

●黄益券

马克编目的质量保障:

计算机辅助编目在 ZNK 系统的实现

ABSTRACT The life line of a system is the fullness and correctness of the data. The application of new technology cannot guarantee the correctness of the machine readable cataloging. An integrated measure is often taken to improve the quality of MARC data; to perfect the softwares is important, but it is not omnipotent. The process of computer-aided cataloging should be the process of the mutual aid between man and machine. The paper reports the effort made in this field by the author including his preerntative help to MARC cataloging rules and the technology and methodology of automatic detection.

SRBJECT TERMS Machine readable cataloging—Preparatilns Cataloging for publications in western languages—Data detection

CLASS NUMBER G254.36

我国图书馆自动化事业经过多年的探索之后,取得突破性进展。中国马克,即《中国机读目录通讯格式》正式出版,北京图书馆马克格式的书目数据的发行是其重要的标志。与此同时,适用于不同规模、机型的马克编目软件纷纷问世。这些都为我国各类型图书馆书目数据库的建立、馆际间书目数据交换创造了良好的环境和条件。然而,由于马克格式的复杂性和编目员对计算机技术的迷惑,使得马克编目的推广很困难。特别是在中小型图书馆,往往缺少训练有素、责任心很强的编目员,即使有了符合马克编目要求的软件,编目质量仍然基本上停留在原先手工编目的

水平上。这是一个相当普遍的问题。

数据质量是系统性能的重要因素。由于缺乏必要的数据项目,或者数据错误,即使系统的设计性能很好,也得不到充分的发挥。

书目数据质量的保障,有各种措施可以采取:

1. 尽量采用国家图书馆和地区性权威的编目机构的书目数据;
2. 加强对本馆编目人员的技术训练,提高他们的马克编目水平;
3. 招聘受过计算机知识训练的图书馆学系毕业生;
4. 实行部门或地区性图书馆间的合作编

目;等等。

让我们分析一下利用权威机构编目成果的问题。由于在版编目不能很快地落实,并且存在很大的时差,使各馆不得不自编许多(目前估计有50%)图书。即使在版编目在全国范围内得以实现,由于发行渠道多,仍然有相当部分(可能有30%)图书要各馆自编。这仅仅就正式出版的图书而言,还有许多非正式出版物,基本上不能依靠国家图书馆和地区编目中心。提高本馆人员的编目水平的问题仍然是不可忽视的。

马克编目确实复杂,有些问题很难弄明白。连结记录的层次代码、丛书项第一指示符,以及连结项4××字段之间的关系都是令人糊涂的例子。即使是训练有素、经验丰富的编目员,也有相当的难度。刚毕业的图书馆专业毕业生,面对出版物的多样性和马克编目的复杂性,仍然是一个新手。

地区或部门的合作编目确有成功的例子,但不是任何地区和部门能实现。

作为系统分析和设计人员所考虑的是通过提高软件的性能,对编目质量进行自动监督,给编目人员以尽可能多的帮助。这就是本文的主题。

马克编目的质量问题

中国农科院文献中心从1992年4月开始正式实行中西文计算机编目。软件是自己研制的,简称ZNK编目系统^[1]。中文书目记录格式采用《中国机读目录通讯格式》,即CN MARC,西文采用美国马克。设计要求的。不仅能够接受北图的书目数据和LC光盘数据BiblioFile,并且要求保留和处理同格式有关的一切数据,包括头标、段标号(tags)、指示符(indicators)和代码数据。本系统的工作库的可变段是准可变长的,主库存储方式是完全可变长。可变长的问题完全靠程序处理。同马克格式有关的数据中,有许多

要求编目员提供。对编目员来说,这正是难点。

本编目系统有3个屏幕格式:

第1屏分为两个部分,上5行是填写定长数据的表格,包括需要编目员提供的部分头标项目和编码数据(100—106)字段的有关项目。这些定长项目设置隐含值。下半部是索书号,ISBN号、题名、作者、出版项、主题标题和关键词等常用项目。为了减少子段标识的击键,尽可能多地设分项,并自动赋予子段标识符。例如,出版地、出版社、出版者都分设栏目。题名项因为子项太多,难以在屏幕上罗列出来,需要编目员在打字时同时输入。第1屏幕的某一项目(例如题名,超过254字节时)可在第2屏继续。

