《中国分类主题词表》的 OWL 表示及其语义深层揭示研究[1]

曾新红[2]

（深圳大学图书馆 深圳 518060）

摘　要　该文阐述了将中文叙词表转换成网上可共享本体的意义。在借鉴国外相关研究成果的基础上，提出了用 OWL（Web Ontology Language）表示《中国分类主题词表》的具体方案，并就词表中存在的大量复合概念的深层语义揭示提出了解决意见。

关键词　中国分类主题词表 本体表示 OWL 复合概念 隐含语义

分类号　G254.24-39

Research on Representation of Chinese Classified Thesaurus in OWL and Its Implied Semantic Reveal

Zeng Xinhong

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

Abstract　This paper discusses the significance of representing Chinese thesauri in OWL as ontologies which are available on the Web. Based on the related research work of other experts, it presents a scheme for converting Chinese Classified Thesaurus(CCT) to an ontology in OWL Lite. It also puts forward a solution for revealing the semantic behind the compound concepts that exist in the CCT largely.

Keywords Chinese Classification Thesaurus, CCT, Ontology representation, OWL, Compound concept, Implied semantic

1　前言

随着网络技术的普及和发展，信息标引和检索的范畴已超出了文献信息机构。人们正在致力于将现有的以 HTML 技术为基础的 Web 网络发展成为以 XML、RDF、本体（ontology）等技术为基础的语义 Web，从而使 Web 上的信息具有计算机可理解的语义，给人们的生活和科学研究工作带来更大的便利。在这个过程中，本体起着为 XML 和 RDF 等技术提供语义支持的关键作用，是语义 Web 的核心。当前，本体已成为信息技术界的研究热点。

本体本来是一个哲学范畴，意为“客观存在的一个系统的解释和说明，客观现实的一个抽象本质”。在信息技术界，本体被赋予了新的含义。它是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，在知识工程、数字图书馆、软件重用、信息检索和 Web 上异构信息的处理、语义 Web 等领域具有广泛的应用前景。对本体的定义有很多，其中得到广泛认可的是 Studer（1998）在 Gruber（1993）和 Borst（1997）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”。这包含四层含义：（1）概念模型（conceptualization），指通过抽象出客观世界中一些现象（Phenomenon）的相关概念而得到的模型。概念模型所表现的含义独立于具体的环境状态。（2）明确（explicit），指所使用的概念及使用这些概念的约束都有明确的定义。（3）形式化（formal），指本体是计算机可读的（即能被计算机处理）。（4）共享（share），指本体中体现的是共同认可的知识，反映的是相关领域中公认的概念集，即本体针对的是团体而非个体的共识。[1][2]

从本体的定义来看，它与图书馆学中的规范化词表（叙词表）有着许多相似之处。笔者

---

[1] 国家自然科学基金项目“本体继承机制研究”（项目号 60373084）和广东省自然科学基金项目“面向语义 Web 的本体构造”（项目号 04011304）成果论文。

[2] 曾新红，女，1968 年出生，1992 年北京大学图书馆学情报学系科技情报专业毕业，硕士，副研究馆员，主要研究领域：数字图书馆相关技术。

认为，有着几十年词表编纂和标引、检索研究和实践经验的图书馆学界，可以在本体的研究中发挥出自己的独特作用。叙词表的知识性和科学性还可以继续发展和提高，其发展方向就是构建可解决网络环境下的信息检索问题的本体。叙词表与本体的研究是相辅相成的，一方面，收录有大量规范术语与词间关系的叙词表可以弥补计算机专家研究本体时在词汇术语研究方面的不足（电子版叙词表已经初步具备了本体的特征），同时，本体技术的引入也可极大地推动叙词表的自动化管理，实现其动态更新完善，丰富其词间关系，并更容易被公众通过网络获得和使用，使其使用范围扩展至网络信息的标引和检索。

《中国分类主题词表》[3]是一部国家级的大型综合性分类主题一体化叙词表，其基础是《中国图书馆图书分类法》和《汉语主题词表》，共收录分类法类目 5 万余个，主题词及主题词串 21 万余条，包括哲学、社会科学和自然科学所有各个领域的学科和主题概念。该词表由北京图书馆等 40 个单位历时 8 年编纂而成，于 1994 年正式出版，现广泛应用于全国各类型图书馆和信息机构的文献标引工作。笔者认为，其权威性、科学性和知识性使其成为中文叙词表本体表示的首选实例。采用面向 Web 的国际标准本体语言 OWL 将该词表表示为本体形式，可以使其成为网络上共享的、与具体的计算机系统实现无关的国际通用中文本体，从而能够在网络信息环境中发挥出更大的作用。同时，也可为其他中文专业叙词表的本体化表示提供有益的借鉴。

