Web of Science机构标引典型错误及其维护策略

魏凤萍 何益华 袁青
(华中科技大学图书馆, 武汉 430074)

摘要: Web of Science依据增强机构信息列表对地址进行机构标引, 实现不同机构署名形式的文献按机构归并。本文分析、总结机构标引错误的类型和影响, 探讨机构标引的特点和标引错误的原因; 介绍华中科技大学通过地址检索与机构扩展检索的对比查找误标文献和漏标文献, 实现机构标引维护的方法; 最后, 提出机构标引维护需要更多机构积极参与, 数据库信息处理技术须进一步提高和优化, 机构列表需要机构和数据库公司的协同管理与维护。

关键词: 增强机构信息列表; 机构标引; Web of Science
中图分类号: G354
DOI: 10.3772/j.issn.1673-2286.2018.05.007

机构列表是大型文献数据库中常见的规范化信息, 通过机构列表的规范文档可以判断和确定文献机构归属, 并且作为检索机构文献的重要途径, 其质量将影响检索效率以及统计分析和计量评价的可信度。在实际应用中, 却常因机构列表与文献中标注的机构信息质量或机构列表本身编制中存在错误, 而导致在计量分析和机构学术评价中出现偏差, 引起误判。以Clarivate Analytics为例, 在机构科研评价和学术分析方面广泛采用InCites和ESI两个分析评价工具[1-2], 而InCites和ESI进行分析和评价的基础是WOS涵盖全球5 000多个名称规范的机构信息[3-4], 即增强机构信息列表(Organization-Enhanced List, 以下简称机构列表)。增强机构列表由WOS收集、整理、归并机构文献中不同的名称拼写形式, 选择最准确规范的名称作为首选名称, 其他拼写形式作为名称变体被关联。庞大的机构收录范围以及跨语种、跨文化、跨来源的机构名称表达差异, 比如缩略语、相似机构、作者机构报告错误等, 都会影响WOS的机构列表编制质量。

在已有的研究中, 机构信息列表通常作为机构文献检索途径或分析角度而被广泛研究或应用[5-7], 但尚没有研究WOS机构标引的特点及如何进行维护的文献。本文将基于若干机构标引错误的实例来分析错误表现、影响、原因、维护方法, 以及对策和建议, 希望对提高机构标引质量提供参考, 使WOS、InCites和ESI指标数据能够准确地体现机构学术水平, 更好地用于学术评价和学科分析。

1 机构标引错误的类型和影响

机构标引是对文献地址进行判断并添加恰当的机构首选名称, WOS通过对文献地址的机构标引实现对文献的机构归并。机构标引可能出现两种错误: 一种是误标, 即张冠李戴, 将属于甲机构的地址标引为乙机构; 另一种是漏标, 即本来属于某机构的地址却没有标引该机构。表1列举机构被误标、漏标的5个实例。

对同一篇文献来说, 若机构还有其他标引准确的地址, 一条地址被误标或漏标不会对该机构造成影响, 否则地址的误标或漏标将造成文献的误标或漏标, 进而导致机构扩展检索的误检或漏检。文献(以下用D表示)1中地址[14]的误标对第二军医大学没有影响, 该校参与研究的上海长征医院标引准确, 但是对第四军医大学的误标导致机构扩展检索的误检; D4中地址[128]

