

# 医学知识组织体系发展现状及研究重点\*

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

**摘要:** 文章基于对200多部活动的医学术语表、分类表、叙词表、本体及一体化语言系统的研究和分析,归纳总结了生物医学领域知识组织体系的发展现状。重点对主流医学知识组织体系MeSH、SNOMED CT、UMLS的构建模式进行剖析,对有代表性的标准化医学术语集及医学领域本体项目进行梳理。最后就医学知识组织体系的研究重点及推进我国医学领域知识组织体系建设进行简要论述。

**关键词:** 医学知识组织体系, 构建模式, 医学术语标准化, 医学领域本体, UMLS, SNOMED CT, MeSH

DOI: 10.3772/j.issn.1673—2286.2012.12.003

## 1 引言

知识组织体系(Knowledge Organization Systems, KOS)是对人类知识结构进行表达和有组织阐述的语义工具,包括传统的建立在文献单元基础上的分类法、叙词表以及更泛指的情报检索语言,也包括网络环境下建立在概念单元基础上的概念地图、语义网及本体等<sup>[1]</sup>。

医学知识组织体系(Medical Knowledge Organization Systems, MedKOS)广泛应用于生物医学研究及实践活动过程中,用来组织和表述各类生物医学数据资源,这些MedKOS是我们既熟悉又处于变革中的知识组织体系。为跟踪国际上快速发展的MedKOS,推进我国MedKOS的建设和应用,本文对200多部活动的医学术语表、分类表、叙词表、本体及一体化语言系统进行了系统分析和研究,并就相关问题进行梳理、归纳和论述。

## 2 MedKOS基本情况分析

(1) MedKOS类型: 根据知识组织体系的结构、语义强弱程度、所实现的功能等要素,目前大致将知识组织体系分为三类: ①词汇总表类: 强调对概念的解释,形式简单,不涉及复杂的语义关系,如权威规范文档、词汇总表、术语表、词典、指南等; ②分类体系: 强调概念间的层级聚合和类别体系,起到范畴归类作用,如分类法、知识分类体系、类目表等; ③关联组类: 强调概念间各种关系的揭示,如叙词表、语义网络、本体等<sup>[2]</sup>。

医学领域最具影响力的一体化语言系统整合了上述三种类型的知识组织体系,充分结合了各种知识组织体系的优点。目前主要一体化语言系统的实践活动包括: 美国国立医学图书馆(NLM)建立的UMLS(Unified Medical Language System, 一体化医学语言系统)<sup>[3]</sup>; 美国国立癌症研究所(NCI)建立的NCIm(NCI Metathesaurus, NCI超级叙词表)<sup>[4]</sup>; 中国医学

\* 本文系国家“十二五”科技支撑计划课题“面向外科技文献的超级科技词表和体系建设”(编号: 2011BAH10B01)子任务的研究成果之一。

科学院医学信息研究所建立的中文一体化医学语言系统<sup>[5]</sup>；以及中国中医研究院中医药信息研究所发展建立的中医药一体化语言系统<sup>[6]</sup>。

(2) MedKOS统计分析：共对203部MedKOS进行统计，其中英文198部，中文5部，总体情况见表1，表2列举了有代表性的MedKOS的基本特征。

表1 医学知识组织体系基本情况

| 描述项  | 描述内容                                                |
|------|-----------------------------------------------------|
| 类型分布 | 一体化语言系统4部，本体22部，叙词表36部，语义网络1个，分类表35部，规范术语表26部，其他79部 |
| 主题领域 | 医学综合20部，基础医学22部，临床医学96部，预防医学22部，药学28部，其他15部         |
| 词表体量 | 超过10万条的23部，1万—10万条的33部                              |
| 释义   | 带释义的85部                                             |
| 多语种  | 多语言版本的50部                                           |
| 更新频率 | 定期更新的83部                                            |

### 3 MedKOS构建模式研究

知识组织体系的构建模式决定了知识组织体系的使用效率和适应性，张琪玉教授认为“情报检索语言的进步主要是结构模式的进步”<sup>[21]</sup>。当前网络环境下，传统MedKOS的结构模式正在寻求改变，新型MedKOS不断涌现，医学领域很多知识组织体系有其独特的构建模式。

#### 3.1 MeSH的构建模式

MeSH (Medical Subject Headings, 医学主题词表) 是生物医学领域最著名的大型综合性叙词表，随着信息环境的变化，MeSH的构建模式也在不断演进。

(1) MeSH的三级概念结构：传统叙词表由“叙词—入口词”二级结构组成，新型MeSH词表由“主题词(叙词)—概念—术语(入口词)”三级结构组成，或称主题词类。一个主题词类可以包含多个概念、一个概念包含多个术语。每个概念内的术语表示精确的同义关系，主题词类中多个概念之间的关系可能为同义、广义、狭义和相关关系。每个主题词类中的一个概念被推选为优选概念，并被指定为主题词，每个概

念中的一个术语被推选为优选术语，如图1所示<sup>[7,22]</sup>。

MeSH的三级概念结构避免了传统叙词表二级结构带来的关系丢失或歧义现象，如主题词与入口词之间的关系可能是同义、近义或反义关系，这些关系无法在二级结构中准确表达出来，只能粗略地表达为用代关系，而MeSH的三级结构可以将这些语义关系明确清晰地标识出来，有利于计算机的理解和处理，同时也保留了传统叙词表二级结构的功能。

