

# 基于知识组织的机构规范文档构建方法研究<sup>\*</sup>

曾建勋 郭红梅

**摘 要** 机构是文献的重要描述要素,对于文献资源的组织和管理起着重要作用。机构数量庞大、层级关系复杂、更名重组现象频繁、各种形式名称交替使用,使得机构名称识别度降低,给以机构为单元的信息检索、知识导航、科研评价等工作造成严重困扰。为确保机构的唯一性,国内外组织和学者不断探索机构规范文档建设实践和名称消歧归一化的方法。本文在综合分析国内外研究现状的基础上,提出构建以机构规范文档为核心的知识组织体系,确保机构名称的唯一性、机构层级关系的有序性和机构间关联关系的规范化。该体系不仅涵盖一级机构,还涉及二、三级或更细粒度机构的建设,进而实现对各层级机构及其机构知识资源的规范控制和组织。本文重点讨论机构规范文档的总体结构、建设框架、建设思路、关键技术、服务应用模式和更新维护策略,以期为大数环境下机构属性和关系的全面描述和揭示提供支撑。图 6。表 1。参考文献 23。

**关键词** 知识组织 规范文档 机构实体

**分类号** G250.74

## Institutional Authority File Construction Based on Knowledge Organization

ZENG Jianxun & GUO Hongmei

### ABSTRACT

The institution is an important descriptive element of the literature. Identifying and tracking these institutions present challenges, as they often merge and split, acquire or are acquired by other institutions, change their names and locations. Institutions may be known by a variety of names over time even within one language and their hierarchical structures are also complex. So it is difficult to analyze institution knowledge unit in information retrieval, knowledge navigation, scientific research management and evaluation etc. In order to ensure the uniqueness of the institution, domestic and foreign organizations and scholars have carried out the construction of institutional authority file and explored institutional name disambiguation and normalization algorithm.

The paper tries to construct knowledge organization system centered on institutional authority file to ensure the uniqueness of institutional name, the orderliness of the institutional inter-hierarchical relationship, and the standardization of the relationship between institutions. The institutional authority file not only includes the first-level organization but also involves fine-grained organizations, such as the second-level and third-

<sup>\*</sup> 本文系国家社会科学基金重点项目“基于知识组织的图书馆资源发现服务体系研究”(编号:17ATQ002)的研究成果之一。(This article is an outcome of the key project “Research on Library Resource Discovery Service System Based on Knowledge Organization” (No. 17ATQ002) supported by National Social Science Foundation of China.)

通信作者:曾建勋,Email: zeng@istic.ac.cn,ORCID: 0000-0002-0432-9618 (Correspondence should be addressed to ZENG Jianxun, Email: zeng@istic.ac.cn, ORCID: 0000-0002-0432-9618)

level sub-units, and then realize the standardized control and organization of all levels of institutions and related knowledge resources.

The paper focuses on the overall construction architecture, construction framework, construction ideas, key problem and technologies of institutional authority file. It analyses usage requirements and provides several possible future service or application models and update and maintenance strategies. We hope knowledge organization system centered on institutional authority file can provide support for the comprehensive description and disclosure of institutional attributes and relationships in the big data environment. 6 figs. 1 tab. 23 refs.

## KEY WORDS

Knowledge organization. Authority file. Institution name.

## 0 前言

互联网发展带来数字化资源激增,如何利用知识组织体系对这些海量资源,尤其是其中蕴含语义的文本资源进行描述、组织、揭示、深度聚合,一直是业界关注的重点。近年来业界加强了对主题词表、分类法、领域本体等知识组织体系的构建与应用研究,从研究主题或领域实践角度对海量资源的知识组织进行关联,形成语义环境下的知识关系网络,推进知识组织体系在知识导航、智能检索、计量分析等服务中的应用。

科技文献中,作者机构作为资源的创造者和发布者,是文献的重要描述要素,也是进行检索入口、资源导航、知识组织和关联聚类的知识单元。机构规范文档是知识组织体系的重要组成部分,它不仅包含对机构名称要素的规范描述,也涉及对机构层级关系的梳理,因此应从知识组织的角度来推进机构规范文档的建设和应用。但由于文献中所包含的机构对象数量庞大、层级关系复杂,机构淘汰、更名、重组等现象频繁,加之曾用名、全称、简称、俗称等不规范名称交替使用,导致现有机构名称识别度降低;相关机构的组织结构和从属关系模糊,机构属性和特征描述欠缺或表述多样,给以机构为知识单元在知识导航、检索、统计评价中的应用造成严重困扰。为此,需要以知识组织的方法和思维来构建机构规范文档,以确保机构名称的唯

一性、机构层级关系的有序性、机构间关联关系的规范化。除了要针对一级机构属性和关系外,也要注重对二、三级机构属性和关系的规范文档建设,即要构建以机构规范文档为核心的知识组织体系,对机构的各种名称及其属性信息进行有序集中,梳理其中的组织关联结构、参照关系,进而实现对机构及其知识资源的规范控制和组织。

## 1 机构规范文档的研究现状

规范控制思想最早由美国图书馆学家卡特于1876年提出。“规范文档”(Authority File)概念来自文献编目领域,指由规范记录组成的计算机文档,通过规范控制来确保记录的唯一性和稳定性<sup>[1]</sup>。随着信息资源数量的增加,其中所涉及的作者、机构、期刊等名称类数据体量也不断增多,类型更为复杂。为有效集中不同名称下的信息资源,实现各类名称的标准化和规范化,规范文档的建设显得更为重要。

作为一种以机构为对象的重要知识组织体系,机构规范文档着重解决机构名称混乱、机构关系模糊、机构信息资源繁琐无序等问题,以实现与其他知识要素的整合与组织。这样,可以丰富机构入口的智能检索和分面导航,提升信息检索查准率和查全率,满足更精确的统计分析需求,优化科研实体的文献计量与评价,助力学科领域发展评估、科研成果绩效管理、机构评价等服务工作。

## 1.1 机构规范文档的建设实践

最早的规范文档由美国国会图书馆创建,当时称为“规范表”,专指规范记录的集合。1940—1980年代,图书馆编目工作都不同程度地涉及规范文档的建设。随着互联网技术的发展,维基百科、百度百科实现了基于互联网的名称规范文档(人名和法人团体)的链接和检索共享。2001年欧洲多个国家的图书馆、档案馆和文献信息中心等机构提出建设 LEAF(Linking and Exploring Authority Files)项目<sup>①</sup>。2003年美国国会图书馆、德国国家图书馆和联机计算机图书馆中心(Online Computer Library Center, OCLC)对机构、会议、地名等多种国际规范文档进行集成,构建了 VIAF(Virtual International Authority File)项目<sup>②</sup>。2003年 Ringgold 公司创建机构标识数据库,具有详细的机构数据与清晰的层次结构,除包含机构的各

