医学领域知识组织体系评价与分析研究*

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

摘要：医学领域知识组织体系已广泛应用于各种医学信息系统中，其质量将对医疗工作及医学研究产生影响。文章采用定性及定量相结合的方法，从多角度出发，对各种医学领域知识组织体系的功能、内容和结构进行详尽的评价与分析，并对其整体质量进行归纳总结。

关键词：医学知识组织体系，叙词表，评价分析
DOI：10.3772/j.issn.1673—2286.2012.12.006

1 引言

知识组织体系是一种对内容概念及其相互关系进行描述和组织的机制，可对各信息对象按照知识内容和知识结构进行描述和组织[1]。网络环境下，知识组织体系在用户信息需求与信息资源之间起着重要的桥梁作用，是解决网络信息服务中信息有效链接的工具。至今，知识组织体系已广泛应用于图书情报、科研教育、商业门户网站、搜索引擎及学科信息门户等众多领域。

医学领域知识组织体系是对各种医学知识内容和知识结构进行描述和有组织阐述的语义工具的统称，在临床决策支持系统及各种医学数据检索及服务系统中发挥着极其重要的作用。因此，医学领域知识组织体系的质量对医疗工作及医学研究将具有直接的影响[2]，而对其质量评估及评价分析研究不仅为实际应用工作的成功开展及优化改进提供依据，亦可对今后知识组织体系的编制及更新维护工作提供借鉴。目前，医学知识组织体系的质量评估方法主要为基于词汇、结构、语义及统计的比较与对比法[3]。本文提出以规范化程度、主题领域、影响力及受控程度为主的定性评估，与以收词量、清晰度、语义关系揭示程度、入口率、类目数及生物医学文献术语命中率为主的定量统计相结合的方法，对各种国际著名医学领域知识组织体系的功能、内容和结构等方面进行系统的评价及分析研究，并对其整体质量进行归纳总结。

2 研究基础

从上世纪七十年代开始，随着计算机在图书情报领域的广泛应用以及医学领域知识组织体系编制技术的逐步完善，国内外相关研究机构制定了大量的医学领域知识组织体系，覆盖了各主题领域。本文调研了两百多个医学领域知识组织体系，对各个体系采用定性及定量相结合的方法进行了详尽的评价与分析；为了重点阐述评价方法与分析过程，本文选取了十个具有较好应用基础及影响力的医学领域知识组织体系，在文章以后的章节中，将以这些知识组织体系为代表。这些知识组织体系代表的类型及中英文名称列于表1。

3 医学领域知识组织体系的定性评价与分析

定性评价是一种对知识组织体系的功能、内容和结构等方面进行

* 本文系国家“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建设”（编号：2011BAH10B01）子任务的研究成果之一。

表1 医学领域知识组织体系示例

类型	中文名称	英文名称	名称缩写
叙词表	医学主题词表	Medical Subject Headings[4]	MeSH
	NCI叙词表	NCI Thesaurus[5]	NCIt
	心理学索引术语主题词表	Thesaurus of Psychological Index Terms[6]	PSY
分类表	国际疾病分类法，第10版，临床修订版	International Classification of Diseases, 10th Edition, Clinical Modification[7]	ICD10CM
	功能、残疾和健康国际分类法	International Classification of Functioning, Disability and Health[8]	ICF
本体	国际系统医学术语集 —— 临床术语	Systematized Nomenclature of Medicine -Clinical Terms[9]	SNOMED CT
	基因本体	Gene Ontology[10]	GO
	解剖学基础模型本体	Foundational Model of Anatomy Ontology[11]	FMA
术语表	MedlinePlus健康主题	MedlinePlus Health Topics[12]	MEDLINEPLUS
	HUGO基因命名表	HUGO Gene Nomenclature[13]	HUGO

分析、判断、归纳与总结的方法。本文通过借鉴传统知识组织体系的评价方法，并充分考虑网络环境下知识组织体系的特点，提出一套对医学领域知识组织体系的定性评价指标，即：规范化程度、主题领域、影响力及受控程度。

3.1 规范化程度

医学领域知识组织体系的规范化程度，主要通过知识组织体系的编制机构、发展历程及版本更新周期来反映。一般而言，正规的编制机构在知识组织体系制定方面经验丰富，所编制的知识组织体系规范化程度较高；知识组织体系发展历史越久远，不仅其规模会逐渐增大，而且体系中存在的各种缺陷及不足亦会逐步得到改善或解决；而知识组织体系的版本更新周期反映了主题新颖性，因为更新周期越短，编制机构便可及时增补新主题、新术语，从而及时满足信息时代的发展要求。表2列出本文所选的十个医学领域知识组织体系的规范化程度评价与分析结果；其中，MeSH的编制机构——美国国立医学图书馆，是世界范围内最大的医学图书馆，其过半个世纪的MeSH编制及维护经验，加之每年及时地更新与增补，从而也决定了MeSH的规范化程度最高。

