

医学知识组织系统构建研究与应用实践^{*}

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

摘要: 医学知识组织系统是对医学领域重要概念及其语义关系的规范化描述, 有助于实现海量医学知识的深度组织与有效利用, 对推动医疗健康大数据的语义互联、共享共用及医药科技创新发展具有重要意义。本文全面阐述代表性中文医学知识组织系统的内容结构、规模体量及发展历程等基本情况, 对其在术语服务、文献主题标引与主题检索、临床医学知识库构建等主要应用实践进行概要介绍, 并对今后的发展趋势和研究重点进行总结和展望。

关键词: 知识组织; 叙词表; 本体; 术语服务; 主题标引

中图分类号: G250 **DOI:** 10.3772/j.issn.1673-2286.2020.07.004

引用格式: 李晓瑛, 李军莲, 邓盼盼, 等. 医学知识组织系统构建研究与应用实践[J]. 数字图书馆论坛, 2020(7): 30-35.

1 医学知识组织系统概述

在图书情报领域, 一般认为知识组织(Knowledge Organization, KO)是以知识为对象的整理、加工、表示、控制等一系列组织化过程及其方法^[1], 目的是使知识有序化。医学知识包括医学科技文献、健康知识、电子病历、健康档案、心跳及血压类生命体征数据、医学影像、解剖部位、基因及核酸序列、药物信息、公共卫生信息等多种类型, 规模十分庞大。通过对无序分散的医学知识按照其内在逻辑联系, 运用一定的组织方法、工具和标准开展有序化、系统化活动, 形成如分类法、叙词表、术语表、一体化语言系统、本体、语义网络、知识图谱等不同结构的医学知识组织系统(Medical Knowledge Organization System, MKOS), 有助于实现海量医学知识的深度组织与有效利用, 对推动医疗健康大数据的语义互联、共享共用及我国医药科技创新发展具有重要意义。

国际上, 世界卫生组织(World Health Organization, WHO)、美国国立医学图书馆(U.S. National Library of Medicine, NLM)、美国国立癌症研究所(U.S. National

Cancer Institute, NCI)、国际健康术语标准化与研发组织(International Health Terminology Standards Development Organization, IHTSDO)等相关机构积极从事医学知识组织系统的研究、制定、更新维护和发布服务。如WHO编制的系列《国际疾病分类法》(International Classification of Diseases, ICD), 依据人类疾病的病因、部位、病理、临床表现这4个特征对临床疾病进行分类组织, 成为国际普遍采用的卫生信息标准; NLM早在1956年就开始构建《医学主题词表》(Medical Subject Headings, MeSH), 现已作为全球最有影响力的医学综合性叙词表和通用生物医学信息组织标准, 广泛用于医学资源的主题标引与检索; NCI开发的《NCI叙词表》(NCI Thesaurus, NCI Thes), 为肿瘤相关的医学资源分类与组织、精准医学研究的开展提供了重要的肿瘤专题术语服务; IHTSDO负责维护的《医学系统化术语表-临床术语》(Systematized Nomenclature of Medicine-Clinical Terms, SNOMED CT), 涵盖疾病、临床发现、操作、微生物、药物等绝大多数临床信息, 成为当今世界上公认的最庞大的临床医学学术语集。上述这些医学知识组织系统, 对世界

^{*}本研究得到NSTL先期研发任务“STKOS超级科技词表内容建设机制和发展研究(医学部分)”(编号: XQYF0101-4)资助。

其他国家本土化术语标准的构建与应用具有一定的典范作用和借鉴意义。

国内相关机构在医学知识组织系统建设方面作出了有力的尝试和突出的贡献。半个世纪以来,中国医学科学院医学信息研究所致力于中文医学知识组织系统建设与应用研究,至今已建成《中文医学主题词表》

(*Chinese Medical Subject Headings, CMeSH*)、《中文一体化医学语言系统》(*Chinese Unified Medical Language System, CUMLS*)、《医学主题分类一体化系统》等一系列有影响力的医学学术标准;“十二五”期间,在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建设”的支持下,构建了具有我国自主知识产权的、有效服务于外文科技文献组织的《科技知识组织体系》(*Scientific & Technical Knowledge Organization System, STKOS*) (医学部分),为我国海量外文科技文献信息的组织和利用提供支撑。本文首先全面阐述若干代表性医学知识组织系统的基本情况、内容结构、规模体量及发展历程,而后概要介绍其在医学术语服务系统、生物医学文献主题标引与检索和临床医学知识库中的应用实践,最后对今后的发展趋势和工作重点进行总结和展望。

