中国图书馆学报(双月刊) 1998 年第 1 期
ZHONGGUO TUSHUGUANXUEBAO

●朱 岩

书目数据的自动校验

ABSTRACT A good computer software can help people to find and correct errors in data according to their data structure, characteristics and regularity, thus improving the quality of bibliographical data. 2 refs

KEY WORDS Bibliographical records. Automatic checking. MARC. Cataloguing system.

CLASS NUMBER G356.1

怎样充分利用计算机软件的功能,检查、发现书目数据库建设过程中可能产生的差错,提高数据质量,这是许多人许多系统都很关心的问题。

要想让计算机查出书目数据中的所有问题与差错是不可能的。如题名或文摘中的某个文字录错,即造成语义错误,目前计算机技术还难以查出来,还要靠人去辨别。

近年国内外的实践表明,计算机的软件确可帮助人们发现数据中的不少错误,而且这种自动校验,即书目数据质量的自动控制,在数据库建设与维护过程中,尤其是每当登录一批新的数据——无论是本馆自编或是接受外来数据,在发挥着非常重要的防范与把关作用。

计算机软件对书目数据实行自动检验,是根据机读目录的数据结构、数据标识、数据特征以及数据出现的一些规律性来进行的。而这些方面数据的差错往往又是人难以校对出来的。校验可分六个层次进行。

1 代码(CODE)校验

当一个数据库系统接收外来源数据时,首先应进行这一检验。

当前,计算机的文字平台即字符集编码体系还处在向 ISO/IEC 10646 国际通用编码字符集过渡的过程中,因此现实的字符集体系是多种多样的。在我国大陆至少有三套体系:

(1) ISO 646,即 ASCII+ GB 2312-80 体系。它有 94 个图形字符(基本拉丁字母与符号)、32 个控制功能符(以上为单字节)和常用汉字 6763 个、非汉字符号 682 个(以上为双字节)。这是多数微机、小型机使用的文字编码体系。

(2) ISO 646(ASCII) + GBK 体系。即 94 个图形字符(基本拉丁字母与符号)、32 个控制功能符(以上为单字节)和汉字扩充内码规范(为双字节)。后者含有 20952 个中、日、韩汉字,28 个汉字构件,13 个汉字结构符,139 个图形符号。GB 2312-80 中的 6763 个汉字为其中的子集,其代码与 GBK 保持一致。这就是 WINDOW 95 中国版的文字平台。目前已被一些应用系统采用。

(3) ISO/IEC 10646 国际通用多八位编码字符集(也称 UNICODE)。即世界上拼音文字与表意文字(汉字)以及各种符号统一编码的文字平台,也称为大字符集的平台。其中含有近万个全世界拼音文字和各种符号,20902 + 6557(扩充) = 27459 个中、日、韩汉字。微软等一些著名计算机公司正在开发这一平台的产品。

此外我们还可能遇到来自台湾的采用 CCCII 或 BIG5 文字平台的书目数据。

接受外来书目数据时，首先应检验与确认数据所采用的字符集编码体系是什么，不然本系统可能识别不了这些数据。

如果源数据采用 CNMARC 格式，则应检查 100 字段第 26～29 位和第 30～33 位（从 0 记位）的字符集标识，并与实际数据的 CODE 核对，判断数据所使用的字符集体系。如果源数据未采用 CNMARC 格式，则应根据数据方提供的格式文本核对，或直接对实际数据测试，判断其文字 CODE 体系。

2 格式结构校验

当一个数据库系统接受外来数据特别是第一次接受某一单位的数据的时候应该进行这种检验，以防止不符合 MARC 结构的数据进库。

校验的内容有：

（1）采用的数据格式。即使源数据说明采用的是 CNMARC 格式，但由于认识、理解上的原因也难以保证外来数据不出偏差。至于来自美国、英国、加拿大的书目数据，一般都说是符合 USMARC 格式，但英、加的 MARC 与美国的格式仍有差别。日本 MARC 结构则更特殊，既不是现在 UNIMARC，也不是 USMARC。格式搞不清楚，是无法利用外来数据的。当对方没有提供格式文本时，应从识读原始数据中确认其格式结构。