种形式名称外,还存储有机构类型、层级关系、学科及行业分类等字段<sup>③</sup>。我国规范文档建设工作起步于 1960 年代,2003 年为促进不同地区资源最大程度的共建共享,国家图书馆、我国台湾汉学研究中心、中国高等教育文献保障系统管理中心、香港特别行政区大学图书馆、澳门特别行政区图书馆联合建设了中文名称规范联合数据库检索系统,包括了对个人、团队、会议等名称的规范<sup>④</sup>。随着对规范文档认识的加深和扩展,不少机构为满足机构知识库或学科领域资源整合的需要,有针对性地创建机构名称规范库,如中科院机构名称规范库<sup>[2]</sup>、中国生物医学机构名称规范库<sup>⑤</sup>等。此外不少数据库商为满足文献资源精确检索和精准服务的需求,也构建了机构数据库,如万方机构信息数据库<sup>⑥</sup>、维普机构信息数据库<sup>⑦</sup>和知网机构信息数据库<sup>⑧</sup>等(见图 1)。

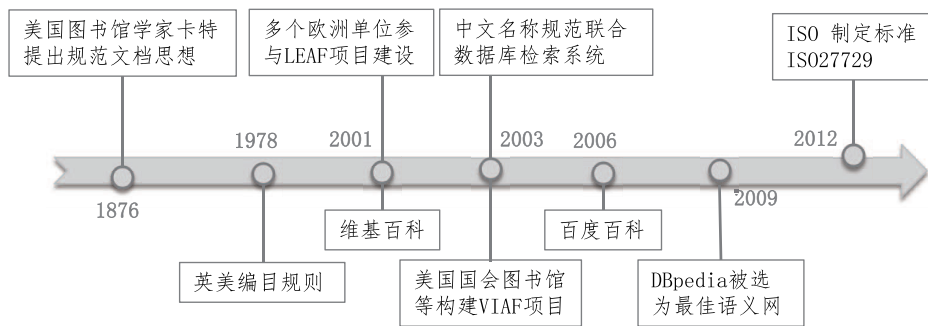


图 1 以图书馆为主体的机构实体编码实践

## 1.2 机构实体编码标准研究

全球法人机构识别编码(Legal Entity Identifier, LEI)是由 20 位数字和字母组成的唯一编

码,用于标识与国际金融交易相关联的法人机构<sup>⑨</sup>。1996 年《中国文献编目规则》出版,具体规定了规范档的编制问题,2005 年为顺应国际

① LEAF Project[EB/OL]. [2020-08-13]. <http://xml.coverpages.org/ni2002-03-28-a.html>.

② VIAF[EB/OL]. [2020-08-13]. <https://viaf.org/>.

③ Ringgold[EB/OL]. [2020-08-13]. <https://www.ringgold.com/identify/>.

④ 中文名称规范联合数据库检索系统[EB/OL]. [2020-08-13]. [http://cnass.ccna.org/jsp/index.jsp?lanval=zh\\_CN](http://cnass.ccna.org/jsp/index.jsp?lanval=zh_CN).

⑤ SinoMed[EB/OL]. [2020-08-13]. <http://www.sinomed.ac.cn/dataService/index.jsp?type=sjfw>.

⑥ 万方[EB/OL]. [2020-08-13]. <http://jc.wanfangdata.com.cn/#/index>.

⑦ 维普[EB/OL]. [2020-08-13]. <http://www.cqvip.com/>.

⑧ 中国知网[EB/OL]. [2020-08-13]. <https://www.cnki.net/>.

⑨ LEI[EB/OL]. [2020-08-13]. <https://lei.global/>.

编目实践与理论变革形势的需要,国家图书馆听取了各方意见后对编目规则进行了重新修订。2008年美国国家信息标准组织(NISO)创建了机构标准,包括机构的各种名称、类型及关联机构等。2012年国际标准化组织(ISO)出版了ISO 27729标准(International Standard Name Identifier, ISNI),主要对各种媒体中公共机构的名称进行标准化,以避免不同身份参与者使用相同名称而造成混淆<sup>①</sup>。NISO I2工作组推荐采用ISNI标准用于标识机构<sup>②</sup>。2012年Ringgold将ISNI标识符用于Ringgold标识数据库中,成为第一个将该标准应用于机构标识的注册机构。开放研究者与贡献者身份识别码ORCID(Open Researcher and Contributor ID)用于确保科研人员发表的论文与名字的唯一关联,为每个研究者保留特定的ISNI识别码<sup>③</sup>。

### 1.3 机构名称归一化方法研究

机构名称是标注机构其他属性和关系的前提,是编目规范、智能检索、科研评价以及知识库建设的基础。国内外学者对机构名称的消歧归一方法进行了诸多探索,包括基于字符串相似度的机构名称消歧方法、基于统计规则的机构名称消歧方法、基于机器学习的机构名称消歧方法和基于元数据的名称消歧方法等。

字符串相似度消歧算法有余弦相似度、欧式距离和编辑距离等,其中编辑距离最为常用。French等第一次将编辑距离的概念应用在机构地址字符串聚类中并取得了不错的效果<sup>[3]</sup>;黄林晟等提出了基于编辑距离的中文机构名称简称和全称的匹配算法<sup>[4]</sup>;蒋勇等利用归一化压缩距离算法对机构名称进行聚类<sup>[5]</sup>;杨昭等从语言学角度提出基于共现关系和相似度的机构名称归一化框架模型,用于中文文献题录数据机构名称归一化<sup>[6]</sup>。由于机构名称存在多种变

体形式,一些字符串层面相似度非常高的机构并非指向同一个机构实体,因此单用名称字符串本身进行匹配是不够的,需要与统计规则或其他特征相结合来进行机构名称的消歧归一。孙海霞等引入地区、类别和命名特征,将匹配规则与编辑距离相结合,对高校、医院与科研院所三类机构进行名称消歧规范<sup>[7]</sup>;黄水清等在文本相似性算法、编辑距离算法和基于规则算法的基础上,提出将统计模型和特征规则相结合来识别机构及其各种变异名称之间的一对多关系<sup>[8]</sup>;杨波等将规则和统计方法相结合实现了多个机构变异名称与规范名称之间的映射<sup>[9]</sup>;相晓敏等提出了基于规则和语义编辑距离的机构简称识别方法<sup>[10]</sup>。

此外,可以利用独立于词汇的通用特征来对机构进行描述,利用机器学习的方法对机构进行分类。Cuxac等提出将监督学习的贝叶斯模型与半监督的聚类算法相结合,对文献数据库中的机构进行消歧<sup>[11]</sup>;Rimmert利用机器学习的方法对德国机构层级关系和结构变化进行标注归类,设计不同的模式以满足不同结构的要求<sup>[12]</sup>;张建勇等将传统字符串匹配方法与词向量表示的深度学习方法相结合,基于SimNet框架搭建了词向量的匹配学习算法<sup>[13]</sup>;曾建勋等通过构建机构名称规范数据的语义模型来实现对机构名称的消歧规范<sup>[14]</sup>;周毅等结合NSTL名称规范数据的特点,构建科研实体名称规范的关联数据模型<sup>[15]</sup>。

为提高机构信息的扩展性及与已有机构资源的关联性,薛秋红等分析了中文机构名称规范与维基百科网络资源整合的优势及必要性,提出了语义网环境下的资源整合模式,以及评价整合结果的相关指标和算法<sup>[16]</sup>;Rimmert构建了从维基百科机构实体信息中抽取机构基本信息的过程,并评估维基百科作为机构消歧工

① ISNI[EB/OL]. [2020-08-13]. <https://www.iso.org/standard/44292.html>.

