

●陈源蒸

中文图书自动编目的实现

摘要 目前一些出版社运行的复合出版系统包含有 ECIP 软件,灵活运用 XML 语言、排版软件、数字文献处理、元数据研究诸方面的技术,在制作复合文件的处理过程中,真正实现了中文图书的自动编目。参考文献 7。

关键词 中文图书 自动编目 复合出版系统 电子在版编目

分类号 G254.362

ABSTRACT At present, the hybrid document publishing systems of some publishers have already included ECIP software, flexibly applied XML, type setting software, digital document processing and metadata techniques, and realized automatic cataloging of Chinese monographs while producing hybrid documents. 7 refs.

KEY WORDS Chinese monograph. Automatic cataloging. Hybrid document publishing system. ECIP.

CLASS NUMBER G254.362

1 实现自动编目需要解决的问题

现有图书馆自动化系统的编目软件,其操作流程是从印本图书上,将编目所需的信息资源,按书目记录的结构形式(中文图书使用的是 CNMARC 格式)逐项录入,生成机读目录。虽然机读目录本身具有印刷版与数字版双重属性,打印卡片与建库检索的处理也是自动化的,但实际上这仍然是手工操作,而非自动编目。因为从印本图书上著录的数据,在图书的电子文本中已经存在,人工又重新录入一次,存在“比特—原子”“原子—比特”的非数字化操作的普遍问题。且由于信息开发的深度不一,内容选取范围不同等因素,很难统一。

完整意义的自动编目指是由出版单位从出版物制作的源头,利用计算机排版文件的相关部分(包含扉页、版本记录页、内容简介、附书名页等,详细著录还有目次页与各种插页),自动提取书目记录所需的元数据,按 MARC 格式供全社会使用。美国国会图书馆的 ECIP 计划正是依照这一原理而提出的。笔者也曾倡议在我国推行 ECIP 计划,实现中文图书自动编目,因为图书馆界自身是不能实现自动编目的^[1]。

由于书目记录所需要的元数据,均在出版物的对象数据之中,出版物利用计算机排版所形成的排版文件(以下称文件)也是计算机可读的,只是结构形式不同。用它自动编制书目记录(以下称书目)是完全可以做到的,主要是对线性文件进行结构化处

理,但在处理过程中需要妥善解决如下几个问题:

(1)文件包含的数据与书目需要的数据有 3 种不同情况:文件与书目一致;文件需要印出,书目不需著录;书目著录需要,但文件上没有。

(2)文件中含有说明字号、字体、排版格式等注解语句,是书目不需要的。

(3)对书目各个项目标识的以 XML 描述的结构标签,是文件不需要的。

(4)文件的数据排列次序与书目数据的次序也不完全一致。

(5)自动生成书目,所有元数据必须惟一性识别,书目格式中一些选择性项目要有所调整。

(6)运用 XML 语言,DC 元数据较好表示,DC 与 CNMARC 需要做到完全对应。

2 HDPS 采取的做法

自 20 世纪末数字图书馆在我国兴起以后,人们即提出了自动编目的问题,并进行了多方面探讨,还有人推出了电子编目员的方法^[3]。

目前在一些出版社运行的复合出版系统(HDPS, Hybrid Document Publishing System),其中包含有 ECIP 处理软件(可单独运行),灵活运用 XML 语言、排版软件、数字文献处理、元数据研究诸方面技术,在制作复合文件(Hybrid Document)的处理过程中,妥善解决了上述问题,真正实现了中文图书的自动编目,填补了国内空白。关于复合文件融印刷版与数字版和对

象数据与元数据于一体的原理与相关技术,本文不作详细叙述,这里只说明实现自动编目的有关问题。

(1)运用 XML 语言混合处理结构化数据与非结构化数据的功能。一般数据库管理系统仅能处理结构化数据,XML 语言以一对相互匹配的起始和结束标识符来标记信息,从而可以只描述结构化数据,而对非结构化数据忽略,实现两者的分离。凡是书目中不需要的数据(含说明字号、字体、排版格式等的注解语句)均分离在 XML 语言的处理范围之外。如“出版 北京图书馆出版社”,只将需要的出版单位数据标识在所定义的结构标签之内,而出版两个字则分离在外,从而使利用线性结构文件制作结构数据成为可能。

