

外文会议规范文档构建研究*

杨代庆, 赵捷, 贾明静

(中国科学技术信息研究所, 北京 100038)

摘要: 外文会议的规范控制是对会议文献实现多维度关联聚类揭示的基础, 多来源的外文会议规范文档建设工作需要突破传统以实物为唯一标准的指导思想, 引入互联网信息源作为重要参考内容, 才能实现真正的规范控制。此外, 文章详细探讨了会议规范文档构建过程中工作流程建立、规范规则制定、协作平台功能设计等内容, 为今后相关规范文档的构建提供有益借鉴。

关键词: 外文会议; 规范文档; 编目

中图分类号: G254.3

DOI: 10.3772/j.issn.1673-2286.2015.07.003

1 引言

学术会议文献是指在科学技术领域各种会议上发表的论文, 包括由商业出版社或学(协)会出版的各种连续或非连续出版物以及会议预印本等^[1]。学术会议文献含有大量的最新情报信息, 是了解世界科学技术发展动向、水平和最新成果的主要渠道, 具有很高的学术参考价值^[2]。长期以来, 由于认可程度低、文献获取困难等^[3]多方面原因, 学术会议文献的利用率不高, 与之相关的研究也相对较少。

随着数字技术的迅速发展, 会议文献的可获得性有了明显改善, 被重视程度有所提高。研究表明, 学术会议文献已经成为计算机科学领域的主要引用文献类型^[4], ACM出版论文的参考文献有40%来自会议文献^[5]。特别是, 近年来我国通过IEEE发表的会议论文数量有了显著增长^[6], 在汤森路透会议文献数据库(CPCI)中所占份额也迅速增大^[7]。尽管如此, 由于会议出版具有较强的不规律性, 并且每次会议的内容及主题都可能存在不同, 主办及承办单位变动较为频繁等

原因, 在传统编目过程中, 一直难于对其进行真正意义上的规范控制, 更多地是采取规则性进行规范控制, 这直接导致了会议文献的关联揭示存在严重不足, 难以形成有效的会议导航、机构索引以及分面筛选等, 从而阻碍会议论文的充分利用。因此, 尽快建立多来源的会议名称规范文档, 对会议书目信息进行规范控制和组织, 实现会议文献的导航聚类以及关联服务, 将有利于提升会议文献的发现机会, 提升会议论文的利用效果。

2 问题提出

规范文档一般包括主题规范文档和名称规范文档, 而会议名称及机构规范则属于名称规范的范畴。图书馆界对于名称规范文档的研究大致可以分为两个方向: 一是研究制定建立规范文档的规则; 二是从实践上开展对名称规范文档库的建设。在规则制定方面, 最早可以追溯到1904年由Cutter编制的《字典编目法》第四版, 随后《美国图书馆协会编目规则》、《图书编目分类介绍》等均从不同角度对规范文档的建立和形

* 本研究得到国家社会科学基金项目“机构规范文档结构及构建方式研究”(编号: 15BTQ015)和国家科技图书文献中心项目“外文会议编目数据清理”(编号: 2013XM47)资助。

式进行了阐述。在规范文档规则研究方面最值得一提的是1978年出版的AACR2, 它提供了建立规范文档的规则, 并一直作为最重要的条例影响到图书馆的规范控制工作。但这些规范规则的对象主要是非电子资源, 90年代后期, 随着电子资源的涌现, AACR2等条例已经逐渐不能满足图书馆对于规范及控制工作的要求。面对新形势的挑战, IFLA于1999年推出了新的书目框架模型FRBR, 随后又分别推出FRAD和FRSAD对FRBR中的名称规范以及主题规范进行统一规定和详细阐述。最终, IFLA将FRBR、FRAD、FRSAD进行融合后, 于2010年正式推出新一代的资源编目规则RDA。在规范名称文档知识库建设方面, 早在1909年美国国会图书馆就开始“美国国会图书馆标题表”

