●朱 芊

全国中文机读书目主题标引格式问题分析

摘 要 中文机读书目主题标引格式的不统一,会产生书目主题格式多样的弊端。其主要原因是由于叙词法后组式标题只能套录标题法先组式标题格式。因此,必须统一中文书目主题标引格式。表 2。参考文献 8。

关键词 MARC 主题标引格式 中文书目数据 计算机检索
分类号 G254.2

ABSTRACT The fact that there does not exist an unified format for the subject indexing of Chinese MARC has caused many problems. In this paper, the author analyses some reasons and restates the necessity of a unified format. 2 tabs. 8 refs.

KEY WORDS MARC. Format of subject indexing. Chinese bibliographical records. Computer retrieval
CLASS NUMBER G254.2

中国图书馆界采用 MARC 格式建立中文书目数据库已走过了十年的历程,这是一个不断总结经验、不断统一认识、不断提高的过程。随着信息网络的发展,MARC 的局限性也逐渐显露出来,尤其是我国图书馆界普遍采用的叙词法后组式标引方式与根据标题表制定的 MARC 先组式标题格式的矛盾日益突出,出现多种主题标引格式并存的现象。主题标引格式的不统一,直接影响书目数据库的质量,不利于文献信息资源共建共享,是一个亟待解决的问题。本文对此加以探讨。

1 我国中文书目主题标引套录 MARC 格式的历史过程

1973 年,我国图书情报部门开始了应用计算机技术的研究。1986 年,UNIMARC 中译本面世。随后,国家图书馆、北京大学图书馆等分别编写了《中国机读目录通讯格式》的讨论稿。1987 年底至 1988 年初,国家图书馆根据新出版的 UNIMARC 手册对讨论稿作了修订。1991 年 2 月,书目文献出版社正式出版《中国机读目录通讯格式》,即 CNMARC。1988 年 4 月,国家图书馆开始采用 CNMARC 格式实施中文图书计算机编目,并规定主题字段无论是学科名称主题,还是专有名称主题,其主标题无论是一个主题词,还是多个主题词组配,均录入“ $a ”字字段。如果“ $a ”字字段由多个主题词组配,仍保留手工标题的组配符号和轮排方式。从 1990 年起国家图书馆正式对外发行中文机读书目数据软盘,以后又以光盘方式发售,全国图书馆界基本采用这种主题标引格式。见下例:

例 1:《国际海上集装箱运输》

机读目录格式

606 $a 国际运输 海上运输 集装箱运输
606 $a 海上运输 国际运输 集装箱运输
606 $a 集装箱运输 海上运输 国际运输

转换成卡片目录格式

①国际运输 海上运输 集装箱运输
②海上运输 国际运输 集装箱运输
③集装箱运输 海上运输 国际运输

例 2:《小麦育种》

机读目录格式

606 $a 小麦 —作物育种
606 $a 作物育种 —小麦

转换成卡片目录格式

①小麦 —作物育种
②作物育种 —小麦

1996 年 7 月 1 日《中国机读目录格式 WH/ T0503 —96》作为文化部行业标准向全国发布。1998 年 4 月、11 月国家图书馆业务处与图书采编部两次召开全国中文编目工作研讨会,邀请全国有关的专家和编目人员针对各编目单位在建立书目数据库的过程中,对 UNIMARC 格式和 CNMARC 格式的理解存在的差异,而使所加工的书目数据出现不统一的几十个字段与子字段进行了专题研讨。其中,确定主题标引格式遵循 CNMARC 标准,以主题字段中的“ $x ”子字段取代“ $a ”词串中的“ :”和“ —”,得到与会专家和各单位编目人员的认可。见下例:

例 1:《国际海上集装箱运输》

机读目录格式

6060 $a 国际运输 $x 海上运输 $x 集装箱运输

转换成卡片目录格式

国际运输 –海上运输 –集装箱运输

例 2:《小麦育种》

机读目录格式

6060 $a 小麦 $x 作物育种

转换成卡片目录格式

小麦 –作物育种

2　目前国内七大编目机构的做法

随着 CNMARC 标准的发布、实施,全国图书馆界在编目工作实践的基础上对 CNMARC 进行深入研究。1998 年以来,国内陆续出版了 CNMARC 使用本的一系列著作,其中与图书编目有关的著作有:国家图书馆图书采编部编的《中文普通图书编目手册》(1998 年);王鹤祥主编的《中文普通图书机读目录著录细则》(1998 年);中国科学院文献情报中心等编的《文献机读目录数据处理手册》(1999 年);熊光莹主编的《计算机编目技术手册》(1999 年);谢琴芳主编的《CALIS 联机合作编目手册》(2000 年);全国图书馆联合编目中心编的《中文图书机读目录格式使用手册》(2000 年);甘琳编著的《中文图书数据处理规程》(2000 年);潘太明等修订的《中国机读目录使用手册(修订版)》(2001 年)。这些著作中主题字段的使用规定不统一,反映在国内几个较大的编目机构编制的中文机读书目数据中主题标引套录 CNMARC 格式上仍存在差异,主要表现在主标题是由两个或两个以上主题词组成主题词串的格式不统一。现以《国际海上集装箱运输》标引的一组具有交叉组配关系的主题词串与《小麦育种》标引的一组具有限定组配关系的主题词串为例说明之,请见表 1。

表 1

编目机构名称	交叉关系的主标题	限定关系的主标题
国家图书馆全国联合编目中心	6060 $a 国际运输 $x 海上运输 $x 集装箱运输 6060 $2CT $3S029132 $a 国际运输 6060 $2CT $3S029852 $a 海上运输 6060 $2CT $3S036370 $a 集装箱运输	6060 $a 小麦 $x 作物育种 6060 $2CT $3S082507 $a 小麦 6060 $2CT $3S100090 $a 作物育种
上海图书馆	6060 $a 国际运输 $x 海上运输 $x 集装箱运输	6060 $a 小麦 $x 作物育种
深圳 ILAS 编目中心	6060 $a 国际运输 $x 海上运输 $x 集装箱运输	6060 $a 小麦 $x 作物育种
中国科学院文献情报中心	6060 $a 国际运输 6060 $a 海上运输 6060 $a 集装箱运输	6060 $a 小麦 $x 作物育种
新闻出版署信息中心(CIP 数据中心)	6060 $a 国际运输 海上运输 集装箱运输	6060 $a 小麦 $x 作物育种
北京地区高校图工委联合编目中心	6060 $a 国际运输 海上运输 集装箱运输	6060 $a 小麦 –作物育种
CALIS 联机合作编目中心	6060 $a 国际运输 $x 海上运输 $x 集装箱运输	6060 $a 小麦 $x 作物育种

3　出现问题的原因分析

上面列举的国内七大编目机构采用的中文机读书目主题标引格式,交叉关系的主题词串出现四种组配形式,限定关系的主题词串出现三种组配形式。为什么会出现上述局面呢?究其原因,我认为是更适于计算机检索的后组式叙词语言与根据标题表(LCSH)制定其主题字段的 MARC 格式之间的矛盾(请看表 2)。

MARC 格式主题字段其子字段的定义和排列次序,源于标题法原理。它规定主标题一般选自标题表主表中的一个词或一个词组,主标题(Subject heading)必须录入“ $a ”子字段。标题法先组式词语标识比较直观,含义比较明确,描述性好。但是从检索的角度讲,这些标题是事先固定好的,它只有一种形式,只能在字顺序列中唯一的一个地方找到它,不能随意从标题中任何一个主题因素入手进行检索,缺少多途径检索的功能。

后组式叙词语言最基本的原理是概念组配。它的标识无论是单一概念词还是复合概念词,都可以互相组配,以表达更为复杂的概念。如表 2 中是用叙词“药源性疾病 ”和“心脏血管疾病 ”组配,表达“药源性心脏血管疾病 ”这一主题概念。读者检索时,可以从“药源性疾病 ”检索,也可以从“心脏血管疾病 ”检索;既可以用“药源性疾病 ”与“心脏血管疾病 ”组合检索,也可以用“心脏血管疾病 ”与“药源性疾病 ”组合检索。从检索的角度讲,叙词语言可以自由组配,可以自由扩大、缩小或改变检索范围,可以实现多途径检索,因而更

表 2

国别及标引方式	标 题 形 式
中国 叙词法后组式标题	例 1:6060 $a 药源性疾病 $x 心脏血管疾病 $x 诊疗 主体面 例 2:6060 $a 班级 $x 学校管理 主体面
美国 标题法先组式标题	例 1:650 0 School management and organization $x Handbooks , manuals , etc. 主标目(Subject heading) 例 2:650 0 History of Medicine , 16th Cent . $v Chronology. 主标目(Subject heading)

