2007 年数字图书馆建设与应用研讨会暨成果展示会“中国数字图书馆十年：回顾与展望”征文

中文叙词表本体共建共享系统 OTCSS 的设计与实现*

曾新红　林伟明

（深圳大学图书馆，深圳，518060）

[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 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。

[关键词] 叙词表，本体，中文叙词表本体，本体构建，网络术语学服务，机辅标引，概念检索

Design and Implementation of OTCSS (OntoThesaurus Co-construction and Sharing System)

Zeng Xinhong　Lin Weiming

(The Library of Shenzhen University, Shenzhen 518060, China)

[Abstract] This paper presents the design and implementation of OTCSS (OntoThesaurus Co-construction and Sharing System), and recommends that the thesauri owners can update existing Chinese thesauri to OntoThesauri and realizes their co-construction and sharing services via Internet quickly by using the OTCSS.

[Keywords] thesaurus ontology OntoThesaurus ontology construction Networked Terminology Service computer-aided indexing concept retrieval

1　前言

国家社科基金项目“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”（05CTQ001）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（Machine to Machine）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。项目成果“中文叙词表本体共建共享系统”应用本体技术实现了中文叙词表的形式化表示及其扩展（中文叙词表本体，OntoThesaurus）以及一致性检查与逻辑/层次结构完善机制，并可在中文叙词表本体的共享应用服务中及时采集主题标引员、领域专家、OPAC 用户及其他用户提出的新词/词间关系和其他修订意见，经统计分析后提供给修订专家参考，结合本体推理技术半自动实现中文叙词表本体的动态完善。中文叙词表本体共建共享系统的总体研究、中文叙词表的 OWL 本体转换方案和实现方法、中文叙词表本体的一致性检查和逻辑/层次结构完善机制、中文叙词表本体的检索实现等相关内容请于近期关注本项目资助发表的系列成果论文。

本文主要介绍 OTCSS 系统的设计与实现方法，并对利用该系统快速实现我国现有中文

* 国家社科基金项目“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”（05CTQ001）；
深圳大学科研启动基金项目“图书馆信息管理系统领域模型的研究开发”（项目编号：200555）

叙词表的本体转换和网络化共建共享提出具体实施建议。

2 *OTCSS* 系统功能

OTCSS 系统通过网络为主题标引员、*OPAC* 用户及其他用户（如领域专家、网络上的一般用户等）提供中文叙词表本体的检索服务和相关信息获取服务；在服务中及时收集这些用户发送的新词、词间关系及其他修订意见；对采集到的修订意见进行统计分析；辅助修订专家提取集成后的知识动态完善中文叙词表本体；定期发布更新版本供共享使用。*OTCSS* 主要由以下子系统构成：

1） 中文叙词表本体网络术语学服务（*OntoThesaurusTerminologyService*，*OTTS*）
2） 基于中文叙词表本体的机辅标引系统（*OntoThesaurusAssistIndexingSystem*，*OTAIS*）
3） 基于中文叙词表本体的机辅检索系统（*OntoThesaurusAssistRetrievalSystem*，*OTARS*）
4） 知识集成与本体动态完善中心（*OntoThesaurusMaintenceSystem*，*OTMS*），包括：
①新词等信息的采集、发送、接收和统计分析功能
②本体的人工干预和动态完善功能

OTTS 实现中文叙词表本体的网络检索和信息获得服务，也是实现 *OTAIS* 和 *OTARS* 的基础。

OTAIS 主要实现利用中文叙词表本体辅助标引员对图书进行主题标引；当标引员进行标引时，可以从分类、主题等多个途径检索本体库，显示相应主题词的相关信息和词间关系等。标引员可从显示的主题词中直接取词对图书进行标引。如果某主题词或词间关系不在该叙词表本体中，并且标引员认为它们有必要加入到中文叙词表本体库中，则将新的主题词/关系信息或其他修订意见发送到 *OTMS* 中，由 *OTMS* 子系统进行集成。

OTARS 主要实现利用中文叙词表本体，帮助 *OPAC* 检索用户构造检索式，规范检索词，实现扩检缩检等，提高检索质量；同时也可以大量收集检索用户使用的自由词（入口词）到 *OTMS* 中，由 *OTMS* 子系统进行集成，为叙词表本体修订专家的修订决策提供参考。