② NISO[EB/OL]. [2020-08-13]. <https://www.niso.org/standards-committees/institutional-identifiers>.

③ ORCID[EB/OL]. [2020-08-13]. <https://orcid.org/>.

具的效果<sup>[17]</sup>;黄永文等提出将机构永久性标识与已存在的名称消歧工具相结合,通过继承一些权威元数据标准来构建机构元数据规范模型,以支持以机构为单元的检索和相关信息的可视化分析<sup>[18]</sup>。

#### 1.4 机构规范文档的应用服务研究

机构规范文档主要应用在文献编目、信息检索、科学评价和机构知识库建设等方面,随着机构所描述属性特征的丰富化和关联关系有序化,机构规范文档也越来越多地与机构本体和机构画像研究相结合。Baldrige 描述框架认为机构画像是机构特征的描述,是自我评估的基础,可以辅助识别差距、规划未来和实现目标<sup>[19]</sup>;Paul 等从角色、规则、权利、责任和过程几个角度对机构进行描述,构建机构本体和概念模型,并指出概念模型不仅依赖机构本体,还需要其他与机构相关的本体,如 REA 本体和服务本体等<sup>[20]</sup>;孟琳在对多种文献和网络资源融合后,利用社团发现和关系抽取技术来识别机构的研究团队和兴趣主题,其偏重对挖掘算法的改进和创新,但并未深入分析属性之间的内在关系<sup>[21]</sup>;Taneja 基于机构网站中的信息和用户的评价,为两百多所高校定义元数据画像标签<sup>[22]</sup>。

总体来看,近年来国内外针对机构规范文档进行了大量研究与实践探索,积累了许多有益的经验,提出了很多先进技术方法,但仍存在需要完善和改进的方面。①一些大型机构或组织根据实际需求构建了自己的机构规范文档,但由于各自的使用目的不同,底层数据存储结构和规范控制流程、标准存在较大差别,各自独立,无法进行共享、重组复用。②目前一些商业化的机构规范库建设相对成熟完善,但由于商业目的和经济原因,并未开放使用,并且分散在不同中外文数据库商处。③机构规范文档建设工作相对复杂,还不能完全自动化、流程化,需要大量人工辅助,即使像科睿唯安这样世界知名的大型数据库商也设有专门的机构规范团队来检查机构地址信息的准确性,定期从机构官

网、Hoovers 企业信息数据库和 uniRank 高校排名数据库等权威来源收集机构信息<sup>[23]</sup>。④大多机构名称消歧归一方法和算法的探索研究,仅是利用小规模数据集进行的实验性研究,主要涉及高校和科研院所等一级科研机构名称的消歧规范,对机构之间的关系和二、三级更细粒度的从属机构特征很少涉及。而且这些方法大多是针对某一具体细节的改进,缺乏对整个机构规范流程的系统性研究,无法对大规模机构数据进行流程化规范,规范效率不高。⑤机构规范文档主要应用在以一级机构为单元的知识组织与导航、科研评价和机构知识库建设中,在机构画像和机构本体方面的探索应用较少。因此,本文尝试从知识组织的角度全面分析科研机构属性特征及其关联关系,构建能对科研机构进行准确描述和画像的机构规范文档。规范文档中包含对多个层级机构的属性、特征和关联关系的注释,以期为不同层级或粒度的科研机构评价、信息检索、知识库构建、对标比超等提供支撑,提升基于机构驱动的知识服务效果。

## 2 机构规范文档的体系结构研究

### 2.1 机构规范文档的内涵

本文的机构特指科研机构,即从事科学研究或与科学研究相关的机构,主要包括高等学校、科研院所、医疗机构、公司企业、政府机关等。文献及相关数据库、机构官网、机构新闻信息等是机构信息的主要来源,经过对机构来源数据的本地解析和网络采集,得到初始的机构名称及机构相关信息库,在初始库的基础上,经过规范整理和计算加工即可形成具有归一化“概念”的机构规范名称库、标识机构属性特征的标签库和揭示机构关联的关系库,三者统称为机构规范文档,具体如图2所示。为实现对更细粒度机构的统一化和规范化,机构规范文档不再局限对一级机构属性和关系的梳理组织,而是延伸对二、三级机构的描述和归一,随着



机构层级有序度和凝聚度的不断细化和机构属性特征的丰富化,机构规范文档自身就具有了

轻量级的本体和画像特征,所以机构规范文档也是构建机构本体和进行机构画像的基础。

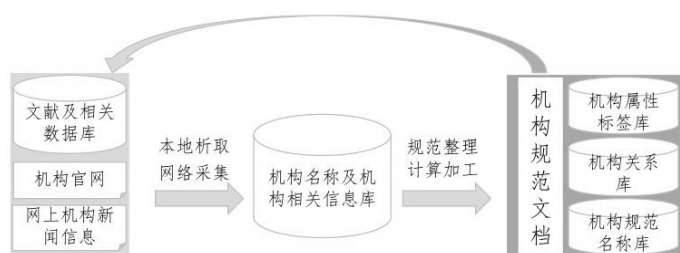
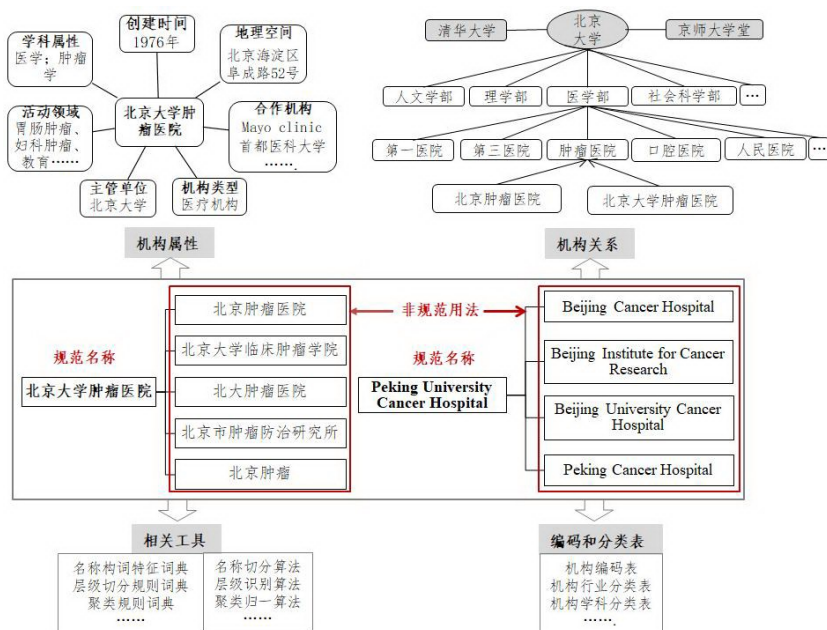


