国家图书馆西文规范数据库更新机制述略

张丽娟
（国家图书馆，北京 100081）

摘要：规范控制工作是编目的重要环节，是实现目录功能的主要途径。规范系统的更新维护是保持规范控制工作生命力的必备保障。经过几十年的建设，国家图书馆西文规范控制业已形成比较完整的体系，特别是在规范数据库的更新维护方面积累了一定的经验。以国家图书馆西文规范数据库的更新机制为切入点，介绍规范数据库定期更新的方法和路径，进而总结规范数据发生新增、修改和删除的各类更新情况以及具体原因。在总结经验的基础上，进一步思考西文规范数据库在更新机制方面存在的问题，对未来工作提出具体思路。

关键词：国家图书馆；西文编目；规范控制；规范数据更新
中图分类号：G254.36；G259.251
DOI：10.3772/j.issn.1673-2286.2019.06.008

1 规范控制概述

规范控制，又称权威控制，是为确保文献信息资源检索点的唯一性和一致性，而建立、维护、使用和评估规范记录（authority record）和规范文档（authority file）的工作过程[1]。

规范控制是编目工作不可或缺的重要环节，是书目系统先进性和完整性的具体体现，是实现目录功能的主要途径。规范控制的作用可归纳为：首先，规范控制能确定统一的检索点形式，汇集质同形异的检索点，从而提高检索的查全率；其次，规范控制能确定唯一的检索点形式，区别形同质异的检索点，从而提高检索的查准率；最后，规范控制能在相关检索点形式之间建立一种逻辑关系，通过参照系统予以揭示，从而将用户从检索点的非规范形式指引到规范形式或相关检索点，最终获得理想的检索结果，即提高检索的便捷性，起到导航作用[2]。

规范控制工作的开展既需要深厚的编目思想的指导，又需要先进的计算机和网络技术的支撑，内涵丰富，环节复杂。1985年，美国伊利诺依州立大学的Burger[3]出版了有关规范控制工作的专著，他按照流程将规范工作的内容归纳为规范记录的创建、规范文档的聚合、规范系统的建立、维护和评估五个环节。这些环节中的核心是规范文档，规范控制工作都是围绕规范文档的建设和利用开展的。规范文档是指规范记录的集合，是将受控检索点的规范形式、变异形式及说明信息按照一定的次序排列而成的统一管理和维护检索点、对书目文档实施规范控制的工具。随着信息技术的飞速发展，规范文档从传统的缩微平片等载体形式转变为电子化的数据库形式。

规范文档是一个生长的有机体，随时保持更新的状态，包括新增记录、修改记录和删除记录3种更新类型。在新增记录方面，如果馆藏中出现新的实体，则需要为其创建新的规范记录。如果以往规范记录由于不能区分而共用一条规范记录，则在获得足够区分信息时就可从共用记录中分离出新规范记录。在修改记录方面，如果规范检索点形式、变异形式或参照说明等发生变化，则需要对规范记录进行修改。或者书目机构获得了更多的限定信息，则可对规范记录进行补充说明。在删除记录方面，如果一条规范记录的规范检索点被废止，则需要删除相应的规范记录。或者由于规范数据库是合作项目共建，如果发现重复记录则予以删除。上述更新类型都是针对规范文档本身的。除此之外，更新还包括另一层含义，即将规范记录的更新变化体现到书目

记录中与之相连的规范检索点中去。本文主要涉及规范文档自身的更新。

国家图书馆（以下简称“国图”）西文规范控制建设始于20世纪80年代，经过30多年的努力，西文规范控制业已形成较完善的体系，不仅在系统内实现了规范记录对书目检索点的有效控制，还自行研发了应用程序，较好地解决了规范数据库自身的更新问题。但是，由于国图西文规范控制属于引用模式，通过直接购买国外成熟的规范文档和相应更新文件用于本地规范控制，因此在规范控制过程中存在一些现实问题。本文以介绍国图西文规范控制更新工作的开展为切入点，总结数据库更新的各类情况，进而思考完全引进模式所存在的问题，提出新的工作思路。