2 医学知识组织系统构建研究

2.1 中文医学主题词表

*CMeSH*是目前我国医学领域规模最大的主题词表。1970年前后,面向我国医学文献主题标引和检索的需要,着手构建中文标准化医学学术语;1978年获得NLMMESH词表翻译授权后,开始翻译《MeSH-注释字顺表》,至1996年初步建成*CMeSH*; *CMeSH*集成了《MeSH·中文版》和《中医药学主题词表》,并且建立了主题词表与《中国图书馆分类法·医学专业分类表》的双向关联映射。1999年,研发出版Windows光盘版,2012年起提供*CMeSH Browser*在线服务,与MeSH Browser具有良好的兼容性。

就内容结构而言, *CMeSH*由主题词、副主题词及增补概念3个部分组成(见图1),各部分均为叙词-概念-术语3级结构^[2]。其中,2018版主题词为34 822个,按学科属性从多个角度进行分类组织,形成一套包含30个大类、层级最深达13级的树状等级结构;2018版副主题词有90个,主要用于对主题词进行限定和组配,

从而提高主题词的专指度;增补概念是对主题词的补充和扩展,规模超过10万个,大多收录自文献中的化学物质和药物、治疗方案、罕见疾病及有机体的名称,并且与主题词建立了术语映射关系^[3]。*CMeSH*每年进行版本更新,现已被广泛应用于中文医学信息资源的自动标注、智能检索、中英文双语检索、学科分类导航,以及搜索引擎、数据挖掘、热点监测、趋势预测等领域。

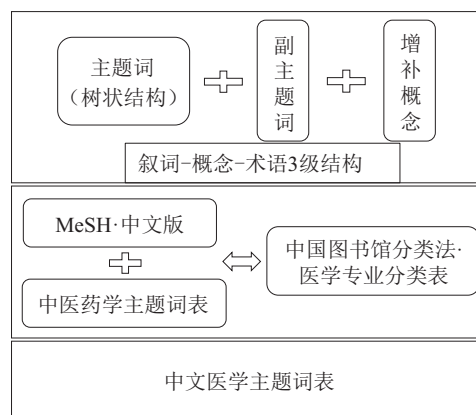


图1 *CMeSH*内容和结构

2.2 中文一体化医学语言系统

*CUMLS*是一个有效服务于医学信息资源组织的、面向计算机应用的知识组织系统。*CUMLS*秉承语义网络和本体的建设理念,通过自然语言-主题语言-分类语言的一体化建设,以及搭建多种知识组织系统间的关联网络,力图构建一个多元化的知识组织系统^[4]。医学文献信息数据量急剧增长,但文献加工能力明显不足,为解决此矛盾,1996年构建了面向计算机应用的知识组织工具“医学关键词-主题词-分类号自动转换系统”;为适应网络环境下复杂医学信息资源组织和利用的需要,2000年正式提出构建*CUMLS*。经过长期努力和建设,目前*CUMLS*已形成一定规模(包含生物医学概念27万个),并成为医学信息领域的一项重要基础设施。

*CUMLS*主要由医学词表系统、语义网络、辅助构建平台共3个部分构成(见图2)。其中,医学词表系统集成20余部医学领域的叙词表、分类表、术语表等来源词表以及医学文献的关键词,以概念为中心、以继承来源词表的语义关系为基础,形成概念集合和知识系统;与NLM研发的一体化医学语言系统(*Unified Medical Language System, UMLS*)相比, *CUMLS*具有概念与分类名、文献关键词相互关联的特点与优势,

从而形成自然语言-主题语言-分类语言相互关联的一体化语言系统。语义网络包含语义类型(主题类目)和语义关系两个部分,为医学词表系统提供了统一的组织框架和顶层类目体系,同时还可深度揭示概念间的语义关系。辅助构建平台涵盖一系列词表构建工具(如同义词识别、语义相似度计算、主题分类自动映射、主题词/副主题词自动组配等),以及支持词表管理与审核的在线加工平台。*CUMLS*有助于提供医学文本主题和分类自动标注、知识导航、智能检索、文献相关性检索等知识服务,对网络环境下医学信息资源的知识组织与知识服务具有良好的实用价值。

