MeSH 增补概念的术语映射分析

李丹亚 胡铁军 李军莲

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 基于对 MeSH 增补概念结构的分析，重点探讨增补概念与 MeSH 主题词的术语映射和转换，及与 UMLS 语言网的术语映射，试图从 MeSH 微观结构的一个侧面对其功能、适应性及使用效率进行分析。

〔关键词〕 MeSH；增补概念记录；术语映射；UMLS 语义网；知识组织系统

Analysis on Terminology Mapping in MeSH Supplementary Concept *LI Dan-ya, HU Tie-jun, LI Jun-lian, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 Based on analyzing the MeSH supplementary concept structure, the paper mainly discusses the terminology mapping and conversion between MeSH supplementary concept and medical subject heading, as well as between the MeSH supplementary concept and UMLS semantic network, so as to illustrate the functionality, adaptability and efficiency of MeSH from the aspect of its microstructure.

〔**Keywords**〕 MeSH; Supplementary concept records; Terminology mapping; UMLS semantic network; Knowledge organization system

1 引言

知识组织系统的术语映射和关联通常是指不同叙词表、分类表、本体等知识组织系统之间的兼容转换，其中也包括知识组织系统内部建立的术语映射，如美国国立医学图书馆（NLM）的 MeSH（Medical Subject Headings）词表。MeSH 由主题词表、副主题词表、增补概念表、树形结构表等组成[1]。增补概念表是对主题词表的补充和扩展，包括 3 个子表或称卫星子表。MeSH 增补概念记录（Supplementary Concept Records，简称 SCRs）[2]的术语不是主题词，多来自文献中的专业术语。对于这类术语，MeSH 建立了增补概念表与主题词表的术语映射，同时也与 UMLS（Unified Medical Language System）的语义类型（或称宏词表）建立了术语映射。这些关联机制的建立提升了 MeSH 词表的整体性能，使其更加适应网络环境下的应用。对 MeSH 的微观结构进行分析，有助于深入探析检索语言结构与功能的关系，探索适应网络环境的知识组织工具的组合模式及应用模式。

2 基于术语映射的 SCRs 表结构

2.1 SCRs 结构的演变

MeSH 于 1960 年正式出版[3]，1980 年 MeSH 在已有主题词表和副主题词表的基础上，基于文献处理中积累的化学物质名称创建了补充化学物质表，建设伊始就设计了化学物质名称与主题词表的术语映射及关联机制。1999 年基于 MeSH 以概念为中心的构建理念和模式，补充化学物质表逐渐从原有的词结构组织方式向基于概念结构的组织方式过渡，并建立起与 UMLS 语义网的术语映射关联，这项工程耗时多年，MeSH 的补充化学

〔收稿日期〕 2011-12-23

〔作者简介〕 李丹亚，研究员，发表论文 40 余篇。

物质表随之更名为增补概念表，1997 年网络版 MeSH 检索系统 MeSH Browser 对外发布了新的概念结构。近几年 MeSH 增补概念表开始增加有关治疗方案方面的术语，2011 年又增加了罕见疾病方面的术语，并延续了与 MeSH 主题词表及 UMLS 语义网的术语映射机制[2,4~6]。目前 MeSH 增补概念表包括 3 个子表：化学物质和药物名称、治疗方案术语及罕见疾病术语子表。

2.2 SCRs 的基本结构

MeSH 增补概念表的术语不是主题词，但结构与 MeSH 主题词表非常类似。MeSH 增补概念表支持与 MeSH 主题词表及 UMLS 语义网的术语映射。包括词结构与概念结构 2 种模式，词结构由入口词、优选词 2 级结构组成；概念结构由类、概念和术语 3 级结构组成。概念和术语之间为严格的同义关系，类和概念之间可以是密切关联的同义、相关、狭义或广义关系[1,7]。SCRs 类中的每个成员共享与 MeSH 主题词和 UMLS 语义类型的映射关系，这种关系多为广义与狭义关系。增补概念记录除标识出词间关系、映射关系外，还包括丰富的词义注释、专业特征信息、使用及管理信息等，表 1 是增补概念表的元数据描述[8]。