适应计算机检索的需要。

我国的中文书目普遍采用叙词标引方式,标引所揭示的文献内容主题中的主体面(标题法中称主标目,亦称主标题)往往需要多个主题词组配。而按 MARC 规定,只允许一个词或是事先固定好的一个词组使用“ $a ”子字段,这种格式束缚了后组式叙词语言。由于存在叙词法后组式标题只能套录标题法先组式标题格式这一难解的问题,才产生了中文书目标题格式多样化现状。

4 关于统一中文书目主题标引格式的建议

据美国有关人士统计(见 Martin Dillon , Is MARC Dead? http/ / www. oclc. org/ institute/ alamarcl. ppt) ,互联网现大约有 4800 万个知识来源,如果要给这些网页编目的话,有人估计需要花去全美编目人员 24 年的时间,按翻一番的速度,第二年就要花 48 年的时间。也就是说,按照传统的信息组织方式,图书馆是无法跟上时代发展需要的。在这样的形势下,MARC 的局限性逐步显露出来:第一,MARC 格式结构复杂,字段大量重复。据有关专家统计,目前中文书目使用了 CNMARC 中的 125 个字段(亦有称 123 个字段)、394 个子字段。编制一条机读书目数据不仅需要经过严格的专业训练(编目人员必须掌握著录规则和熟记几百个字段、子字段、指示符及代码的定义) ,而且需要花费一定的时间。第二,人机界面不友好,只适合图书馆专业编目人员使用,难以推广到图书馆以外的行业中去。第三,MARC 格式的描述手段往往只适用于完整的、静止的信息内容的处理,不易处理动态的多媒体信息。第四,MARC 需要在专门的软件系统中使用,不太适应互联网的环境。第五,修订程序相当复杂,特别是联机合作编目,一个代码的改变,就会影响整个集成网络化系统。所以,具有数据结构简单,人机界面友好的组织网上信息的元数据元素集(Dublin Cord Element Set)等检索工具的研究与应用,在全世界方兴未艾。那么 MARC 是否过时,应摒弃不用?我认为一种信息组织工具的优劣,应该根据其检索功能来评价。MARC 是用于描述、存储、交换、控制和检索机读书目数据的标准,其详尽描述的书目信息,可以使读者较快而准确地找到所需文献线索。因此,只要纸质出版物存在,MARC 仍有不可替代的检索作用。

由于 MARC 在组织和检索书目信息上的作用,在相当一段时期内中文书目主题标引格式仍须套录 MARC 格式。我国七大编目机构采用的主题标引格式中哪种标题形式更好?我认为应从是否有利于检索的角度看待这个问题。分析我国七大编目机构主题标引格式,可得出两点意见:第一,我认为在机读目录中仍采用“ :”和“ —”作为主题词间的组配符号是不妥的。MARC 是计算机能够识别的一种目录,文献的标识和信息单元是通过计算机能够识别的代码组织成一条条书目记录,以至集合成书目数据库。“ :”和“ —”是文字符号,是手工目录使用的组配符号,不利于计算机识别、处理。用“ $x ”替代“ :”和“ —”,生成手工目录时,并不影响主题词的前后次序,而且计算机已经实现布尔逻辑检索,读者不必了解主题词之间组配符号的涵义,以及是否有组配符号,只要在检索框中填入主题词,计算机就会自动提供相应的书目记录。机检系统中继续使用“ :”和“ —”已没有任何意义。第二,将“国际运输 ”、“海上运输 ”、“集装箱运输 ”三个复合概念主题词组配表达“国际海上集装箱运输 ”这一主题概念的主题词串,分拆成“6060 $a 国际运输 ”、“6060 $a 海上运输 ”、“6060 $a 集装箱运输 ”三个标题,这种做法也存在比较大的问题。当然,这种标引方式不影响计算机检索的效果。但是,它首先是混淆了单主题与多主题的界限,破坏了叙词法概念组配以表达一个专指主题的基本原则;其次是削弱了主题标引揭示文献内容的作用,影响了浏览检索的功能。例如《科技进步与现代领导》一书的内容主题是“领导科学学 ”,机读目录格式应采用“6060 $a 领导学 $x 科学学 ”,转换成手工目录标题格式应是“领导学 —科学学 ”。如果选择单个词标题,机读目录格式是“6060 $a 领导学 ”、“6060 $a 科学学 ”,转换成卡片目录格式(计算机浏览性检索界面也提供这种数据格式) 是“ ①领导学 ”、“ ②科学学 ”,读者会误认为这本书讲了两个问题。所以,机检系统只有使用了后组控制符号或手检标目符号,避免或减少出现二义性问题,才可以采用单个主题词标引方式。

