Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

联合编目系统中文书目库增量数据批更新中的数据转换分析与实现

□ 张楠 / 国家图书馆 北京 100081

摘要：文章结合全国图书馆联合编目中心业务特点，对其联合编目系统中文书目库增量数据批更新过程中的数据转换环节进行梳理、分析，针对不同类别的转换规则，用perl语言高效实现数据转换处理。

关键词：全国图书馆联合编目中心，联合编目系统，中文书目库，数据转换
DOI：10.3772/j.issn.1673—2286.2012.07.007

在全国图书馆联合编目中心[1]（Online Library Cataloging Center，以下简称“联编中心”）新版联合编目系统（Union Cataloging System[5]，以下简称“UCS系统”）项目建设过程中，为了充分利用国家图书馆丰富的馆藏和优质的书目数据资源，将国家图书馆业务自动化集成系统Aleph500系统[2]（以下简称“Aleph系统”）中的数据作为UCS系统的初始书目数据，并与旧版UCS系统中的各成员馆存量数据进行整合，形成新版UCS系统的书目和馆藏数据库。系统正式投入运行之后，为了降低成员馆及用户的编目成本，提高编目工作质量，避免书目数据资源的重复建设，需将国家图书馆Aleph系统中每日生产的书目数据，作为UCS系统的增量数据，及时更新至UCS系统中，方便各成员馆下载使用，实现书目数据资源的共建共享。UCS系统增量数据的批更新涉及数据下载、数据转换、数据查重、数据装入等各环节。由于业务流程的差异，国家图书馆Aleph系统生产的书目数据，不能原样导入UCS系统中，需要按照UCS系统的书目数据格式及业务要求，进行相应的数据转换。本文对UCS系统中文书目库增量数据更新过程中的数据转换规则进行研究、分析，结合perl程序实例，给出相应的实现方法。

1 UCS系统数据组成概述

UCS系统是联编中心在参考了国内外联合编目系统的特点和成员馆用户的意见后，基于艾利贝斯公司的Aleph系统开发的新的联合编目系统。UCS系统旨在建立全国范围的联合书目数据库和联合馆藏数据库，为各成员馆和广大读者提供服务，并为今后馆际互借和文献传递的进一步发展提供条件。

联合书目数据库分为中文书目库ucs01和外文书目库ucs09，分别存放联编中心各成员馆上传的书目数据，并供各成员馆下载，以实现书目资源的共建共享。联合馆藏数据库ucs60，存放书目数据下链接的馆藏记录。馆藏记录是描述成员馆收藏文献情况的概要数据，包括相关的书目数据库代码和系统号、馆藏机构代码、索取号、收藏卷期、本地文献唯一标识符、缺期馆藏信息、馆际互借方式等数据内容。UCS系统中一条书目记录可链接多个馆藏记录。

2 ucs01库增量数据批更新过程介绍

联编中心UCS系统ucs01库的初始书目数据取自国家图书馆的中文书目库nlc01。在UCS系统建设过程中已将nlc01库的存量数据整体装入UCS系统的ucs01库中。UCS系

统建成投入使用后，对ucs01库数据更新采用了增量更新方式，即将国家图书馆Aleph系统nlc01库每日生产的书目数据，作为UCS系统ucs01库的增量数据，及时更新至UCS系统中，方便各成员馆下载使用。批更新程序每日凌晨3：00自动从nlc01库下载增量数据，按照既定的数据转换规则，转换处理后上传至UCS系统中；再经过查重后，批量装入ucs01库中，同时生成相应的馆藏记录。批更新过程如下：

2.1 增量数据的下载

根据Aleph系统中文书目库nlc01的Z13表Z13_UPDATE_DATE字段获取每日更新的系统号，由系统号根据p_print_03[3]获取书目顺序文件，包括增加、修改、删除的日更新数据。

将下载的所有数据，按照下述条件生成3个文件：

（1）删除记录文件：*.del。条件：记录中有“DEL字段”。

（2）错误文件：*.err。条件：记录中没有200字段，或200字段没有$$b子字段。下一步将错误文件返交国家图书馆检查和修改。

（3）一般文件：*.new，条件：剔除（1）（2）之后的所有剩余记录。下一步进行数据转换处理。

2.2 数据转换处理

将章节2.1中（3）步生成的一般文件*.new，按照数据转换需求进行数据转换处理。由于数据转换需求繁多、复杂，因此需对这些需求进行梳理、分析，分类出不同的转换规则，再按照不同的规则进行编程实现。本文章节3将对这一步进行详细介绍。

