

东北师范大学计算机科学与技术学院

# 图书自动编目

---

## 使用指南

图书自动编目项目组

2012 年 4 月

本文档作为我项目组对已完成系统《图书自动编目》的简明使用手册。涵盖本系统的简介、使用方法和系统设置等主要信息。

# 目录

第一章 概述 .....	2
第二章 录入 ISBN 处理 .....	3
一、    默认自动处理方式 .....	3
1.    登录网站后，首页就是录入 ISBN 处理的页面。 .....	3
2.    默认情况下录入 ISBN，将会自动进行分析处理 .....	3
二、    改变数据源 .....	4
三、    关闭自动分析获取 .....	4
四、    获取状态说明 .....	5
1.    列队中 .....	5
2.    处理中 .....	5
3.    处理失败 .....	6
4.    处理成功 .....	6
第三章 历史记录 .....	8
一、    查看历史记录 .....	8
1.    布局介绍 .....	8
2.    翻页 .....	9
3.    搜索 .....	9
二、    编辑条目 .....	9
三、    删除记录 .....	11
四、    导出数据 .....	11
第四章 系统设置 .....	12
一、    XPath 参数设置 .....	12
1.    中国国家图书馆 XPath 参数设置 .....	13
2.    美国国会图书馆 XPath 参数设置 .....	14
二、    字段名到记录号参数设置 .....	14
三、    记录号到字段名参数设置（MARC 信息） .....	15
第五章 参考文献 .....	16

# 第一章 概述

图书编目是图书馆工作的重要组成部分，传统的手工编目耗时耗力，如果能尽可能的实现自动化、数字化、信息化，能极大的提高工作效率。目前虽然采取了信息化的手段，对许多信息进行了数字化，但是，编目人员进行编目时，不仅仍需手动录入编目信息，并且对网上现有的编目数据没有进行充分利用。我们注意到，可以通过区分书籍的 ISBN 号，在网上查询到书目信息，减少重复人工录入的工作量。

本系统提供一套接口，可以获取[中国国家图书馆](#)，[美国国会图书馆](#)的编目资料，减少重复录入资料的繁琐性，进一步提高信息化程度和提高工作效率。

本项目是 2011~2012 年度，东北师范大学科研立项项目，指导教师杨贵福，项目负责人程颖宇，项目组成员：刘美君、文毅、袁小康、何泽林。

系统分为首页、历史记录和系统设置三个部分。

其中，首页是 ISBN 号录入和获取的工作页面，录入的 ISBN 会被处理，然后添加的数据库中；历史记录显示了所有曾经录入并且正确获取的书目信息，同时可以对这些信息进行编辑、删除、导出等功能；系统设置主要是整个系统的一些工作参数设置，包括数据源 XPath 参数设置，字段名到记录号的对应以及记录号和字段名的对应。

项目核心分析技术使用的是 XPath，能够在数据源网页放生变化的情况下，仅需要修改部分 XPath 参数即可完成系统更新，无需改变代码。表现层使用 AJAX 技术和 HTML5/CSS3 的一些新特性，不但保证了工作效率整个系统也提供了优美简洁的界面。整个系统基于开源免费的 LAMP 体系，部署成本低，可靠性强。

