[文章编号] 1004-325X(2008)08-0095-05

近年来领域本体的应用新进展

Review of Application of Domain Ontology from 2006 to 2008

余 倩（中国科学院国家科学图书馆 中国科学院研究生院 北京 100190）

[摘 要] 近年来，领域本体的发展非常迅速，不仅得到了广泛的应用，在实际应用中也取得了积极的作用。通过查找ISI、Elsevier等数据库查询到，近两年来国外文献中涉及的领域本体，包括化学领域、生物领域（分为基因领域和生物医学领域）、地理学领域和其他领域，并选取其中有代表性的领域本体并对其应用及进展进行分析研究，总结出领域本体应用进展的特点：涉及学科领域广；更加专业化、针对性更强；涉及多个学科的领域本体增多。领域本体的建设发展将有力推动数字图书馆的进步。

[关键词] 本体 领域本体 本体构建

[中图分类号] TP391; G250.76 [文献标识码] A

[Abstract] In recent years, the domain ontology has developed quickly, and has been used in practical application widely and also made a positive role. Through searching ISI, Elsevier database, the author finds out that the domain ontology involved in foreign literatures in recently two years includes chemical field, biological field(into genetic field and biomedical field), geography and other fields. Through the analysis and research on the representative domain ontology, we could get the features of the application development of the domain ontology, such as wider disciplines; more specialized and targeted; more domain ontology of more disciplines. Therefore, the development of the domain ontology will effectively promote the progress of digital libraries.

[Key words] Ontology; Domain ontology; Ontology construction

1 引 言

数字图书馆中包含了大量的信息，各种不同学科、不同存储格式、不同来源的信息资源交汇在一起，如何将各个这些数据信息有效获取、转换和利用，是数字图书馆普遍面临的问题。

N.Guarino提出将本体划分为顶级本体(top-level ontology)、领域本体(domain ontology)、任务本体(task ontology)和应用本体(application ontology)。其中领域本体是指描述特定领域(电信、汽车等)中的概念以及概念之间的关系。

对数字图书馆而言，领域本体在数字图书馆对其知识进行语义层面的组织中扮演着至关重要的角色，因此可以说，领域本体的构建是语义网络环境下数字图书馆知识组织不可或缺的关键步骤[1]。本文选取了2006至2008年间国外领域本体的应用发展作为分析对象，从化学、生物、地理学等多个学科领域的角度介绍了近两年来领域本体的新进展及应用。

2 化学领域的领域本体研究及应用

化学领域的本体出现的比较早，应用也比较广泛、成熟，早在上个世纪就有化学本体出现并应用。近年来，随着本体自身的发展和完善，化学领域的本体更是对化学学科自身的发展起着非常重要的作用。

2.1 本体在化学领域中的应用

目前有很多本体应用于化学领域，应用较为广泛的包括：

ChEBI[2](Chemical Entities of Biological Interest)是一个免费的分子实体字典，也是一部关于“小”的化学化合物的字典。在化学研究中，被讨论的分子实体不是天然产品就是合成产品，这些产品是用于干预活的生物体进程的。在ChEBI中，基因组编码的大分子(来自蛋白质卵裂的核酸、蛋白质和多肽)都不能作为一项规则。除了分子实体外，ChEBI还包括实体的群(部分分子实体)和等级。ChEBI包含了本体的分类，即分子实体或实体等级和制定它们的父/子实体之间的关系。

CO[3](The Chemical Ontology)是一种自动地基于化学官能团的新型化学本体，它客观地通过计算机程序分配对小分子进行分类。CO作为一种新型化学本体已经应用于PubChem数据库和一些描述小分子相互作用的数据库。利用CO，我们可以通过语义相似度评分的方法来比较分子，

这种评分方法是基于功能团的分配而不是3D形式，此方法可以成功地识别小分子，称为约束一个共同的结合位点。CO将作为一个搜索化学数据库和识别与生物活动相关的重点功能团的强有力工具。

