数字图书馆

编者按： 数字图书馆栏目由同方知网技术有限公司协办。同方知网与电子杂志社以网络出版和知识情报服务为主要业务方向，依靠自主开发的全文数据库管理、知识挖掘与数字出版等先进技术，与社会各界通力合作，坚持打造可为全社会提供各种知识服务的《中国知识资源总库》。其中，国家重点出版项目——《中国学术文献网络出版总库》，大规模集成整合了我国学术期刊、博硕士学位论文、会议论文、报纸、年鉴、工具书、学术图书、专利、标准、科技成果等各类文献资源。尤其是基于《总库》的行业、专业与个性化数字图书馆，融合了各类先进的知识服务模式，为高效率创新、学习和管理决策创造了理想的信息化环境。

中文叙词表本体——叙词表与本体的融合*

曾新红

（深圳大学图书馆　深圳 518060）

【摘要】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和迫切性。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。

【关键词】叙词表　本体　知识组织系统　OWL　共建共享　OTCSS

【分类号】G254　TP18

OntoThesaurus（Chinese - Thesaurus - Ontology）——An Integration of Thesaurus and Ontology

Zeng Xinhong

（Shenzhen University Library, Shenzhen 518060, China）

【Abstract】This paper expatiates the importance and emergency of OntoThesaurus formalization and its application in Internet, with the network information society requires for Knowledge Organization System incrementally and the KOS - related societies are making great efforts to meet the needs. Definitions of thesaurus and ontology are studied comparatively and feasibility of integrating them to form a new kind of KOS (OntoThesaurus) is verified. The reasons for representing OntoThesaurus with OWL rather than SKOS are given, and the OWL classes and properties for OntoThesaurus are defined and listed. The realization of comprehensive functions and several prototypes of OTCSS (OntoThesaurus Co - constructing and Sharing System) demonstrates that the definition of OntoThesaurus is scientific, feasible and universal for Chinese thesauri.

【Keywords】Thesaurus　Ontology　KOS　OWL　Co - constructing and Sharing　OTCSS

1　网络信息社会对知识组织系统的需求

搜索引擎的问世使自然语言成为网络信息检索的主力语言，传统的叙词表、分类法、规范档等受控语言似乎

收稿日期：2008 - 11 - 24

* 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”（项目编号：05CTQ001）的研究成果之一。

正在被网络信息社会所抛弃。自然语言真的可以取代受控语言吗？或者说，网络信息社会只需要大众化的标签(Tag)吗？笔者认为答案是否定的。网络信息的海量和无序已使越来越多的人在思考网络信息资源的有效组织和高效检索时，重新把目光投向了传统的知识组织系统——以叙词表和分类法为代表的情报检索语言。网络知识组织系统(NKOS)，就是在这样背景下产生和发展起来的。

NKOS(Networked Knowledge Organization Systems/Services)网站致力于讨论功能和数据模型，以使知识组织系统(例如分类系统、叙词表、地名表和本体)能够作为网络化的交互式信息服务，通过Internet来支持多种信息资源的描述和检索[1]。NKOS有两种类型：一种是来自信息科学界(Information Science，即国内所称的图书馆学情报学界或图书情报界)的传统知识组织系统的延伸和发展，如分类法、叙词表、主题标题表、规范档等的网络应用；另一种则是在网络环境中产生和发展起来的语义工具，如本体(Ontology)和语义网络(Semantic Network，如WordNet(也被称为词汇数据库))等。关于NKOS领域的研究现状见参考文献[2]已作了较为全面的综述，最新动态则可参见NKOS网站[1]，本文不再赘述。在此着重介绍在NKOS框架下KOS的类型，探讨主要研究对象——叙词表和本体在其中的位置。

笔者对参考文献[2-5]中的相关内容进行了综合、修改和补充，给出KOS的类型分布如图1所示。

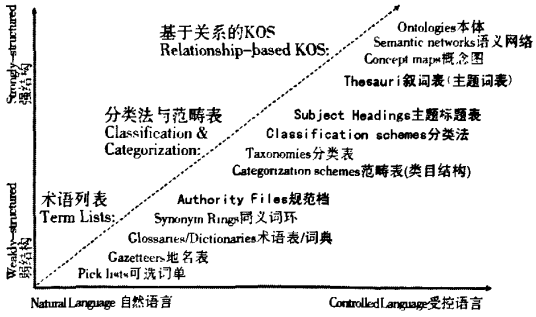

图1 KOS类型分布图

从图1中可以看到，NKOS框架下的KOS是一种广义的知识组织系统概念，包括所有的(广义的)受控词表，大致可分为：从最简单的线性结构和提供多义性(Ambiguity)控制的各种术语列表，到具有等级关系控制和树状结构的分类法和范畴表(有些含少量横向相关关系，如LCSH)，再到基于关系(含纵向等级关系和横向相关关系)和网状结构的知识组织系统类型。

叙词表和本体同属于这种结构最复杂、控制程度最高的类型。有理由相信：只要解决了最高端类型的知识组织系统的形式化表示和网络应用问题，其他低端KOS的这些问题只是他的简化，自可迎刃而解。

2 信息科学界(图书情报界)的应对研究

传统的手工编制和纸本服务方式显然不能满足网络时代用户对叙词表的需求。为用户提供交互式或自动术语学支持的前提是叙词表的数字化和网络化。国际信息科学界已为此作出了巨大努力。

参考文献[2,5-9]通过大量实例介绍了叙词表、分类法等传统知识组织系统在国际上的网络应用，展示了传统的知识组织系统在网络环境下所具有的蓬勃生命力。

王军等在参考文献[2]中将KOS在网络环境下的表示和发展大致划分为三个阶段：KOS的电子化；HTML表示的KOS；用语义网(Semantic Web)的相关技术(例如XML、RDF、OWL以及W3C最新推出的SKOS)表示KOS。笔者对他们的表述作了一些补充和修改，并结合参考文献[4,10]等的相关内容，对KOS在网络环境下的表示和发展作以下综合评述。

