肿瘤本体构建研究*

李晓瑛，李丹亚，夏光辉，李军莲，胡铁军
（中国医学科学院医学信息研究所，北京 100020）

摘要：在借鉴已有疾病本体描述框架的基础上，复用权威医学知识组织系统中肿瘤相关概念及内容结构，从肿瘤（名称）、病因、诊断、治疗四个维度构建肿瘤本体；此外，探讨一种基于生物医学文献主题标引词的语义关系发现方法，用于丰富肿瘤本体类间关系及扩充肿瘤本体知识库；最后，利用Protégé本体构建工具，尝试构建呼吸系统肿瘤本体，以期为构建大规模肿瘤本体及其它领域本体提供一些有价值的参考与实践。

关键词：领域本体；肿瘤本体；知识组织系统；语义关系发现
中图分类号：G254
DOI：10.3772/j.issn.1673-2286.2015.08.007

1 引言

在信息科学中，普遍认为本体是共享概念模型的明确的形式化规范说明[1]。本体连同叙词表、分类表、词典等其它类型的语义工具，统称为知识组织系统；不同之处在于，本体兴起于网络信息化时代，其核心作用不只限于定义某一特定学科领域的权威概念名称（术语）及其之间的相互关系，并且需要无障碍地在人、计算机等不同主体之间进行对话、互操作、共享等语义交流，是语义网发展的基础与核心。通常，依据本体中概念的主题领域这一维度，可将本体分为四种类型：领域本体、通用本体、应用本体及表示本体[2]。其中，领域本体是一种描述特定领域知识的专用本体，旨在对某一领域的重要概念、属性以及概念间关系给出一种形式化说明；在网络信息资源管理中，领域本体起到语义导航、语义检索、语义标注及术语服务等多种重要作用[3]。例如，在生物医学领域，疾病本体可看作一种利用本体描述语言所建立的疾病知识库，其在很大程度上提高了计算机解读和理解与疾病知识相关的词汇和语义的能力。

2 国内外疾病本体研究现状

目前，随着本体理论研究的逐步完善，国内外许多研究机构尝试开展各种类型的疾病本体构建研究。其中，最著名的疾病领域本体是由美国西北大学基因药物中心与马里兰大学医学院基因组科学研究所联合开发的人类疾病本体（Disease Ontology，DO）[4]；DO通过对人类疾病的名称、表型特征进行本体化描述，旨在提供一个具有高度一致性、可重用性及可持续发展的医学疾病类词典；此外，DO已完成与《医学主题词表》（Medical Subject Headings，MeSH）[5]、《国际疾病分类法》（International Classification of Diseases，ICD）、《NCI叙词表》（NCI thesaurus，NCIt）、《国际系统医学术语集》（Systematized Nomenclature of Medicine，SNOMED）、《在线人类孟德尔遗传数据库》（Online Mendelian Inheritance in Man，OMIM）之间的语义精确互映射，从而促进了各种疾病及相关健康知识向特定医学代码的映射。澳大利亚科廷科技大学研发的人类疾病类本体（Generic Human Disease Ontology）[6]，从疾病类型、表型、病因、治疗四个维

* 本研究得到十二五国家科技支撑计划课题“面向外文科技文献的超级科技词表和本体构建研究”（编号：2011BAH10B01）资助。

度进行建模，旨在向医师及医学研究人员提供可供计算机直接操作的人类疾病信息，以支持其开展各种医学分析及应用研究。在国内，军事医学科学院解放军医学图书馆的郭会雨等人[7]依据斯坦福大学医学院的本体构建七步法，通过复用《医学知识库》、一体化医学语言系统（Unified Medical Language System，UMLS)[8]、《医学主题词表》及《ICD-10 中国疾病诊断标准数据库》（ICD-10 Diagnosis Standard Database of China，DSDC)等已有的领域资源，在Protégé本体构建及编辑平台上构建出包含疾病领域重要术语（本体类）、术语间等级关系、等同关系、部分整体关系（本体类间关系）、优选名称、别名、定义、代码等属性（本体属性）的疾病领域本体；然而该领域本体构建过程较多依赖领域专家手工参与完成，随着当前大数据时代的到来，这种手工化本体构建工作急需向自动化处理转变。

