

档案数字化成果质检公共组件

(绍兴市越城区档案馆 2025 年 5 月 20 日)

一、概述

1. 组件简介

组件名称：档案数字化成果质检公共组件

建设单位：绍兴市越城区档案馆

开发上线时间：2024 年 6 月

2. 应用场景

本组件用于档案资源建设体系，依托 OCR 识别、算法模型、特征分析等技术，对档案条目和档案原文进行质量检查，规范性、准确性、一致性、可读性等质量问题进行批量检测，自动识别出著录项不规范、图像扫描质量、偏斜度、分辨率、空白页、污点、黑边等问题，并生成质检报告。不仅为进馆档案数字化加工质量提供有力保障，也便于各立档单位数字化加工项目标准化自检，确保高质量完成档案数字化任务。

3. 解决问题

近几年，各级档案馆不断加快纸质档案数字化进程，推动实现馆（室）藏传统载体档案数字化率高水平跃升，质检工作量大幅增加，人工逐页检查耗时耗力。目前档案质检往往采用按比例抽查方式，存在人工检测效率低、未抽查部分质量无法保障等问题。而长时间重复检查工作易引发视觉疲劳，漏检率随工作时长会显著上升。

通过建设档案数字化成果质检应用，完成数字化成果批量自动质检，极大提升检查效率、规范质量标准、降低人工依赖，保障档案数

字化成果的完整可用性。在应用的基础上将其组件化，实现质检功能快速复用。

二、当前运行使用情况

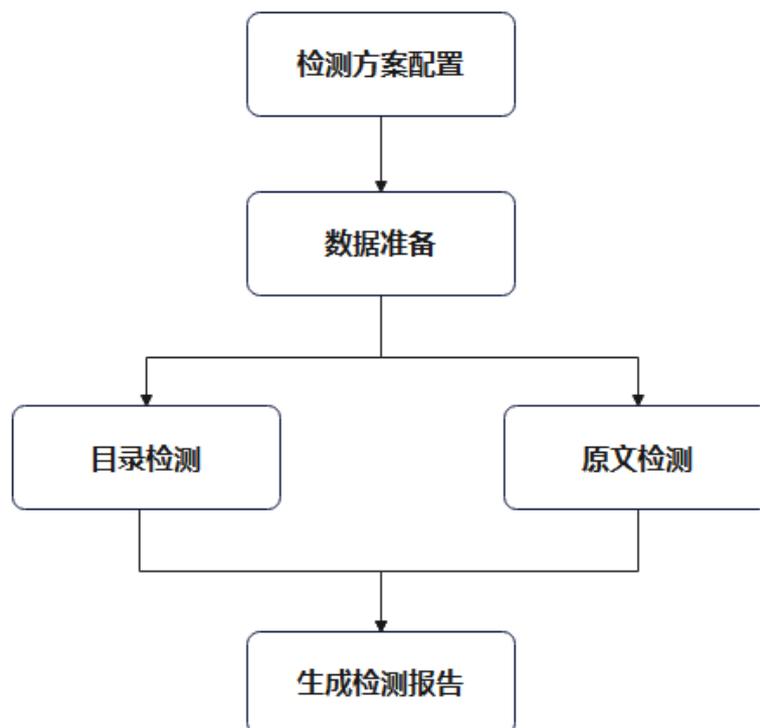
目前越城区档案馆对进馆单位 2023 年文书档案进行质检试运行测试，共计检查 63 家单位档案 16250 件 236604 页，系统检测结果与人工质检出的问题进行一一印证，结果准确率约为 90%。

