• 医学信息研究 •

学术论文作者机构规范文档构建*

孙海霞 李军莲

（1 南京大学信息管理学院 南京 210046
2 中国医学科学院医学信息研究所 北京 100020）

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 以中国生物医学文献数据库为基础，面向基于学术论文开展机构检索、分析与评价相关知识服务需要，对学术论文作者机构名称规范目标与内容、体系结构与组织方式以及构建过程与实现策略进行研究、实践总结。

〔关键词〕 中国生物医学文献数据库；机构名称规范；规范文档结构；社会化协作；计算机辅助环境
〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2015.11.010

Construction of Authority Files of Affiliations of Academic Paper Authors *SUN Hai－xia, 1 School of Information Management, Nanjing University, Nanjing 210093, China, 2 Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China; LI Jun－lian, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 Based on Chinese Biomedical Literature Database (CMB), as required by providing relevant knowledge services of affiliations retrieving, analyzing and evaluating based on the academic papers, the paper studies and makes a practical summary on the goal and content, the system architecture and organization form, the construction process and implementation strategy of name standardization of affiliations of academic paper authors.

〔**Keywords**〕 China Biomedical Literature Database (CBM); Affiliations name authority, Authority file structure; Social collaboration; Computer aided environment

1 引言

随着国家在科研领域资源投入的持续增加，各类学术成果的产出量逐年上升，以科研机构为中心的各种知识服务理论研究与实践工作越来越受到图书情报领域的重视。学术论文作为核心知识载体之一，已成为开展知识组织、知识检索、科学计量分析、关联挖掘、学科发展、最新科研动向、科研评价等知识服务研究和实践活动的主要依据[1－3]。“作者机构”作为学术论文的重要标目项，是开展相关知识服务活动时科研机构与论文衔接的纽带。但由于机构自身因为更名、合并、拆分等带来的同一实体机构名称的多样性和复杂性，不同作者发文时常常对同一机构名使用不同的表达形式，甚至同一作者在不同时间也会如此等客观现实的存在，使得目前各类数据库很难保证作者机构检索点的查准率和查全率[4－5]，尤其在当前学术论文快速增长的

〔投稿日期〕 2015－06－30
〔作者简介〕 孙海霞，助理研究员，发表论文 20 余篇；通讯作者：李军莲，副研究馆员。
〔基金项目〕 中国医学科学院医学信息研究所基本科研业务专项“中国生物医学文献服务系统发展关键问题研究”（项目编号：13R0103）。

背景下，在很大程度上影响和制约着各项知识服务研究和实践活动的开展。对此，开展论文作者机构规范控制研究，构建作者机构规范文档，实现统一机构不同著录形式的汇聚，揭示不同机构名称之间的变更、隶属等语义关系，用于学术论文的组织，是提高数据库作者机构检索点的文献查准率和查全率、最大程度解除基于学术论文开展以科研机构为中心的各种知识服务理论研究与实践制约因素的重要手段之一[1]。

规范文档（Authority File）的概念在文献编目领域中由来已久，是指由规范记录组成的计算机文档。长久以来，规范文档建设相关理论研究与实践主要围绕知识内部特征（知识内容）进行，如各种主题词表、一体化语言系统等[6-8]；作者、机构等知识外部特征项的规范研究与实践则主要集中在书目规范控制方面，致力于图书编目与检索的一致性以及不同书目系统之间的互操作。如国际图书馆协会和机构联合会（International Federation of Library Associations and Institutions，IFLA）从服务角度对规范项提出了要求[9]，德国国家图书馆、美国国会图书馆和 OCLC 启动了虚拟国际规范文档（Virtual International Anthority File，VIAF）项目[10-11]，国家图书馆和 CALIS 制定了作者、团队作者等著录规则，构建了系列名称规范库[12-13]，并进行语义表达与关联研究等[14-15]。面向学术论文的机构名称规范控制研究还比较少，唐金玲从检索角度对当前 3 大数据库中论文作者机构名称问题进行了分析与总结[4]，曾建勋等从知识评价角度提出了学术论文机构著录要求[1]，董琳从学科评价角度提出了机构名称清洗需求与策略[3]，吴英杰等进行了学术论文数据库作者机构名称非规范著录形式自动检测研究[5]，高星等人进行了论文机构规范名和别名对应关系自动发现技术研究[16]，杨奕红等进行了多层级机构表编制与应用实践[17]，总体看还处于起步与探索阶段。

