●张丽娟

CNMARC 存在的问题与 ECIP 计划的实施*

摘 要 CNMARC 存在着使用烦琐、数据著录重复、字段冗余等缺点。要解决这些问题,应实施 ECIP 计划。ECIP 利用了一次录入生出两种产品、对线性文件作结构化处理等原理,实现自动编目,可全面提高数据质量。实施 ECIP 计划,必须认真执行有关标准,建立综合书目服务网络,做好系统开发工作。参考文献 3。

关键词 CNMARC ECIP CIP 自动编目
分类号 G254.36

ABSTRACT There are some problems in CNMARC, such as overelaborate description, repeated contents and redundant fields. The author thinks that ECIP is a solution to these problems, and proposes some principles for the implementation of the proposed ECIP project. 3 refs.

KEY WORDS CNMARC. ECIP. CIP. Automatic cataloging.
CLASS NUMBER G254.36

自 1990 年《中国机读目录通讯格式》出版以来,我国编制的书目记录已达 100 多万种。十几年来,CNMARC 格式在数据整合、资源共享、业务管理和公共服务中起了很重要的作用。但它和任何一种社会现象一样,也不可避免地有其历史的局限性。随着信息技术的不断发展,它的不足之处,日显严重,应采取措施,予以解决。

1 CNMARC 格式存在的问题

1.1 数据处理复杂,使用烦琐

CNMARC 记录结构分为记录头标、地址目次区、数据字段、记录分隔符四部分组成。它共有 10 个功能块、479 个字段,子字段有近千个,其间对应的关系比较复杂,反映格式也比较复杂。特别是在字段指示符的设置上最为明显,很多指示符设置了却很少使用甚至从未使用过。例如 200 字段的指示符 1,"0"表示无检索意义,"1"表示有检索意义。在我们调查的河北省 34 所高等院校图书馆中实际使用情况是 100% 的书目记录在 200 字段的指示符 1 均选择了"1"。

1.2 CNMARC 格式难以掌握

CNMARC 基本上是 UNMARC 的译文。由于 UNMARC 自身就存在一些弊端,而 CNMARC 又未作调整,因此就有许多不适应中文文献的地方。许多字段、子字段没有中文例证。尤为严重的是没有一个权威性的实用版本。已见的 10 余种手册中,处理中文图书共使用 123 个字段,用法不同的有 36 个字段,于是就出现了几千个图书馆编制的几千万条书目记录(实际反映的图书只有 100 多万种)难以整合到一起,不同系统之间不能做到等同使用等现象。例如关于译著书名原文的著录格式,这次调查的结果是有不下四种的著录方式,有用 300 字段的,有用 454 字段的,有用 510 字段的,也有 454 和 510 字段同时使用的。而每一种著录方式当事馆均可以找到相关的著录手册来印证。

1.3 数据著录重复

主要是著录数据与检索数据的重复。例如 200 字段的 $a、$c、$d、$e、$f、$g、$i、$z 等 8 个子字段都有重复对应的字段:$a 相同作者的另一合订书名、$c 不同作者的合订书名对应 423 字段;$d 并列正书名对应 510 字段;$e 其他书名信息对应 517 字段;$f 主要责任者对应 701、711 字段;$g 其他责任者对应 702、712 字段;$i 分辑名对应 517 字段;$z 并列正书名语种对应 510 字段。而在 200 所有这些子字段中,除了 $a 有检索意义外,其他的 7 个子字段均无检索意义。因此造成了这样一种局面:编目人员为了遵循 CNMARC 著录格式而在 200 字段中录入了这些并无检索意义的子字段,这些重复的子字段占据大量内存空间,且毫无意义地增加了编目人员工作量。

1.4 字段冗余,数据冗长

CNMARC 中的冗余字段较多,如 3—块中 301、302、303、304、305、306、307、308、312、313、314、315、316、317、318、320、322 等字段,在 UNMARC 已说明 301 - 315 字段均可用 300 一个字段代替,CNMARC 却设置了 19 个字段。数据制作复杂,每一条完整的书目记录至少需要十几个甚至几十个字段方可比较全面规范地反映出该文献的基本内容。尤其是对一些附属丛书图书的综合著录,其著录数据更加庞大。