三、核心功能

1. 功能清单

整个组件由检测方案配置、数据准备、目录检测、原文检测、生成检测报告五个部分组成。检测项配置作为档案数字化成果质检的前置条件，质检前要进行初始化设置。

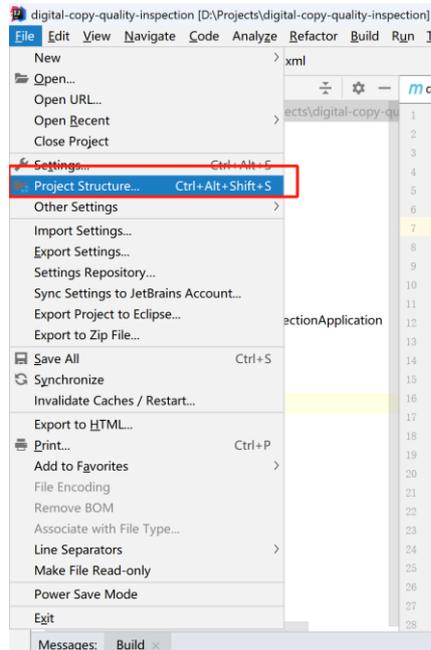
2. 业务流程图



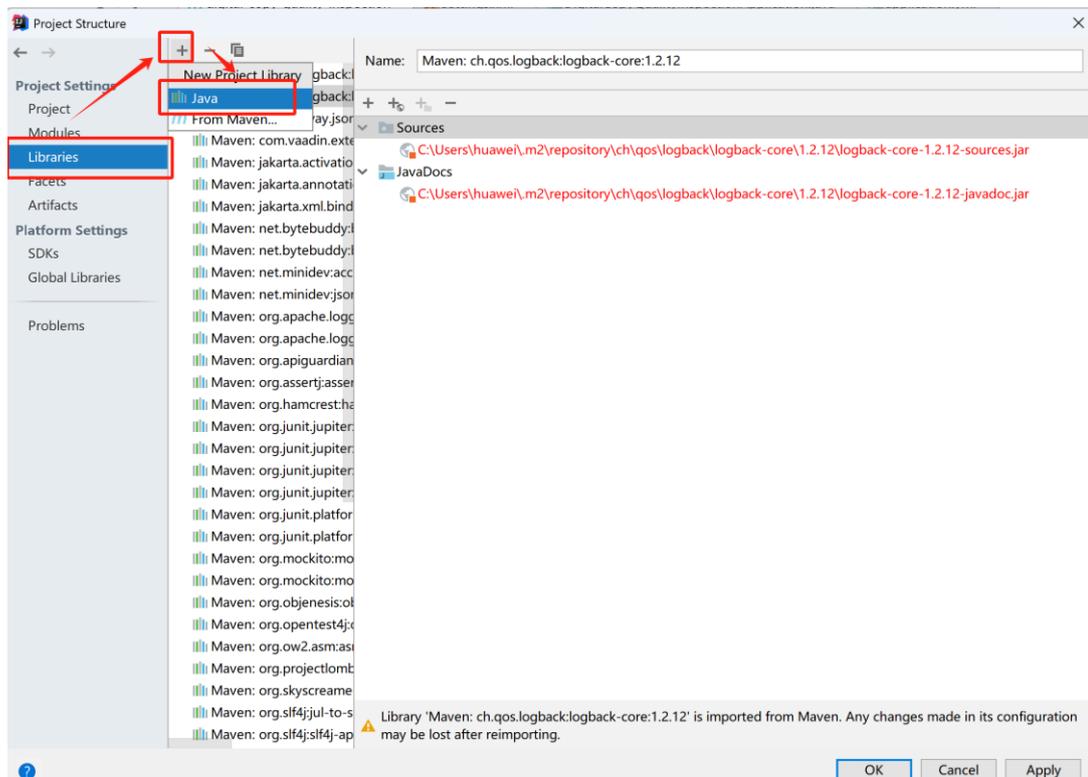
四、技术要点

1. SDK 引入说明

第一步：打开 File，点击 Project Structure 选项。



第二步：选择 Libraries，点击“+”按钮，选择 Java 选项。



第三步：选择 SDK -> OK

2. 接口参数及返回值说明

入参说明:

参数参数	参数类型	是否必传	参数说明
excelPath	String	true	条目 excel 文件服务器存放绝对路径, 用于获取待质检数字化副本的相关信息。 例如: /data/excel/J001.xlsx。该文件应包含数字化副本的关键标识、元数据等信息, 是质检流程的重要输入依据。
originalFolder	String	true	原文存放文件夹服务器存放绝对路径。例如: /data/original。SDK 将在此文件夹中查找对应的数字化副本文件进行质检。确保路径准确无误, 否则可能导致无法找到待质检文件
categeCode	String	true	待质检的档案门类代码, 用于区分不同类型的档案, 以便根据不同的规则进行质检。例如 WS, 不同的档案门类可能有不同的质检重点和标准。

出参说明:

参数名称		参数说明
reportPath	String	质检完成后生成的质检报告路径

3. SDK 调用示例

在 Java 代码中, 通过以下步骤调用 SDK 进行质检:

1. 初始化 SDK: 在使用 SDK 前, 需要先进行初始化操作, 加载必要的配置文件和资源

2. 构建质检请求: 根据入参说明, 构建包含必要参数的质检请求对象。示例代码如下:

```
String reportPath = QualityInspectionUtil.star(excelPath, originalFolder, categeCode);
```

3. 获取质检结果: 从质检响应对象中获取质检报告文件路径, 根据该路径读取质检报告, 查看质检结果

```
system.out.print(reportPath )
```

例, 输出:/data/excel/2024.4.4 质检报告-dkdtgyzhyke.xlsx

4. 质检项说明

序号	检测项	检测说明
1	图像倾斜度检测	判断数字化扫描图片的倾斜度是否大于设置的倾斜度阈值
2	题名一致检测	校验题名在首页是否一致
3	文号一致检测	校验文号在首页是否一致
4	责任者一致检测	校验责任者在首页是否一致
5	图像存储格式检测	图片格式是否非标准格式：JPG、PNG、TIF、TIFF、JPEG、JPE
6	图像扫描分辨率检测	图片分辨率 dpi 是否低于设定的阈值
7	清晰度检测	污点判断，阴影判断，订书机孔痕，黑框马赛克等
8	图像给空白页检测	图像或 PDF 单页是否为空白
9	保管期限检测	保管期限是否永久、30 年、10 年
10	图像重复页检测	PDF 内是否存在相同页
11	元数据必著录项检测	必填著录项内容是否为空检测
12	文件格式安全检测	文件存储的格式是配置的危险格式
13	TIF 压缩格式检测	TIF 压缩格式是否与设定一致（LZW, JPG, 无压缩）
14	文件页数一致性检测	著录项“页数”值是否和实际文件页数相同
15	档案挂接匹配度检测	是否挂接原文检测
16	涉密档案挂接检测	“密级”元数据值是否包含秘密、机密、绝密
17	目录重复进行检测	配置的元数据项（如：档号）是否重复
18	著录内容规范性检测	著录项配置的格式、长度 是否符合实际值
19	文件病毒检测	文件是否包含病毒检测

