中国图书馆学会专业图书馆分会 2009 年学术年会征文
见：专业图书情报机构的知识服务创新 / 中国图书馆学会专业图书馆分会编. —北京：国家图书馆出版社，2010.8

敦煌学叙词表本体共建共享系统的研究与实现*

Research and Implementation of Dunhuangnology_OTCSS

曾新红 林伟明 胡振宁

(深圳大学图书馆，深圳，518060)

[文摘] 本文简要介绍了敦煌学检索词表、中文叙词表本体 OntoThesaurus 和中文叙词表本体共建共享系统 OTCSS (OntoThesaurus Co-construction and Sharing System)。从构建国际敦煌学知识库的各种具体需求出发，详细论证了建设敦煌学叙词表本体共建共享系统 Dunhuangnology_OTCSS 的意义，然后介绍了 Dunhuangnology_OTCSS 原型系统的实现过程及其功能，并对进一步的开放应用和共建工作提出了建议。

[关键词] 敦煌学 叙词表 本体 中文叙词表本体 知识组织系统 OWL 共建共享 Dunhuangnology_OTCSS
[分类号] G254 K870.6

1 敦煌学检索词表简介

敦煌学检索词表（又称为敦煌学分类检索词表）是敦煌研究院院级项目“敦煌学信息检索系统”的重要组成部分，包括：敦煌学分类号检索词对应表、字顺表（主表）、族首词索引和英汉译名对照表。有 word 格式的电子版。它是一部规范敦煌学及其相关联学科术语概念的词表，全表共收词 8000 余条，其中正式叙词 7813 个，非正式叙词 787 个。编制历时七年之久，收词主要来自《敦煌学大辞典》与《敦煌石窟知识辞典》，以专业研究人员惯用语为标准，并参考汉语主题词表、中国分类主题词表、社会科学检索词表。新收词目以专业人员习用俗成为主，即专业文献中出现频率较高的词语，同时也是专业研究人员检索的常用词语。[1-2]

从网络等公开信息渠道和 2005 年 11 月在上海举办的“敦煌学知识库国际学术研讨会”论文集[3]中所透露出来的信息可知，该词表是目前国内外唯一的一部敦煌学专业主题词表。

2 OntoThesaurus 和 OTCSS 简介

中文叙词表本体（OntoThesaurus）和中文叙词表本体共建共享系统（OTCSS）是国家社科基金项目“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”（编号 05CTQ001）的成果。

中文叙词表本体是中文叙词表与本体的融合，采用 W3C 推荐标准、网络本体表示语言 OWL 来表示中文叙词表，并可进一步演化为包含概念间更详细关系的细粒度本体。

中文叙词表本体共建共享系统提供对中文叙词表本体的全面支持，功能包括：

*本文系国家社科基金项目“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）的研究成果之一。

（1）可将已有中文叙词表文本自动转换为 OWL 文件（初始 OntoThesaurus）。

（2）实现了 OntoThesaurus 的网络共享应用功能，包括供人使用的 OntoThesaurus-TS 和供应用系统使用的 Web Service API（OntoThesaurus-API，目前可提供 16 个通用服务函数）。

（3）实现了 OntoThesaurus 的一致性推理检测机制。可对初始 OntoThesaurus 进行一致性检测，找出并修改中文叙词表的原有错误；在共建和修订过程中运用一致性检测，保证 OntoThesaurus 在整个生命周期中的健康运行。

（4）实现了 OntoThesaurus 的网络化用户共建和修订专家维护所需的各项功能。解决了中文叙词表本体的及时更新问题。

该系统的整体研究和若干功能的实现请参见参考文献[4-7]。

3 建设 Dunhuangnology_OTCSS 的意义

敦煌学知识库国际学术研讨会是一次高规格的国际学术研讨会，日本京都大学、日本国际佛教学大学院大学、美国耶鲁大学、德国海德堡大学和中国的北大、清华、人大、兰大、武大、台湾南华大学、中国社科院历史所等高等学府和研究机构的相关学者，以及敦煌学国际联络委员会、中国敦煌吐鲁番学会、中国唐史学会、中国魏晋南北朝史学会和上海市历史学会的主要负责人等 50 余位学者参加了会议。在此次会议上宣读的论文和展示的敦煌知识库软件内容涉及敦煌知识库的框架和技术支持，敦煌知识库需要遵守的原则、规范，各种专题敦煌数据库软件的编辑，地区和单位敦煌知识库的建设等诸多方面，反映了国际敦煌学知识库的最前沿和最新的研究成果。