1.5 数据完成滞后

在图书出版领域,特别是计算机、语言类的图书,由于其生命周期短,因此有的图书还没有分编出来,更新的版本已经面市了。另一方面,由于每一条数据的编制都要遵循一定的原则,因此编目积压现象普遍存在,数据完成滞后。

由此可见,对 CNMARC 的书目实体描述部分做重大改进已是势在必行。自 2000 年以来,国内图书馆界的诸多资深专家,深入分析了我国中文图书在版编目和图书馆界联

* 本文为河北师范大学 2003 年社科基金资助课题论文。

合编目的现状、取得的成绩与存在的问题，结合《图书书名页》和《图书在版编目数据》两项国家标准的修订，吸收各方面对数字文献处理的研究成果，提出了在我国推进ECIP计划的构思。

2 ECIP的技术原理

所谓ECIP(Electronic Cataloging in Publication)计划，其内容即是利用印前电子文本加注结构标签的方法进行"自动编目"。

2.1 利用"一次录入，两种产品"的原理

图书编目实际上是将图书扉页(书名页正面)、版本记录页(书名页背面)等处的有关数据元素，按一定著录规则进行描述，形成书目记录。计算机编目则是用数据库方式把有关数据元素录入相应字段、子字段，通过计算机编目软件进行处理，同时产生目录卡片和机读目录两种产品。

2.2 对线性文件进行结构化处理

"自动编目"的实现，是利用中文图书印刷版的印前电子文本，从中直接提取书目记录所需要的描述性数据元素，在图书生产过程中自动产生该书的书目记录。但印前电子文本是线性文件，要使之成为数据库文件，就需要对印前电子文本进行结构化处理。有两种做法：一种是以数据库方式录入，定稿后转换为线性方式，以产生印前电子文本；一种是以线性方式录入，对有关数据元素加注结构标签，定稿后按其结构标签转换为数据库文件。

2.3 属性数据元素在图书物理结构之中

图书馆的书目记录与CIP数据，均包含著录数据与检索数据两部分内容。其中著录数据取自文献本身，也就是书目记录的著录数据存在于该书的印前电子文本之中。通过剖析印刷版图书的物理结构，可确定书目记录所需要的数据元素。书目记录与CIP数据，主要取开端部分的内容，有时涉及结尾与外表部分。《图书书名页》规定了扉页、版本记录页与附书名页有关事项，《中文普通图书著录条例》规定以扉页与版本记录页为主要信息源，确定这些部分的数据元素并对其进行结构化处理，是实现自动编目的基础。

3 实施ECIP的重要性

3.1 可全面提高书目数据质量

首先从CIP数据本身来说，目前只是根据出版单位报来的数据进行核定，出版单位未报的缺项，则无法判断。如现有的CIP数据基本没有并列书名，原因是出版单位未报，CIP中心无法增补。分类、主题标引，也因可参考的材料很少，不能保证标引质量。实现ECIP以后，除必要的书目数据外，尚有目次页、内容简介、出版说明、作者简介等内容，对提高标引质量将起到很大作用。对于正式书目记录而言，原先关于全面反映责任者、进行内容分析著录、介绍图书内容等要求，均因工作量太大而难以做到。现在由图书的电子文本自动生成，问题迎刃而解，书目质量将大大提高。

3.2 可以实现自动编目

目前，全社会重复进行大量图书编目的机构(不包括为本单位编目者)，少说也有一二百家，不仅浪费人力物力，而且影响标准化的推行，书目质量也得不到保证。由图书的电子文本自动产生新书预报与正式书目数据以后，全社会都可以享用这一成果，发行界和图书馆界都不用另行编目，其经济效益极其可观，每年最少可减少数千万元的编目费用。