2 更新机制

2.1 概况

国图西文规范控制工作起步较早，但比较系统地开展西文规范控制工作则归功于2003年引进的Aleph 500图书馆集成管理系统。规范控制的核心是规范数据库的建设，国图西文规范控制采用直接引用模式，即引进了美国国会图书馆（Library of Congress，LC）的名称规范文档（LC Name Authority File，LCNAF）和主题规范文档（LC Subject Authority File，LCSAF），并将其装载至Aleph 500系统，通过系统功能与书目数据相连，实现对书目检索点的有效控制。为保持规范数据库的活力，国图还配套引进了LC规范数据库的周更新文件，用于对规范数据的维护。

系统使用初期，为确保系统安全，周更新文件无法即时对规范数据库进行更新，国图只能暂时采用集中更新的方式，将更新文件按照时间的先后顺序合并，再用合并后的最终文件对规范数据库进行整体更新。LC的规范数据每年有二三十万的增长量，而集中更新每三年才进行一次，西文规范数据库的时效性严重滞后。2010年，国图信息技术部门研发了专门的更新文件装载程序，可通过外部应用软件直接将更新文件中的规范记录灌装至Aleph 500系统[4]，至此终于实现西文规范数据库与LC规范数据库的同步更新。

目前西文名称规范数据已达10 610 526条，主题规范数据量已达433 018条。2010年1月—2018年12月，国图共完成名称和主题规范数据库更新各469期，名称规范数据库更新记录6 579 610条，主题规范数据库更新记录153 912条，两个数据库记录新增、修改、删除3种情况的数据更新量如图1和图2所示。

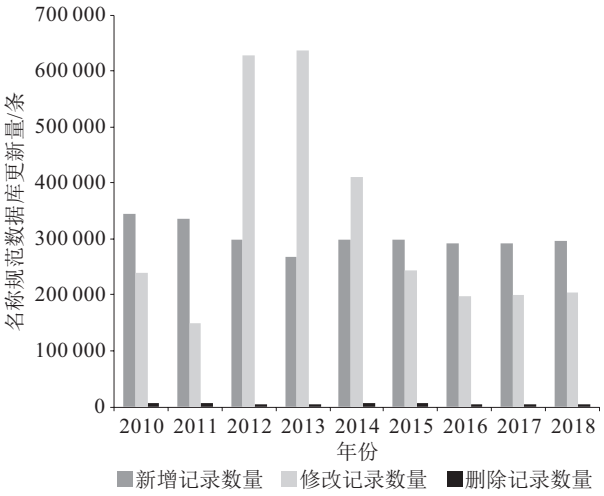

图1 2010—2018年名称规范数据库各年更新量对比图

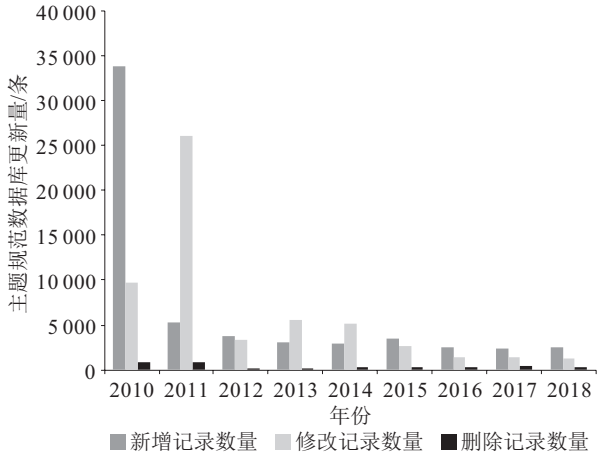

图2 2010—2018年主题规范数据库各年更新量对比图

由此可见，名称规范数据库的数据总量和更新量都远远高于主题规范数据库。原因在于，名称规范数据库主要由LC领衔的名称规范合作项目（Name Authority Cooperative Program，NACO）成员馆基于增加馆藏的情况予以建设，由于馆藏数量迅猛增长，所以名称规范记录的数量也随之大幅增长。而主题规范数据库是基于美国国会图书馆标题表（Library of Congress Subject Headings，LCSH）的内容，通过主题规范合作项目（Subject Authority Cooperative Program，SACO）参与者以提议的方式并由LC政策专家审核批准后才能用于更新记录，所以主题规范记录的增长十分有限。在新增、修改、删除3种更新类型中，两个规范数据库都是删除记录的情况最少，这主要得益于LC和