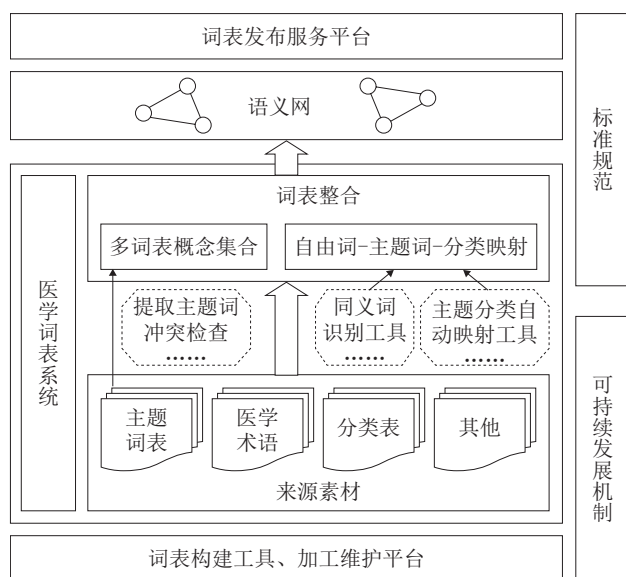


图2 CUMLS基本组成框架

2.3 医学主题分类一体化系统

图书情报领域的主题词表和分类法是两种不同的信息检索语言。主题词表通过主题词来揭示文献,具有较强的专指性;而分类法通过学科知识体系对文献进行划分和组织,体现了学科的系统性。两者各有优缺点,既有区别,又有联系,且互为补充。主题分类一体化系统将主题词表与分类法结合为一体,通过建立主题词表的主题词与分类法的分类名的映射关联,实现主题词表与分类法的相互兼容、相互转换,进而实现标引和检索的一体化。

《医学主题分类一体化系统》通过建立*CMeSH*的主题词与《中国图书馆分类法·医学专业分类表》的分类名的双向对应和映射关联^[5],实现一表两用,可方便标

引人员进行主题标引和分类标引,并易于用户实现族性检索和特性检索。此外,《医学主题分类一体化系统》充分参考了MARC21元数据规范等新技术标准,很大程度上弥补了传统医学知识组织系统无法及时揭示医学新学科、新主题的不足,基本符合医学信息资源组织处理的现实需要。因此,《医学主题分类一体化系统》对我国医学文献工作者具有一定的参考和实用价值。

2.4 科技知识组织体系(医学部分)

*STKOS*是一部面向外文科技文献信息的超级科技词表,通过基于多来源的术语同义汇聚及基于范畴的概念组织,覆盖理工农医4个领域的核心概念。*STKOS*词表建设采取计算机自动处理与人工审核相结合的方式(见图3),构建以科技术语为基本单元、以概念为核心、以继承来源词表的已有关系为基础,通过概念与多来源术语的语义关系进行关联,形成概念统领同义术语、范畴类目统领概念的科技知识组织系统。在国家科技图书文献中心组织的“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”支持下,*STKOS*共建成规范概念61.5万个(含术语232.1万个),其中医学概念26万个、术语100万个;“十三五”期间,*STKOS*医学概念进一步扩充、更新与完善,确保*STKOS*医学领域知识覆盖的完整性,提升语义标注、智能检索、数据挖掘和知识发现等知识服务能力。

2.5 精准医学本体

近年来,随着现代生物医药技术的快速发展,精准医学相关研究受到全世界的广泛关注,共同目标为基于患者遗传图谱中个体变异,实现对疾病最佳的诊断、治疗和管理。其中,对精准医学相关大数据和知识的组织与管理是一项重要工作,成为开展对深层次分子生物过程的认识和理解、RNA靶向药物研发、肿瘤早期干预、分子靶向治疗等研究的支撑条件和基础。因此,在国家“十三五”重点研发计划项目“疾病研究精准医学知识库构建”的支持下,“精准医学本体和语义网络构建”课题提出一种基于本体的精准医学知识组织和管理框架。

《精准医学本体》(*Precision Medicine Ontology, PMO*)从患者表型、疾病、化学物质和药物、细胞机制、分子机制以及遗传机制共6个维度对相关概念进行梳