作为疾病知识库，上述大规模的疾病领域本体比较全面地覆盖了重要的疾病概念，但就某一具体的疾病而言，其所揭示的领域知识可能并不深入；再者，对于疾病的描述，是一门综合了临床医学、解剖学、药学等多个领域的交叉学科，疾病知识中所涉及概念的分类与界定、概念间关系的梳理与获取存在很大的困难；另外，随着疾病知识扩充、更新、删除等知识演化，疾病本体的版本更新及维护亦将花费不菲的工作量。因此，越来越多地研究者开始关注体量较小但内容精细的疾病专题本体构建工作，例如传染病本体。其中，国际上具有较高影响力的传染病本体有日本国家信息研究所开发的公共卫生领域疫情监测本体（BioCaster）[9]、美国开放生物医学本体（Open Biomedical Ontologies，OBO）项目中的传染病本体（Infectious Disease Ontology，IDO）[10]；BioCaster通过规范多种语种的传染病术语，构建传染病知识库，用于对各地区关于疾病传播的网络信息及其它网络资源进行持续跟踪与数据挖掘。而IDO本体通过构建一套相互之间可互操作的传染病知识库，促进了医学领域知识的整合。国内中国医学科学院医学信息研究所方安等人[11]针对当前传染病本体构建中存在的一致性差和共享困难等问题，在借鉴UMLS语义网络的基础上构建传染病本体，进而搭建知识服务平台，提供与传染病相关的知识浏览和知识检索等知识服务。

与传染病相比，肿瘤也是人类疾病中一个不容忽视的组成部分，国外最为著名的肿瘤本体即美国开放生物医学本体中以网络本体语言（Web Ontology Language，OWL）描述的《NCI叙词表》[10]。然而，就目前文献调研情况来看，国内尚未深入开展肿瘤本体构建研究。作为国家“十二五”科技支撑计划课题“面向外文科技文献的超级科技词表和本体构建研究”的主要内容和任务之一，本研究借鉴已有疾病本体的构建模式[6]，在复用《国际系统医学术语集–临床术语》（Systematized Nomenclature of Medicine – Clinical Terms，SNOMED CT）[12]、UMLS等权威医学知识组织系统资源的基础上，以呼吸系统肿瘤为例，对构建肿瘤领域本体的完整过程进行研究与实践。此外，对于已有知识组织系统中所缺少的语义关系，本研究尝试从生物医学文献的主题标引词中进行发现，以期完善肿瘤本体知识库，全面呈现肿瘤知识。

3 肿瘤本体模型构建

在充分考虑临床上常用的肿瘤信息及借鉴已有疾病本体描述框架[6]的基础上，从肿瘤（名称）、病因、诊断、治疗四个维度构建肿瘤本体模型（见下图1）；其中，肿瘤及肿瘤子类指肿瘤概念及下位概念名称。病因与肿瘤之间存在引发的关系，包含遗传因素及环境因素等；遗传因素包括基因、DNA序列，环境因素又分为化学物质、微生物（如细菌、病毒）等。诊断指从医学角度对人体健康状况所做出的判断，与肿瘤之间存在临床发现关系，一般涉及两方面：发现部位和体征；发现部位即人体解剖结构，体征主要指人体主观上的异常感觉或客观上身体解剖部位的病态改变。治疗通常指临床上干预或改变人体健康状态的手段或过程，包括药物治疗、手术、放射疗法等。

