·信息管理与用户研究·

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

李慧佳 马建玲 张秀秀 杨丽娜
(中国科学院兰州文献情报中心 甘肃兰州 730000)

摘 要:由于体制改革更名和中西文名称的简写、缩写等原因,许多中文机构名称存在不统一和不规范表述的问题,这对这些机构相关资源、学术成果的定位检索、共享与统计将造成一定的影响。文章通过概述和分析“中科院机构名称规范库”建设的思路、内容与服务,认为我国的中文机构名称规范库建设应在规范控制的全面性、规范标准的制定、与前沿技术的融合及可持续性建设等方面予以重视与发展。

关键词:规范控制;名称规范;关联数据;知识资源;中科院机构名称规范库
中图分类号:G250.74 **文献标识码**:A **DOI**:10.11968/tsyqb.1003-6938.2016020

The Practice and Analysis of the Construction of Chinese Institution Name Library
——*“Instiution Name authority of Chinese Academy of Science”as example*

Abstract Because of system change and the different expressions in Chinese and foreign languages, many Chinese institutions have different foreign language names or irregular expressions, and this phenomenon brings about negative influences on the statistics, sharing, and retrieval of resources and academic achievements of concerned institutions. Based on the practice of “Institution Name authority of Chinese Academy of Science”, the author believes that attention should be paid to the comprehensiveness of specification control, the development of standards, the fusion with advanced technology, and sustainability of construction in the construction of Name Library of Chinese Institutions.

Key words specification control; name authority; concerned data; knowledge resources; Institution Name authority of Chinese Academy of Science

随着历史变迁和体制改革,机构名称特别是团体名称因其基本职能、组织结构的变化也经常发生变化,而团体名称的中西文全称、简称的错写、漏写等问题又导致机构之间的关系错综复杂。从图书馆学发展视角来看,这种不规范、不统一的中文机构名称表述现象容易造成信息资源(如机构学术成果、机构人本资源、机构的实时动态信息、机构的社会网络数据)检索点的选择困难和错误,也不利于对其相关数据的统计和挖掘分析。于是,为机构建立规范文档,以实现对机构名称的规范控制便成为了自上世纪 70 年代以来各个国家和机构陆续采取的办法与措施。但在国际一体化的发展形势和背景下,不同国家资源之间的利用和共享在很大程度上依赖于数据的一致化和规范化,目前国内外这种名称规范文档的建设仍处于分散状态,采用的描述规范格式各有不同,从而导致数据之间的关联和交换非常困难。因此,如何将规范文档以关联数据的形式发布,实现跨平台的资源自由共享便成为了当前名称规范工作密切关注的问题。本文主要围绕名称规范控制这一核心问题,重点对“中科院机构名称规范库”的建设思路和建设内容进行了概述与分析,一方面以为中科院机构知识库的知识关联、知识发现开发提供协助支撑,另一方面也以求所得到的实践经验能为我国中文机构名称规范库的建设提供一些启示。

* 本文系中国科学院文献情报能力专项基金项目“开放知识资源登记系统(二期)”(项目编号:Y30005100)研究成果之一。
收稿日期:2015-12-06;责任编辑:魏志鹏

1 名称规范控制面临着新的课题

1.1 名称规范控制的发展与实践

本世纪初期是规范控制和规范文档发展的萌芽阶段,此时,规范控制仅限图书馆内部工作章程,并没有具体的相关规范控制实践活动。到了上世纪中期,图书馆界逐渐意识到规范控制的重要性,因此,在规范控制的理论发展和实践应用方面都有了初步探索。特别是上世纪70年代初随着计算机的普及应用,图书馆计算机化、自动化得以推广,从而促使图书馆自动化规范控制得到了进一步发展。尤为重要的里程碑是美国图书馆首先实现了图书目录的自动化生成,并实现了自动化的目录规范控制[1]。1977年,美国化学文摘社(CAS)研制出自动化的著者索引生产系统(AIMS),取代了原有的手工系统。该系统明确区分相似的人名,保证某一个人或团体的论文和专利都采用正确形式标引[2]。之后,图书馆界以及学者相继出版或发表了有关规范控制的图书、文章,涉及规范控制的理论、方法等内容;以及相关标准、文档,如”A little Brief Authority”(1978)、“Name Authority Control for Card Catalogs in the General Libraries”(1983)等。

