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）

文 摘 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 Xinhong[1]，Ming Zhong[2]，Jiang Ying[3]，Lin Weiming[1]，
Hu Zhenning[1]，Zhang Shuiying[1]

(1. The 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

我国现有一百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

作者简介：**曾新红**，女，1968年生，1992年毕业于北京大学，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，副研究馆员，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）。

临着矛盾的处境：理论上，网络上的海量信息使得信息检索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]，NCI[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/内嵌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。

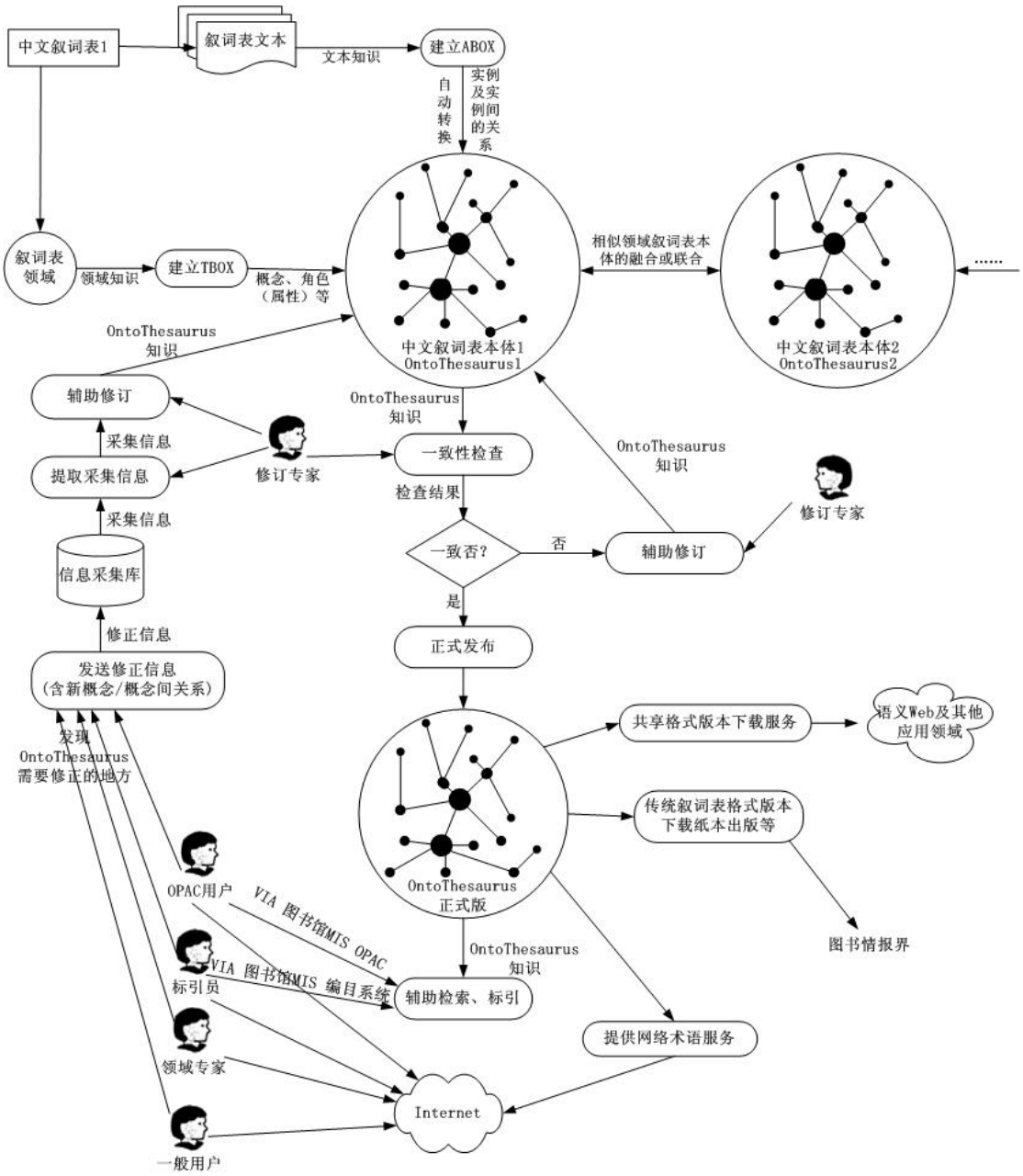

图 1 系统总体结构

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力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着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，与上位、下位和相关关系对应的属性名称改为与 SKOS 一致），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 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	
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：

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）。

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 部分的片断。

```
……
 <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 部分的片断

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/扫描转换为此格式），然后用统一的转换程序将其转换为 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 部分的片断。