(1) KOS的电子化：KOS网络化发展的前期阶段是KOS的电子化，代表特征是KOS的MARC描述和数据库化。用数据库存储和MARC表示方便了对KOS的管理和访问，也便于将他们与相应的电子资源集成在一起。例如：在英国科学文摘(INSPEC)和工程索引(EI)数据库中分别集成了INSPEC词表和EI词表，以方便查询词的选取、扩检和缩检等操作。20世纪90年代以后，我国采用计算机技术编制的许多专业性叙词表，也是和检索系统集成在一起开发的。

MARC格式是图书情报界用来对书目、分类法、主题词表、规范档等进行交换的标准格式。例如LCSH、LCC都提供MARC版本并可以在网上查询。《中图法》编委会也在2002年5月至10月根据UNIMARC Classification Format并结合《中图法》的结构特点设计了“中国分类法数据机读格式”(CNMARC Format for Classification Data)，并依据此格式建立了“《中图法》

机读数据库"[11]。可惜的是,这一版本的《中图法》未提供开放的网上查询服务。

(2) 基于网页技术的 KOS:代表特征是通过网页技术提供传统 KOS 的网上浏览和查询功能,主要供人使用,这是目前 KOS 在网络上的主要表现方式。其中包括两种类型。

①是基于静态 HTML 网页技术制作的网络 KOS,仅提供浏览界面(少量有简单的查询界面)。HTML 只是一种描述网页显示格式和布局的语言,并不便于计算机理解和自动处理,因此 KOS 的 HTML 表示,只相当于纸本式 KOS 在网络上的翻版。

②是采用动态网页技术(如 ASP,PHP,JSP 等)将数据库存储的 KOS 展示在 Web 上,可提供灵活的检索功能和良好的交互界面。

参考文献[12]调查分析了 40 个英文网络叙词表,详细列出了其网址和用户界面内容。《中国分类主题词表》第二版的电子版也用到了一些 HTML 网页技术来方便主题词款目和分类—主题对照表的显示,但还未提供 Internet 上的网络服务。

(3) 基于语义网技术表示的 KOS:在语义网框架下发展出来的一系列描述语言,包括以 XML 为代表的用于内容和结构描述的标记语言、以 RDF 为代表的描述语义和关系的资源描述框架,以及以 OWL 为代表的可满足逻辑和证明要求的本体表示语言等,其目标是使计算机能够更好地理解网络上的信息,从而进行知识发现、数据集成、信息导航等活动[13]。这些工具被用来表示 KOS 标志着 NKOS 的真正产生。

应用 XML 描述大型词表的例子有 DDC 和 Mesh。比 XML 更进一步的资源描述框架 RDF(S)可以用来描述词表中的概念及其之间的关系,例如联合国粮农组织(FAO)用 RDF 表示多语种词表 AGROVOC,阿姆斯特丹大学用 RDF(S)表示的艺术与建筑叙词表 AAT。W3C 于 2005 年 11 月发布的 SKOS(简约知识组织系统)标准草案[14],也是基于 RDF 设计的。

W3C 于 2004 年 2 月发布的正式推荐标准 OWL(Web Ontology Language)是一种用于在语义 Web 上发布和共享本体的语义置标语言,他代表了面向 Web 的本体表示语言的最新发展趋势。他面向 WEB,相对于 XML、RDF 和 RDF Schema 拥有更多的机制来表达语义,而又与他们兼容。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。因此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,如美国国家癌症研究所发布的 NCI 叙词表的 OWL 版本[15,16]等。

在用 OWL(或其前身 OIL + DAML 等)表示 KOS 的研究和实践中,几乎都涉及叙词表原有关系的细化和扩展,而且一般不再保留原有的几种粗粒度关系,而直接代之以细粒度的各种具体词间关系。即叙词表的原有结构已被抛弃。因此,笔者认为,这些活动似乎更倾向于解决本体的构建问题,而不是叙词表的形式化表示问题。

叙词表的标准建设也在顺应时代的发展。在叙词表的发展过程中已形成了很成熟的标准和规范,这些标准大多制定于 20 世纪 80 - 90 年代。随着网络环境下词表种类的扩展和服务方式的改变,要求制定数字环境下相关标准的呼声越来越高。英美两国已对其受控词表的编制标准进行了修订:2005 年,美国国家信息标准化组织 NISO 发布了 Z39.19 的第四版 Z39.19 - 2005[17],而同年英国标准协会(British Standards Institution, BSI)也发布了 BS 5723 的升级版本 BS8723[18]。这两个升级版本的标准都扩展了词表类型的适用范围,并对原有的叙词表关系提出了子关系细化方案,以及词表在网络环境下的显示建议等。

我国图书情报界也对叙词表的电子化发展付出了巨大的努力,取得了一些重要的成果。20 世纪 90 年代以来编制的叙词表都运用了计算机编表技术,有些在编表的同时建立了比较完善的词表管理系统,其中最具代表性的有《军用主题词表》、《农业科学叙词表》和《国防科学技术叙词表》管理系统等[19]。2005 年出版的《中国分类主题词表》二版电子版代表了目前我国叙词表电子化发展的最高水平。

中文叙词表在网络化发展方面还比较薄弱,到目前为止,整体网络应用者还未见到实用例子。近年来,这方面的研究已呈上升趋势,但对于中文叙词表的整体转换和开放网络服务方面的研究和实践还十分欠缺。

由此可见,我国叙词表电子化网络化发展的整体水平还基本处于上述三个阶段中的第一阶段,与国际先进水平相比还有一定的差距。由于词表编制技术和应用技术没有根本性的改观,更新依然缓慢,成本依然高昂,极大地限制了中文叙词表的发展和应用。从另一