表1 机构误标和漏标举例

编 号	入藏号	被引频次	地址标引	分 析
1	000262216900018	2 166	⊟ [14] Mil Med Univ，SW Hosp 3，Chongqing，Peoples R China 增强组织信息的名称 Fourth Military Medical University Third Military Medical University Second Military Medical University	第四军医大学和第二军医大学属误标；WOS机构扩展检索，两所学校误检；InCites和ESI中同时被3所学校统计
2	000353233500012	47	⊟ [2] Jiaotong Univ，Ruijin Hosp，Infect Dis，Shanghai，Peoples R China 增强组织信息的名称 Southwest Jiaotong University	上海交通大学漏标，西南交通大学误标；WOS机构扩展检索，上海交通大学漏检，西南交通大学误检；InCites和ESI中，未被上海交通大学统计，被西南交通大学统计
3	000369024300002	30	⊟ [5] Univ Sci & Technol，Wuhan Natl Lab Optoelect Huazhong，Britton Chance Ctr Biomed Photon，Wuhan 430074，Peoples R China 增强组织信息的名称 Huazhong University of Science & Technology University of Science & Technology of China	中国科学技术大学属误标；WOS机构扩展检索，中国科学技术大学误检；InCites和ESI中同时被两所学校统计
4	000325111200009	1 151	[128] Shanghai 6th Peoples Hosp，Shanghai，Peoples R China ⊞ [129] Beijing Univ，Affiliated Hosp 1，Beijing 100871，Peoples R China [130] Suchow Univ，Affiliated Hosp 2，Suzhou，Peoples R China	[128] 上海交通大学、[130] 苏州大学均漏标；WOS机构扩展检索均漏检；InCites和ESI中均未被两所学校统计
5	000305942200030	157	[3] Univ Sci & Technol China，Sch Life Sci，Hefei 230026，Anhui，Peoples R China [4] Huazhong Univ Sci & Technol，Coll Life Sci & Technol，Wuhan 430074，Hubei，Peoples R China [5] Peking Univ，Coll Chem & Mol Engn，State Key Lab Rare Earth Mat Chem & Applicat，Beijing Natl Lab Mol Sci，Beijing 100871，Peoples R China	中国科学技术大学、华中科技大学和北京大学均漏标；WOS机构扩展检索均被漏检；但InCites和ESI中被3所学校统计

注：被引信息均为2017年12月17日数据。

对上海交通大学的漏标导致机构扩展检索的漏检。

InCites和ESI中，文献没有统计到机构名下或统计到其他机构名下将造成学术成果的遗失或流失；若文献数量较多、被引次数较高，多个指标数据将受到影响。以2017年11月ESI临床医学领域排名数据为例[8]，上海交通大学以总被引频次175 954次排名全球第149名（其高被引论文163篇，篇均被引频次11.2次），排名第148的根特大学总被引频次为176 534次，比上海交通大学多580次。D2和D4的被引频次分别是47次和1 151次，且均属于临床医学领域高被引论文。若两篇文献中上海交通大学被准确标引，那么该校在临床医学领域可能超过根特大学排名，且高被引论文为165篇，篇均被引频次也应高于11.2次。

机构标引直接影响机构扩展检索效率，还可能影响InCites和ESI指标数据的精准度，而机构扩展的检索效率也影响WOS模拟ESI指标数据进行学科预测和分析的精准度[9]。

2 机构标引特征与标引错误原因识别

2.1 机构标引特征

机构是判断文献归属的依据，机构列表是对地址进行机构标引的依据，机构标引的结果反映地址与机构列表匹配的结果。WOS是基于英文内核的检索系统，自动忽略大小写和字符“&”等，此外机构标引还具有

以下特点。

第一，地址中机构名称和二级机构名称均作为判断依据。D3的地址[5]中机构名称为Univ Sci & Technol，二级机构名称为Wuhan Natl Lab Optoelect Huazhong，该地址同时标引华中科技大学和中国科学技术大学。华中科技大学的机构列表没有关联与UNIV SCI TECHNOL完全一致的名称变体，但关联的WUHAN NATL LAB OPTOELECT与地址中二级机构名称的前半部分完全一致；中国科学技术大学的机构列表中恰好关联UNIV SCI TECHNOL，与地址中的机构名称完全一致。

第二，采用基于文本的精确匹配技术，地址与机构列表关联的名称变体在拼写上的细微差别都将导致匹配不成功。D2的地址[2]中机构名称为Jiaotong Univ，二级机构名称为Ruijin Hosp，上海交通大学机构列表关联的名称变体JIAO TONG UNIV与机构名称里的Jiaotong有空格的区别，关联的SHANGHAI RUIJIN HOSP与地址中的二级机构名称Ruijin Hosp前方不一致；D4地址[128]机构名称Shanghai 6th Peoples Hosp中的数字是序数词，该校机构列表关联的名称变体SHANGHAI 6 PEOPLES HOSP中的数字是基数词，虽然是相同的数字，但因为基数词和序数词的不同拼写形式也导致匹配失败。