OTMS 主要实现对收集到的新的主题词/关系信息或其他修订意见进行统计分析和集成，对候选的新主题词/关系或其他修订意见进行一致性检查，然后向中文叙词表本体修订专家提供决策参考，最后由修订专家决定是否将新的主题词/关系或其他修订意见加入本体中。修订专家决定发布新版本的叙词表本体时，*OTMS* 可以对叙词表本体进行一致性及逻辑性检查，并向修订专家提供决策参考，辅助修订专家修改错误，确认无错误时才把新版本发布出去，供 *OTAIS* 以及 *OTARS* 使用。

3 *OTCSS* 需求分析

本系统采用软件工程方法指导整个设计和开发过程，具体方法详见参考文献[1]。需求分析是整个软件开发中必不可少的一部分。我们采用用例说明从用户的角度来捕获系统的功能并建立用例模型。对用例模型进行分析，找出用例中执行事件流的各个类，将用例的行为分配给类，并定义类的属性、行为以及类与类之间的关系，建立分析模型。

根据需求可将本系统划分为检索叙词表本体，采集发送新词/词间关系，集成发送的知识，管理叙词表本体，检查叙词表本体，发布叙词表本体，登录系统，修改密码以及管理用户九个用例，主要的角色有：一般用户、标引员、修订专家、管理员。他们在本系统中的权限见表 1。图 1 为本系统的 *Use Case* 图。图 2 为检索叙词表本体用例文档的局部。

角色名	角色权限
一般用户	拥有使用主题词检索、新词/关系采集发送（只能发送入口词知识）、修改密码的权限

标引员	拥有使用主题词检索、新词/关系采集发送、修改密码的权限
修订专家	拥有使用主题词检索、新词/关系采集发送、修改密码、提取采集信息、主题词管理、词表检查、词表发布更新的权限
管理员	拥有对各用户进行管理、修改密码的权限