3.2 主题领域

主题领域反映了知识组织体系覆盖学科领域的广度和深度，在一定程度上决定了知识组织体系的适用范围。本文采用《中国图书馆分类法》对医学知识组织体系进行分类，其结果列于表3；相比而言，“医学综合”领域的三个知识组织体系MeSH、ICD10CM及SNOMED CT，学科覆盖范围较广，涵盖了医学各个子学科领域，预示着这些知识组织体系将会有广泛的应用前景。

3.3 影响力

至今，医学知识组织体系已普遍应用于各种医学术语服务系统中，包括电子病例系统、书目数据库、图像数据库、事实数据库及专家系统等；此外，医学知识组织体系之间的交叉映射，及其经翻译转化后的其他语言版本，亦可反映知识组织体系的影响力。表4列出一些医学领域知识组织体系的影响力；其中，SNOMED CT不仅已成功应用于多种医学术语服务系统，而且

表3 医学领域知识组织体系的主题领域

名称缩写	主题领域
MeSH	医学综合
NCIt	肿瘤学
PSY	心理学
ICD10CM	医学综合
ICF	健康和身体缺陷
SNOMED CT	医学综合
GO	遗传学
FMA	解剖学
MEDLINEPLUS	卫生保健
HUGO	遗传学

表2 医学领域知识组织体系的规范化程度评价结果

名称缩写	编制机构	发展历程	更新周期
MeSH	美国国立医学图书馆	最早发布于1960年，至今已经过多次更新并完善	1年
NCIt	美国国立癌症研究所	1998年在NCI超级叙词表的基础上衍化而成	1月
PSY	美国心理学会	至今已收录8200多个标准并相互参照的术语	1年
ICD10CM	世界卫生组织	1998年在ICD-9-CM的基础上发展而成	1年
ICF	世界卫生组织	从2001年起，逐渐成为一种描述和度量健康及残疾的国际标准	1月
SNOMED CT	国际健康术语标准发展组织 美国病理医师学会	2002年，通过对国际系统医学术语集——参考术语集（SNOMED Reference Terminology，SNOMED RT）与英国国家健康服务部的临床术语3（Clinical Terms3, Read Codes）的合并、扩充并重组结构而形成	半年
GO	基因本体联盟	从1998年起，至今已收录近10万个基因结构化术语	1月
FMA	美国华盛顿大学结构信息研究组	从1995年至今，已收录13万的英语、拉丁语、法语、西班牙语及德语术语	1周修改
MEDLINEPLUS	美国国立医学图书馆	1998年发布第一版，至今已从最初的22个健康主题发展到900多个	1日
HUGO	HUGO基因命名委员会	从1996年至今，已批准近33000对人类基因及符号	1月

表4 医学领域知识组织体系的影响力

名称缩写	医学术语服务系统中的使用情况	交叉映射	语言版本
MeSH	1. MEDLINE、PubMed、CBM数据库标引、检索 2. 美国国立医学图书馆的馆藏编目	与UMLS语义网[14]等体系	多语种
NCIt	为众多NCI及其他系统提供癌症术语，为临床数据交换标准协会术语表、美国食品药品监督局、联邦药物疗法术语表和国家处方药项目委员会提供生物医学编码和参考标准	与GO、HUGO等5个体系存在映射	英文
PSY	用于PsycINFO等检索系统的数据主题检索	无	英文
ICD10CM	用于医院临床病案记录的疾病统一编码与命名	与ICD-9-CM[15]存在映射	多语种
ICF	用于HGNC数据库中	与SNOMED CT存在映射	多语种
SNOMED CT	1. 提供电子健康记录使用的核心常用术语集 2. 提供药品信息系统的药品概念与编码信息 3. 为英国国民健康信息基础架构提供术语标准	与ICF、ICD-9-CM等5个体系存在映射	多语种
GO	解决不同数据库中基因产品的一致描述问题	与EC[16]等存在映射	英文
FMA	1.为美国国家心理健康研究所的人类大脑项目提供原型；2. 美国国立癌症研究所人类癌症鼠模型的演示项目	无	多语种
MEDLINEPLUS	为用户卫生信息提供高水准的主题分类	无	多语种
HUGO	用于HGNC、PubMed等数据库中	与OMIM[17]等存在映射	英文