(LCSH)的编制, LCSH中包含部分名称性标题, 但对于人名/机构/会议名称的收录量较少, 并且没有对同名异义的名称或异名同义的名称进行消歧, 难以直接用于名称规范的控制。美国国会图书馆牵头的“名称规范合作计划”(NACO), 通过成员机构的合作共建方式对人名、机构名、地名、统一题名等建立规范文档。日本国立图书馆从1964年开始构建主题及名称规范文档, 目前规范记录超过100万条^[8], 并于2012年加入“国际虚拟规范文档”项目(VIAF)。此外, 德国、挪威等国家也都构建了相应的名称规范文档。我国的名称规范文档构建主要体现在中文名称规范知识库上, 如国家图书馆、CALIS联机编目中心、北京地区高校图书馆分别建立的名称规范库以及香港中文名称规范数据库(HKCAN)。

近年来, 名称规范文档除作为规范控制外, 在语义网构建方面也起到重要作用, 一些重要的名称规范文档已经发布为关联数据, 并通过与各种外部异构数据集之间的关联, 正在成为数据网络(Web of Data)的重要组成部分, 例如: VIAF发布的三元组已经超过2亿条, 并且与Dbpedia数据集建立了1万条关联信息, 与其他数据集建立了4百万条关联信息^[9]。

构建多来源的会议规范文档, 对会议文献进行控制, 进而建立会议导航等知识聚类服务, 以及实现会议活动识别消歧以及关联发布具有重要意义。尽管国内外已经建立了较多的名称规范文档, 但以人名、机构名、以及题名等方面的内容居多, 会议名称规范文档则分散在各种相关规范文档之中, 没有形成集中的规范文档库。此外, 图书馆在会议文献的揭示过程中, 由于会议名称以及主承办机构的不稳定性等原因, 基本上没有

图书馆能够对多来源的会议录形成真正语义上的规范控制, 进而实现完整的会议文献多维度导航, 而仅有部分出版社及学协会对其自身出版或主办的会议形成了真正的导航。因此, 构建多来源的会议规范文档, 并形成稳定的工作流程及更新机制, 已经成为当前图书馆外文会议文献组织揭示的重要任务。

国家科技图书文献中心(NSTL)是国内订购外文文献较多的信息服务机构, 每年订购外文会议文献约5000种, 涉及学协会及出版社约1000多家。为了提高外文会议文献的使用成效, 在编目工作中需要对多来源的会议录文献构建会议名称及会议机构规范文档, 对会议文献进行语义层面的规范控制。在构建过程中, 如何确定会议名称规范库的元数据、名称消歧规则、以及工作平台构建, 是会议规范文档构建要解决的重要问题。此外, 如何形成规范文档的维护流程及机制, 实现规范文档的长期可靠维护也是会议规范文档构建的重要内容。

3 会议规范文档元数据标准及规范规则

3.1 元数据标准

会议规范文档构建, 首要工作是对编目规则进行修订, 并制定规范文档的元数据标准。修订过程中参考了AACR2以及RDA, 其中比较重要的元数据项有: 会议规范名称、会议名称、会议录正题名、其他题名(可以是缩略题名、识别题名或变异题名)、责任者名称、责任者规范名称。会议名称、会议录正题名、其他题名以及责任者名称等元数据项, 根据RDA条款的设计原则按照会议录上的会议名称进行原样著录, 不直接更正会议录实物中的错误; 会议规范名称根据查询会议录、会议官网以及VIAF等规范文档库后, 经过名称主干提取等方式形成同一连续会议的统一名称; 责任者规范名称根据查询会议录、会议官方网站以及相关学协会网站后, 将责任者规范为一级机构名称。

3.2 元数据规则

根据元数据标准的要求, 在规范文档建立过程中采用相应的规则对现有元数据进行规范处理, 较为重要规则有:

- 1) 对必备元数据的缺失项补充著录

a.会议名称(711 12 \$a)与会议时间(711 12 \$f)参照会议录实物进行补录,实物缺失相关字段的,通过互联网查找权威网站进行参考补录。

b.主承办机构(711 02 \$a)与责任方式代码(711 02 \$4)参照会议录实物进行补录,实物缺失相关字段的,通过互联网查找权威网站进行参考补录,若最终无法补录的书目,将被归结为非会议文献。

2) 元数据项的规范整理

a.会议名称(711 12 \$a)不允许含有届次、年代、文集名、卷、会议简称等信息,著录内容要包含有会议字样。多个会议名称对应同一册会议录时,要分别著录在多个711(12)字段,默认会议时间、地点相同;多个会议名称多个会议届次、会议时间、会议地点的情况,要分别著录在多个711(12)字段,将各自相应正确的届次、时间、地点匹配准确著录。

例: 711(12) \$a10thInternational Congress on Noise as a Public Health Problem 2011: Proceedings of the Institute of Acoustics Volume 33 Pt.3

修改后为: 711(12) \$a International Congress on Noise as a Public Health Problem

b.会议名称前有定冠词/冠词的不要著录,会议名称中带有“&”,要换成“and”

例: The American Ceramic Society Annual Meeting

修改后为: American Ceramic Society Annual Meeting

c.会议名称按全称著录,机构名称按全称展开,不许著录机构简称、会议字样缩写等。

例: NACE International Conference

修改后为: National Association of Colleges and Employers International Conference

d.团体会议名称常用两种表达形式:团体名称+会议字样或会议字样 of 团体名称,以第二条规范名称采取的格式为主。

例: Annual meeting of the Association for Computational Linguistics

规范名称为: Annual meeting of the Association for Computational Linguistics

第二条数据: Association for Computational Linguistics Annual meeting

规范名称为: Annual meeting of the Association for Computational Linguistics

e.主承办机构名称

会议主承办机构自身正式的官方公开的称谓,应视为其规范名称,内容包括主办机构全称与其固有的名称缩写,除非机构名称本身带有定冠词外,规范名称要去掉定冠词,机构名称中带有“&”,要换成“and”。会议主办者规范名称可通过以下具体途径获得:从会议文献版权页中获得;从会议文献出版物关于会议情况介绍中获得;从发布会议信息的官方网站中获得。

f.会议简称(711 12 \$c)著录在711 12 \$c里,有则必备,会议简称不包含会议时间、会议届次等相关信息,如果原始数据中将会议全称与简称分别著录两个711 12字段的情况,处理办法为将简称的711 12字段删除,并将简称添加到全称的711 12字段的\$c中。

g.主承办机构简称(711 02 \$c),主承办机构简称著录在711 02 \$c里,有则必备。

4 会议规范文档工作流程与平台构建

4.1 工作流程

会议名称规范文档构建是一个动态过程,需要嵌入到日常编目工作当中,不断实时进行更新维护,并且调整传统的编目工作流程和习惯才能起到好的作用和效果。会议名称规范文档构建过程中,建立的工作流程主要包括三个部分:根据待规范名称采用相似性算法进行推荐、人工通过模糊检索方式进行判断规范、手动建立规范名称。

具体流程为:编目人员输入待规范控制的初始名称后,系统根据相似性算法从规范文档库进行自动推荐,若编目人员确认,则将初始名称与推荐的规范名称建立关联;若在规范文档库中推荐条目不正确,编目人员可以通过关键词模糊查询等方式,在规范文档库中进行查找,以确认规范名称;若待规范的初始名称在规范库中不能发现与之相对应的规范名称,则编目人员可以通过会议官网、VIAF等规范库等互联网资源进行查找后,人工建立新规范条目,并与初始名称建立关联。