项目推进过程中，得到了各位老师和同学的大力相助，特别是 2010 级图书馆学专业的潘雪丽同学在 MARC 方面的支持，在此表示诚挚感谢。

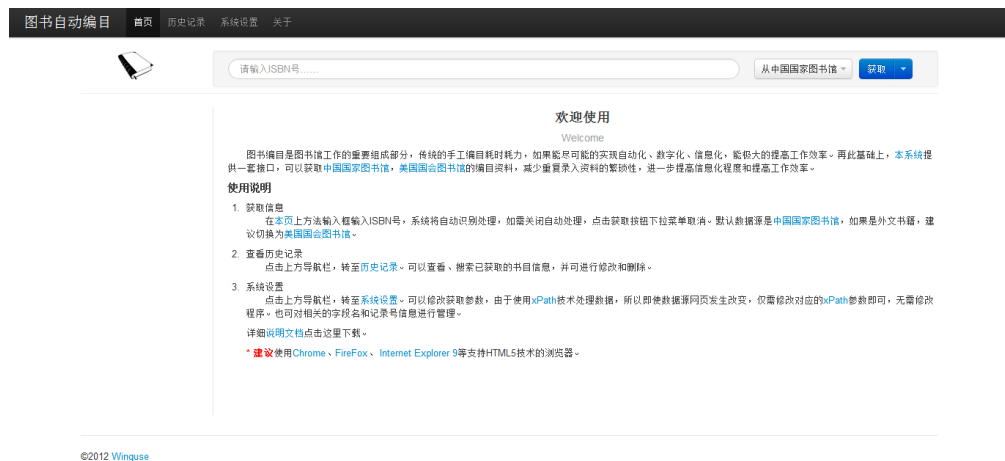
项目前端使用了 jQuery、Bootstrap 项目成果，感谢原有开源工作者的努力。