4 肿瘤本体构建过程

类、关系、属性是本体中最重要的三个元素，亦是

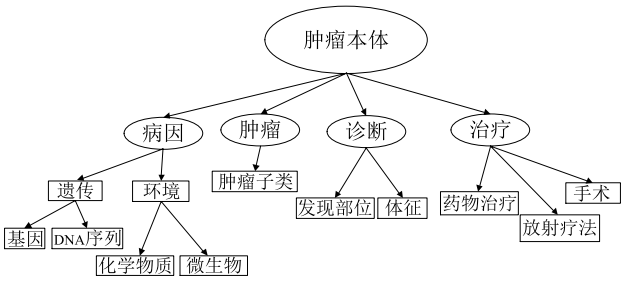

图1 肿瘤本体模型

本研究中构建肿瘤本体的三个关键步骤（见下图2），期间涉及从已有知识组织系统中复用与肿瘤相关的概念及概念关系，从PubMed/Medline生物医学文献数据库[13]中发现语义关系等多个环节。

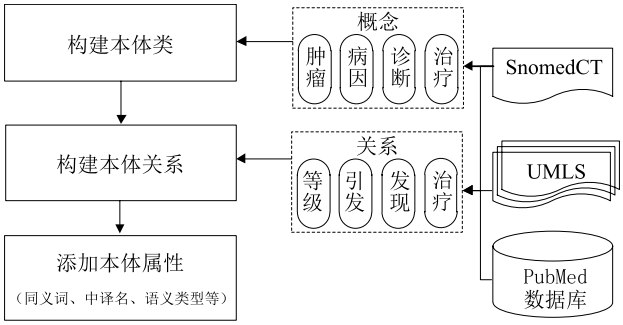

图2 肿瘤本体构建过程

4.1 构建肿瘤本体类

与叙词表、分类表等类型的知识组织系统相比，本体的一个显著特征是对领域知识的共享与复用。在肿瘤本体的构建过程中，肿瘤相关概念并非利用关键词抽取技术从生物医学资源中收集和筛选，而是直接复用已有的医学知识组织系统SNOMED CT。至今，SNOMED CT是国际上公认的最全面、最准确的临床术语集，是一部经过系统组织编排的、易于计算机处理的医学术语集，涵盖了绝大部分的临床信息。构建SNOMED CT的初衷，是使不同的临床医生、研究人员、医疗卫生机构及其它使用者在指陈同一临床事物时所采用的临床术语能进行交互，从而实现临床信息交换。就历史沿革而言，SNOMED CT发展于1974年美国病理学会编著出版的SNOMED，并逐渐形成SNOMED RT（参考术语集），之后于2002年和英国国家卫生服务部的《临床术语》（Clinical Terms，又称Read Codes）进行融合形成；自2007年开始，SNOMED CT版权归国际卫生术语标准制定组织。

在数据结构上，概念表、描述表、关系表是SNOMED CT体系结构中的三个重要部件[14]；其中，概念表收录了包括身体结构、疾病、临床发现、操作、有机体、药品等19个顶级概念轴中约30万条具有唯一含义、并且经过逻辑定义的概念；描述表收录了约80万条能够代表某个具体概念的术语（概念优选名称）及其概念同义词；关系表包括约136万条以三元组格式存储的语义关系；这三种数据文件，为肿瘤本体构建中本体类的获取提供了有力支持。具体而言，首先通过关系表中的直接上下位关系及概念表，获取SNOMED CT中肿瘤及其下位肿瘤概念数据，用于构建肿瘤类；其次，通过关系表中已有的引发、发现、治疗等语义关系，同时结合直接上下位关系及概念表，获取相应的病因、诊断、药物、手术等概念（及其下位概念），这些概念共同构成了肿瘤本体类；其中，肿瘤是本研究的核心类。