中国对规范控制的研究和实践起步较晚,其中具有代表性的是中国国家图书馆。中国国家图书馆于1995年成立了中文名称规范组,使得中文名称规范工作有了组织保障。之后,名称规范控制得到不断发展,2003年国家图书馆应用Aleph500集成管理系统后,实现了中文名称规范数据对中文书目数据的实时控制。同年,中国高等教育文献保障系统(CALIS)联机合作编目中心启动了中文名称规范工作,其中文名称规范工作在遵从国际标准和模型的基础上还建立了适应不同文字/字体、不同MARC格式要求的共享机制。

为了实现中国大陆与港澳台地区对中文书目的规范管理,最大程度的实现资源共享共建,2003年,国家图书馆、CALIS管理中心、香港地区大学图书馆协作咨询委员会(Joint University Librarians Advisory Committee,JULAC)联合发起成立了“中文名称规范联合协调委员会”。联合协调委员会于2010年发布了“中文名称规范联合数据库检索系统”[3]。

1.2 名称规范控制的挑战与机遇并存

名称规范控制是根据一定的规范控制规则将名称相关信息统一标准化展现,以达到规范控制的目的。名称规范又包括个人、家族、团体、地理名称、题名规范等内容,规范数据展现了机构通过某一特定个人、家族、团体或具有统一题名不同版本来组织作品的受控检索点和其他信息。美国伊利诺大学的罗伯特名称规范控制工作有5个流程程:建立规范记录;将规范记录集中,形成规范文档;将规范文档和书目文档连接;对规范文档和规范系统进行维护;对规范文档和规范系统进行评估。简而言之,名称规范控制的建设就是建立规范记录、形成规范文档、及时对规范维护的过程。

目前,名称规范控制面临着一定的挑战和发展机遇,一方面,已有的研究者研究主要对个人名称规范控制(包括关于个人名称规范著录规则和标准的制定,个人名称规范控制理论与方法的研究,以及个人名称规范控制实践应用的探索)进行了一定的研究,而对较之于个人名称规范控制更加复杂的团体名称规范控制工作则研究者不多。这对现在的相关工作者来说,要搞清楚一个机构的历史变更情况、上下层级关系,以及不同的书写习惯等,需核实大量资料方可确定。但通常所能借鉴的研究成果和实践经验相对有限,使得这一工作极具挑战性。同时,大数据时代的到来,以及语义网技术、搜索引擎、元数据等新的网络信息资源组织方法和手段的出现,也都对名称规范控制提供了新的技术支持。如在冗余、虚假、错误信息较多的大数据时代,语义网技术的充分应用与数据规范控制工作相辅相成。一方面,语义网通过数据结构化与语义表征使得分散无联系的数据资源逐渐具备关联化的基础,规范控制能够保证语义数据之间的一致性,降低冗余度。另一方面,规范数据的权威性能够为可信网络服务提供支持[6]。因此,就有学者提出利用日益成熟的语义Web技术对各种名称标识进行规范控制[5]。

2 “中科院机构名称规范库”的建设实践与分析

随着近几年中科院机构知识库功能的完善和应

用的推广，越来越多的中科院院所及其他科研单位均开始应用机构知识库，机构知识库中的参与机构逐渐增加，资源的种类和数量也越来越丰富，如何实现资源的语义化关联、知识分析和知识发现便成为亟待解决的问题。同时，机构知识库中提交的各类资源所属单位名称参差不齐，缺乏统一的著录规范。为了解决诸如此类的问题，就必须对机构知识库中各类机构进行规范控制，最终实现用户的知识关联、知识发现等更高需求。“中科院机构名称规范库”就是这样一个旨在为中科院机构知识库中知识关联、知识发现提供基础保障的机构名称规范控制平台与工作机制。