(2) MeSH的表内和表间术语映射机制：MeSH由主题词表、副主题词表、增补概念表和树形结构表四个部分组成。MeSH增补概念表的术语多来自文献，是对主题词表的补充和扩展，MeSH增补概念词不是主题词，但它至少与一个MeSH主题词建立映射关系。另外，MeSH中的每个概念都标识有UMLS语义网的语义类型，通过语义类型建立起与外部语义网高层类目间的映射关联，共享UMLS语义网的丰富语义关系<sup>[7,23]</sup>。

#### 3.2 SNOMED CT的构建模式

SNOMED CT (Systemized Nomenclature of Medicine—Clinical Terms, 国际系统医学术语集——临床术语) 是医学临床广泛应用、结构独特的大型医学知识本体。本体是一种对共享概念体系的明确的形式化规范描述<sup>[24]</sup>，从本体的构建思想来看，SNOMED CT可称之为医学领域一例成功利用知识本体的研究成果。

(1) SNOMED CT的概念分类体系：基于医学特定的概念分类理论，SNOMED CT构建了顶级概念分类框架，将32万多个具有唯一含义并经过逻辑定义的概念，分别编入19个顶级概念类(轴)，通过上下位语义关系“is\_a”，形成以某一特征属性为核心的概念等级结构。每个概念都有一个完全指定名称(FSN)及相对应的一个首选术语和若干个同义术语，FSN由概念名称和括号中的限定词组成，如“Pneumonia (disorder), 肺炎(疾病)”，括号中的内容为概念的语义标签，表明其在语义上所属的类别<sup>[25,26]</sup>。

(2) SNOMED CT的连接概念(linkage concept)：连接概念用于概念间语义关系的揭示和表达，如“概念1+连接概念+概念2”三元组表示形式，图2中的“Finding site(发现部位)”、“Causative agent(治病因素)”为连接概念。连接概念是

表2 典型MedKOS特征一览表<sup>[7-20]</sup>

| MedKOS名称                        | 类型          | 主题领域 | 释义   | 体量<br>(术语/概念)*      | 语种                 | 更新         | 版权机构                                        |
|---------------------------------|-------------|------|------|---------------------|--------------------|------------|---------------------------------------------|
| 医学主题词表 (MeSH)                   | 叙词表         | 医学综合 | 有    | 75.8万/32.1万         | 21种语言              | 1年/1周      | 美国国立医学图书馆                                   |
| NCI叙词表 (NCIt)                   |             | 肿瘤学  | 有    | 23.8万/9万            | 英文                 | 1月         | 美国国立癌症研究所                                   |
| 医学主题词表 (中文版)                    |             | 医学综合 | 有    | 11万/5.5万            | 中英文                | 1年         | 中国医学科学院医学信息研究所                              |
| 中国中医药学主题词表                      |             | 中医药  | 有    | 0.83万/0.56万         | 中英文                | 不定期 (2007) | 中国中医研究院中医药信息研究所                             |
| UMLS语义网和导航图                     | 语义网络        | 医学综合 | 有    | 语义类型133个<br>语义关系54个 | 英文                 | 不定期        | 美国国立医学图书馆                                   |
| 国际系统医学术语集——<br>临床术语 (SNOMED CT) | 本体          | 临床医学 | 逻辑定义 | 118万/32.4万          | 英文、西班牙、<br>丹麦、瑞典语等 | 半年         | 国际健康术语标准发展组织-<br>SNOMED标准研发组织; 美国病理<br>医师学会 |
| 解剖学基础模型本体<br>(FMA)              |             | 解剖学  | 有    | 14.9万/8.2万          | 英法德、西<br>班牙语等      | 每周修改       | 美国华盛顿大学结构信息学研究组                             |
| 中国图书馆分类法-<br>医学专业分类表            | 图书<br>分类法   | 医学综合 | 类目注释 | 0.5万个类名             | 中文                 | 不定期 (1999) | 《中国图书馆分类法》编委会/医科<br>院信息所图书馆                 |
| 国际疾病分类法-第10<br>修订版 (ICD10)      | 知识<br>分类表   | 临床医学 | 类目注释 | 1.15万个类名            | ICDs包括<br>43种语言    | 1年 (10年修订) | 世界卫生组织                                      |
| NCBI分类表<br>(NCBI Taxonomy)      |             | 生物学  | 类目注释 | 63.4万个类名            | 英文                 | 1天         | 美国生物技术信息中心                                  |
| 处方药物标准术语表<br>(RxNorm)           | 规范<br>术语表   | 药学   | 逻辑表示 | 49.7万/20.4万         | 英文                 | 1天         | 美国国立医学图书馆                                   |
| 观测指标标识符逻辑命名<br>与编码系统 (LOINC)    |             | 检验学  | 逻辑表示 | 36.4万/14万           | 英法德、<br>中文等        | 半年         | 美国印第安纳大学医学中心<br>Regenstrief研究院              |
| 人类基因命名表 (HUGO)                  |             | 遗传学  | 无    | 14万/3.3万            | 英文                 | 1天         | HUGO基因命名协会                                  |
| 一体化医学语言系统<br>(UMLS)             | 一体化语<br>言系统 | 医学综合 | 有    | 1080万/266万          | 包括21种<br>语言        | 半年         | 美国国立医学图书馆                                   |
| NCI超级叙词表 (NCIm)                 |             | 医学综合 | 有    | 360万/140万           | 英文                 | 1月         | 美国国立癌症研究所                                   |
| 中文一体化医学语言系统                     |             | 医学综合 | 有    | 60万/30万/3万<br>(叙词)  | 中英文                | 不定期        | 中国医学科学院医学信息研究所                              |
| 中医药一体化语言系统                      |             | 中医药  | 有    | 60万/15万             | 中英文                | 不定期        | 中国中医研究院中医药信息研究所                             |