相关参与者高质量的工作及稳定的规则体系。名称规范记录每年新增记录的数量基本保持在30万条上下，比较平稳。虽然LC在2013年3月31日开始实行资源描述与检索（Resource Description and Access，RDA）规则，但是LC的RDA化进程不是一刀切的，而是从2008年完成RDA测试之后逐步开始的，到“RDA切换日”这天，RDA数据的比例达到100%。LC实现RDA本地化之后，规范记录也没有迅猛增长，这是因为新规则的启用引发了对历史数据的修改，但并不会造成实体数量的增加，因此，在RDA全面实施的2012年和2013年，名称规范记录的修改量达到高峰，但是增长量保持稳定。

2.2 工作流程

LCNAF和LCSAF周一至周六每天更新，内容包括LC编目员在前一天创建和修改的记录，以及由NACO参与者提供的记录，而上一周的删除记录在每个周末统一删除。LC政策专家进行审核通过的需更新的主题记录每周进行更新。LC将这些更新文件按周制作成LC规范文档的周更新文件，供其他机构下载使用。2010年国图在完成LC规范数据装载软件测试后，也制定了更新工作规范，并设计了更新工作流程。

首先由专人负责将这些周更新文件定期从LC提供的FTP地址上下载至本地，并上传至指定的FTP进行长期保存。LC规范文档的周更新文件为ISO 2709数据格式，以卷、期组合的方式命名，包括5种类型的文件，分别是XML文件、report文件、property list文件、records文件、UTF8文件。这些文件除可用于对本地规范记录的更新外，还包含更新量汇总，每条被更新记录的完整记录信息等，对于了解每期更新动态，减少操作失误，严格保证数据库更新操作的安全，具有非常重要的意义。

随后由专人通过专门研发的装载软件对Aleph 500系统中的本地规范数据库进行更新。装载软件基于Web的B/S模式，将功能实现的核心部分集中到服务器上，仅需在IE浏览器上输入相应的地址即可方便快捷地使用该软件，更新只需五个步骤：确定要更新的目标库，即是对名称库还是主题库进行更新；将LC原始规范数据文件上传到Aleph 500服务器，并对该文件进行第一步的转换，即加回车换行符到文件中，使其分行；将上一步生成的已分行MARC21格式文件转换为Sequence格式，输入的文件为上一步输出的文件；将生成的Sequence文件转换字符集为UTF；将生成的UTF文件中的数据更新到相应的西文规范数据库。最后，将上述各步骤生成的文件下载保存，将第五步生成的文件中实际更新的记录总量与LC更新文件中记录更新量对比是否一致，数量相同，则从本期更新中随机抽取有代表性的规范记录，在本地规范数据库中查看这些记录的更新状态及更新时间，核对无误后，更新人员认真填写更新日志。

规范数据库更新是一项严谨的工作，一旦误操作就可能造成规范数据库的混乱，纠正这些错误将会耗费更新人员大量的时间和精力。因此，由专人对整个操作进行监督及文档管理至关重要。

2.3 更新规则

更新文件中的规范记录如何与目标库进行匹配是准确完成更新的重要前提。在LC规范记录中，为每条记录都分配了唯一的记录控制号，同时记录在001字段和010字段。两个字段的区别在于，如果规范记录进行了本地化修改，则001字段所记录的LC控制号（Library of Congress Control Number，LCCN）将被本地控制号所取代，但是010字段所记录的LCCN在任何情况下都保持不变。此外，本地修改之后，规范记录中还会增加一个表示操作员代码的字段“OWN”。在开展更新工作时，主要依靠010字段LCCN来匹配数据，同时兼顾本地修改的情况。具体更新规则如下。

对于LC更新文件中的新增记录，直接装入目标数据库。对于LC更新文件中的修改记录，用目标库010字段LCCN进行匹配，如果目标库中无同号记录，则作为新增记录直接装入；如果数据库中已有同号记录，且未进行过本地修改，则用更新文件中的修改记录覆盖库中的同号记录。对于更新文件中的删除记录，用目标库的010字段LCCN进行匹配，如果数据库中无同号记录，则该记录不必写入；如果数据库中已有同号记录，且未进行过本地修改，则用删除记录覆盖数据库中同号记录。对于编目员已修改而无法进行同号覆盖的LC更新文件中的规范记录，装入临时库暂存，装载时同样先用010字段LCCN进行匹配，如果临时库中无同号记录，则直接写入；如果临时库中已有同号记录，则覆盖。