出版■<P,name>北京图书馆出版社</P,name>

对书目需要但文件中没有的数据,作隐藏文字或不排版处理,如:

<P,place>北京</P,place>

(2)发挥排版软件的特点,以不同方法加入结构标签。目前排版软件有多种形式,各有特点。人们按其自身具备的功能,用不同的方法,制作复合文件。

方正排版软件属于批处理排版方式。这种排版方式排版过程不直观,是其缺陷,但人们利用其注解语句,说明 XML 语言描述的语义,使方正小样文件成为复合文件。

WORD 文件,操作方便,形象直观,不少出版单位已用它进行排版印刷。人们利用其隐藏文字的功能,说明 XML 语言描述的语义,既不影响它“所见即所得”的功能,也不影响正文的打印输出。同样可成为复合文件。

从我国出版业的实际情况出发,利用现有出版软件的特点,与 XML 语言结合使用,不仅免除编写 XSL 表达软件的大量工作,又使复合文件具有实现的可行性。

(3)选择适用的全文检索系统进行优化。目前全文数据处理系统很多,但多数系统是以原始文本用于浏览,而另建一文本用于检索。而海文全文处理系统只用一个文件,可去除排除文件中含有的说明字号、字体、排版格式等的注解语句,直接以排版文件建立索引,以大样文件浏览,符合自动编目的基本要求。

优化工作主要是增加系统对 XML 语言的处理功

能和在元数据的标识中采用国家标准格式。优化后的海文系统,能处理含有以 XML 语言描述的结构标签的排版文件,从而产生复合文件。以其含有的 XML 语言描述的结构标签,产生符合标准化要求的书目数据,既有 DC 元数据表示格式,也有 CNMARC 表示格式,CNMARC 格式还可转换为 ISO-2709 交换格式。

(4)编制中文图书元数据表。虽然不少人主张书目记录采用 DC 元数据,而且以 XML 语言描述印前电子文本的结构化数据,也以采用 DC 元数据为好。但 MARC 的数据结构严密,其著录格式遵循国际标准,有严格的语义规则和完整的信息描述手段,能够精确完整地记录文献资源^[4]。它所生产的元数据虽然是半个世纪科学财富的积累,但 DC 元数据在某些功能上仍不够完善。国内外图书馆专家大都主张,DC 元数据与 CNMARC 应以共存互补的形式,处理数字与非数字信息资源,OCLC 的 CORC 计划就是使联机编目向数字化发展^[5]。

因此,在编制中文图书元数据表时要统筹考虑。一方面要调整 CNMARC 格式,克服它重复、烦琐、冗长的缺陷,另一方面又要扩展 DC 元数据的修饰词,解其过于简单的问题,做到相互之间能有全面的、惟一性的对应。在排版文件进行结构化处理时没有二义性障碍,所产生的数据可满足多方面需求。

HDPS 系统中的“中文图书元数据表”,与《图书书名页》、《图书在版编目数据》、《普通图书著录规则》、《中国机读目录格式》等国家标准保持一致,列出了 CNMARC 与 DC 元数据两种格式的全面对应关系^[6]。其中与中文图书物理结构不完全适应的地方,均作了妥善处理,既从中文图书出版的实际情况出发,又认真执行了国家有关标准,使所产生的书目记录达到标准化的要求。

3 HDPS 处理流程

中文图书自动编目,可以取排版文件中的相关部分(扉页、版本记录页、内容简介、附书名页等),利用“ECIP 处理软件”专门进行编目处理,也可在制作整本图书的复合文件时一并处理。两者的处理流程稍有不同,以下是进行自动编目处理的流程:

进入 HDPS 系统的自动标识模块→取已印刷出版图书排版文件的相关部分(制作 CIP 数据使用未印刷时的文件)→对编目所需数据元素标识结构标签形成复合文件→在排版软件中进行排版处理,确认

大样文件无误→将复合文件装入 HDPS 系统的数据库→装库成功后即可浏览该文件的 MARC 屏幕显示格式→可以个别或批量将 MARC 屏幕显示格式转换为 ISO-2709 通信格式。