个角度来看,这也说明中国的学者在这一领域大有可为。笔者希望,通过本研究,可以为中文叙词表的网络化发展提供一种基于最前沿技术的、切实可行的解决方案,为实现我国 KOS 领域的跨越式发展尽一份力量。

3 来自知识组织系统其他相关各界的启示

目前,除了信息科学界(图书情报界)之外,知识组织系统也已成为语义 Web、人工智能、知识工程等领域共同研究的课题。从这些相关领域的研究进展中得到一些启示,以使研究能够尽可能地博采众长,并满足不同领域的应用需求。

在语义 Web 界,不仅发展了上一小节中提到的各种描述语言,还制定了支持这些描述语言的各种技术标准,由 W3C 推荐发布,形成了广泛的国际共识。如 SPARQL 检索语言和 Web Service 系列标准等。基于语义 Web 技术表示的 KOS 可直接利用这些标准和技术来实现其网络服务、检索等功能。

对本研究产生重要影响的本体是语义 Web 的核心。从目前国内外相关研究的动态来看,本体理论和本体实用技术的发展都已日趋成熟。学术界对本体表示和推理的逻辑基础描述逻辑(Description Logics)和框架逻辑(Frame Logic)、本体的表示语言、开发本体的方法、本体的底层结构以及本体的应用等都进行了比较深入的研究[20]。各大研究机构和 IT 公司对本体的构建、检索和推理工具的研发投入巨大,已推出了若干实用开发工具,例如,本体构建工具 Protégé 和本体开发工具包 Jena 等,这些都为应用本体技术推进叙词表的网络化发展提供了良好的技术环境。

2002 年以来,本体的研究逐渐引起了大陆图书情报界学者的注意,他们就本体在图书情报界的应用前景、基于本体的信息处理模式和检索模式、本体的开发思路和方法等问题提出了各自的看法。这些研究也对本研究的实施产生了一定的影响。

在知识科学界,本体被认为是一种深层次上的知识,可以为各种不同的知识系统、乃至其他系统之间的知识(或资源)共享和互操作提供手段[21]。我国著名的理论计算机专家、中国科学院数学与系统科学研究院院士陆汝钤先生提出,知识是结构化的信息,正如概率论是研究信息论的基本数学工具一样,本体就是知识结构性的基本描述,这一点与国际上有关专家产生共识。他建议我国要大力发展知识工程,构筑海量的知识库,并积极参与到语义网建设的国际努力中去,参与制定相关的国际标准,贡献本体知识库,这样才能使在新一代的因特网世界中占有一席之地[22]。

4 叙词表与本体的比较研究

国内外已有多位学者对叙词表与本体进行了比较研究,其中比较有代表性的见参考文献[23-28]。这些文献从多个方面对本体和叙词表进行了比较,基本总结出了这两者之间的差异,但也存在一些不太准确的结论。在全面参考了二者的权威定义和相关文献的基础上,笔者对叙词表和本体的关系给出以下评述。

国际标准 ISO 2788-1986[29]给叙词表下的定义是"The vocabulary of a controlled indexing language, formally organized so that the a priori relationships between concepts (for example as "broader" and "narrower") are made explicit."(受控的标引语言词表,被正式地(也可译为"形式化地")组织,以便概念之间的推理关系(如"上位"和"下位")明确化)。这个早期的较抽象的定义与以下介绍的本体定义表述有颇多相似之处,之后人们又将叙词表的结构、用途等因素加入进来,给出了多种不同的定义表述。如 ANSI/NISO Z39.19-1993[30]为叙词表下的定义是"一种受控词表,以一种众所周知的顺序排列和结构化,以便术语之间的等同、同形异义、等级和相关关系被明确显示,并通过相互使用的标准关系指示符进行标识"。我国《文献叙词标引规则》(GB/T3860-1995)[31]则认为,叙词表是"自然语言中优选出来的语义相关、族性相关的科学术语所组成的一种规范化词典。在文献标引与情报检索过程中,它是用以将文献、标引人员及用户的自然语言转换为统一的系统语言的一种术语控制工具。"

在人工智能界,最早给出本体定义的是 Neches 等,他们认为:"本体定义了构成一个主题领域的词汇表的基本术语和关系,以及将这些术语和关系结合起来定义词汇表扩展的规则(An Ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary)"[32]。可以看出,这个早期的本体定义与叙词表的定义没有本质的区别。斯坦福大学的 Gruber 给出的本体定义被引用得

最为广泛,他认为,"本体是概念模型的规范说明(An Ontology is a specification of a conceptualization)"[33]和"本体是概念模型的明确的规范说明(An Ontology is an explicit specification of a conceptualization)"[34]。Borst在此基础上对本体的概念进行了引申,认为"本体是共享概念模型的形式化的规范说明(An Ontology is a formal specification of a shared conceptualization)"[35]。得到广泛认可的是Studer(1998)在Gruber(1993)[33]和Borst(1997)[35]的定义基础上提出的"本体是共享概念模型的明确的形式化规范说明(An Ontology is a formal, explicit specification of a shared conceptualisation)"[36]。其中,概念模型(Conceptualization),指客观世界中某一现象的抽象模型,通过已经识别的该现象的相关概念而得到;明确(Explicit),指所使用的概念的类型以及对其用法的约束都有明确的定义;形式化(Formal),指本体应该是计算机可读的这个事实,将自然语言排除在外;共享(Shared)则反映了这样一个观念:本体捕获的是共同认可的知识,即他不是某一个个体专用的,而是被一个团体所接受的知识[36]。

Alexandra Moreira等在文献[24]中采用一种分析-综合(Analytical-Synthetic)方法对叙词表和本体在信息科学和计算机科学文献中出现的各种(英文)定义进行了全面研究,得出的结论是:从计算机科学的角度来看,本体和叙词表一样,是一个概念系统,因此它属于认识论(Epistemological)层次,而非(哲学的)本体论层次。即计算机科学的本体和信息科学的叙词表都在认识论层次上起作用,区别在于所使用的语言、形式化水平和用途上(某些计算机科学界的研究者也将叙词表称为是非形式化的本体(Informal Ontology))。本体针对领域概念注册,目标是自动推理;而叙词表针对的是用户和文献语言之间的交流。叙词表完成了计算机科学企图用本体完成的部分目标,因此它们也被称为术语学本体(Terminological Ontologies)[24]。