第2屏幕是输入其他项目的设计:每行分为段标号、指示符和字段数据3栏。数据栏只能在开头处自动赋值“\$a”;其他必要的子段,标识,也需要编目员在打字时同时输入。一行写不完的可下续一行。续行没有限制。屏幕填满之后,自动翻页,页数也没有限制。

第3屏是为多卷集的子目项设计的,有分卷号、分卷名、页数、价格、ISBN号和ISSN号等项目,每一卷占用一屏(页),按换页键(PageDown)著录一个卷册,没有限制。入库程序对上述数据进行编辑处理,编制各种索引(检索点)之后,对各项数据进行去除无用空格、密集存储。

上述性能,对于一个编目水平较高、管理严格的图书馆已经符合要求了。

然而,在我馆使用1年多之后,发现其中绝大多数的书目记录不符合要求。主要问题如下:

1. 没有提供正确的定长(代码)数据。如著录字典、辞典时,“内容性质”没有填写“e”。又如,图表(例证)设隐含值“a”,说明有图表。但在形态载体项中没有这一说明。如此等等。

2. 缺少必要的子段标识。如题名项中经常需要用“\$e”、“\$v”来作为副题名和说明

卷册的分隔符。而书目记录中有副题名或不同卷册时, 缺少这些符号。

3. 缺少必要的指示符。如丛书项在第 2 屏幕输入, 没有指示符说明该丛书名是否作检索点。

缺少这些东西对卡片目录并无妨碍, 然而对于机读目录来说是缺少马克格式的必要数据。缺少这些数据将会减少书目数据的效用和系统性能的发挥。例如, 在头标中的书目等级代码 m 表示该记录属一般图书, s 表示连续出版物, j 表示录音资料(音乐)。有了这些代码, 以一个字母 m 或 s 就可以从中抽出全部图书或期刊。以“e”作为内容性质的识别条件, 就可以从全部馆藏中把字典、辞典从中区分出来。在 CN MARC 中, 团体名和会议名的字段号都是 710, 区别两者的符号是第 1 指示符的值, 0 代表团体, 1 表示会议。缺少第 1 指示符, 就无从区别。层次等级代码、丛书项的指示符对处理丛书是重要的, 否则就无法把分散著录的一套丛书集中起来。有的编目员把总丛书名作为正题名, 分卷书名作为卷册题名, 固然可以达到集中的目的, 却忽视了必要的分散。丛书是以一个学科范围或读者群为目的作为一套书出版的, 其中不同的卷册在专业上可以是完全不同的对象。集中, 主要是管理上的需要; 分散, 主要是读者的需要。甚至多卷集中各卷也有自己的专门论题。这时候, 也只有以分卷名作为正题名时才能突出其个性。美国的机读目录中, 有许多多卷书是作为丛书予以著录的, 不无道理。

产生上述问题的原因之一就是马克编目比手工著录卡片增加许多工作量, 而且其中有些问题确实比较麻烦。既然没有这些数据也可以打印出符合要求的卡片来, 编目员也就感到满足了。而负责质量把关的人也没有发现这些问题, 这就值得引起足够重视了。

ZNK 编目系统的改进

考虑到这个问题的普遍性, 解决这个问题具有普遍意义, 我们决定从提高系统性能方面作出努力。

提示

原设计的第 1 屏幕格式上只提示著录项目名称, 如“书目等级”、“日期类型”等定长数据项; 基本著录事项中, 有的有子字段(分项)提示, 有的则没有。这些都以屏幕格式出现。

新版以窗口的方式增加了对马克项目的详细揭示:

1. 提示(定长)代码的码值及其含义。其中有: 书目等级, 出版形式, 图表/例证。日期类型和出版期的填写是最为复杂的项目, 根据 UniMark handbook 一书中的例子予以说明, 供编目员参考。(《中国机读目录通讯格式》一书没有这些提供资料。)改进版对于层次等级的用法也有说明。

