服务专指交互式的、面向用户的服务,功能包括:浏览、检索、将入口词(自由词)转化为标引词(受控概念)、查询扩展(英文扩展、同义词扩展、上/下位词扩展、指定关系扩展等),获得。用户利用这些服务,可以发现、获得所需的概念术语,并可复制到任何界面(如编目标引、OPAC检索、搜索引擎、数据库检索等界面)进行检索、查询扩展、查询重构、分类标引、翻译等工作。以上功能的实现依赖于OntoThesaurus的检索功能和推理功能的实现。本系统基于SPARQL本体检索语言实现了OntoThesaurus的一般检索和高级检索,并通过推理功能可获取选中主题词、相应入口词(同义词扩展)、相应英译名(英文扩展)、上下位词等,详见《中文叙词表本体的检索实现及其术语服务研究》一文^①。

内嵌式机辅标引/机辅检索是M2M的、应用场景之后的术语学服务,即将上述服务功能嵌入图书馆信息管理系统编目模块(主题标引)和OPAC检索模块,并与这些模块产生信息交互,为标引员和检索用户提供更直接、更全面的服务。根据我国图书馆信息管理系统的使用现状和信息技术的发展趋势,本系统拟开发通用的Web Service服务接口。如有必要,也将开发用于C/S系统的Java程序接口以及用于B/S系统的Java Web应用程序接口^[12]。

5.2 共享格式版本和传统叙词表格式版本下载/打印服务

本系统可提供OWL、SKOS等共享格式版本的下载服务,以满足其他M2M或面向用户的交互式服务的应用需求(直接使用或经二次开发)。例如可用于搜索引擎的扩展检索(目前搜索引擎的术语服务还大都停留在简单的同义词环层次上),自动分类和标引、文本挖掘和信息抽取、自动映射、术语或文献的自动翻译、语义推理等。也可用来对Web2.0的以用户为中心的语义标注和内容管理进行适当的规范控制。

为兼顾传统叙词表用户的使用习惯和某些欠发达地区的使用要求,本系统也拟提供传统叙词表格式版本的下载、打印功能。

6 中文叙词表本体的共建与动态完善

本系统通过提供Internet网络界面(需注册),在中文叙词表本体的共享使用过程中及时采集用户在使用本体时动态(瞬间)产生的修改意见。这种方法能够快速反映科学技术和社会的发展变化,其时效性远胜于对静态标引数据进行统计的方法(如《中国分类主题词表》二版修订时采用的方法)。可视作应用Web2.0技术中的社会标记法(social tagging)和民间分类表(或称自由分类法,folksonomy)^[3,13,14]的类似方法来改进词表的升级/维护工作的具体实践,可以为评估社会标记法和民间分类表方法在创建、升级和维护词表中可能扮演的角色提供实践经验。

参考《中国分类主题词表》二版的修订规则^[15],本系统将需要采集的用户知识(修改意见)分为以下五种类型(其中所用术语尽量便于各领域人士理解):

- (1) 新增补叙词(正式主题词/概念);
- (2) 为原叙词增加入口词(同义词);
- (3) 修改原叙词款目信息;
- (4) 原叙词款目整条删除;
- (5) 新增相关关系子关系种类(向细粒度本体演化)。

笔者将OntoThesaurus的用户大致分为以下三类。

标引员:在主题标引第一线工作的标引员(编目

① 文中提到的其他论文也是本课题的系列成果论文,敬请关注。

员)是叙词表的专业使用者,应用叙词表进行主题标引是他们的日常工作,在主题分析过程中经常会发现新的概念术语,现有的机制规定只能采用靠词标引等非直接方式来标引这些含新概念的文献。若采用增词标引则只能是本地的,这会引起与规范叙词表的不统一,因为没有一个好的机制来接收分析这些新的概念术语并及时修订叙词表。笔者曾做过六年的主题标引工作,对此颇有感触,深为这些白白浪费掉的宝贵的修订意见而惋惜。标引员在工作中可以接触大量的新文献,其主题分析过程是一个丰富的知识分析和发现过程,易于发现新概念和概念间的具体关系,以及词表中存在的错误,他们应是中文叙词表本体共建的主力军。笔者认为,可以仿照CALIS联合编目上载数据和编目员上载资格认证的机制来管理和激励标引员的共建行为。

领域专家对于某些专业性很强的领域(如生物医学、敦煌学等),领域专家的作用非常重要。中文叙词表本体的共享应用可以为领域专家的研究工作带来很大的方便,他们的共建参与也可为中文叙词表本体的完善贡献宝贵的领域知识。另外,增加定义相关关系的子关系是将叙词表扩展为细粒度本体的关键手段,此时领域专家的意见尤其值得重视。