2　研究现状

从目前国内外相关研究的动态来看，已经有了一些比较成熟的本体实现和应用技术。如本体的建模元语、本体的描述语言、构造本体的规则等都已日趋成熟并逐渐取得了共识。[1][2]近年来，国内外也已有一些学者采用构建叙词表（thesaurus）的方式来开发本体。[4][5]

2002 年以来，本体的研究也逐渐引起了大陆图书馆界学者的注意，他们就本体在图书馆界的应用前景、本体与词表的关系、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[6]-[13]。中国科学院文献情报中心的毛军在文献[6]中研究了叙词表的 RDF 表示方法，提出将叙词表的微观结构（叙词+关系）作为一个基本的语义单元进行处理，并且将叙词用概念和词汇两个层次的资源来描述，将原来的“用、代、属、分、参”关系分别净化和简化为“属和参”和相应的 RDF 属性。

最近有两项与本文密切相关的最新研究成果。一项是美国国家癌症研究所（National Cancer Institute, NCI）于 2003 年公布的 NCI Thesaurus 的 OWL 版本[14][15]，其深层次的语义关系揭示、科学的维护和更新流程很值得我们借鉴。另一项是用于语义 Web 的 TIF（Thesaurus Interchange Format，叙词表交换格式），它是在针对 W3C 欧洲语义 Web 先进发展计划（SWAD-Europe）的 Workpackage 8（Thesaurus Research Prototype） Deliverable 8.1/8.2 的一篇投稿中提出来的，版本时间为 2003 年 7 月 31 日，作者是英国 CCLRC 的 Brian Matthews 等[16]。这篇题为 Modelling Thesaurus for the Semantic Web（为语义 Web 建模叙词表）的文章提出了一项“基于概念”的初始草案标准，该标准与该领域中的 ISO 标准兼容。该文首先描述了为叙词表建模的两种选择：面向概念（concept-oriented）和面向术语（term-oriented）模式，并将它们与叙词定义的 ISO 标准关联。然后给出了将 TIF 建立在面向概念模式之上（并带有一个扩展，允许简洁的面向术语表示法）的理由。该文建立了一个 TIF RDF Schema（面向概念）和一个 TIFS RDF Schema（面向术语），并讨论了将它们定义成 OWL 本体的可能效益，分别给出了 OWL 本体表示版本：TIF OWL Ontology 和 TIFS OWL Ontology。我们还将在第 4 节分析它们的相关内容。

3　OWL 简介

OWL（Web Ontology Language）[17]-[21]是一种用于在语义 Web 上发布和共享本体的语

义置标语言，由 W3C 的本体工作组开发，2004 年 2 月 10 日成为 W3C 正式推荐标准。它代表了面向 Web 的本体表示语言的最新发展趋势。它是 RDF/S 的一种扩展，并源自 DAML+OIL Web 本体语言。OWL 能够被用来清晰地表达词汇表中的词汇含义以及这些词汇之间的关系（这些词汇和它们之间的关系的表达就称作本体）。OWL 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。选择 OWL 来表示和扩展《中国分类主题词表》和其他中文叙词表可以保证所建立本体的高质量和国际通用性。

OWL 包括三个子语言：OWL Lite, OWL DL 和 OWL Full。OWL Lite 包含 OWL 的某些基本要素（feature），并且做了限制，对于工具开发者来说比较容易支持。OWL DL 包括 OWL 的所有结构，但是设置了许多约束，适用于那些需要在推理系统上进行最大程度表达的用户，即推理系统能够保证计算完全性（所有结论都能够被计算出来）和可判定性（所有的计算都在有限的时间内完成）。OWL Full 支持那些需要在没有计算保证的、语法自由的 RDF 上进行最大程度表达的用户，目前没有任何推理软件可以完全支持它。从目前已有的研究来看，一般选择使用 OWL Lite 来表示叙词表[14][16]。本文所涉及的内容也落在 OWL Lite 的表达范围之内，因此也选用 OWL Lite。但笔者认为，OWL Lite 的表达能力有限，从发展的眼光来看，应采用 OWL DL 来表示中文叙词表，这样可以给中文叙词表本体的深层语义扩展留下余地。OWL DL 的良好推理能力也是检索系统所需要的，现在已有工具开发者开发出了支持 OWL DL 的强有力的推理系统。随着时间的推移，支持 OWL DL 的推理软件必然会大量涌现，届时将无需再为了实现的方便而束缚手脚。