## 第二章 录入 ISBN 处理

本章节主要介绍 ISBN 录入的处理方式，所有正确处理的结果都会存于数据库中。

### 一、 默认自动处理方式

#### 1. 登录网站后，首页就是录入 ISBN 处理的页面。

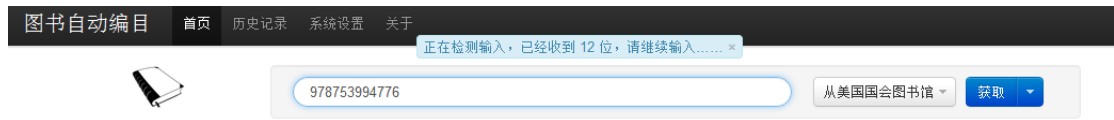


#### 2. 默认情况下录入 ISBN，将会自动进行分析处理

没有输入时，提示输入 ISBN：

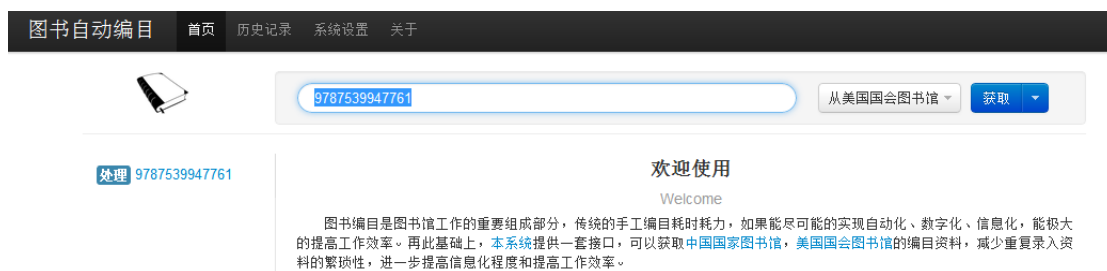


每输入一个数字，开始检测校验：



如果 ISBN 合法<sup>1</sup>，则添加到队列进行处理：

<sup>1</sup>程序进行自动判读的时候，限制 ISBN 的开始不能省略 978，所以请完整输入 13 位 ISBN。



如果获取成功，则显示对应的信息页面：

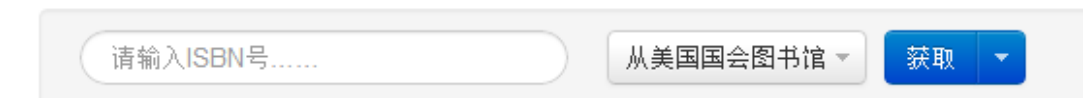


## 二、 改变数据源

点击输入框右边第一个按钮，是个下来菜单，可以切换数据源：



选择对应的数据源，按钮文字改变，表示切换完毕：



## 三、 关闭自动分析获取

点击获取按钮右侧的小三角：



请输入ISBN号..... 从中国国家图书馆 获取

打开自动识别处理  
关闭自动识别处理

欢迎使用

选择关闭自动识别获取，获取按钮变为浅蓝色：



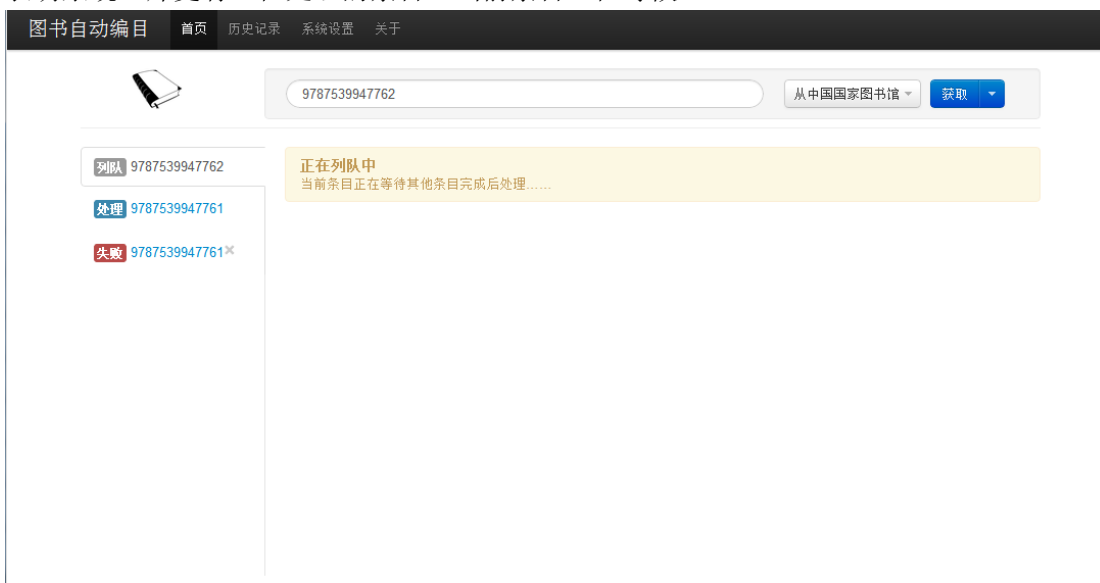
请输入ISBN号..... 从中国国家图书馆 获取

自动获取关闭，每次输入完后，需要点击获取按钮才会开始获取。

## 四、 获取状态说明

### 1. 列队中

表明系统一斤更有正在处理的条目，当前条目正在等候。



图书自动编目 首页 历史记录 系统设置 关于

9787539947762 从中国国家图书馆 获取

正在列队中  
当前条目正在等待其他条目完成后处理.....

列队 9787539947762  
处理 9787539947761  
失败 9787539947761✖

### 2. 处理中

表明正在处理。因为网络以及数据源的问题，所以这个也会需要一点时间。



### 3. 处理失败

有可能数据源并无相关的记录, 例如中国国家图书馆就通常没有外文书籍的记录, 或者是数据源服务器出错, 或者是或许参数需要修改等。



### 4. 处理成功

获取成功会显示相关获取后的结果, 点击查看详细信息, 可以转到历史记录看到相关情况。

图书自动编目

首页

历史记录

系统设置

关于

获取成功!

9787539947761

从中国国家图书馆

获取

成功 通天人物 [专著]  
/ 李佩甫著\*

通天人物 [专著] / 李佩甫著  
ISBN: 9787539947761 / 2012-4-15 13:16:1 / 中国国家图书馆

显示详细信息

记录号	字段名	值
#004	FMT	BK
#003	头标区	-----nam0 22----- 450
001\$a	控制号	005598372
#002	记录最后处理	20120222164346.0
010\$d	价格	978-7-5399-4776-1 : CNY39.90
100\$a	通用数据	20120208d2011 em y0chiy50 ea
101\$a	正文语种	chi
102\$a	出新国别	CN320000
105\$a	专著编码	y z 000ay
215\$a	页数册数	r
215\$a	页数册数	388页 ; 24cm
200	题名及责任者项	通天人物 [专著] / 李佩甫著

注：获取结束后的结果，可以点击右侧标签栏标题后的小叉关闭。



# 第三章 历史记录

本章主要介绍历史记录的管理，包括对书目信息的查看、编辑、删除、导出等操作。

图书自动编目

首页

历史记录

系统设置

关于

查询成功! 当前第 1 页, 共 3 页, 22 条记录 >>

请输入搜索关键词.....