从论文集中所收录的论文可以看出，与会者对构建统一的敦煌知识库时要提供主题词（关键词）检索途径以及规范化标引表现出了极大的关注（有 14 篇论文涉及此方面内容）[3]。

其中首篇论文，由台湾醒吾技术学院蔡忠霖，南华大学郑阿财提交的“关于‘敦煌学知识库’建构的设想”一文认为：知识库的特质要结合一般大众（知）与研究学者（识）的要求；在资料的分类和浏览上，建议以《敦煌学大辞典》作为基本数据，参考其分类条目来整理所有数据，并制定出统一的“分类主题词表”，用多种浏览方式在网上加以呈现，并提供中英专有名词对照。检索的设计上，至少应提供题名、关键词、作者、出版社、出版地（收藏地）出版日期及洞窟编号、文献编号等多种方式。[3]

由敦煌研究院樊锦诗和张元林提交的“关于‘敦煌知识库’的构想”一文则提到，敦煌知识库要面向整个社会，面向整个学术界；必须建立新的运行机制，成立一个打破各自隶属关系的全国性的统一协调机构或类似于全国范围的敦煌知识库共建共享协调委员会，来统筹安排、组织实施这项工作，并制定相关的全国通用的标准，及对文献标引人员进行培训，掌握编制标准化数据的技能，将已建成的非标准化的数据库改为标准化数据库，最终形成一个高效丰富、多边共享的敦煌知识数据库网络系统。[3]

上海人民出版社李伟国在“简论敦煌学知识库的基本框架和搜索引擎”一文中建议，知识库的搜索不只是面向专业的工作者，更应面向普通需要敦煌资料的学者和社会公众，因此

搜索的设计要尽量提供更宽的检索形态和主题词；他认为非常重要的任务是编制一个以“同义语义场”为主的搜索引擎，将《敦煌学大辞典》中涉及的大量同义语梳理出来就可以了，如有遗漏，在使用过程中可以不断加以补充。[3]

中国社科院历史所杨宝玉在“敦煌文书目录知识库构建设想”中呼吁，关键词检索使用频率极高，故构建知识库时一定要予以高度重视。[3]

上海师范大学汤勤福在“古典文献数据库的困境与敦煌学知识库的对策”一文中所分析的困境与对策对于日后中文叙词表本体共建共享系统推广应用中可能遇到的困难及其解决方案有极高的参考价值。[3]

中国国家图书馆史睿在“古籍文献索引与知识发现——知识库基础理论研究之一”则将关键词的建设提高到知识体系的高度，认为知识库应以知识体系为核心组织全部信息，底层是具有严格规范控制的各学科关键词，这是支撑全部知识库的基础。[3]

兰州大学的吕娟和董翔在“关于敦煌学数据库中检索字段的探讨”中对在 CALIS 项目之一的敦煌学数据库建库过程中所进行的新增词标引和关键词标引进行了详细说明，反映出《汉语主题词表》在敦煌学方面的收词不足和编制敦煌学专业词表的必要性。[3]

中国国家图书馆林世田和萨仁高娃在“国家图书馆善本特藏部敦煌资源库的建设”中对敦煌吐鲁番学论著数据库和国际敦煌项目的建设作了说明，前者在国图 OPAC 检索系统中可通过主题词途径检索，对于后者，作者认为“我们现在所做的工作只是对纸本资料的网上再现，只称作未加格式化的资料群，并无知识库应具备的纵横连接和更深层的标引。随着 IDP 项目的完善和改进，我们将其建设成成熟的敦煌知识库。” [3]

新疆吐鲁番学研究院的李肖和汤士华在“吐鲁番学研究院资料信息中心发展规划及目前工作进展”一文中对图书资料的数字化建设提出了规划方案：参照英国国家博物馆“国际敦煌学项目”（IDP）的做法，将建立所有的资料篇目索引，这个索引中的文书应该有主题（或关键）词，题名（定名），形制，遗址、语言文字等方面的说明，其他文物则应有形制与遗址的说明；在电子检索中附加相应的照片与地图。[3]

兰州大学图书馆的韩春平等在“敦煌学数字图书馆中石窟艺术库的图版关键词及其提取”一文中论述了为图版进行关键词标引和提供主题检索的重要性，并说明了具体的关键词提取方法。[3]