与多个医学知识组织体系之间的映射，其英文原版亦在世界众多国家广泛翻译，表明SNOMED CT是一种当前国际上具有很高影响力的临床医学术语标准。

3.4 受控程度

受控程度指医学领域知识组织体系的编制机构对其版权的保护或限制级别，从一定程度上起到约束并限制该知识组织体系使

用条件及范围的作用，是知识组织体系评价中一个极其重要的内容。例如，本文重点评价的十个医学知识组织体系中，MeSH、NCIt、MEDLINEPLUS、HUGO、GO及FMA受控程度最低，对任何组织和个人均完全开放、免费使用；PSY允许授权单位内部科研、产品开发及分析使用；ICD10CM及ICF仅供授权单位使用，且不可进行翻译及内容修改；SNOMED CT的受控程度最高，任何组织和个人对其任何方式的下载、访问或使用，均需得到其版本持有机构（国际健康术语标准发展组织SNOMED标准研发组织）的授权。

4 医学领域知识组织体系的定量评价与分析

定量评价是对知识组织体系的功能、内容和结构等方面进行一系列量化与统计的方法。相比而言，定性评价相对主观，是评价人根据自己对知识组织体系的调研、理解与判断所作出的主观性分析，其评价结果在很大程度上因人而异；而定量评价更为客观，是评价人应用科学的方法构造数学模型所计算出的数值结果，其评价结果的可信度、可靠性较高。本文对医学领域知识组织体系进行评价与分析时所采用的定量指标，除了传统的收词量、清晰度、入口率与类目数之外，还包括语义关系揭示程度与生物医学文献术语命中率。

4.1 收词量

收词量指知识组织体系包含概念及术语的数量，决定了知识组织体系的规模[18]。收词量与主题领域相结合，可反映知识组织体系的主题覆盖度与专指度，是衡量知识组织体系主题完备性与专指性的重要指标。表5列出一些医学领域知识组织体系的收词量；其中，SNOMED CT规模最大，收录了324494个概念，共计1181154个术语，是一部主题完备的医学综合领域的知识组织体系。

表5 医学领域知识组织体系的收词量

名称缩写	概念数	术语数
MeSH	321367	758306
NCIt	90135	238385
PSY	6741	7961
ICD10CM	98178	102764
ICF	1435	1521
SNOMED CT	324494	1181154
GO	58270	104241
FMA	82079	139095
MEDLINEPLUS	1858	2909
HUGO	31206	136029

4.2 清晰度

清晰度形象地表达了知识组织体系中的术语能被用户理解并正确使用的可能性，亦即知识组织体系对术语注释的详尽程度[18]。一般而言，清晰度越高，知识组织体系的注释信息越丰富，越有助于使用者明确术语的含义及术语间的关系，以及掌握术语的使用方法。清晰度的计算公式为：

$$\text{清晰度} = \frac{\text{含有释义的术语总数}}{\text{术语总数}} \quad [1]$$

显然，清晰度越接近1，知识组织体系的术语表示越清晰。如表6所示，在本文所选取的十个医学领域知识组织体系中，清晰度最高的是ICF，从具体的数值可反映出该体系中超过一半的术语具有注释信息；而清晰度最低的是ICD10CM、HUGO，表明这些体系中的术语无注释信息，从而将会对使用者从释义角度理解这些体系中的术语或概念造成障碍。

表6 医学领域知识组织体系的清晰度

名称缩写	清晰度
MeSH	0.036
NCIt	0.275
PSY	0.278
ICD10CM	0
ICF	0.504
SNOMED CT	0.0006
GO	0.328
FMA	0.008
MEDLINEPLUS	0.303
HUGO	0

4.3 语义关系揭示程度

知识组织体系的语义关系揭示程度主要通过概念之间、术语之间关系类型与关系数量来衡量，关系类型越丰富，揭示深度越精细；关系数量越多，揭示程度越高。传统知识组织体系中的语义关系，通常分为三种类型：等同、等级和相关关系[1]。本文在这三种语义关系类型的基础上，将医学领域知识组织体系中概念之间、术语之间的关系细化为直接上下位、广义与窄义、限定与被限