2. (变长)著录项(除版次外)每项都有详细提示(包括指示符、子段标识及其用法说明)。

第 2 屏幕因为是选择项, 原设计只列出较为常用的字段号, 例如 101, 225, 701, 710 等段号(其他由编目员自由增加), 没有提示字段名。新版增加了对当前字段名的提示。只要光标所在的行有段号, 屏幕的右上角就显示当前字段名称并说明是否可重复。新版还以窗口的方式增设了字段提示(段号和字段名总表)及子段提示(包括指示符、子段标识及其用法说明)。

操作提示使原有性能得到充分发挥。

1. 本系统自动检测每项定长数据是否在取值范围。例如, 书目等级只能是 m, s, a, c 几个字母中择其一而用之, 如果(那怕是无意中)输入其他字符, 光标就不动了, 新用户对此会茫然不知所措。现在新版给予提示: “光标不动是因为输入非法值, ……”同时自动

显示该项的取值范围,以及不同代码的不同含义,使编目员能够选择一个合法的字符,程序得以往下进行。

2. 题名项中(“\”之后)的责任者子项是用来做显示的,而 700,701 项的著者项才是用来做检索点的。为减少重复击键,软件容许编目员只在题名责任者栏目中打上 700,701 或 710 等段号,就自动把它复制过来。新用户对这些性能不了解而未加利用。现在,只要光标一达到有关栏目,就同时显示这种性能的提示。

3. 对子段标识方便设施的提示。在汉字输入中,小写字母被用作汉字输入代码,要输入一个小写字母就必须改变汉字方式为英文方式,这就大大地降低了打字速度,原设计中容许用“\”代潜“\$”符号,可用大写字母(不必改变汉字方式)输入。打回车键之后,系统自动变为“\$”和小写字母。因为没有提示,编目人员未曾加以利用。新版不断提示这种方便可供利用。

4. 新版增加了自动把题名复制到关键词字段的功能。并提示编目员提取关键词的原则和方法。

5. 如果 ISBN 号打错了,光标不动,改对了就是了。然而,有时错误出于印刷错误,系统根据第 10 位字符检测,不对头,光标也不动。新版提示告诉编目员把该 ISBN 号在 \$z 处(错误的 ISBN 号)栏目输入,之后,程序得以往下进行。

如此等等,一方面使新手在操作上得到及时帮助,同时也保证了数据的质量。

检测

尽管有提示和程序运行控制,编目员疏忽某些问题仍然是可能的。在存盘(完成一款目)之前先对刚才的著录做一些检测,才能达到保证数据质量的目的。

检测的内容有代码数据、指示符、子段标识和著录的方法。举例说明如下:

1. 代码数据检测的项目有内容性质(本

系统称之为文献类型)、文艺作品(文学体裁)代码、图表代码、层次等级代码、日期类型代码。

标测的主要根据是题名。题名中常常有说明文献类型(内容性质或体裁)的词语,例如“字典”、“词汇”、“标准”、“述评”等,可作为判断的根据。会议文献也如此。文学作品可以根据分类号进行测定,也相当可靠。图表的种类在题名中能得到体现者为数不多。检测的根据只凭图表代码值(隐/含)“a”而在形态载体项(本系统叫物理描述项)中又没有 \$C 字段(关于图表例证的说明)时,则要求编目员根据该书的实际情况,或者加以图表说明,或者把“a”改成“y”(y 表示无图表)。日期类型的判断根据很少。层次等级代码主要是根据题名项和丛书项来测定。

2. 指示符的测定针对最容易出错的字段:710(团体/会议)、225(丛书项)、327(内容附注项)、500(统一题名)。

3. 子段标识。著者项中缺少著作方式 \$4,题名中附题名标识 \$e,上下册等卷册标识 \$v,都是常有的。判断过程比较复杂。

4. 著录方法的检测,主要是丛书和多卷集的处理。如果发现 200 项(题名)中有“丛书”字样,系统就向编目员发问:“你在做丛书总片吗?”因为按分散著录的原则,只有做总片时,才能把丛书名著在 200 项。否则,应该把分卷名保留在此项,把总题名等写在 225 项。接着,系统又问该丛书是否做检索点,得到是或否的回答之后,再一步一步往下进行。如果在第 2 屏存在 225 项,系统就检测其指示符和层级代码;等等。