表 1　系统角色列表

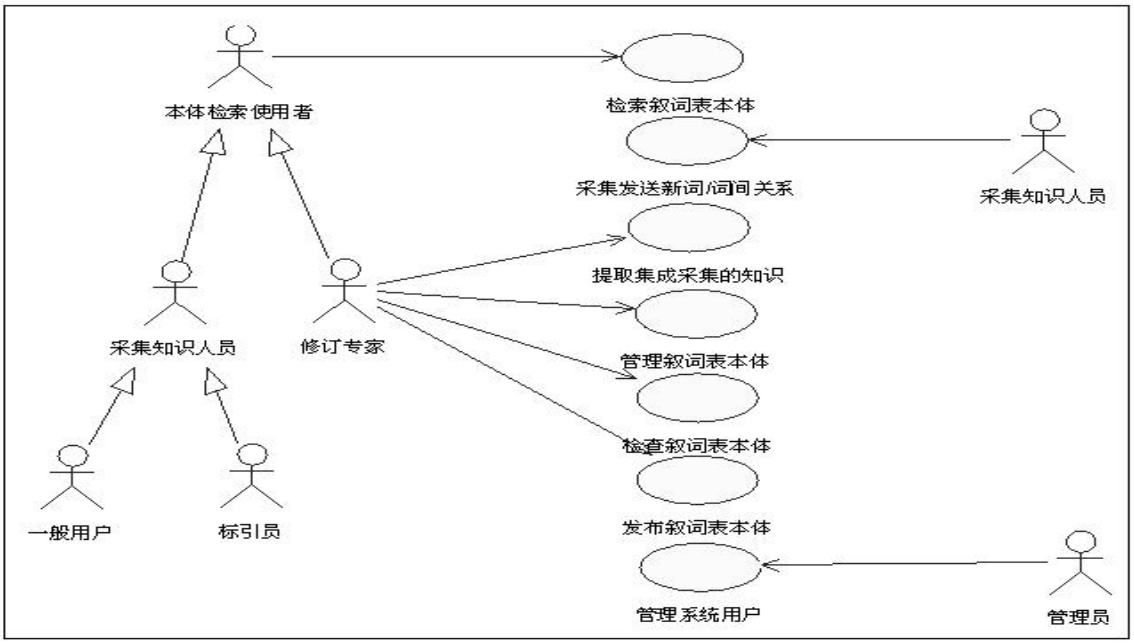

图 1　*OTCSS* 系统 *Use Case* 图

UC0_0：检索中文叙词表本体
主角：叙词表本体使用者（标引员，一般用户，修订专家等）
涉众：
简要说明：为中文叙词表本体使用者提供检索服务。包括从各种途径查找主题词，显示某一主题词的直接相关信息（款目内容），并可在这些信息点上作扩展检索。本功能可以独立使用（与 MIS 系统无交互），也是实现 UC0_1 至 UC0_3 的基础。
前置条件：在版权允许的范围内服务
基本路径：1)选择检索途径和检索方式，输入检索点查找，系统显示命中主题词（主题词串）
2)选中某一主题词，系统显示其直接相关信息（主题词款目内容）(唯一命中时自动显示)
……
扩展路径：3a1）……
补充规约：……
1a) 检索途径包括：任意途径，主题词（默认途径，含入口词），分类号（种类与本体 TBox 中分类相关 dataproperty 同步，若同步有困难，可固定为：中图分类号，科图分类号，LC 分类号，UDC，DDC），主题词英译名。
1b）应支持多种检索方式：精确，前方一致（默认），任意一致（仅针对主题词）；
……

图 2　用例文档示例

◆　检索叙词表本体用例为叙词表本体使用者（包括一般用户、标引员、修订专家）提供检索服务，包括从各种途径查找主题词，显示某个主题词的直接相关信息（款目内容），并可在这些信息点上作扩展检索。

◆ 采集发送新词/词间关系用例为采集知识人员（包括一般用户、标引员）提供服务，采集他们所发现的新词/词间关系知识。

◆ 提取集成采集的知识用例接收采集的知识并进行统计分析，为修订专家提供修订依据，辅助修订专家把根据采集知识修订叙词表本体。

◆ 管理叙词表本体用例为修订专家提供对叙词表本体进行新增、修改、删除的服务。

◆ 检查叙词表本体用例为修订专家提供在修订时对叙词表本体进行一致性检查及逻辑/层次结构完善服务。

◆ 发布叙词表本本体用例为修订专家提供发布叙词表本体的服务，经发布后的叙词表本体供检索叙词表本体用例以及采集发送新词/词间关系用例使用。

◆ 管理系统用户用例为管理员提供对用户进行新增、修改、删除的服务。

用例文档编写完毕后，对每个用例进行用例分析，从中找出分析类，并结合 *CRC*（*class-responsibility-collaboration*）卡片来定义类的属性、行为及类与类之间的关系。图 3 为 *OTCSS* 的类图之一，图 4 为 *OTCSS* 的顺序图之一。