OWL 进行交换的标准语法是 RDF/XML，它具有与 RDF 和 RDF Schema 的最大兼容性，这些 XML 和 RDF 格式和标准同样是 OWL 标准的组成部分。

下面介绍我们将要用到的 OWL Lite 词条的含义和语法（注意，我们在所举实例中采用的是面向概念模式，与传统的面向术语模式有所不同）：

3.1 类

• Class

定义了一组共享了某些相同属性的 individual。这些 individual 又称为这个类的实例（instance）。在 OWL Lite 及 DL 中，一个 individual 不能同时又是一个 Class。Class 能够通过 subClassOf 定义出一个特定的类层次。有一个内置的公共类 Thing，它是所有 individual 的 Class，也是所有 class 的 superclass。例如：

```
<owl:Class rdf:ID=”Concept”/>
```

定义了一个名为 Concept 的类，具体的概念“考古学”、“考古技术”都是这个 Class 的成员（individual）。

• rdfs:subClassOf

定义一个 Class 的子类。子类继承了父类的所有属性，子类的 individual 同时也是父类的 individual。类层次可以通过给出一个类是另一个或多个类的子类这样的声明来创建。例如：

```
<owl:Class rdf:ID=”PersonConcept”>
    <rdfs:comment>The class of person concept</rdfs:comment>
    <rdfs:label>人物</rdfs:label>
    <rdfs:subClassOf rdf:resource=”#Concept”/>
</owl:Class>
```

定义了一个类 PersonConcept（人物概念），它是 Concept 类的子类。同样的道理，我们也可以将 CompoundConcept（先组的主题词串（复合概念））、RegionConcept（地名概念）定义为 Concept 的子类。

可以使用以下语法来定义一个 Class 的成员：

```
<Concept rdf:ID=”考古学”/>
```

以上语法定义了 Concept 类的一个实例“考古学”。

• owl:equivalentClass

用于创建同义类，即将两个类声明为相同，它们拥有不同的名字却拥有相同的 individual 集合。我们可以用它来在不同词表本体之间为不同名称的相同类之间建立映射。两个 individual 之间的等同关系则需要使用 owl:SameAs 来定义。

3.2 属性

• rdfs:Property

OWL 的属性，主要包括以下两种：

①Object property：表达 individual 之间的关系。例如词表中概念与术语之间的关系（具有正式主题词、非正式主题词）就可以定义为 Object property：

```
<owl:ObjectProperty rdf:ID=”HasPreferredTerm”>
  <rdfs:comment>has preferred term</rdfs:comment>
  <rdfs:label>正式主题词</rdfs:label>
  <rdfs:domain rdf:resource=”#Concept”/>
  <rdfs:range rdf:resource=”#PTerm”/>
</owl:ObjectProperty>
```

②Datatype property：表达 individual 和数据值（data value）之间的关系。例如词表中的概念与分类号、概念和范围注释（scope note）之间的关系就可以定义为 Datatype property：

```
<owl:DatatypeProperty rdf:ID=”CLCCode”>
  <rdfs:comment>Chinese Library Classification code</rdfs:comment>
  <rdfs:label>中图法分类号</rdfs:label>
  <rdfs:domain rdf:resource=”#Concept”/>
  <rdfs:range rdf:resource=”&rdfs;literal”/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID=”ScopeNote”>
  <rdfs:comment>scope note</rdfs:comment>
  <rdfs:label>注：</rdfs:label>
  <rdfs:domain rdf:resource=”#Concept”/>
  <rdfs:range rdf:resource=”&rdfs;literal”/>
</owl:DatatypeProperty>
```

