数字时代图书馆学情报学研究论丛

（第二辑）

多语种叙词本体

MULTI-LINGUAL THESAURI-ONTOLOGY

邓仲华　赵又霖　黎春兰　汤　平　编著

WUHAN UNIVERSITY PRESS
武汉大学出版社

数字时代图书馆学情报学研究论丛

（第二辑）

本书得到教育部人文社会科学重点研究基地重大项目（11JJD630001）资助

多语种叙词本体

MULTI-LINGUAL THESAURI-ONTOLOGY

邓仲华　赵又霖　黎春兰　汤　平　编著

WUHAN UNIVERSITY PRESS
武汉大学出版社

图书在版编目(CIP)数据

多语种叙词本体/邓仲华,赵又霖,黎春兰,汤平编著. —武汉:武汉大学出版社,2011.10
数字时代图书馆学情报学研究论丛. 第2辑
ISBN 978-7-307-09287-7

Ⅰ.多… Ⅱ.①邓… ②赵… ③黎… ④汤… Ⅲ.叙词法—研究 Ⅳ.G254.24

中国版本图书馆CIP数据核字(2011)第227265号

责任编辑:詹 蜜 责任校对:黄添生 版式设计:马 佳

出版发行:武汉大学出版社 (430072 武昌 珞珈山)
(电子邮件:cbs22@whu.edu.cn 网址:www.wdp.com.cn)
印刷:湖北省京山德兴印务有限公司
开本:720×1000 1/16 印张:29 字数:410千字 插页:2
版次:2011年10月第1版 2011年10月第1次印刷
ISBN 978-7-307-09287-7/G·2306 定价:46.00元

《数字时代图书馆学情报学研究论丛》
编 委 会

序

“图书馆学情报学”是我国的习惯用法，是涵盖图书馆学、情报学、档案学、出版发行学等学科的名称。在我国台湾被称为“图书馆与资讯科学”，英文为 Library and Information Science。美国也用 Library and Information Studies 来称谓这一学科。

1807 年，德国学者马丁·施莱廷格（Martin Schrettinger，1772—1851）首次使用了“图书馆学”这一概念，1808 年他又在《试用图书馆学教科书大全》中建立了以图书馆整理为核心的学科体系，标志着图书馆学学科正式诞生。

自 1887 年美国学者杜威（Melvil Dewey，1851—1931）在哥伦比亚大学创办世界第一所图书馆学校，1930 年在卡内基基金的资助下芝加哥大学设立第一所图书馆学博士班课程以来，图书馆学开始走进大学殿堂，成为高等教育中的一个专业。

图书馆学教育在美国的兴起带动了全球图书馆教育的发展。1919 年英国在伦敦大学建立了图书馆学院。目前，美国有 56 所美国图书馆学会（ALA）认可的图书馆学院，每年招收图书馆与情报学学生 26 000 人左右。

在施莱廷格后的两个世纪，图书馆学科不断变化。特别是在 20 世纪 50 年代以来的冷战期间，美苏军备竞赛，两大阵营形成。苏联卫星上天，美国实施阿波罗计划，科技文献激增。科学家对文献信息的获取变得困难。一门新型学科——情报学应运而生。1963 年美国文献工作学会正式更名为美国情报学会（ASIS）。大量增设图书馆学与情报学硕士点、博士点。图书馆学课程表中也增加了大

量的情报学课程。

20 世纪 70 年代，计算机技术在图书馆与信息工作中广泛应用，自动化、地区性图书馆网络形成，机读目录广泛应用，国际图联将世界书目控制列为核心计划。图书馆学（Library Science）发展为“图书馆与情报学”（Library and Information Science），后来又进一步演变为“图书馆与情报研究”（Library and Information Studies）。

20 世纪 80 年代高新技术迅速发展，信息时代到来。美国里根政府实施星球大战计划，欧洲实施尤里卡计划等。联机图书馆系统广泛建立，并扩展至世界主要发达国家。商业性联机数据库如 ORBIT，DIALOG 发展迅速，图书馆与情报职业面临挑战。为适应信息时代要求，国际上图书馆学情报学专业开始调整。国际上有较多大学将图书馆学院易名为图书馆与情报学院或信息研究学院，图书馆学、情报学在硕士、博士层次合二为一。