2.3 数据查重

(1)先用035字段来进行查重，一对一重直接覆盖，不重进入(2)查重，如有一对多重的保留待人工核查。

(2)035不重数据，进行UCS-1查重，不重直接新增，一对一重增加馆藏，如有一对多重的保留待人工核查。

2.4 数据装入

将经过转换处理、查重后的文件用p_manage_18[3]装入UCS系统的ucs01库，同时根据书目数据910字段生成馆藏记录。

3 数据转换规则分析

UCS系统和国家图书馆Aleph系统业务流程差异，使得从Aleph系统获取的日更新书目数据不能直接导入UCS系统的书目库中，需要按照UCS系统的业务要求，对源数据进行转换处理。笔者在实际工作中，通过与联编中心业务人员反复分析、讨论，将纷繁、复杂的数据转换需求进行梳理、分类，整理出如下部分的数据转换规则，包括文献类型转换、特定字段的增加、删除和修改等规范化方面的数据转换规则。

3.1 转换文献类型

中文书目数据采用CNMARC格式，CNMARC记录的文献类型放在200$$b子字段。对于同一文献，ucs01库定义可能和nlc01库的定义不同，例如，对于地方志文献，nlc01库的200$b包括专志、地方文史、地方史志，而在ucs01库中200$b统一著录“地方志”。

通过对nlc01库的书目记录200$b进行整理、分类，总结出如表1所示的文献类型转换规则。

3.2 增加字段

为适应联合编目业务需求，UCS系统为中文书目数据构建了一些特殊字段，这些字段是nlc01库中的书目数据不具备的，如记录上载馆信息和中心唯一控制号的049字段，记录上载馆代码及其书目数据系统号的035字段，用于自动生成馆藏记录的910字段，记录书目数据状态的QUA字段等。对于这些特殊字段，需要按照联编中心的业务规则，增加到从nlc01库下载的书目记录中。增加的具体字段及其内容如表2。

3.3 删除字段

nlc01库书目记录的一些字段对于UCS系统是不需要的，如排架分类号090字段、索取号096字段等，转换处理中需要将这些字段删除。需要删除的字段具体如下：

（1）删除除“FMT”、“LDR”、“OWN”、“STA”和“QUA”外的所有非数字字段；

（2）删除090和096字段；

（3）删除除“905”外的所有9XX字段；

（4）删除下列字段：890、817、810、808、803、800、703、770、771。

表1 文献类型转换规则表

文献范围	转换前nlc01库的文献类型	转换后ucs01库的文献类型
台港期刊	期刊	海外中文期刊
台港报纸	报纸	海外中文报纸
学位论文	博士后论文	博士后报告
电子资源	电子资源.CD	电子资源
电子资源	电子资源.VCD	电子资源
电子资源	电子资源.DVD	电子资源
电子资源	电子资源.MP3	电子资源
缩微文献	缩微制品	缩微品
缩微文献	缩微胶片	缩微品
地方志	专志	地方志
地方志	地方文史	地方志
地方志	地方史志	地方志
家谱	家谱	专著

3.4 修改字段

nlc01库书目记录中的一些字段排列位置、子字段内容等不同于ucs01库中的书目记录，转换处理中需要将这些字段按照UCS系统中文书目记录的字段格式进行修改。需要修改的字段具体如下：

（1）将200、5XX、7XX拼音子字段$$9紧接在$$a子字段之后；

（2）将7XX字段$$f子字段的“~”改为“-”；

（3）将100字段$$a26-29位改为“50^^”；

（4）将OWN字段的$$a改为“OLCC”。

4 数据转换程序实现说明

考虑到数据转换规则的可变性，在程序实现时，将本文章节3中不同类别的数据转换规则定义在一文本文件，即转换规则文件中。这样，一旦转换规则发生变化，只需修改规则文件，无需改动程序，从而保证了程序的可扩展性。