4.2 构建肿瘤本体关系

语义关系是对领域概念知识的组织，也是知识组织系统构建中的重要内容。本研究中，肿瘤本体类间关系首先从SNOMED CT关系表中已有的语义关系中获取，包括等级关系、引发、发现、治疗等；图3是从SNOMED CT关系表中获得的与肺肿瘤（lung neoplasm）有关的语义关系，包括等级关系（is_a）、发现部位（finding_site_of）等。之后，复用UMLS超级叙词表（Metathesaurus）中与肿瘤有关的语义关系数据。UMLS超级叙词表是生物医学领域概念、术语、涵义及语义关系的广泛集成，整合自SNOMED CT、MeSH、NCIt等160多部知识组织系统，概念数达300万；超级叙词表除对多部异构来源表中表达同一内涵的多个术语以相同概念唯一标识符（Concept Unique Identifier，CUI，见MRCONSO概念表）进行整合外，亦保留和继承了来源表中的其它语义关系（见MRREL关系表）。因此，对肿瘤本体构建工作而言，可借助超级叙词表CUI，查找到UMLS其它来源词表中肿瘤相关概念的术语表达形式以及相应的语义关系，进而复用到肿瘤本体关系中。

尽管SNOMED CT与UMLS中已有一些语义关系，但经过分析后发现，疾病与药物之间重要的治疗关系

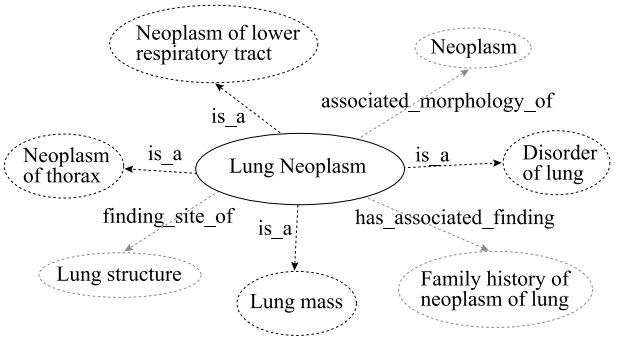

图3 SNOMED CT中关于肺肿瘤的语义关系

相对较少。因而，本研究将通过语义发现技术，从文献科学数据中获取更多的语义关系，用于丰富肿瘤本体关系。PubMed/Medline是由美国国立医学图书馆研发的大型开放型生物医学文献数据库，公众可自由获取全文文献及其基于MeSH词表的主题标引词。MeSH词表中约有包含疾病、病因、诊断、治疗等在内的2万个主题词，以及药物疗法、投药&剂量、治疗应用等90个副主题词，PubMed/Medline主题标引结果便是通过MeSH主题词与副主题词组配实现；例如一篇论证吉非替尼（Gefitinib）治疗肺癌的文章，在利用MeSH词表进行主题标引后，标引词即为“吉非替尼/投药&剂量（或治疗应用）”以及“肺癌/药物疗法”。而基于这种MeSH主题词与副主题词组配的主题标引词便可发现肿瘤与药物之间的治疗关系。另外，在对生物医学文献进行主题标引时，标引员通常采用为所标引的文献主题词打星号的方式区分主要标引词，即加权标引；亦即，带有星号的标引词为文献重点讨论内容，最能表达文献主题；进而基于带星号的标引词所推导出的语义关系不仅关键而且准确，因为文献最核心概念一般很少标错。进一步结合发现某一对具体关系的文献数，即该关系的出现频次（或称共献率），可对肿瘤治疗关系的发现结果进行筛选及过滤，提高肿瘤本体关系的可靠性与准确性。图4是从PubMed主题标引文献中，发现的与肺肿瘤相关的语义关系，包括引发（cause_of）、诊断（diagnose_of）、药物治疗（drug_therapy_of）、手术（surgery_of）、放射疗法（radiotherapy_of）等多种类型。

4.3 添加肿瘤本体属性

对知识组织系统而言，属性是对概念深层次的描述。在肿瘤本体中，本体属性对理解肿瘤概念内涵、揭示肿瘤领域知识起到非常重要的作用。经过对疾病领域知识的分析，本研究着重构建的肿瘤本体属性包括：（1）概念同义词：通常也称为入口词（概念优选名称已作为本体类名）；（2）中文名称：即肿瘤本体类名的中译名，中英文双语对照的本体类名有助于理解肿瘤领域知识及掌握医学术语；（3）语义类型：即肿瘤（名称）、病因、诊断、治疗等相关概念所对应的范畴类目，用于对肿瘤本体中所有概念进行统一分类，例如肺肿瘤的语义类型为肿瘤发生（Neoplastic Process）。

