Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

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*

□ 杨奕虹 李雅萍 张立丽 / 中国科学技术信息研究所 北京万方数据股份有限公司 北京 100038
林霄剑 / 北京万方数据股份有限公司 北京 100038

摘要：文章总结了当前海量二次文献中机构名称多样性现状及由此产生的问题，介绍了机构多层级词表编制方法，以及机构多层级词表在文献计量评价与机构科研绩效管理中的应用效果。通过应用效果可以看出：机构多层级词表的应用，解决了海量数据中机构名称归一化问题，从而提高了文献检索的查全率、查准率，保证了文献计量结果的准确性；同时，通过多层级词表的应用，可以解决一个机构对其多层级下属机构的科研绩效管理问题。

关键词：机构多层级词表，机构名称规范，机构名称词表，机构名称归一化
DOI：10.3772/j.issn.1673—2286.2013.06.010

引言

在采用文献计量学的方法进行学科评价及机构的科研能力评价过程中，机构名称对于机构特征统计指标是一个非常重要的统计单位[1]。如果需要采用文献计量的方法了解一个机构科研能力的历史变化趋势，其文献计量的数据时间跨度可以达到十年、二十年。在这个时间范围内，由于同一机构可能会经历重组或更名等种种原因，造成了同一机构存在合并名称、曾用名、简称、相似名称等多种表述方式的情况[2]，给统计工作造成一定困难。虽然目前采用“包含”检索已可以解决主要问题，但由于不能同时保证检索结果的查全率和查准率，所以检索统计结果将严重影响评价结论。

在文献主题检索中，为保证检索的查全率和查准率，通常会在文献处理时先根据《汉语主题词表》、《中国分类主题词表》等工具词表进行文献分类标引和主题标引，同时在检索系统后台挂接相应的词表工具；在《中国分类主题词表》中的附录一为“组织机构”词表，但词表中收录的机构名称有限[3]，而且没有机构款目名称之间的“用代关系”。也就是说，到目前为止，机构名称及其机构名称变迁关系的词表国家还没有统一规范，因此需要在文献计量统计工作开展的同时建立一套机构多层级词表，以确保评价数据归属的准确性。

为此，万方数据公司开创性地提出了“机构多层级词表”模型，并投入大量人力物力，通过对海量二次文献的机构名称数据的统计分析，并以机构自身官方公布的下属机构设置为事实依据，建立了《中国机构多层级词表》，包括《中国本科院校多层级词表》、《中国大型集团公司多层级词表》、《中国主要科研机构多层级词表》，希望通过多层级词表，高效解决机构名称归一化问题。

1 机构多层级词表的编制

1.1 机构名称的数据准备

机构名称的数据来源于《万方软件公司期刊信息库》。我们将期刊信息库中的作者机构名称整理，并按机构名称“完全一致”的方式归并，整理后得到552,826个机构名称，这些机构名称涉及17,036,075篇论文。样例见表1。

1.2 机构名称的多样性分析

在上述55万个机构名称中，我

* 本文系国家高科技发展计划（863计划）“云计算一期”重大专项课题“以科技文献为主的搜索引擎研制”子课题（编号：2011AA01A206）成果之一。

表1 期刊机构名称论文数量统计

序号	期刊机构名称	发表论文数量
1	上海交通大学	34,828
2	上海交通大学机械与动力工程学院	8,425
3	上海交通大学管理学院	4,468
4	上海交通大学电子工程系	4,377
5	上海交通大学计算机科学与工程系	3,477
6	上海交通大学自动化系	3,352
7	上海交通大学电气工程系	2,910
8	上海交通大学安泰管理学院	2,888
9	上海交通大学电子信息与电气工程学院	2,762
10	上海交通大学农业与生物学院	2,747
…	……	……
总计	552,826（55万）个机构名称	17,036,075篇论文

们以中国本科院校、中国大型集团公司、中国主要科研院所共4,309家机构为重点，对各类机构名称的多样性进行了分析，总结出机构名称多样性主要呈现以下几种规律：

（1）合并名称

从1992年开始教育部进行高校管理体制改革和布局结构调整，国内数百所高校走向合并。截止到2009年底，教育部已成功将304所高校合并组成125所新规模的综合性大学[2]。例如：1998年，“浙江大学”、“杭州大学”、“浙江农业大学”、“浙江医科大学”合并组建新的“浙江大学”。

（2）曾用名

上述1,700万篇期刊论文，发表时间跨度为1983-2012年。为此，需要了解一个机构的名称变化，以便更加精确地检索。例如：2004年“北京广播学院”更名为“中国传媒大学”。