```
……
<Concept rdf:ID="知识">
    <CLCCode rdf:datatype="http://www.w3.org/2001/XMLSchema#string">[B023]</CLCCode>
    <CLCCode rdf:datatype="http://www.w3.org/2001/XMLSchema#string">G302</CLCCode>
    <Related rdf:resource="#基本知识"/>
  </Concept>
  <CompoundConcept rdf:ID="知识-传播">
    <CLCCode rdf:datatype="http://www.w3.org/2001/XMLSchema#string">G2</CLCCode>
  </CompoundConcept>
  <Concept rdf:ID="知识表达">
    <CLCCode rdf:datatype="http://www.w3.org/2001/XMLSchema#string">TP18</CLCCode>
    <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">
    <CLCCode rdf:datatype="http://www.w3.org/2001/XMLSchema#string">[D912.1]⑦</CLCCode>
    <CLCCode rdf:datatype="http://www.w3.org/2001/XMLSchema#string">D913⑦</CLCCode>
    <CLCCode rdf:datatype="http://www.w3.org/2001/XMLSchema#string">D(9)5⑦</CLCCode>
  </Concept>
……
```

图 3 《中国分类主题词表》的 OWL 文件 ABox 部分的片断

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[2]。

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（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

[2]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。

应用提供机器可理解的术语学工具。

5.1 网络术语学服务（Terminology Service）

广义的网络术语学服务可以是 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（如经由术语学服务注册中心）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 开发的网络术语学服务专指交互式的、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学服务研究》一文[3]。

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的编目模块（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学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）。可视为应用 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

[3]见注 3

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）：

⑴ 新增补叙词（正式主题词/概念）；

⑵ 为原叙词增加入口词（同义词）；

⑶ 修改原叙词款目信息；

⑷ 原叙词款目整条删除；

⑸ 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，因为没有一个良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[4]。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

[4] 见注 3

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）。

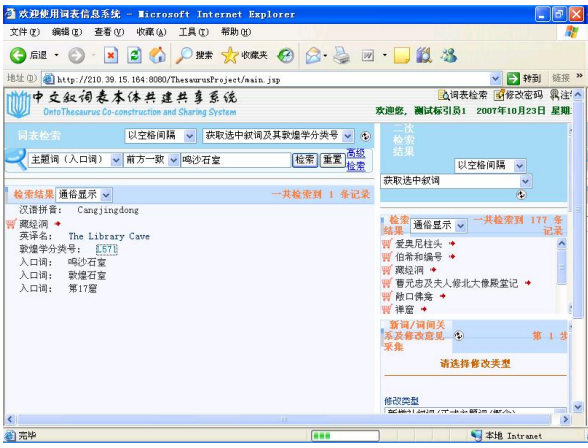

图 4 用户界面

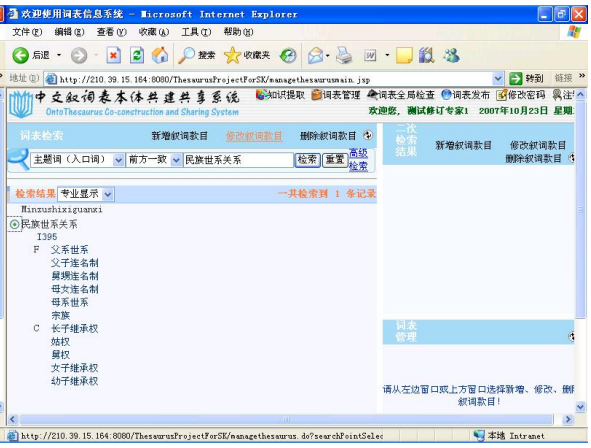

图 5 修订专家使用的界面

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，并正在寻求与其它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 Termology 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人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学服务，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学服务的实践经验，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. http://www.haworthpress.com/web/CCQ (Accessed Jul. 1, 2005)

2 唐静. 叙词表转换为Ontology的研究. 情报理论与实践，2004(6):642-645

3 Udo Hahn and Stefan Schulz. Building a Very Large Ontology from Medical Thesauri. Handbook on Ontologies, Springer-Verlag, 2004, pp.133-149

4 Jennifer Golbeck et al.. The National Cancer Institute's Thesaurus and Ontology. http://www.mindswap.org/papers/webSemantics_NCI.pdf (Accessed Mar 16, 2004)

5 Douglas Tudhope, Traugott Koch, Rachel Heery .Terminology Services and Technology: JISC state of the art review. 15-09-2006. http://nkos.slis.kent.edu/ (Accessed Oct. 17, 2006)

6 刘耀，穗东方. 领域Ontology概念描述体系构建方法探析. 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 31 (1999)183-220

8 SKOS Core Guide: W3C Working Draft 2 November 2005. http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102 (Accessed Jun. 8, 2007)

9 SKOS Use Cases and Requirements: W3C Working Draft 16 May 2007. http://www.w3.org/TR/2007/WD-skos-ucr-20070516/ (Accessed Jun. 8, 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. 情报学报，2005(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. 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. 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. 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. 国家图书馆学刊，2000(4):68-74

16 OCLC. Terminology Services. http://www.oclc.org/research/projects/termservices/ (Accessed May. 5, 2007)