一般用户:可以贡献大量的入口词,能够“无主要信息损失地”反映用户语言。

用户的修订意见发送之前需通过系统的检测,以降低无效信息的干扰。成功发送的修订意见被知识集成服务器统一接收,经分类统计分析,可按频率(接收频率/发送者频率/标引频率)排序,提供给修订专家参考或直接写入新版本。人工干预和自动处理之间的适当平衡需要谨慎的考虑,系统在多个阶段提供这两种选择。叙词表是受控的,而本体更是具有当前知识组织系统中最准确和最形式化的关系定义,因此修订专家的人工干预是必要的。借助本体推理技术推理出来的隐含知识,如隐含于关系中的新概念、推出的互逆/对称关系等,在系统运行的初级阶段也应经过人工判断后才能写入,待本体相对成熟之后方可设为自动写入。修订专家也有可能出错,因此新版本发布之前应该进行一次全面的一致性检测和修正,以将错误率降至最低(参见第4节)。

当然,如果全程选择自动处理,得到的将是一个完全以用户为中心的知识库(类似于 folksonomy),这在某些应用背景下也是可行的。并且,本系统也可以用来从零开始建设全新的中文叙词表本体及其共

建共享服务。

上述所有工作都可以通过 Internet 远程进行,通过注册控制权限,具有极大的灵活性。本系统的设计与实现详见《中文叙词表本体共建共享系统 OTCSS 的设计与实现》一文^①。图4为供用户使用的界面(可使用叙词表本体检索、获得和新词/关系采集发送等功能),图5为供修订专家使用的界面(可使用知识提取、词表管理、词表全局检查和词表发布等功能)。

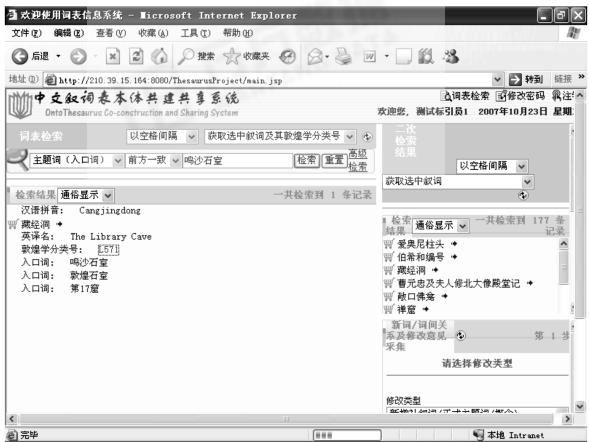


图4 用户界面

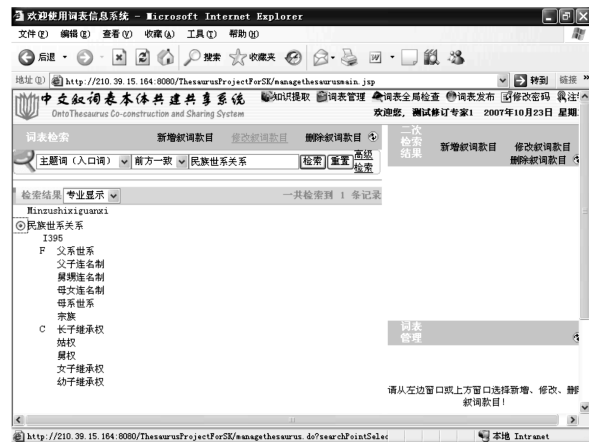


图5 修订专家使用的界面

7 对我国现有中文叙词表实现网络化共建共享的建议

本系统目前以《敦煌学检索词表》、《中国分类主题词表》(一版局部)和《社会科学检索词表》(局部)为例实现了原型系统,主要功能已接近实用水平,并正在寻求与其他叙词表编纂机构的合作,以获取更

多的需求来改进系统,使其具有更广泛的适应性。本系统的实践表明,我国现有的中文叙词表可以快速地实现本体化升级和网络化共建共享。OCLC Terminology Service Project^[16]的做法值得我们借鉴。

本文所提出的中文叙词表本体共建共享方案是一种低投入、高效率的可自增值方案,可以充分利用我国图书馆界传统的知识组织系统建设优势、现有的大量标引员/领域专家等专业人才储备,借助目前在我国广泛运行的图书馆自动化系统和其他信息服务系统进行融入用户环境、可持续、日常性的中文本体建设。在共建中共享,在共享中共建,换一种开放的网络经济模式来发展我们的事业,也许会让我们整个行业获得新生。