另有若干论文提出采用 DC 元数据，或采用 DC 元数据与其它元数据标准相结合（如 CDWA）来建设敦煌学文献信息库、敦煌遗书数据库、敦煌学数字图书馆等[3]。而 Subject and Keywords（主题和关键词）是 DC 的核心元素。

从以上这些论文中我们可以感觉到敦煌学领域对建设敦煌学专业规范词表的迫切要求。

若要满足以上论文中提出的需求，词表必须提供通过网络可自由访问的电子版本。2006 年 6 月，敦煌研究院资料中心表达了与深圳大学图书馆合作建设该词表网络电子版的意向，并提供了 word 格式的局部电子版本作研究之用。当时我们正在进行国家社科基金项目“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”（项目编号：05CTQ001）的研究，于是决定将这部词表作为该项目的一个实际案例进行深入分析。根据国际最新发展动态，我们未仿照国内其它词表电子版的建设方式，而是在对其 word 格式的版本进行格式改造和错误修正之后，直接将其转换升级为形式化（OWL）表示的本体，并建设成敦煌学检索词表本体共建共享系统的原型系统。标引员、敦煌学领域专家可以通过该系统提供的网络共建功能，对词表本体进行进一步的规范建设，将其建设成为具有严格规范控制的敦煌学叙词表本体共建共享系统（Dunhuangnology_OTCSS）。

Dunhuangnology_OTCSS 几乎可以满足上述论文作者提出的所有期望：标引员、研究专家和一般公众可以通过网络检索和获取所需的敦煌学主题词及其相应的分类号、英译名、同义词以及上位、下位和相关主题词，从而达到规范标引、引导检索选词和扩展检索词形态的目的；OntoThesaurus-API 的使用可以帮助各种已建数据库实现基于知识组织系统的辅助标引和智能检索；叙词表本体中丰富的入口词和便捷的入口词检索及同义词获取方式可以满足“同义语义场”的需求；标引员和研究专家在使用过程中如发现遗漏或错误，可以发送修订意见甚至作为修订专家直接参与叙词表本体的修订（需要申请修订专家权限），一般公众也可以就添加同义词入口词提出自己的建议，从而依靠大家的力量使叙词表本体紧跟标引和检索的实践需求，在一致性检测机制的辅佐下，共同建设一个严格规范控制、入口词丰富、具有严密的知识网络结构、常建常新的敦煌学知识组织系统，为敦煌学知识库的建设以及敦煌学的研究和知识普及提供强有力的支持。

4 Dunhuangnology_OTCSS 原型系统的实现及其功能

国家社科基金 05CTQ001 课题组已实现了 Dunhuangnology_OTCSS 的原型系统：

1）对敦煌研究院资料中心提供的敦煌学检索词表的主表（word 版）进行了格式改造，逐个添加分隔符，以方便计算机识别和转换。在此过程中对照《敦煌学大辞典》对词条进行了校对和补充：改正文字和标点、格式错误；理顺同义词指引关系，删除非叙词款目，将其中含有的入口词归入相应叙词款目，若不存在相应叙词款目则新增（分类号默认为 L 者一般为新增的叙词款目）；将“M.??窟”改为“第？？窟”，“Y.??窟”改为“榆林窟第？？窟”等等。

2）对敦煌研究院资料中心提供的英汉译名对照表进行格式改造，逐个添加分隔符，以方便计算机识别和合并入主表。

3）制定具体的转换规则，利用本系统的自动转换程序将以上两个改造好的带分隔符的文件转换和合并为一个完整的 OWL 文件（初始的 Dunhuangnology_OntoThesaurus）。

4）以该 OWL 文件为基础实现了敦煌学叙词表本体共建共享原型系统。功能包括：

- **OntoThesaurus_TS**：术语学服务功能，即通过 Internet 网络向敦煌学研究专家、标引员和一般公众等用户提供开放式服务，用户可以通过各种途径检索叙词表本体，并可

以根据自己的需要获取所需的主题词及其相关的信息（如分类号、同义词、英译名、上下位词和相关词等），再粘贴至任意的编目著录文本框、检索词输入框或其他文本编辑栏内，进行标引、检索、翻译等活动。（详见参考文献[6]）

图 1　OntoThesaurus_TS 界面

- **OntoThesaurus_API**：面向应用程序的术语学服务接口。已有的敦煌学数据库系统、知识库系统、数字图书馆系统等可以利用此接口将以上术语学服务功能嵌入到本地系统中，实现本地系统的基于敦煌学叙词表本体的智能服务。