4 制定各项操作规范

为保证编目数据符合国家标准的要求,制定了在标识过程中要执行的各项操作规范。

(1)规定数据元素的回车规则。目前系统对数据元素的识别以回车为标志,因每一数据元素均需回车,与一般录入规则有所不同。如:

<T,proper>傅雷家书</T,proper>
<Co,s-personal name>傅敏</CO,s-personal name>
<D,s-personal r>编</D,s-personal r>

(2)对文件与书目数据排列次序作不同处理。印本文件的数据排列与书目数据的排列不尽相同,如著作方式,在印本上有时排列在著者之前,而书目数据规定著作方式必须在著者之后,因此作如下处理:

主编<C,p-personal name>杨医亚</C,p-personal name>
<D,p-personal d>主编</D,p-personal d>

(3)仅为书目需要的数据预先设置。对于印本文件中没有,而书目数据需要的,可预先设置,插入到文本中,可减少大量标识操作。如:

<L,text l>chi</L,text l>
<Id,publishing codes of country>CN</Id,publishing codes of country>
<Id,publishing codes of region>110000</Id,publishing codes of region>

(4)为适应著录规则的需要,作必要的技术处理。如《大众用药手册》扉页列有主编张熙增等5个人名,则在张熙增后加注“等主编”标识,其他4人在书目记录中则不显示。

<C,p-personal name>张熙增<C,p-personal name>
<D,p-personal r>主编</D,p-personal r>
<D,p-personal d>等主编</D,p-personal d>

(5)逐步研制各种录入模板。每一书目数据需要标注的结构标签的数据元素约40个左右,虽有元素表可自动标识,操作量仍然很多。

若采取模板方式录入,则可大大简化操作。每个出版社的版本记录页大都具有固定模板,可在原有录入模板上预先标注结构标签后进行书稿录入,则丝毫不增加录入人员操作,即可完成版本记录页的

加注结构标签工作。

此外,内容简介、目次页和印本文件中没有的机读数据,也可预制模板,只需填写有关参数即可。这样,每种图书主要是对扉页的元数据进行标识处理,一般只有6~7项。因此,在排版文件上的编目工作比手工编目要减少很多,而所产生的书目数据比手工编目要准确得多。

5 实现中文图书自动编目的意义

中文图书实现自动编目对整个文献信息系统的各个环节(出版、发行、图书馆)都有重大意义。

(1)从中文图书制作的源头进行自动编目,在图书付印的同时,即可将书目记录在网上发布,以最快的速度供所有书商与图书馆使用。目前到处重复编目的现象可以有效地避免,将节省大量人力物力(目前每年全社会用于编目的费用在1亿元以上)。它体现了系统工程的思想,上道工序为下道工序创造提高效率的条件。

(2)如在“在版编目”阶段使用本软件,ECIP功能即可得到实施,整个新书预报工作将出现完全崭新的面貌。ECIP问世将对编目界再一次产生革命性影响^[7]。

(3)推进标准化,提高数字管理水平。在出版、发行、图书馆之间普遍存在的“原子—比特”、“比特—原子”的反复交换的问题有望获得解决。各家分别编目,因信息开发的深度不一,内容选取的范围不同,严重影响书目资源的共享。由于在出版物制作的源头产生的是符合标准的书目数据,各个环节就可直接以“比特—比特”方式交换信息,实现信息资源共享,真正做到“一家编目,大家使用”。

(3)CNMARC与DC元数据的全面对应,电子书的属性描述与书目记录的内容保持一致,为数字与非数字信息资源统一元数据格式创造了条件,有利于各方面制定元数据方案,实现数字与非数字信息资源的一体化处理。

(4)“自动编目”软件的研制成功,为自动标引的实现创造了条件。在自动编目过程中所提供的书目数据内容,可满足自动标引软件试运行的需求,从而加快开发进程。

(5)自动编目的实现,对金版工程、网上书店和数字图书馆的建设,均将产生积极的推动作用。尤其是“复合出版系统”的全面应用,其影响将更加深远。

(下转第73页)