8 结 语

知识组织系统(叙词表、分类法、规范档等)是图书馆界几十年智慧的结晶,是图书馆界最值得骄傲的宝贵财富。许多手工制作的词表可能有这样那样的缺陷,但毕竟凝结了图书情报专家或领域专家若干年甚至数百人年的脑力劳动,是极富价值的。目前,在 IT 界构建本体、Taxonomy 等知识组织系统的过程中,领域专家构建的样本常用作评估机器自动构建结果的标准,图书馆界的分面分析方法、同义词聚合和概念间的关系识别方法等仍然被认为是极有价值的方法。几乎毫无规范的 Folksonomy 都可以很好地提供服务,何况精心制作的叙词表。因此我们应该解放思想,顺应网络信息时代的要求,让中文叙词表走出图书馆,为广大用户提供更直观、更快捷、更先进的网络中文术语服务,并利用使用者的集体智慧实现其共建和动态更新完善。

任何一项理论和规范都需要在实践中验证,才能逐渐完善。目前国际上语义 Web 界的研究普遍缺乏实践支持,亟需大量的实践来推动理论的进展和规范的成熟。本系统的研究和实践可以为国内外学术界提供中文本体构建和网络术语服务的实践经验,缩短我国知识组织系统建设与国际最新规范建设之间的距离,实现共同发展,共同进步。

参 考 文 献

[1] Jean Aitchison, Stella Dextre Clarke. The Thesaurus : A Historical Viewpoint, with a look to the Future.[2005-07-

01]. <http://www.haworthpress.com/web/CCQ>.

- [2] 唐静. 叙词表转换为 Ontology 的研究 [J]. 情报理论与实践, 2004(6) : 642-645.
- [3] Udo Hahn, Stefan Schulz. Building a Very Large Ontology from Medical Thesauri. Handbook on Ontologies, Springer-Verlag, 2004 : 133-149.
- [4] Jennifer Golbeck, et al. The National Cancer Institute's Thesaurus and Ontology. [2004-03-16]. http://www.mindswap.org/papers/webSemantics_NCI.pdf.
- [5] Douglas Tudhope, Traugott Koch, Rachel Heery. Terminology Services and Technology : JISC state of the art review. 15-09-2006 [2006-10-17]. <http://nkos.slis.kent.edu>.
- [6] 刘耀, 穗东方. 领域 Ontology 概念描述体系构建方法探析 [J]. 大学图书馆学报, 2006(5) : 28-33.
- [7] Aldo Gangemi, et al. An overview of the ONIONS project : Applying ontologies to the integration of medical terminologies. Data & Knowledge Engineering, 1999, 13 : 183-220.
- [8] SKOS Core Guide : W3C Working Draft 2 November 2005. [2007-06-08]. <http://www.w3.org/TR/2005/WD-swbpskos-core-guide-20051102>.
- [9] SKOS Use Cases and Requirements : W3C Working Draft 16 May 2007. [2007-06-08]. <http://www.w3.org/TR/2007/WD-skos-ucr-20070516>.
- [10] 曾新红.《中国分类主题词表》的 OWL 表示及其语义深层揭示研究 [J]. 情报学报, 2005, 24(2) : 151-160.
- [11] ANSI/NISO Z39.19-2005. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. Developed by National Information Standards Organization, approved July 25, 2005 by the American National Standards Institute.
- [12] 肖洪, 余锦凤. 基于 Ontology 的数字图书馆知识管理系统的设计与实现 [J]. 现代图书情报技术, 2006(3) : 20-26.
- [13] 刘炜, 葛秋妍. 从 Web2.0 到图书馆 2.0 : 服务因用户而变 [J]. 现代图书情报技术, 2006(9) : 8-12.
- [14] 毛军. 元数据、自由分类法(Folksonomy)和大众的因特网 [J]. 现代图书情报技术, 2006(2) : 1-9.
- [15] 卜书庆.《中国分类主题词表》的修订技术与规范 [J]. 国家图书馆学刊, 2006(4) : 68-74.
- [16] OCLC. Terminology Services. [2007-05-05]. <http://www.oclc.org/research/projects/termservices>.

(责任编辑 芮国章)

作者：[曾新红](#)，[明仲](#)，[蒋颖](#)，[林伟明](#)，[胡振宁](#)，[张水英](#)，[Zeng Xinhong](#)，[Ming Zhong](#)，[Jiang Ying](#)，[Lin Weiming](#)，[Hu Zhenning](#)，[Zhang Shuiying](#)

作者单位：[曾新红](#)，[林伟明](#)，[胡振宁](#)，[张水英](#)，[Zeng Xinhong](#)，[Lin Weiming](#)，[Hu Zhenning](#)，[Zhang Shuiying](#) (深圳大学图书馆, 深圳, 518060)，[明仲](#)，[Ming Zhong](#) (深圳大学信息工程学院, 深圳, 518060)，[蒋颖](#)，[Jiang Ying](#) (中国社会科学院图书馆, 北京, 100732)

