中文叙词表本体——叙词表与本体的融合[1]

曾新红

(深圳大学图书馆，深圳，518060)

[文摘] 本文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，详细阐述了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了深入的比较研究，论证了将它们合二为一的可行性。阐述了直接采用 OWL（而不用 SKOS）表示中文叙词表本体（OntoThesaurus）的原因，并列出了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，证明了这些定义的科学性、可行性和通用性。

[关键词] 叙词表　本体　中文叙词表本体　知识组织系统　OWL　SKOS　共建共享　OTCSS

[分类号] G254　TP18

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

Zeng Xinhong

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

[Abstract] The networked information society calls for Knowledge Organization Systems (KOS) incrementally. The information science society and other KOS-related societies have made and are making great efforts to meet the needs. It is important and urgent to realize the formal representation and access via Internet for Chinese thesauri. The paper studies the definitions of thesaurus comparatively with those of ontology, then comes to a conclusion that establishing a new kind of KOS (OntoThesaurus) by integrating thesaurus with ontology is possible. The reasons for representing OntoThesaurus in OWL rather than in 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　OntoThesaurus　KOS　OWL　SKOS　Co-constructing and Sharing　OTCSS

1　网络信息社会对知识组织系统的需求

搜索引擎的面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。网络信息的极度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，就是在这样一个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

[1]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。

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。在此我们着重介绍在 NKOS 框架下 KOS 的类型，看看我们的主要研究对象——叙词表和本体在其中的位置。

笔者对参考文献[3][2][4][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。

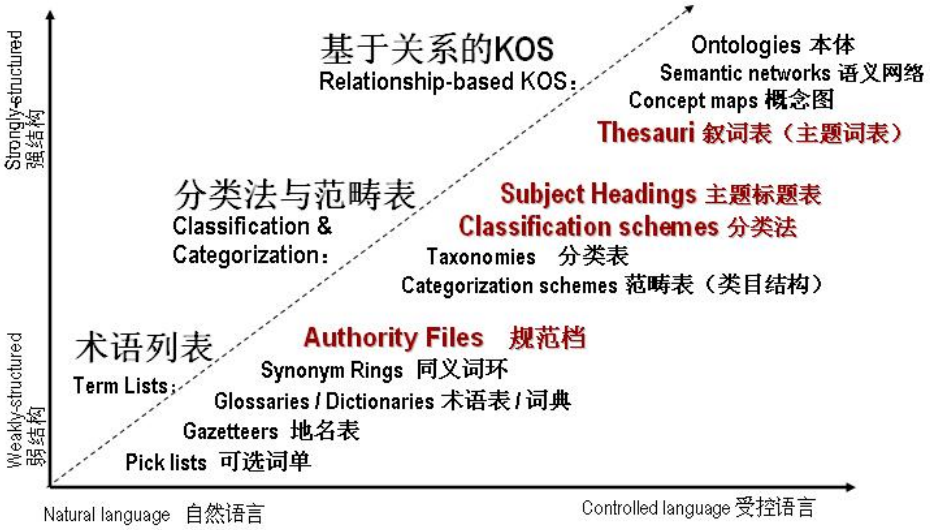

图 1　KOS 类型分布图[4][3][2][5]

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的、提供多义性（ambiguity）控制的各种术语列表，到具有等级关系控制的、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）的、网状结构的知识组织系统类型。

叙词表和本体同属于最后这种结构最复杂、控制程度最高的类型。我们有理由相信：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，其它低端 KOS 的这些问题只是它的简化，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[6][2][5][7][8][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：1) KOS 的电子化；2) HTML 表示的 KOS；3) 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出以下综合评述。

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，也便于将它们与相应的电子资源集成在一起。例如：在 INSPEC（英国科学文摘）和 EI（工程索引）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。上世纪九十年代以后，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。

一种是基于静态 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。

另一种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》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，它代表了面向 Web 的本体表示语言的最新发展趋势。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。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。也就是说，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是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上世纪八九十年代。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。上世纪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不小的差距。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。我们可以从这些相关领域的研究进展中得到一些启示，以使我们的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、Web Service 系列标准等。基于语义 Web 技术表示的 KOS 可以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，本体开发工具包 Jena 等，这些都为我们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自己的看法。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，这一点已经成为国际上有关专家的共识。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，贡献我们的本体知识库。这样才能使我们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有参考文献[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共同和不同之处，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/T3860-95）[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。”

我们再来看一看本体的定义。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，也就是说，它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。也就是说，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，它们的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，我们从它们的定义可以看出，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，因此它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，我们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，它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。我们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，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。

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）。它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（见图 1），例如中文叙词表（尤其是当我们希望将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，是因为它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，我们不可能依赖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，它遵从OWL DL规范，可实现完全的推理。主要创新点如下：

•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）

•参考ANSI/NISO Z39.19-2005[17]的第8节（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。

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（我们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		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					
Concept	CLCCode	&xsd;st	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	ring	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。

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一百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. 15-09-2006[EB/OL]. [2006-10-17]. http://nkos.slis.kent.edu/ .

[4] Marcia Lei Zeng and 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[演示文稿 ppt 打印件]. 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]. 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, Editors. 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]. 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(7):6-10.

[24] A. Moreira, L. Alvarenga, 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 3860-95, 文献叙词标引规则[S]. 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 .