多卷集的处理也有类似的判断。这些都具有辅导编目的性质。

处理技术

提示和判断都要花时间。目前许多单位还用 286 型微机,其本身的速度就较慢,但又

不能不考虑这些用户的实际条件。这是在系统分析和设计过程中要考虑的问题。解决的办法是容许自己决定是否要提示,检测是在编目过程中实时处理,还是在编目结束后批式处理。这样,用户可以根据各自设备的条件决定运行方式。决定的时机有三:一是一次性的系统定义,二是每次开机时决定,三是编目过程中随时决定。这三种方式,从一至三,技术难度逐渐增加,性能也逐级提高,一个熟练的编目员,在一般情况下,不必提示;但不常用的项目,也常常需要帮助。我们采用设置功能键的办法来实现第三方式。在第 1 屏幕设 F5 为自动提示,光标到什么项目,就提示什么项目(有些项目没有提示,例如版次项很简单,不必提示)。F6 为按键提示,即,第一次按下 F6 告诉系统:现在进入按键提示状态;光标到达某一项目,再按下 F6,系统就提示,不按 F6 就不提示。F5 适用于初学者,F6 为熟练的编目员所设置。两者可以随时切换使用,方便灵活。

第 2 屏幕设 F5 为字段提示,F6 为子字段提示。两者都是按键显示。按 F5 时,屏幕显出可能用到的项目名称和段号,共 20 多个项目。这些项目在一般情况已经够用。在个别情况下需要其他项目时,就得查马克表了。按 F6 时,系统根据当前字段(光标处)的段号提示该字段的指示符、各子字段标识及其含义。

第 1 屏幕有两个窗口,一个操作提示窗口,一个提示子字段表。第 2 屏幕也开辟两个窗口,分别用于字段提示和子段提示。

讨论和结论

新程序完成后,恰逢我馆资料工作部门要上机编。资料著录原来很简单,许多地方不符合著录规则。资料员是学农的,未曾受过专

门的编目训练。但由于新系统具备了帮助编目和自动控制的性能,因此比用旧版容易上手得多。可以说新版具有实用和教学工具的两重性能,对于马克编目的推广和机读目录数据的质量保证具有新的意义。同时,我又要说,它不是万能的。因为质量的问题不可能完全由系统的智能来解决。比如说,数据项的提供,凡是必备项,程序可以控制。码值限于字母的代码项目,系统不接受数据数值。没有题名,程序不往下运行,等等。这些都可以控制。但对于具体出版物“有则必备”的项目,以及其他条件更宽的选择项,决定权完全在编目员。在代码数据中,例如,图表代码要查阅全书之后才能决定,日期类型和出版期 I、出版期 II 的著录由编目员查阅书中有关内容或一套书的情况之后才能准确进行。系统只能给予引导和帮助,在人机对话中逐步做出正确的处理。不能武断地决定如何进行。本来,马克格式的许多设置,包括代码数据、字段号、字段标识和指示符等,都是要求编目人员提供一些必要的数据,供系统做进一步的处理的判定条件或根据。这是要求人帮助机器。系统提示和自动纠错,基本上是机器帮人的一种设计意图。由于编目情况的复杂性,有些问题很难由设计人员做简单的决策,因此其中也需要人的帮助。这种机帮人、人帮机的工作模式贯穿于整个系统的运行之中,体现了系统工程中人类因素的设计思想。

参考文献

1 黄益券等. 马克编目和公共查询系统的研制. 农业图书馆情报学刊, 1992, (6): 57~65

黄益券 1961 年武汉大学图书馆学本科毕业。现为中科院文献中心副研究馆员。已发文 10 余篇。通讯地址: 北京白石桥路 30 号, 邮码: 100081

(来稿时间: 1993-11-09。编发者: 刘喜申。)