尽管叙词表和本体有不同的起源和用途,从他们的定义可以看出,它们都是通过受控词汇来表达概念的概念系统,都提供了对领域知识的共同理解与描述,都追求概念及其之间关系的明确化(Explicit)和描述的形式化(Formal)。只是因为叙词表是设计来使标引人员所用的词汇和检索人员所用的语言相匹配,即供人工使用而达到领域知识共享的,而本体从一开始就是设计给计算机理解和使用的,因此他们采用了不同的方法来达到明确化、形式化和推理的目标。叙词表通过严格的词汇控制等构建方法学来保证概念的无歧义,通过统一的、规范化的指示符和显示格式来正式(供人工使用的形式化)地、明确(Explicit)地表示概念间的推理关系(A Priori Relationships,如IS-A关系(上位、下位关系)、同义关系等[37]),并通过人工干预来完成推理;而本体则通过具有严格数学基础的形式化方法来保证概念的明确无歧义和实现自动的推理。

因此,可以对叙词表与本体的关系得出以下结论:叙词表是一种非数学意义上形式化的特殊本体,他们之间的主要区别在于概念间关系的粒度(叙词表用"用、代、属、分、参"等来揭示较粗粒度的等同关系、等级关系和横向的相关关系,而本体可以容纳任意种类和粒度的关系)和形式化的程度(人工方法和数学方法之区别)。参考文献[23-28]中所列举的很多不同之处归根结底都来源于这两个基本的区别。因此笔者认为,叙词表和本体是可以融合的。可以引入本体的形式化方法来表示叙词表,提高叙词表的科学性,使叙词表能够被计算机理解和实现自动推理;然后可以在叙词表原有概念体系的基础上扩展更具体的关系种类,使其能够具备细粒度本体的功能,这样,叙词表的网络化发展和本体的构建就可合二为一。基于这种理解,笔者创建了中文叙词表本体OntoThesaurus——一种新型的、同时具备叙词表和本体特征的知识组织系统。

5 中文叙词表本体的构建

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(Working Draft)[14],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,在很多方面还需要完善和扩展。笔者认为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。

笔者的感觉是,SKOS的定义较为松散,表示能力较强而推理能力不足,比较适合用来表示图1所示的"分类法与范畴表"、"术语列表"分组中的一些规范化程度不高的KOS(对于中文KOS仍需扩展)。他过于拘泥于传统词表的表面形式,没有从概念及其关系的实质来表示知识组织系统,因此,不太适合用来表示控制程度高、对形式化和推理要求也较高的基于关系的KOS,例如中文叙词表(尤其是当中文叙词表进一步向本体发展时)。

另外还有一个问题应引起我国研究者的重视。在现有的SKOS实现案例中,几乎都采用一个没有任何实质性语义的序号(类似于数据库的控制号,如:http://example.com/Concept/0001[38,39])来表示URI后面的概念,占据着一个概念描述的核心位置,然后用prefLabel表示首选词(叙词),用altLabel表示非首选词(非叙词,入口词),这种缺乏人类可读性(Human Readable)[40]的URI导致了概念间关系的揭示极不直观。事实上,SKOS并未规定要用序号来表示URI中的概念部分,而是建议URI去除前面部分之后应具有人类可读性[40]。笔者认为,概念应该通过文字来表述,首选词和非首选词是一个概念的若干不同的词汇表述形式,首选词之所以被选为首选词,是因为他被认为是这个概念最正规、最合适的表述形式,且具有唯一性,因此完全可以在URI中直接用作概念的文字表述(中文叙词简短连续,尤其适合),而首选词与非首选词之间是一种等同关系,在它们之间定义一个表示等同(入口)关系的属性即可。序号是人为指定的符号,主要用于实现,本身并没有语义,确实需要的话可以作为概念的一个属性出现(如NCI叙词表OWL版本的做法[16])。这样,概念之间的关系可以直接在具有人类可读性的概念表述之间定义和表示,好处是:大幅提高文件的人类可读性,易于发现概念间的错误关系,同时可简化检索、一致性检测等功能的实现复杂度。

从W3C于2007年5月发布的工作草案SKOS Use Case and Requirements[41]中可以看出,为显示和检索目的而实现概念之间关系的表示(Representation of relationships between concepts)已被列为已接受需求(Accepted Requirements)之首(R-ConceptualRelations),同时,使用OWL来扩展特殊概念类型、用OWL为词表进行编码等在Use Case案例中具有强烈的需求,实现SKOS与OWL-DL的兼容被列入了下一版本SKOS的候选需求(R-CompatibilityWithOWL-DL)。由此看来,SKOS还有很长的路要走。但SKOS于2008年8月29日发布的最后征求意见草案[42]中对这些需求的解决并不彻底,该稿已于2008年10月3日截止意见征求,预示着SKOS已基本定型,正在向成为W3C推荐标准作最后的努力。

其实,在2004年OWL成为W3C推荐标准之后,为了兼容那些规范化程度不高的KOS,SKOS仍选择基于RDF来制定标准,就已确定了自己的低调定位,面对NKOS界对形式化和推理越来越高的要求,自然有些力不从心。中文KOS有自己相对独立的发展历程,人们不可能依赖SKOS来解决所有的问题,必须依靠自己的力量对中文KOS的形式化表示问题进行深入的研究,积极参与到相关标准的制定过程中去,才能最终建立起能够满足我国中文KOS形式化表示要求的标准体系。