3.3 为制定EBOOK标准奠定基础

ECIP的实现，涉及一系列标准与规范的执行，将有力地推进出版行业标准化的进程。目前出版界执行《图书书名页》国家标准很不规范，项目也不一致，有的把翻译书的原著书名印在扉页上，和并列书名混淆。这种情况不仅给编目工作带来困难，对图书电子文本的利用也很不便。《图书书名页》标准的认真执行，可为制定EBOOK的标准奠定基础。ECIP的实现，不仅对于统一机读目录格式有极大的影响，对于出版物元数据的研究工作也将产生积极作用，促使EBOOK的属性数据元素与书目记录的属性数据元素趋向一致。从而为大量开发与利用中文数字化信息资源创造了条件。

3.4 加快出版数字化进程

在我国的具体条件下，ECIP的全面实现，也就是图书出版从选题、审稿、录入排版到印刷、发行的全过程都将有所反映，也就产生了一个非常完善的书目信息数据库，从而形成出版数字化系统的完整解决方案，包含出版物自身数字化、出版过程数字化、出版管理数字化、出版商务数字化。可用以加强出版行业的科学管理，增加信息技术附加值，提高经济效益。还可在每一环节上实施底账核查的部分，以打击造假制假、盗版侵权的非法行为。

3.5 促进数字图书馆的建设

ECIP的软件技术，不仅实现了中文图书的"自动编目"，而且推动了EBOOK的"自动制作"，即在出版中文图书印刷版的同时，"自动制作"其电子版。这对于中文数字信息资源建设具有难以估算的价值，将大大加快我国数字图书馆建设的进程。

4 实施"自动编目"的方案设计

ECIP是在出版过程中，利用排版印刷的电子文本产生书目数据。只有实施ECIP计划，才是真正意义上的"自动编目"。我国的计算机编目经过了十多年的实践，已经有了一支专业队伍，积累了宝贵经验，在机读格式、著录规则、规范文件、分类与主题工具等方面正逐渐形成统一的文本，联机编目网络已开始运作。这是实现中文图书"自动编目"的最基本条件。此外还需要认真做到以下几点。

4.1 认真执行国家有关标准

(1)信息处理自动化的过程，也就是标准化的过程。信息处理数字化，就需要在更大的范围内实施标准化。2001年，《图书书名页》和《图书在版编目数据》两项国家标准已经修订，修订后的《图书书名页》规范了书名页与附书名页的有关内容；修订后的《图书在版编目数据》进一步明确了著录数据和检索数据的两部分内容。执行修订后的两项标准、推行ECIP计划，是实现"自动编目"的契机。可要求出

版社在印前电子文本中,对《图书书名页》国家标准所规定的内容加注结构标签。这样,在一种书的印前电子文本成型以后,即可以从中提取CIP所需的描述性元数据。经上级单位审定后的CIP数据及印前电子文本再返回出版单位,在图书印刷出版时即可产生正式的书目记录。

(2)加强协调相关标准。ECIP的实施,除上述两项标准外,还涉及到出版业、图书馆界和计算机领域的有关标准,如《图书流通信息标准》、《普通图书著录规则》、《中国机读目录格式》、《信息技术——通用多八位编码字符集》等。ECIP的结构处理还要与EBOOK的结构保持一致,才能在网络环境中实现更大范围的资源共享。

(3)统一元数据格式。印前电子文本结构标签的格式,应与目前中文图书的计算机编目格式保持一致,应以文化部发布的CNMARC标准为基础,按"自动编目"的特点,做适当调整。

4.2 建立综合书目服务网络

(1)建立社会化书目服务体系。ECIP是在图书生产过程中,从图书印前电子文本中以软件方法提取相关数据元素,自动产生数字化的书目记录,因而要在网络环境中运作,方可有更好的效果,书目网络中心的建立势在必行。网络中心一方面为出版社实施ECIP提供技术支撑,例如各出版社不可能配备齐全辅助编目的工具书,中心应在网上提供给各社使用,以保证CIP数据的正确制作与及时处理;另一方面要最大限度地向全社会发布,充分发挥其宣传报道新书的作用,扩展图书发行工作。只有建立这样一个社会化的书目服务机构,才能推进ECIP的顺利实施。

(2)全面考虑文化、经济、管理功能。在书目网络建设时,不仅要考虑扩展图书发行的经济效益,还要把ECIP作为出版数字化经常建设的一项重要工作,与提高出版业自动化管理水平、推动电子商务模式、促进EBOOK制作、扩大中文甚至资源开发等等紧密联系,还要为数字图书馆建设创造良好条件,使ECIP的实施产生最大的社会效益。