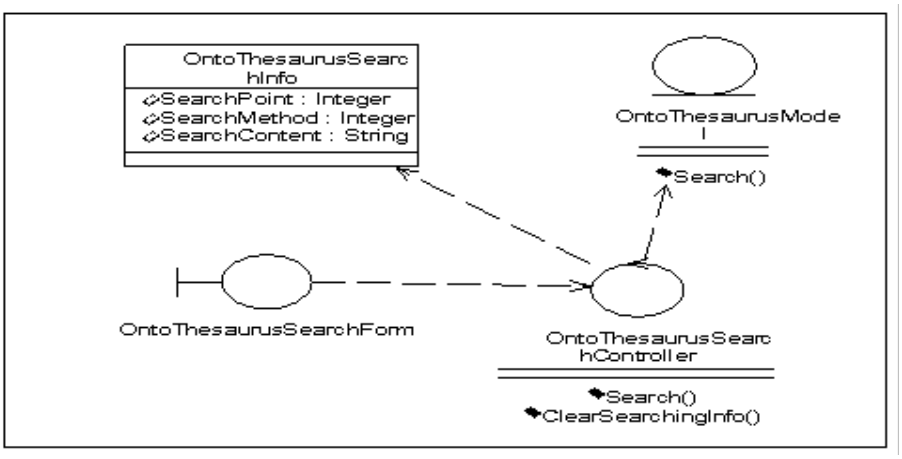

图 3　*OTCSS* 检索叙词表本体用例的类图之一

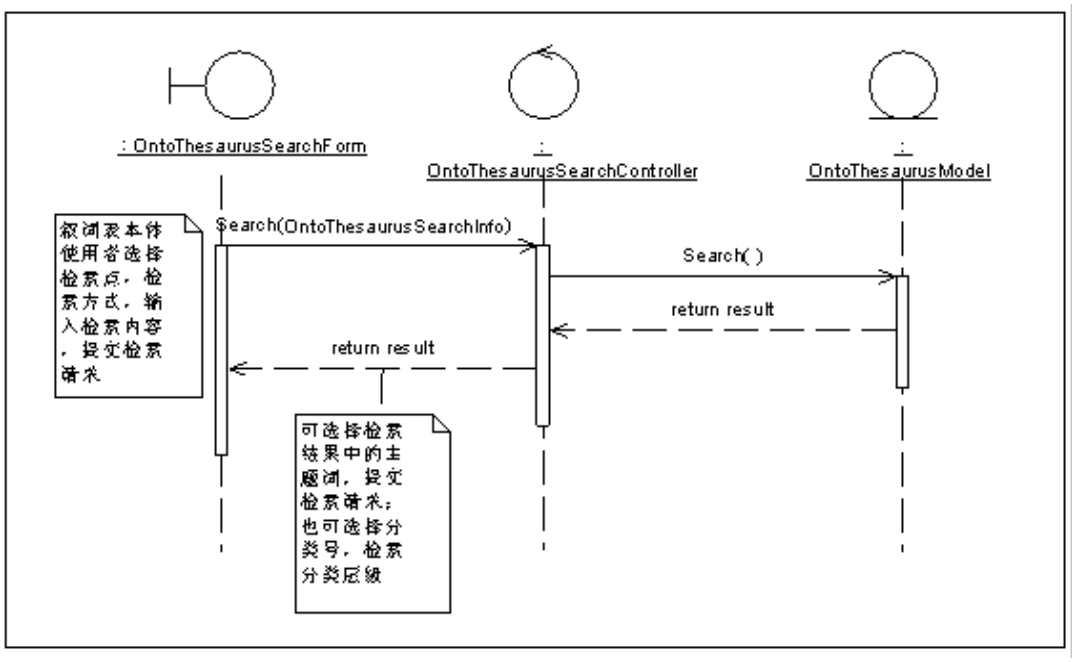

图 4　*OTCSS* 检索叙词表本体用例的顺序图之一

4 *OTCSS* 体系结构

OTCSS 系统是基于 *B/S* 模式使用 *J2EE* 的 *JSP* 以及 *JavaBeans* 技术开发的。本文结合 *Struts* 框架和 *DAO*（*Data Access Object*）设计模式提出了 *OTCSS* 的体系结构如图 5。

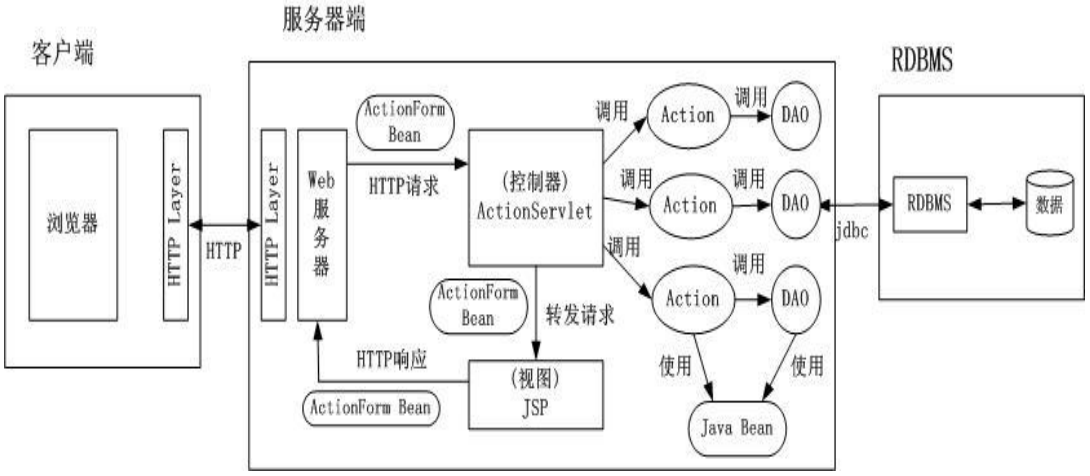

图 5　*OTCSS* 体系结构

OTCSS 把应用划分为三层：表示层，业务逻辑层和数据层。

1） 表示层：主要包括各 *JSP* 页、各静态页面等。

2） 业务逻辑层：主要包括控制器 *ActionServlet*、业务逻辑操作类 *Action*、模型类（*ActionFormBean* 和 *JavaBean*）以及各个数据访问对象 *DAO*。