笔者的前期研究成果"《中国分类主题词表》的OWL表示及其语义深层揭示研究"[43]撰写于2004年7月,直接采用当时刚刚成为W3C正式推荐标准的OWL来表示叙词表,目的是为了能够利用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中文叙词表本体的映射、集成。

本项目正式实施时笔者对参考文献[43]中的类定义和属性定义作了修改和扩展,使其能够较普遍地适应我国现有中文叙词表的形式化表示要求,并参考SKOS标准草案的原语对若干属性名称作了一些修改,尽量与SKOS保持一致。中文叙词表本体的详细类定义和属性定义见表1和表2。这些类定义和属性定义构成了OntoThesaurus的TBox,他遵从OWL DL规范,可实现完全的推理。主要创新点如下:

(1)采用面向概念模式[44],以概念为中心,关系属性只针对概念而声明,并直接以叙词作为概念的表述形式,取消非叙词款目及"用"关系的相应表述。此举大大缩小了OntoThesaurus的容量并简化了实现过程。(非叙词的入口作用通过检索实现,书本格式中的非叙词款目可通过程序自动生成);

(2)参考ANSI/NISOZ39.19-2005[17]的第8节

表1 类定义

类名称	含义	OWL 定义或说明(以《中国分类主题词表》为例)
Concept	概念。词表中所有概念(将正式主题词视为概念)都是这个类的 individual(成员,实例)	<owl:Class rdf:ID = "Concept"/> (定义 Concept 类) <Concept rdf:ID = "考古学"/> (定义 Concept 类的实例"考古学")
CompoundConcept	复合概念。它是 Concept 类的子类,本定义中专指主题词串,词表中所有主题词串都是这个类的 individual	<owl:Class rdf:ID = "CompoundConcept" > <rdfs:subClassOf rdf:resource = "#Concept" > </owl:Class > < CompoundConcept rdf:ID = "考古—中国— 明代 "/ >
GeneralConcept	一般通用概念,是 Concept 的子类	总论复分对照表中列出的主题概念是这个类的 individual
PersonConcept	人物概念,是 Concept 的子类	附录二"人物"中列出的主题概念是这个类的 individual
RegionConcept	地名概念,是 Concept 的子类	包括世界地区表、中国地区表中列出的主题概念及辅助表九"通用时间、地点复分表"中列出的通用地点概念
WorldRegionConcept	世界地名概念,是 RegionConcept 的子类	辅助表二"世界地区表"中列出的主题概念是这个类的 individual
ChinaRegionConcept	中国地名概念,是 RegionConcept 的子类	辅助表三"中国地区表"中列出的主题概念是这个类的 individual
InstituteConcept	机构概念,是 Concept 的子类	附录一"组织机构"中列出的主题概念是这个类的 individual
EraConcept	时代概念,是 Concept 的子类	包括国际时代表、中国时代表表中列出的主题概念及辅助表九"通用时间、地点复分表"中列出的通用时间概念
WorldEraConcept	世界时代概念,是 EraConcept 的子类	辅助表四"国际时代表"中列出的主题概念是这个类的 individual
ChinaEraConcept	中国时代概念,是 EraConcept 的子类	辅助表五"中国时代表"中列出的主题概念是这个类的 individual
ChinaNationalityConcept	中国民族概念,是 Concept 的子类	辅助表六"中国民族表"中列出的主题概念是这个类的 individual
NTerm	Non - preferred term,非正式主题词(非叙词)	所有非叙词(入口词)都是这个类的 individual

(注:Concept 的子类可根据需要扩展。)

表2 属性定义

Domain	Property	Range	属性特征	Label	叙词表对应标识	
	ObjectProperty				英	中
Concept	HasNTerm	NTerm		入口词	UF	代 D
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	在不扩展子关系时具有对称性	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	
	* Process_CounterAgent		Related 的子属性,表示处理/反工具(Process/ CounterAgent)关系	处理/反工具相关词(扩展)	RTPCA	
	* CounterAgent_Process		Related 的子属性,Process_CounterAgent 的逆属性	反工具/处理相关词(扩展)	RTCAP	

	* Action_Product		Related 的子属性,表示行为/产品(Action/Product)关系	行为/产品相关词(扩展)	RTAPd	
	* Product_Action		Related 的子属性,Action_Product 的逆属性	产品/行为相关词(扩展)	RTPdA	
	* Action_Property		Related 的子属性,表示行为/属性(Action/Property)关系	行为/属性相关词(扩展)	RTAPp	
	* Property_Action		Related 的子属性,Action_Property 的逆属性	属性/行为相关词(扩展)	RTPpA	
	* Action_Target		Related 的子属性,表示行为/目标(Action/Target)关系	行为/目标相关词(扩展)	RTAT	
	* Target_Action		Related 的子属性,Action_Target 的逆属性	目标/行为相关词(扩展)	RTTA	
	* ConOrObj_Property		Related 的子属性,表示概念或物体/性质(Concept or Object/Property)关系	概念或物体/性质相关词(扩展)	RTCOP	
	* Property_ ConOrObj		Related 的子属性,ConOrObj_Property 的逆属性	性质/概念或物体相关词(扩展)	RTPCO	
	* ConOrObj _Origins	Concept	Related 的子属性,表示概念或物体/来源(Concept or Object /Origins)关系	概念或物体/来源相关词(扩展)	RTCOO	
	* Origins_ ConOrObj		Related 的子属性,ConOrObj _Origins 的逆属性	来源/概念或物体相关词(扩展)	RTOCO	
	* ConOrObj _Measure		Related 的子属性,表示概念或物体/度量单位或机制(Concept or Object /Measurement Unit or Mechanism)关系	概念或物体/度量单位或机制相关词(扩展)	RTCOM	
	* Measure_ ConOrObj		Related 的子属性,ConOrObj _Measure 的逆属性	度量单位或机制/概念或物体相关词(扩展)	RTMCO	
	* RMaterial_Product		Related 的子属性,表示原材料/产品(Raw material / Product)关系	原材料/产品相关词(扩展)	RTRMP	
	* Product_RMaterial		Related 的子属性,RMaterial_Product 的逆属性	产品/原材料相关词(扩展)	RTPRM	
	* DiscOrField_ObjOrPrac		Related 的子属性,表示学科或领域/对象或从业者(Discipline or Field / Object or Practitioner)关系	学科或领域/对象或从业者相关词(扩展)	RTDFO	
	* ObjOrPrac_DiscOrField		Related 的子属性,DiscOrField_ObjOrPrac 的逆属性	对象或从业者/学科或领域相关词(扩展)	RTODF	
	DatatypeProperty					
	CLCCode		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	
Concept	UDCCode	&rdfs;literal	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		杜威十进制图书分类法类号,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	&xsd;string	Cardinality =1(只能出现一次)	汉语拼音	PY	
Concept	EngCounterpart	&xsd;string		英译名		E
Concept	ScopeNote	&xsd;string		范畴注释	SN	注:

(注:表中带“ * ”号的是扩展的子关系属性,参考 ANSI/NISO Z39.19 - 2005 制定。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。)

(Relationships),对 Broader,Narrower 和 Related 三个属性分别进行了子属性扩展,同时保留原有的三个父属性,并规定两个概念之间的这三种关系不能既声明为父属性又声明为子属性,只能二者择一(通过一致性检测机制保证)。此举利于将初始的粗粒度 OntoThesaurus 逐渐演化为细粒度本体,从而支持基于概念间具体子关系的推理。

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。Concept 的子类类型和相关关系子关系类型在未来的发展中也可以根据需要增加。

笔者认为,对于中文叙词表这种控制程度高、结构

严谨的 KOS 而言，适合定义一个 OWL 应用子集来满足其形式化表示和进一步扩展要求。以上定义比较简洁、严谨，能够基本满足国内现有的 100 多部中文叙词表的形式化表示和一般扩展细化要求，可以作为基本应用子集使用。在将来的实践中若发现有其他的需求，可以再吸收 SKOS 的一些原语（将其改造成符合 OWL DL 规范）或扩展定义相应的 OWL 类和属性。

6 中文叙词表本体共建共享系统功能简介

以上述定义为基础的中文叙词表本体共建共享系统（OntoThesaurus Co - construction and Sharing System，OTCSS）已实现以下功能：

（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 17 个服务函数）。

（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本项目资助发表的其他成果论文，本文不再赘述。这些功能的顺利实现表明，中文叙词表本体的定义是科学的、可行的。

7 结 语

中文叙词表本体 OntoThesaurus 保留了叙词表的原有结构和内容，使叙词表几十年来的理论成果和实践经验得以保持和延续；同时引入了具有严格数学基础的形式化的本体表示方法，为实现自动推理和容纳更多种类、更具体的概念间关系提供了可能。以《敦煌学检索词表》、《社会科学检索词表》（局部，含民族学、宗教学、逻辑学部分）和《中国分类主题词表》（一版局部，含 D 类、K 类和 B9 类的所有叙词款目）为基础建成的多个 OTCSS 原型系统[45,46]，证明了这种表示法具有广泛的适应性，有利于快速实现现有中文叙词表的本体化升级和网络共建共享。

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。笔者认为，我国图书情报界有能力在中文知识组织系统的构建和服务方面，依托已有的理论、实践成果和人才优势，大力推进中文知识组织系统的网络构建和网络服务，逐步建立起具有中国特色的、同时与国际标准兼容的中文知识组织系统规范体系。

参考文献：

[1] NKOS Network, Networked Knowledge Organization Systems/Services[EB/OL]. [2008 - 07 - 06]. http://nkos.slis.kent.edu/.

[2] 王军，张丽. 网络知识组织系统的研究现状和发展趋势[EB/OL]. [2008 - 03 - 06]. http://eprints.rclis.org/archive/00010939/01/review_on_the - development_of_NKOS.pdf.

[3] Douglas Tudhope, Traugott Koch, Rachel Heery. Terminology Services and Technology: JISC State of the Art Review. [EB/OL]. [2006 - 10 - 17]. http://nkos.slis.kent.edu/.

[4] Marcia Lei Zeng, Athena Salaba. Toward an International Sharing and Use of Subject Authority Data[EB/OL]. [2008 - 03 - 07]. http://www.oclc.org/research/events/frbr - workshop/presentations/zeng/Zeng_Salaba.ppt.

[5] 徐晓梅，牛振东. 数字图书馆的知识组织研究[J]. *现代图书情报技术*，2007(10)：1 - 6.

[6] 康艳，张虹，侯汉清. 情报检索语言不是"明日黄花"[J]. *图书情报工作*，2007(10)：139 - 142.

[7] 韩志萍，韩志敏. 叙词表在网络环境下的新应用及对我国的启示[J]. *情报理论与实践*，2003(5)：462 - 465.

[8] 曹树金，郭菁. 网络叙词表的组织结构及优化模式研究[J]. *图书情报工作*，2005(3)：31 - 35.

[9] 焦玉英，李法运. 网络环境下信息检索语言的优化研究[J]. *情报学报*，2003(3)：291 - 296.

[10] 曾蕾. Types of Knowledge Organization Systems/Structures/Services(KOS) & How KOS are Used[R]. 2004 数字图书馆前沿问题高级研讨班，深圳大学城图书馆，2004.

[11] 国家图书馆《中国图书馆分类法》编辑委员会. 中国分类主题词表：第二版[K]. 北京图书馆出版社，2005：12 - 26（第二版修订说明部分）.

[12] 司莉，陈红艳. 网络叙词表用户界面设计策略[J]. *现代图书情报技术*，2008(5)：14 - 20.

[13] 宋炜，张铭. 语义网简明教程[M]. 北京：高等教育出版社，2004.

[14] SKOS Core Guide: W3C Working Draft 2 November 2005[EB/OL]. [2007-06-08]. http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102.