\*统计数据截止到2012-5-10

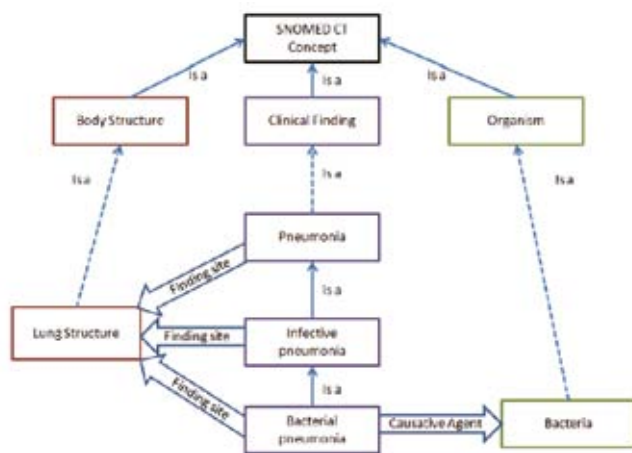
SNOMED CT 19个顶级概念轴中的一类, 包括约60个概念, 每个连接概念都进行域、范围、角色层级结构、角色组的明确限定, 以保证连接概念的正确使用<sup>[25, 27]</sup>。

(3) SNOMED CT的语义关系和概念逻辑化定义: SNOMED CT的语义关系包括定义性、限定性、历史性和附加性, 其中最重要的定义性语义关系包括两类: “is\_a”关系和“Attribute”关系, SNOMED CT通过纵向的上下位关系和横向的属性关系确定两个概

念间的具体语义关系, SNOMED CT大约包括136万个语义三元组。SNOMED CT的属性关系是通过上述连接概念表达的, 通过连接概念, 将归属于同一概念轴和不同概念轴的概念连接起来, 形成具有内在语义关联、能清晰表达临床信息、便于计算机理解的短句。图2显示了细菌性肺炎 (Bacterial pneumonia) 概念的逻辑化定义图, 表明该病是一种由细菌 (Bacteria) 导致的发生在肺部 (Lung structure) 的疾病; 此外还展示了各自所属的3个顶级概念轴路径<sup>[25, 27]</sup>。

|                          |                                                                                                                |                                                                                                                                                                                                                                                                                                                                          |                 |                  |                                   |
|--------------------------|----------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|------------------|-----------------------------------|
| MeSH                     | Fatty Liver (脂肪肝)                                                                                              |                                                                                                                                                                                                                                                                                                                                          |                 |                  |                                   |
| Concept 1<br>(Preferred) | Fatty Liver(脂肪肝)                                                                                               |                                                                                                                                                                                                                                                                                                                                          |                 |                  |                                   |
|                          | Concept UI                                                                                                     | M0008275                                                                                                                                                                                                                                                                                                                                 |                 |                  |                                   |
|                          | Scope Note                                                                                                     | Lipid infiltration of the hepatic parenchymal cells resulting in a yellow-colored liver. The abnormal lipid accumulation is usually in the form of <a href="#">TRIGLYCERIDES</a> , either as a single large droplet or multiple small droplets. Fatty liver is caused by an imbalance in the metabolism of <a href="#">FATTY ACIDS</a> . |                 |                  |                                   |
|                          | Semantic Type                                                                                                  | T047 (Disease or Syndrome)                                                                                                                                                                                                                                                                                                               |                 |                  |                                   |
|                          | Term(Preferred)                                                                                                | Fatty Liver                                                                                                                                                                                                                                                                                                                              | Term UI:T016028 | Date:01-JAN-1999 | Thesaurus:NLM (1966)              |
| Concept 2<br>(Narrower)  | Steatohepatitis(脂肪性肝炎)                                                                                         |                                                                                                                                                                                                                                                                                                                                          |                 |                  |                                   |
|                          | Concept UI                                                                                                     | M0533572                                                                                                                                                                                                                                                                                                                                 |                 |                  |                                   |
|                          | Scope Note                                                                                                     | Inflammation of the liver related to lipid accumulation in fatty liver.                                                                                                                                                                                                                                                                  |                 |                  |                                   |
|                          | Semantic Type                                                                                                  | T047 (Disease or Syndrome)                                                                                                                                                                                                                                                                                                               |                 |                  |                                   |
|                          | Term(Preferred)                                                                                                | Steatohepatitis                                                                                                                                                                                                                                                                                                                          | TermUI:T747355  | Date:10-MAR-2009 | Thesaurus:NLM (2010)              |
|                          | Term                                                                                                           | Liver Steatosis                                                                                                                                                                                                                                                                                                                          | TermUI:T747367  | Date:10-MAR-2009 | Thesaurus: NLM (2010)             |
|                          | Term                                                                                                           | Steatosis of Liver                                                                                                                                                                                                                                                                                                                       | TermUI:T768838  | Date:07-APR-2010 | Thesaurus: NLM (2011), ORD (2010) |
| Tree Number              | C06.552.241                                                                                                    |                                                                                                                                                                                                                                                                                                                                          |                 |                  |                                   |
|                          | BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH UR US VE VI |                                                                                                                                                                                                                                                                                                                                          |                 |                  |                                   |
| Date of Entry            | 19990101                                                                                                       |                                                                                                                                                                                                                                                                                                                                          |                 |                  |                                   |
| Unique ID                | D005234                                                                                                        |                                                                                                                                                                                                                                                                                                                                          |                 |                  |                                   |

图1 MeSH的三级概念结构

图2 SNOMED CT概念逻辑化定义示例图<sup>[25]</sup>

(4) SNOMED CT的数据结构：概念表、描述表、关系表是SNOMED CT体系结构的三个重要部件。概念表收录了19个顶级概念轴中约32万条具有唯一含义、并经过逻辑定义的概念；描述表收录了80多万条能够代表某个具体概念的术语及该概念的同义词；关系表包括约136万个三元组语义关联<sup>[28,29]</sup>。