国家图书馆全国联合编目中心目前采用的主题标引格式是:一个主题概念若是需要两个或两个以上主题词组配

标引,先用组配好的主题词串套录 MARC 先组式标题格式;如果这个标题中的主体面是用多个主体因素主题词组配,这几个主体因素主题词要依次重复著录主题字段,并在“ $a ”子字段前面标识“ $2 主题规范表代码 ”与“ $3 主题规范记录号 ”。以《国有企业公司化改造研究》为例:

机读目录格式

6060 $a 国营企业 $x 经济体制改革 $x 研究 $y 中国

6060 $2CT $3S029301 $a 国营企业

6060 $2CT $3S040518 $a 经济体制改革

转换成卡片目录格式

①国营企业 —经济体制改革 —研究 —中国

②国营企业

③经济体制改革

全国联合编目中心采用这种主题标引格式是基于这样的考虑:一,主题字段、子字段的使用遵循 MARC 标准;二,保留主题词串,继续实现机检系统简单显示界面浏览性检索功能与卡片或书本式目录手工检索功能;三,坚持叙词法概念组配原则,主题词串中多个主体因素主题词依次重复著录相应主题字段,并挂接主题规范数据标识,实现计算机自动扩检、缩检和多途径检索功能。上述主题标引格式既能发挥计算机检索的优势,又照顾了传统目录的需要,比前面两种格式更规范,也比前面两种格式提供了更多的检索途径。但是这种标引方式也存在一定问题,值得商榷。

首先,主题字段的所有正式主题词应该全部挂接主题规范数据。主题字段著录的正式主题词,不论是主体因素主题词,还是空间因素主题词、时间因素主题词等等,均选自主题规范数据库。主题词挂接规范数据是为了充分发挥计算机组织和检索文献信息的作用,其前景是网上挂接词表(亦称主题规范数据库),实现机辅标引、自动标引和词表助检。因此,只要是选自规范数据库的主题词都应标有规范数据标识。如上例中的“研究 ”和“中国 ”,也是选自主题规范数据库中的正式标引用词,并不是编目人员自拟的非控主题词,应与“国营企业 ”与“经济体制改革 ”一样对待。其次,主题词挂接规范数据是用于计算机检索,在软件设计中应采用“610 非控主题词 ”字段的处理方式,挂接规范数据的单个主题词不在卡片目录上显示。因此,进一步完善国家图书馆目前采用的主题标引格式,具体的做法是:遵循 MARC 标准,保留组配标题形式;以“ $x ”子字段符号取代组配符号“ :”和“ —”,然后将标题中的主题词串拆成单个主题词,每个主题词根据自身的属性,选择相应的主题字段著录,并按 MARC 标准挂接各自的规范数据;但是挂接规范数据的主题词只用于计算机检索,卡片目录上不显示。如果采用这一建议,上例显示的两种格式应是:

机读目录主题标引格式

6060 $a 国营企业 $x 经济体制改革 $x 研究 $y 中国

6060 $2CT $3S029301 $a 国营企业

6060 $2CT $3S040518 $a 经济体制改革

6060 $2CT $3S086121 $a 研究

607 $2CT $3S096218 $a 中国

卡片目录主题标引格式

国营企业 —经济体制改革 —研究 —中国

元数据的研究与开发正成为当今信息网络发展的一个热点,MARC 也在进一步适应新的发展环境,像 USMARC 目前已发展成为 MARC21,开始被用来对电子文本进行描述。CNMARC 主题字段的使用更应着眼于网络信息的组织与检索技术的发展,使中文信息资源更好地为全世界所利用。

参考文献

1 中国机读目录格式 WH/ T0503 - 96
2 北京图书馆《中国机读目录格式使用手册》编委会编. 中国机读目录格式使用手册. 北京:华艺出版社,1995
3 张琪玉. 张琪玉情报语言学文集. 北京:北京图书馆出版社,1999
4 马张华,侯汉清. 文献分类法主题法导论. 北京:北京图书馆出版社,1999
5 沈迪飞. 图书馆信息技术工作. 北京:北京图书馆出版社,2000
6 吴建中. DC 元数据. 上海:上海科学技术文献出版社,2000
7 曹树金,罗春荣. 信息组织的分类法与主题法. 北京:北京图书馆出版社,2000
8 刘湘生,汪东波. 文献标引工作. 北京:北京图书馆出版社,2001

朱　芊　国家图书馆词表组副研究馆员。通讯地址:北京中关村南大街 33 号。邮编 100081。

(来稿时间:2001-08-01)