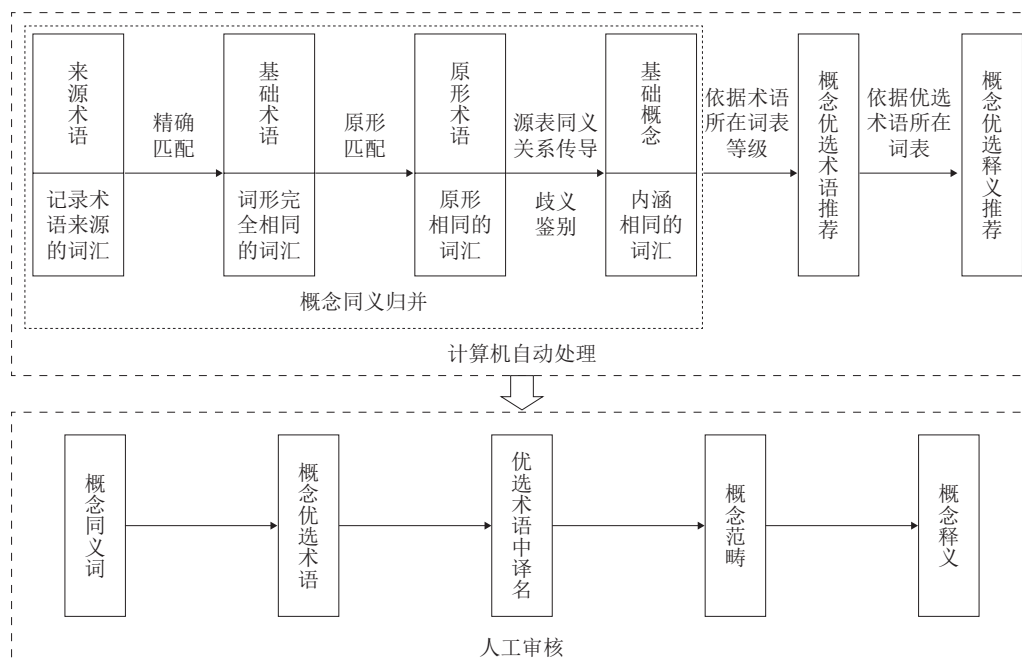


图3 STKOS概念建设流程

理,涵盖临床医学、基础医学、表型、药物、基因、分子、小RNA、基因突变等多个主题领域;在调研、分析、遴选与精准医学密切相关的20余部医学词表的基础上,核定本体概念类6.7万个(见图4),并将其自顶向下

按树状结构进行组织,层级最深达13级。此外,通过挖掘和界定本体概念间的语义关系,最终确立了53种关系类型,并在等级上进一步细化组织,以支持对精确医学知识的深度揭示。

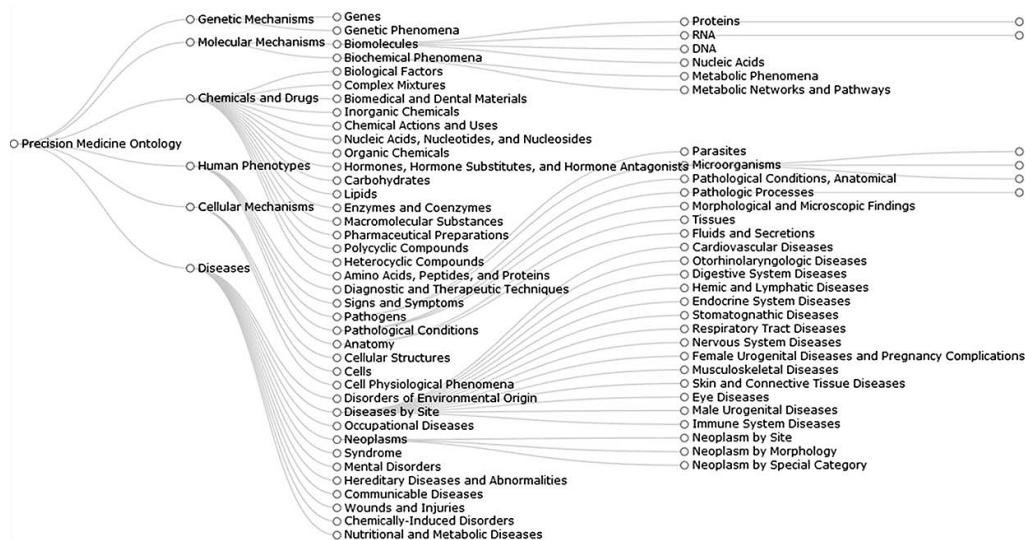


图4 精准医学本体顶层结构

2.6 中文临床医学术语系统

中国医学科学院医学与健康科技创新工程“中文临床医学术语系统构建研究”项目旨在面向我国电子病历、个人健康档案、国家医疗卫生服务、临床研究等应

用需求,开展中文临床医学相关的专业术语体系建设。通过加强临床医学术语的标准化建设,构建具有中国特色和自主知识产权的《中文临床医学术语系统》^[6],提升临床医学信息的描述、组织、整合、共享和利用等工作质量和效率。该系统不仅对实现临床电子病历信息