3 更新情况分析

规范数据库发生日常更新的原因很多。资源种类的

丰富、出版方式的变化、版本形式的多样化等使馆藏资源与日俱增。在这些增加的馆藏中可能会出现新的责任者、新的作品或者原有作品的衍生品，这些都需要在编目时为其构建新的规范检索点。此外，编目员从这些新馆藏资源中可能获取到更多的有用信息用于优化旧的规范记录。再则，编目界日新月异，新规则层出不穷，规则变化会导致规范数据发生批量更新。如RDA取消了检索点选取的“3原则”，书目记录中检索点的范围大幅扩展，促使大量新规范记录产生；RDA规则鼓励“如实转录”，因此规范检索点取消了大量的人为缩写，而以用户容易理解的全拼形式记录，增强表达性，从而引发大量规范检索点的形式变化；RDA采用“首选名称+附加成分”构建规范检索点的方法，对首选名称和变异名称的选择，以及为区分同一名称的不同实体添加附加成分的顺序都与AACR2存在差异。格式上的变化也是造成规范数据库更新的原因之一。MARC21规范数据格式为适应RDA做了相应修订，增加字段近40个。

规范数据库日常更新的3种类型一般通过MARC21规范数据格式的头标/05字符位代码予以表示，即分别用“n”“c”“d”表示“新增”“修改”和“删除”。

3.1 新增规范记录

如果某实体在规范数据库中找不到对应的规范记录，就需为其创建规范记录，常见的是为新增的个人、团体创建规范记录，还可以为新增的作品或内容表达建立新记录。如2014年出版的《习近平谈治国理政》英文版就是一个新的内容表达，可为其创建规范记录如下。

LDR/05 n

100 1#$aXi, Jinping.$tXi Jinping tan zhi guo li zheng.$lEnglish

400 1#$a 习近平.$t 习近平谈治国理政.$lEnglish

400 1#$aXi, Jinping.$tXi Jinping, the governance of China

430#0$a 习近平谈治国理政

670##$aXi Jinping, The governance of China, 2014: $bcolophon (Xi Jinping tan zhi guo li zheng--English)

此外，在LC规范控制实践中还存在一种需要新增规范记录的情况，即分离未区分的规范记录。当多个实体拥有相同的名称，但是用于区分它们的信息不足时，这些实体可暂时共用一条规范记录，待后续编目员获得的信息足以将它们区分开时，再重新创建规范记录。在LC规范数据库中目前存在51 785条名称未区分的规范记录[5]。未区分的规范记录用008字段32字符位代码“b”表示。

例如，LC控制号为“nr2001024383”的规范记录就是一条未区分的记录，其规范检索点形式为“Li, Qiang”。《国家图书馆藏民国军事档案文献初编》的责任者之一“李强”和《大型公共场所人员疏散策略模拟与应用》的著者“李强”是两个不同的实体，却共用同一规范检索点形式。

LDR/05 c

008/32 b

100 1# $aLi, Qiang

400 1# $a 李强

670##$aDa xing gong gong chang suo ren yuan shu san ce lüe mo ni yu ying yong, 2011: $bt.p. (李强, Li Qiang)

670##$aGuo jia tu shu guan cang Minguo jun shi dang an wen xian chu bian, 2009: $bt.p. (李强 = Li Qiang)

由于规范形式通过添加附加成分可区分，为后者“李强”创建规范记录。RDA规定个人名称附加成分的优先顺序为：出生日期和（或）死亡日期、名称的更完整形式、个人活跃期、职业或工作添加。由于无时间信息，名称也完整，所以根据资源的题名推断职业或工作信息作为附加成分。

LDR/05 n

008/32 a

100 1#$aLi, Qiang$c(Writer on evacuation of civilians)

400 1#$a李强$c(Writer on evacuation of civilians)

667##$aFormerly on undifferentiated name record: nr2001024383.