图2 机构规范文档的构成

## 2.2 机构规范文档的总体结构

机构规范文档建设是对机构信息进行规范化、迭代化、丰富化的过程。机构规范文档体系是由机构规范名称库、机构属性特征标签、机构关联关系以及建设过程中使用的相关工具(词典和算法等)和各种分类标准、分类体系等共同构成,如图3所示。在机构属性和关系标注方面,借鉴目前比较成熟的性质、领域、行业、主题等分类体系、规则、标准以及算法和技术,并结合数字图书馆资源中机构的特征,对重要的属性和关系进行定义和描述,并

设计相应的数据库结构来扩展性存储已定义的属性特征和关联关系,其中属性特征包括机构行业、学科、地理信息等,关联关系包括层级关系(北京大学—医学部—肿瘤医院)、沿革关系(北京大学—京师大学堂)、同一关系(北京大学肿瘤医院—北京肿瘤医院)和相关关系(北京大学与清华大学)等。此外,机构规范文档建设过程中也会积累比较实用的算法、特征词典等工具,描述机构唯一标识编码和行业领域分类体系、类目,它们共同决定机构规范文档的建设和使用效果。



2.3 机构规范文档元数据描述框架

元数据及其标准是机构规范文档建设的前提,前期必须制定完善准确的元数据描述框架来定义和描述机构实体的所有属性和关系,以准确定位和描述机构实体。

属性是对对象的抽象刻画,具有相同属性的事物就形成一类,关系是基于属性特征的关联来构建的,两个或多个机构实体可通过特定属性而产生关联关系。将机构实体作为概念,属性及关联关系作为概念特征,机构实体间可通过多种属性和关系关联在一起,形成立体化的机构网络图谱。机构的主要属性包括名称、研究领域、地域归属、联系方式等,进一步基于某个或某些属性构建机构间的关联关系。对机构各种曾用名的变更过程分析可以得到机构沿

革关系,对机构名称的切分可以得到层级隶属关系,基于研究领域的关联可以得到机构的学科关系、行业关系。基于成果关联和领域关联,结合相同的计量评价指标可以进一步识别某机构的相关或对标机构。

描述性元数据是应用范围最广的元数据,可对机构的内容、属性、外在特征等进行描述。为构建完整的机构实体描述元数据集,本研究综合百度百科、维基百科、Ringgold 数据库等成熟的机构集属性,在全面调研国内外机构规范文档元数据结构和机构描述相关规则、深入分析机构规范文档应用场景、咨询相关专家与机构数据库管理人员的基础上,最终设计了机构规范文档元数据集,表 1 中列出了主要元数据集。

表 1 机构规范文档主要元数据集

| 序号 | 元数据类别 | 元数据名称    | 字段标识       | 元数据表示                   | 必备性说明 |
|----|-------|----------|------------|-------------------------|-------|
| 1  | 机构标识符 | 本地唯一标识符  | UnifiedID  | dct:Identifier          | 必备    |
| 2  |       | 全国组织机构代码 | CHNII      | Registered Code         | 有则必备  |
| 3  |       | ISNI 码   | ISNI       | Registered Code         | 有则必备  |
| 4  |       | ORCID 码  | ORCID      | Registered Code         | 有则必备  |
| 5  |       | NISO 标识符 | NISOII     | Registered Code         | 有则必备  |
| 6  | 机构名称  | 规范名称     | Name       | Formal Name             | 必备    |
| 7  |       | 译称       | Name       | Translated Name         | 有则必备  |
| 8  |       | 别称       | Name       | Another Name            | 有则必备  |
| 9  |       | 简称       | Name       | Abbreviation            | 有则必备  |
| 10 |       | 曾用名      | Name       | Former Name             | 有则必备  |
| 11 |       | 语种       | Language   | Language                | 有则必备  |
| 12 | 研究方向  | 学科归属     | Discipline | Discipline              | 有则必备  |
| 13 |       | 行业归属     | Industry   | Industry                |       |
| 14 |       | 活动领域     | Fields     | Fields                  |       |
| 15 | 机构日期  | 成立日期     | Time       | dct:createdDate         | 必备    |
| 16 |       | 变更日期     | Time       | dct:modifiedDate        | 有则必备  |
| 17 |       | 撤销日期     | Time       | dct:revocationDate      | 有则必备  |
| 18 | 地域归属  | 国家       | Country    | Vcard:country-name      | 必备    |
| 19 |       | 行政区划     | CADC       | Administrative division |       |

续表

| 序号 | 元数据类别 | 元数据名称 | 字段标识        | 元数据表示                | 必备性说明 |
|----|-------|-------|-------------|----------------------|-------|
| 20 | 描述信息  | 机构简介  | Description | Description          | 有则必备  |
| 21 |       | 发展历史  | History     | History              |       |
| 22 |       | 机构水平  | Level       | Level                |       |
| 23 |       | 机构性质  | Type        | Type                 |       |
| 24 |       | 主管部门  | Department  | Competent department | 必备    |
| 25 | 关联关系  | 层级关系  | Hierarchy   | Hierarchy            | 有则必备  |
| 26 |       | 沿革关系  | Change      | Change               |       |
| 27 |       | 相关关系  | Related     | Related              |       |
| 28 |       | 同一关系  | Equal       | Equal                |       |
| 29 | 评价评议  | 评论    | Reviews     | Reviews              | 有则必备  |

## 2.4 机构规范文档的存储结构

机构规范文档存储字段包括两部分内容,一是机构属性字段,二是机构间的关系字段。机构基本属性主要有唯一标识、规范名称、变异名称(如简称、译称、别称、曾用名、外文名称等)、类型、层级、学科、行业、简介、成立日期、更名日期、所在国家与省市、网址等。机构间的关系主要有同一关系、层级关系、沿革关系和相关关系。在遵循统一性、实用性、可扩展性的原则

下,设计机构规范文档的存储结构,将定义的属性和关联关系拆分成不同的子表结构,各表都通过唯一标识 ID 与规范名称表建立关联,如图 4 所示,共包含 7 张主表。除主表之外,机构还可通过规范名称或唯一标识 ID 与文献、专利等成果数据库关联,进一步基于成果信息挖掘推理机构之间的相关、相似或对标关系。这种数据存储结构可减少信息冗余,提高数据处理效率,未来如果新增属性和关系也易于扩展。