样服务(how),即服务的方式方法问题,这是服务人本化的保障。为读者服务的思想不牢,服务态度不好,方式方法不对路,就是馆藏再好也很难发挥作用。只有牢固树立为读者服务的思想和科学合理的方式方法,才能充分发挥馆藏作用,满足读者的需求。服务时间(when),即何时服务。当今社会是信息社会,信息瞬息万变,其时效性很强,要求服务准确、及时和快捷,否则,信息就会失效。因此,图书馆人必须研究读者群的工作和学习规律,充分利用现代通讯技术和网络技术,为读者提供便捷、及时、周到的服务。

4.2 图书馆内部管理的以人为本

图书馆功能的发挥和实现,要以科学的管理为保障,要体现服务的人本化。管理必须人本化。服务人本化是管理人本化的目的,管理人本化是服务人本化的保障,没有管理的人本化,就很难实现服务的人本化。图书馆管理人本化,从宏观管理层面而言,应从以下两方面着手:一是要营造一个适应全体馆员发展的工作和学习环境。现代图书馆就其业务规律和内部管理而言,不论是高校馆还是公共馆,均具有相对的独立性和系统性。要实现管理的人本化,全体图书馆人,特别是馆领导班子一班人,要努力营造一个良好的工作和学习环境,在实现图书馆功能的同时,尽力满足馆员的工作、学习、生活和发展的需要,使不同类型的人才能够留得住、用得上、发展好。在留人方面,应做到“情感留人、待遇留人、事业留人、发展留人”。在用人方面,要尽力做到知人、用人、发展人,做到人尽其才。二是要为每个馆员创造一个良好的工作平台和发展平台。人是社会发展的历史存在,是人类社会物质文明和精神文明的继承者和创造者。人之所以为人,其不仅具有自

然属性,更具有社会属性,个人聪明才智的发挥及需要的满足,要依托于其在岗位和组织的奉献。管理的人本化就是将馆员个人的工作和发展与其所在组织功能的实现紧密相连,形成馆员个人与组织相辅相成的有机整体。在全馆上下形成“我尽职、馆发展、馆发展、我发展”的坚强信念。图书馆是全体馆员发展的重要平台,要根据读者和社会发展的需要,结合本馆的实际情况,明确办馆的指导思想,制定本馆的发展规划和实施计划,不断进行机制和体制创新,适时调整馆内的组织机构和管理机构,为馆员构建良好的发展平台。掌握每位馆员的具体情况,确定馆员的工作岗位和职责,加强对馆员的工作考核,并客观公正地进行评估,将每个馆员的工作实绩与其待遇和发展相联系,促进个人和社会的全面发展。

参考文献

- 1 马克思,恩格斯.马克思恩格斯全集(第3卷).北京:人民出版社,1960:514
- 2 马克思,恩格斯.马克思恩格斯全集(第46卷).北京:人民出版社,1979:393
- 3 周三多等.管理学—原理与方法.上海:复旦大学出版社,1999
- 4 [美]斯蒂芬·P·罗宾斯.组织行为学.北京:中国人民大学出版社,1997
- 5 张德.组织行为学.北京:高等教育出版社,1999
- 6 吴建中.21世纪图书馆新论.上海:科学技术文献出版社,1998
- 7 余源培.人的需要和人的全面发展.学术季刊,2002(11)

张治理 安徽财贸学院图书馆馆长,副教授。通信地址:安徽省蚌埠市宏业路255号。邮编233041。

(来稿时间:2004-07-10)

(上接第40页)

参考文献

- 1,2 陈源燕.推行 ECIP 计划 实现中文图书“自动编目”.图书馆学刊,2002(1)
- 3 张轴材等.数字图书馆的实践探索.见:数字图书馆:新世纪信息技术的机遇与挑战国际研讨会论文集.北京:北京图书馆出版社,2002
- 4 刘炜等.数字图书馆引论.上海:上海科学技术文献出版

社,2000

- 5,7 国家图书馆发展研究院.转型期图书馆工作研究.北京:北京图书馆出版社,2003
- 6 陈源燕.关于 CNMARC 格式调整的构想.图书馆学刊,2002(6)

陈源燕 中宣部出版局离休干部。通信地址:北京市东厂北巷4号楼553室。邮编100006。

(来稿时间:2004-04-19)