670##$aDa xing gong gong chang suo ren yuan shu san ce lüe mo ni yu ying yong, 2011: $bt.p. (李强, Li Qiang)

3.2 修改规范记录

规范记录的数据内容部分一般包括规范检索点、单纯参照、相关参照、参考数据源等信息，当这些信息发生变化时，就要对规范记录进行修改，使规范记录的信息更完整，更方便识别和区分。

规范检索点一般由“首选名称+附加成分”组成，当首选名称或附加成分发生变化时，即需对规范记录进行修改。例如，将LC控制号为“n 79133113”的规范记录“Ba, Jin, 1904-”由于补充了卒年信息，规范检索点形式修改为“Ba, Jin, 1904-2005”。规则变化也常常引发记录的修改。例如，RDA规则要求不应人为

地进行缩写，如用“approximately”取代了拉丁文缩写“ca”，个人活跃期用“active”，取代“fl.”，所以当对原AACR2的规范检索点进行“RDA化”修改后，拉丁缩写要转化为完整英语形式。此外，参照或数据源信息的补充也会造成数据的修改。例如，为LC控制号为“n 00011963”的规范记录“Sargent, John F.”增加了670字段的来源信息“$aPhone call to author, Feb. 10, 2012 $b (prefers full name, John Francis Sargent, Jr.; b. 1962)”。

3.3 删除规范记录

LCCN是不可重复使用的，一旦为某实体创建规范记录，并将LCCN分配给该记录，则不能将该控制号用于其他实体。如果规范记录本身发生变化，可将规范记录连同LCCN一并删除。MARC21规范数据格式头标/05字符位除了代表“d”表示删除之外，还有两个代码“s”和“x”也表示删除的情形。“s”表示一个规范检索点因被拆分成两个或多个规范检索点而删除的记录，该规范检索点在被拆分后新增的规范记录中以单纯参照形式出现。“x”表示由于一个规范检索点被另一个规范检索点取代而被删除的记录，该规范检索点也会以单纯参照形式出现在另一条规范记录中。当代码“s”和“x”皆不适用或者编目机构不需要细分删除的情形，则用代码“d”表示已删除的记录。LC规范记录即是如此，仅用代码“d”表示已删除的记录。对于未区分的规范记录，如果获得了可区分的信息实现了所有未区分记录的分离，均新建了规范记录，则原始的那条共用记录则需要删除。

4 小结

国图西文规范控制直接引用模式能充分共享国外的先进经验和成果，大大节省了建设规范数据库的人力和物力，还通过摸索实现与引用规范数据库的同步更新。尽管取得了一定的成绩，但是国图西文规范数据库的更新建设还存在一些有待解决的问题。

首先，更新虽然及时，但是缺乏自建的模式仍然不能完全满足西文规范控制建设的需要。例如，馆藏书目数据的检索点不能实现与规范数据库检索点的完全匹配。由于LC规范数据库是基于其多个成员馆的馆藏情况而共同建设，馆藏资源的获取途径、发行限制和受众群体等不同，国图编目员经常遇到对书目数据的检索点进行规范控制时，在西文规范数据库中找不到匹配规范记录的情况，编目员只能凭经验使用资源上的检索点形式，造成书目数据库中非控检索点的存在，尤其是在编目亚洲发行的资源时这种情况尤其突出。大量非控检索点的存在无疑为今后数据库维护增加了工作负担。再如，对于中国名称，LC规范检索点采用了汉语拼音的拉丁化形式。近年来，LC在建设名称规范数据时加大了对中国实体信息的补充，为不少中国名称增加了中文形式的单纯参照，以便用户更好地识别实体。但是，由于中国名称的特殊性，仅靠拼音很难区分实体，而LC在中文信息方面不具备优势，因此添加的单纯参照十分有限。国图编目员如果能利用自身信息优势，在规范记录更新时尽可能为中国名称添加可靠的单纯参照，将大大提升LC名称规范数据库的质量。但是完全引进的更新模式使编目员无法通过自建来完善规范记录。