本文以中国生物医学文献数据库（China Biomedical Literature Database，CBM）为基础，在借鉴现有书目规范控制、各类知识组织系统构建与整合理论和实践基础上，面向基于学术论文开展机构检索、分析与评价及相关知识服务需要，对学术论文作者机构名称规范目标与内容、体系结构与组织方式以及构建过程与实现策略进行研究、实践与总结。

2 CBM 作者机构名称规范目标与内容

2.1 作者机构规范文档知识服务目标

CBM 是国内生物医学领域最早、最权威的期刊论文数据库之一，是一个集题录检索、引文检索和学术分析于一体的知识服务型数据库。学术分析包括引文分析、作者分析、机构分析、基金分析等。CBM 作者机构规范文档知识服务目标包括两个方面：一是提升 CBM 自身的知识组织、机构检索、分析与评价等服务能力；二是为实现与其他服务系统之间资源和服务整合提供支撑。具体通过 3 个阶段来逐步实现，见表 1。

表 1 CBM 作者机构规范文档知识服务目标

知识服务目标		知识服务实例
第 1 阶段	知识组织	机构库建设、机构数字图书馆建设、机构竞争情报系统建设
	知识检索	机构知识导航（类型、学科、地区等角度）、机构检索、机构链接、作者消歧检索
第 2 阶段	知识分析与评价	机构影响力、机构学科分布、机构科研合作、机构优势学科、机构研究动态与关注点、机构核心研究人员（团体）、机构科研投入、作者机构变迁、基金机构分布、期刊论文机构分布
第 3 阶段	个性化服务	机构用户行为分析与研究、机构数字图书馆建设、相关机构推荐、机构科技监测
	互操作服务	互操作对象包括：图书编目系统；其他文献数据库、科学数据库等资源库；卫生信息化系统

2.2 作者机构名称规范内容

CBM 作者机构规范包括 3 个方面：形式规范、一般性描述属性规范和关系属性规范。形式规范的目标是实现一个机构的不同著录形式能够汇聚在一起，用同一个表达形式（下称规范机构名称）进行表达。一般性描述属性规范是对机构基本信息的揭示与控制，主要指机构类型、所属领域、等级、所在地区、地址等一般描述信息的规范。关系属性规

范可分为系统内部作者机构间关系规范和系统与外部机构规范文档之间的关系规范两个层面。系统内部作者机构关系规范包括机构变更、隶属、挂靠、附属、相关等关系的规范；与外部机构规范文档之间的关系规范主要指与外部机构规范文档的互操作，直接表现为各种映射关系。形式规范是构建作者机构规范文档和实现各类机构知识服务活动的基础，一般属性和系统内部作者机构关系属性规范是实现深度知识检索与评价的基础，外部关系规范是实现不同系统之间资源和服务整合的基础。

3 CBM 作者机构规范文档体系结构设计

3.1 概述

CBM 作者机构规范文档体系的设计不限于最终规范内容本身，还考虑了规范控制过程和边建设边服务需要。CBM 作者机构规范文档体系由 7 大类文档组成，分别为作者机构名称来源文档、预规范作者机构名称文档、辅助规范文档、作者机构规范文档、作者机构名称索引文档、映射文档和管理文档，见图 1。内部规范文档主要通过规范作者机构名称 ID、预规范作者机构名称 ID、原始作者机构称 ID 进行关联；与外部规范文档的映射主要基于规范作者机构名 ID 进行；与 CBM 文献库和论文其他知识要素的语义关联主要基于上述 3 类唯一标识符与 CBM 文献 ID 的映射关系进行。