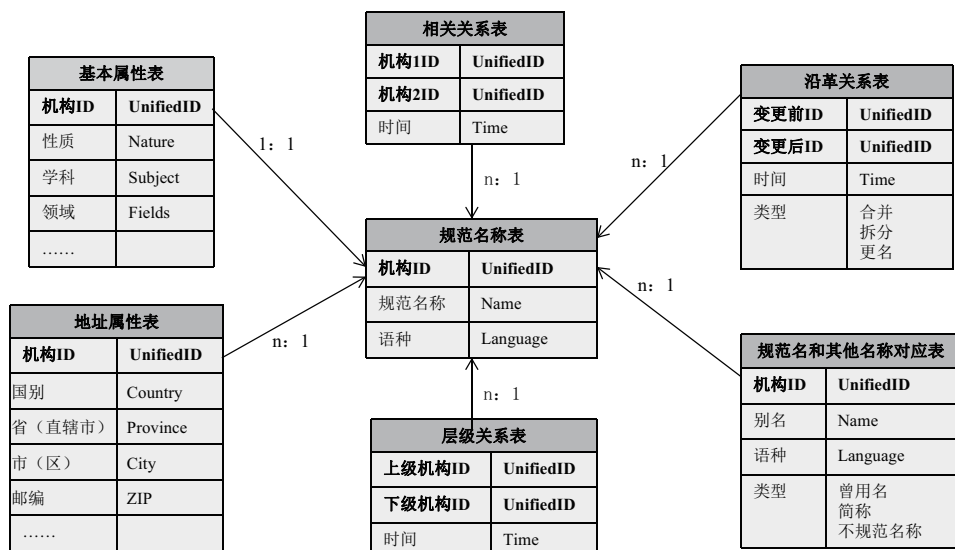


图 4 机构规范文档的存储结构



### 3 基于数字资源的机构规范文档构建

#### 3.1 机构数据特点分析

机构名称结构不一,形式多样,来源广泛。机构名称是对现实存在的机构实体的符号化表示。尽管机构实体具有唯一性,但由于在发展过程中机构淘汰、更名、重组与合并等现象频繁,加之语言的丰富性,导致有多种多样的名称形式与同一个机构实体对应,如简称、全称、曾用名、正式名称等。这些累积下来的大量机构数据,既存在于专门的机构库中,又出现在机构相关的成果库、百度百科、维基百科以及机构官网和机构名录中。此外,为提升机构规范文档的内容丰富度和用户参与度,用户也可注册与登记相关机构信息。

在机构来源数据中,成果数据库最多,尤其是中外文文献数据库,常用的有 CNKI 和 Web of Science 等,除此之外数字图书馆中还有大量专利数据、基金项目数据等,它们包含的机构数据相对结构化,可以直接下载或通过本地收割后对数据进行解析、切分,进而提取机构数据,还可以从

机构官网、权威规范库和权威网站等采集机构信息,进行解析、清洗规范后提取机构属性特征。

#### 3.2 机构规范文档构建框架

机构规范文档构建框架主要包括数据层、逻辑层、业务层和应用层,如图 5 所示。数据层表明机构数据的来源,主要包括已成熟的权威规范库、各类开源网站、订购与自建的文献数据库等,经过数据收割和远程采集来获取初始机构信息。逻辑层主要是采用算法逻辑和业务逻辑对初始机构数据进行加工规范,对机构名称进行聚类归一,对机构属性和关系进行标注等,算法逻辑主要指机构数据的切分算法、分级算法和名称消歧归一算法等,业务逻辑主要包括对规则、词典词表和采集、加工、质检任务的管理维护等。业务层是在数据层和逻辑层的基础上构建以机构名称规范库为核心的规范体系,包括机构名称库、机构属性标签库以及机构关联关系库等。应用层主要面向机构规范文档的数据使用与呈现,将整合集成、加工规范产生的规范数据通过集成调用接口为相关应用提供机构数据服务。

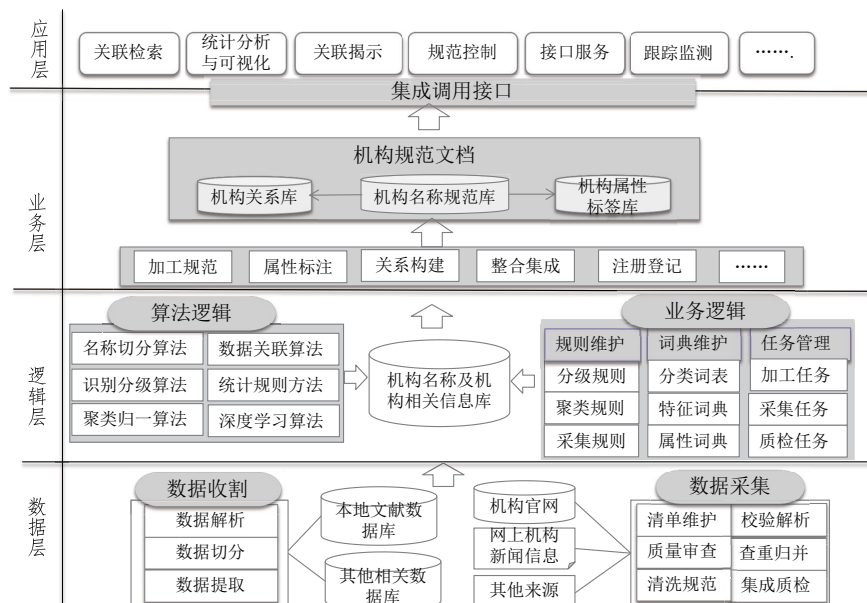


图 5 基于数字资源的机构规范文档构建框架

### 3.3 机构规范文档的构建思路

笔者结合海量学术论文数据库加工过程,尝试设计一套富有实践性、扩展性、自动化、可持续发展的机构信息规范化加工流程。借助数据清理规范、关联分析、深度挖掘、机器学习等大数据技术,设计一个智能化、自动化、流程化、协同化的名称规范管理系统。一方面能自动识别、监测机构名称变更情况,并对其关联关系进行揭示与管理控制;另一方面面向大众提供机构名称注册与登记系统应用服务,形成机构规范文档的开放式利用与可持续发展机制。在借鉴已有实践经验基础上,分析机构评价、精准检索等场景中对机构属性和关系的需求,凝练并

设计机构规范文档中应包含的属性和关系,基于机构信息规范化加工平台,完成对机构名称、机构属性和机构关系的标注和规范。构建思路如图6所示。在机构规范文档元数据模型的基础上,首先要完成机构各种名称的消歧归一,然后根据元数据框架中描述的定义和规则,完成对空间位置、学科、行业、领域、性质、功能、水平、学术地位、人才培养、科研成果等属性的标注。科研机构间除了沿革、层级关系外,还包含基于计量评价或主题分析推荐的相关、相似或对标机构等,应根据关系类型设计不同的提取和构建方法。