表 1 增补概念元数据描述

英文名称	缩写	中文名称	描述信息
Name of Substance	NM	优选名称	原来称物质名称，现在指增补概念的优选名称
CAS Type 1 Name	N1	CAS1 型名称	首选 Chemline 数据库 CAS（化学文摘社）1 型名称，为 CAS 分配的系统名称；如无则选择 Medline 数据库中使用的系统名称
Concept	-	概念名称	与 NM 字段和 N1 字段术语有非严格意义上的同义关系，即密切关联的同义、相关、狭义或广义关系
Term	-	术语名称	表示与概念之间为严格的同义关系
CAS Registry/EC Number	RN	CAS 登记号/EC 编号	含有美国化学文摘社（CAS）分配的化学物质登记号，以及国际生物化学与分子生物学联盟命名委员会分配的酶编号。
Related Registry Number	RR	相关登记号	含有 NM 字段化合物的盐类、光学异构体或同位素标记形式的 CAS 登记号
Heading Mapped-to	HM	映射主题词	优选名称及同义词映射的主题词、主题词/副主题词
Semantic Type	ST	语义类型	UMLS 语义网的语义类型
Pharm. Action	PA	药理作用	含有体现该化学物生物活性的映射药理作用主题词
Indexing Information	II	映射标引说明	含有供标引人员考虑使用的、除 NM 和 PA 字段中有关术语以外的相关映射主题词或映射主题词/副主题词
Previous Indexing	PI	标引回溯说明	含 NM 字段术语曾使用过的映射主题词或主题词/副主题词
Source	SO	术语来源	术语首次出现在 Medline/ PubMed 数据库的文献出处
Frequency	FR	出现频率	含有 Medline/PubMed 标引期刊文献中标识术语出现频率
Thesaurus Id	TH	词典标识	标识术语的权威来源参考，如 USAN（美国药名）、Merk 索引、NEGWER（有机化学药物及同义词化学词典）
Note	-	注释	注释信息
Date of Entry	DE	建立日期	记录建立日期
Revision Date	RD	修订日期	记录修订日期
Unique Identifier	UI	唯一标识符	增补概念记录唯一标识符。
Record Type	RY	记录类型	标识 MeSH 的记录类型，增补概念记录用“C”标识

3 SCRs 与主题词表的术语映射

3.1 子表的术语映射

增补概念表包括 3 种类型的子表，不同子表其术语映射形式也不尽相同。子表的组织方式与卫星子表相类似，卫星子表通常由一组专门的词汇构成，子表又与某个范围更广的母表相融合，作为其中的一个组成部分。

3.1.1 *化学物质和药物名称子表* 该表收录了文献中出现的化学物质和药物的各种名称，其优选名称的确定通常基于权威工具书或药典中的通用名称，每条记录包括化学物质详细的专业特征信息和映射主题词。术语与主题词的映射基于不同的维度，如体现化学物质名称的映射主题词或映射主题词/副主题词、体现化学物质生物活性作用的映射药理作用主题词以及可能与化学物质相关的其他映射主题词。图 1 中，Tiadenol（硫地醇）的映射主题词是脂肪族醇类，映射药理作用主题词是降血脂药，该例中未提及其他的相关映射主题词。

NM	Tiadenol (硫地醇)	
Concept 1 (Preferred)	Tiadenol	
	ST	T109 (Organic Chemical)
	N1	2,2' -(1,10-decanediylbis(thio)) bis-ethanol
	RN	6964-20-1
	Term	Tiadenol
	Term	2,2' -(decamethylenedithio)-diethanol
	Term	bis(2- hydroxyethylthio)-1,10-decane
	Term	bis(hydroxy-2-ethylthio)-1,10-decane
	Term	Thiadenol
	Term	Tiadenolo
Concept 2 (Narrower)	BS 530	
	Term	BS 530
Concept 3	……	
HM	*Fatty Alcohols (脂肪族醇类)	
PI	SULFIDES (73-76)	
SO	Arch Int Pharmacodyn Ther 222(1):166;1976 …	
PA	Hypolipidemic Agents (降血脂药)	
FR	60	
UI	C004568	