（3）简称

很多机构都有其常用的简称，例如“北大医院”就是“北京大学第一医院”的简称。

（4）机构独立名称

大学作为学科领域科研的主力机构，每所大学都有其自身特色的院系设置，同时下设各类分校、附属研究所、附属公司产业，其附属研究所或者公司的名称是可以脱离上级机构而独立存在的。例如：“中国药物依赖性研究所”就是“北京大学”的附属研究所，其对外名称既可以是“北京大学中国药物依赖性研究所”，也可以是“中国药物依赖性研究所”。

（5）相似机构名称（非规范名称）

由于期刊论文中的作者单位均为作者个人提供，因此作者个人标注习惯不同会造成机构名称多样性，尤其是数字编号命名的机构名称尤为突出。例如：“704所”是采用阿拉伯数字标识，但也有用大写汉字标识：如“七零四所”，或“七〇四所”。

1.3 机构名称的多层级关系确定

上述同一机构的各类名称，都存在着一定的层级关系。但确定这些层级关系不能依靠简单的推理，而是需要基于客观事实。因此上述4,309家机构其各种机构名称之间的层级关系的确定有两种途径：一是依靠万方数据公司建设的《中国企业公司产品数据库》、《中国高等院校数据库》、《中国科研机构数据库》。上述数据库已建设近20年，积累了大量的机构名称的历史沿革，因此机构的曾用名、简称、合并名称与现在正在使用的机构名称就能产生映射关系；二是查询了解上述机构自身官网公布的院系设置、附属研究机构、附属医院等机构设置信息，通过分析了解最新机构及下属机构的独立名称，确定各名称之间的层级关系。

1.4 机构多层级词表结构

机构多层级词表的结构设计主要是采用叙词表编制模型中的“用”、“代”、“属”、“分”、“族”的思想体系[4]：把现有本科高校、研究院所、集团公司的机构名称作为“一级机构”，也就是“族”的概念；然后依据各机构的官方网站上的信息，对各机构及其下属分支机构进行层级分析，通过建立“属分”关系来表达机构与下属机构的层级关系；将文献中的机构名称作为“机构款目名称”，该机构在网站上正式公布的名称作为本机构的“机构规范名称”，相当于“用”的概念；与该机构相关的“曾用名”、“合并名称”、“简称”、“相似名称”，这四类名称都映射到“机构

规范名称”，相当于“代”的概念；通过建立“用”、“代”关系，将各类非规范的机构名称映射到“机构规范名称”。

通过这种词表模型，将本科高校相关的直属院系、附属研究院所、研究中心、重点实验室、附属医院以及下属独立单位、研究院所及其下属分支机构、集团公司及其下属分公司等机构名称，进行了分层处理，机构的层级将依据本机构的实际情况设置，最多可以扩展到五级。同时，也将同一机构的多种表达方式与“机构规范名称”建立了“用代关系”，解决了机构归一化问题。

1.5 机构多层级词表样例

机构多层级词表款目机构样例如下：

1.6 机构多层级词表编制平台

机构规范平台是元数据机构规范工作中的一个重要环节，同时也是机构词表质量的后台保障。机构规范平台主要是针对机构词表的编辑加工人员而设计的，所以比较注重操作的方便性，一些简单的数据处理，在机构规范平台上即可实现，将技术人员从繁杂的数据处理中解放出来。机构规范平台不仅具有对机构词表进行导入、导出、查重、查错等功能，同时也可直接对机构词表进行编辑、修改，以保证对机构词表进行持续的补充完善。

2 机构多层级词表在文献计量中的应用

2012年万方数据-南京大学“创新力评估”联合实验室推出了《2012年中国高校科技创新力排行榜》，其中“五大核心刊发表论文”是87个评价指标之一。该指标的数据范围是：2001-2011年由中国科学技术信息研究所、中国科学院、中国社会科学院、北京大学、南京大学五大期刊评价单位确定的核心刊并集，共3,112种（其中不包含49种文学类期刊），构成9,961,947篇中文核心期刊评价数据集。在这个数据集里，检索需要评价的高校期刊论文数量，在检索的过程中，需要依据《中国高校机构多层级词表》，检索目标高校的各类名称，这样就能保证目标高校期刊论文数量的查全率。

以“上海交通大学”[5]为例：

通过构建“上海交通大学”多层级机构词表，可以将“上海交通大学”及其简称（“上海交大”）、上海交通大学的合并名称（“上海第二医科大学”、“上海农学院”、“上海海洋水下工程科学研究院”）以及确定为上海交通大学的下属机构，但期刊作者单位并没有标注“上海交通大学”字样的独立名称，如“上海市第一人民医院”期刊论文全部检索出来，形成“上海交通大学”五大核心刊数据集（见表2）。