（在面向概念模式中，分类号和范围注释属性只与概念相关，见第 4 节）

OWL 另外还有两种属性：注释属性（owl:AnnotationProperty）和本体属性（owl:OntologyProperty）。OWL DL 允许对 class、property、individual 和 ontology header 进行注释，但规定不能为注释属性定义子属性或 domain/range 限制。注释属性对象必须要么是一个数据字符串（data literal）或一个 URI 引用（URI reference），要么是一个 individual。OWL 有五种预先定义的注释属性：owl:VersionInfo，rdfs:label，rdfs:comment，rdfs:seeAlso，rdfs:isDefinedBy，可以直接使用。如我们前面已经用到了 rdfs:label 和 rdfs:comment 来规定属性的显示名称和解释属性的含义。本体属性用于表达本体之间的关系，如本体的引进（owl:imports）可以用来将一个已存在的本体引入当前本体，从而可以重用本体和分布式构造本体。本体属性还有几个实例分别用来表示本体之间的版本关系和兼容关系，在此就不详细介绍了。

OWL 的属性之间的关系有：子属性（owl:subPropertyOf），等同属性（equivalentProperty），翻转性（inverseOf）。属性特征有：传递性（TransitiveProperty），对称性（SymmetricProperty），

值唯一性（FunctionalProperty），翻转值唯一性（InverseFunctionalProperty）等。

如中文叙词表中概念与术语之间的 HasPreferredTerm 和 IndicateFormally 关系就是互为翻转的属性（若 A P B，则 B Q A，那么 P 和 Q 就是互为翻转的属性）：

```
<owl:ObjectProperty rdf:ID="IndicateFormally">
    <rdfs:comment>indicate formally</rdfs:comment>
    <rdfs:label>正式表示</rdfs:label>
    <owl:inverseOf rdf:resource="#HasPreferredTerm"/>
</owl:ObjectProperty>
```

而概念间的属、分关系则既是互为翻转的属性，又都具有传递性（A 属 B，B 属 C，则 A 属 C）：

```
<owl:ObjectProperty rdf:ID="HasBroaderConcept">
    <rdfs:comment>has broader concept</rdfs:comment>
    <rdfs:label>属</rdfs:label>
    <rdf:type rdf:resource="&owl;TransitiveProperty"/>
    <rdfs:domain rdf:resource="#Concept"/>
    <rdfs:range rdf:resource="#Concept"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="HasNarrowerConcept">
    <rdfs:comment>has narrower concept</rdfs:comment>
    <rdfs:label>分</rdfs:label>
    <rdf:type rdf:resource="&owl;TransitiveProperty"/>
    <owl:inverseOf rdf:resource="#HasBroaderConcept"/>
</owl:ObjectProperty>
```

概念之间的参见关系则是具有对称性的属性（A 参 B，则 B 参 A）：

```
<owl:ObjectProperty rdf:ID="HasRelatedConcept">
    <rdfs:comment>has related concept</rdfs:comment>
    <rdfs:label>参</rdfs:label>
    <rdf:type rdf:resource="&owl;SymmetricProperty"/>
    <rdfs:domain rdf:resource="#Concept"/>
    <rdfs:range rdf:resource="#Concept"/>
</owl:ObjectProperty>
```

具体的类和属性定义见第 4 节。限于篇幅，第 4 节将不再一一列举所有类和属性的具体定义语法，其他类和属性的定义可参照以上示例进行。

4 《中国分类主题词表》的 OWL 表示

本节主要讨论以下几个问题：1.概念模式的确立；2.类和属性的确立；3.复合主题中语义的深层揭示。

4.1 概念模式的确立

Brian Mathrews 等[16]强烈建议采用面向概念模式，认为这符合 ISO2788:1986 的要求。因为 ISO2788:1986 指出，术语组是一个首选词（preferred term，也称为标引词或叙词）和它的非首选词（non-preferred term，也称为同义词）选项为存在于人类思想领域的某个抽象概念构成的一组可能的标签，换句话说，就是术语代表概念。在面向概念模式中，叙词表的等级和相关关系在概念之间声明。概念是这个概括层次结构（generalisation hierarchy）中的节点。在一个多语种叙词表中，等同关系在概念之间声明。相对于传统的面向术语模式，面