通天人物 [专著] / 李佩甫著

Beginning EJB 3 application development : from novice to professional / Raghu R. Kodali and Jonathan Wetherbee ; with Peter Zadrozny

Java程序设计教程 [专著] = Textbook for programming in Java / 雍俊海编著

嵌入式系统基础教程 [专著] / 俞建新, 王健, 宋健编著

信息资源编目 [专著] / 段明远编著

Struts 2完全学习手册 [专著] / 王伟平等编著

TCP/IP illustrated, Volume 1, The protocols = TCP/IP 详解, 卷1, 协议 / W. Richard Stevens. [monograph]

生命与使命同行 [专著] 走近商榷全国教书育人楷模 / 教育部新闻办公室编撰

模拟电子技术基本教程 [专著] / 华成英主编

C++ master reference / Clayton Walnum.

通天人物 [专著] / 李佩甫著

数据来源: 中国国家图书馆 [库] / ISBN#: 9787539947761 / 获取时间: 2012-04-15 13:16:01 / 最后检验时间: 2012-04-15 13:16:01 / UserID: 1 男

记录状态: 不需更新 备注否: 否

导出数据

高级查询处理 编辑 删除

记录号	字段名	值
#004	FMT	BK
#003	头标区	---nam0 22--- 450
001\$a	控制号	005598372
#002	记录最后处理	20120222164346.0
010\$d	价格	978-7-5399-4776-1: CNY39.90
100\$a	通用数据	20120208d2011 em y0chiy50 ea
101\$a	正文语种	chi
102\$a	出版国别	CN320000
105\$a	专著编码	y z 000ay
215\$a	页数册数	r
215\$a	页数册数	388页; 24cm
200	题名及责任者项	通天人物 [专著] / 李佩甫著
210\$a	出版地	南京: 江苏文艺出版社, 2011
300\$a	附注内容	读客知识小说文库 015
606\$a	普通主题	长篇小说 -- 中国 -- 当代
606\$a	普通主题	长篇小说
690\$a	分类号	I247.57
701\$a	人名	李佩甫 (1953~) 著
905\$c	排架号	I247.57 lpf
905\$c	排架号	I247.57 lpf
801\$a	国家代码	CN YNL 20120207
#012	未知国图字段:	A530000YNL UCS01005183416
#001	OWN	ZB301
#005	系统号	005598372

第一页 上一页 下一页 最后一页

©2012 Winguse

## 一、 查看历史记录

### 1. 布局介绍

历史记录左侧标签是一个个条目题名，点击该标签可以切换不同的书目：

Beginning EJB 3 application development : from novice to professional / Raghu R. Kodali and Jonathan Wetherbee ; with Peter Zadrozny.

Java程序设计教程 [专著] = Textbook for programming in Java / 雍俊海编著

嵌入式系统基础教程 [专著] / 俞建新, 王健, 宋健编著

每个书目信息头部，是这个书目信息的头部控制区，可以对这个条目的管理控制：

**Java程序设计教程 [专著] = Textbook for programming in Java / 雍俊海编著**

导出数据

数据来源：中国国家图书馆[书] / ISBN：9787302155799 / 获取时间： 2012-04-13 09:58:08 / 最后修改时间：2012-04-14 10:21:23 / UserID：1 [开]

