全国馆藏管理与服务平台改进

索晶

2020-11-25

“联合目录建设带来的变化是革命性的。”

全国联合目录中的馆藏建设

索晶

2014-6

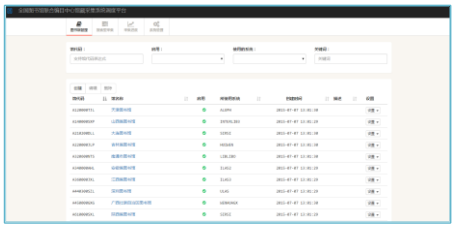
全国联合目录

- 2010年10月全国图书馆联合编目系统上线试运行，
- 2010年10月对分中心和主要的成员馆发布了馆藏收割的公函，呼吁建立全国联合目录
- 2011年联合编目中心对全国公益性图书馆实行书目数据免费政策，并把提交馆藏作为成员馆的对等义务写在了联编中心的协议书内；
- ……
- 截止6月底，联合编目中心揭示全国图书馆的馆藏量为**27376367。**

馆藏收割的难点：

- 目前国内的公共图书馆不到3000家。
- 截至2013年6月21日，全国普通高等学校（不含独立学院）共计2198所，其中独立设置民办普通高等学校424所；全国成人高等学校共计298所，其中民办成人高等学校1所。——百度百科
- 其他图书馆……
- **国内海量的图书馆和图书室**

2.3系统截图——图书馆管理



Knowlesys

2.3系统截图——采集进度

Knowlesys

实体资源馆藏工作的背景

- 待处理馆藏量巨大，存量数据超过1.2亿条书目数据，每年新增约1000万条书目数据；
- 数据质量参差不齐，数据来源复杂。

- 基础查重算法：十四字段查重；010$a、
- 改进方案：结合八字段查重，对不同的查重结果设置区间，减少查重者需要消耗的精力

- 查重数据量巨大导致传统查重方法无法达到实际效果；
- 查重工作人员的经验复用问题；
- 查重工作流程标准化问题。

- 查重工作界面

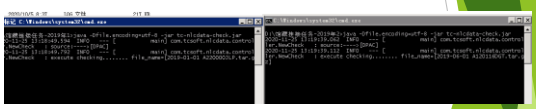
新的查重算法

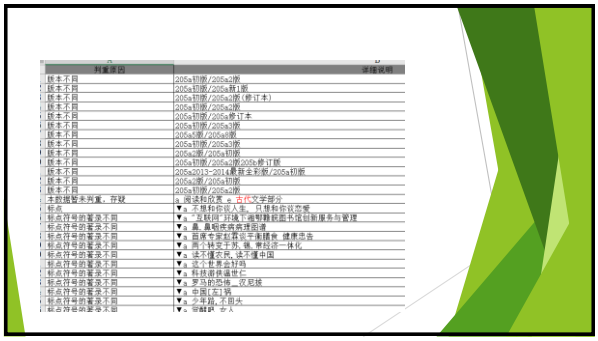
- 预处理：
- ISBN号：取自010$a，9位和13位互检，有无小短横互检。
- 正题名：取自200$a，所有非文字内容都忽略，比如标点符号、特殊符号、图形、图标等，不包括阿拉伯数字和英文字母。简繁体互检。
- 责任者：只取第一责任者，也就是200$f。责任者前代表朝代、国别的括号（）和括号里的内容忽略，名字中的符号（比如中圆点、下圆点、小短横）忽略。
- 页码：取自215$a，只取数字，忽略标点符号（（），；-）和文字描述（册、页、）等。