五、使用指南

1. 获取方式

在 IRS 平台的指定下载区域，找到对应“数字化副本质检 SDK”的下载链接。下载完成后，解压压缩包，可获取 SDK 的相关文件，包括 JAR 包、配置文件等。

2. 操作截图

(1) 检测方案配置

操作	序号	名称	编号	版本号	是否启用	变量值	描述
编辑 删除	2	文号一致检测	2	-	是	-	校验文号在首页是否一致
编辑 删除	3	责任者一致检测	3	-	是	-	校验责任者在首页是否一致
编辑 删除	4	缺页漏页检测	4	-	是	-	PDF实际页数和条目数是否相等
编辑 删除	5	倾斜度检测	5	-	是	-	图片倾斜度是否在配置范围内
编辑 删除	6	分辨率检测	6	-	否	2	"图片分辨率dp>是否大于设置数值"
编辑 删除	7	空白页检测	7	-	是	-	PDF空白页检测
编辑 删除	8	重复页检测	8	-	是	-	PDF重复页检测
编辑 删除	9	清晰度检测	9	-	是	-	污点判断, 阴影判断, 订书机孔痕, 黑框...
编辑 删除	10	目录内容检测	10	-	是	-	元数据长度、格式、必填项等校验
编辑 删除	11	保管期限检测	11	-	是	-	保管期限是否永久、30年、10年、100年。
编辑 删除	12	档号题名重复检测	12	-	是	-	档号+题名两者不允许同时重复
编辑 删除	13	原文格式检测	13	-	是	TIF、PDF	原文格式是否TIF、PDF、JPG等格式
编辑 删除	14	TIF压缩格式检测	14	-	是	JPG	TIF压缩格式是否与设定一致(LZW、JPG...
编辑 删除	15	目录文件格式	1	v1	是	XLS XLSX DBF	检测数字化成果目录文件格式, 有效格式...
编辑 删除	16	目录数据重复性	2	v1	是	-	检测数字化成果目录文件是否存在重复的...
编辑 删除	17	文件页数一致性	3	v1	是	-	检测数字化成果目录文件中, "页数"字段...
编辑 删除	18	文件命名及存储规则	4	v1	否	-	检测数字化成果文件的命名是否规范, 应...
编辑 删除	19	文件格式安全性	5	v1	是	-	对照"文件格式管理"模块的定义, 检查上...
编辑 删除	20	必著项	6	v1	是	-	检测数字化成果目录文件中是否包含不...

(2) 数据准备——目录原文上传

目录原文上传

点中或将文件拖拽到这里上传

支持文件格式: xls, xlsx

选择工作表: 数据表 (已匹配), 卷内级 (未匹配), 文件级 (未匹配)

数据校验: 重复校验, 合规性校验

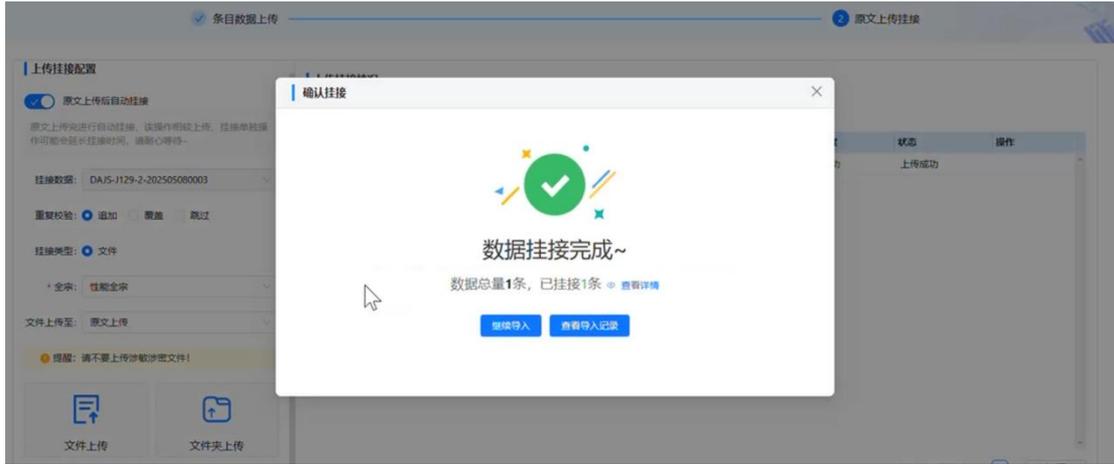
导入位置及元数据匹配

导入系统位置: * 全宗: 性能全宗, * 档号门类: 文书档案, * 著录期限: 档内级

excel元数据 (全部 25 | 已匹配 21 | 未匹配 4)

excel元数据	系统元数据
<input checked="" type="checkbox"/> 全宗号	全宗号
<input checked="" type="checkbox"/> 档号	档号11
<input checked="" type="checkbox"/> 案卷档号	案卷档号
<input checked="" type="checkbox"/> 题名	题名
<input checked="" type="checkbox"/> 全宗名称	全宗名称
<input checked="" type="checkbox"/> 年度	年度
<input checked="" type="checkbox"/> 开放预审结果	开放预审结果
<input checked="" type="checkbox"/> 开放标识	开放标识
<input checked="" type="checkbox"/> 保管期限	保管期限