记录状态：

不需要更新

信任否：

否

离线重新处理

编辑

删除

书目的主要属性内容是接下来的部分，包含记录号、字段名和对应的值：

记录号	字段名	值
#004	FMT	BK
#003	头标区	-----nam0-22----- 450-
001\$a	控制号	003551137
#002	记录最后处理	20070929163157.0
010\$d	价格	978-7-302-15579-9 : CNY49.00
905\$c	排架号	TP312JA yjh

## 2. 翻页

由于书目信息可能很多，为了方便，提供分页功能，在整个网页的最下端：

第一页

上一页

下一页

最后页

## 3. 搜索

为了方便查阅，用户可以在搜索框中输入信息进行那个检索，例如：



EJB

检索

重置

Beginning EJB 3 application development : from novice to professional / Raghu R. Kodali and Jonathan Wetherbee ; with Peter Zadrozny.

导出数据

数据来源：美国国会图书馆[书] / ISBN：9781590596715 / 获取时间： 2012-04-14 08:56:28 / 最后修改时间：2012-04-14 10:17:08 / UserID：1 [开]

记录状态：

不需要更新

信任否：

否

离线重新处理

编辑

删除

记录号	字段名	值
I006	国会图书馆记录控制号	2007271644
I005	书目信息固定链接	http://lccn.loc.gov/2007271644
105\$a	专著编码	Book (Print, Microform, Electronic, etc.)

## 二、 编辑条目

书目信息是可以再编辑的，点击控制区按钮：

编辑

即可进行相关操作，点击后，效果如图：

9



EJB

搜索

重置

Beginning EJB 3  
application development :  
from novice to professional  
/ Raghu R. Kodali and  
Jonathan Wetherbee ; with  
Peter Zdrozny.

EJB & JSP Java on the  
edge / Lou Marco.

Beginning EJB 3 application development : from novice to professional / Raghu

导出数据

数据来源: 美国国会图书馆[<#>] / ISBN: 9781590596715 / 获取时间: 2012-04-14 08:56:28 / 最后修改时间: 2012-04-14 10:17:08 /UserID: 1 [<#>]

记录状态: 不需要更新

信任否: 否

离线重新处理

编辑

删除

关闭编辑

记录号	字段名	值	
1006	国会图书馆记录控制号	2007271644	删除记录
1005	书目信息固定链接	http://lccn.loc.gov/2007271644	删除记录
105\$a	专著编码	Book (Print, Microform, Electronic, etc.)	删除记录
200\$f	作者	Kodali, Raghu R.	删除记录
200	题名及责任者项	Beginning EJB 3 application development : from novice to profes:	删除记录
210\$a	出版地	Berkeley, CA : Apress , c2006.	删除记录
701\$a	人名	Wetherbee, Jonathan.	删除记录
701\$a	人名	Zdrozny, Peter.	删除记录
215\$a	页数册数	xxviii, 481 p. : ill. ; 24 cm.	删除记录
010\$a	ISBN	1590596714	删除记录
010\$a	ISBN	9781590596715	删除记录
010\$a	ISBN	9781590596715	删除记录
300\$a	附注内容	Includes index.	删除记录
610\$a	非控主题	JavaBeans.	删除记录
610\$a	非控主题	Java (Computer program language)	删除记录
680\$a	分类号	QA76.73.J38 K617 2006	删除记录
676\$a	DDC	005.13/3	删除记录
1001	其它系统编号	(OCoLC)ocm73514574	删除记录
1004	质量代码	lccopycat	删除记录
1002	电子文件信息	Table of contents only http://www.loc.gov/catdir/toc/fy0712/2007/	删除记录
1003	链接主题	Table of contents only	删除记录
1003	链接主题		删除记录

提交

增加一项

重置

关闭编辑

第一页

上一页

下一页

最后一页

用户可以对书名、字段信息进行修改，也可以标记这个书目信息是否被信任，或者是否需要再更新。所有的编辑都在本地完成，提交用 AJAX 实现，不用等待。在某些必要的时候，比如你修改了处理的参数，可以点击离线重写处理进行重新对目标网页分析。如果需要重新载入，则需要你把记录删除，回到首页重新获取。