ActionServlet 是中心控制器，用来截获用户的 *HTTP* 请求，然后把这个请求映射到具体的业务操作上，从而获得业务操作结果，并把该结果返回页面。

Action 是用来处理某一项具体任务或进行一个业务操作，它在客户请求、页面和业务逻辑之间起到一个桥梁的作用。

DAO 抽象和封装了所有对数据源的访问，分离了底层的数据访问操作与高层的业务逻辑，提供了面向对象的数据访问接口。

模型类与 *DAO* 一起完成具体的业务逻辑操作，模型类包括持久化对象（*PO*，*Persistent Object*）类，值对象（*VO*，*Value Object*）类以及 *ActionFormBean*。*PO* 是持久层的数据表示，反映的是数据库中某条记录对应的实体，而 *VO* 是业务层的数据表示，用于业务层之间的数据传递，*ActionFormBean* 则是 *Web* 层的数据表示，对应着 HTML 的表单。

3） 数据层：主要包括数据库和数据访问层。*OTCSS* 的数据库采用 *SQL Server* 2000，数据访问层则直接使用微软提供的 *JDBC API*。

5　*OTCSS* 的实现

根据上述的体系结构，本文把 *Windows XP Professional* 作为开发平台，使用 *SQL Server*2000 数据库服务器，采用 *Eclipse+Tomcat* 5.5 开发环境编码实现 *OTCSS*。图 6 为 *OTCSS* 的登录界面，图 7 和图 8 为供 *OPAC* 用户、标引员和其他用户使用的界面（可使用叙词表本体检索和新词/关系采集发送功能），图 9 为供修订专家使用的界面（可使用提取与集成采集信息、管理叙词表本体、检查叙词表本体和发布叙词表本体功能）。