（2）检验记录头标区、地址目次区、数据字段的结构。它包括校验记录头标区的长度是否为 24 倍，地址目次区的字段号、字段长度是否与数据字段区的实际字段的字段号和长度一致，并由此核算字段起始字符位置的记数是否正确。当然，这项检查并不一定对一批记录中的每个记录都进行检查，可以抽样检查。还应校验字段分隔符和记录分隔符是否采用了机读格式规定的 ISO 646 中的信息区分控制符 IS_2（1/13）IS_3（1/12）。

3 必有字段检验

一个书目记录必须具备必有字段才可称为正式记录，否则只能是非正式记录或不完整记录。这是应该进行的一项基本检验。

根据 CNMARC 格式规定，一个书目记录必有的字段是：

001 记录标识号
100 通用处理数据
101 作品语种
120 编码数据字段（测绘资料必备）
123 编码数据字段（测绘资料必备）
200 题名与责任说明
206 资料特殊细节项（测绘资料必备）
801 记录来源

其它格式的机读目录也都做了相应的必有字段规定，一些数据库系统内部也会增加必要的管理信息字段，如馆藏 905 字段。对于一个系统而言，经过预编准备正式进库的数据也应增加这种检验。

4 数据标识检验

所谓数据标识是指机读格式规定的字段标识（字段号）、字段指示符和子字段标识。

校验的内容是：

（1）字段号校验。字段号必须是三位十进制数字型数据，否则就是错误的，应加以提示。所有的字段号都应是机读格式文本中规定的记录号，超出其范围的字段号则是错误的。

（2）字段指示符校验。除 001 和 005 两个字段不设指示符外，其余字段均设两位指示符。此外还可以根据格式规定检查具体的指示符数据：如 010，011，014，020，021，022，040，091，092，094 各字段，除 327 字段外的 3- - 各字段，100、102～121、124～192 各字

段, 205、206、208～215、230 各字段, 602～605、607、615～692 各字段, 720～722、801、802 和 905 字段, 这些字段的两位指示符全部为空格, 如若不是则为错误。

再如 101 字段第 1 位指示符只能是 0 或 1 或 2, 第 2 指示符只能是空格, 若实际数据超出这一范围, 则视为出错。依此类推, 对其它有具体规定指示符数据的字段, 均可以这样检查。在一些单位机读编目中, 常有忽略字段指示符的现象, 进行检查是必要的。

(3) 子字段标识检验。除 001 和 005 两个字段无子字段标识外, 其余全部字段都有子字段标识。子字段标识符号为两倍, 第 1 位按机读格式规定, 只能是 ISO 646(ASCII)的信息区分功能符 IS$_1$(1/15), 而不应是文字图形符号。第 2 位标识符取值范围为小写字符 a～z, 或A, 或 1～4。超出上述范围的值则为出错。

在特定字段内, 子字段第 2 位标识符的取值均规定了具体范围。

如 200 字段, 子字段第 2 位标识符的范围为: a、b、c、d、e、f、g、h、i、z、v、A。超出这个范围则为错误。

如 210 字段, 子字段第 2 位标识符的范围为: a、b、c、d、e、f、g、h。超出这一范围则为错误。

如 600 字段, 子字段第 2 位标识符的取值范围为: a、b、c、d、f、t、x、y、z、2、3。超出这一范围则视为错误。

如 700 字段, 子字段第 2 位标识符取值范围为: a、b、c、d、f、g、p、3、4、A。超出此范围则为错误。

如 710 字段, 子字段第 2 位标识符取值范围为: a、b、c、d、e、f、g、h、p、3、4。超出此范围则为错误。

5　数据特征校验

该项校验主要是利用数据在长度、使用字符种类等方面有一定限度的特点实施检查, 辅助人们发现错误。

(1) 非法空格校验。所有数据字段的数据在子字段标识符(两位)之后均不允许出现空格。若出现空格则为错误。出错可能有两种情况:

第 1 种: 子字段标识符之后全部为空格, 即该子字段没有具体数据, 这种字段属垃圾, 应予清除。

第 2 种: 子字段标识符之后出现了一个或几个空格, 然后才是具体数据。这也是不正常数据。

上述两种状况均视为不允许的错误, 应加以检查与提示改正。北图在过去对 ISDS 数据库检索时就发现此类数据错误, 曾造成检索软件不能正常运行。

(2) 定长数据检验。在机读目录中, 虽然相当数量字段的数据是变长的, 即使用文字的多少(字节的多少)是不固定的, 但仍有一部分字段, 如号码性字段(子字段)或编码性字段是固定长的。因此可以利用这一特点检验。防止数据加工中可能产生的错误。CNMARC 规定有固定长字段。

001 字段北图规定为 10(字节)。

010 字段的 $a ISBN 子字段, 不含连接符为 10 位, 含连接符为 13 位(字节)。

011 字段的 $a ISSN 子字段, 不含连接符为 8 位, 含连接符为 9 位(字节)。

020 国家书目号字段的 $b 子字段, 中国为 10 位(字节)。

040CODEN 字段的 $a 子字段, 数据长度为 6 位(字节)。

100 字段的数据长度为 36 位(字节)。

101 字段中的 $a, $b、$c、$d、$e、$f、$g、$h、$i、$j 每个子字段中的数据均为 3 位(字节)。

102 字段的 $a 出版或制作国别数据, CNMARC 规定为 2 位(字节)。

105 字段的数据长度为 13 位(字节)。

110 字段的数据长度为 11 位(字节)。

115 字段 $a 子字段为 20 位(字节), $b 子字段为 15 位(字节)。

116 字段为 18 位(字节)。

117 字段为 9 位(字节)。

120 字段为 13 位(字节)。

121 字段的 $a 字段为 9 位(字节), $b 的子字段为 8 位(字节)。

125 字段为 4 位(字节)。

126 字段的 $a 为 15 位(字节), $b 为 3 位(字节)。

127 字段的 $a 为 6 位(字节)。

128 字段的 $a、$b、$c 各子字段中的数据均为两位字符(字节)。

130 字段的 $a 为 11 位(字节)。

191 字段为 7 位(字节)。

192 字段的 $a、$b、$c, 分别为两位字节。

软件均可以利用该长度限定实施检验, 违犯者即属错误。

(3) 数据类型校验。机读目录格式规定了一些子字段的数据, 或一些固定长数据中特定位置的数据仅限于使用小写字母型数据, 或数字型数据, 或大写字母型数据等, 或者一些号码型数据是按照规定的算法生成的, 如果数据加工过程中录入了不符合规定的错误数据, 则可以通过软件查出。

例如: 100 字段中前 8 位, 只能是 0～9 数字型数据, 第 9 位数据只能是 a～j 小写英文字母中的 1 位数据, 第 10～17 位只能是 0～9 数字型数据或空格, 第 18～20 位为 a、b、c、d、e、k、m、u 范围的小写英文字母型数据, 其它见格式有关规定。

101 字段各子字段中的数据只能是 3 位小写英文字母。

105 字段第 1～8 位只能是小写英文字母。第 9～11 位为 0 或 1 数字。第 12～13 位为小写英文字母。

还有 010 ISBN 和 011 ISSN 中的 $a 子字段的数据, ISBN 前 9 位(不含连接符)和 ISSN 的前 7 位(不含连接符)数据为 0～9 数字型的数据, ISBN, ISSN 的最末一位数据可能是 0～9, 也可能是 X(表示值为 10)。ISBN 与 ISSN 的每位数据构成均有规定(国际和国家标准), 应按其规定检验, 发现有错号应加以提示, 判断是分配错还是印刷错还是录入错。

(4) 数据项重复性检验。机读目录格式规定, 一些字段和子字段不可以重复出现, 或可以重复出现, 凡规定不可重复出现的, 就可以通过软件检查, 发现重复即提示为错误。

(5) 出现规律校验。有些数据字段出现是伴随性的, 即当一个字段或子字段出现, 就必然有另一个字段或子字段伴随出现。软件可以利用这类伴随现象检验, 防止漏编某一数据字段。