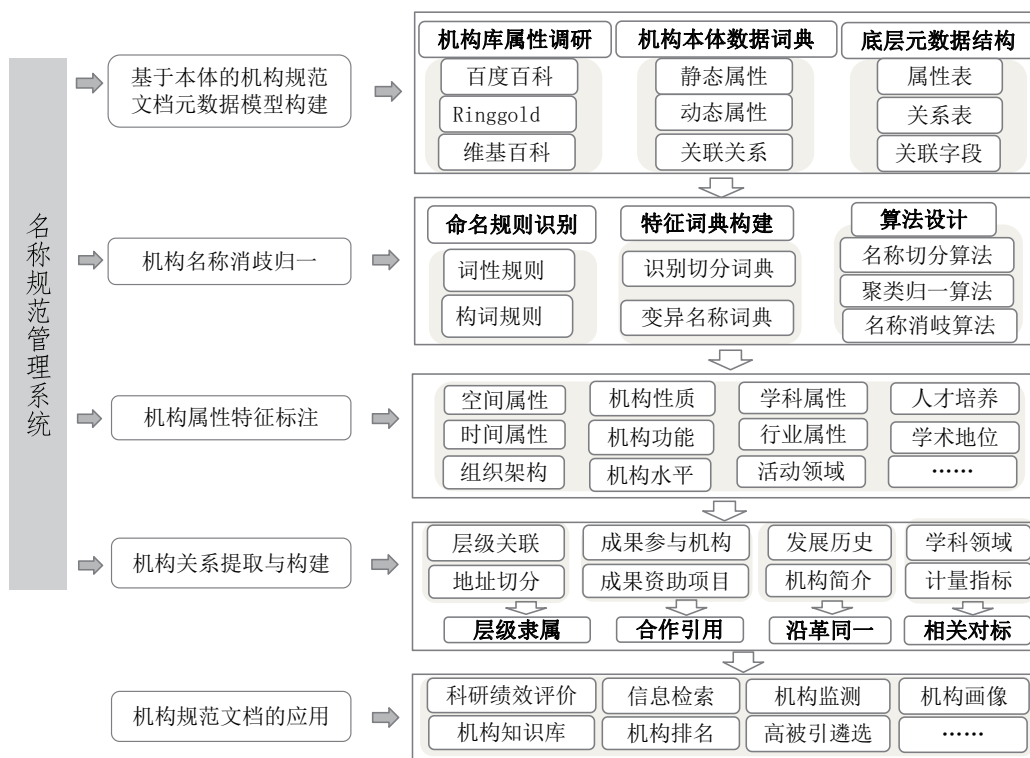


图6 机构规范文档的构建思路

机构规范系统按照功能类型分为前台和后台两部分,前台主要包括数据采集管理(本地数据收割和远程数据采集)、知识库管理(名称特征词典、分级规则、相似性规则、采集规则等的

管理)、协同规范(名称切分、识别分级、聚类归一等)和集成质检(质检任务管理、属性和关系的自动检查、人工抽查等)等功能模块,后台主要分用户管理、使用统计、系统管理等功能模

块。机构规范文档建设的最终目的是支撑对机构属性和关系描述的服务工作,因此需要验证机构规范工作的质量,根据实际用户需求实时完善、调整和更新加工流程和技术方法。

### 3.4 机构规范文档构建的关键技术

机构名称是收集和标注其他属性和关系的前提和基础,必须对机构更名、重组与合并等过程中产生的大量变异名称进行归一,建立映射关系,才能全面、准确地组织管理与机构相关的资源。因此,在机构规范文档建设中机构各种名称的消歧归一是最核心的工作。此外,对机构特征的识别应放在能揭示机构特性的中心词上,由中心词构成的特征词典是实现机构不同层级切分和名称规范的前提和保障。除了明确的属性和关系外,机构间还有些隐性属性和关系需要依据相应的规则或逻辑关系推理得到。下文重点介绍机构名称消歧归一、特征词典构建和关系构建三个关键问题的解决思路。

#### 3.4.1 机构名称消歧的相似度算法

机构名称归一的目的是集中同一机构各种形式的名称,构建各种变体名称与规范名称之间的映射关联,并为每个机构规范名称赋予唯一标识符以确保机构的唯一性,实现对机构的规范控制。相似度算法的选择直接关乎分类结果的好坏,目前比较常用的有编辑距离算法、地址邮编加字符串相似度算法和分布式词向量算法。

编辑距离算法的核心思想是计算将一个字符串转换成另一个字符串所需要进行的插入、删除、替换三种操作的最小次数。在编辑距离算法中,插入、删除、替换三种操作的权重完全相同,但对中文机构名称来说,插入操作一般对应的是机构全称与其缩写形式间的转换,删除操作是插入操作的逆操作,而替换操作意味着两个名称间存在不同的字,不同的字越多,语义差别越大。在对中文机构名称消歧时,三种操作所产生的影响是不同的,因此在计算名称相似度时应赋予不同的权重,增加替换操作的权重,

降低插入与删除操作的权重。

根据机构名称和地址相关信息来判断两个机构名称的相似性,从而将变异名称与规范名称进行对应。首先计算几个机构名称字符串相似度,地区与邮编两个特征在机构名称相似度计算过程中起辅助作用,当一个或多个样本数据的机构名称与待匹配记录的机构名称之间的相似度高于一定阈值时,才对两者的地区和邮编进行相似比较,在相似度计算过程中,优先考虑名称的相似度并赋予更高的权重,对地址和邮编的相似度赋予较低的权重。

机构名称是具有语义属性的文字按照一定逻辑关系构成的,在对名称消歧归一时,除了比对名称字符串相似性之外,还应考虑它们的语义逻辑。分布式词向量算法在利用特征词典对机构名称消歧的过程中,将机构名称分解为具有语义属性的词向量,而且在计算过程中词向量的语义信息还可进一步关联和加强,克服传统仅基于字符串匹配的缺点。

#### 3.4.2 机构名称特征词典的构建

机构名称是由中心语和修饰词组成,中心语是能反映机构实体基本特征的词或短语,用于限定机构的活动领域、类型、性质等。修饰语是对中心语的补充说明,从时间、空间、次序等方面对中心语的外延进行限定,可能存在,也可能不存在,具有一定的不确定性。广泛收集机构名称特征词以及地区、学科、行业等属性,构建机构名称特征词词典和属性词典,是实现机构不同层级名称切分、机构名称归一化、机构属性规范化的前提与基础。

首先采用汉语分词系统对机构名称进行分词处理和词性标注,其中名词所占比例最大,其次是动词和形容词。然后通过词频遴选出高频的、能标识机构特征的词语作为特征词,又可细分为无限定成分的、不可拆分的单位特征词和有限定成分的限定型特征词两种,如公司、研究所、学院、系等单位特征词,有限公司、有限责任公司、附属医院等为限定型特征词。虽然不同层级机构名称特征词之间存在差异性,但特

征词和机构层级之间并不存在关联,包含某个特征词的机构名称可能属于不同的层级,比如“学院”这一特征词,既可以出现在一级机构名称中,也可以出现在二级或三级名称中,因此在特征词典构建中只按照机构性质进行整理,而不区分层级。