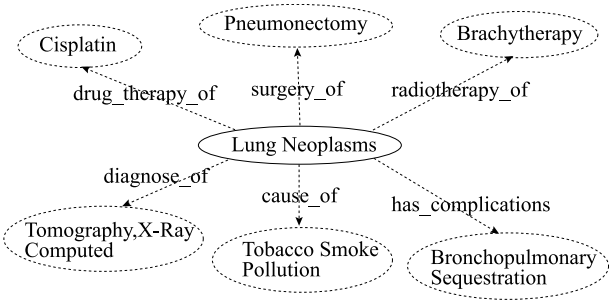

图4 从PubMed主题标引文献中发现肺肿瘤关系

5 肿瘤本体构建平台及成果

5.1 肿瘤本体构建平台

经过10多年的发展，本体构建编辑工具已逐步成熟，目前存在多种具有较高影响力的本体构建工具，如美国南加利福尼亚大学于1990年发布的Ontosaurus、英国开放大学于1997年开发的WebOnto等；其中，使用最广泛、最受关注的工具是美国斯坦福大学生物医学信息研究中心开发的Protégé[15]。相比而言，Protégé具有以下优势：

（1）开放资源，支持在线及本地两种使用模式，且用户可免费、轻松获取其本地版工具；

（2）支持平台手工编辑本体及基于Java编程语言的本体自动生成，界面及Java源码风格简单友好，易学易用；

（3）支持以OWL、RDF、XML等多种方式存储本体文件，且可在不同本体格式之间相互转化；

（4）集成了OWLViz、OntoGraf等多种可视化插件，便于直观浏览本体元素，且支持将可视化结果保存为图形格式；

（5）持续更新，功能日益完善并增多，受到全球24.5万用户的信赖。

然而，Protégé亦存在一些不足，主要体现在：（1）同时只能打开一个本体，不支持多个本体之间的匹配、合并、复用；（2）基于Java编程语言的本体自动生成，所能处理的数据量相当有限，极大地影响了大型本体的开发工作。因此，本研究选择以肿瘤的一个分支，以呼吸系统肿瘤本体为例，生成轻量级专题领域本体，从而对构建肿瘤本体的整个过程进行实践探索，以期为今后大规模肿瘤本体及其它领域本体构建工作提供一些有价值的借鉴。

5.2 肿瘤本体构建成果

鉴于Protégé本体构建平台无法支持大数据量的肿瘤本体构建工作，本研究最终选用呼吸系统肿瘤这一分支进行尝试，以期为大规模肿瘤领域本体的构建做好前期实践探索。经统计，呼吸系统肿瘤本体共包含肿瘤（Tumor）、身体系统结构（Body System Structure）、身体组织结构（Body Tissue Structure）、基因（Gene）、细胞结构（Cell Structure）、物质（Substance）等9个本体大类，以及良性呼吸系统肿瘤（Benign neoplasm of respiratory system）、恶性呼吸系统肿瘤（Malignant neoplasm of respiratory system）等3900个下位类（见下图5本体类等级导航）。