### 3.3 UMLS的构建模式

UMLS是生物医学领域、跨语言多表集成的知识组织体系，整合了上百个异构异质MedKOS中的上千万个概念名称及相关信息。UMLS由三个密切相关

的知识源构成<sup>[30,31]</sup>。

(1) 超级叙词表 (Metathesaurus)：超级叙词表整合了166部生物医学领域的叙词表、分类表、本体、疾病编码集、标准化术语表等单一知识组织体系，2012AA版UMLS包括术语10,810,680条，概念单元2,669,792个，涵盖21种语言<sup>[32]</sup>。

超级叙词表以概念为核心进行组织，每个概念代表不同来源知识组织系统同义词的集合，超级叙词表概念结构 (Metathesaurus concept structure)<sup>[33]</sup>的设计为同义概念的组织、归并搭建了基本架构，基于概念唯一标识符 (CUI)、原形术语唯一标识符 (LUI)、字符串唯一标识符 (SUI)、来源术语唯一标识符 (AUI)，将不同MedKOS同一概念的各种名称和变体形式连在一起，搭建起知识组织体系间的语义关联。超级叙词表概念结构示例见图3。

超级叙词表通过对数据文档、元数据的合理设计和布局，将各来源词表纷繁庞杂的信息进行有序整合，通过概念名称文档、关系属性文档、描述属性文档等20多个文档的巧妙设计，实现了异构异质资源的同构化表示，并充分保留和继承了各来源表的原始信息。超级叙词表的每个来源表既是全局中的一个组成部分，又是一个独立的知识组织体系，既可整合利用，也可以被单独调用。

(2) 语义网 (Semantic Network)：UMLS语义网既是UMLS的一个组成部分，也是一个独立的顶级医

| CUI      | LUI      | SUI      | AUI      | 字符串                           | 来源表      |
|----------|----------|----------|----------|-------------------------------|----------|
| C0001403 | L0001403 | S0354372 | A4367951 | Addison's disease             | MTH      |
| C0001403 | L0001403 | S0354372 | A2922421 | Addison's disease             | SNOMEDCT |
| C0001403 | L0001403 | S0010794 | A0019740 | Addison's Disease (优选名称)      | MSH      |
| C0001403 | L0494851 | S2164152 | A2018589 | Primary adrenal insufficiency | MDR      |
| C0001403 | L3246333 | S3773545 | A3996251 | Addison, maladie              | MSHFRE   |

图3 超级叙词表概念结构示例图

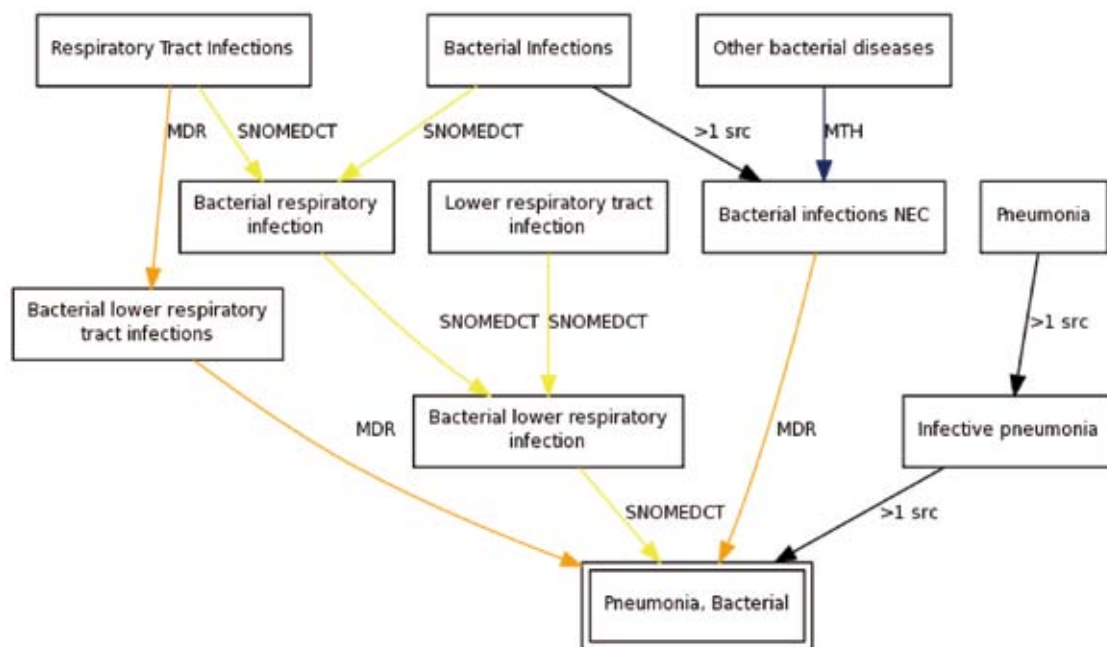
学语义网络。语义网由语义类型和语义关系构成，语义类型是语义网络中的节点，语义关系是节点之间的链接，如“后天畸形（语义类型1）+affects（语义关系）+细胞功能（语义类型2）”。语义类型有133个，涵盖了许多领域宽泛词汇的范畴；语义关系有54个，包括1个等级关系链（isa）和53个非等级的相关关系链<sup>[11]</sup>。基于UMLS语义网，为知识组织体系中的概念标识语义类型，可使其继承语义网的丰富语义关系。UMLS语义网具有全局统控作用，它可使完全相异的知识组织体系使用一个共同的语义网络，是一种基于全局本体统控、各来源表语义互联的模式，图4显示了概念“Pneumonia, Bacterial（细菌性肺炎）”在超级叙词表多个知识组织系统中的语义互联情况<sup>[12]</sup>。