### 3.4.3 基于统计规则的机构关系构建

机构间存在层级、合作、引用、相关、相似或对标等关系,它们是基于统计规则和知识推理的方法从已标注的属性特征和机构相关联的科研产出字段中挖掘推断的。层级隶属关系可通过解析科研产出中作者单位的字段获得,科研人员在成果机构署名中通常会从顶层机构标识到最细小单元,因此依据切分规则词典对署名机构进行切分可得到不同粒度机构的层级隶属关系,又可按照机构性质将层级关系细分为科研直属关系、教学直属关系、产业关系等。同一科研成果作者的署名机构之间存在合作关系,标识同一基金项目的不同科研成果的署名机构之间也视为合作关系。科研机构间的引用关系主要基于科研产出的参考文献来构建,又细分为直接引用关系、共被引关系和耦合关系。相关机构是具有紧密合作或引用关系的机构,相似或对标机构是指同一研究领域人员规模、科研产出、学术水平和影响力相当的机构,可从所属研究领域、人员规模、科研产出、学术影响和科研合作几个角度设计相应的计量评价指标,通过计量评价指标来辅助科研机构识别相关、相似和对标机构。

## 4 机构规范文档应用与维护

### 4.1 机构规范文档的应用模式

机构规范文档除了包含对机构各种形式名称进行有效组织管理的规范名称库外,还包含能组织和揭示机构属性和关联属性的标签库和关系库。因此该机构规范文档除支撑传统的以机构名称为入口的检索导航和一级机构的统计分析、科研评价管理外,还可基于机构的语义特

征和关联关系探索更多的服务应用模式。

(1)推进以机构为入口的智能检索、推荐和导航。具有多种语义属性和关联关系的机构规范文档具有轻量级的本体功能,不仅对机构的属性和关系进行了明确的定义和规范的语义描述,还可对属性和关系进行语义聚合,依据全局信息和知识推理来挖掘机构之间的隐形关系,形成具有语义关联的网络。它除了满足传统的基于机构名称字段的检索外,还可支持按照某种属性或关系对机构进行关联检索,或者将属性和关联关系进行组合检索。因此,可实现基于整个机构语义关联网络的机构链接分析,可准确定位机构信息,来有效呈现机构的组织架构、发展脉络和关联关系等。

(2)推进更细粒度机构单元的统计分析、计量与评价。机构规范文档不局限于对一级机构的深度规范加工,而是扩展到二、三级机构,并构建了机构间的层级关系,标注了不同层级机构的研究主题、学科、行业等属性特征。因此,依据规范文档可全面准确地收集各层级机构的科研成果,满足不同需求的统计评价工作,实现同领域、同等规模机构间的对比分析。此外,对细粒度机构的属性和关系进行标注,也有助于深度了解机构科研主题的变迁和学术影响力的演化,从不同维度揭示机构科研人员、学科、行业领域的发展特征,辅助机构识别未来可能的竞争对手和合作团队。

(3)促进以机构为核心的多源知识集成与融合。网络环境下,资源分散性较大,即使同一机构的资源也分散在不同数据库中,同一类型的资源也分散在不同数据库商处,虽然它们的存储结构和标准不同,但都具备机构属性。机构规范文档中对机构名称进行统一及标准化处理,可将机构规范文档与文献、专利等成果数据库建立关联,构建机构信息之间的映射,快速消除不同数据库中机构名称的歧义,实现同一机构相关资源的融合集成。此外,随着开放获取理念的广泛认可,越来越多的机构开始构建开放资源库,开放资源的收集和建设更离不开与



国际接轨的机构规范文档的支持。

#### 4.2 机构规范文档的更新维护机制

受机构数量众多、属性多样、关系复杂的影响,对机构数据进行全面清理和规范不能一蹴而就,需要不断迭代循环,逐步扩充,分阶段分目标完成。本研究通过对文献数据库的调研发现,大量文献集中在少量高产出、高影响力的机构中。对2017年SCI论文中国第一作者的情况进行统计,结果显示94%以上的论文集中在1800多个机构中,其中发文量10篇以上的机构有1200多个,覆盖了97%以上的论文;对2010—2019年“中国科技论文统计源期刊”进行统计,结果显示81%以上的论文集中在3800多个机构中;对2010—2019年万方数据库中的论文机构进行统计,结果表明80%的论文集中在高产出、高影响力的6000多个机构中。因此,在机构规范文档前期建设中,重点对遴选出的几千个高水平科研机构进行深度的整理、规范和标注,能初步满足对80%以上论文的统计分析需要。后期根据机构发展水平和实际服务需求,依据科研产出和学术影响力的计量指标,逐步扩大机构遴选范围,最终形成覆盖全面、规范统一、内容丰富的机构规范文档。

随着机构自身的发展和学术环境的变化,新机构会产生,旧机构会被淘汰,机构的某些属性特征和关联关系也会发生改变。在机构规范文档服务应用的过程中,用户也会提出越来越多新的需求。因此,需要制定合理的更新维护机制,来及时响应机构规范文档中机构的新增、字段的更新、用户的反馈等问题,推进机构规范文档的持续更新和使用。对于已遴选进入机构规范文档中的机构,系统需要制定自动跟踪扫描策略,对已有的属性和关系进行监控,定期采集机构官网和其他网络资源信息,经解析、清洗、集成质检后与机构规范文档进行对比,如果发现冲突或识别出某种属性或关系发生了变化,则自动提示数据管理人员进行查核,依据不同属性和关系的更新规则,及时

修正冲突数据,更新属性和关联关系。对于用户,尤其是机构用户,他们对所在机构的属性特征、组织架构、层级关系更为熟识,能及时发现自动监控机制中无法识别出的错误或变化。因此要制定登记注册反馈机制,方便用户通过线上或线下登记注册的方式对错误、更新、应用需求进行反馈,继而再由管理员对反馈消息进行审核和响应。对还未遴选进入规范文档中的机构,系统选取能测度科研产出和学术影响力的计量指标并设定相应的阈值,定期从文献等相关成果库中统计各机构的指标结果,将产出量和影响力指标超过设定阈值的机构自动筛选出来,提醒数据管理人员将其遴选到机构规范文档中进行深度加工。

#### 5 结语

机构实体是文献情报领域重要的分析对象和知识组织单元,因此要加强对机构实体的属性与关系描述,优化机构实体的识别、消歧、匹配等规则与算法。机构规范文档是基于机构实体进行大数据知识组织的工具,是支撑科研机构快速展示、分析评估、管控成果以及构建机构知识库的基础,可以满足以机构为入口的智能检索、多维导航、计量评价等需求。机构规范文档建设需要在大数据导入、审核、维护、更新过程中不断完善,不断优化机构实体识别、消歧、匹配等规则与算法,既要解决机构名称归一、演化问题,又要解决机构层级划分、特征分类和属性标识问题。大数据环境下,更需要构建机构规范文档,以实现复杂数据中的机构名称进行消歧,对机构属性和关系进行准确揭示。机构规范文档的智能构建,更需要充分发挥机构主体的积极性,形成机构参与和应用的协同共建共享机制。在强化机构信息查询服务的基础上,强化以机构为入口的智能检索和导航,支撑机构知识库构建和同类机构的对标分析,推进支持系统嵌入的机构规范文档接口应用,继而全面提升机构规范文档的开放服务能力。