其次，大量的更新也造成问题数据量的增长。数量如此庞大的数据库难免存在记录质量问题。例如，LC一些规范记录的008字段，代码应记录为小写字母，但是经常会出现大写字母，这样的记录无法实现对书目文档相关检索点的自动更新。在这种情况下，编目员只能将大写字母改为小写，虽然只是简单修改，但是保存之后会形成本地控制号和操作员代码。这样的记录更新时不能依靠控制号的匹配自动覆盖，只能暂时将它们放置在临时数据库中。随着更新的进行，这部分记录的数量不断增长，更新信息无法在规范数据库中体现，造成更新无效。

最后，更新的一个重要方面是为规范检索点增加了诸多单纯参照，但是国图Aleph 500系统并未将这些增加参照信息即时抽取索引，造成更新内容不能快速在检索机制中体现，而大大降低了更新的效果。

解决上述问题的关键是逐步在引进模式中增加自建环节。随着对规范记录创建内容及格式标准认识和理解的深入，国图编目员已经具备一定的规范自建能力，他们希望在共享国外规范成果的同时，尽可能多地参与到项目建设的愿望愈发强烈。按照NACO对成员馆的要求，加入项目的前提条件是参加一个联机合作编目系统，以便能够提交联机规范记录。国图2010年正式加入OCLC实现书目记录的上传，已具备提交规范记录的基本条件。国图应积极申请加入NACO等国际规范控制合作项目，在引进LC成熟规范数据库的同时，逐步增加自建环节，这样可以解决规范数据库没有相关检索点的

问题，同时还可以在发挥中文信息优势以及完善中国有关实体规范记录方面做出更多的贡献。在参与国际规范控制共建过程中，国图还能充分吸收各国规范控制建设经验，打造一支素质优良、具备参与国际项目建设能力的编目员队伍。随着规范控制工作的深入，国图应设置规范管理综合岗，用于专门解决规范控制中的各类问题。对于临时数据库中的更新记录，应由专人负责比对，用人工或半人工的方式将更新的重要信息合并至规范数据库中。国图也应加强对检索点抽取工作的力度，即使不能实现时时抽取，也应加大批量抽取的频率，使规范记录更新中的参照形式能尽快在索引中体现，从而提升OPAC的检索效果。

规范数据库的更新是规范控制建设中的重要一环。除国图之外，我国图书馆界在外文资源信息组织方面也多采用引进国际上成熟规范数据库的方式，因此，国图在西文规范数据库更新方面的经验对于其他图书馆建设外文书目系统具有一定的参考借鉴意义。我国中文文献信息编目中的规范控制环节虽完全采用自建模式，但也需要建立科学合理的更新维护机制，因此对于LC规范数据库相关情况的研究在一定程度上也希望能给予中文规范控制系统建设以很好的启发。

参考文献

[1] 黄俊贵. 规范控制概说 [J]. 高校图书馆工作，1999，19（3）：1-8.

[2] 罗翀. 国家图书馆外文书目规范控制的实践探索 [J]. 图书馆学研究，2011（16）：30-34.

[3] BURGER R H. Authority work：the creation，use，maintenance，and evaluation of authority records and files [M]. Littleton：Libraries Unlimited，1985：39-41.

[4] 罗翀. 国家图书馆外文规范控制的理论研究与实践探索 [D]. 北京：北京师范大学，2010.

[5] The Library of Congress. Linked Data Service [EB/OL]. [2019-03-21]. http://id.loc.gov/search/.

作者简介

张丽娟，女，1980年生，硕士，馆员，研究方向：西文编目，E-mail：zhanglijuan@nlc.cn。

A Brief Review on the Updating Mechanism of Western Languages Authority Database in National Library of China

ZHANG LiJuan
(National Library of China, Beijing 100081, China)

Abstract: Authority control is an indispensable part of cataloging. The update of authority control system is the necessary guarantee to keep the vitality of authority control work. After decades of construction, the western languages authority control system of the National Library of China has formed a relatively complete system, and some experiences had been accumulated in the update of the authority database. This paper introduces the updating mechanism of the authority database of western languages in the National Library of China, the methods and paths of the regular updating of the authority database, and then summarizes the specific reasons of new records, correction records and deletion records. At the end, the paper thinks about the problems existing in the update mechanism of the western languages authority database, and proposes specific ideas for future work.

Keywords: National Library of China; Western Languages Cataloging; Authority Control; Authority Data Update

（收稿日期：2019-05-21）