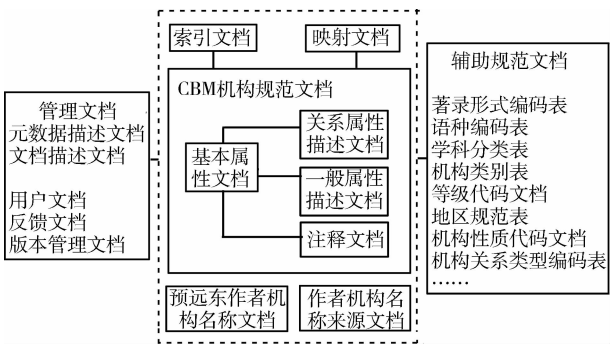

图 1 CBM 机构规范文档整体体系结构

3.2 作者机构名称来源文档

作者机构名称来源文档存放的是从 CBM 中采集过来的原始作者机构名称及相关描述信息，内容包括 CBM 文献 ID、原始作者机构名称 ID、原始作者机构名称、邮编、作者、所在地等。

3.3 预规范作者机构名称文档

预规范作者机构名称文档存放的是对作者机构名称来源文档中相关信息清洗和初步规范后的结果，内容包括预规范作者机构名称 ID、原始作者机构名称 ID、预规范作者机构名称、语种、所在国家、所在地区、机构类型等。

3.4 作者机构规范文档

CBM 作者机构规范文档由基本属性文档、一般属性描述文档、关系属性描述文档和注释文档组成。基本属性文档里用于存储 CBM 作者机构规范名称的基本信息，核心内容包括规范作者机构名称 ID、预规范作者机构名称 ID、原始作者机构名称 ID、CBM 文献 ID、规范作者机构名称、优选规范作者机构名称标识。一般属性描述文档用于存储 CBM 作者机构规范名称的类型、机构分类、所属学科与领域、所在地区、分级、语种、性质等一般描述信息。关系属性描述文档用于存储 CBM 作者机构之间关系信息。CBM 关系属性描述文档中关系可以是规范作者机构名称之间的关系、预规范作者机构名称之间的关系，也可以是规范作者机构名称与预规范作者机构名称之间的关系。关系类型分为变更关系（拆分、合并、更名等）、层级关系（隶属、挂靠、附属等）、相关关系（作者相关、文献相关、基金相关、领域相关、分级相关等）和其他关系 5 大类。注释文档是对 CBM 作者机构规范名称各个规范项的说明和其他信息的补充说明，既是建设成果，也反用于辅助 CBM 作者机构规范名称文档的构建。

3.5 作者机构名称索引文档

作者机构名称索引文档分为 CBM 作者机构名称索引文档、作者机构规范名称索引文档和作者机构预规范名称索引文档，前者是对后二者的综合。索引方式上包括字索引、词索引和综合索引。作者

机构规范名称索引文档主要服务于外部系统，CBM 作者机构名称索引文档和和作者机构预规范名称索引文档主要服务于 CBM。

3.6 映射文档

映射文档主要用于存储和揭示 CBM 作者机构规范名称与其他机构规范文档中规范机构名称之间的映射关系，实现 CBM 作者机构规范与外部系统和服务的互操作。核心元数据项有 CBM 规范作者机构名称 ID、外部机构规范名称唯一标准符、外部机构规范文档名称编码、映射关系类型。CBM 作者机构规范名称与其他机构规范文档中规范机构名称之间的映射关系类型主要分为等同映射、向上映射、向下映射、相关映射和其他映射 5 大类，其中相关映射又分为行政相关、地区相关、学科相关、类别相关等。