20 世纪 90 年代，全球进入后信息时代——数字时代到来。克林顿政府开始实施国家信息基础设施计划（NII）、全球信息基础设施计划（GII）。新一代互联网投入使用。欧美初步建成信息社会，全球进入无缝信息环境。世贸组织建立和一揽子贸易协定生效，使全球经济一体化并逐步进入知识经济时代。各国继续加强图书馆学、情报学学科调整。图书馆学、情报学学科内容向情报科学汇集。

进入 21 世纪以来，国际上信息管理学科变化很快。自雪城（SYRACUSE）大学将学院更名为信息研究学院（The School of Information Studies）后，在美国立即出现了 iSchool 的浪潮。伊利诺依斯大学、华盛顿大学、密歇根大学、匹兹堡大学、加州大学伯克利分校、北卡罗来纳大学等知名大学的图书馆与情报学院宣称自己为 iSchool。这些 iSchools 通过宪章组成 I-Schools 联盟（ISG）。目前共有 20 所美国的大学加入联盟（联盟宪章不允许超过 25 个）。iSchool 强调信息、技术与人的关系（relationship between information, technology and people）。iSchool 的标准包括：必须有杰出的研究和杰出的博士教育；必须能在科学、企业、教育与文化进步过程

中提供任何形式的信息所需的专门技术；必须能提供信息技术及其应用、信息使用与用户方面的专门知识。2004～2006年的联盟领导委员会协调人是雪城大学信息研究学院的Raymond von Dran院长，2006～2007年将由匹兹堡大学信息学院院长Ron Larsen担任。联盟成员的标准主要强调研究即实质性承担研究活动（三年中每年研究支出达到100万美元），同时，致力于培养未来的研究者（通常通过研究型的博士点），引领推动信息职业领域。

国际上图书馆与情报学科的发展表现出明显的特征：研究范围由传统的图书馆领域扩大到信息领域（information field），研究视野由实体的图书情报机构扩大到虚拟空间，研究对象由图书文献转向了信息内容。一系列相关学科如图书馆学、情报学、档案学、出版科学、信息管理与系统乃至数字商务汇集于信息科学（Information Sciences）下，从而使图书馆学情报学研究发生了根本的变化。

武汉大学图书馆学科起源于1920年美国学者韦棣华女士创办的武昌文华大学图书科，档案专业起源于1940年的文华图书馆学专科学校的档案管理科。1978年武汉大学创办科技情报学专业，后改为情报学专业。1983年创办图书发行学专业，2002年创办电子商务专业。1984年经教育部批准建立武汉大学图书情报学院。2001年更名为信息管理学院。图书馆学和情报学两个二级学科被国务院学位委员会批准为国家重点学科。“图书馆、情报与档案管理”被国务院学位委员会批准为一级学科博士学位授权点。教育部批准“武汉大学信息资源研究中心”为国家人文社会科学重点研究基地。信息产业部批准成立“国家信息资源管理（武汉）研究基地”。新闻出版总署批准建立“新闻出版总署武汉大学高级出版人才培养基地”。“网络信息资源开发与数字图书馆建设”被国家计委、教育部等批准为“十五”211重点学科建设项目。建立一级学科博士后流动站。武汉大学信息资源研究创新基地被列为国家“985二期工程”建设项目。一批院内校级重点研究基地如武汉大学四库学研究所、武汉大学中国科技评价中心、武汉大学政府信息

研究中心、武汉大学数字图书馆研究所、武汉大学出版发行学研究所、武汉大学图书馆学情报学国际合作研究中心也在科研和人才培养中发挥着重要平台作用。

强调一级学科内学科群建设和学科协调发展是武汉大学图书馆与情报学科建设的基本目标。以图书馆学、情报学两个国家重点学科为龙头促进图书馆学、情报学、档案学、信息资源管理、出版发行学等学科的协调发展。