向概念模式将概念和术语分离，更容易维护和更新，因为对术语的修改不会干扰概括层次结构本身。而且 ISO5964:1985 强烈建议，多语种等同关系只在首选词之间设定。面向概念模式通过只设定概念之间的等同关系而执行了这个建议，这也可以简化在同一领域的两个相似叙词表之间创建映射的过程，这一点对于叙词表在语义 Web 上的应用尤为重要，因为语义 Web 上可能会有许多不同的个体创建他们自己的、涉及同一领域的标引体系。

《中国分类主题词表》[3]形式上是面向术语的，但其每一个正式主题词（即首选词）都可以被视为一个概念，因此可以采用面向概念的模式将其表示为本体。即词表中的每一个正式主题词都既表示为 Concept 类的 individual，也表示为 PTerm 类的 individual。属、分、参等关系在概念与概念之间声明，分类号和范围注释（Scope Note）属性也面向概念定义而不再面向术语定义。

4.2 类和属性的确立

我们希望所建立的本体既要符合 OWL 语言的规范，具有良好的可推理性，又要兼顾词表的更新维护，能够方便地转换成高效率的数据库结构和生成传统的书本式叙词表和各种对照表。因此，类和属性的确立就显得尤为重要。

笔者仔细分析了 NCI Thesaurus OWL 版本[15]和 TIF OWL 版本[16]的结构，从中发现了许多可借鉴的东西，也摒弃了一些不合理的设计。例如，NCI Thesaurus 的 OWL 版本中，将每一个具体的概念都定义为一个单独的类，许多属性表示的是类与类之间的关系，而不是 individual 之间的关系。而 OWL 的类可以定义一组共享了某些相同属性的 individual，叙词表中的每一个具体概念其实都可以定义为 Concept 类的 individual。又例如在 TIF OWL 版本中，在 Concept 类和 Term 类之上还定义了 Thesaurus 类和 ThesaurusObject 类。定义 Thesaurus 类的目的是为了方便添加描述叙词表名称、创建者、修改日期等的属性，这实际上可以通过定义 OWL 头和注释属性来实现；定义 ThesaurusObject 类的目的是为了定义一个所有叙词表类的超类，这似乎也无必要，OWL 中已有一个预定义的超类 OWL:Thing，它是所有 individual 的超类。TIF 中这两个类的定义会增加类层次的复杂性，但并没有多少实际意义。另外，TIF OWL 版本还将一些注释也定义为类，如 ScopeNote，GeneralNote, HierarchyNote 等，笔者也认为不太合理。

笔者对《中国分类主题词表》的结构进行了深入的分析，认为定义以下类和属性较为合适，既能够较好地表示出词表的语义结构和概念间关系，同时也可以兼顾本体形式和传统词表形式之间的互相转换。

表 1 类定义

类名称	含义	OWL 定义或说明
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。词表中所有主题词串（复合主题）都是这个类的 individual。我们将在 4.3 节讨论复合概念的深度语义表示。	<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	附录二“人物”中列出的主题概念是这个类的

	类。	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。
World RegionConcept	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。
Term	概念的具体表现形式：术语。词表中所有的主题词都是这个类的 individual。	包括正式主题词、非正式主题词。
PTerm	Preferred term，正式主题词，它是 Term 类的子类。	
NTerm	Non-preferred term，非正式主题词，它也是 Term 类的子类。	

表 2 属性定义

Domain	Property	Range	属性特征
	ObjectProperty		
Concept	HasPreferredTerm	PTerm	与 IndicateFormally 互为翻转属性。
Concept	HasNonpreferredTerm	NTerm	与 IndicateInformally 互为翻转属性。
Concept	HasBroaderConcept	Concept	具有传递性。与 HasNarrowerConcept 互为翻转属性。
Concept	HasNarrowerConcept	Concept	具有传递性。
Concept	HasRelatedConcept	Concept	具有对称性。
PTerm	IndicateFormally	Concept	与 HasPreferredTerm 互为翻转属性。
NTerm	IndicateInformally	Concept	与 HasNonpreferredTerm 互为翻转属性。
	DatatypeProperty		
Concept	CLCCode	&rdfs;literal	
Concept	ScopeNote	&rdfs;literal	