如 200 题名与责任说明字段中, 当出现 $d 并列题名子字段, 则一定要有 $z 并列题名语种子字段。并且还一定要有 510 并列题名作为检索点的字段。许多系统还规定必须有 $a 正题名的汉语拼音子字段 $A 同时出现。

又如 200 题名与责任说明字段中, 当出现 $f 子字段, 则一定会出现 7- - 责任者作为检索点字段。如果没有, 则为漏编。反之, 如果 200 字段中没有 $f 子字段, 也可能有 7- - 字段, 因编目员可能从题名页以外的其它信息源发现了责任者而补做在这里。这种情况是少见的。

当一个编目单位决定对丛书、多卷集采取分层著录, 而记录中有 225(丛书)字段时, 则书目记录中就应有 411 或 462、463 字段。若没有, 则属于漏做。软件应提示编目员补做。

6 记录排它性校验

对于一部文献, 在一个数据库中只能有一个描述该出版物特征的书目记录, 不能有

一个以上书目记录, 不然就造成书目记录重复, 影响检索。因此对于书目数据库必须进行查重与剔重的维护作业。

在数据库建设过程中, 总会有更新的记录进入库中, 或存在库中的是出版前预编记录(在版编目记录), 当出版物正式出版后, 根据出版物编制的正式完整记录要登录到库中来。上述这些记录都必须取代库中原有记录, 以保证库中的数据质量和记录的惟一性。

进行这项检验是根据记录头标中第 5 位(从 0 记位)记录状态的代码来判定的。当该状态为 c(经修改的记录)、d(被删除的记录)、p(出版前记录, 现已为完整的正式记录)时, 要将这些状态的书目记录取代库中原有状态为"n"的记录。当然, 在检验记录状态代码的同时, 还必须核对记录中的题名或 ISBN 或责任者等项, 看两个记录是否为同一出版物的书目记录, 相同时才可替代, 以便保证准确地替代记录。当一个书目数据库经常接受外来数据时, 这项检验与替代作业是必不可少的。

利用软件对书目数据校验, 可以采取批处理方式, 也可以采取实时处理方式进行。接受外来数据或接收本系统成批的在编记录经过审校正式进库时, 均应采取批处理方式对数据进行校验。凡校验中发现数据有误或有疑问, 均应加以提示, 供人工核对与修改。对于正确的记录也应统计出数量提供数据库管理者备案。所谓实时处理, 即在编目过程中编目员每送一个字段(或子字段)或一个记录, 软件就对其检查, 检查结果当时就提示给编目员, 以便及时修改。采取哪种方式, 由系统设计者根据系统用户的实际需求决定。但无论哪种方式, 软件检验都应对数据中的问题与差错有简洁、明了而又具体的提示, 这样才便于编目人员判断、改正。一个成熟的图书馆编目软件系统都应具备这种数据自动校验的功能, 由于经验不足可能短时间校验功能还不周全, 但应根据实践的不断总结、积累逐步完善起来。这样才可减轻编目人员繁琐的校对劳动, 提高数据的质量, 受到用户的欢迎。

参考文献

1　中国机读目录格式使用手册编委会. 中国机读目录格式使用手册. 北京: 华艺出版社, 1995
2　ISDS. 记录的校验和数据库管理. 1986. 12

朱　岩　北京图书馆教育中心主任。通讯地址: 北京白石桥路 39 号。邮编: 100081。

(来稿时间: 1997. 8. 21 编发者: 赵薇)

(上接第 17 页)

信息服务业及其技术发展动向. 情报学报, 1996, 15(1)
4　马费成. 情报学的进展与深化. 情报学报, 1996, 15(1)
5　马费成. 论网络时代的图书情报教育. 图书情报知识, 1996(4)
6　陈光祚等. 科技文献检索. 武汉: 武汉大学出版社, 1987

马费成　武汉大学图书情报学院院长, 教授, 博士生导师。主要从事情报学理论方法和信息经济学领域的教学科研工作。出版著作 7 部(含合作), 在国内外发表论文 60 余篇。通讯地址: 武汉市, 邮编 430072

陈　锐　武汉大学图书情报学院情报学专业博士生。在国内外发表论文 40 余篇。通讯地址同上

(来稿时间: 1997. 8. 12。编发者: 李万健)