## 参考文献

- [1] 曾建勋. 知识链接及其服务研究[M]. 北京:科学技术文献出版社,2012:67. (Zeng Jianxun. Knowledge link and its service research[M]. Beijing:Science and Technical Documentation Press,2012:67. )
- [2] 李慧佳,马建玲,张秀秀,等. 中文机构名称规范库建设的实践与分析——以“中科院机构名称规范库”建设为例[J]. 图书与情报,2016(1):133-139. (Li Huijia, Ma Jianling, Zhang Xiuxiu, et al. The practice and analysis of the construction of Chinese institution name library—“Institution Name Authority of Chinese Academy of Science” as example[J]. Library & Information, 2016(1):133-139. )
- [3] French C J, Powell I A, Schulman E. Using clustering strategies for creating authority files[J]. Journal of the American Society for Information Science and Technology, 2000, 51(8):774-786.
- [4] 黄林晟,邓志鸿,唐世渭,等. 基于编辑距离的中文组织机构名简称—全称匹配算法[J]. 山东大学学报(理学版),2012(5):43-48. (Huang Linsheng, Deng Zhihong, Tang Shiwei, et al. Chinese organization's full name and matching abbreviation algorithm based on edit-distance[J]. Journal of Shandong University (Natural Science), 2012(5):43-48. )
- [5] Jiang Y, Zheng T H, Wang X, et al. Affiliation disambiguation for constructing semantic digital libraries[J]. Journal of the American Society for Information Science and Technology, 2011, 62(6):1029-1041.
- [6] 杨昭,任娟. 中文文献题录数据机构名称归一化研究[J]. 图书情报工作, 2020, 64(10):95-102. (Yang Zhao, Ren Juan. Research on institution name normalization based on Chinese bibliographic data[J]. Library and Information Service, 2020, 64(10):95-102. )
- [7] 孙海霞,王蕾,吴英杰,等. 科技文献数据库中机构名称匹配策略研究[J]. 数据分析与知识发现, 2018, 2(8):88-97. (Sun Haixia, Wang Lei, Wu Yingjie, et al. Matching strategies for institution names in literature database[J]. Data Analysis and Knowledge Discovery, 2018, 2(8):88-97. )
- [8] Huang Q S, Yang B, Yan L S, et al. Institution name disambiguation for research assessment[J]. Scientometrics, 2014, 99(3):823-838.
- [9] 杨波,杨军威,阎素兰. 基于规则的机构名规范化研究[J]. 现代图书情报技术, 2015(6):63-69. (Yang Bo, Yang Junwei, Yan Sulan. Research on rule-based normalization of institution name[J]. Data Analysis and Knowledge Discovery, 2015(6):63-69. )
- [10] 相晓敏. 中文组织机构名识别和消歧技术研究与应用[D]. 上海:华东师范大学,2016. (Xiang Xiaomin. Research and application of Chinese organization name recognition and disambiguation technology[D]. Shanghai: East China Normal University, 2016. )
- [11] Cuxac P, Lamirel C J, Bonvallet V. Efficient supervised and semi-supervised approaches for affiliations disambiguation[J]. Scientometrics, 2013, 97(1):47-58.
- [12] Rimmert C, Schwechheimer H, Winterhager M. Disambiguation of author addresses in bibliometric databases-technical report[EB/OL]. [2020-08-13]. <https://pub.uni-bielefeld.de/download/2914944/2914947/DisambiguationOfAuthorAddressesInBibliometricDatabases.pdf>.
- [13] 张建勇,钱力,于倩倩,等. 科研实体名称规范的研究与实践[J]. 数据分析与知识发现, 2019, 3(1):27-

37. (Zhang Jianyong, Qian Li, Yu Qianqian, et al. Constructing name authority for research entities[J]. Data Analysis and Knowledge Discovery, 2019, 3(1): 27-37. )
- [14] 曾建勋, 贾君枝. 机构名称规范数据的语义模型构建[J]. 大学图书馆学报, 2019, 37(1): 42-47. (Zeng Jianxun, Jia Junzhi. The construction of semantic model of organization name data[J]. Journal of Academic Library, 2019, 37(1): 42-47. )
- [15] 周毅, 张建勇, 刘峥, 等. 科研实体名称规范的关联数据模型构建[J]. 图书情报工作, 2020, 64(10): 109-117. (Zhou Yi, Zhang Jianyong, Liu Zheng, et al. Research on the construction of linked data model for research entity's name authority data[J]. Library and Information Service, 2020, 64(10): 109-117. )
- [16] 薛秋红. 中文名称规范档与维基百科的资源整合研究[D]. 太原: 山西大学, 2017. (Xue QiuHong. Research on the resource integration of Chinese name authority files and wikipedia[D]. Taiyuan: Shanxi University, 2017. )
- [17] Rimmert C. Institutional disambiguation for further countries: an exploration with extensive use of wikidata (project report) [EB/OL]. [2020-08-13]. [https://pub.uni-bielefeld.de/download/2930347/2930348/Institutional\\_disambiguation\\_for\\_further\\_countries\\_with\\_wikidata.pdf](https://pub.uni-bielefeld.de/download/2930347/2930348/Institutional_disambiguation_for_further_countries_with_wikidata.pdf).
- [18] Huang Y W, Li J, Sun T, et al. Institution information specification and correlation based on institutional PIDs and IND tool[J]. Scientometrics, 2020, 122: 381-396.
- [19] Baldrige organizational profile [EB/OL]. [2020-08-13]. <https://www.nist.gov/baldrige/baldrige-organizational-profile>.
- [20] Paul J, Maria B, Owen E. Institution aware conceptual modelling[C]//Proceeding of the ER Forum and the ER 2017 Demo track, 2017.
- [21] 孟琳. 多源信息融合的机构画像的方法研究[D]. 北京: 北京邮电大学, 2018. (Meng Lin. The research of organization profiling using the combination of multisource information[D]. Beijing: Beijing University of Posts and Telecommunications, 2018. )
- [22] Taneja G. How are higher education institutions defining their meta-description tags?[J]. International Journal of Educational Management, 2018, 32(7): 1293-1306.
- [23] Organization name unification procedures: unifying address variants [EB/OL]. [2020-08-13]. <http://ips.clarivate.com/m/pdfs/Clarivate-NameAddressUnificationProcess.pdf>.

曾建勋 中国科技信息研究所研究馆员。北京 100038。

郭红梅 中国科技信息研究所馆员。北京 100038。

(收稿日期: 2020-09-14; 修回日期: 2020-11-04)