2.1 建设思路

“中国科学院机构名称规范库”中所涉及的控制范围包括：研究单元、学校及公共支撑单位、共建单位、院直接投资的全资及控股企业、“四类机构”、院设非法人单元；以及中国科学院创新单元，包括国家实验室、国家重点实验室、中国科学院重点实验室、国家工程研究中心、国家工程技术研究中心、国家工程实验室、野外台站网络等内容。平台在建设思路设计上以中科院机构为突破口，设计具有普适性的机构名称规范控制业务流程和应用功能。

在规范控制的实现过程中，主要根据以下原则进行控制：①对于机构存在隶属关系的，应对所属关

表1 “中科院机构名称规范库”的元数据表

标签	可重复性	著录内容	必备性	说明
机构描述				
* 标识符			[1,1]	系统自动生成
ORCID			[1,1]	用于识别机构的ID号
ISNI			[1,1]	用于识别机构的ID号
RinggoldID			[1,1]	用于识别机构的ID号
* 机构名称(中文)			[1,1]	机构中文正式名称
* 机构名称(英文)			[1,1]	机构英文正式名称
机构缩写			[0,∞]	机构正式缩写名称
机构URL	+		[0,2]	机构网站首页的网址
机构描述	+		[0,1]	机构的简介或说明
* 机构类型	单选	下拉菜单(取值范围见参数表9)	[1,∞]	机构的类型划分
* 行政区域	单选	下拉菜单(取值范围：中国省份)	[1,1]	机构所在地理位置
* 学科主题	多选	取值范围见参数表5	[1,∞]	机构内容的主题描述
研究方向(中文)		多个方向用空格分开	[0,5]	
研究方向(英文)		多个方向用；分开	[0,5]	
机构地址			[0,1]	
电子邮件	+		[0,2]	
电话	+		[0,2]	
传真			[0,2]	
关联关系				
机构隶属关系	+		[0,∞]	和当前机构相关的上下级机构集合
机构变更关系	+		[0,∞]	当前机构的前向或后向集合
管理元数据				
元数据创建者			[1,1]	元数据记录的创建者
元数据创建时间			[1,1]	元数据创建日期
元数据修改者			[0,1]	元数据记录的修改者
元数据修改时间			[0,1]	元数据修改时间
元数据状态		待审核，已审核，已发布，已退回	[1,1]	记录的状态

注：登记字段中：字段名后带+号，说明字段可以重复，字段前带 * 号，说明字段必备。

系、共建关系、依托关系等进行名称规范控制;②对于机构存在历史变更的,应对更名变更、合并拆分、转移变更等进行名称规范控制;③对于机构存在其他名称的,要明确机构的主标目名称,而其他所有与机构主标目名称不同,但指向同一机构的名称作为连接标目名称。

为了实现团体名称规范控制,“中科院机构名称规范库”的元数据在依据都柏林核心元数据(DC)的基础上针对机构实体的特点加入了特色元数据(见表1)。

2.2 “中科院机构名称规范库”建设的内容

2.2.1 机构别名词表

“中科院机构名称规范库”主要根据《中国机读规范格式》CNMARC对机构名称进行规范化描述,地层操作采用MARC数据。当主标目出现两种或两种以上名称时,只需建立一条规范记录,第一与第二及其他主标目的关系视为同等。将同一机构名称的等同标目均记录于同一条记录,只要1**栏的主标目和7**栏的等同标目与资源平台上的目录链接起来,进行检索时便能查获同一机构的全部资源。

如:中国科学院西安光学精密机械研究所(见图1)对机构的中文、英文名称,简称,以及来源于Web of Science的各种该机构名称缩写组合。

通过检索机构的主标目,可恢复目录数据库中机构主标目下的所有标目及资源,并通过链接获得等同标目;反之,也可通过检索机构的等同标目来恢复目录库中与其相关的主标目和其他等同标目以及相关资源。