刊名：[情报学报](#) [ISTIC](#) [PKU](#) [CSSCI](#)

英文刊名：[JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION](#)

年，卷(期)：2008，27(3)

引用次数：0次

参考文献(18条)

1. [Jean Aitchison](#). [Stella Dextre Clarke](#) [The Thesaurus: A Historical Viewpoint, with a look to the Future](#) 2005

2. [唐静](#) [叙词表转换为Ontology的研究](#) [期刊论文] - [情报理论与实践](#) 2004 (6)

3. [Udo Hahn](#). [Stefan Schulz](#) [Building a Very Large Ontology from Medical Thesauri](#) 2004

4. [Jennifer Golbeck](#) [The National Cancer Institute's Thesaurus and Ontology](#) 2004

5. [Douglas Tudhope](#). [Traugott Koch](#). [Rachel Heery](#) [Terminology Services and Technology: JISC state of the art review](#) 2006

6. [刘耀](#). [穗志方](#) [领域Ontology概念描述体系构建方法探析](#) [期刊论文] - [大学图书馆学报](#) 2006 (5)

7. [Aldo Gangemi](#) [An overview of the ONIONS project: Applying ontologies to the integration of medical terminologies](#) 1999

8. [SKOS Core Guide: W3C Working Draft 2 November](#) 2005 2007

9. [SKOS Use Cases and Requirements: W3C Working Draft 16 May](#) 2007 2007

10. [曾新红](#) [《中国分类主题词表》的OWL表示及其语义深层揭示研究](#) [期刊论文] - [情报学报](#) 2005 (2)

11. [ANSI/NISO Z39.19-2005. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies](#) 2005

12. [肖洪](#). [余锦凤](#) [基于Ontology的数字图书馆知识管理系统的设计与实现](#) [期刊论文] - [现代图书情报技术](#) 2006 (3)

13. [刘炜](#). [葛秋妍](#) [从Web 2.0到图书馆2.0: 服务因用户而变](#) [期刊论文] - [现代图书情报技术](#) 2006 (9)

14. [毛军](#) [元数据、自由分类法\(Folksonomy\)和大众的因特网](#) [期刊论文] - [现代图书情报技术](#) 2006 (2)

15. [卜书庆](#) [《中国分类主题词表》的修订技术与规范](#) [期刊论文] - [国家图书馆学刊](#) 2000 (4)

16. [OCLC Terminology Services](#) 2007

17. [文中提到的其他论文也是本课题的系列成果论文, 敬请关注](#)

18. [文中提到的其他论文也是本课题的系列成果论文, 敬请关注](#)

相似文献(9条)

1. 期刊论文 [曾新红](#). [林伟明](#). [明仲](#). [Zeng Xinhong](#). [Lin Weiming](#). [Ming Zhong](#) [中文叙词表本体一致性检测机制研究与实现 - 现代图书情报技术](#) 2008 (5)

研究中文叙词表本体(OnoThesaurus, 即基于中文叙词表建立的本体知识库)的一致性检测机制, 并将其应用在中文叙词表本体共建共享系统(OTCSS)的修订意见提交、叙词表本体更新和全局检查等相关过程的实现中, 取得了良好的应用效果。

2. 学位论文 [谷建军](#) [基于叙词表的中医古籍文献领域本体建模方法研究](#) 2006

1. 前言随着20世纪90年代中医药文献数字化研究的开展, 中医古籍文献数字化工作已经走过了几个阶段。从2000年国家中医药管理局设立的重点研究专项“中医药古代文献资源数字化关键问题研究”的起步阶段, 到2001年国家科技部基础工作重大项目“中医药科技信息数据库建设”项目, 再至2003年国家科技部医学科学数据共享服务系统“中医药学科学数据共享服务系统”建设项目, 中医古籍文献数字化已成功研制出“中医本草文献数据库”、“中医方剂文献数据库”, 在全国三十余家中医院校和研究机构的参与下, 成功构建了我国第一个中医古籍文献知识库, 目前已收录了本草、方剂类古籍260余种, 6000余万汉字, 并于2003年实现了网络运行。在数字化工作的研究中, 导师柳长华教授提出的基于“知识元”的中医古籍计算

机知识表示方法在知识库建设中取得了进展, 基本形成了一套较成熟的建库技术。以这种技术建立的数据库使知识的查询更加精确, 避免了大量冗余信息的出现, 使用户最大限度地摆脱了信息爆炸的困扰。但随之而来的另一个问题又出现在查询者面前, 这就是所谓的“信息孤岛”现象。古籍数字化的功能不仅在于一般的信息查询, 更重要的是古籍文献中的知识发现。普通的数据库难以达到知识挖掘的深层次要求, 古籍数字化的目标是建设知识库。