3.7 辅助规范文档

CBM 作者机构辅助规范文档主要用以辅助机构一般描述项内容的规范，有些是面向所有类型机构，有些则是面向特定类型机构。表 2 是主要辅助规范文档及用途说明。所有辅助规范文档均可动态更新与维护。

表 2 CBM 机构规范主要辅助规范文档及用途

序号	辅助规范文档名称	主要用途
1	著录形式编码表	作者机构名称著录形式揭示，如简写、缩写、作者常用写法、官方名称等
2	语种编码表	作者机构名称语种揭示
3	学科分类表	作者机构所在领域揭示
4	机构类别表	作者机构分类规范。不同类型机构（区分高校、卫生机构、实验室等）采用不同的类别表
5	地区规范表	作者机构所在地区规范
6	等级代码文档	作者机构等级规范。不同类型机构（区分高校、卫生机构、实验室等）采用不同的等级代码规范文档
7	机构性质代码文档	作者机构经营性质揭示
8	机构关系类型表	机构关系规范，是对同一类型机构内部和不同类型机构之间可能具有的关系的总结
9	映射关系类型表	CBM 作者机构规范名与其他机构规范文档中规范机构名称之间的映射关系规范
10	机构名称文档编码表	外部机构名称文档登记与管理

3.8 管理文档

管理文档用于各类数据的管理与说明，包括元数据描述文档、文档描述文档、用户管理文档、反馈文档和版本管理文档。元数据描述文档用于解释各类 CBM 机构规范文档涉及的元数据内涵和外延；文档描述文档是对各类 CBM 机构规范文档内容的说明；用户管理文档是对 CBM 机构规范文档的构建、维护和应用等各类型用户的统一管理；反馈文档用于记录 CBM 机构规范文档的使用反馈信息和反馈信息处理情况；版本管理文档用于记录 CBM 机构规范文档更新变化情况。

4 CBM 作者机构名称规范文档实现

4.1 作者机构名称规范过程

CBM 作者机构名称规范过程整体分为原始作者机构名称采集、清洗、形式规范控制、一般性描述属性规范控制和关系规范控制 5 个主要阶段。（1）原始作者机构名称采集阶段主要是从 CBM 数据库中获取完整的原始作者机构著录信息。（2）清洗阶段主要是对采集过来的原始作者机构名称进行拆分，生成原始作者机构名称唯一标识符；对拆分后的作者机构名称进行形式检查、提取有效片段、去重，生成预规范作者机构唯一标识符；完成机构类型、所在地区、语种等部分非关系属性的初步规范等。（3）形式规范控制阶段主要是将同一机构的不同作者机构著录形式的汇聚在一起，并从中推荐出 CBM 规范机构名，生成规范作者机构名称唯一标识符。（4）一般性描述属性规范控制阶段主要是完成机构类型、所属领域、所在地区、分级等所有非关系属性的规范。（5）关系规范控制阶段主要是进行机构变更、隶属、挂靠、相关、映射等关系的规范，生成各类关系唯一标识符。

4.2 作者机构规范文档主要实现策略

CBM 作者机构规范文档构建原则之一就是边建设边服务，因此重点强调构建过程的阶梯式循环，保证中间规范成果可用和可复用，注重计算机辅助

和社会化协作。

4.2.1 阶梯式循环建设 如图 2 所示，CBM 机构规范文档整体构建路线不是线性的，而是循阶梯式循环进行的。首先启动核心类型机构规范，且只考虑形式规范；随后在上述基础上进行一般性描述属性规范，并引入非核心类型机构规范控制；接着启动 CBM 作者机构名称内部关系规范，同样是核心类型机构优先；最后着手构建 CBM 作者机构名称与外部机构规范文档映射关系。具体实施时还考虑年代范围和机构类别因素。