（3）专家词典（SPECIALIST Lexicon）和工具：

是超级叙词表建立和维护所需的生物医学词汇库和一组工具集，主要用于超级叙词表同义概念的自动归并，减少人工构建词表的复杂程度。

#### 4 面向卫生信息标准的MedKOS

卫生信息标准化的重要内容之一是保证不同健康系统交换数据时对术语理解的一致性，这需要各种标准化知识组织体系的基础支撑作用。在统计分析的203个MedKOS中，已有60多个成为国际、区域、国家或行业的标准，涉及临床、基础医学、健康指标、电子处方、实验室检查、医疗设备、医疗保险、收费系统等方面。目前，卫生信息标准化术语集正在走向多语种多知识组织体系的交叉映射和集成融合。

图4 UMLS语义网概念导航图<sup>[12]</sup>

## 4.1 临床标准化MedKOS

(1) ICDs (International Classification of Diseases, 国际疾病分类法): ICDs是目前国际上共同使用的关于临床病案记录中疾病术语的标准分类编码体系。ICDs有43种语言版本,由世界卫生组织(WHO)负责编订、推广和应用。ICDs早在1900年出版第一版,约每10年修订一次,1994年开始启用ICD-10。在我国,ICD-10的中文版本早已被列为国家标准,于2001年开始实施,成为疾病与健康相关问题统计分类的重要标准和基础工作。ICD-11的修订工作已于2010年启动<sup>[34,35]</sup>。

(2) LOINC (Logical Observation Identifiers Names and Codes, 观测指标标识符逻辑命名与编码系统): LOINC由美国印第安纳大学医学中心Regenstrief研究院开发维护,是一套用于标识实验室检验项目和临床观测指标的标准化术语编码集,观测指标涉及血液学、微生物学、化学、生命体征、放射诊断报告、肿瘤登记指标、调查问卷等多类观测指标。LOINC中还包含近10种临床上已被广泛使用的标准化评价指标,如Braden (Braden压疮危险因素评估量表)、FLACC (FLACC疼痛量表)、OASIS (护理效果评估量表)等。目前LOINC已在世界范围内广泛使用,作为临床实验室结果电子信息交换的标准,包括近10种语言版本<sup>[19,36]</sup>。

(3) “NNN” 护理分类系统<sup>[37]</sup>: “NNN”代表目前在国际护理专业上已经普遍采纳的三个标准化分类系统: NANDA (NANDA nursing diagnoses, NANDA护理诊断表)<sup>[38]</sup>、NIC (Nursing Interventions Classification, 护理措施分类表)<sup>[39]</sup>、NOC (Nursing Outcomes Classification, 护理结局分类表)<sup>[40]</sup>。三个系统相互关联,目前已有包括中文在内的多种语言版本。

## 4.2 标准术语集的映射和集成融合

(1) 标准术语集的交叉映射: 不同术语集间的映射是医学资源集成融合的重要基础任务,如美国国家健康和人类服务部(HHS)在全美范围内实施健康信息技术互操作目标,强调标准化术语和代码要与病人、医疗保健提供者、实验室和诊断设施等数据资源紧密关联,以提高医疗保健的质量和效率。LOINC观测指标编码体系与CPT (通用医疗操作术语集)<sup>[41]</sup>账

单编码体系间的映射项目就是该目标的一个实例<sup>[42]</sup>。

(2) 标准术语集的集成融合: 目前已有越来越多的标准化术语集实现了不同层面的集成和融合,如美国国立医学图书馆建立的RxNorm (处方药物标准术语集)<sup>[18,43]</sup>,是美国联邦政府指定的进行临床电子数据交换采用的处方药物标准命名表。它采用UMLS的方法和技术,整合了包括Alchemy金标准、FDA国家药品代码目录(NDC)、VHA国家药品文件-参考术语(NDF-RT)在内的11个词表中的标准药物术语,对药物通用名和商品名采用标准规范的描述方式,命名中包括药物的成分、剂量和剂型。

(3) 标准化知识表示系统: 这类系统为不同来源医学学术语建立了标准化的知识表示架构,如前面提到过的UMLS项目,它不仅提供各种临床数据的集成融合,也提供临床数据与文献数据、基因数据等在更大领域范围内的一致性和兼容性,可帮助解决医学信息交换中的许多问题。

## 5 基于本体架构的MedKOS

本体的开发和利用是实现语义网、支撑全球网络化知识交流的核心要素,也是MedKOS建设与发展的主流趋势。本体构建是一项极其庞大的复杂工程,基于传统MedKOS构建本体已成为业界的共识。通过XML、RDF、OWL等描述语言进行本体转换的例子举不胜举,较典型的有基于XML的MeSH叙词表的表示,基于OWL的NCI叙词表的描述等等。在转换基础上通过添加领域知识,添加概念、类和实例,最终形成具有丰富语义关系的本体。此外,本体构建方法和工具、本体描述语言、本体映射、本体集成等都是研究的热点。目前已产生大量生物医学本体,下面是几个有代表性的研究成果。

### 5.1 生物医学领域本体建设

(1) FMA (Foundational Model of Anatomy Ontology, 解剖学基础模型本体): FMA由美国华盛顿大学结构信息学研究组(SIG)编制和维护。FMA是解剖学领域的一部参考本体,以符号形式表达解剖学信息,由四个相互关联的部分组成,解剖学类目、解剖学结构抽象表达、解剖学变化抽象表达,以及对类和关系的原则、规则和定义进行解释的Metaknowledge