在上述表中，“上海交通大学”的数据集为116,555篇论文。

“上海交通大学”为规范名称，通过包含检索作者单位字段“上海交通大学”，可以得到91,657篇论文，占“上海交通大学”数据集的78.64%；

“上海交大”为“上海交通大学”的简称，包含检索“上海交大”，检索结果有204篇，占上述数据集的0.18%；

“上海第二医科大学”在2005年7月与“上海交通大学”合并，成为“上海交通大学医学院”，因此我们称“上海第二医科大学”是“上海交通大学”的合并名称，因此在2001-2011年核心刊论文数据集中，标注“上海第二医科大学”的期刊论文需要归入“上海交通大学”。在上表中，“上海海洋水下工程科学研究院”是在2001年3月正式并入“上海交通大学”，但在2003年仍然有一篇论文仅标注“上海海洋水下工程科学研究院”；“上海农学院”在1999年9月正式并入“上海交通大学”，成为“上海交通大学农学院”，已不独立存在，但在2001-2003年，仍然有4篇论文仅标

表2 上海交通大学机构科研产出统计表

上海交通大学	科研产出总量	百分比
规范名称	91,657	78.64%
上海交通大学	91,657	
简称	204	0.18%
上海交大	204	
合并名称	16,870	14.47%
上海第二医科大学	16,865	
上海海洋水下工程科学研究院	1	
上海农学院	4	
下属独立名称	7,824	6.71%
上海市第一人民医院	2,487	
上海市第六人民医院	1,747	
上海市精神卫生中心	1,164	
上海市胸科医院	711	
上海市儿童医院	334	
上海市第九人民医院	176	
上海第六人民医院	137	
上海胸科医院	107	
上海第一人民医院	104	
上海第九人民医院	84	
上海市高血压研究所	68	
……	……	
总计	116,555	

注了“上海农学院”。这类合并名称发表的论文数为16,870篇，占上述数据集的14.47%；

“上海交通大学”下属研究所、医院，其机构名称可以脱离“上海交通大学”而独立存在，我们称之为独立名称。例如：“上海交通大学附属第一人民医院”，同时也可以对外称为“上海市第一人民医院”、“上海第一人民医院”。因此通过检索可以看到，期刊论文中期刊作者单位以“上海市第一人民医院”名义发表2,487篇论文，以不规范的相似名称“上海第一人民医院”发表104篇，因此单独以“上海市第一人民医院”发表的2,591篇论文数量也应该进入“上海交通大学”。这类以独立名称发表的论文总量为7,824篇，占上述数据集的6.71%。

上述统计说明：如果不应用机构多层级词表，仅检索“上海交通大学”，其数据的查全率仅为78.64%。

通过上述样例，可以表明，通过应用机构多层级词表，可以确定一个机构的规范名称，然后将这个机构相关的简称、曾用名、合并名称、相似机构名称、下级独立机构名称等多种形式的机构名称与该机构的规范名称建立映射关系，从而解决了机构名称的归一化，保证了文献计量评价数据的查全率和查准率。

3 机构多层级词表在科研绩效管理中的应用

随着信息服务走向知识服务，机构要素已成为知识导航五大要素之一[6]；同时，大数据概念的普及，“数据说话，让决策有理有据”的管理理念也在各大科研机构中逐渐落实到行动。目前随着各大机构内部信息化系统的建立，其内部科研绩效的量化管理已成为系统功能之一，因此建立机构科研产出等相关知识库[7]已成为重点高校、大型科研院所、集团公司等组织机构的日常所需；但由于高校合并、院所转制、集团公司重组，形成了各类机构名称多样性，因此建立一个特定的机构知识库，需要识别一个机构的多种名称及其附属机构，这需要耗费一定的人力物力。针对上述需求，应用机构多层级词表就可以高效满足机构所需。

仍以上述“上海交通大学”116,555篇论文数据集为例。首先需要对上述数据中的同一篇论文多个作者单位进行分拆处理，然后根据机构多层级词表，对机构进行分级、去重，得到“上海交通大学”二级、三级机构科研产出数据集131781篇论文总量；再次对上述论文集合中同一篇论文不同作者的二级机构进行拆分、分级、去重，得到“上海交通大学”二级机构科研产出数据集为117,618篇论文；用同样的方法对“上海交通大学”的三级机构（以医学院为例）进行拆分、分级、去重，得到“上海交通大学医学院”科研产出数据集为66,200篇论