2. 知识库系统的原理从知识的使用角度来看, 知识库是由知识和知识处理机构组成, 知识库形成一个知识域, 该知识域中除了事实、规则和概念之外还包含各种推理、归纳、演绎等知识处理方法。知识库系统的核心组成部分是知识库和推理机构。知识库对知识进行存储和管理, 推理机构是推理机使用知识库内的知识执行推理的机构。如果一个系统具有能用计算机所存储的知识对输入的数据进行解释和推理, 并有对其进行验证的功能, 则该系统称为知识库系统。知识库系统的实现涉及到两个关键问题: 知识表示和知识推理。知识库的处理过程分为二个层面: 先将知识由底层数据经过一系列加工, 如分类、归纳、综合等处理过程而得到上层信息, 称为知识表示。这种信息再经过解释、比较、推理得到我们所获取

的知识，即知识推理的过程。为了实现知识推理，一种基于本体的知识表示方法成为各个领域构建知识库推理系统的首选。3. 本体的概念、作用与分类本体(Ontology)起源于哲学领域，古希腊哲学家亚里士多德(Aristotle)定义Ontology为“对世界客观存在物的系统的描述，即存在论”。Ontology是客观存在的一个系统的解释或说明，它关心的是客观现实的抽象本质。Ontology这个哲学范畴，被人工智能界赋予了新的定义，从而被引入信息科学中。目前普遍接受的本体定义为：共享概念模型的形式化规范说明。从内涵上来看，本体是领域(可以是特定领域的，也可以是更广的范围)内部不同主体(人、机器、软件系统等)之间进行交流(对话、互操作、共享等)的一种语义基础，即由本体提供一种明确定义。Ontology自身所要实现的目标，即：“在人类和应用系统之间实现共享和相互理解”。Ontology能够将领域中的各种概念及概念之间的关系显式化、形式化地表达出来，从而将术语的语义表达出来，因而在语义查询方面发挥着重要作用。自W3C主席TimBermee-Lee在1998年首先提出了语义web的概念之后，Ontology正在成为人工智能和信息处理领域的研究热点之一。本体强调相关领域的本质概念，同时强调这些概念之间的关系。本体论可以有效地表达知识和知识之间的关系，基于本体论的知识库系统可以建立有靠的知识表达体系，揭示知识之间的内在关系。本体技术主要在以下几个方面提高知识库系统的性能：可重用性、知识获取、查找智能性、可靠性、规范定义、任务解析、可维护性。本体通常可分为以下几类：领域本体、通用本体、应用本体、表示本体。本文关注的是本体类型中的领域本体，主要讨论如何运用Ontology技术构建中医古籍领域本体。4. 本研究的意义、方法与创新点本文通过对本体的国内外研究与发展现状的考察，根据中医古籍数据库的实际情况，在知识推理层面提出了建设面向中医古籍数据库应用的中医古籍文献领域本体的设想。参考国内外领域本体的建设方法，论述了利用叙词表建设领域本体的优势，提出了基于叙词表的适合中医古籍数据库应用的中医古籍文献领域本体建设方法。最后通过一个实例阐述了中医古籍文献领域本体的具体建设方法，为中医古籍数据库的进一步建设提供了理论与实践的双重参考。