图 1 化学物质和药物术语（概念结构）

针对化学物质和药物名称子表，MeSH Browser 除提供主题词、副主题词、增补概念 3 个检索途径外，还专门设计了特色检索路径，如映射主题词（HM）、映射药理作用主题词（PA）、映射相关主题词（II）、CAS 登记号或/EC 编码（RN）及相关 CAS 登记号（RR）[6]。

3.1.2 *治疗方案名称子表* 治疗方案指疾病的联合用药方案，每条治疗方案术语标识了治疗方案中每种药物映射的主题词及治疗方案名称自身的映射主题词。图 2 显示了 MOPP 治疗方案所映射的能够体现联合用药方案中实质内容的多个主题词，治疗方案的注释信息及在 Medline/PubMed 数据库文献题录中出现的词频数。

NM	MOPP protocol (MOPP方案)
HM	*Antineoplastic Combined Chemotherapy Protocols (*抗肿瘤联合化疗方案)
HM	Mechlorethamine (氮芥)
HM	Prednisone (泼尼松)
HM	Procarbazine (丙卡巴肼)
HM	Vincristine (长春新碱)
FR	740
Note	chemotherapy protocol consisting of above cpds; used in treating Hodgkins disease; MOP is combination without prednisone

图 2 疾病治疗方案术语

3.1.3 *罕见疾病名称子表* 2010 年 NLM 开始增加美国国立卫生研究院（NIH）罕见疾病研究办公室（ORDE）编制的罕见疾病术语表到 MeSH 词表，总计处理了 10 379 个 ORDR 术语，其中有些是 MeSH 中已经存在的，与 MeSH 主题词表进行合并，其余 ORDR 术语于 2011 年加入到 MeSH 的增补概念表[9]。图 3 是罕见疾病术语与主题词的映射实例[6]，罕见疾病术语通常需要映射多个主题词才能揭示疾病的特征和本质。增加这类术语有助于提高对罕见疾病的认识、鉴别，及提高文献检索效果。

NM	KBG syndrome (KBG综合征)
ST	T047 (Disease or Syndrome)(疾病或综合征)
HM	*Abnormalities, Multiple (畸形,多发性)
HM	*Bone Diseases,Developmental(骨疾病,发育性)
HM	*Mental Retardation(精神发育迟滞)
HM	*Tooth Abnormalities(牙畸形)
HM	*Facies(面容)

图 3 罕见疾病术语

3.2 术语映射和转换过程

增补概念表与主题词表的术语映射遵循最相近的语义原则、最精确的匹配原则及最广泛的兼容原则。映射源与映射目标可以是一对一或一对多形式，映射主题词可以采用主题词/副主题词组配形式，如“tuftsin-AZT conjugate”映射到 2 个主题词/副主题词组配形式“Tuftsin/ * analogs & derivatives（促吞噬肽/类似物和衍生物）、Zidovudine/ * analogs & derivatives（齐多夫定/类似物和衍生物）”，通过多种映射形式最终达到表间术语的语义最为接近、匹配最为精确。最广泛兼容原则是指在无法精确匹配的情况下，可通过上位匹配、下位匹配或相关匹配为映射源和映射目标术语找到映射关系，映射源与映射目标之间多为狭义与广义关系。术语映射的结果通常是指一组术语及其关系的集合。

增补概念名称通过与主题词的映射转换，使其具有等同于 MeSH 主题词的标引和检索功能，记录中的任何术语可直接用于 PubMed/Medline 及相关数据库文献题录的标引和检索，可以像主题词一样进行各种操作，如与副主题词的组配及加权等。