4.2 平台主要功能

整个会议规范文档的建设工作由NSTL各成员馆在协同工作平台上统一完成。平台主要功能包括对编目数据中的会议名称以及主办单位进行名称规范,并根

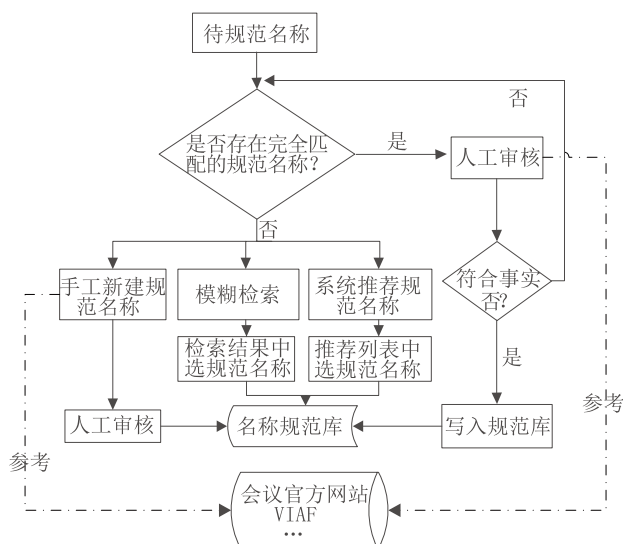


图1 会议规范文档工作流程

据会议的层级建立父子关系。平台架构采用客户-服务器模式，数据统一保存在服务器中，各参建单位在本地安装的客户端对分配的任务进行工作，结果实时保存到服务器。

4.2.1 名称规范功能

对进行规范控制的会议名称，系统首先到“会议规范文档库”中查找与待规范会议名称完全匹配的记录，如果存在，则显示在列表的规范名称栏中，进行人工确认审核。如果审核不能通过，则由专家统一处理；不能完全匹配的会议名称，系统自动在“系统推荐”选项卡的列表中根据相似性算法自动推荐规范名称列表。系统推荐的规范名称是系统根据已有的“会议规范文档库”，查找与待规范会议名称相似的记录，结果按匹配度高低进行排序；系统提供对“会议规范文档库”模糊检索功能，检索的结果可以作为规范名称的参考；系统提供手动创建新规范记录功能，手工录入该会议的规范会议名称及简称。

4.2.2 相似度计算推荐

在会议名称和机构名称规范中，为了提高效率，在系统中增加了名称相似度计算推荐的功能。主要采用的方法是通过计算基于单词词干编辑距离与计算基于单词词干字符串余弦相似度方法。当基于单词词干的编辑距离与名称字符串单词个数的比值大于20%时，优先



图2 协同工作平台中规范控制界面

采用单词词干字符串余弦相似度排序，否则优先采用编辑距离排序。

1) 预处理

对待规范名称字符串进行去停用词处理，停用词主要包括常用冠词、介词、标点符号、以及Conference、Symposium、Book、Society等对于规范会议名及机构名意义不大的实词。去掉停用词后的字符串，采用了Porter算法进行词干提取，并形成名称词干字符串。例如：Society of Automotive Engineers of Japan Annual Congress 在经过去停用词及取词干处理后，转变为字符串：automot engine japan annual congress。

2) 基于词干编辑距离相似度计算

基于词干编辑距离的Distance (S,T) 的计算方法为：S表示待规范名称字符串形成的词干串，T表示规范库中规范名称字符串形成的词干串，Distance表示两个词干字符串之间的编辑距离。S=(S₁...S_i)，0≤i≤m，m表示S字符串中包含的词干个数；T=(t₁...t_j)，0≤j≤n，n表示T字符串中包含的词干个数。

当 i=0, 且 j=0 时， Distance_{i,j}=0

当 i>0 或 j>0 时， Distance_{i,j}=

$$\text{Min} \left(\begin{array}{l} \text{Distance}_{i-1,j-1} + N \text{ (若 } s_i = t_j \text{ 则 } N=0, \text{ 否则 } N=1) \\ \text{Distance}_{i-1,j} + 1 \\ \text{Distance}_{i,j-1} + 1 \\ \text{Distance}_{i-2,j-2} + 1 \text{ (当且仅当 } s_{i-1} s_i = t_{j-1} t_j) \end{array} \right)$$