表7 医学领域知识组织体系的语义关系揭示程度

名称缩写	语义关系类型	语义关系数量
MeSH	直接上下位、广义与窄义、限定与被限定、同位、同义、其他	1392472
NCIt	直接上下位、广义与窄义、同义、其他	406245
PSY	直接上下位、广义与窄义、相关或可能同义、其他	27149
ICD10CM	直接上下位、同位、相关或可能同义	325125
ICF	广义与窄义、同位、同义、	5527
SNOMED CT	直接上下位、广义与窄义、同义、相关或可能同义、其他	1522813
GO	直接上下位、广义与窄义、同位、同义、相关或可能同义、其他	838091
FMA	直接上下位、同位、其他	552499
MEDLINEPLUS	直接上下位、同位、同义、相关或可能同义、其他	46268
HUGO	同义	67610

定、同位、同义、相关或可能同义及其他关系。各个知识组织体系中的语义关系类型及数量列于表7；其中，MeSH与SNOMED CT的语义关系揭示程度最高，因为这两个体系中语义关系的类型不仅丰富，而且语义关系的数量也很多。

4.4 入口率

入口率也称等同率，指叙词表中入口词与叙词的比率[19]，是衡量叙词表中入口词的丰富程度的重要指标，其计算公式为：

$$\text{入口率} = \frac{\text{入口词}}{\text{叙词}} \qquad [2]$$

从术语控制、检索语言自然语言化的角度考虑，公式2所计算出的入口率越高越好。在本文所选择的医学领域知识组织体系中，MeSH、NCIt与PSY均为叙词表，其入口率分别为2.30、1.83、0.46；显然，因MeSH的入口率最高，所以最适合用于生物医学文献检索系统中。

4.5 类目数

类目是分类法的基本单元，类目数是衡量分类表完备性和专指度的一个重要指标[20]，一般要求类目划分详细，展开充分，数量众多，从而达到分类组织的目的。例如，在ICD10CM分类表中，共有22080个类目（占总术语数的21.5%），这些类目将整个体系划分为21类，且纵向最深展开到7级，反映出该体系具有较好的分类组织结构；相比而言，ICF的类目数量较少（共265个，占总术语数的17.4%），整个体系只有4类，纵向最深展开到6级，表明该体系的类目划分较粗，不利于实现分类组织的目的。

4.6 生物医学文献术语命中率

生物医学文献术语命中率指生物医学文献中所使用的术语，与医学领域知识组织体系所收录术语的符合度，可用于衡量医学知识组织体系在文献中的适用度及实际使用度（对某些面向健康数据的医学知识组织体系而言，如RxNorm[21]，因其术语含有明确的剂型与剂量，所以这些体系中的术语很少出现在文献中）。本文收集了2007至2010年国家科技图书文献中心所收录的全部英文生物医学文献，并从中提取出514737个生物医学术语；通过字符串精确匹配，发现其中的512886个术语出现在已调研的医学知识组织体系中，命中率达到99.64%。表8列出这些文献中出现的生物医学术语与部分知识组织体系的命中数；由于知识组织体系的总收词量在一定程度上将影响文献术语命中数，对收词量较大的知识组织体系而言，其文献术语命中数相对较多；因此，本文提出将知识组织体系的文献术语命中数与其总收词量的比值，作为文献术语命中率，以比较各知识组织体系的绝对文献术语命中程度。从表8可知，MeSH与生物医学文献术语的命中数最多，近三分之一的文献术语来自MeSH，这与MeSH的大规模收词量有关；但从绝对的文献术语命中率来比较，

表8 医学领域知识组织体系的文献术语命中率

名称缩写	命中数	命中率
MeSH	171891	0.23
NCIt	65864	0.28
PSY	5016	0.63
ICD10CM	3976	0.04
ICF	386	0.25
SNOMED CT	141387	0.12
GO	16680	0.16
FMA	11623	0.08
MEDLINEPLUS	2040	0.70
HUGO	34366	0.25

MEDLINEPLUS更适合应用于文献处理，因为该体系所收录的70%的术语出现在生物医学文献中；而ICD10CM的文献术语命中率最差，只有4%的术语出现在生物医学文献，根本原因在于该体系的主要应用是为临床疾病的病案分类提供统一的编码与命名标准。