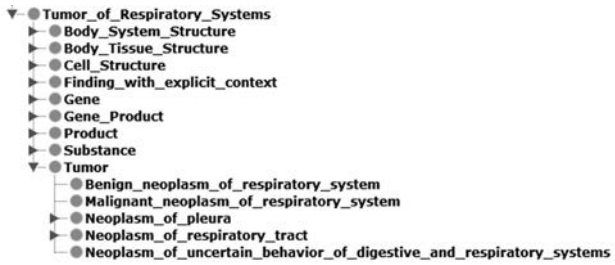

图5 呼吸系统肿瘤本体类等级导航

此外，呼吸系统肿瘤本体包含等级（is a）、发现部位（has finding site）、治疗（be treated by）等9种本体关系类型（见下图10），共3000多条本体关系。借助本体可视化插件，可直观浏览这些关系。如下图6所示，呼吸道肿瘤（Neoplasm of respiratory tract）的直接上位（is a）是肿瘤（Tumor），直接下位类共有8个，包括良性呼吸道肿瘤（Benign neoplasm of respiratory tract）、上呼吸道肿瘤（Neoplasm of upperrespiratory tract）、下呼吸道肿瘤（Neoplasm of lowerrespiratory tract）等，发病于（has finding site）呼吸道结构（Respiratory tract structure），与肿瘤细胞（Neoplastic Cell）之间存在异常细胞（disease has abnormal cell）关系，通常用药物喹唑啉（Quinazoline）进行治疗（be treated by）。

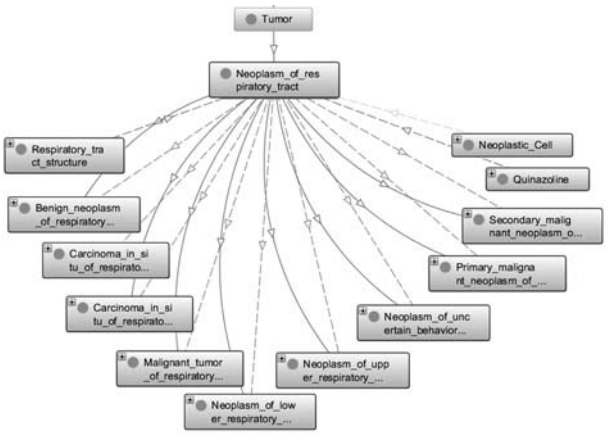

图6 呼吸系统肿瘤本体可视化示例

另外，可在Protégé本体平台上，浏览呼吸系统肿瘤本体的概念同义词、中文名称、语义类型等本体属性信息。如下图7所示， Neoplasm of lower respiratory tract的概念同义词有Tumor of lower respiratory tract、Tumour of lower respiratory tract、Neoplasm of lower respiratory tract (disorder)，中文名称为下呼吸道肿瘤，语义类型为肿瘤发生（Neoplastic Process）。这些重要属性可看作对呼吸系统肿瘤本体类的辅助说明，有助于用户更好地理解呼吸系统肿瘤概念。

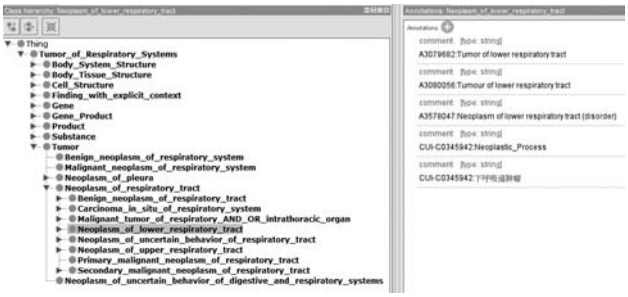

图7 呼吸系统肿瘤本体属性示例

6 结语

领域本体是领域概念及概念间详细关系的一种形式化描述，是语义网发展的重要基础。鉴于从领域资源中重新获取领域概念是一件十分困难且效果未必理想的工作，复用权威知识组织系统中的概念及其语义关系是当前领域本体构建中的重要内容。本研究即是在规划肿瘤领域本体基本模型的基础上，探讨了通过复用SNOMED CT、UMLS等具有较高影响力的医学知识组织系统中肿瘤相关概念及内容结构，构建肿瘤专题领域本体。此外，对于已有知识组织系统中存在较少的关于肿瘤治疗方面的语义关系，本研究提出了一种基于PubMed/Medline生物医学文献主题标引词的语义关系发现方法。最后，以呼吸系统肿瘤本体为例，在Protégé本体构建平台上，完成了呼吸系统肿瘤本体构建工作，这为大规模肿瘤本体及其它领域本体构建工

作提供了一定的实践经验。然而，就目前肿瘤本体构建情况而言，在今后的工作中，还需进行更深入的研究和拓展，例如全面构建大规模肿瘤本体，探讨客观有效的本体性能和质量评价指标及方法等。

参考文献

[1] Studer R, Benjamins V R, Fensel D. Knowledge engineering: principles and methods [J]. Data & Knowledge Engineering, 1998, 25(1):161-197.