文。“上海交通大学”二级机构对比统计结果见图1，三级机构（上海交通大学医学院）对比统计结果见图2。

从图1中可以看出：从2001年到2011年11年间，“上海交通大学”在五大期刊评价机构确定的核心期刊中发表的论文总量为117,618篇。

其中，“学院系”里面的“医学院”的科研产出总量最多，为62,556篇，占“上海交通大学”科研产出总量的53.19%；其次是“电子信息与电气工程学院”，科研产出总量为12,122篇，占“上海交通大学”科研产出总量的10.31%。

“研究院所”里面的“微纳科学技术研究院”的科研产出总量最多，为825篇，占“上海交通大学”科研产出总量的0.70%；其次是“高等教育研究院”，科研产出总量为186篇，占“上海交通大学”科研产出总量的0.16%。

“仅以上海交通大学名义发表”的科研产出总量为4,884篇，占“上海交通大学”科研产出总量的4.15%，这部分数据可以根据论文作者的信息细分到“上海交通大学”各下属机构。

“其他机构”的科研产出总量为1,963篇，占“上海交通大学”科研产出总量的1.67%。

在图2中，“上海交通大学医学院”的科研产出总量为66,200篇，从科研绩效的对比统计中很容易看出：

“附属单位”中的“附属瑞金医院”科研产出总量最多，为13,132篇，占“医学院”科研产出总量的19.84%；其次是“附属仁济医院”，科研产出总量为9,599篇，占“医学院”科研产出总量的14.50%。

“院系”中的“基础医学院”的科研产出总量最多，为1,644篇，占“医学院”科研产出总量的2.48%；其次是“公共卫生学院”，科研产出总量为371篇，占“医学院”科研产出总量的0.56%。

“仅以上海交通大学医学院名义发表”的科研产出总量为319篇，占“上海交通大学医学院”科研产出总量的0.48%；“仅以上海第二医科大学名义发表”的科研产出总量为252篇，占“上海交通大学医学院”科研产出总量的0.38%。

“其他机构”的科研产出总量为1,246篇，占“上海交通大学医学院”科研产出总量的1.88%。

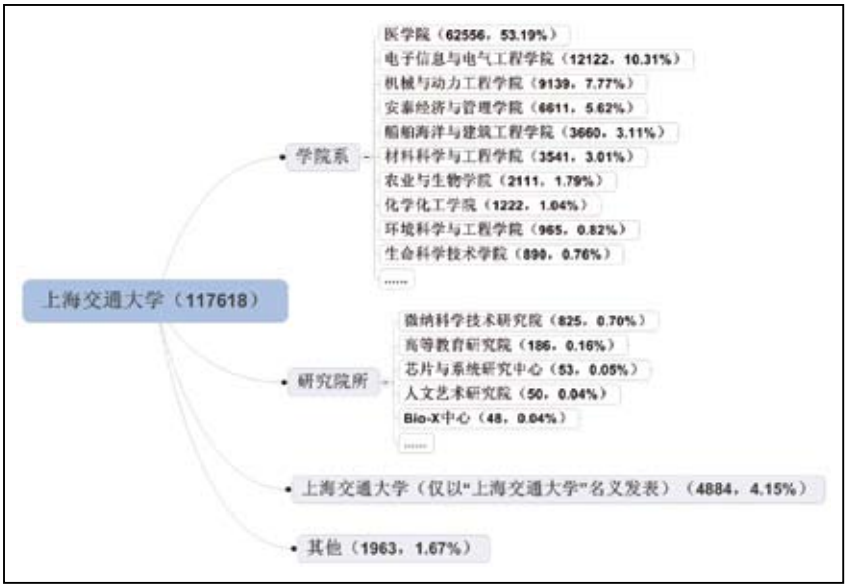

图1 “上海交通大学”二级机构科研绩效对比统计

图2 “上海交通大学医学院”三级机构科研绩效对比统计

4 结论及建议

4.1 机构多层级词表解决了机构名称归一化的问题

通过建立机构多层级工具词表，可以将机构的规范名称、合并名称、曾用名、简称以及同一机构的各种形式的相似名称建立映射

关系，从而解决机构名称归一化问题。再将工具词表与海量文献中的机构数据表进行链接，采用计算机处理加人工干预结合的方法，完成海量文献中机构名称信息的清洗规范工作，因此提高了文献计量结果的准确性，保证了评价结果可信度。

4.2 机构多层级词表可以满足机构对其下属机构科研产出绩效管理需求

机构多层级词表中，不仅将同一机构多种形式的名称与规范名称建立了映射关系，还将同一机构及其下属机构根据客观事实建立了层级关系，将具有层级关系的机构词表与海量文献中的机构数据进行映射处理，可以很容易分别统计出各大高校院系及其附属研究所（医院）、大型科研院所及其下属研究中心、集团及其下属分公司（研究院）等机构的科研产出总量，从而解决了各大机构科研处对本单位及其下属机构科研产出的绩效对比与管理的需求。