BAO[4]（BACIIS本体）是一个为BACIIS（Biological And Chemical Information Integration System，生物和化学信息集成系统）建立的领域本体，其主要目标是：①引导用户构建有效的疑问；②方便解决各种数据格式和资料来源的多变性；③方便整合存储生物和化学数据的Web数据库。BACIIS实现了多个异构生命科学Web数据库的一体化，提供全球范围的跨数据库的免费获取[5]。

其它例如早期的描述陶瓷物质化学成分的本体Plinius Ontology，描述化学元素的化学元素本体，目前化学元素Chemical-Elements Ontologg等等。下面将详细介绍近年化学领域方面的新本体OntoCAPE的应用研究与进展。

2.2 OntoCAPE

OntoCAPE[6]（Ontology for Computer-Aided Process Engineering）是一个在化学处理工程领域正式的、重要的本体。OntoCAPE诞生于为CAPE（Computer-Aided Process Engineering，计算机辅助过程工程）建立一个本体的想法，最早开发于COGents项目，COGents项目是为化学处理的数字仿真开发一个基于代理的架构。OntoCAPE发展完善于IMPROVE项目，该项目着重关注新的概念和软件工程以解决支持协作完成的工程设计处理。OntoCAPE是基于一个用于过程工程综合的数据模型——CLiP，而在以往表现过程工程知识方面，特别是支持计算机辅助过程模型（Computer-Aided Process Modeling，CAPM）方面，成功的例子很少[7]。

OntoCAPE也第一次通过自然语言和图例非正式地描述。图例的格式是UML类型图表，用于提供概念的层次及它们间主要关系的图示。通过使用DAML+OIL将这种非正式的表达方式进一步转化为一种正式的表达方式。随着OntoCAPE的发展，本体模型语言将会更加规范地用于语义网。目前，OntoCAPE 1.0已经正式发布，并用于COGents工程。OntoCAPE由600个概念和400个关系组成。迄今为止，OntoCAPE已经被用于大量过程建模和设计应用。所有这些应用都有一个共同点，即已存在的软件工具均是通过一系列本体工程开发的，这些软件工具已经应用于正式规格的OntoCAPE的安装、处理和推理。

2.3 FGO

FGO[8]（Functional Group Ontology）实质是一个化学功能团的结构分类，这个化学功能团作为已有知识的一个重要信息来源可自动集成支持识别、分类和预测数据分析的任务。我们拥有了一个新的注解方法，这种方法可以从已有的本体表达法中选择正确的等同信息来构建源结构。

FGO在化学功能团注释化学成分方面是一个有效的、简单明了的知识表示方法。FGO的应用已经超出了单纯的注解和信息检索应用，这表明，FGO可能有助于在代表生物活性的功能的分部和分类方面的大规模预测应用。FGO对群集的小分子还采取了语义相似度评分的方法，群集的小分子远快于基于图形相似性测度，但在相似群集中可大量提取、挖掘和容纳药效构象（pharmacophoric）模式。更多化学结构共享类似注释，我们可以从中通过分析提取出FGO的共同的模式和坐标信息内容，这有助于寻找药效团模式。基于通过语义相似性和注释搜索得出的内容，有助于在匹配数据库中的亚结构搜索，以及在匹配的分类数据库中进行信息内容检索，上述同样可能有助于在碎片数据库（fragment data bases）中使用组合设计和重新匹配发展。

3 生物领域的领域本体研究及应用

生物学领域涉及比较广，与医学、化学等多个学科多有交叉，相关本体也多是与其他学科相结合，其应用也比较成熟、广泛。

选取近两年来比较热门的生物领域本体，分别介绍基因领域、生物医学领域等的本体应用及进展。

3.1 基因领域

GO（Gene Ontology）已经成为生物信息领域中一个极为重要的方法和工具，并正在逐步改变着我们对生物数据（Biological Data）的组织和理解方式，它的存在已经大大加快了我们对所拥有的生物数据的整合和利用。