(元知识)<sup>[14]</sup>。FME是一个用于浏览和组织FMA的软件工具<sup>[44]</sup>。

(2) HPO (Human Phenotype Ontology, 人类表型本体): HPO是一个描述人类疾病表型异常的受控词表, 如“atrial septal defect (心房间隔缺损)”, 术语采用逻辑定义的方式描述。HPO初期基于OMIM (在线人类孟德尔遗传数据库) 建立, 现在增加了医学文献中的信息, 由柏林Charité医院和英国剑桥大学联盟建立<sup>[45]</sup>。

(3) GO (Gene Ontology, 基因本体): GO由基因本体联盟建立和维护。GO的主要目标是对跨物种和跨数据库的基因表达和基因产物属性进行结构化及定义精确的描述。GO建立了三个独立的结构化的受控词表, 分别为生物学过程、细胞成分和分子功能词表, 用于描述生物体的基因产物<sup>[46]</sup>。

## 5.2 生物医学本体集成系统

(1) NCBO BioPortal (National Center for Biomedical Ontology BioPortal, 美国国家生物医学本体中心BioPortal): BioPortal是一个包括8个来源的300多个生物医学本体, 如图5所示, 包括OBO Foundry、UMLS等8个。它是一个开放的生物医学本体仓库及集成系统, 通过Web浏览器和服务器提供对OBO格式、OWL、RDF、RRF格式、Protégé框架开发

的本体的访问。图5大致可以看出BioPortal的主要功能, 包括浏览、搜索、可视化显示、建立不同本体术语间的映射、添加注释到本体; 另外它通过Annotator和Resource Index工具集成了诸如PubMed、PubChem、OMIM等不同领域的20多个异构生物医学数据资源。Annotator是一个利用本体对自由文本进行标注的工具, Resource Index可通过本体关联生物医学数据资源<sup>[47]</sup>。

(2) OBO Foundry (Open Biological and Biomedical Ontologies, 开放生物医学本体): OBO Foundry官方网站由美国加州大学伯克利生物信息学开放资源项目组 (BBOP) 建立和维护。OBO Foundry是一个基于生物医学数据本体的联合实验项目, 目的是建立一套适合生物医学领域、可互操作的参考本体的构建规则, 实现跨生物学和医学研究领域间的语义整合。OBO Foundry中比较重要的生物医学本体有FMA、HPO、GO、CL (细胞类型本体)、IDO (传染病本体), 以及NCBITaxon (NCBI生物分类法) 等几十种资源<sup>[48]</sup>。OBO-Edit是一个开源的跨平台的本体编辑软件, 主要用于浏览和编辑OBO文件<sup>[49]</sup>。

此外还有国际合作项目OBI (生物医学调查本体)<sup>[50]</sup>、欧盟的开放临床术语通用框架OpenGALEN<sup>[51]</sup>, 以及美国梅奥医学中心生物医学信息学部设计的Lexical Grid (词汇网格)<sup>[52]</sup>等项目。

| ONTOLGY NAME                     |        | REVIEWS | PROJECTS | UPLOADED   |
|----------------------------------|--------|---------|----------|------------|
| Gene Ontology (GO)               |        | 0       | 1        | 11/22/2012 |
| Gene Ontology Extension (GO-EXT) | Public | 34,400  | 0        | 06/24/2011 |
| ICD10 (ICD10)                    | Public | 12,318  | 0        | 07/21/2010 |
| ICD10CM (ICD10CM)                | Public | 91,590  | 0        | 09/29/2011 |

图5 NCBO BioPortal浏览界面<sup>[47]</sup>

## 6 结语

### 6.1 MedKOS研究重点

通过对医学领域知识组织体系发展现状的系统梳理及对有代表性的MedKOS的分析,当前医学领域知识组织体系的发展和研究主要表现在以下几个方面。

(1) 一体化医学语言系统在未来MedKOS建设与发展过程中继续发挥着引领与示范作用,UMLS的构建及在医疗信息系统、病案系统、自然语言处理、文本自动标注、智能检索及搜索引擎等领域的广泛应用,推进了医学领域知识组织体系的集成整合发展进程。目前,多种知识组织体系关联集成模式也是知识组织技术的发展方向。

(2) 领域本体已成为当前MedKOS发展的主流活动,生物医学本体的开发一直走在领域本体建设的前列,目前已取得一些有代表性的成果。研究重点集中在传统知识组织体系向本体转化、本体构建方法与工具集的建立、本体表示语言、本体映射等方面,并且向更大规模的集成和融合发展。一些医学知识本体的特有结构模式如SNOMED CT得到了广泛的认同和应用。本体由于概念及语义关系揭示的明确化、模型化和机器可理解等特点,成为语义网环境下不可或缺的知识组织工具,呈现开放、关联、集成、国际化的发展态势。

(3) 医学知识组织体系之间的映射,是实现不同MedKOS集成、融合,进而实现分散异构数据资源互操作的方法,也是今后MedKOS建设与发展的重要基础任务。医学领域大型知识组织体系间的映射研究以国际疾病分类法(ICDs)为代表。ICDs的映射涉及ICD家族不同版本间的映射,如ICD-9到ICD-10的映射转换;以及ICDs与外部知识组织体系之间的映射,如ICD-10-CM与SNOMED CT的映射等多种知识组织体系,此外还包括多种语言版本ICD的映射,如中文等40多种语言。