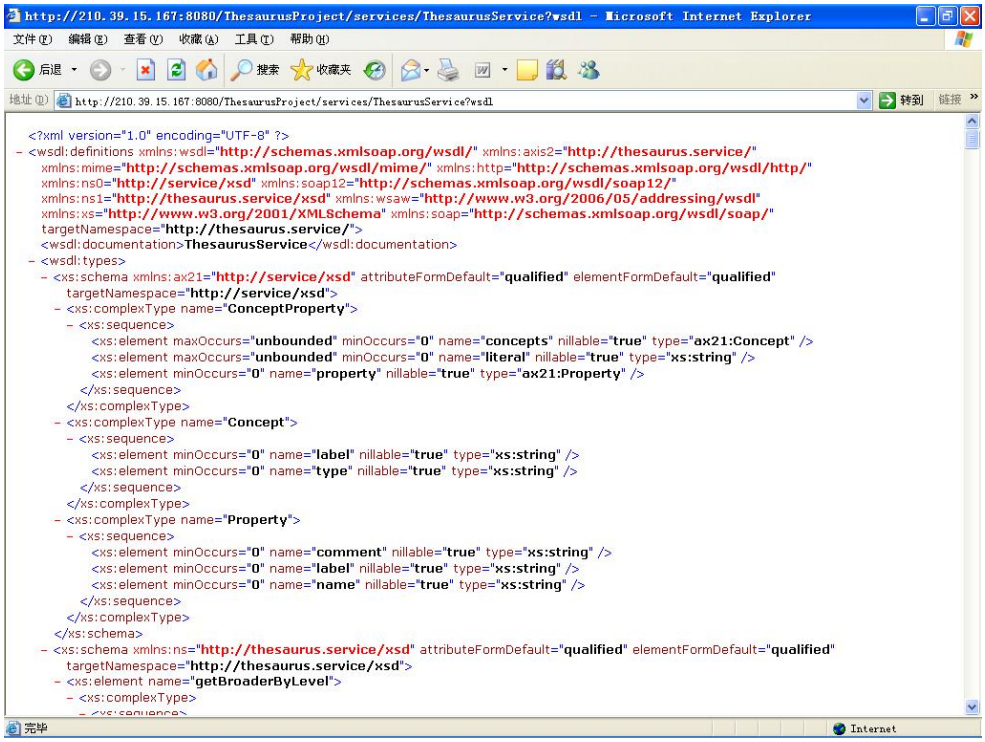

图 2　OntoThesaurus_API 的 WSDL

- **中文叙词表本体的共建功能**：领域研究专家、标引员和一般公众等用户在使用本系统提供的术语学服务功能的同时，可以在线填写和发送对该叙词表本体的修改意见，如新增叙词（正式主题词/概念）、为原叙词增加同义词入口词、修改原叙词款目信息、删除原叙词款目，甚至为相关关系增加更具体的子关系等，不同的用户有不同的发送权限。系统将接收到的修改意见信息进行统计后，修订专家可以通过网络界面提取这些修改意见进行修改确认，将合理的意见加入到叙词表本体中。修订专家也可以直接对叙词表本体进行编辑管理，通过全局检查和分步检查对叙词表本体进行一致性检测和修改，并可以对修改完善后的新版本进行发布。