研究意义：中医古籍知识库建设的要求；中医古籍知识深入整理研究的要求；便于网络中医古籍文献资源的统一管理。研究方法：文献调研法、概念分析法、本体构建法。创新点：在中医古籍文献数字化领域提出建立本体系统的设想；分析了适合中医古籍文献数据库的本体表示语言和编辑工具；提出中医古籍文献领域本体的建设目标；设计了中医古籍文献领域本体的建设方法；建立了一个以“病证”概念为核心的中医古籍文献领域本体模型。5. 本体的国内外研究现状国外主要研究现状：①理论深化研究；②信息系统中应用；③本体作为一种能在知识层提供知识共享和复用的工具在语义网中的应用。国外较为知名的本体知识系统：WordNet、FrameNet、GUM、SENSUS、OntoSeek、Cyc、HowNet和SUMO等。国内主要研究现状：我国本体的研究尚处于起步阶段，一个是对W3C发布的关于本体的外文资料的翻译，一个是主要为面向应用的研究，无论是理论还是实际应用都相对落后于国外。面向中医药领域的研究主要有：浙江大学网格计算实验室开发的基于语义的中医药信息本体虚拟组织模型——DartGrid服务栈；北京中医药大学和中国科学院计算机研究所开发的基于本体的中医专家临床病案知识库。6. 领域本体的构建20世纪50年代叙词表得到了很大发展，成为主题检索的主要语言，各国拥有的叙词表数以千计，并涵盖了各个领域。从一定意义上讲，叙词表可以说是一种轻量级本体(Light-weightOntology)。基于叙词表构建领域本体有诸多的优越性，目前人工智能界普遍推荐利用叙词表构建领域本体。中医古籍文献叙词表与本体的关系：中医古籍文献叙词表反映了中医古籍文献中包含的概念，概念来自于古籍内容与古籍本身，是对中医古籍文献的客观反映。叙词表表示的是树状结构，这种树状结构表示了古籍文献内部的自然构成方式。叙词表的结构是可见的、清晰的，可称为显性结构。领域本体继承了叙词表的树状结构特征。本体更重在表示一种概念之间的隐含关系，这种关系是模糊的、不明显的，可以称为隐性结构。相对来说，本体的反映更细致，更深入，为文献中的知识关联提供了可实现的途径。叙词表或本体是对于体现古籍内涵的概念的集合。领域本体的建模元语：(概念)类、属性、函数、公理、实例。建模语言：选用OWL语言。本语言的优势在于：基底层语法符合XML标准格式；为W3C推荐的标准本体编辑语言，便于与数据库之间的数据交换；支持多种语言输入，并支持中文；网络中有免费教学手册，便于下载学习。编辑工具：选用Protégé-2000。其优势在于：界面友好，具有图形化的用户界面；版本更新速度快，目前已发布了3.1.1版；支持多种语言格式，支持中文编辑；本体文档可以不依赖于本体编辑器进行代码修改，方便与数据库的连接；网络开放资源：是W3C推荐的本地编辑器；是基于XML的本体标记语言，多种存储格式，可以适应不同需要。构建方法：选用斯坦福大学医学院开发的七步法。7. 中医古籍文献领域本体模型(病证模型)的构建元数据(Metadata)就是数据之数据，或描述原始数据的独立数据。元数据是针对网络信息标引发展起来的，它以Web页为背景，通过元数据将Web信息组织起来，构成基于元数据的有序信息系统，为网络信息资源的组织提供了重要手段。其主要学术意义和应用价值在于信息处理。根据中医文献数字化研究室的最新研究，中医药古籍元数据包括三类概念：一是表达古籍外部特征的元数据，称为书目元数据；二是表达古籍内部篇、卷、章、节层次特征的元数据，称为书体结构元数据；三是表达古籍知识单元内容的元数据，称为语义元数据。本领域本体模型以“语义元数据”为核心概念集，以“病证”语义元数据及其包涵的概念为中心建立本体模型。

有关病证与其他概念间的关系主要有二类：等级关系，包括上下位关系和实例关系；非等级关系，包括同义关系、交叉关系、排斥关系等。以《诸病源候论》“风痉候”为例，为本体添加类和实例：“风痉候”条文：“风痉者，口噤不开，背强而直，如发痫之状。其重者，耳中策策痛；卒然身体痉直者，死也。由风邪伤于太阳经，复遇寒湿，则发痉也。诊其脉，策策如弦，直上下者，风痉脉也。”“风痉候”的概念等级链为：病证——风病——风痉。条文中与与本候概念相关的其他概念有：证候表现、预后、病因、病位、脉象。添加到本体中，如图所示：8. 讨论中医古籍文献领域概念十分丰富，概念间关系错综复杂，难以在短时间内完成本体系统的建设，应根据实际需要分阶段完成。本文将中医古籍文献领域本体的研究目标分为二个阶段：长期目标：建立相对完整的中医古籍文献领域本体系统平台。建立本体的中英文对照词表，便于与世界接轨。短期目标：根据数据库建设的需要，分别以本草、方剂、病证为中心概念，开始本体系统的建设。

3. 期刊论文 曾新红.Zeng Xinhong 中文叙词表本体——叙词表与本体的融合 -现代图书情报技术2009(1)从网络信息社会对知识组织系统的需求、来自信息科学界和其他相关各界的应对发展现状等方面,详细阐述实现中文叙词表的形式化表示和网络应用的重要性 and迫切性.对叙词表和本体的概念进行深入的比较研究,论证将他们合二为一的可行性.阐述直接采用OWL(而不 用SKOS)表示中文叙词表本体(OntoThesaurus)的原因,并列出具体的类定义和属性定义.中文叙词表本体共建共享系统OTCSS的多项功能和若干原型系统的实现,证明这些定义的科学性、可行性和通用性.