由于转换规则中涉及复杂的字符处理，为了保证程序执行效率，采用perl语言[4]实现程序开发。不同类别的数据转换规则对应的程序实现如下：

4.1 转换文献类型的处理

按照文献类型转换规则表（表1），进行文献类型的转换，具体程

表2 增加字段表

字段号	指示符	子字段数据	备注
035	^^	$$a(A100000NLC)nnnnnnnnn	(1) 第一个035字段； (2) “nnnnnnnnn”为当前记录NLC01系统号，“^”表示空。
049	^^	$$aA100000NLC$$bUCS01000000000 $$cnnnnnnnnn$$dNLC01	(1) 应删除原记录所有049字段； (2) “nnnnnnnnn”为当前记录NLC01系统号，“^”表示空。
801	^2	$$aCN$$bOLCC$$cyyyymmdd	(1) 最后一个801字段； (2) $$c是批处理当天日期。
910	^^	$$aA100000NLC$$lNLC01 $$nnnnnnnnnn$$z2	(1) 应先删除原记录所有910字段； (2) 生成910字段的条件：（a）对于有Z30字段的记录，存在有效单册，即：Z30字段有“$$I$$”或“"$$I $$”字串；（b）或有“LOC”字段；（c）或有“905”字段； (3) “nnnnnnnnn”为当前记录NLC01系统号，“^”表示空。
CAT	^^	$$aOLCCAUTO $$b51$$cyyyymmdd $$lUCS01$$h0900	(1)在OWN字段之前； (2)$$c是批处理当天日期。
QUA	^^	$$aA	(1)记录最后一个字段。

序实现如下：

```
…
…
$key=substr($_,10,3);//取出记录的字段
    if($key eq '200'){//判断是否为200字段
//将200字段的$b子字段按照转换规则进行转换
if(s/(\$\$b)(.*?)($|\$)/$1.replace($2).$3/e){
            if($log){print LOG $_;}
        }

sub replace{
   my $t=shift;
   $t=$b200{$t}||$t;
   return $t;
}
…
…
```

4.2 增加字段的处理

按照增加字段表（表2），将相应字段增加到原始数据中，具体程序实现如下：

```
…
…
$line=$_;
    $id=substr($line,0,9);//取出记录的系统号
    $key=substr($line,10,3);//取出记录的字段
     if($oid ne '' && $id ne $oid){$new=1;&output;}
……
……
sub output{
   print $doc;
my $f910='';
//判断原始记录是否满足生成910字段的条件，满足则构造910字段，用于自动生成馆藏记录
    if(exists $types{$fs{200}}){
      $s=$types{$fs{200}};
        if($s eq '' || ($s eq 'S1' && $fs{z30}) || ($s eq 'S2' && $fs{loc}) || ($s eq 'S3' &&substr($fs{head},8,1) ne '1' &&substr($fs{head},7,1) ne 'a')){
$f910="$oid 910 L \$\$aA100000NLC\$\$lNLC01\$\$n$oid\$\$z2\n";
            if($log){print L910 $f910;}
        }
    }
//构造035、049、QUA等字段，增加到原始记录中
print qq#$oid 035 L \$\$a(A100000NLC)$oid
$oid 049 L \$\$aA100000NLC\$\$bUCS01000000000\$\$c$oid\$\$dNLC01
$oid 801 2 L \$\$aCN\$\$bOLCC\$\$c$today
$f910$oid CAT L \$\$aOLCCNLC\$\$b51\$\$c$today\$\$lUCS01\$\$h0005
$oid QUA  L \$\$aA
#;
   $doc='';%fs=();
}
…
…
```

4.3 删除字段的处理

按照章节3.3中的删除字段处理规则，将相应的字段从数据中删除，具体程序实现如下：

```
…
…
%del=("049",1,"090",1,"096",1,"890",1,"817",1,"810",1,"808",1,"803",1,"800",1,"703",1,"770",1,"771",1);
%nodel=("FMT",1,"LDR",1,"OWN",1,"STA",1,"QUA",1);
$line=$_;
   $id=substr($line,0,9);//取出记录的系统号
   $key=substr($line,10,3);//取出记录的字段
…
…
   $oid=$id;
    if($del{$key}){print LOG "$line";next;}//删除$del中的字段
    $key1=substr($key,0,1);
    if($key1 eq '9' && $key ne '905'){print LOG "$line";next;}//删除除“905”外的所有9XX字段
//删除除“FMT”、“LDR”、“OWN”、“STA”和“QUA”外的所有非数字字段
if(!$nodel{$key} && $key=~/\D/){print LOG "$line";next;}
…
…
```