[2] Van Heijst, Th.Schreiber, Wielinga J. Using explicit ontologies in KBS development [J]. International Journal of Human Computer Studies, 1997, 46(2-3):183-292.

[3] 何琳. 领域本体的半自动构建及检索研究[M]. 南京：东南大学出版社，2009:32-36.

[4] Disease Ontology[EB/OL]. [2015-06-22]. http://www.disease-ontology.org/.

[5] Medical Subject Headings [EB/OL]. [2014-09-08]. http://www.nlm.nih.gov/mesh/MBrowser.html.

[6] Hadzic Maja, Chang Elizabeth. Ontology-based Support for Human Disease Study[C]. Proceedings of the 38th Hawaii International Conference on System Sciences, 2005.

[7] 郭会雨, 张文举, 李娜. 疾病领域本体模型构建研究[J]. 预防医学情报杂志, 2011,27(6):460-465.

[8] UMLS Home [EB/OL]. [2015-05-11]. http://www.nlm.nih.gov/research/umls/.

[9] Collier N, Kawazoe A, Jin L, et.al. A multilingual ontology for infectious disease surveillance: rationale, design and challenges [J]. Language resources and evaluation, 2007, 40(3-4):405-413.

[10] The Open Biological and Biomedical Ontologies [EB/OL]. [2015-07-02].http://www.obofoundry.org/.

[11] 方安,洪娜,高东平,等. 传染病本体构建及其在知识服务平台中的应用[J]. 现代图书情报技术, 2012(1):7-12.

[12] SNOMED CT[EB/OL]. [2015-07-01].http://www.ihtsdo.org/snomed-ct.

[13] PubMed[EB/OL]. [2015-07-01].http://www.ncbi.nlm.nih.gov/pubmed/.

[14] 李丹亚,李军莲,李晓瑛,等. 医学知识组织体系发展现状及研究重点[J]. 数字图书馆论坛, 2012(12):13-21.

[15] Protégé[EB/OL]. [2015-07-01].http://protege.stanford.edu/.

作者简介

李晓瑛，女，博士，中国医学科学院医学信息研究所助理研究员，研究方向：知识组织，E-mail：lixiaoying@imicams.ac.cn。
李丹亚，女，中国医学科学院医学信息研究所研究员，研究方向：知识组织、资源建设。
夏光辉，男，硕士，中国医学科学院医学信息研究所助理研究员，研究方向：知识组织。
李军莲，女，硕士，中国医学科学院医学信息研究所副研究馆员，研究方向：知识组织、资源建设。
胡铁军，男，中国医学科学院医学信息研究所研究员，研究方向：资源建设。

Research on the Construction of Tumor Ontology

LI XiaoYing, LI DanYa, XIA GuangHui, LI JunLian, HU TieJun
(Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China)

Abstract: Based on the known disease ontologies and tumor related concepts in several famous medical knowledge organization systems, this paper intends to construct the tumor ontology from 4 dimensions: tumors (names), pathogenic factors, diagnoses and treatments. In order to increase the relations of tumor ontology as well as the tumor knowledge, this paper also presents an algorithm to find the semantic relations from indexed biomedical papers. Finally, this paper uses protégé to build the tumor of respiratory system ontology, and aims to provide some helpful information for the construction of generic tumor ontology and other domain ontologies.

Keywords: Domain Ontology; Tumor Ontology; Knowledge Organization System; Semantic Relation Finding

（收稿日期：2015-07-10；编辑：雷雪）