我们深刻认识到信息资源与自然资源、人力资源共同构成支撑现代经济社会发展的资源体系。信息资源是知识经济时代重要的国家战略资源，是实现经济和社会全面、可持续发展的基础条件。对信息资源的拥有、开发和利用水平，是衡量一个国家综合国力和国际竞争力的重要标志之一。消弭信息鸿沟、实现信息公平，是消除贫困、促进经济发展、构建和谐社会的重要条件之一。

信息资源管理人才培养是学院的基本任务。学院每年为国家培养本科生 260 名，硕士研究生 150 名，博士研究生 55 名左右。学院有一支知识结构和年龄结构合理的优秀学术队伍。这支队伍中有武汉大学人文社会科学资深教授 1 人，博士研究生导师 26 人，国务院政府特殊津贴专家 6 人，教育部新世纪优秀人才支持计划 3 人，武汉大学珞珈特聘教授 2 人。作为实现研究型学院建设目标的一部分，在教学的同时，广大教师承担了大量的科学研究任务。为了推动本学科领域的前进，分享他们的见解，在武汉大学出版社的大力支持下，并报有关部门批准，我们拟出版《数字时代图书馆学情报学研究论丛》（简称《论丛》）。

为了编辑这套丛书，武汉大学邀请了国内外知名学者担任《论丛》的学术顾问，组建了主要由信息管理学院的博士研究生导师担任委员的编辑委员会。

《论丛》拟用 4 年时间出版著作共 20 卷。20 卷著作将分为三个系列：（1）学科年度进展。主要约请信息管理学院图书馆学系、档案与电子政务学系、信息管理科学系、现代出版系、信息系统与

电子商务系的有关教师和校外专家共同编写本学科的年度研究进展，主要有《图书馆学研究进展》、《情报学研究进展》、《档案学研究进展》、《出版学研究进展》、《信息资源管理学研究进展》；（2）个人学术专著。涉及图书馆、情报与档案管理基本理论研究、信息组织与检索、信息资源管理、信息资源建设与信息服务、文献编纂与出版、数字图书馆与信息系统工程等研究方向；（3）研究报告系列。我院研究人员共承担教育部哲学社会科学研究重大攻关项目、国家社会科学基金重点项目、教育部人文社会科学重点研究基地重大招标项目、国家自然科学基金项目、国家社会科学基金项目多项。特别是211项目和985项目，围绕数字信息资源开发与管理、数字信息资源服务与保障、信息资源公共获取与知识产权协调管理、数字图书馆关键技术与系统、资源与服务整合、信息构建与知识管理等主题正在进行探索。在信息构建的理论与方法、信息系统与资源整合、元数据知识表达、网络计量与参考、信息服务集成机制、信息资源与服务集成技术、媒体及数字出版、数字内容分销、信息资源的长期保存、商务信息流等关键领域力图实现图书馆学科在数字图书馆领域、情报学科在数字资源管理领域、档案学在数字化政务信息管理领域、出版发行学在数字出版与数字化分销、信息系统科学在集成系统以及数字化商务信息流研究方面取得研究结果。本系列将对部分研究结果进行报告。

丛书的出版是学院广大教师和研究人员辛勤探索的结果，在此，谨向严谨治学、辛勤耕耘的各位著作者表示感谢！对武汉大学出版社的支持表示感谢，对责任编辑严红女士在策划编辑过程中付出的艰辛劳动表示感谢。同时，还望广大读者不吝批评指正，共同推动图书馆学、情报学、档案学、出版发行学和信息资源管理学科的进步！