GO[9]项目已发展成为3个结构控制词表（本体），在独立物种的前提下，用词表对物种相关的生物过程、细胞成分和分子功能方面描述其基因产物。GO项目的3个目标是：①开发和维护本体；②诠释基因产物；③发展工具以助创造、维修和使用本体。GO的应用非常广泛，与基因领域相关的大量数据库、数据仓库以及系统，都是基于GO开发或者建立的。

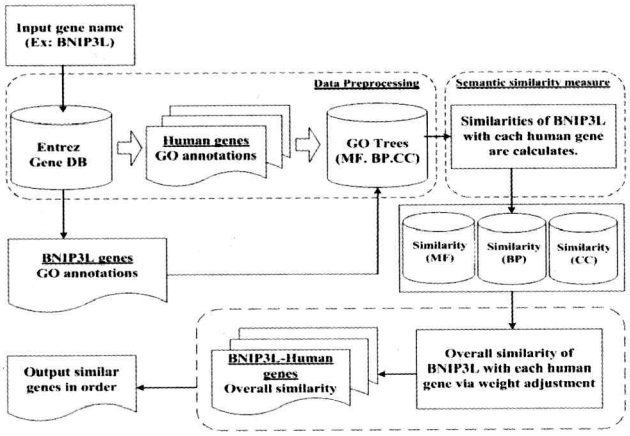
3.1.1 SGDS

SGDS（Similar genes discovery system，相似基因发现系统）[10]是基于GO和Entrez gene语义相似性措施建立的，用于确定一组相似的基因。SGDS由3个模块组成（如下页图1所示）：数据预处理模块、语义形似匹配模块和基因量化模块。SGDS在Entrez基因数据库中获取基因名称并寻找它的功能注解，被人们怀疑的基因和人类基因的语义相似会通过GO的3个方面分别地计算，全部记录是通过根据他们的不同分量得到的相似值来获取，最后，系统会根据语义相似程度输出基因列表。

SGDS的发展为说明基于基因本体和Entrez基因的语义相似测量提供了一种新颖的方法，这种方法有赖于基因本体分等级后的结构以及通过路径长度和深度的非线性功能量化后，两种GO术语之间的语义相似。与其他文献中提及的方法（如Risnik，Jiang，和Lin's测量）相比，文中的试验结果显示语义相似测量不比其它方法差。

3.1.2 GOHSE

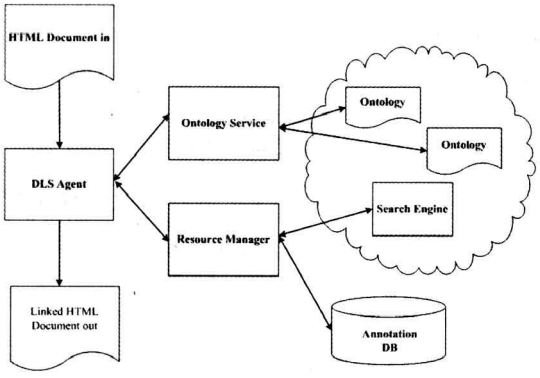
图 1　SGDS 示意图[10]



GOHSE[11] 系统是一个支持浏览生物资源的应用程序。GOHSE 通过动态的附加超文本链接来增强网络资源。GOHSE 将 COHSE（The Conceptual Open Hypermedia Service，概念开放式超媒体服务）应用于生物信息学，使用 GO 作为本体及相关的关键词映射，以 GO 联合体作为链接目标。

图 2 显示的是 GOHSE 系统的逻辑结构。GOHSE 代理接纳文件 / 页面，并产生页面以加强联系，为此 GOHSE 有赖于服务商提供的本体和资源方面的资料。GOHSE 代理商 OS 和 RM 使用的都是简单的 CGI 界面，这便使得该系统能够在一个相对宽松、松散耦合和开放式体系结构中利用现有的协议和结果。

图 2　GOHSE 系统的逻辑结构[11]