查重算法

- 有ISBN号的，用判新程序1查重；没有ISBN号的数据，用判新程序2查重
- 程序1：
  - 有ISBN号的数据，用程序1判新，查重点为：ISBN、题名。
  - 如果内容全相同，为匹配数据。
  - 如果ISBN不同，视为新数据。
  - 如果ISBN相同，题名不同，再用补充程序A判新。
- 程序2：
  - 没有ISBN号的数据，用程序2判新，查重点为：题名、第一责任者、出版者。
  - 三个查重点全相同，才视为馆藏可聚合，只要有一个查重点匹配不上，就视为新数据。
  - 如果责任者或者出版者一方数据有，另一方缺失，视作此查重点相同。

全国实体馆藏管理与服务平台（联合目录）查重



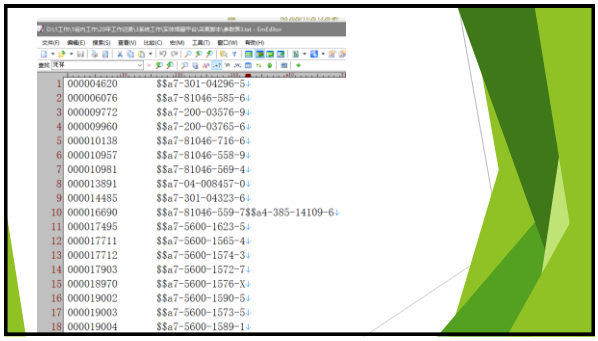


判重原因	详细说明
版本不同	20Sa初版/20Sa2版
版本不同	20Sa初版/20Sa修1版
版本不同	20Sa初版/20Sa2版(修订本)
版本不同	20Sa初版/20Sa2版
版本不同	20Sa初版/20Sa修订本
版本不同	20Sa初版/20Sa3版
版本不同	20Sa2版/20Sa3版
版本不同	20Sa初版/20Sa3版
版本不同	20Sa2版/20Sa初版
版本不同	20Sa初版/20Sa2版/20Sb修订版
版本不同	20Sa2013-2014教基全彩版/20Sa初版
版本不同	20Sa2版/20Sa初版
版本不同	20Sa初版/20Sa2版
本数据暂未判重，存疑	a_阅读科效果_e_古代文学部分
标点	▼a_不畏和你说人生，只想和你说恋爱
标点符号前著录不同	▼a_互联网+环境下调整数据图书馆创新服务与管理
标点符号前著录不同	▼a_鼎_晨报老师教师版
标点符号前著录不同	▼a_首席专家赵晋识平衡膳食_健康患者
标点符号前著录不同	▼a_两小种变千势_强_策略统一体化
标点符号前著录不同	▼a_这个建交民_这个建中国
标点符号前著录不同	▼a_这个世界会好吗
标点符号前著录不同	▼a_科技前供骗世仁
标点符号前著录不同	▼a_罗马的恐惧_仅尼拔
标点符号前著录不同	▼a_中国(五)销
标点符号前著录不同	▼a_少年班_不回头
标点符号前著录不同	▼a_智能时代_中人