2.2.2 机构名称标识

在名称规范控制方面,为了对已规范数据进行识别和区分,避免规范控制工作的重复进行,也便于其他领域更加便捷的使用规范数据,众多机构的做法是对规范数据进行唯一标识。目前,名称唯一标识的主要研究集中在个人名称标识方面。例如美国宾夕法尼亚州立大学的YoojinH等人为解决数字文献作者名称的变动(包括姓名的变更、机构名称的分离和合并等),通过建立人物规范文档,赋予每个作者唯一的ID号,当名称发生变动时,将当前名称作为规范名,而将旧名作为属性字段予以保存[6]。Web of Science为提交论文的作者建立Researcher ID,以对数据库中的作者进行唯一标识,让作者管理自己的出版物列表,并跟踪出版物的被引频次和作者的h指数,从而识别潜在合作关系,以及避免作者的混淆。因此,可以说Researcher ID是关于学术研究社区中作者名称歧义问题的有效解决方案[7]。另外,ORCID (Open Researcher and Contributor ID)作为一个非盈利的组织,同样专注于解决学术研究中研究者名称混淆的问题,其通过为研究者配置唯一的并可链接的标识符,从而提升科学发现的进程并提高科研资助和合作效率[8]。

在机构名称标识方面,相对于个体名称规范的研究较少,较为突出的有ISNI(International Standard Name Identifier)和Ringgold identifier。ISNI(International Standard Name Identifier)是ISO认证的国际标

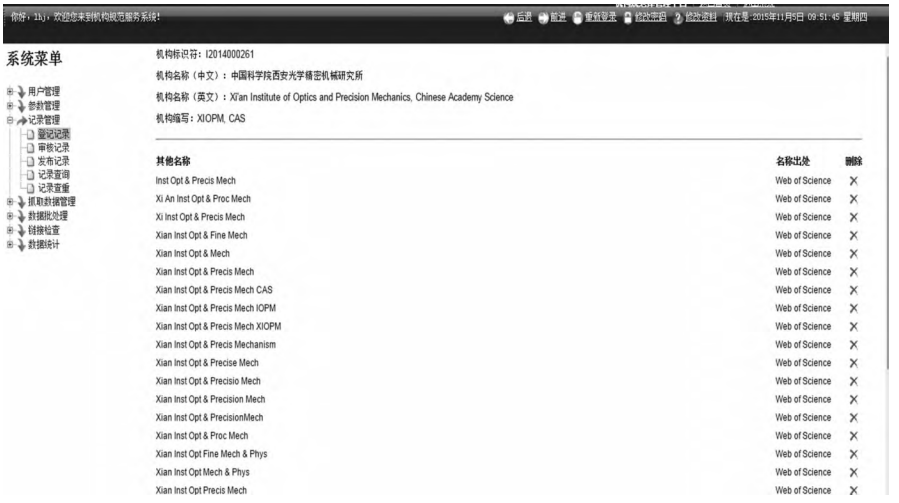

图1 机构其他名称规范

准代码，用于对成千上万的机构进行唯一标识，从而发现创造性的工作和行为。ISNI 是连接不同学科领域的桥梁，并且是关联数据和语义网技术应用的关键组件[9]。ISNI 是包含个人名称、团体名称等多个实体的名称规范。而 Ringgold identifier 是专门针对机构名称进行规范化标识。Ringgold identifier 数据库中存储了超过 370,000 个机构和联盟的唯一标识。Ringgold 是 ISNI 的登记机构，美国国家信息标准化组织推荐采用 Ringgold 来识别科研机构。另外，ORCID 利用 Ringgold 来实现个体科研人员与科研机构之间的关联。除此之外，The DUNS Number 是 The Dun & Bradstreet 公司的产品，主要是商业机构标识的国际标准代码，在 DUNS 数据库中存储了超过 2 亿的全球商业机构。DUNS 不仅仅是九位数字，而且还是一个用于保证机构信息准确、全面、及时的机构标识系统[10]。