第三，关联相同名称变体的机构不一定都会被标引。中国科学技术大学和北京科技大学的机构列表都关联BEIJING UNIV SCI TECHNOL，这一拼写形式是北京科技大学英文名称的简称，具有较高的专指度。如图1所示的检索试验说明：近5年有94篇地址为Beijing Univ Sci Technol的SCI论文标引为北京科技大学，但没有论文同时标引中国科学技术大学。

#2	0	AD=（Beijing Univ Sci Technol）AND OG=（University of Science & Technology Beijing）AND OG=（University of Science & Technology of China） 索引=SCI-EXPANDED 时间跨度=最近5年
#1	94	AD=（Beijing Univ Sci Technol）AND OG=（University of Science & Technology Beijing） 索引=SCI-EXPANDED 时间跨度=最近5年

图1 检索试验（检索时间2017年12月17日）

2.2 机构标引错误原因识别

论文地址机构标引错误，即地址被漏标或误标的直接原因是地址与机构列表匹配不成功或匹配错误，而匹配不成功或匹配错误的原因可能是出现以下问题。

第一，原文地址拼写不规范。WOS地址信息的著录可能会出现机构名称表达或英文拼写不规范等错误[10]，如D2中[2]在期刊原文中的地址为Infectious Diseases，Ruijin Hospital，Jiaotong University，Shanghai，China[11]，数据库转录为Jiaotong Univ，Ruijin Hosp，Infect Dis，Shanghai，Peoples R China。机构名称按照原文提取为Jiaotong Univ，该变体检索专指度低，属于作者拼写不规范。

第二，地址信息转录时出错。D1中[14]在期刊原文中的地址为Southwest Hospital of Third Military Medical University，Chongqing，China[12]，而数据库转录为Mil Med Univ，SW Hosp 3，Chongqing，Peoples R China。原文中机构名称“Third Military Medical University”是准确和规范的，但在转录时仅提取“Mil Med Univ”，该名称检索专指度低。

第三，机构列表关联的名称变体不全面。D4的[130]在期刊原文中的地址为Second Affiliated Hospital of Suchow University，Suzhou[13]，被数据库转录为Suchow Univ，Affiliated Hosp 2，Suzhou，Peoples R China，机构名称Suchow Univ提取准确。但苏州大学的机构列表关联有SOOCHOW UNIV和SUZHOW UNIV等相似的拼写，却没有SUCHOW UNIV。

第四，机构列表关联关系不清晰，未能如实反映机构的结构或历史。江苏大学（2001年）的前身江苏理工大学与江苏科技大学（2004年）具有相同的英文拼写形式，即Jiangsu University of Science and Technology[14]，江苏大学的机构列表中没有关联与其前身江苏理工大学相关的变体，而江苏科技大学关联多个与Jiangsu University of Science and Technology相关的变体。根据机构历史沿革，2004年以前的英文名称Jiangsu Univ Sci & Technol应该属于江苏理工大学，即现在的江苏大学。

第五，标引机制出现问题。新入库的文献需要时间加工，机构拼写准确却被漏标通常只是短期问题。但D5发表于2012年，被漏标的几个地址中机构名称拼写准确规范。进一步研究发现多篇类似文献仅通信作者的地址被标引，推测可能是某段时期内数据库标引机制出现问题，仅标引某种类型作者的地址。

3 基于地址检索的机构标引维护

3.1 机构误标和漏标文献的查找方法

WOS开放提交修改意见的通道，若确认论文信息著录错误，公司会补录或修改[15]，对机构标引进行维护就是发现标引错误的论文、提交修改建议。机构标引的误标或漏标直接影响机构扩展检索的误检或漏检，因此可通过地址途径检索与机构扩展检索的对比来查找标引错误的文献。