武汉大学信息管理学院院长　**陈传夫**
武汉大学信息资源研究中心主任　**马费成**

前　言

叙词表是“将文献、标引人员和信息用户的自然语言转换成规范化语言的术语控制工具，它是概括各门或某一学科领域并由语义相关、族性相关的术语组成的可以不断补充的规范化词表”，在文献标引和信息检索方面有着广泛的用途。叙词表突出的特点是丰富的关系定义，如 ANSI Thesaurus 标准（Z39.19—1980）中规定有 13 种词汇间关系（这些关系完全包括了中国《汉语主题词表》的“用、代、属、分、参”结构），使其在表达语义方面具有优良的性能。多语种叙词表则是在普通叙词表的术语及关系中加入了不同语种的映射。早在 1985 年，国际标准化组织规范了多语种叙词表的制定和修改规则，随后，许多国家也制订了自己国家的专业叙词表，并附带有其他语种叙词，目前编制完成的叙词表超过 2 000 种。多语种叙词表不仅仅是跨语言信息检索的重要工具，而且是一个多语种的语义词典，在语义网、跨语言知识组织与管理、全球信息资源组织等方面有广阔的应用前景。但就目前的使用情况来看，这些重要的作用受诸多因素的影响还没有得到很好的发挥，如：大多叙词表都是基于数据库构建，但在操作、检索、显示等方面未达到简单直观的程度，限制了应用的推广；叙词表的构建是一项综合系统工程，需要国家或大型机构支持方能完成；科学技术的发展、认识的深入，促进新术语的产生，随之对叙词表的维护功效提出了更高的要求。

本体是一个概念模型明确的规范说明。在应用上，本体对概念及其关系的描述更加精细，这一点特别适合于多语种的映射。因此，将叙词表构建成多语种本体是综合了本体与叙词表两方面的优势，它不仅完全保留了叙词表在功用方面的优势，而且利用本体的

学习功能使叙词表具有自动维护的功能。以本体的形式直接构建多语种叙词表对扩大叙词表应用范围与提高其利用率具有深远的意义。通过采用本体理论构建多语种叙词本体，充分发挥叙词表在功用上的优势；通过利用本体的学习功能对叙词本体自动修复、自动优化、自动进化，增强叙词表的智能性，促进叙词表在知识交流与创新层次的实现与发展。事实上，研究多语种叙词表本体的构建理论和多语种叙词本体的自动维护模型，不仅提高了叙词本体的应用效能，而且扩大了其社会功用。例如，多语种叙词本体可以成为跨语言信息检索的语料库，实现多种信息的高效组织，提高用户检索的检全率和检准率；在实际的运用中，多语种叙词本体还可以作为网络搜索引擎以及网络百科全书信息组织的词汇工具，同时还可以为商业信息数据库服务。

本书在研究和撰写过程中，按照研究方法、技术路线和研究方案进行。全书共10章。第1章是叙词表及其应用，包括叙词表的起源及发展历程、叙词表定义和特点、叙词语言的词汇控制和叙词表的编制与维护；第2章是本体及其应用，包括本体起源及其发展历程、本体的相关性质、应用及构建；第3章是叙词本体，包括叙词本体的定义、功能、结构和多语种叙词本体；第4章是叙词本体的构建，包括叙词本体构建的理论基础、建模方法、词间关系及多语种方案和多语种叙词本体的构建；第5章是叙词本体演化，包括叙词本体演化概述、过程、基本方法和本体学习工具以及叙词本体演化框架KAON介绍；第6章是叙词本体信息源获取，包括网络爬虫原理、开源项目分析和资源下载器设计；第7章是叙词本体概念获取，包括叙词本体中的概念词、HTML解析、概念获取的流程、中文分词算法和中文分词系统介绍；第8章是概念关系抽取，包括关系抽取研究、概念关系抽取的方法和概念关系的抽取过程；第9章是专家评审机制，包括叙词本体演化的管理、专家评审机制的必要性、专家评审机制、专家评审内容与规则和专家评审结果的表达；第10章是系统设计与实现，包括系统总体结构、本体库构建、数据检测子系统、概念获取子系统、关系抽取子系统、本体演化管理和实验结果。本书适合从事情报学、信息资源管理、图书馆学中