[15] Jennifer Golbeck, et al. The National Cancer Institute's Thesaurus and Ontology[EB/OL]. [2004-03-16]. http://www.mindswap.org/papers/webSemantics_NCI.pdf.

[16] nciOntology.owl(Version03.09d)[EB/OL]. [2004-03-16]. http://www.mindswap.org/2003/CancerOntology.

[17] ANSI/NISO Z39.19-2005, Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies[EB/OL][S]. [2005-12-28]. http://www.niso.org/standards/resources/Z39-19.html.

[18] BS8723, Structured Vocabularies for Information Retrieval. BSI Public Draft, 2004.

[19] 戴维民. 中国情报检索语言50年研究论纲[EB/OL]. [2005-10-14]. http://www.chinalibs.net.

[20] Staab S, Studer R. Handbook on Ontologies[M]. Springer-Verlag, 2004.

[21] 陆汝钤. 世纪之交的知识工程与知识科学[M]. 北京:清华大学出版社, 2001.

[22] 陆汝钤. 研究知识科学,发展知识工程,推进知识产业[EB/OL]. [2008-02-18]. http://www.mscenter.edu.cn/blog/shilion/archive/2008/01/31/873.html.

[23] 戴维民. 从情报检索语言到本体[J]. *图书情报工作*, 2005,49(7):6-10.

[24] Moreira A, Alvarenga L, A. de Paiva Oliveira. "Thesaurus" and "Ontology": A Study of the Definitions Found in the Computer and Information Science Literature, by Means of an Analytical-Synthetic Method[J]. *Knowledge Org*,2004, 31(4):231-243.

[25] 李景,钱平. 叙词表与本体的区别与联系[J]. *中国图书馆学报*, 2004(1):36-39.

[26] 王素芳. Ontology与叙词表的整合初探[J]. *大学图书馆学报*, 2005(1):74-78.

[27] 甘利人,李岳蒙. 主题法、分类法与Ontology的比较研究[J]. *现代图书情报技术*, 2005(12):1-6.

[28] 赵焕洲,唐爱民. 对两种知识组织系统——叙词表与Ontology的比较[J]. *情报理论与实践*, 2005(5):469-471.

[29] ISO 2788-1986, Documentation - Guidelines for the Establishment and Development of Monolingual Thesauri, 2nd ed[S]. Geneva: International Organization for Standardization, 1986.

[30] ANSI/NISO Z39.19-1993, National Information Standards Organization, Guidelines for the Construction, Format, and Management of Monolingual Thesauri[S]. Bethesda, Md.: NISO Press, 1994.

[31] 中华人民共和国国家标准. GB/T 3860-1995, 文献叙词标引规则[S]. 北京:中国标准出版社,1995.

[32] Neches R, et al. Enabling Technology for Knowledge Sharing[J]. *AI Magazine*, 1991, 12(3):36-56.

[33] Gruber, T R. A Translation Approach to Portable Ontology Specifications[J]. *Knowledge Acquisition*, 1993,5(2):199-220.

[34] Gruber T R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing[J]. *International Journal of Human-Computer Studies*,1995,43(5-6):907-928.

[35] Borst W N. Construction of Engineering Ontologies for Knowledge Sharing and Reuse[D]. Enschede: University of Twente, 1997.

[36] Studer R, Benjamins V R, Fensel D. Knowledge Engineering: Principles and Methods[J]. *Data & Knowledge Engineering*, 1998, 25(1-2):161-197.

[37] C S G Khoo, Jin-Cheon Na. Semantic Relations in Information Science[J]. *Annual Review of Information Science and Technology*, 40(2006):157-228.

[38] Skos API[EB/OL]. [2008-04-05]. http://www.w3.org/2001/sw/Europe/reports/thes/skosapi.html.

[39] DREFT SKOS Thesaurus API Demonstrator[EB/OL]. [2008-04-05]. http://www.w3.org/2001/sw/Europe/reports/thes/dreft/.

[40] Best Practice Recipes for Publishing RDF Vocabularies[EB/OL]. [2008-10-13]. http://www.w3.org/TR/2008/NOTE-swbp-vocab-pub-20080828.

[41] SKOS Use Cases and Requirements: W3C Working Draft 16 May 2007[EB/OL]. [2007-06-08]. http://www.w3.org/TR/2007/WD-skos-ucr-20070516/.

[42] SKOS Simple Knowledge Organization System Primer[EB/OL]. [2008-10-13]. http://www.w3.org/TR/2008/WD-skos-primer-20080829/.

[43] 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(2):151-160.

[44] Brian Matthews, et al. Modelling Thesauri for the semantic Web [EB/OL]. [2003-07-31]. http://www.w3c.rl.ac.uk/SWAD/thesaurus/tif/deliv81/final.html.

[45] 社科检索词表本体共建共享系统SST_OTCSS. http://210.39.15.167:8080/ThesaurusProjectForSST/login.jsp; Web Service API地址: http://210.39.15.167:8080/ThesaurusProjectForSST/services/ThesaurusService?wsdl.

[46] 中国分类主题词表本体共建共享系统CCT_OTCSS. http://210.39.15.167:8080/ThesaurusProjectForCCT/login.jsp; Web Service API地址: http://210.39.15.167:8080/ThesaurusProjectForCCT/services/ThesaurusService?wsdl.