(3)CIP-MARC-BIP综合处理。在图书正式出版前,提取CIP数据,也可以从最后的图书电子文本中产生完整的MARC记录,这是实现中文图书"自动编目"的最佳途径。这一书目记录的充分利用将产生极大的社会效益,每年减少的重复编目费用就有几千万元。书目网络中心还可用以建立全国性的图书现货目录。在各出版社的配合下,可以逐步建成规范的"可供书目"(Book in Print)。

(4)提高出版社的自动化管理水平。实施ECIP计划并自动编制完整的MARC记录,还要和出版社建立本社书目信息中心联系来考虑,实现编辑管理、出版管理、发行管理的高度集成化,全面提高出版社的自动化管理水平。

4.3 做好系统开发工作

(1)选择软件工具。目前北大方正电子有限公司和海文电子信息系统公司合作,在方正系统的注解方式中,以内部嵌套方式加注XML语言结构标签,实现线性结构文件的数据库化,以海文全文检索系统进行数据处理。

(2)优化处理流程。采取数据库方式录入,录入人员只需按项目录入数据内容,由软件生产为加注排版语句和结构标签的形式。

(3)规范操作方法。操作步骤如下:作者填写"作者书稿登记表";责任编辑在书稿上按ECIP著录细则进行标注后,录入人员据此用数据库方式录入至相应工作单;书稿录入完成后,按文本中的XML结构标签,生成书目记录;出版社将CIP数据连同电子文本对报送的CIP数据及图书书名页审定后加注CIP中心审定号,形成印刷型文本形式,返回出版社;出版社按CIP数据在《中国图书在版编目快报》上刊登,并在网上发布;图书正式出版后,出版社在缴送样书的同时,将改善书印刷版电子文本的前端、封面、结尾几部分传送至CIP中心;CIP中心根据出版社缴送的样书与所附有关的电子文本进行核对,生成MARC记录。

(4)依靠作者参与,做好主题标引。目前是采用人工赋值方法,请作者或责任编辑把6—块的数据加到印前电子文本中。作者对自己的著作是最了解的,因而能准确揭示其主题内容。作者给出的主题词可能不符合图书馆规范化的要求,但从另一方面说,则更接近读者对主题的理解。书目记录审定人员可据以修正,给出规范化的主题概念。

改进后用于中文图书ECIP与自动编目的MARC格式只有46个字段,153个子字段。相对CNMARC减少了大量冗余的字段。取消了CNMARC中一些没有检索意义的字段,增设了一些具有检索意义的子字段。如关于责任者的著录,取消了200字段中无检索意义的$f、$g子字段,由7—字段直接产生书目记录的作者项。再如交替题名的著录,在200字段增设了$t子字段,著录交替题名,并设为可检子字段,不在517重复著录;特别值得一提的是在《普通图书著录规则》、CNMARC格式及各手册均无说明的、未在扉页出现的合订书名,改进后的MARC格式采取了在304字段作附注,同时记入517字段,在423字段连接215子字段的著录方式,说明此合订书名未在扉页出现,既生成附注,又是检索点。由于改进后的MARC其中有13个子字段可以由计算机自动生成,只有127个子字段需要在ECIP操作中标识处理,因此这样一个数据量,是便于编目人员学习和掌握的。

总之,"自动编目"的实现,不仅可以推进我国书目网络系统跨越式发展,从落后走向世界先进行列,还可为EBOOK的大量生产和数字图书建设奠定坚实的基础,将产生极大的社会效益和经济效益,因而期待引起国内各方面的重视和投资者的兴趣。

参考文献

1 陈源蒸.关于CNMARC格式调整的构想.图书馆学刊,2002(6)
2 陈源蒸.中文图书ECIP与自动编目手册.北京:北京图书馆出版社,2003
3 都平平,武保民.从USMARC看CNMARC.山东图书馆季刊,2002(4)

张丽娟 河北师范大学图书馆副研究馆员。通讯地址:河北石家庄。邮编050016。
(来稿时间:2003-12-10)