机构漏标属于本机构的文献没有标引该机构；机构误标是标引的张冠李戴，既可能是其他机构的文献误标为本机构，也可能是本机构文献误标其他机构。漏标文献不能通过机构扩展途径检索到，其检索思路为“地址检索 NOT 机构扩展检索”；其他机构误标本机构的文献能被机构扩展检索到，但地址途径检索不到，其检索思路为“机构扩展检索 NOT 地址检索”；误标其他机构的文献机构署名不规范且没有辨识度，应先找到署名不规范的文献，其检索思路为“署名不规范的文献 NOT 署名规范的文献”，然后逐篇核实、排查。

3.2 机构标引维护实践

使用简称、正常词序、其他词序、医学院与附属医院、实验室等角度编写机构名称相对规范的地址检索式，根据名称中核心字词缺失编写机构名称不规范的地址检索式。以华中科技大学为例，按照误标和漏标文献的检索思路分别检索并保存检索历史，2018年4月28日运行结果如图2所示。#1和#2分别检索名称规范和不规范机构文献；#3可视为完整的地址检索式；#4是机构扩展检索；#5是机构扩展漏检的文献，可能存在机构漏标；#6是地址检索漏检的文献，可能是其他机构的文献误标本机构；#7是本机构署名不规范的文献，可能误标其他机构。

#7	9	#2 NOT #1 索引=SCI-EXPANDED 时间跨度=最近5年
#6	0	#4 NOT #3 索引=SCI-EXPANDED 时间跨度=最近5年
#5	97	#3 NOT #4 索引=SCI-EXPANDED 时间跨度=最近5年
#4	26 693	OG=Huazhong University of Science & Technology 索引=SCI-EXPANDED 时间跨度=最近5年
#3	26 790	#1 OR #2 索引=SCI-EXPANDED 时间跨度=最近5年
#2	26 268	AD=(Univ* Sci* Tech* SAME (Hu*z$ong OR Hu* z$ong)) OR AD=(Univ* Sci* Tech* SAME (430074 OR union OR tongji) SAME (wuhan or hubei)) 索引=SCI-EXPANDED 时间跨度=最近5年
#1	26 781	AD=(cent* china tech* univ*) OR AD=((h* z$ong OR h*z$ong OR h*$hong) un* s*) OR AD=((hu*z$ong OR hu* z$ong) univ* SAME (tongji OR union hosp* OR xiehe hosp* OR 430030 OR 430074)) OR AD=((Hu*z$ong OR Hu* z$ong) S* T*) OR AD=((H*z$ong OR h* z$ong) Univ* Tech*) OR AD=(Hu*z$ong Tech* Sci*) OR AD=(Sci* Tech* Huazhong Univ*) OR AD=(Huaz$ong (Sci* OR Tech*) Univ*) OR AD=(hust SAME (tongji OR wuhan OR hubei)) OR AD=(Huazhong Polytech* Univ OR Univ Hua Zhong Li Gong) OR AD=((ton*j* OR ton* j* OR tong$i) SAME (coll* OR sch* OR med* OR univ* OR hosp*) SAME (w*han OR 430030 OR hubei)) OR AD=((xiehe OR xie he OR union OR li yuan OR liyuan) hosp* SAME (tongji OR wu han OR w*han OR hubei)) OR AD=((Wuhan OR wu han) Nat* High Magnet* Field) OR AD=(WNLO OR Wuhan Nat* Lab*) OR AD=((wu han OR wuhan) SAME Nat* Lab* O*) 索引=SCI-EXPANDED 时间跨度=最近5年

图2 通过地址检索查找漏标和误标文献（检索时间：2018年4月7日）

如#5所示，华中科技大学有97篇SCI论文被机构扩展检索漏检，即机构漏标。其中，7篇发表于2018年，可能稍后会被标引；其他年份发表的文献则需要修改。

如#6所示，地址检索途径漏检文献数量为零。该检索既可以对机构标引进行维护，还可以对地址检索式进行维护。若地址途径存在漏检文献，如果文献不属于本机构，说明是其他机构文献被误标为本机构，需提交修改建议；如果属于本机构，说明出现新的地址表达，应更新地址检索式。