“中科院机构名称规范库”考虑到为了使机构名称规范工作可以持续进行，并为今后科研机构在关联数据和语义网技术的应用方面起到决定性作用。因此，在机构元数据设置是充分考虑到机构标识的重要性，从而增加了 ISNI 字段和 Ringgold ID 字段。

2.2.3 机构关联关系

(1)机构历史关系：机构历史关系主要包括等级关系、共建关系、依托关系三类。其中：等级关系是双向关系(见图 2)。两个机构 A 和 B，其中 B 机构是 A 机构的分支机构(见表 2)，反之，A 机构是 B 机构的所属机构(见表 3)。

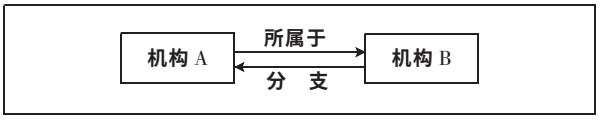

图 2 机构等级关系

表 2 分支机构示例

机构名称	分支机构
国家天文台	云南天文台
	新疆天文台
	南京天文光学技术研究所
	长春人造卫星观测站

表 3 所属机构示例

机构名称	所属机构
北京正负电子对撞机国家实验室	高能物理研究所
核探测与核电子学国家重点实验室	
纳米生物效应与安全性重点实验室	
粒子天体物理重点实验室	
核分析技术重点实验室	
北京市射线成像技术与装备工程中心	