### 三、 删除记录

点击删除按钮，可以删除一个条目：



删除会有确定提示，如图：



### 四、 导出数据

点击导出数据按钮，可以完成数据导出：



导出的数据以对应的数据格式下载。目前支持原始的 MARC、文本；以及本系统处理后的文本和 CSV（Excel）格式导出。

# 第四章 系统设置

本章主要介绍系统参数的设置，包括数据源 xPath 参数设置，字段名到记录号的对应以及记录号和字段名的对应。

## 一、 xPath 参数设置

XPath 即为 XML 路径语言（XML Path Language），它是一种用来确定 XML 文档中某部分位置的语言。XPath 基于 XML 的树状结构，提供在数据结构树中找寻节点的能力。本节简要介绍系统设置中，几个 XPath 参数的设置。

点击系统设置后，点击数据源分析参数，然后点击对应的两个数据源标签，加载后如下图所示：

图书自动编目

首页

历史记录

系统设置

关于

数据源分析参数

字段名 到 记录号

记录号 到 字段名

中国国家图书馆 XPath 等参数

interfaceUrl: http://opac.nlc.gov.cn/F/?func=find-m&find\_code=l

failXPath: string(//html/head/title)

failString: 中文及特藏文 - 多库检索

set\_numberXPath: string(//html/body//form/input[@id='set\_number']/a

sessionString: string(//html/head/meta[@http-equiv='REFRESH']/

detailUrl: ?func=full-set-set\_body&set\_entry=000001&forma

catalogueNamesXPath: //tr/td[1]

catalogueValuesXPath: //tr/td[2]

importURL: string(//div[@id='operate']/a[@title='保存/邮寄']/attr

importURLReplaceFrom: full-mail-0

importURLReplaceTo: full-mail

textParameter: &format=002

MARCPParameter: &format=997

fileDownloadURL: string(//html/body/p[@class='text3']/a[1]/attribute::

提交

重置

美国国会图书馆 XPath 等参数

## 1. 中国国家图书馆 XPath 参数设置

参数名	默认值	说明
<b>interfaceUrl</b>	http://opac.nlc.gov.cn/F/?func=find-m&find_code=ISB&request=	这个是国家图书馆网页查询的第一个请求 URL 地址，后面紧跟 ISBN 号码。
<b>failXPath</b>	string(//html/head/title)	初次查询返回后，界定返回结果的特征信息的 XPath，默认是截取标题。
<b>failString</b>	中文及特藏文 - 多库检索	上一个特征信息为此值时，表示查询失败。数据源无此记录。
<b>set_numberXPath</b>	string(//html/body//form/input[@id='set_number']/attribute::value)	国图查询信息后，对于一个特定的数目，用一个 setnumber 标记，此处提取该信息的 XPath。
<b>sessionString</b>	string(//html/head/meta[@http-equiv='REFRESH']/attribute::content)	对于每次 http 请求，需要一个 session ID，这里提前该值，如果没有，有可能数据不对。
<b>detailUrl</b>	?func=full-set-set_body&set_entry=000001&format=002&set_number=	对于一个书目信息，我们需要返回的字段记录的二元组，这个是对应的下个网页的信息。
<b>catalogueNamesXPath</b>	//tr/td[1]	对于字段记录二元组网页的字段名信息特征
<b>catalogueValuesXPath</b>	//tr/td[2]	对于字段记录二元组网页的字段名对应的值的信息特征
<b>importURL</b>	string(//div[@id='operate']/a[@title='保存/邮寄']/attribute::href)	导出数据的 URL 特征地址。
<b>importURLReplaceFrom</b>	full-mail-0	对于导出信息的修正，从这个体会为下一个
<b>importURLReplaceTo</b>	full-mail	上一个替换为这个
<b>textParameter</b>	&format=002	文本的导出参数
<b>MARCPParameter</b>	&format=997	MARC 的导出参数
<b>fileDownloadURL</b>	string(//html/body/p[@class='text3']/a[1]/attribute::href)	定位文件的下载 URL