下面我们给出一个词族的示例：

<PTerm rdf:ID="地热"/>

<Nterm rdf:ID="地热资源"/>
<PTerm rdf:ID="大地热流"/>
<Pterm rdf:ID="地热蒸汽"/>
<Pterm rdf:ID="地下热水"/>
<Pterm rdf:ID="热矿水"/>
<Nterm rdf:ID="天然蒸汽"/>
<Concept rdf:ID="地热">
<rdfs:comment>Terrestrial heat</rdfs:comment>
<HasPreferredTerm rdf:resource="#地热"/>
<HasNonpreferredTerm rdf:resource="#地热资源"/>
<HasNarrowerConcept rdf:resource="#大地热流"/>
<HasRelatedConcept rdf:resource="#地热能"/>
<CLCCode>P314</CLCCode>
</Concept>
<Concept rdf:ID="大地热流">
<rdfs:comment>Terrestrial heat flow</rdfs:comment>
<HasPreferredTerm rdf:resource="#大地热流"/>
<HasBroaderConcept rdf:resource="#地热"/>
<HasNarrowerConcept rdf:resource="#地热蒸汽"/>
<HasNarrowerConcept rdf:resource="#地下热水"/>
<CLCCode>P314.2</CLCCode>
</Concept>
<Concept rdf:ID="地热蒸汽">
<rdfs:comment>Geothermal vapour</rdfs:comment>
<HasPreferredTerm rdf:resource="#地热蒸汽"/>
<HasNonpreferredTerm rdf:resource="#天然蒸汽"/>
<HasBroaderConcept rdf:resource="#大地热流"/>
<HasBroaderConcept rdf:resource="#蒸汽"/>
<HasRelatedConcept rdf:resource="#地下热水"/>
<CLCCode>P314.1</CLCCode>
<CLCCode>TK521^{+}.31</CLCCode>
</Concept>
<Concept rdf:ID="地下热水">
<rdfs:comment>Geothermal water</rdfs:comment>
<HasPreferredTerm rdf:resource="#地下热水"/>
<HasBroaderConcept rdf:resource="#大地热流"/>
<HasNarrowerConcept rdf:resource="#热矿水"/>
<HasRelatedConcept rdf:resource="#地热蒸汽"/>
<HasRelatedConcept rdf:resource="#温泉"/>
<CLCCode>P314.1</CLCCode>
<CLCCode>TK521^{+}.33</CLCCode>
</Concept>
<Concept rdf:ID="热矿水">
<rdfs:comment>Thermal mineral water</rdfs:comment>

```
  <HasPreferredTerm rdf:resource="#热矿水"/>
  <HasBroaderConcept rdf:resource="#地下热水"/>
  <CLCCode>P314.1</CLCCode>
  <CLCCode>TK521+.33</CLCCode>
</Concept>
```

“蒸汽”和“温泉”是其他词族中的概念，也会在该本体中定义。资源定义和引用没有绝对的顺序要求，甚至可以引用存在于其他文件中的资源，从而可以构造分布式的本体。英文注释则取自《汉语主题词表》[22]。词族层次关系可以通过属性 HasBroaderConcept 和 HasNarrowerConcept 推理出来，所以没有定义 HasTopConcept（族首词）属性。

4.3 复合主题中语义的深层揭示

《中国分类主题词表》中存在着大量的先组主题词串（主题款目）。它们是一些复合概念，由多个简单概念组配而成，采用“：”、“—”、“，”等组配符号。这些符号，尤其是“—”隐含着十分丰富的语义关系。我们可以利用 OWL 的属性定义对其进行深层的揭示。

“：”和“，”的含义比较纯粹，分别表示交叉组配和限定组配，其语义比较好理解，而且是国际通用的符号，因此在本体表示中可不必做进一步的语义揭示。当“—”用于表示联结组配，表达事物与事物之间的关系、比较、影响、作用、应用等类型的联系时，两个事物之间的关系语义已由中间的功能词表达出来了，也可以不再做进一步的语义揭示。

但“—”还可以表示其他的语义。《文献主题标引规则》（GB3860-83）规定，主题构成的因素及其序列可分为主体因素（研究对象、材料、方法、结果、条件等）、通用因素、位置因素、时间因素、文献类型因素等五种。主题因素构成的五种因素即五个范畴。主题词表中的每个主题词必定属于其中的某个范畴，每个范畴分别表示属于该范畴的主题词在主题款目（组配标题）中的职能，而主题款目中的各个主题词都须按照五个范畴的规定次序进行排列。该标准的起草人之一刘湘生对此做了进一步的说明，如下表所示：