共建关系是单项关系(见图 3)。机构 C 是由机构 A 和机构 B 共同建设(见表 4)。

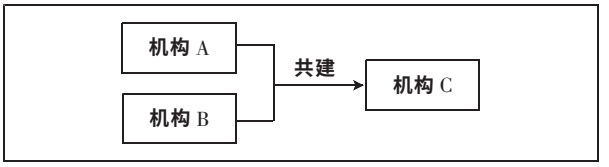

图 3 机构共建关系

依托关系为单项关系(见图 4)。机构 C 是由机构

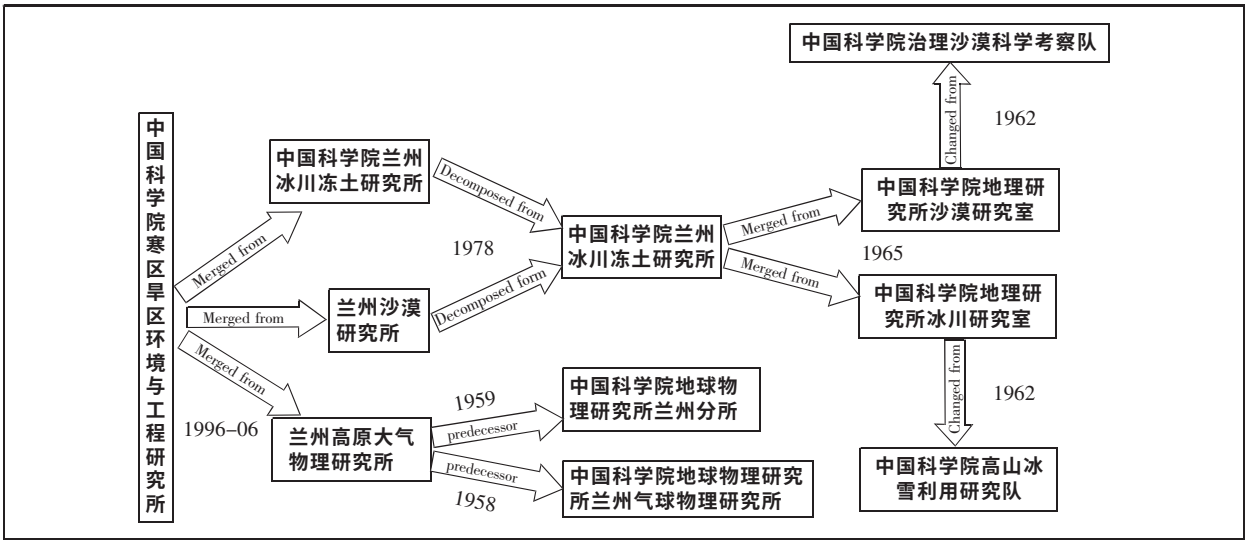

图 5 机构变更关系示例

A 和机构 B 共同建设(见表 4)。

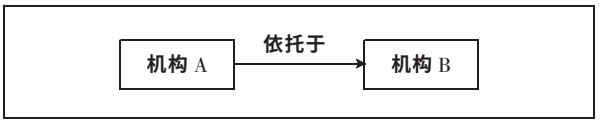

图 4　机构依托关系

表 4　共建机构示例

机构名称	共建机构
中国科学院水土保持与生态环境研究中心	中国科学院
	水利部
北京生命科学研究院--中美生命科学研究中心	中科院北京生科院
	美国内布拉斯加大学医学院

依托关系(见图 4)为单项关系。机构 A 是依托机构 B 建立的(见表 5)。

表 5 依托机构示例

机构名称	依托机构
中国科学院档案馆	文献情报中心
机器人技术国家工程研究中心	沈阳自动化研究所
中国科学院藏药研究重点实验室	中国科学院数学与系统科学研究院

(2)机构变更关系:机构的变更关系主要是原机构与现机构之间的相继关系。相继关系是原机构按时间顺序变更产生的现机构，包括更名(Changed from/to,现机构从原机构更名而来)、拆分(Decomposed from/to:现机构(一个以上)由原机构拆分而来,其原机构不存在;)、合并(Merged from/to,现机构由原机构(一个以上)合并而来)、转移(Transferred from/to,现机构由原机构中的部分转化而来,其原机构仍然存在)和前身(对于变更关系不明确的默认为现机构前身)五种关系(见图 5)。

2.3　“中科院机构名称规范库”的应用服务

(1)精准发现知识资源。“中科院机构名称规范库”建设的初衷是规范机构名称,包括机构的历史性名称、不同来源名称等各种不规范名称,并对每个科研机构赋予唯一标识、规范名称对应关系。其目的在于为知识资源系统提供接口服务，实现各类型知识资源系统中资源发现的准确性和全面性，保证科研机构对回溯知识资源的充分发现。

(2)规范存储知识资源。由于机构名称不规范导致存储知识资源时，属于同一机构的不同机构名称下的知识资源经常会作为独立的资源进行存储。从而导致知识资源存储过于分散，不利于知识资源的整合利用。对机构名称的规范控制可以最大程度的保证知识资源的有效整合，既减少数据库存储的数据冗余,也可提高数据库中知识资源读取的效率。

(3)有效利用知识资源。如今众多的知识资源系统均引入关联数据、语义网等前沿技术和理论。为了使前沿技术和理论得到充分发挥,必须在后台有完整的、准确的规范文档进行支撑。机构名称规范控制还可以有效降低知识资源统计分析的误差。另外,机构名称的规范控制有利于科研机构有效利用知识资源发现潜在合作领域或潜在的研究领域。

3　不足与建议

“中科院机构名称规范库”的建设仍存在许多问题,如:部分机构条目信息不完整,尤其是对于新成立的机构和历史机构缺失信息较多，难以保证机构元数据录入的完整性；目前只考虑所级研究单元和创新单元,没有涉及到研究部室、研究团队等更细小的单元,机构名称规范的覆盖率还有待提高;机构关联关系和变更关系的整理虽有流程可进行机器批处理,但是由于机构网站布局结构不同,在提取关联关系和变更关系时难免有所出入，因此需要人工干预机器处理结果,以保证规范数据的准确性;在进行跨系统平台的 OAI-PMH 元数据收割,由于不同平台的元数据标准方案不同而产生的机构名称的书写和表达有各不相同，从而对采集数据的存储和后期的分析利用造成重重困难；由于同一机构名称存在多种不同的名称标识,机构名称除全称、简称等正式名称外,还存在中英文名称,不同书写习惯名称等;另外还存在机构的隶属关系导致机构名称组合方式不同,机构历史变化导致机构名称迁移等问题。在今后的工作需在以下几个方面予以改进和加强，以进一步实现用户的机构知识库建设需求：①规范控制全面性。以点到面的形式全面扩展,力图包含中国所有科研机构(如高等院校、省级研究所、政府、企业研究院等);②前沿技术的融合。为了迎合大数据时代科研工作对知识资源的需求，将关联数据和语义网等技术应用在机构名称规范控制工作上。③规范标准的制定。包括规范格式、规范标识等内容建设,为今后实现中国机构名称规范文档的国际化共享奠定基