3.2　生物医学领域

毋庸多言，生物和医学是结合非常紧密的两个学科，因此有关这两个学科领域的本体多有交集。

3.2.1　FMA

FMA[12](the Foundational Model of Anatomy)是一个生物医学信息学方面的参考本体。领域参考本体论所代表的知识是通过领域理论独立于具体对象的。FMA 的目的是要提供一个实物和空间的概念用以命名人体。

我们可以用公式来表示FMA的组成：FMA=(AT，ASA，ATA，Mk)[13]，其中AT指解剖分类法(Anatomy Taxonomy)，ASA指解剖结构的抽取(Anatomical Structural Abstraction)，ATA 指解剖转换抽取(Anatomical Transformation Abstraction)，以及 Mk 引用自元知识(Metaknowledge)。

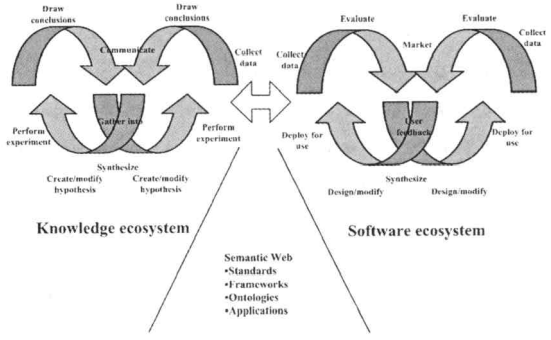
FMA 是用 Protégé构建，Protégé 框架语言是 FMA 的典型代表性语言[14]。目前我们将 FMA 的最初 Protégé框架表示法翻译成 OWL。我们致力于可以尽早实现在 OWL 中描绘 FMA，并注重两个目标：①只代表在 FMA 的框架表示法中是目前明确的或可直接从语义 protégé 框架中推断出来的信息；②代表目前 FMA 的框架表示法中所有的信息，从而产生相应 OWL 代表完整的 FMA。

3.2.2　SWAN

SWAN[15](Semantic Web Applications in Neuromedicine，语义网应用于神经医学)是一个跨学科项目，通过语义网技术启用社会的能源和自组织能力，为阿尔茨海默病(AD)研究界制定一个有效的专门知识基地。目前，SWAN 本体已经建立，它的建立有助于满足大量生物医学研究者工作的需求，也同大量的生物本体团体有着广泛的讨论与合作。

SWAN 建立的目的是对阿尔茨海默病实现分布式知识库模式，以及将知识库中有关阿尔茨海默病方面的资料与其他生物医学信息链接起来。

图 3[15]　SWAN 中知识和软件生态系统的相互作用



3.2.3　OBO

OBO(Open Biomedical Ontologies，开放生物医学本体) 本体包括 50 多个正交的词表，这些词表涵盖了不同领域，包括基因组学(例如 GO， Molecule Role)、化学(Physico-chemical Process，ChEBI)、解剖学(如 C. elegans Gross Anatomy，Mouse Adult Anatomy)和显性(如Human Disease，Mammalian Phenotype)[16]。OBO 本体可再利用以创造代表特定领域的特定本体。

OBO[17]是共同使用跨生物和医学领域的、结构良好的受控词表的保护地址栏。成为 OBO 的本体应该是开放的，只要来源是公认的，它可以自由使用，它在同一个名称下是不编辑和重新分配的，在同一个语法(OBO-format，OBO-format 的扩展格式，或 OWL)下，它是可以被描述或有可能描述的。此外，还应该有唯一空间名称并正交其它 OBO 本体，并对术语进行类型定义。经过巨大的努力，OBO 本体已经建立并开始维护，这已经能满足生物研究者不断增加的使用需求。

4 地理学领域的领域本体研究及应用

GIS(Geographic Informatiom System, 地理信息系统)是地理信息服务不可缺少的一个重要系统，在它诞生的30多年里，其应用已经涉及到多个学科领域中，但是由于不同GIS软件的使用，导致不同空间数据格式的不兼容，使得GIS一度被人们认为是信息孤岛。地理领域的本体的诞生，为这一难题带来了光明。