如#7所示，华中科技大学有9篇文献机构署名不规范，其中可能有文献误标其他机构。2017年9月初，检索发现一篇ESI高被引论文（即D3）同时标引华中科技大学和中国科学技术大学。9月4日提出修改意见，9月22日检索发现记录被更正，仅标引华中科技大学。

4 改进机构标引信息质量的对策和建议

4.1 机构标引维护需多机构协同参与

被误标或漏标的文献虽然是少数，但维护工作仅依靠个别人、个别机构肯定不够，需要更多的机构积极参与。从技术难度上来说，查找漏标和误标本机构的文献可以通过地址检索与机构扩展检索的对比，方便快捷；而查找本机构误标其他机构的文献需要构造复杂的检索式和人工逐篇查看，烦琐低效。从共建共享的角度来说，一篇标引错误的文献在一家机构提出修改意见后，可能多家机构、多篇文献都会被修改。

4.2 数据库信息处理技术需进一步提高和优化

数据库公司要进一步提高和优化相关的信息处理技术，以减少机构错误标引，提高标引质量。首先，要优化地址信息转录技术，准确提取机构名称信息，不遗漏机构名称的核心单词、不改变核心单词的顺序，避免名称或拼写的不准确而导致漏标和误标。其次，采用更智能的匹配技术，如忽略基数词和序数词的区别，避免书写习惯的差异而导致的匹配失败；再次，提高机构列表的全面性和准确性，包括使变体名称具有唯一性和专指性、及时关联机构的新名称或新变体、对容易混淆的名称进行限定等；最后，建立自检机制和定期回溯标引机制，通过及时自检和定期回溯标引，及时发现并纠正错误，进行查漏补缺。

4.3 机构列表需要机构和数据库公司协同管理

机构列表是机构标引的依据，机构列表的完整性和准确性直接影响机构标引的质量。对机构标引错误的论文进行维护是“治标”，对机构列表的维护才是“治本”，做好机构标引需要“标本兼治”。机构标引是一项长期而复杂的工程，其管理和维护不是一蹴而就的，需要机构和数据库公司的共同努力、协同工作。

在现阶段，WOS应开设快捷通道接受修改意见和建议，对机构列表中存在的错误能快速反应。而未来，可以借鉴维基百科所采用的协同内容创作和协同信息质量控制技术[16]，对机构用户开放编辑和修改权限，实现机构用户和数据库公司的协同管理与维护。但是数据库对信息质量的要求会更高，需要对操作人员进行培训和资格审查，并制定严格的规则和流程进行制约和规范。

5 结语

本文基于机构漏标和误标的文献实例与相关研究成果，分析、总结WOS机构标引的特点和标引错误的原因。在对华中科技大学文献进行维护的探索和实践中，发现通过地址检索与机构扩展检索的对比能快速查找漏标和误标本机构的文献，但查找本机构误标其他机构文献的效率较低。

英文名称检索专指度不高的机构既难以通过地址途径全面准确地检索机构文献[14]，也难以通过地址检索与机构扩展检索进行对比来查找漏标和误标文献。因此，本文提出的方法仅适用于英文名称检索专指度高的机构，希望本文的探索能引起对机构标引的研究与关注。只有当论文被准确归属到所在机构时，文献计量指标才能准确体现机构学术水平并进行机构科研评价。

参考文献

[1] 董政娥，陈惠兰. 基于ESI和InCites数据库的东华大学学科发展预测 [J]. 东华大学学报（自然科学版），2013，39（5）：689-694.

[2] 张善杰，陈伟炯，李军华. 基于ESI数据库的高校学科发展决策方法及应用研究 [J]. 现代情报，2013，33（2）：32-35.