4.3 机构多层级词表更广泛地应用

机构多层级词表不仅可以应用在文献计量评价和机构科研产出绩效管理，还可以在建立机构-人名词表中发挥作用。可以通过机构多层级词表中的规范机构名称，建立多层级机构-人名知识库。同时，根据多层级机构-人名知识库可以反过来细分仅以一级机构名义发表的论文：例如，在上述“上海交通大学”多层级绩效管理统计中，如果需要细分以“上海交通大学”名义发表的论文，就可以根据期刊论文的作者名称，利用“上海交通大学”的作者及多层级机构-人名词表，将上述仅标注“上海交通大学”的论文细分到作者所在院系或者研究所，这样就可以更加精确地统计“上海交通大学”各下属机构的科研产出。机构多层级词表与多层级机构-人名词表相互补充，能更好地满足各大型机构对其多层级下属科研绩效精确管理的需求。

4.4 机构多层级词表还需要不断地更新

建立机构多层级词表，其层级和各类名称的变化都是以机构的官方设置及变迁事实为依据的，因此机构多层级词表需要每年保持不断地更新，以确保各机构层级及各种名称之间变化情况的时效性和准确性。

4.5 机构多层级词表需要增加英文名称词表

随着中国科技创新能力不断提升，中国学术机构在国内外期刊发表的英文论文被SCI、EI等期刊论文统计分析数据库收录的数量逐年增多。为了管理好各大学术机构的科研产出，也需要了解各学术机构的英文论文数量，因此需要在机构多层级词表中增加各机构名称的英文名称词表。

4.6 机构多层级词表需要国家投入建设

很显然，机构多层级词表在文献计量评价、下属机构科研产出管理等知识组织工作中起到举足轻重的作用，编制机构多层级词表需要投入大量的人力物力，而且由于机构变化较多，完成的机构多层级词表都需要定期更新，因此建议此项工作由国家投入资金来编制、推广、应用，以便更加深入、更加规范地做好机构知识组织工作。

参考文献

[1] 董琳.学科评价之文献计量数据准备[J].情报理论与实践,2010(6):49-52.
[2] 中国高校紧锣密鼓大合并[OL]. [2010-02-03]. http://edu.ifeng.com/zhuanti/jin30nianlaidaxue/doclist/201002/0203_9450_1536226.shtml.
[3] 刘萍.对《中国分类主题词表》组织机构ā人物主题设置及其计算机标引之己见[J].四川图书馆学报,1998(4):12-13.
[4] 中国科学技术情报研究所.汉语主题词表[M].北京:科学技术文献出版社,1991.
[5] 上海交通大学[OL]. [2013-04-20]. http://www.sjtu.edu.cn/xbdh/yjdh/yx.htm.
[6] 张建聪,吴广印.面向知识导航的机构要素元数据规范及互操作[J].情报学报,2010(1):84-92.
[7] 贾倩,毕经元,王立伟,等.面向大型科研机构的知识管理系统设计[J].现代情报,2012(12):143-148.

作者简介

杨奕虹（1965-），女，中国科学技术信息研究所情报学硕士，研究馆员。研究方向：信息管理、知识组织、数据库建设。
E-mail: yangyh@wanfangdata.com.cn

The Compilation of Multi-Echelon Thesaurus of Organization Names and Its Application in the Document Measurement and Evaluation and in the Management of Achievements in Scientific Researches

Yang Yihong, Li Yaping, Zhang Lili / Institute of Scientific and Technical Information of China; Beijing Wanfang Data Co., Ltd., Beijing, 100038
Lin Xiaojian / Beijing Wanfang Data Co., Ltd., Beijing, 100038

Abstract: The article summarizes the variety in the names of organization in mass secondary document arising from that situation. It also summarizes the compilation of multi-echelon thesaurus of organization names and its application in the document measurement and evaluation as well as in the management of achievements in scientific researches. From the result of such application we can conclude that: the application of organization thesaurus helps to solve the problem of unifying names of organization in mass data, which improved the recall ratio and precision ratio of the document retrievals; Meanwhile, the use of multi-echelon thesaurus enables an organization to solve the issue of managing the achievements of scientific researches within multi-echelon subsidiary organizations.

Keywords: Multi-echelon thesaurus of organization names, Standard of organization names, Thesaurus of organization names, Organization names normalization

（收稿日期：2013-05-09）