实体资源馆藏建设中海量数据的查重方案

报告人：秦晶

- 实体资源馆藏工作的背景
- 联编主要的查重方法
- 实体资源馆藏查重工作需要解决的问题
- 实际的方案

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，；
- 实体资源馆藏建设与联合编目建设的差异。

联编主要的查重方法

- 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

实体资源馆藏查重工作需要解决的问题

- 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。

查重工作的本质

- 查重工作是一个管理问题；
- 查重工作是一个工程问题。

工作总量的测算

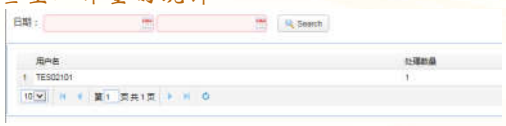
- 年均查重1500万条书目数据；
- 一个常规的工作人员每年完成10万条以上的人工查重工作。

工作流程的管理

- 查重工作的分配

- 查重工作界面

- 查重工作量的统计



工作逻辑的封装

项目代号		项目名称			
文档名称	国家图书馆馆藏揭示平台接口说明书				
产品版本		页数		密级	无

国家图书馆馆藏揭示平台
查重接口说明书

V 1.0.0

工作逻辑的封装

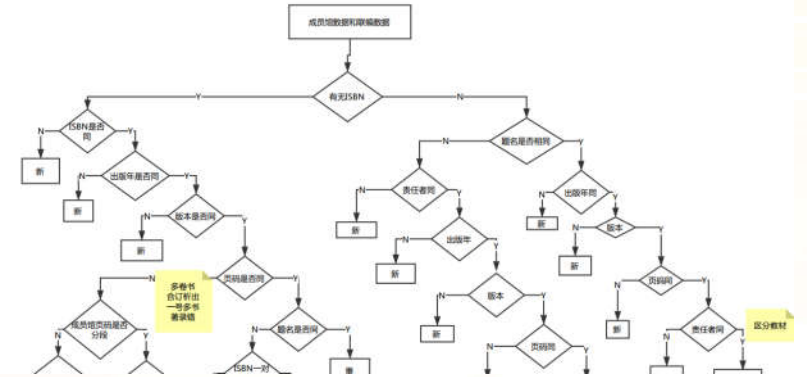
- 解析aleph数据接口
- 统计查重数量接口
- 查重结果集
- 数据判断处理：
  ① 新建书目
  ② 新建馆藏
  ③ 人工判断

工作接口的封装

- 网络访问的封装
- 日志的封装
- 数据解析的封装

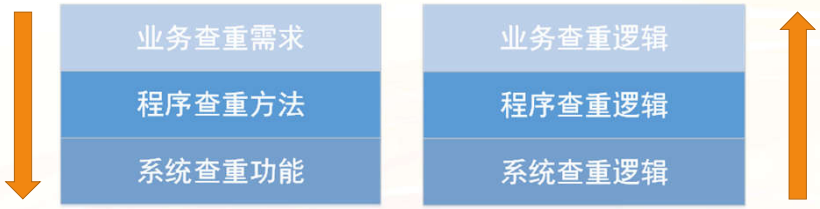
实际程序流程

- 详见PDF



```
if (isbnRetCount == 0) { //isbn无
    System.out.println("无isbn重复");
} else {
    //增加条件 出版年
    String year = null;
    if (s.getJSONArray("210d") != null) {
        year = s.getJSONArray("210d").getString(0);
    }
    int yearRetCount = 0;
    if (year != null) {
        System.out.println("继续查重出版年:" + year);
        filterQuery.put("210d", year);
        yearRetCount = Check.BoolCount(filterQuery);
    }
```

- 在设计时，通过不断的整合把业务需求抽象为系统功能；
- 在实现时，通过不断的封装把系统逻辑实现为业务方法



- *谢谢！！*