## 2. 美国国会图书馆 XPath 参数设置

参数名	默认值	说明
homepage	http://catalog.loc.gov	美国国会图书馆首页地址，用于获取 Session。
interfaceUrl	http://catalog.loc.gov/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=First	查询入口网页，用于定位查询 URL。
queryUrlXPath	string(//html/body/form[1]/attribute::action)	定位查询的 URL。
PidXPath	string(//html/body/form[1]/table/tr/input[1]/attribute::value)	查询的 Session（Pid）的特征 XPath。
SeqXPath	string(//html/body/form[1]/table/tr/input[2]/attribute::value)	获得 SeqID 信息的 URL。
queryFailXPath	string(//html/head/title)	判别查询成败的特征位置。
queryFailString	Library of Congress Online Catalog	失败的特征值。
detailUrlXPath	string(//html/body/form/center[2]/a[2]/attribute::href)	详细信息链接的地址 XPath。
catalogueNamesXPath	//html/body/form/table[1]//tr/*[1]	字段名的 XPath。
catalogueValuesXPath	//html/body/form/table[1]//tr/*[2]	对应的值的 XPath。
RIDXPath	string(//html/body/form/input[@name='RID']/attribute::value)	RID 信息，用于导出数据。

## 二、 字段名到记录号参数设置

点击第二个标签，可以设置字段名和记录号的对应关系。这里主要是为了统一两个数据源之间的字段对应关系到标准的 MARC 数据中。例如：书目的“题名及责任者项”在国家图书馆可能是“题名”，而国会图书馆可能是“Main title:”，我们对应的 200，然后翻译为“题名及责任者项”。

图书自动编目

首页 历史记录 系统设置 关于

编目名到编目号码数据获取成功

数据源分析参数

字段名到记录号

记录号到字段名

字段名

记录号

增加

049	#012	删除
004	#010	删除
096	#009	删除
本地 LC 类型索	#008	删除
索取号	#007	删除
LC 索取号	#006	删除
系统号	#005	删除
OWN	#001	删除
记录最后处理	#002	删除
头标区	#003	删除
FMT	#004	删除

系统会检测修改，自动提交服务器。  
如果需要增加，可以填写最上面的增加表单，然后点增加。  
如果编目处理时，发现未定义的字段名，这里也会自动加入类似的记录，请注意修改：

删除

默认情况下，对应的编目号码以“未定义编目号[×××]\*\*\*”的格式出现。

### 三、记录号到字段名参数设置（MARC 信息）

此处我们已经导入了大部分的 MARC 数据<sup>[1][2]</sup>，如果需要个性化修改，同上小节所述进行修改，系统将自动检测并提交服务器。

图书自动编目 首页 历史记录 系统设置 关于

数据源分析参数

字段名 到 记录号

记录号 到 字段名

增加

#012	未知国图字段: 048	删除
#011	最近处理日期	删除
#010	未知国图字段: 004	删除
#009	未知国图字段: 096	删除
#008	本地 LC 类型索	删除
#007	索取号	删除
#006	LC索取号	删除
005\$a	未定义编目名[005\$a]***	删除
200	题名及责任者项	删除

不过有些字段是无法跟这的记录好对应上的，我们建议的处理方法是，用自定的记录好处理。例如我们目前找到的有国图的“最近处理日期”对应号码是“#011”，国会图书馆的“LC control no.:" 对应为“!006”，翻译对应为“国会图书馆记录控制号”。

我们建议，国家图书馆的前面以#打头，国会图书馆的以!打头。

对于那些未定义的号码，系统处理的时候，也会自动加入，例如：

删除

系统定义的格式为：未定义编目名[×××]\*\*\*，请注意修改。



## 第五章 参考文献

- [1] 段明莲.信息资源编目[M].北京：北京大学出版社，2008.
- [2] 刘小玲.CNMARC 书目数据编制方法及操作实例[M].北京：国家图书馆出版社，2008.