础。④规范控制可持续性建设。机构名称不断变化，科研需求不断提升，这就需要机构名称规范控制工作持续进行。因此，要制定机构名称规范控制的可持续建设方案，并使其得到确实地执行。

参考文献：

[1] S. Michael Malinconico, James A. Ri zzolo. The New York Public Library Automated Book Catalog Subsystem[J].Journal of Library Automation, 1973(6):3-36.

[2] 林明.规范控制的发展历程[J].图书馆工作与研究，2001(5):2-6.

[3] 中文名称规范联合协调委员会网站[EB/OL].[2015-11-09].http://www.cccna.org.

[4] 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(14):47-51.

[5] 孙立杰.中文名称规范的发展与应用研究[J].图书情报工作,2012(1):173-175,239.

[6] HongY,OnBW,LeeD.System Support for Name Authority Control Problemin Digital Libraries:Open DBLP Approach[J].Lecture Notesin Computer Science,2004(3232):134-144.

[7] Researcher ID[EB/OL].[2015-11-08].http://isiwebofknowledge.com/researcherid/.

[8] What is ORCID?[EB/OL].[2015-11-09].http://orcid.org/node/47.

[9] International Standard Name Identifier (ISO 27729)[EB/OL].[2015-11-09].http://www.isni.org/.

[10] The D&B D-U-N-S Numbe[EB/OL].[2015-11-09].http://www.dnb.com/content/dam/english/dnb-data-insight/duns_number_overview_2011.pdf.

作者简介 李慧佳，女，中国科学院兰州文献情报中心馆员；冯建玲，女，中国科学院兰州文献情报中心研究馆员；张秀秀，女，中国科学院兰州文献情报中心馆员；杨丽娜，女，中国科学院兰州文献情报中心助理馆员。

（上接第132页）

[7] 罗卫.电子政务“信息孤岛”新探—基于信息生态的视角[J].情报科学,2013,31(1):31-35.

[8] 杨文士,焦叔斌,张雁,等.管理学原理(第二版)[M].北京:中国人民大学出版社,2004.

[9] 李辉.论协同型政府[D].长春:吉林大学,2010.

[10] 北京市海淀区突发事件应急委员会办公室.平战一体 融合发展—北京市海淀区创新推进应急管理体系建设[J].中国应急管理,2013(10):30-35.

[11] 魏建龙.网格化城市管理的探索与思考——以福州市鼓楼区城市网格化管理为例[J].福州党校学报,2015(3):34-36.

[12] 胡衡华.创新应急管理工作 构建公共安全体系[J].中国应急管理,2015(5):22-24.

[13] 陈鹏.中国社会管理创新体制模式研究——基于四种模式的案例分析[J].北京师范大学学报(社会科学版),2015(4):5-24.

[14] 北京市东城区信息化工作办公室.北京东城区网格化的工作模式、精细化的城市管理[J].信息化建设,2011(9):10-12.

[15] 樊永婷.创新社会服务管理中的城市网格化模式研究[D].呼和浩特:内蒙古师范大学,2013.

[16] 晋中市委讲师团课题组.晋中网格化社会服务管理模式的现状、问题与对策[J].探索与研究,2013(7):41-44.

[17] 赵文明,阮占江.长沙创推网格化社会服务管理[N].法制日报,2013-01-03(4).

作者简介 崔顺爱，女，北京市海淀区城市服务管理指挥中心；刘霆，男，北京市海淀区城市服务管理指挥中心科员；汪柏弟，女，北京大学信息管理系硕士研究生。