4.1 BUSTER

Klien E[18]的文章用一个实际的案例研究什么程度的基于本体服务的发现可以解决那些语义异构问题。这种方法是将基于本体的元数据和基于本体的搜索结合起来。基于寻找地理信息服务用于评估森林潜在的风暴危险的假想，表明通过术语学的推论，需要可以寻找到一个与风暴危险等级匹配的可能合适的服务。

BUSTER(Bremen University Semantic Translator for Enhanced Retrieval，用于高级检索的布莱梅大学语义译者)方法[18]是使用基于本体的元数据结合术语推论，以确保信息检索中的语义协调工作能力。这种方法已经在明确定义的领域里得到发展，少数的人或机构已开始专心致力于一个共享的词汇表。文中，作者将这种方法在更广的环境下用于地理信息网络服务发现。

4.2 ONKI-Paikka

ONKI-Paikka是在芬兰地方本体SUO (Suomalainen Paikkaontologia)基础上，开发于FinnONTO(National Semantic Web Ontology Project in Finland，芬兰国家语义网本体项目)项目中[19]。ONKI-Paikka中位置的可视化使用的是Google Maps API。

ONKI-Paikka是用来解决位置寻找和位置名称的解疑问题。因为地理位置名称被广泛使用，但是它们的语义高度不明确。在ONKI-Paikka中，位置可以通过多方面的搜索引擎找到，结合语义自动实现，以及包含约束搜索和可视化结果的地图服务[20]。

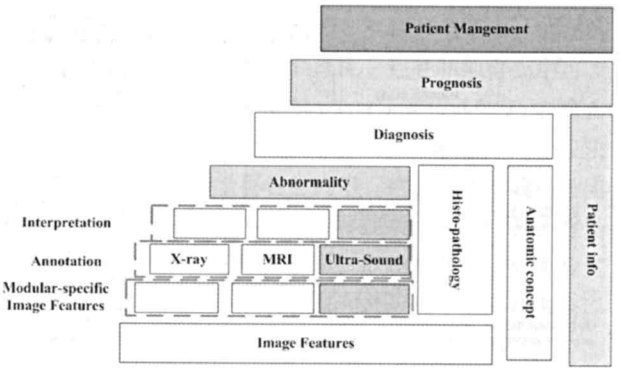
5 其他领域的领域本体研究及应用

5.1 MIAKT

MIAKT[21](Medical Imaging with Advanced Knowledge technologies，先进知识技术下的医学成像)是一个两年的项目，其目的是为乳腺癌筛查程序和乳腺癌三层评估(Triple Assessment，简称TA)提供一个高级知识服务的原型。图4是MIAKT项目的分层结构。

乳腺癌的诊断很大程度上依赖于视频和触觉信息，比如各种医学图像、物理检查结果，这些视频和触觉信息不能精确和全面地用数学或逻辑模型来描述。我们期望有一个实用的领域本体不过分拘泥于规定的程序，它影响在人类可读性的花费上有自动扣除的优点。我们不能通过实验希望临床医师、BCIO(Breast Cancer Imaging Ontology，乳腺癌成像本体)的潜在用户去本能地使用机读数据，如同他们使用纸张和铅笔那样或像在TA会议中处理口述的阅读报告那样。

图4 是MIAKT项目的分层结构[22]



5.2 SwetoDblp

SwetoDblp[22]是一个浅结构的大密度本体，但拥有大量真实的实例数据。SwetoDblp是在之前创建和适用SWETO (Semantic Web Technology Evaluation Ontology，语义网技术评价本体)的基础上建立的。SwetoDblp是通过SAX-剖析过程(SAX-parsing process)创建的，SAX-剖析过程可执行各种特定领域的庞大XML文件(可通过DBLP网站获取)到RDF的转换。本体的架构词表部分是一个本体的子集，被用于LSDIS实验室的出版物图书馆的后端系统。此架构采用的主要概念和关系来自于FOAF和都柏林核心以及必要的他们的扩展形式。此外，我们使用OWL词表以说明与其它6个词表(如AKTors出版本体)的分类和关系的等同[23]。