5 结语

本文提出了一种定性与定量相结合的方法，对各种类型的医学领域知识组织体系进行了详细的评价与分析；而通过各个角度的评价与分析结果，可总结归纳出各体系的整体质量。例如，医学综合领域的叙词表MeSH，不仅规范化程度、语义关系揭示程度、入口率及生物医学文献术语命中率高，而且对任何组织和个人均完全开放、免费使用，因此MeSH是目前最权威、最常用的标准医学主题词表；相比而言，SNOMED CT虽然也是一部非常重要的医学综合领域术语表，规模庞大，规范化程度及语义关系揭示程度较高，但其使用受限，从而在一定程度上限制了其在各类医学系统中的自由应用。此外，在实际应用医学领域知识组织体系时，应结合具体使用场景而充分考虑其编制目的。例如，医学综合领域的分类表ICD10CM，因其编制目标是为临床疾病的病案分类提供统一编码与命名标准，所以该体系更适合应用于医院临床疾病分类，而非提供生物医学文献检索的服务系统。综上所述，医学领域知识组织体系的评价与分析是一项极其复杂的工作，除了本文所提出的各种定性及定量的评价指标之外，实际应用需求亦为重要的参考因素；而这种综合的评价结果不仅为医学领域知识组织体系在文献及相关医学系统中的成功应用及优化改进提供了依据，亦为日后知识组织体系的设计、编制及维护更新提供了借鉴与启示。

参考文献

[1] 李育婷.网络数字环境下知识组织体系的发展现状及未来趋势[J].情报资料工作,2009(2):45-8.
[2] BODENREIDER O. Quality Assurance in Biomedical Terminologies and Ontologies[J]. A Report to the Board of Scientific Counselors, 2010.
[3] ZHU X, FAN J W, BAORTO D M, et al. A Review of Auditing Methods Applied to the Content of Controlled Biomedical Terminologies[J].J Biomed Inform, 2009,42(3):413-25.
[4] MeSH Browser [OL].[2011-08-28]. http://www.nlm.nih.gov/mesh/MBrowser.html/.
[5] NCI Thesaurus [OL].[2012-05-18]. http://ncit.nci.nih.gov/.
[6] Thesaurus of Psychological Index Terms [OL].[2012-05-18].http://www.apa.org/pubs/databases/training/thesaurus.aspx/.
[7] ICD-10-CM[OL].[2011-12-19].http://www.cdc.gov/nchs/icd/icd10cm.htm.
[8] ICF[OL].[2011-12-19].http://www.who.int/classifications/icf/en/.
[9] SNOMED Clinical Terms[OL].[2012-05-15].http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html/.
[10] Gene Ontology[OL]. [2012-05-18]. http://www.geneontology.org/.
[11] FMA[OL]. [2012-05-18].http://sig.biostr.washington.edu/projects/fm/index.html.
[12] MedlinePlus[OL].[2010-04-27].http://www.nlm.nih.gov/medlineplus/healthtopics.html/.
[13] HUGO Gene Nomenclature[OL]. [2012-05-18]. http://www.genenames.org/.
[14] The UMLS Semantic Network[OL]. [2011-02-11]. http://semanticnetwork.nlm.nih.gov/.
[15] ICD-9-CM[OL].[2011-09-23].http://www.cdc.gov/nchs/icd/icd9cm.htm/.
[16] Enzyme Commission (EC) Numbers[OL]. [2012-05-18].http://bitesizebio.com/articles/enzyme-commission-ec-numbers/.
[17] OMIM[OL]. [2012-05-18]. http://www.ncbi.nlm.nih.gov/omim/.
[18] 薛春香,侯汉清.网络环境中知识组织体系构建与应用研究[M].南京:东南大学出版社,2009:151-80.
[19] 吴雯娜,曾建勋.叙词表微观结构的描述与评价——EI叙词表与中文叙词表的对比分析[J].图书情报工作,2009,53(8):12-6.
[20] 尚加宁,韩露盈.图书分类法性能的定量评价测评初探[J].情报理论与实践,1999,22(5):356-8.
[21] RxNorm[OL]. [2011-12-05]. http://www.nlm.nih.gov/research/umls/rxnorm/.

作者简介

李晓瑛（1982-），博士，研究方向：医学知识组织体系建设与自然语言处理。E-mail: lixiaoying@imicams.ac.cn

Research on Assessment and Analysis of Medical Knowledge Organization System

Li Xiaoying, Li Danya, Li Junlian, Hou Li, Hu Tiejun / Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, 100020

Abstract: Medical knowledge organization system has been widely used in the medical information system, and its quality may have a direct impact on healthcare and biomedical research. This paper applies a qualitative and quantitative method to deeply assess and analyze the function, content and structure of all kinds of medical knowledge organization system, and also summarize their quality in general.

Keywords: Medical knowledge organization system, Thesaurus, Assessment and analysis

（收稿日期：2012-08-01）