4. 期刊论文 曾新红.林伟明.明仲.Zeng Xinhong.Lin Weiming.Ming Zhong 中文叙词表本体的检索实现及其术语服务研究 -现代图书情报技术2008(2)在简单介绍中文叙词表本体共建共享系统OTCSS项目背景的基础上,阐述实现中文叙词表本体网络术语服务(OntoThesaums -TS)的意义.详细描述OntoThesaums的检索实现方法,以及其术语服务应用场景典型案例,并对OntoThesaurus的术语服务提出进一步研究计划.

5. 学位论文 付佳佳 基于叙词表的领域本体建模研究 2006众所周知，叙词表是一种为解决信息主题排序而创造的人工语言，它的本质是对自然语言中的词汇进行选择、规范，并揭示其间相关关系，由此形成受控词汇的集合，它的出现主要是为了解决大量的文献如何被方便科学检索的问题。然而，WWW是当今主要的网络信息的集散地，不仅汇聚了海量的信息，而且信息数量正在以指数级的速度增长。随着数据量的激增，WWW上大量分布的无结构和半结构化数据日益加剧信息检索的困难，因此，如何组织海量的数字信息，并为用户提供精确高效的网络检索服务成为重要而迫切的研究课题，这引起了人们对传统知识组织工具如叙词表、分类表等在网络环境中适应性的争论。尽管叙词表和分类法等传统知识工具已开始在网上发展，但是对机器语言来说，其互操作性和表达性仍比较差，为此人们提出了本体这种能在语义和知识层次上描述信息系统的概念模型建模工具。领域本体构建的重要意义主要体现在：首先，领域本体的目标是捕获相关领域的知识，确定该领域内共同认可的词汇，并从不同层次的形式化模式上给出这些词汇之间相互关系的明确定义。从而实现人们对同一客观事物的共识，形成一个统一的认识事物的标准。即为人类认识活动构建顶层概念框架。其次，本体更加突出知识共享的功能，尽管二者都对概念间等级关系、相关关系进行了揭示，但本体更着眼于给出人类事物认识的知识(或领域知识)总框架，因为在本体中的一个实例中每个概念都有其属性信息、实例信息，而这些在词表系列中则少有展示，很多已经涉及到专业词典中的知识，因此说一个本体是一个人类知识(或领域知识)体系的汇总毫不夸张。最后，本体的出现还是为了设计一种机器可以理解的语言。通过本体可以克服计算机系统之间的语义鸿沟，实现某个领域内不同主体(人、机器、软件系统等)之间的对话、互操作、知识共享等目的，于是它被认为是一种共享的概念模型的形式化的规范说明。其中形式化就是指应该是机器可读(可理解、可操作)的意思，而这也成为了在计算机网络环境下应用研究的主题之一。领域本体的构建体现了目前的趋势，但是原本属于本领域的概念信息是丢失还是融合?这是本文探讨的问题。笔者认为，由于叙词表和领域本体之间有许多相同和不同之处，使得基于叙词表来构建领域本体具有一定的优越性。由于某学科领域的叙词表包括本学科领域中相对比较完整的术语(叙词)，因此这些术语(叙词)可以为本领域本体中的概念的创建提供指导；另外，叙词表中的限义词、涵义注释、等级关系、词间关系，为领域本体中概念的属性、实例以及关系的创建可以提供线索和指导，这些指导将为本领域本体的创建者节省大量的时间和精力。基于叙词表构建的领域本体至少在本领域的概念方面应该还是比较完整的。叙词表可以说是在图书情报界为信息检索提供的知识财富，其作用和原理与本体有异曲同工之妙。如果能利用现存的叙词表，将其转换为相应的领域本体，必将使领域本体的建立事半功倍。本文在第一章中，研究了本体在改善对知识管理方面的作用，论述了建立领域本体这一课题的意义，阐述了本文的研究内容和本文的章节安排。在第二章中，系统地研究了本体的理论，从本体的定义、分类、描述语言和建模工具等方面进行了论述。而在第三章中，研究了叙词表的概念、应用现状，并分析了叙词表在表达语义方面的局限性和本体在此方面的优势。为了提高论述的有效性，本文还以具体的例子来说明了这一点。在第四章，根据前文的论述，总结并分析了叙词表和领域本体的区别与联系，论述了基于叙词表建立领域本体的可行性和优越性。第五章，本文又研究了当前本体建模的主要的方法，并在总结这些方法的特点的基础上，提出了基于叙词表建立领域本体的方法。在第六章，本文通过对食品安全领域本体的建模这一实例，详细地说明了这一方法。在这个实例中，笔者自行开发了一个由叙词表的词间关系向领域本体的概念间关系转化的系统，从而实现了基于叙词表建立领域本体的关键一步，这也是本体的创新之处。第七章，文章对全文作了一个总结，提出了本文的不足之处以及对未来工作的展望。本文采取了文献调查、案例佐证、技术对比等方法，从理论和实践的角度研究了基于叙词表的领域本体的建模。但是，由于国内对于本体的开发方法以及如何构建领域本体研究的较少，对基于现有的叙词表构建领域本体的研究，也还处于起步、探索阶段。同时限于个人的能力和水平，笔者仅对本体及叙词表的理论，基于叙词表构建领域本体的可能性及方法进行了相当粗浅的研究；另外，由于任何一个领域本体的构建都是相当复杂的，而且需要该领域的专家的参与，同时