(4) 当前医学术语标准化活动更加关注卫生信息

标准化建设,特别注重临床疾病、药物、实验室等多种标准化术语集的融合使用,强调临床数据与文献数据、基因数据等在更大领域范围内的一致性和兼容性,以促进医学领域各种复杂数据资源的知识整合与知识共享,这也是未来医学术语标准化建设的重要命题。

(5) 知识组织体系正处于快速变革的重要机遇期,我们既熟悉又陌生。以MeSH为代表的传统知识组织体系结构模式的不断变革带给了我们新的启示。MedKOS的创新发展、持续性和动态性、大众参与性,都将促进知识组织体系的持续健康发展。

### 6.2 推进我国MedKOS建设

研究分析显示,我国医学领域知识组织体系建设与国外有很大差距,要使国外先进经验变成符合我国国情的有益经验,要从认真学习做起。

引进、消化、吸收国际上广泛使用的MedKOS,进行翻译和本地化改造,符合国际上的通用做法,如权威知识组织体系的多语言版本,美国、澳大利亚对ICD-10的本地化改造等等。在这些方面我国虽也进行了一些有益的尝试,如ICD10中文版的发布、MeSH中文版的发布等,但这些还远远不能满足医学研究和医学实践的需要,MedKOS的本地化改造也与国外有很大差距。国家“十二五”科技支撑计划“面向外科技信息的知识组织体系建设与应用示范”项目,为推进我国MedKOS的建设提供了难得的发展机遇。我们要认真分析研究国际上著名的医学知识组织体系,结合我国实际情况,将医学知识组织体系建设融入医学科研和医学实践的整个活动之中,切实推进我国MedKOS的建设与发展。

## 致谢

本研究得到了侯丽、孙海霞、冀玉静、陈颖、李芳、钱庆等同志的参与支持,在此表示感谢。

## 参考文献

- [1] 闫巍,曾民族.构筑知识基础结构的关键技术[J].现代图书情报技术,2005(8):1-6,31.
- [2] 曾蕾.用于标引浏览检索的语义工具[EB/OL].[2007-12-06].<http://www.libnet.sh.cn/upload/htmleditor/File/071213121547.pdf>.
- [3] Unified Medical Language System® [EB/OL].[2012-05-28].<http://www.nlm.nih.gov/research/umls/>.
- [4] National Cancer Institute. NCI Metathesaurus[DB/OL].[2012-05-28].<http://ncim.nci.nih.gov/ncimbrowser/>.
- [5] 李丹亚,胡铁军,李军莲,等.中文一体化医学语言系统构建与应用[J].情报杂志,2011,30(2):147-151.
- [6] 曾召,王小平.UMLS与中医药一体化语言系统的建立[J].中华医学图书情报杂志,2006,15(3):1-3.