4.4 修改字段的处理

按照章节3.4中的修改字段处理规则，对数据进行修改，具体程序实现如下：

```
…
…
    %fixed=("LDR",1,"001",1,"005",1,"100",1,"105",1,"110",1,"207",1,"461",1,"462",1,"327",1 );
    $line=$_;
```

```
        $id=substr($line,0,9);//取出记录的系统号
        $key=substr($line,10,3);//取出记录的字段
        $key1=substr($key,0,1);
    …
    …
    if($key1 eq '7'||$key1 eq '5'||$key eq '200'){$line=~s/\$\$9[\^](\$|$)/$1/;}
        if($key1 eq '5'||$key eq '200'||$key1 eq '7'){
    #将200、5XX、7XX拼音子字段$$9紧接在$$a子字段之后，但若200字段有多个$$9时，则不作修改，仅给出报告
        @t=$line=~/\$\$9/g;
        if($#t>0){if($log){print M9 $line;}}
    elsif($#t==0 && $line=~s/(\$\$9.*?)(\$|$)/$2/){
            $sf9=$1;
            $line=~s/(\$\$a.*?)(\$|$)/$1$sf9$2/;
        }
        }

        if($key eq '200' && $line=~/\$\$b(.*?)($|\$)/){$fs{200}=$1;}
    //修改原始数据中的100、OWN、7XX字段等
    elsif($key eq '100'){$line=~s/(\$\$a)([^\$]{26})[^\$]{4}/$1${2}50^^/;}
    elsif($key eq 'OWN'){$line=~s/(\$\$a).*?($|\$)/${1}OLCC$2/;}
    elsif($key1 eq '7'){$line=~s/(\$\$f[^\$]*?)~/$1-/;}
    elsif($key eq 'LOC'||$key eq '905'){$fs{loc}=1;}
    elsif($key eq 'Z30'){
            if($line=~/\$\$I(SB|BD|TE|IP|QD|DT|CT)?\$\$/){$fs{z30}=1;}
    }
    elsif($fixed{$key}){
        $line=~s/\-/^/g;
    }
    …
    …
```

5 结束语

截至2011年底，联编中心的成员馆已接近1000家，书目数据年下载量已达百万级。UCS系统作为全国图书馆联合编目的系统平台，其增量数据的更新不但要及时、高效，而且要保证数据质量。数据转换作为数据更新过程中的重要一环，其处理效果对联编中心的数据质量起到至关重要的作用。通过对UCS系统ucs01库中文书目数据与国家图书馆Aleph系统nlc01库的中文书目数据进行反复比较、研究，总结出UCS系统中文书目增量数据更新中的数据转换规则，并在实践中利用perl语言[4]高效实现了数据转换处理程序。经运行实践证明，完成UCS系统上万级中文增量数据转换，只需不到30秒的时间，从而大大提高了增量数据更新的总体效率。本文特将这部分工作中的研究、实践整理成文，以便与同行分享。

参考文献

[1] 全国联合编目中心简介[EB/OL].[2012-03-20].http://olcc.nlc.gov.cn/about-zxjj.html.
[2] 艾利贝斯有限公司北京代表处-公司产品-ALEPH[EB/OL].[2012-03-23].http://www.exlibris.com.cn/product/aleph/index.asp.
[3] ExLibris Ltd.ALEPH 20 User Guide[M].2009.
[4] SCHWARTZ R L, CHRISTIANSEN T.Perl语言入门[M].洪峰,译.第二版.北京:中国电力出版社, 2000.
[5] ExLibris Ltd.Union Cataloging System Usage and Setup [M]. 2010.

作者简介

张楠（1979-），国家图书馆信息网络部。研究方向：计算机技术在图书馆领域的应用。E-mail:zhangnan@nlc.gov.cn

Discussion and Implementation of the Data Conversion in Batch Updating Incremental Data from Chinese Bibliographic Base of Union Cataloging System

Zhang Nan / National Library of China, Beijing, 100081

Abstract: This paper discusses the data conversion in batch updating incremental data from Chinese bibliographic base of Union Cataloging System, and converts the incremental data of Chinese bibliographic base highly efficient by developing the perl programs for each type incremental data conversion.

Keywords: Online Library Cataloging Center (OLCC), Union Cataloging System (UCS), Chinese bibliographic base, Data Conversion

（收稿日期：2012-04-13）