标准化及卫生管理、医保支付等医院信息化系统之间的数据互操作具有重要意义,而且有助于推动临床信息标准化建设、支持临床决策和临床医学科技创新发展。目前,该系统初步构建了14个顶层概念大类和65种语义关系,涉及规范概念25万个、术语85万个,并将陆续完善与发布。此外,考虑到我国疾病谱(特别是高发传染病和癌症方面)与欧美等发达国家有所差异,《中文临床医学术语系统》与国际普遍采用的临床医学术语SNOMED CT,在临床用药品器械、卫生管理、医保支付、人口统计学等主题上亦别有中国特色。

2.7 术语加工平台

立足于医学知识组织系统的长期构建实践,研发出稳定高效的计算机辅助词表加工平台及术语同义归并、新词发现推荐、术语歧义鉴别等系列工具,支持从零构建、词表复用等编表场景;提供单表及整合表的逐条加工与批式审核等多种模式,提升了词表构建效率,推动了计算机先进技术在传统以人为主的词表构建中的应用与发展。此外,积累了一系列词表构建的关键技术与算法,包括以概念为中心的多来源术语同义归并与整合、概念分类体系质量自控、叙词表到本体转化、文本概念与医学实体的自动识别、文献主题标引等,有助于推动医学知识组织体系的自动构建及其在生物医学文献中的应用研究。

综上,依托术语加工平台有效地构建科学、规范的医学知识组织系统,在一定范围内实现医学用语的统一(如PMO对相关概念的规范描述与组织),有助于促进医学信息资源的深度组织和有效利用(如CMeSH、CUMLS及《医学主题分类一体化系统》对中文生物医学文献的主题标引和组织检索、STKOS对外文科技文献的自动标注与智能检索),实现医疗健康信息和数据的共享互通(如《中文临床医学术语系统》的建设目标),在一定程度上推动了我国医疗卫生信息化和医学科技创新。

3 医学知识组织系统应用实践

3.1 医学术语服务系统

医学术语服务系统(Chinese Medical Terminology Services, CMTS)旨在构建中国最专业、最权威、规模

最大的生物医学领域术语服务平台。CMTS服务的术语资源包括CMeSH(见2.1节)、CUMLS(见2.2节)及《医学主题分类一体化系统》(见2.3节)、《医学专业分类表》、《中文临床医学术语系统》(见2.6节)等不同类型的医学知识组织系统,服务内容涉及概念中英文名称、同义术语、等级结构、范畴分类、语义关系、释义及重要属性等术语信息,服务方式包括智能检索、等级浏览、分类导航、可视化展示、数据定制、格式下载、API接口访问等多种模式。此外,CMTS提供医学知识组织系统内容构建与相关构建工具服务。

3.2 医学文献主题标引与主题检索

文献主题标引的目标是给出揭示文献主题内容的语词,主题检索则为以表达文献主题内容的语词来检索文献的方式。目前,中文生物医学文献普遍采用CMeSH词表(见2.1节)的主题词进行文献的主题标引与主题检索。例如,中国生物医学文献服务系统SinoMed在提供智能检索之外,支持对文献按照CMeSH主题词及其对应的分类名进行主题检索(见2.1节)与分类检索(见2.3节)^[7],面向全国提供服务,周访问量在30万次以上;万方医学网在其高级检索功能中,亦支持导航浏览CMeSH等级结构体系(见2.1节)的方式选取主题词,进而查找医学文献;国家科技图书文献中心NSTL的外文科技文献资源则基于STKOS超级科技词表(见2.4节)开展主题(概念)标引,NSTL网络服务系统则提供STKOS词表的范畴体系导航及关联的主题(概念)检索。相对而言,借助于CMeSH主题词和STKOS概念名称的规范性,上述这些文献服务系统的检索结果更加全面、准确。