还要耗费大量的人力和物力，不是一个人在短期内就能完成的，因此，笔者开发的系统还没有广泛地得到验证，所构建的领域本体模型也比较简单，一部分功能还没有完全实现，还需要进一步的完善。这些都是笔者将来要做的工作。

6. 学位论文 [鲜国建](#) [农业科学叙词表向农业本体转化系统的研究与实现](#) 2008

在当今信息时代和知识经济社会，信息资源已成为重要的战略资源，在国家科技进步与创新、经济和社会可持续发展过程中发挥越来越重要的作用。现代信息技术和通信技术为信息的收集、加工、存储、传输和利用提供了强有力的技术保障，信息资源呈指数级增长。大量的信息给人们的工作、生活和学习提供丰富的信息资源的同时，又使人们淹没在信息的海洋之中。如何组织、管理和维护海量的信息资源并为人们提供高效优质的信息服务成为一项重要而迫切的任务。本体(ontology)作为一种能在语义 and 知识层次上描述信息系统的概念模型建模工具，为解决这一问题提供了新的途径，已受到国内外研究人员的广泛关注，成了研究的热点，而本体构建也是其中一个重要的研究方向。本文对本体和叙词表的相关知识进行了详细论述，分析得出了叙词表向本体转化的必要性和可行性，并使用当前最新的本体描述语言—网络本体语言(WebOntologyLanguage，简称OWL)，成功地将《农业科学叙词表》(以下简称《农表》)中的叙词(包括正式叙词及非正式叙词)及词间关系进行了表示和描述。在此基础上，设计和实现了一个转化系统，能够自动批量地将词表中的知识结构和语义关系转化到农业本体中。基于叙词表来构建领域本体，不仅为构建领域本体提供了一种较好的方法，也可以加快本体的构建进程，还能提高本体的科学性、规范性和权威性。本文还在本体应用方面进行了探索，基于转化得到的农业本体构建了一个智能检索原型系统，在智能导航、自动扩大检索范围和跨语言检索等方面都进行了初步尝试。实验结果表明，该系统能提供较友好的导航功能，检全率也有一定的提高，还可以实现简单的跨语言检索。如果能进一步丰富和完善这些功能，将能大大提高传统检索系统的性能，也将会有广阔的应用前景和实际的使用价值。

7. 期刊论文 [李娜](#).[任瑞娟](#) [叙词表、分类法与分布式本体 -现代情报](#)2007, 27 (12)

本文分析叙词表、分类法与分布式本体概念的内涵与外延及各自的属性,探讨了三者相互关系,在此基础上提出了建立基于叙词表、分类法与分布式本体模型的设想.这种分布式本体是在语义和知识层次上描述信息系统的概念模型建模工具.通过对这种分布式本体的机理与实现方法的分析与总结得出结论:基于叙词表、分类法构建的分布式本体是在分布异构的网络环境下探索知识发现、知识组织、知识检索、知识服务的有效途径,是智能网络服务的必然归宿.

8. 会议论文 [曾新红](#).[林伟明](#) [中文叙词表本体共建共享系统OTOSS的设计与实现](#) 2007

阐述了中文叙词表本体(OnoThesaurus,即基于中文叙词表建立的本体知识库)共建共享系统OTCSS的设计与实现方法,并对我国叙词表编纂机构利用本系统快速实现现有中文叙词表(主题词表)的本体转换和网络化共建共享提出了建议。

9. 期刊论文 [Choi Suk-Doo](#),[王一丁](#) [利用叙词表开发本体 -数字图书馆论坛](#)2007 (5)

文章提出了一种构建大规模韩语叙词表的方法,这种叙词表可用于在各种不同领域内提高检索性能.目前它主要用于标引以及检索过程,新的词汇也正源源不断地添加进来.随着韩语中对于检索性能的新需求的不断增加,开发一个大规模的本体系统应当是必要的,因而一个正在进行的项目的目标就是把现有叙词表转变为一个本体系统.文章将描述叙词表是如何构建的,并指出如何将其演变成为一个本体系统的基础.

本文链接: http://d.g.wanfangdata.com.cn/Periodical_qbxb200803010.aspx

下载时间：2009年10月27日