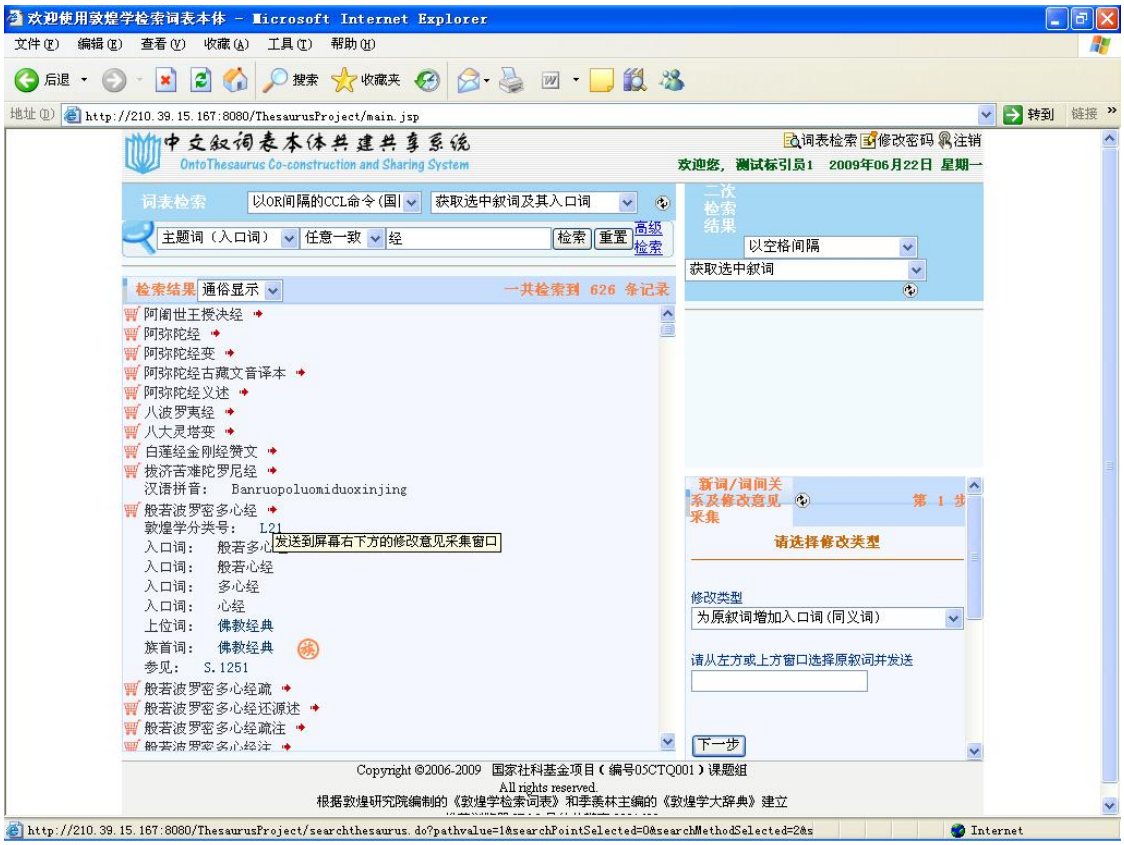

图 3　OntoThesaurus 共建功能的用户界面

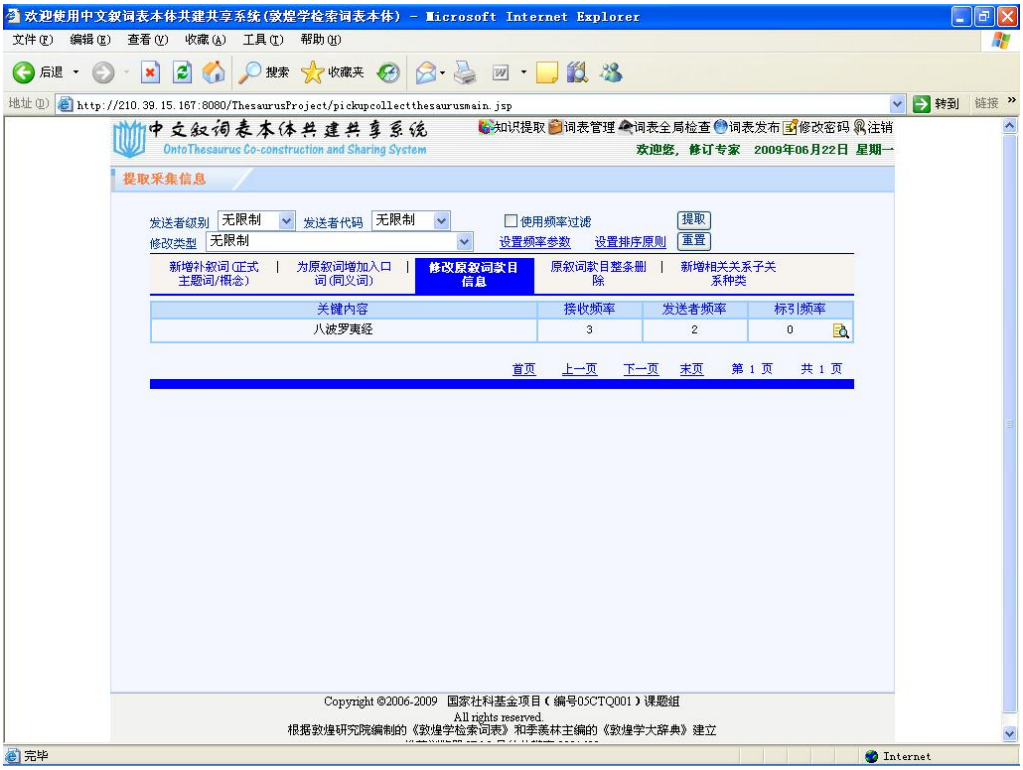

图 4　OntoThesaurus 共建功能的修订专家界面

- **中文叙词表本体的一致性检测功能**：叙词表的编制有严格的国际和国家标准规范。但和我国现有的一百多部综合性或专业性主题词表（叙词表）中的大部分一样，敦煌学检索词表主要是靠手工编纂的，因此难免出现错误，而且有些错误通过肉眼很难发现。在将其进行了本体化升级之后，可以借助本体语言的推理能力对其进行严格的一致性检测，理清和补全相关信息，从而建立起严格的概念体系和词间关系体系，极大地提高叙词表的科学性。初始的 Dunhuangnology_OntoThesaurus 首先要进行一次全局一致性检测，以修正叙词表中原有的错误；在发布供共享应用后，使用者提交修订意见时应用一致性检测机制可以提示错误和自动补齐缺失信息，从而减少提交者的输入工作量和错误信息量；在修订专家提取修订意见修订或直接编辑管理 OntoThesaurus 时，一致性检测机制的应用可以保证更新后的 OntoThesaurus 不出现一致性冲突；而正式发布新版本之前进行一次全局检查，也给了修订专家最后补救的机会。因此，一次性检测机制的实施对保证 OntoThesaurus 在整个生命周期中的质量至关重要。（详见参考文献[7]）

图 5　Dunhuangnology_OTCSS 的全局检查功能

5 进一步的开放应用和共建建议

Dunhuangnology_OTCSS 原型系统经过两年的测试和改进，功能已达到实用要求。如果进一步补充、修订数据并正式开放使用，敦煌学领域专家、标引员和对敦煌学感兴趣的网络用户就可以利用该系统提供的 TS 功能，检索、浏览和获取敦煌学概念术语及其相关信息，