3.3 临床医学知识库

临床医学知识库(Clinical Medicine Knowledge Base, CMKB)是一站式的临床知识在线服务系统,涵盖疾病库、药物库及检查库3个知识库。CMKB依据CMeSH(见2.1节)、CUMLS(见2.2节)等医学知识组织系统中的概念及其语义关系,对疾病、药物、检查项、症状等临床医学知识进行语义化描述,并构建了这些知识点之间的语义关联,易于快速了解疾病诊断、药物治疗、实验室检查等方面的系统性知识,以及实时获取与疾病的病因、诊断、预防、治疗等相关的中外文文献。

4 总结与展望

CMeSH、CUMLS、《医学主题分类一体化系统》、STKOS（医学部分）等中文医学知识组织系统是中文信息组织与知识服务领域的基础性设施，促进了术语服务、文献主题标引与主题检索、医学知识库等应用实践，推动了大数据时代下中文医学资源的深度组织和有效利用，以及医药卫生科技创新发展。今后，中文医学知识组织系统应注重以下方面的建设和发展。

（1）充分借鉴人工智能技术，加快中文MKOS的动态化、持续性发展。伴随着各种医学资源的爆炸式增长，新术语、新概念不断涌现，传统以人工为主的词表更新方式远不能适应新时代的要求。人工智能、深度学习等先进技术与方法已成功应用于自然语言处理领域，并将提高新词发现、质量控制等词表构建效果，极大缩短词表更新与完善周期，促进中文MKOS动态化、持续性发展。

（2）提倡本体、语义网络、知识图谱等语义丰富的知识组织系统构建，实现中文MKOS从机器可读到机器可理解，提升中文医学知识服务能力。知识网络环境对知识组织系统建设提出了新的要求，知识表达、知识组织已从文献单元上升到细粒度的知识单元（如纳米初版），主题法、分类法等简单处理方式转变为语义

处理^[8]；为此，构建本体、语义网络、知识图谱等语义丰富的中文MKOS，将有助于实现中文医学资源从机器可读到机器可理解，最终提供精准的中文医学知识服务。

参考文献

- [1] 蒋永福. 论知识组织[J]. 图书情报工作, 2000 (6): 5-10.
- [2] 李晓瑛, 李军莲, 冀玉静, 等. 基于叙词表及其语义关系的本体构建研究[J]. 情报科学, 2018, 36 (11): 83-87.
- [3] 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.
- [4] 李丹亚, 胡铁军, 李军莲, 等. 中文一体化医学语言系统的构建与应用[J]. 情报杂志, 2011, 30 (2): 147-151.
- [5] 李军莲, 李丹亚, 阮学平, 等. 医学分类主题一体化系统建设[J]. 医学信息学杂志, 2011, 32 (3): 69-73.
- [6] 冀玉静, 孙海霞, 邓盼盼, 等. 中文临床医学术语系统分类框架设计原则与构建思路[J]. 中华医学图书情报杂志, 2018, 27 (11): 27-32.
- [7] 李军莲, 夏光辉, 孙海霞, 等. 中国生物医学文献服务系统(SinoMed)的研制思路及功能概述[J]. 医学信息学杂志, 2011, 32 (12): 36-39.
- [8] 毕强. 数字图书馆知识组织系统建构的发展趋势[J]. 国家图书馆学刊, 2010, 71 (1): 12-17.

作者简介

李晓瑛, 女, 1982年生, 博士, 副研究员, 研究方向: 医学知识组织。

李军莲, 女, 1972年生, 硕士, 研究馆员, 通信作者, 研究方向: 医学知识组织与信息处理, E-mail: lijunlian@imicams.ac.cn。

邓盼盼, 女, 1987年生, 硕士, 助理研究员, 研究方向: 医学知识组织。

冀玉静, 女, 1977年生, 副研究馆员, 研究方向: 医学知识组织与主题标引。

夏光辉, 男, 1983年生, 硕士, 助理研究员, 研究方向: 医学知识组织与信息处理。

李丹亚, 女, 1954年生, 研究员, 研究方向: 医学知识组织。

胡铁军, 男, 1949年生, 研究员, 研究方向: 数字资源建设。

Research on Medical Knowledge Organization System Construction and Its Applied Practice

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

Abstract: Medical knowledge organization system is the standardization of biomedical concepts and the semantic relationships among them, which will contribute to the deep organization and effective utilization of health big data. In this paper, the well-known and representative Chinese medical knowledge organization systems, as well as their significant applications and practices are systematically introduced, followed by the development trends and research focuses in the future.

Keywords: Knowledge Organization; Thesaurus; Ontology; Terminology Services; Subject Indexing

(收稿日期: 2020-04-10)