(作者E-mail:zengxh@szu.edu.cn)

中文叙词表本体——叙词表与本体的融合

作者： 曾新红， Zeng Xinhong
作者单位： 深圳大学图书馆,深圳,518060
刊名： 现代图书情报技术 PKU CSSCI
英文刊名： NEW TECHNOLOGY OF LIBRARY AND INFORMATION SERVICE
年，卷(期)： 2009，""(1)
被引用次数： 0次

参考文献(46条)

1. NKOS Network. Networked Knowledge Organization Systems/services 2008

2. 王军. 张丽 网络知识组织系统的研究现状和发展趋势 2008

3. Douglas Tadhope. Tmugott Koch. Rachel Heery Terminology Services and Technology:JISC State of the Art Review 2006

4. Marcia Lei Zeng. Athena Salaba Toward all International Sharing and Use of Subject Authority Data 2008

5. 徐晓梅. 牛振东 数字图书馆的知识组织研究[期刊论文]-现代图书情报技术 2007(10)

6. 康艳. 张虹. 侯汉清 情报检索语言不是"明日黄花"[期刊论文]-图书情报工作 2007(10)

7. 韩志萍. 韩志敏 叙词表在网络环境下的新应用及对我国的启示[期刊论文]-情报理论与实践 2003(05)

8. 曹树金. 郭菁 网络叙词表的组织结构及优化模式研究[期刊论文]-图书情报工作 2005(03)

9. 焦玉英. 李法运 网络环境下信息检索语言的优化研究[期刊论文]-情报学报 2003(03)

10. 曾蕾 Types of Knowledge Organization Systems/Structures/Services(KOS) & How KOS are Used 2004

11. 国家图书馆《中国图书馆分类法》编辑委员会 中国分类修订说明部分).主题词表 2005

12. 司莉. 陈红艳 网络叙词表用户界面设计策略[期刊论文]-现代图书情报技术 2008(05)

13. 宋炜. 张铭 语义网简明教程 2004

14. SKOS Core Guide:W3C Working Draft 2 November 2005 2007

15. Jennifer Golbeck The National Cancer Institute's Thesaurus and Ontology 2004

16. nciOntology.owl(Version03.09d) 2004

17. ANSI/NISO Z39.19-2005.Guidelines for the Construction,Format,and Management of Monolingual Controlled Vocabularies[EB/OL] 2005

18. BS 8723.Structured Vocabularies for Information Retrieval 2004

19. 戴维民 中国情报检索语言50年研究论纲 2005

20. Staab S. Studer R Handbook on Ontologies 2004

21. 陆汝钤 世纪之交的知识工程与知识科学 2001

22. 陆汝钤 研究知识科学,发展知识工程,推进知识产业 2008

23. 戴维民 从情报检索语言到本体[期刊论文]-图书情报工作 2005(07)

24. Moreira A. Alvarenga L. A.de Paiva Oliveira "Thesaurus"and "Ontology":A Study of the Definitidons Found in the Computer and Information Science Literature,by Means of an Analytical-Synthetie Method 2004(04)

25. 李景. 钱平 叙词表与本体的区别与联系[期刊论文]-中国图书馆学报 2004(01)

26. 王素芳 Ontology与叙词表的整合初探[期刊论文]-大学图书馆学报 2005(01)

27. 甘利人. 李岳蒙 主题法、分类法与Ontology的比较研究[期刊论文]-现代图书情报技术 2005(12)

28. 赵焕洲. 唐爱民 对两种知识组织系统--叙词表与Ontology的比较[期刊论文]-情报理论与实践 2005(05)

29. ISO 2788-1986. Documentation-Guidelines for the Establishment and Development of MonolingnM Thesauri 1986

30. ANSI/NISO Z39.19-1993. National Information Standards Orgnnization, Guidelines for the Construction, Format, and Managemerit of Monolingual Thesauri 1994

31. GB/T 3860-1995. 文献叙词标引规则 1995

32. Neches R EnablingTechnologyforKnowledge Sharing 1991(03)

33. Gruber, T R A Translation Approach to Portable Ontology Speeifications 1993(02)

34. GruberT R Toward Principles for the Design of Ontologies Used for Knowledge Sharing 1995(5-6)

35. Borst W N Construction of Engineering Ontologies for Knowledge Sharing and Reuse 1997

36. Studer R. Benjamins V R. Fensel D Knowledge Engineering:Principles andMethods 1998(1-2)

37. C S G Khoo. Jin-Cheon Na Semantic Relations in Information Science 2006

38. Skes API 2008

39. DREFT SKOS Thesaurus API Demonstrator 2008

40. Best Practice Recipes for Publishing RDF Vocabularies 2008

41. SKOS Use Cases and Requirements:W3C Working Draft 16 May 2007 2007

42. SKOS Simple Knowledge Organization System Primer 2008

43. 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(02)

44. Brian Matthews Modelling Thesauri for the semantic Web 2003

45. 社科检索词表本体共建共享系统SST_OTCSS

46. 中国分类主题词表本体共建共享系统CCT_OTCSS

相似文献(10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(OntoThesaurus,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. 期刊论文 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 中文叙词表本体共建共享系统研究 -情报学报2008, 27(3)

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.

3. 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中心”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贮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示地、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TimBermem-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联。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效的知识表达体系，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信息系统中的应用；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表示的是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，这种树状结构反映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，本体的反映更微细，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；本体文档可以不倚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推荐的本体编辑器；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：选用斯坦弗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页作背景，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。

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学服务研究 -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学服务(OntoThesaums -TS)的意义.详细描述OntoThesaums的检索实现方法,以及其术语学服务应用场景典型案例,并对OntoThesaurus的术语学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的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，因为在本体的一个实体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，但是原本属于本邻域的叙词表是丢弃还是融合?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的相同和不同之处，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，这些指导将为领域本体的创建者们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是比较完整的。叙词表可以说是图书馆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，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，这也是本文的创新之处。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.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.

7.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(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。

8.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)作为一种能在语义和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。

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为一个本体系统的基础.

10.期刊论文 仓定兰.CANG Ding-lan 基于叙词表的领域本体半自动构建的研究和实现 -科学技术与工程

2009,9(24)

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,本体(ontology)作为一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,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.对本体和叙词表的相关知识进行了详细论述,并利用网络本体语言(Web Ontology Language,OWL),将叙词表中的叙词及词间关系进行了表示和描述.设计和实现了一个转化系统,能够自动地将叙词表中的知识结构和语义关系转化到领域本体中.