系统基于增补概念表自动对文献数据库题录进行主题词的转换及标注，并将标注时所用术语的优选词提取出来存放在文献题录的增补概念字段，对于化学物质和药物术语，还提取记录中的 CAS 登记号或酶 EC 编号[5,8]加入文献题录。

4 SCRs 与 UMLS 语义网的术语映射

UMLS 语义网由语义类型和语义关系组成[11]，语义类型的组织与医学领域宏词表的组织有异曲同工之处，均为术语体系的通用上层结构，这种结构支持多种词表的相互关联。MeSH 的增补概念表，每条记录至少与 UMLS 语义网的一个语义类型建立映射，如图 3 中的“T047（Disease or Syndrome）（疾病或综合征）”表示语义类型及其树形结构号。

UMLS 语义类型将所有具有同样上层结构的词表术语相互关联，包括医学领域的叙词表、分类表、本体等知识组织工具，增补概念表通过语义类型与多个外部知识组织工具建立起了关联。此外 UMLS 语义类型之间的语义关系形成了上层结构术语间的丰富语义关系，这种关系可以向下传递，也丰富了增补概念术语之间的语义关系。

5 SCRs 术语映射与词表的功能

5.1 支持受控语言与自然语言的融合

MeSH 的增补概念结构是实现主题语言与自然语言结合的一种值得借鉴的模式。NLM 的标引人员可随时利用期刊文献中新出现的有关化学物质、药物、治疗方案、罕见疾病名称等方面的术语对文献进行标引，使用者亦可用其进行检索。基于术语映射，增补概念表具有了主题语言的某些特征。增补概念表的设计、结构及映射关联，从一个侧面体现了情报检索语言对自然语言的规范控制功能。

5.2 词表内与词表间术语映射的集成融合

MeSH 增补概念表通过与 MeSH 主题词表内的术语映射、词表外与 UMLS 语义网的术语映射，使非主题词的增补概念术语具备了更多的属性和功能，一方面增补概念术语继承了主题词的关系和特征，另一方面通过语义网丰富了词间关系，并与外部知识组织工具建立起广泛的关联，其联动效应提升了 MeSH 词表在知识组织和知识服务中的使用效率。

5.3 基于增补概念表的专业实体知识库

增补概念表收录了大量与生物医学有关的非主题词的专业术语，通过 MeSH 词表的映射关联机制，术语表达更加规范、词间关系更加细化。经过领域专家和使用者的共同努力，增补概念表的术语量不断增加，词义注释与专业特征信息更加翔实，成为专业领域具有较高权威性的知识工具，并逐渐形成网络百科的某些特征和功能，使 MeSH 主题词表的功能不断延伸。

5.4 基于增补概念表的词表扩展特征

基于增补概念表可很方便地对词表进行扩展和维护，如支持随时添加新的术语，无需控制词量，

增补概念表每周更新，不同于主题词表、副主题词表的年度更新。增补概念表有助于新增主题词的推荐及词间关系的提取，也有助于专业领域子表的扩展和外部词表的加入，并可为领域本体构建提供支撑。增补概念表的设计为 MeSH 词表的扩展提供了发展空间，具有良好的兼容性和灵活性。

5.5 支持文本信息处理及智能检索应用

增补概念表既包括来自文献的术语，又与主题词、语义类型关联。根据不同的应用目的，可用于分词、信息抽取、聚类、自动标注等文本信息处理工作；基于术语的不同级别，支持不同等级的扩检和缩检，能够实现不同颗粒度的智能检索；此外还可用于研究热点领域监测、专业知识挖掘及领域知识聚类等系列应用。

5.6 主题词表的性能及易用性

主题词表的性能通常可采用等同率、关联比、参照度等主要指标进行评价[12]。词表等同率也称入口词率，是指非主题词与主题词的比率；关联比是指有参照项的主题词与词表总词量的比例。MeSH 增补概念表的建立大大提高了 MeSH 词表的入口词率和关联比，提升了主题词表的性能及易用性，使其更加适合网络环境下海量信息资源知识组织的需求。