主题因素		组配次序	代码符号
主体因素	对象（学科、事物、问题）	1	A1
	方面（材料、成分、性质、过程、状态、特征、作用、现象）	2	A2
	方法	3	A3
	结果	4	A4
	条件	5	A5
通用因素		6	B
位置因素		7	C
时间因素		8	D
文献类型		9	E

表 3 主题构成因素及其序列

中文文献复合主题的次序公式是：A(A1—A2—A3—A4—A5)—B—C—D—E。[23]

由此可见，组配符“—”的背后隐含着相当丰富而复杂的语义关系。在向本体转换的过程中，有必要对某些隐含语义进行深层次的、明确的揭示，这样才能消除理解中可能出现的歧义，也有助于在检索系统中实现更专指和更深层次的语义推理。

我们可以通过以下属性定义来实现中文复合主题（复合概念）的深层语义揭示（这些属性的 Domain 均为 CompoundConcept）：

ObjectProperty	Range	说明
PrincipalFactor	Concept	主体因素。含义：复合概念 X 具有主体因素 Y（Concept）。

ObjectFactor	Concept	主体因素中的对象因素，是 PrincipalFactor 的子属性。
DisciplineFactor	Concept	对象因素中的学科因素，是 ObjectFactor 的子属性。
ThingFactor	Concept	对象因素中的事物因素，是 ObjectFactor 的子属性。
IssueFactor	Concept	对象因素中的问题因素，是 ObjectFactor 的子属性。
AspectFactor	Concept	主体因素中的方面因素，是 PrincipalFactor 的子属性。
MaterialFactor	Concept	方面因素中的材料因素，是 AspectFactor 的子属性。
IngredientFactor	Concept	方面因素中的成分因素，是 AspectFactor 的子属性。
QualityFactor	Concept	方面因素中的性质因素，是 AspectFactor 的子属性。
ProcessFactor	Concept	方面因素中的过程因素，是 AspectFactor 的子属性。
StateFactor	Concept	方面因素中的状态因素，是 AspectFactor 的子属性。
CharacterFactor	Concept	方面因素中的特征因素，是 AspectFactor 的子属性。
FunctionFactor	Concept	方面因素中的作用因素，是 AspectFactor 的子属性。
PhenomenonFactor	Concept	方面因素中的现象因素，是 AspectFactor 的子属性。
MethodFactor	Concept	主体因素中的方法因素，是 PrincipalFactor 的子属性。
ResultFactor	Concept	主体因素中的结果因素，是 PrincipalFactor 的子属性。
ConditionFactor	Concept	主体因素中的条件因素，是 PrincipalFactor 的子属性。
GeneralFactor	GeneralConcept	通用因素
LocationFactor	RegionConcept	位置因素
TimeFactor	EraConcept	时间因素
DocumentType	Concept	文献类型

表 4 隐含语义的属性定义

主体因素是复合概念中必不可少的因素，其他四种因素都是对主体因素起修饰限定作用的因素，但不是每个复合概念中都含有这四种因素，应视具体情况而定。对象因素和方面因素的子属性细分比较复杂，表中没有也不太可能枚举所有可能因素，因此除非有多个子属性并存共同构成对象因素或方面因素而需要细化到子属性，一般情况下可以直接使用父属性（ObjectFactor, AspectFactor）。《中国分类主题词表标引手册》[23]对各种因素进行了详细的说明，在具体的本体转换表示过程中可以参考这些说明进行区分和判定。例如，复合概念“建筑—空气净化—原理”可以表示为：

```
< CompoundConcept rdf:ID="建筑—空气净化—原理">
  <ObjectFactor rdf:resource="#建筑"/ >
  <AspectFactor rdf:resource="#空气净化" />
  <GeneralFactor rdf:resource="#原理"/>
  <CLCCode>TU834.8</CLCCode>
< /CompoundConcept>
```