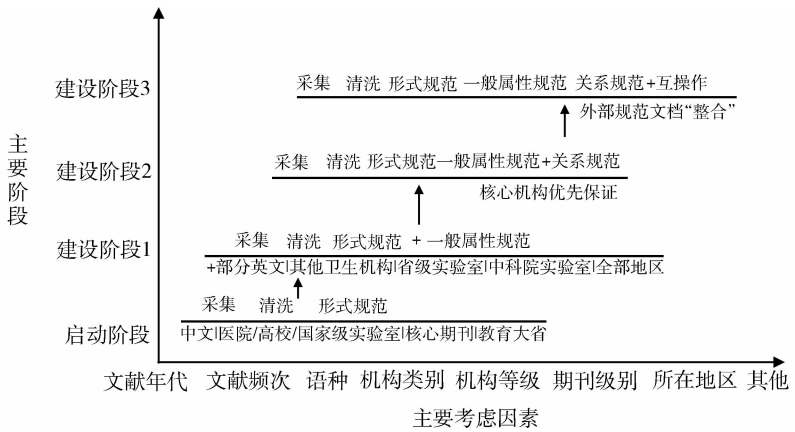

图 2 CBM 机构规范文档整体构建路线

4.2.2 计算机辅助 图 3 是 CBM 作者机构规范文档构建与维护的计算机辅助环境，整体分为应用层、软件层和技术层 3 个层次，贯穿 CBM 论文作者机构名称采集、清洗（预规范）、形式规范、关系规范、互操作和服务各个阶段。

图 3 作者机构规范文档构建计算辅助环境

采集和清洗主要基于各类离线工具进行，涉及的核心技术主要为不同类型机构特征词的总结与规则库的构建。形式规范、关系规范和互操作则以在线协同加工工具为主，这 3 个阶段也是最需自动化处理技术和语义资源支持的阶段。其中，形式规范主要基于相似度技术、规则库构建、自动聚类和分类技术、同名消歧技术进行；关系规范主要基于自动关系发现技术进行，包括作者共现、文献共现、语义相似度技术、规则库构建和各类辅助规范文档的支持。互操作阶段是离线与在线相结合，主要基于语义相似度计算和规则库进行计算机推荐。服务模式主要有 3 种：通过发布工具提供检索和浏览服务；通过定制工具提供定制服务，通过规范接口提供数据调用服务。

4.2.3 社会化协作 学术论文机构规范文档的构建与维护是一个复杂、耗时的工程，需要开放与社会协作，建立社会化协作机制和工作模式。图 4 是 CBM 从工具、技术、标准与内容 4 个层面对需要参与的社会角色及分工进行了思考与总结。需要参与的社会角色应该包括 7 个社会角色，即作者、信息服务人员、用户、期刊编辑部、期刊采编系统、专家和其他机构规范编制机构，不同角色在不同层次需要发挥的作用各有侧重：CBM 主要负责提供技术和协同软件支持；作者、信息服务人员、用户重点参与内容规范与修正；期刊编辑部、期刊采编系统、专家和其他机构规范编制机构主要负责相关标准规范的制定与实施。