特别要说地是，SwetoDblp是开源可下载的，连同其附带的数据表都可以用于其他的创建工作①。附带数据表有助于SwetoDblp内的整合以及添加许多关系和实体，由此产生的本体充分地整合了其它数据源。

目前，SwetoDblp正在用于测试iSPARQL查询引擎。SwetoDblp和SWETO本体正被用于测试新一代超过AquaLog的工具，AquaLog利用的是大规模的本体知识库。

5.3 Kumbang

Kumbang[24]是一个代表软件产品家族可变性的领域本体。Kumbang包含来自观点的特点和结构要点以及这两个观点间的相互关系的建模变化性的概念。即使不是不同于全部，至少也是不同于大多数已经存的建模方法，此方法为软件产品家族构造变化性。Kumbang是通过自然语言和UML 2.0文件严格描述的。

Kumbang基于3个抽取层次构建。在最高级别的抽取是元层(metalayer)，包含建模的概念或元类(meta classes)。第2层是模型层(model layer)，包含Kumbang模型。出现在Kumbang模型的实体被称为类(classes)，并获取元类实例。第3层是实例层(instance layer)，包含出现在模型层的类的实例。

6 近两年领域本体应用进展的特点

通过上述介绍，我们可以总结出近两年来领域本体应用进展的几个特点：①涉及学科领域广泛，相当多的科学领域都有了自己的领域本体，其发展也很迅速；②领域本体更加专业化、针对性更强，相当多的某领域的热门问题都有了属于自己的领域本体，比如专为阿尔茨海默病研究而建立的SWAN；③涉及多个学科的领域本体增多，这点对于交叉学科尤为突出，比如生物、医学、化学等，这些学科由于自身研究的需要，其本体也大多涉及多个学科领域。

7 结 语

数字图书馆正日益得到人们的重视，在人们的生活工作中，特别是科研中扮演着很主要的角色。如何将形形色色、各式各样的信息资源有机整合起来，并深度挖掘信息资源自身的属性以及信息资源之间的联系，以便更加有效地实现基于语义的个性服务、统一认证和检索，本体成为建设未来数字图书馆不可缺少的一部分。

随着信息技术的发展和完善，本体也相应地在各个学科领域推广应用，各个学科领域本体的发展有助于揭示该学科领域本身以及与其他学科领域之间的关系，可以帮助科研工作者发现问题、解决问题。而对于数字图书馆工作而言，领域本体的建设发展不仅可以大大推动信息资源建设，还可以大大提高信息检索的效率和准确性。

注释：

① Sweto Dblp 下载地址:http://lsclis.cs.uga.edu/projects/Semdis/swetodblp.

参考文献：

[1] 牟冬梅, 范 轶. 数字图书馆领域本体的构建与推理——以医学领域本体为例[J]. 图书情报工作, 2007 (8):26-30.

[2] Degtyarenko K, De Matos P, Ennis M. ChEBI: A Database and Ontology for Chemical Entities of Biological Interest[J]. Nucleic Acids Research, 2008 (Database issue): D344-D350.

[3] Feldman Howard J, Dumontier M, Ling Susan, et al. CO: A Chemical Ontology for Identification of Functional Groups and Semantic Comparison of Small Molecules[J]. FEBS Lett, 2005 (21):4685-4691.

[4] Webster Y W, Nianhua L, Bukhres O. BAO, A Biological and Chemical Ontology For Information Integration[J]. Online Journal of Bioinformatics1, 2002 (1):60-73.

[5] Miled B Z, Li Nianhua, Bukhres O. BACIIS: Biological and Chemical Information Integration System[J]. Journal of Database Management, 2005 (3): 72-85.