图 6　*OTCSS* 的登录界面

图 7　*OTCSS* 的叙词表本体检索界面

图 8　*OTCSS* 的新词/关系采集发送界面

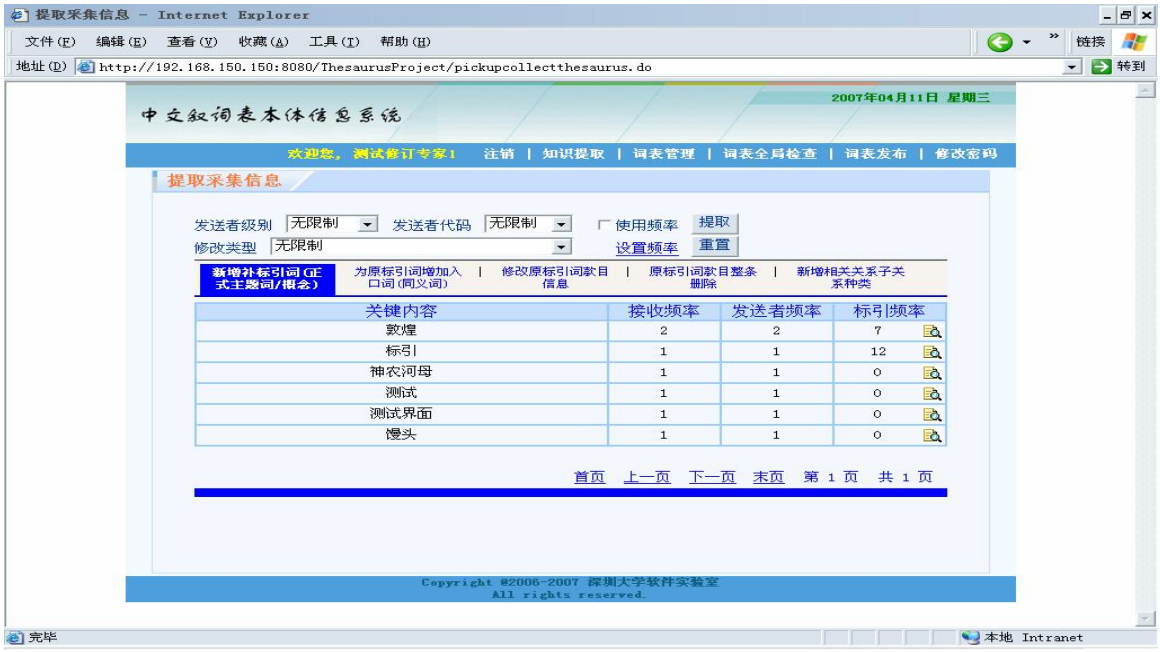

图 9　修订专家使用的界面

6 对我国现有中文叙词表实现本体化升级和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）为例实现了原型系统，并正在寻求与其它叙词表编纂机构的合作，以获取更多的系统需求来改进系统并使其尽快实用化。利用本系统，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，具体建议如下：

① 以一家中文叙词表编纂机构为单位，或多家相近领域词表编纂机构联合运作。
② 提供中文叙词表文本，自行或委托我们将其转换为本系统指定中介格式。
③ 协商确定 OWL 文件转换的类和属性（即 TBox）。本系统根据此 TBox 将已转为中介格式的叙词表文本自动转为 OWL 文件（初始叙词表本体）。
④ 利用本系统的一致性检测机制对初始叙词表本体进行检测，找出并由修订专家修正逻辑错误，完善层次结构。
⑤ 利用本系统提供的网络术语学服务功能为该中文叙词表的用户群提供服务。例如最简单的，在机构主页上设一服务菜单入口即可。
⑥ 根据该领域特点建立适合该叙词表本体的共建和动态完善机制，利用现有专业用户群实现网络共建和动态完善。
⑦ 与图书馆自动化系统合作实现内嵌式机辅标引/机辅检索功能。
⑧ 与科研机构、公司合作，开拓其他增值服务。如提供共享格式版本下载服务、融入用户环境的专门合作服务、点击率提升之后的代言广告等。
⑨ 在此系统的运行基础上可进行大量相关研究和开发，例如不同叙词表本体的集成、分布式服务，构建专业细粒度本体，实现各种衍生服务等。

以上① - ⑥步是基本步骤，投资小见效快，单从技术角度推测，理论建成周期在 2－6 个月之内。如果能由国家设立专门机构统一规划实施，则可以避免很多诸如知识产权归属、资金等非技术因素的制肘，极大地提高效率。OCLC Terminology Services Project[2]的做法也值得我们借鉴。

笔者认为，知识组织系统本身应作为一种标准规范，由国家出资建设，融入用户环境，让读者随处获得随处使用，此举在我国可以起到良好的知识规范、全民教育作用。与欧洲几十亿欧元的大规模本体建设投资相比，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。

7 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，图书馆界的分面方法、分类方法和词表编纂方法等仍然被认为是最具有理论体系和最严谨的方法。几乎毫无规范的 Folksonomy 都可以很好地运行，何况精心制作的叙词表。因此图书馆界的同仁不必妄自菲薄，尽可大胆地亮出我们的成果，让叙词表走出图书馆，为更多的网络用户服务。

参考文献：

1 曾新红等. RUP/UML 在图书馆领域软件开发中的实用化定制. 现代图书情报技术，2006(12):67-71

2 OCLC Terminology Services Research website. http://www.oclc.org/research/projects/termservices/ (Accessed May. 2, 2007)

3 Philippe Kruchten. Tutorial: Introduction to the Rational Unified Process. Proceedings of the 24th international conference on Software engineering. New York , USA: ACM Press ,2002

4 Unified Modeling Language Specification. OMG, 2001-09

5 James Rumbaugh, Ivar Jacobson, Grady Booch. The Unified Modeling Language User Guide. Beijing: China Machine Press, 2001
6 Rational Unified Process. Rational Software Corporation, 2000-02
7 丁峰, 梁维泰. RUP 软件工程过程研究及应用. 计算机工程, 2000, 26(10):112-114
8 Alistair Cockbum 著, 王雷, 张莉译. 编写有效用例. 北京：机械工业出版社, 2002
9 张恂, 沈备军. 统一用例方法的研究. 计算机应用与软件, 2005,22(9):6-9
10 Mike McClure 著. Engine-Collection-Class,一种用来建立可重用企业组件的设计模式. Microsoft，2000
11 Philippe Kruchten 著，麻志毅等译. RUP 导论. 北京：机械工业出版社，2004