[3] 新一代InCites™平台 [EB/OL]. [2018-01-15]. https://0e720eb46ecc531acd64-57adaee1b2704b4b47c9f3be14cb4ebf.ssl.cf6.rackcdn.com/wp-content/uploads/2017/10/InCites_fs_20170830_v2.pdf.

[4] Essential Science Indicators℠快速使用指南 [EB/OL]. [2018-01-15]. https://0e720eb46ecc531acd64-57adaee1b2704b4b47c9f3be14cb4ebf.ssl.cf6.rackcdn.com/wp-content/uploads/2017/11/ESI快速使用指南.pdf.

[5] 房文革，王丽君，张红. 基于Web of Science的机构检索方法 [J]. 农业图书情报学刊，2015，27（1）：64-66.

[6] 赵勇，李晨英，韩明杰. 中外高水平涉农高校的学科结构特征比较——基于QS世界大学农业学科排名的科学计量学分析 [J]. 情报杂志，2015，34（5）：92-97.

[7] 邓珮雯. 2008—2012年上海地区胃肠病学与肝病学SCI论文计量学分析 [J]. 胃肠病学，2013，18（10）：600-604.

[8] InCites Essential Science Indicators [EB/OL]. [2018-01-27]. https://esi.incites.thomsonreuters.com.

[9] 管翠中，范爱红，贺维平，等. 学术机构入围ESI前1%学科时间的曲线拟合预测方法研究——以清华大学为例 [J]. 图书情报工作，2016，60（22）：88-93.

[10] 丁海德，庞芳芳，李德成. SCI数据库中地址信息著录差异与错误分析 [J]. 现代情报，2008（4）：173-174，77.

[11] BROUWER W P，XIE Q，SONNEVELD M J，et al. Adding pegylated interferon to entecavir for hepatitis B e antigen-positive chronic hepatitis B：a multicenter randomized trial（ARES study）[J]. Hepatology，2015，61（5）：1512-1522.

[12] CHENG A L，KANG Y K，CHEN Z，et al. Efficacy and safety of sorafenib in patients in the Asia-Pacific region with advanced hepatocellular carcinoma：a phase III randomised，double-blind，placebo-controlled trial [J]. Lancet Oncology，2009，10（1）：25-34.

[13] SCIRICA B M，BHATT D L，BRAUNWALD E，et al. Saxagliptin and cardiovascular outcomes in patients with type 2 diabetes mellitus [J]. New England Journal of Medicine，2013，369：1317-1326.

[14] 梁桂英，袁润. 基于Web of Science数据库的非特异性机构论文检索模式构建 [J]. 情报杂志，2015，34（4）：176-180，175.

[15] 魏健波，周杰. Web of Science数据录入原则及对作者的启示 [J]. 科技管理研究，2016，36（11）：202-204.

[16] 张薇薇. 社群环境下用户协同信息行为研究述评 [J]. 中国图书馆学报，2010，36（4）：90-100.

作者简介

魏凤萍，女，1979年生，硕士，馆员，研究方向：信息检索、学科服务，E-mail：weifp@hust.edu.cn。
何益华，女，1969年生，副研究馆员，研究方向：信息检索、情报分析，E-mail：hyh@hust.edu.cn。
袁青，女，1962年生，硕士，副研究馆员，通信作者，研究方向：图书馆管理与服务，E-mail：yuanq@hust.edu.cn。

Typical Errors and Maintenance of Organization Indexing of Web of Science

WEI FengPing HE YiHua YUAN Qing
(Library of Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: Web of Science indexes organizations to addresses according to the organization-enhanced list, in order to collect the documents of an organization with different institution names. The types and influences of indexing errors are analyzed, and the characteristics of indexing and the causes of errors are discussed. As a case of HUST, the method of maintenance is proposed, locating documents indexed inaccurately or unsuccessfully by contrasting the results of organization enhanced search with those of address search. Finally, some suggestions are putforward: the maintenance needs more participation, the company should improve and optimize the information processing technology, and organization-enhanced list need collaboration management and maintenance of the company and institutions.

Keywords: Organization-Enhanced List; Organization Indexing; Web of Science

（收稿日期：2018-04-07）