6 结语

MeSH 增补概念与 MeSH 主题词表及 UMLS 语义网的术语映射不完全等同于知识组织系统间的术语映射，它是知识组织系统内与知识组织系统间术语映射的集成融合，其构建模式有独特之处。检索语言的结构模式决定了检索语言的使用效率和适应性，MeSH 增补概念表的设计理念、结构模式、术语映射及转换机制，大大提升了 MeSH 词表的使用性能及易用性。对其进行分析，有助于探索适应网络环境的知识组织系统的构建模式，解决网络环境中遇到的新问题。

参考文献

1 Nelson S J, Johnston D, Humphreys B. Relationships in Medical Subject Headings [EB/OL]. [2010 - 06 - 10]. http: //www. nlm. nih. gov/mesh/meshrels. html.

2 MeSH® Vocabulary Updated for 2011. NLM Technical Bulletin, 2010 (11/12) [EB/OL]. [2011 - 12 - 01]. http: //www. nlm. nih. gov/pubs/techbull/nd10/nd10_ me-dline _ data_ changes_ 2011. html.

3 Celebrating MeSH: 50 years of Medical Subject Headings [EB/OL]. [2010 - 06 - 10]. http: //www. nlm. nih. gov/mesh/mesh_ at_ 50/mesh_ at_ 50. html.

4 Changes in MeSH Data Structure. NLM Technical Bulletin, 2000 (10) [EB/OL]. [2010 - 06 - 10]. http: //www. nlm. nih. gov/pubs/techbull/ma00/ma00_ mesh. html.

5 PubMed® Notes 2011 [EB/OL]. [2011 - 12 - 02]. http: //www. nlm. nih. gov/pubs/techbull/nd10/nd10 _ pm _ notes. html.

6 MeSH Browser (2011 MeSH) [EB/OL]. [2011 - 12 - 02]. http: //www. nlm. nih. gov/mesh/MBrowser. html.

7 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，(11)：53 - 58.

8 XML MeSH Conversion Table. ELHILL MeSH Unit Record - > XML Format [EB/OL]. [2011 - 11 - 20]. http: //www. nlm. nih. gov/mesh/xmlconvert. html.

9 Rare Diseases. NLM Technical Bulletin, 2009 (11/12) [EB/OL]. [2011 - 11 - 22]. http: //www. nlm. nih. gov/pubs/techbull/nd09/nd09_ mesh. html.

10 PubMed [EB/OL]. [2011 - 11 - 22]. http://www. ncbi. nlm. nih. gov/pubmed/.

11 Unified Medical Language System? [EB/OL]. [2011 - 11 - 22]. http: //www. nlm. nih. gov/research/umls/.

12 Lancaster F W. 情报检索词汇控制 [M]. 上海：同济大学出版社，1992：85 - 86.

MeSH增补概念的术语映射分析

作者： 李丹亚， 胡铁军， 李军莲
作者单位： 中国医学科学院医学信息研究所北京m0020
刊名： 医学信息学杂志
英文刊名： Journal of Medical Informatics
年，卷(期)： 2012,33(4)

参考文献(12条)

1. Nelson S J;Johnston D;Humphreys B Relationships in Medical Subject Headings 2010

2. MeSH Vocabulary Updated for 2011,NLM Technical Bulletin,2010(11/12) 2011

3. Celebrating MeSH:50 years of Medical Subject Headings 2010

4. Changes in MeSH Data Structure NLM Technical Bulletin,2000(10) 2010

5. PubMed Notes 2011 2011

6. MeSH Browser(2011, MeSH) 2011

7. 李丹亚;李军莲;胡铁军 MeSH的概念结构及其意义[期刊论文]-医学信息学杂志 2010(11)

8. XML MeSH Conversion Table ELHILL MeSH Unit Record-》XMLFormat 2011

9. Rare Diseases NLM Technical Bulletin,2009(11/12) 2011

10. PubMed 2011

11. Unified Medical Language System 2011

12. Lancaster F W 情报检索词汇控制 1992