应用于自己的科研、工作和学习中，如扩展检索、标引、翻译、解疑等。在应用此系统的过程中，可以利用系统提供的共建功能及时发送自己的修订意见。修订专家则可以通过网络界面提取和处理用户的修订意见，对词表进行增、删、改等管理和一致性检测（全局检查），以及发布新版本。其他的敦煌学应用系统，如敦煌学篇名数据库、敦煌学数字图书馆、敦煌学知识库等，可以利用系统提供的 Web Service API，实现基于此叙词表本体的机辅标引、智能检索等功能。

敦煌学研究遍布全球，Dunhuangnology_OTCSS 的开放使用，可以集全球之敦煌学研究力量，共建一个常建常新的敦煌学知识组织系统，为敦煌学知识库的建设以及敦煌学研究和知识服务提供强有力的支持。

我们愿与敦煌研究院密切合作，争取早日实现向全球提供开放服务的目标。

附：原型系统登录网址：http://210.39.15.167:8080/ThesaurusProject/login.jsp

可以匿名登录或在线申请其他登录帐号。输入登录网址时请严格区分大小写。

参考文献

[1] 李鸿恩. 敦煌学信息检索系统介绍及使用说明[C].见：郝春文主编. 敦煌学知识库国际学术研讨会论文集. 上海古籍出版社，2006

[2] 李鸿恩.《敦煌学检索词表》编制及使用说明，2006

[3] 郝春文主编. 敦煌学知识库国际学术研讨会论文集. 上海古籍出版社，2006

[4] 曾新红. 中文叙词表本体——叙词表与本体的融合[J]. 现代图书情报技术，2009(1):34-43

[5] 曾新红等. 中文叙词表本体共建共享系统研究[J]. 情报学报，2008(3):386-394

[6] 曾新红，林伟明，明仲. 中文叙词表本体的检索实现及其术语学服务研究[J]. 现代图书情报技术，2008(2):8-13

[7] 曾新红，林伟明，明仲. 中文叙词表本体一致性检测机制研究与实现[J]. 现代图书情报技术，2008(5):1-9

作者简介：

曾新红，女，生于 1968 年。深圳大学图书馆研究馆员。研究方向：知识组织与知识管理，数字图书馆相关技术。

Email：zengxh@szu.edu.cn

广东深圳南山区　深圳大学图书馆　邮编：518060

林伟明，男，生于 1982 年，深圳大学图书馆助理馆员，研究方向：计算机应用技术。胡振宁，男，生于 1966 年，深圳大学图书馆副研究馆员，研究方向：图书馆自动化。