上述计算包含删除、插入、替换、交换四种操作，该算法是从两个词干串的左边开始基于词干进行比较，并记录下已比较过的词干串编辑距离，然后进一步得到下一个词干位置时的编辑距离。矩阵Distance_{i,j}从Distance_{0,0}开始逐行逐列计算编辑距离，最终计算出基于词干的S和T的编辑距离。

3) 基于词干余弦相似度计算

待规范名称词干串S与规范名称词干串T的定义与前面相同, S与T的余弦相似度计算方法描述为: 对S和T合并去重后的n个词干, 建立n维文本向量; 分别统计S与T中词干的词频, 并根据词频分别形成对应的两个n维向量A,B; 根据余弦夹角公式计算相似度。

4.2.3 建立会议父子关系

在会议父子关系构建时, 对于同一会议录文集的所有会议, 只创建父与子两级关系, 不再建立子会议的子会议, 并同时以关系树形式显示当前会议已经存在的父子关系, 供编目人员参考(图3)建立规范会议父子关系时, 不检查规范会议父子关系是否存在其它矛盾的关系, 将统一批量由专家审核和修订父子关系的正确性。



图3 协同工作平台父子关系构建

5 推进外文会议规范文档建设的对策

外文会议规范文档的建设是一件工程量大, 需要高度协同的实践性工作, 常常需要解决各种新问题, 例如: 如何区分具有相同规范会议名称的不同会议; 如何统一不同编目人员对标准尺度的一致性认识等。这其中相当部分问题的解决都有赖于目前的工作流程和方法的改变, 以及提升编目人员的认识及能力, 要真正形成适应数字信息环境的图书馆资源揭示与描述新工作环境。

1) 规范文档的构建与维护需要新的工作流程

传统的印本会议录编目工作流程, 注重以实物为准, 不重视甚至不著录会议之间的层级关系, 很少将互联网作为重要的信息来源进行信息著录。编目工作与作为前序的采访工作联系不紧密, 存在较为严重的脱节现象, 甚至编目人员看不到订购信息, 订购人员做的订购信息规范和维护, 无法直接被用到编目数据中, 容易造成订购信息与编目规范数据不一致的情况。数字环境下的会议规范文档建设要求必须改变现有的工作流程, 才能做到真正意义上的名称规范, 而非局限于形式

上的规范(例如: 不同主题的不同会议, 这就要求从多种渠道及方法进行综合判断, 从而得到最终结果)。此外, 规范文档建设与利用应嵌入到信息资源生命周期的全流程之中, 从采访、编目、加工、服务等各个环节都要能共享、利用规范文档信息, 并实现畅通的沟通反馈渠道。

2) 应积极与国内外机构合作开展规范文档库建设

当前, 我国图书馆在规范文档建设中取得了一定成绩, 但总体来说, 还处于初级阶段, 而发达国家在规范文档建设理论和实践上均已取得了较为丰硕成果, 积极与国外机构合作, 可以实现信息共享, 为我国的规范文档建设工作提供借鉴, 达到事半功倍的效果。规范文档建设工作是一项艰巨的工程, 单靠一个或几个图书馆难以实现可持续性建设及维护, 走合作共建道路, 将是规范文档建设的发展方向。在合作共建过程中, 与国外机构的合作固然重要, 但更要加强国内图书馆之间的共建共享, 才是实现规范文档建设的根本之路。

3) 编目人员观念与能力需要全面提升

新形势下, 资源揭示对于图书馆编目工作提出了新要求, RDA颠覆了传统AACR2的编目理念和指导思想。RDA不仅要求编目人员懂得标题表和传统分类体系, 还要掌握超文本(HTML)及扩展标记语言(XML)。随着编目工具的不断更新, 联合编目和编目外包日趋增加, 编目人员习惯的传统编目流程也不断更新, 编目员要积极改变心态, 积极走出编目小圈子, 与其他工作部门加强合作, 并且积极参加培训和学习, 加强现有知识的更新, 提高自身专业技术水平以及信息获取能力, 才能跟上数字时代的发展步伐。