[6] Morbach J, Yang Aidong, Marquardt W G. OntoCAPE——A Large-Scale Ontology for Chemical Process Engineering[J]. Engineering Applications of Artificial Intelligence, 2007 (2):147-161.

[7] Yang Aidong, Braunschweig B, Fraga E S. A Multi-Agent System to Facilitate Component-Based Process Modeling and Design[J]. Computers & Chemical Engineering, In Press, Corrected Proof, Available online 17 November 2007.

[8] Varadwaj P K, Lahiri T. FGO: A Novel Ontology for Identification of Ligand Functional Group[J]. Bioinformation, 2007(3):113-118.

[9] The Gene Ontology. What Does the Gene Ontology Consortium Do [R/OL]. [2008-04-29]. http://www.geneontology.org/.

[10] Chiang Junghsien, Ho Shinghua, Wang Wenhung. Similar Genes Discovery System (SGDS): Application for Predicting Possible Pathways by Using GO Semantic Similarity Measure[J]. Expert Systems with Applications, 2008(3):1115-1121.

[11] Bechhofer S K, Stevens R D, Lord R W. GOHSE: Ontology Driven Linking of Biology Resources[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2006(3):155-163.

[12] Burgun A. Desiderata for Domain Reference Ontologies in Biomedicine[J]. Journal of Biomedical Informatics, 2006(3):307-313.

[13] Rosse C, Jos é L V Mejino Jr. A Reference Ontology for Biomedical Informatics: the Foundational Model of Anatomy[J]. Journal of Biomedical Informatics, 2003(6):478-500.

[14] Noy N F, Rubin D L. Translating the Foundational Model of Anatomy into OWL[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2008 (2):133-136.

[15] Ciccarese P, Wu Elizabeth, Wong Gwen. The SWAN Biomedical Discourse Ontology[J]. Journal of Biomedical Informatics, 2008, In Press.

[16] Marquet G, Mosser J, Burgun A. A Method Exploiting Syntactic Patterns and the UMLS Semantics for Aligning Biomedical Ontologies: The Case of OBO Disease Ontologies[J]. International Journal of Medical Informatics, 2007 (Supplement 3), S353-S361.

[17] Bada M, Hunter L. Enrichment of OBO Ontologies[J]. Journal of Biomedical Informatics, 2007 (3):300-315.

[18] Klien E, Lutz M, Kuhn W. Ontology-Based Discovery of Geographic Information Services — An Application in Disaster Management[J]. Computers, Environment and Urban Systems, 2006 (4):102-123.

[19] Robin L,Tomi K, Riikka H,et al. ONKI-Paikka: An Ontology Service for Geographical Data. [R/OL]. [2008-05-09]. http://www.seco.tkk.fi/publications/2007/lindroos-kauppinen-et-al-onki-paikka-2007.pdf .

[20] Eero H,Robin L,Tomi K, et al. An Ontology Service for Geographical Content. [R/OL]. [2008-04-23]. http://www.seco.tkk.fi/publications/2007/hyvonen-et-al-ontology-2007.pdf .

[21] Hu Bo, Dasmahapatra S, Dupplaw D. Reflections on a Medical Ontology[J]. International Journal of Human-Computer Studies, 2007 (7):569-582.

[22] Aleman-Mezaa B, Hakimpoura F, Arpinara I B, et al. SwetoDblp Ontology of Computer Science Publications[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2007 (3): 151-155.

[23] Shadbolt NR, Gibbins N, Glaser H,et al. Walking through CS AKTive Space: a Demonstration of an integrated Semantic Web Application [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2004(4):415-417.

[24] Asikainen T, Mannisto T and Soininen T. Kumbang: A Domain Ontology for Modelling Variability in Software Product Families[J]. Advanced Engineering Informatics, 2007(1):23-40.

[作者简介]

余 倩 女，1982 年生，中国科学院研究生院在读硕士。

[收稿日期：2008-05-27]