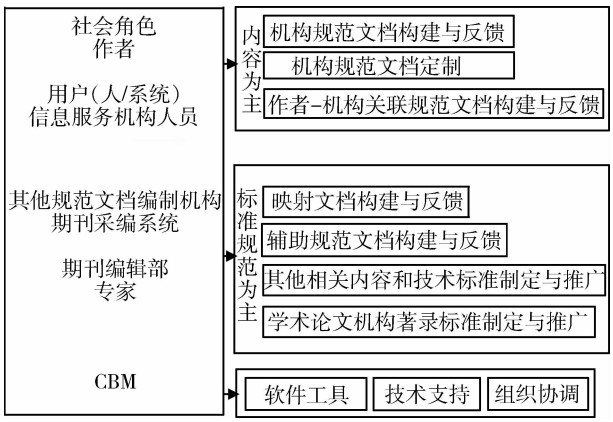

图 4 作者机构规范文档构建与维护社会化协作模式

5 结语

对主题、学科、作者、作者机构、期刊、基金等知识要素进行规范控制和语义关联，构建学术论文规范文档，用于学术论文的组织，是最大程度解除基于学术论文开展知识服务制约因素的重要手段之一[1]。本文重点对中国生物医学文献数据库 CBM 作者机构规范内容、规范文档组织方式、规范文档的构建过程与策略进行了介绍。目前 CBM 已完成近 190 万原始作者机构名称的形式规范，形成 9 万余条优选作者机构规范名、近 34 万优选作者机构规范名对应的其他形式，开始进入机构间关系规范和构建阶段，其中高等院校均已规范至学院级和系级，医院已规范至科室级。所有规范成果已在 CBM 数据库的机构检索、机构链接、作者消歧检索、引文分析、作者（第一作者）分析、机构分析、基金分析和期刊分析等服务中进行了应用。

诚然，目前各种关系发现与不同机构规范文档间机器互操作技术的研究还不够成熟，有些刚处于设计和试验阶段，其工程化应用还需要在 CBM 作者机构关系规范实践中不断优化。此外，还需进一步加强语义存储与描述技术研究，提高规范文档的语义化程度，积极参与到作者机构著录规范的制定、数字化表达等相关标准规范的制定中，更大范围内进行社会化协作实践，促进社会化协作环境的搭建，提高作者结构规范文档更新的动态性和实时性，接受更广范围的应用检验。

参考文献

1 曾建勋，王立学. 面向知识评价的规范文档建设方法[J]. 图书情报工作，2012，56（10）：101－106.

2 苏新宁. 图书馆、情报与文献学学术影响力研究报告（2000—2004）[J]. 情报学报，2006，25（2）：131－153.

3 董琳. 学科评价之文献计量数据准备[J]. 情报理论与实践，2010，33（6）：49－52.

4 唐金玲. 国际三大检索系统论文作者机构名称问题研究——以高校机构名称为例[J]. 情报探索，2014，（9）：80－84.

5 吴英杰. 孙海霞. CBM 数据库作者机构非规范著录数据自动检测研究[J]. 医学信息学杂志，2011，32（5）：38－40.

6 W3C. SKOS Simple Knowledge Organization System Reference：W3C Proposed Recommendation 15 June 2009[EB/OL].[2015－01－25]. http：//www. w3. org/TR/2009/PR－skos－reference－20090615/.

7 贾君枝. 简单知识组织系统与汉语主题词表[J]. 中国图书馆学报，2008，34（173）：75－78，84.

8 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）：1－2，9.

9 国际图书馆协会和机构联合会（IFLA）. 规范数据的功能需求[EB/OL]. [2014－12－15]. http：//www. ifla. org/files/cataloguing/frad/frad_ 2009－zh. pdg.

10 崔春，毕强. 虚拟国际规范文档（VIAF）项目进展[J]. 图书情报工作，2014，58（6）：129－134.

11 贾君枝，石燕青. 中文名称规范文档与虚拟国际规范文档的共享问题研究[J]. 中国图书馆学报，2014，41（214）：83－92.

12 卜书庆，郝嘉树. 国家图书馆中文书目规范控制现状及研究[J]. 图书馆论坛，2010，30（6）：209－213.

13 谢琴芳. CALIS 中文名称规范数据库建设方案及其实施进展[J]. 新世纪图书馆，2005，（1）：3－5.

14 郝嘉树，王广平. 中文人名规范的语义描述与关联探讨[J]. 图书情报工作，2012，56（14）：47－51.

15 陈金星，祝忠明. 责任者名称规范控制研究及进展[J]. 现代图书情报技术，2009，（12）：12－17.

16 高星，戴玮，黄利辉，等. 中文生物医学文献机构名称规范化研究[J]. 医学信息学杂志，2010，31（12）：56－60.

17 杨奕红，李亚萍，张立丽. 机构多层级词表的编制及在文献计量评价与科研绩效管理中的应用[J]. 数字图书馆论坛，2013，（6）：57－63.