一个复合概念在本体中出现一次就可以了，其轮排形式不必重复定义。

5 结束语

要真正实现完善的本体，现有词表中的词汇及词间关系也存在着局限性，因此，实现初始本体的自动更新完善也是非常重要的。从图书馆界的角度来看，词表更新速度赶不上标引工作需要的问题也长期困扰着我们，也是一个亟需解决的问题。如《中国分类主题词表》出版至今已近 10 年。这十年当中由于科学和社会的发展出现了许多新词汇，因词表不能及时增补新的概念和新的名词术语，导致某些新领域的文献标引过粗、不够简练或过于勉强。随着时间的推移，这个矛盾越来越突出，许多单位开始尝试增补新的主题词或增加自由词标引，但没有一个行之有效的机制来广泛收集、甄别和利用这些来自标引实践第一线的新术语。值得欣慰的是，1999 年国家图书馆将该词表的修订提上了议事日程，可以想象，这项庞大和

复杂的词表修订工程又将耗费专家们多年的心血。笔者认为，面对日新月异的社会和科学发展，词表和本体的完善只靠少数专家在指定的时间内突击进行是远远不够的。图书馆界的计算机主题标引和读者公共检索系统已经是成熟和普及的技术，这种标引和检索实践几乎每天都在进行。标引员对文献进行主题标引和读者构造主题检索式进行文献检索，这两种实践本身就是复杂的知识、概念分析过程，我们应该利用面向 Web 的、独立于具体实现的方式来充分采集这些分布式的标引结果和检索需求，建立一种科学的集中统计分析机制，提取新的词汇和词间关系，集中公众的智慧来完善词表和本体。这样既能彻底解决困扰图书馆界多年的词表更新和维护问题，又能建立动态完善的共享本体，从而满足图书馆界和信息技术界的共同需求。限于篇幅，笔者将另外撰文详细探讨这个问题。

参考文献：

1 刘昕鹏. Ontology 理论研究和应用建模——《Ontology 研究综述》、W3C Ontology 研究组文档以及 Jena 编程应用总结. URL: http://gis.pku.edu.cn/Resources/TR/, (Accessed Feb 1, 2004)

2 邓志鸿等. Ontology 研究综述. 北京大学学报（自然科学版），2002，38（5）：730-738

3 中国分类主题词表. 华艺出版社，1994

4 Beck HW, Lin N, Xin J, Otuka A, Laurenson M. Web taxonomy: A thesaurus and database, Proceedings of the World Congress of Computers in Agriculture and Natural Resources, 2001

5 Soo VW, Lee CY, Li CC, Chen SL, Chen CC. Automated semantic annotation and retrieval based on sharable ontology and case-based learning techniques. 2003 JOINT CONFERENCE ON DIGITAL LIBRARIES, PROCEEDINGS

6 毛军. 基于 RDF 的叙词表研究. 情报学报，2003，（2）

7 董慧，杜文华. 基于本体和多代理的数字图书馆信息检索模型. 中国图书馆学报 ,2004, (2)

8 刘柏嵩. 面向语义网的本体表示，中国图书馆学报，2004（2）

9 贺纯佩、李思经. 农业叙词表在中国的发展和农业本体论展望. 现代图书情报技术，2003，（4）

10 常春. Ontology 在信息管理领域的研究背景. 现代图书情报技术，2003，（6）

11 陈文彬. Ontology 在图书服务网络中的应用. 现代图书情报技术，2003，（6）

12 董慧. 基于本体论和数字图书馆的信息检索. 情报学报，2003，（6）

13 楼向英. Ontology：概念及其在数字图书馆中的应用. 图书馆杂志，2002, (11）

14 Jennifer Golbeck et al.. The National Cancer Institute's Thesaurus and Ontology. http://www.mindswap.org/papers/webSemantics_NCI.pdf, (Accessed Mar 16, 2004)

15 nciOncology.owl(version 03.09d). http://www.mindswap.org/2003/CancerOntology, (Accessed Mar 17, 2004)

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

17 OWL Web Ontology Language Overview. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-owl-features-20040210/

18 WL Web Ontology Language Reference. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-owl-ref-20040210/

19 WL Web Ontology Language Guide. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-owl-guide-20040210/

20 OWL Web Ontology Language Semantics and Abstract Syntax. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-owl-semantics-20040210/

21 OWL Web Ontology Language Use Cases and Requirements. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-webont-req-20040210/

22 中国科学技术情报研究所《汉语主题词表》自然科学部本维护组.《汉语主题词表》自然科学（增订本）. 科学技术文献出版社，1991

23 陈树年主编. 《中国分类主题词表》标引手册（上）. 北京图书馆出版社. 1998