参考文献

- [1] 朱江, 陈漪红. 科技会议文献管理与服务现状及改进对策[J]. 四川图书馆学报, 2005(4): 40-43.
- [2] 袁军鹏. 科学计量学高级教程[M]. 北京: 科学技术文献出版社, 2010.
- [3] Godin B. Measuring knowledge flows between countries: the use of scientific meeting data [J]. Scientometrics, 1998, 42(3): 313-323.
- [4] Bar-Ilan J. Web of science with the conference proceedings citation indexes: the case of computer science [J]. Scientometrics, 2010, 83(3): 809-824.
- [5] Wainer J., Oliveira H.P., and Anido R. Patterns of bibliographic references in the ACM published papers [J]. Information Processing and Management, 2011, 47(1): 135-142.

- [6] Shirakawa N., Furukawa T., Nomura M., and Okuwada K. Global competition and technological transition in electrical, electronic, information and communication engineering: quantitative analysis of periodicals and conference proceedings of the IEEE [J]. Scientometrics, 2012, 91(3): 895-910.
- [7] Wolfgang Glänzel, Balázs Schlemmer, András Schubert and Bart Thijs. Proceedings literature as additional data source for bibliometric analysis [J]. Scientometrics, 2006, 68(3): 457-473.
- [8] Web-NDL-Authorities-Report[EB/OL].[2015-06-30].<http://www.oclc.org/content/dam/oclc/events/2013/viaf-council/Web-NDL-Authorities-Report.pptx>.
- [9] VIAF: The Virtual International Authority File[EB/OL].[2015-06-30].<http://datahub.io/dataset/viaf>.

作者简介

杨代庆, 男, 1975年生, 中国科学技术信息研究所高级工程师, 研究方向: 信息资源建设。

赵捷, 男, 1959年生, 中国科学技术信息研究所研究馆员, 研究方向: 信息管理系统。

贾明静, 女, 1985年生, 中国科学技术信息研究所馆员, 研究方向: 信息资源组织。

Research of Foreign Conference Name Authority File Construction

YANG DaiQing, ZHAO Jie, JIA MingJing

(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: The authority control conference name is the basis of the realization of proceedings navigation. Foreign conference name authority file construction must break through traditional catalogue ideology, and take Internet as an important reference. This paper guides the work flow, regulation and system platform of foreign conference name authority file construction in detail.

Keywords: Foreign Conference; Authority File; Catalogue

(收稿日期: 2015-06-30; 编辑: 雷雪)

■ 书 讯 ■

《汉语主题词表》(工程技术卷)

《汉语主题词表》自1980年问世以后, 经1991年进行自然科学版修订, 在我国图书情报界发挥了应有的作用, 曾经获得了国家科学技术进步二等奖。为了适应网络环境下知识组织与数据处理的需要, 2009年由科学技术信息研究所主持, 并联合全国图书情报界相关机构, 完成了《汉语主题词表(工程技术卷)》的重新编制工作。

全书共收录优选词19.6万条, 非优选词16.4万条, 等同率0.84。在体系结构、词汇术语、词间关系等方面进行改进创新。为了方便工程技术领域不同专业用户使用, 《汉语主题词表》(工程技术卷)按专业分13个分册出版, 同时建立《汉语主题词表》服务系统, 提供在线概念检索和辅助标引服务, 通过可视化技术展示各类概念关系, 是图书馆、档案馆、出版社、期刊杂志社、文献信息中心等专业工作者及科研、教育及工程技术领域人员必备的参考书。

《汉语主题词表(工程技术卷)》已于2014年由科学技术文献出版社出版, 全书2300余万字, 总定价3880元, 可分册购买。