- [7] National Library of Medicine. MeSH Browser [DB/OL]. [2012-05-28]. <http://www.nlm.nih.gov/mesh/MBrowser.html>.
- [8] National Cancer Institute. NCI Thesaurus [DB/OL]. [2012-05-28]. <http://ncit.nci.nih.gov/>.
- [9] 李丹亚,胡铁军,诸文雁,等.中文医学主题词表检索系统[J].中华医学图书馆杂志,2001,10(4):1-2,9.
- [10] 中国中医研究院中医药信息研究所.中国中医药学主题词表[M].北京:中国古籍出版社,1996.
- [11] 5 Semantic Network. UMLS® Reference Manual [EB/OL]. [2009-09-01]. <http://www.ncbi.nlm.nih.gov/books/NBK9679/>.
- [12] UMLS Semantic Navigator [DB/OL]. [2012-05-28]. <http://mor2.nlm.nih.gov:8000/perl/auth/semnav.pl>.
- [13] SNOMED CT [EB/OL]. [2012-05-28]. <http://www.ihtsdo.org/snomed-ct/>.
- [14] Foundational Model of Anatomy [EB/OL]. [2012-05-28]. <http://fma.biostr.washington.edu/>.
- [15] 《中国图书馆分类法》编辑委员会,中国医学科学院信息图书馆.中国图书馆分类法-医学专业分类表[M].北京:北京图书馆出版社,1999.
- [16] ICD-10 Version:2010 [DB/OL]. [2012-05-28]. <http://apps.who.int/classifications/icd10/browse/2010/en>.
- [17] NCBI Taxonomy [DB/OL]. [2012-05-28]. <http://www.ncbi.nlm.nih.gov/taxonomy>.
- [18] RxNorm [EB/OL]. [2012-05-28]. <http://www.nlm.nih.gov/research/umls/rxnorm/index.html>.
- [19] Logical Observation Identifiers Names and Codes (LOINC®) [EB/OL]. [2012-05-28]. <http://loinc.org/>.
- [20] HUGO Gene Nomenclature [DB/OL]. [2012-07-01]. <http://www.genenames.org/>.
- [21] 张琪玉.寻找更加结合模式是情报检索语言创新的主流[J].图书馆杂志,2005,24(2):16-18.
- [22] 李丹亚,李军莲,胡铁军.MeSH的概念结构及其意义[J].医学信息学杂志,2010,31(11):54-58.
- [23] 李丹亚,胡铁军,李军莲.MeSH增补概念的术语映射分析[J].医学信息学杂志,2012,33(4):45-49.
- [24] STUDER R, BENJAMINS V R, FENSEL D. Knowledge Engineering Principles and Methods [J]. Data and Knowledge Engineering, 1998, 25(122): 161-197.
- [25] International Health Terminology Standards Development Organisation. SNOMED CT® User Guide January 2012 International Release [EB/OL]. [2012-06-04]. [http://ihtsdo.org/fileadmin/user\\_upload/doc/download/doc\\_UserGuide\\_Current-en-GB\\_INT\\_20120131.pdf](http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_UserGuide_Current-en-GB_INT_20120131.pdf).
- [26] 郭玉峰,刘保廷,尹爱宁.SNOMED CT术语分类体系设定学科背景的探讨[J].世界科学技术-中医药现代化,2007,9(4):86-90.
- [27] 郭玉峰,刘保廷,周雪忠.SNOMED CT的语义关系与连接概念[J].中华中医药学刊,2008,26(10):2206-2209.
- [28] SNOMED CT Component [EB/OL]. [2012-06-04]. <http://www.ihtsdo.org/snomed-ct/snomed-ct0/snomed-ct-components/>.
- [29] 2011AB SNOMED CT Source Information [EB/OL]. [2012-06-04]. <http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/SNOMEDCT/>.
- [30] 2 Metathesaurus [EB/OL]. [2009-09-01]. <http://www.ncbi.nlm.nih.gov/books/NBK9684/>.
- [31] 李丹亚,胡铁军,李亚子,等.UMLS多词表整合机制研究[J].数字图书馆论坛,2012(4):28-46.
- [32] National Library of Medicine. Statistics - 2012AA Release [EB/OL]. [2012-05-28]. [http://www.nlm.nih.gov/research/umls/knowledge\\_sources/metathesaurus/release/statistics.html](http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html).
- [33] NELSON S J, POWELL T, HUMPHREYS B L. The Unified Medical Language System® (UMLS®) Project [EB/OL]. [2010-08-22]. <http://www.nlm.nih.gov/mesh/umlsforelis.html>.
- [34] International Classification of Diseases (ICD) [EB/OL]. [2012-06-10]. <http://apps.who.int/classifications/icd/en/>.
- [35] WHO Business Plan for Classifications [EB/OL]. [2012-06-10]. <http://www.who.int/classifications/BusinessPlan.pdf>.
- [36] MCDONALD C, HUFF S, MERCER K. LOINC® Users' Guide - December 2011 [EB/OL]. [2011-06-15]. <http://loinc.org/downloads/files/LOINCManual.pdf>.
- [37] 詹森,等.护理诊断、结局与措施(第2版)[M].吴袁剑云,等译.北京:北京大学医学出版社,2010.
- [38] NANDA-I Nursing Diagnoses: Definitions and Classification 2012-2014 [EB/OL]. [2012-06-15]. <http://www.nanda.org/Home.aspx>.
- [39] CNC-Overview: Nursing Outcomes Classification (NOC) [EB/OL]. [2012-06-15]. <http://www.nursing.uiowa.edu/cncce/nursing-outcomes-classification-overview>.
- [40] CNC-Overview: Nursing Interventions Classification (NIC) [EB/OL]. [2012-06-15]. <http://www.nursing.uiowa.edu/cncce/nursing-interventions-classification-overview>.
- [41] CPT - Current Procedural Terminology [EB/OL]. [2012-06-10]. <http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt.page>.
- [42] LOINC To CPT Mapping [EB/OL]. [2012-06-10]. [http://www.nlm.nih.gov/research/umls/mapping\\_projects/loinc\\_to\\_cpt\\_map.html](http://www.nlm.nih.gov/research/umls/mapping_projects/loinc_to_cpt_map.html).
- [43] RxNorm Technical Documentation [EB/OL]. [2012-06-10]. [http://www.nlm.nih.gov/research/umls/rxnorm/docs/2012/rxnorm\\_doco\\_full\\_2012-3.html](http://www.nlm.nih.gov/research/umls/rxnorm/docs/2012/rxnorm_doco_full_2012-3.html).
- [44] Foundational Model Explorer [DB/OL]. [2012-05-28]. <http://fme.biostr.washington.edu/FME/index.html>.
- [45] Human Phenotype Ontology Website [EB/OL]. [2012-06-28]. <http://www.human-phenotype-ontology.org/>.
- [46] Gene Ontology [DB/OL]. [2011-06-15]. <http://www.geneontology.org/>.
- [47] NCBO BioPortal [DB/OL]. [2012-06-28]. <http://www.bioontology.org/BioPortal>.
- [48] The Open Biological and Biomedical Ontologies [DB/OL]. [2012-06-10]. <http://www.obofoundry.org/>.
- [49] The OBO Ontology Editor [DB/OL]. [2012-06-10]. <http://oboedit.org/>.
- [50] Ontology for Biomedical Investigations [EB/OL]. [2012-06-28]. <http://obi.sourceforge.net/resources/index.php>.
- [51] OpenGALEN [EB/OL]. [2012-06-10]. <http://www.opengalen.org/>.
- [52] Mayo Clinic Informatics [EB/OL]. [2012-06-28]. <http://www.lexgrid.org/>.

## 作者简介

李丹亚, 研究员。研究方向: 医学知识组织系统, 医学信息资源管理与利用。E-mail: li.danya@imicams.ac.cn

## A Survey of Medical Knowledge Organization System

Li Danya, Li Junlan, Li Xiaoying, Xia Guanghui, Hu Tiejun / Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, 100020

Abstract: In this paper, by studying more than 200 active medical glossaries, taxonomies, thesauri, ontologies, and unified language systems, the current state of the biomedical knowledge organization systems was summarized. Further, the paper deeply dived into the construction models of some popular medical knowledge organization systems, such as MeSH, SNOMED CT, UMLS and analyzed the key projects in standardized medical terminologies and medical ontologies. Finally, it briefly discussed the research directions of medical knowledge organization systems and the future development in China.

Keywords: Medical knowledge organization systems, Construction model, Biomedical terminology standardization, Medical ontologies, UMLS, SNOMED CT, MeSH

(收稿日期: 2012-08-01)