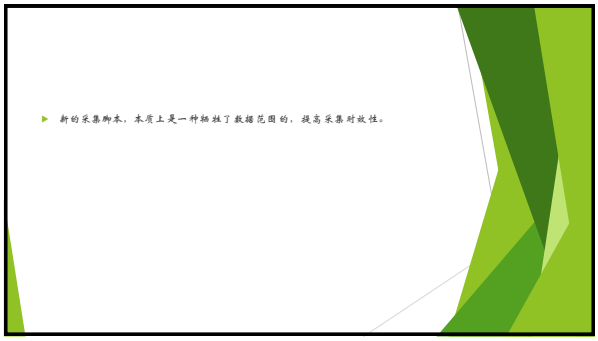
采集脚本

ENABLED	GLC	LIBNAME	SYS	
1	A100000GLC	国家图书馆	Aleph.500	http://opac.nlc.cn
1	A110000CLC	首都图书馆	Aleph.500	http://alephl.clcn.net.cn:8991
1	A120000TJL	天津图书馆	Aleph.500	http://opacwb.tjl.tj.cn:8991
0	A120104MAL	天津市南开区图书馆	ILAS.3	http://60.30.89.36:81
1	A120116DGT	天津市滨海新区大港图	Interlib.3	https://60.29.199.64:443
1	A120233JHZ	天津市静海县图书馆	Interlib.3	http://60.29.33.142
1	A130000HBS	河北省图书馆	ILAS.3	http://ilas.helib.net
1	A130100ZJZ	石家庄市图书馆	Interlib.3	http://123.182.227.70:8088
1	A130207GTL	唐山市图书馆	Interlib.3	http://27.156.195.3:7089
1	A130207FWL	唐山丰南区图书馆	Interlib.3	http://www.fnlib.net:8080
0	A130225LUT	乐亭县图书馆	Interlib.3	http://60.2.24.198:8082
1	A130283QAS	迁安市图书馆	Interlib.3	http://60.2.185.194:8523
1	A130300QHD	秦皇岛图书馆	Interlib.3	http://121.22.16.106:2100
0	A130481WTA	武安市图书馆	ILAS.3	http://221.193.241.57:8020
1	A130500TJS	邢台市图书馆	Interlib.3	http://60.6.198.252:2101
1	A130900CZL	沧州市图书馆	Interlib.3	http://tuchuang.czlib.com.cn
1	A131000LFL	廊坊市图书馆	Interlib.3	http://60.10.135.62:90
1	A131082SAB	三河市图书馆	Interlib.3	http://opac.shstag.com:2001
1	A140000SXF	山西省图书馆	Interlib.3	http://opac.lib.sx.cn
1	A150200ELC	包头市图书馆	Interlib.3	http://1.30.233.68
1	A150400CPG	赤峰市图书馆	Interlib.3	http://116.117.94.102:8081
1	A150500MWT	通辽市图书馆	Interlib.3	http://111.56.44.38:44
1	A150600ESL	鄂尔多斯市图书馆	Interlib.3	http://58.18.34.157

SYS_VERSION	REQ_METHOD	PARAM_STR	PARSE_DIREC	RECKEP
Aleph.500	GET	F?func=find-b&find_code=IBBrequest=$$word&local_base=NLC01	1	
Interlib.3	GET	opac/searchOpt$$word&searchKey=ibn	1	class="t
ILAS.3	GET	MTRxBookKotr.aspx?strType=isbn&strKeyValue=$$word1	1	class="t
MUIWEB	GET	opac/openlink.php?strSearchType=isbn&match_flag=full&HistoryCount=1		
ILASPAC	GET	ILASPAC/MTRxBook=&Mstr.do?searchType=1&mustSearchKey=$$word&	1	
ILAS	GET	Search/searchshow_jsp?y_tablearray=bibliison%2Cserbibm%2Capabibbm%1		
MEBBMAGX	GET	InQiLib/Opac$sr=SearchSolr|simpleSearch_actionSelect1=isbn&text1=1		
Wahulwen	GET	opac/opac/fulltext_result.action?meta=isbn&q=$$word1	1	



1	000004620 $$a7-301-04296-5
2	000006076 $$a7-81046-585-6
3	000009772 $$a7-200-03576-9
4	000009960 $$a7-200-03765-6
5	000010138 $$a7-81046-716-6
6	000010957 $$a7-81046-558-9
7	000010981 $$a7-81046-569-4
8	000013891 $$a7-04-008457-0
9	000014485 $$a7-301-04323-6
10	000016690 $$a7-81046-559-7$$a4-385-14109-6
11	000017495 $$a7-5600-1623-5
12	000017711 $$a7-5600-1565-4
13	000017712 $$a7-5600-1574-3
14	000017903 $$a7-5600-1572-7
15	000018970 $$a7-5600-1576-X
16	000019002 $$a7-5600-1590-5
17	000019003 $$a7-5600-1573-5
18	000019004 $$a7-5600-1589-1



- 新的采集脚本，本质上是一种缩小了数据范围的，提高采集时效性。