的分类学和本体论等相关领域的理论研究者和实际工作者使用。

本书的写作汇集了项目组成员的集体智慧。参与撰写的主要有邓仲华、赵又霖、黎春兰、汤平、刘抒、刘青、艾世浩、陈礼国、喻越、钱文静和杨明松。

本书在研究和撰写过程中，得到了多方的支持，包括教育部、武汉大学人文社会科学研究院、武汉大学信息资源研究中心、武汉大学信息管理学院、武汉大学出版社的大力支持。本书在撰写过程中广泛参阅和引用了国内外众多文献和开源项目，这些文献和开源项目为本书的写作提供了有力的支撑。值此书出版之际，编者一并表示衷心的感谢。

由于时间仓促，加之研究水平有限，书中难免存在一些疏漏和错误，敬请诸位读者批评指正。

编者

2011 年

目　录

1 叙词表及其应用

1.1 叙词表的起源及发展历程

叙词（thesaurus）也称描述词、叙述词，是经过规范化处理的，以基本概念为基础的表达文献主题的词和词组，具有概念性、描述性、组配性。经过规范化处理后，还具有语义的关联性、动态性、直观性，它是描述文献资料主题的一种标识符号。

从信息检索的角度来说，叙词是作为一种检索语言使用，所以在一些场合也称为叙词法。叙词法起源于20世纪50年代末，我国也称为主题词语言，是在单元词法的基础上发展起来的一种新的检索语言。它以单元词语言作为直接基础，又综合了标题法、分面组配式分类法等多种标引语言的原理和方法，包括：

①在单词组配上保留了单元词法的基本原理，在单元词法的完全后组和反记法的技术基础上发展出后组和倒排文档技术。

②采用标题法的基本方法，对语词进行严格控制的规范化方法，以确保语词能够与概念一一对应。此种规范方法在单元词法对单元词的控制中也得以发展。

③适当地采用了标题法中先组方法，以及分面组配式分类法中的概念组配原理，取代单元词法中的单纯字面组配，克服了某些复合主题分拆和组配后产生的意义失真的缺点；在采用标题法的参照系统的基础上，进行了进一步的完善，包括用、代、属、分、参参照。

④采用体系分类法的学科分类和等级结构技术编制了叙词范畴

索引、词族索引和词汇分类表；在关键词语言的轮排技术的基础上编制了轮排索引，从多个方面显示了叙词间的相关关系，以确保准确、全面地选用叙词进行标引和检索。

叙词法在吸取了这些信息标引语言的优点之后，已经发展成为一种性能优越的现代检索语言类型。

第一部叙词表出现在20世纪50年代末期，此后，随着计算机在信息检索领域的应用，叙词表的编制得以迅速增长，使叙词语言逐步发展成为受控检索的主要语种。

1.1.1 叙词表的发展

1.1.1.1 国外叙词表的发展

在国外，叙词法的研究开展得较早，可以追溯到19世纪下半叶。在一些欧美国家，叙词表是作为主要的检索语言。欧美研究叙词法的著名人物主要包括：

卡特（A. Cutter，1837—1903），美国图书馆事业上最杰出的贡献者之一，19世纪下半叶图书馆事业的领袖、标题法的始祖。其代表作《词典式目录编制规则》（*Rules for a Dictionary Catalogue*）（1876年第1版，1904年第4版）为图书馆学文献中的一个里程碑，其中“议题款目”规则部分是世界上第一部字顺主题目录的法典。

卢恩（H. P. Luhn，1896—1964），情报科学的先驱，关键词索引和自动标引的创始人。被公认为以下几个方面的创始人：题内关键词索引（1958），计算机定题服务（SDI）系统（1958），计算机编码系统，利用统计学方法进行自动标引和自动编制文摘。

克莱夫顿（C. W. Cleverdon，1914—1997），英国图书馆员和情报学家，主持领导了世界上首次大规模的、以测试语言在检索中的效益及对情报检索系统的影响因素为目的的克兰菲尔德试验，并在试验中首次采用的“检全率”、“检准率”、词汇的“专指性”、标引“详尽程度”等概念以及试验方法，之后被情报界广泛采用。