继续导入 | 确认导入



(3) 点击“去质检”



(4) 选择检测项，开始质检

质检详情

编号	名称	参数配置	是否启用
1	目录文件格式	XLS XLSX DBF	<input checked="" type="radio"/> 是 <input type="radio"/> 否
2	目录数据重复性	-	<input checked="" type="radio"/> 是 <input type="radio"/> 否
3	文件页数一致性	-	<input checked="" type="radio"/> 是 <input type="radio"/> 否
4	文件命名及存储规则	-	<input type="radio"/> 是 <input checked="" type="radio"/> 否
5	文件格式安全性	-	<input checked="" type="radio"/> 是 <input type="radio"/> 否
6	必选项	-	<input checked="" type="radio"/> 是 <input type="radio"/> 否
7	著录内容规范性	-	<input checked="" type="radio"/> 是 <input type="radio"/> 否
8	图像扫描分辨率	-	<input checked="" type="radio"/> 是 <input type="radio"/> 否
9	图像存储格式	-	<input checked="" type="radio"/> 是 <input type="radio"/> 否

(5) 生成检测报告

质检记录

DAJS-J129-2-202505080003

开始时间: 2025-05-16 13:37:10
完成时间: 2025-05-16 13:38:05
检测耗时: 55秒

开始时间: 2025-05-16 13:21:04
完成时间: 2025-05-16 13:22:12
检测耗时: 38秒

质检报告

检测批次: DAJS-J129-2-202505080003

检测结果: 未通过

下载报告

基本信息

检测类型: 数字化成果质检 抽检目录数(条): 1 已挂接文件总大小: 297.18KB
检测开始时间: 2025-05-16 13:37:10 检测结束时间: 2025-05-16 13:38:05 检测耗时: 55秒

检测结果

检测项	检测结果
目录文件格式	✘
目录数据重复性	✔
文件页数一致性	✔
文件格式安全性	✔
必选项	✔
题名一致检测	✔
文号一致检测	✔
责任者一致检测	✔

元数据检测	著作内容规范性	×
	保管期限检测	✓
数字图像检测	图像扫描分辨率	✓
	图像存储格式	×
	图像倾斜度	×
	图像空白页	✓
	图像重复页	✓
	TIF压缩格式检测	✓
	清晰度检测	✓

不合格档案																				
序号	档号	目录数据检测					元数据检测		数字图像检测				数据挂接检测							
		目录文件格式	目录数据重复性	文件页码一致性	文件格式安全性	必著项	题名一致性	文号一致性	责任者一致性	著录内容规范性	保管期限检测	图像扫描分辨率	图像存储格式	图像倾斜度	图像空白页	图像重复页	TIF压缩格式检测	清晰度检测	档案挂接匹配度	涉密档案挂接
1	J219-WS-2025-0508-001	×							×			×	×							

六、特点和效益

1. 组件特点

(1) 检测标准定制化。依据《纸质档案数字化规范》要求，规范设置检测内容。可按照档案馆进馆标准，个性化定制检测方案，对通用检测规则进行开启或关闭，自定义配置检测规则，确保数据符合当地归档或进馆要求。

(2) 检测内容覆盖全面。检测项包括自动校验图像清晰度、文件格式、图像倾斜度、元数据完整性等核心指标。同时对档案原文进行文字识别，通过元数据一致性校验，降低档案目录错误率。

(3) 加固数据安全。对档案进行格式检测，避免数据不可读问题。在组件调用时，根据需求进行私有化部署集成，有效保障档案数据安全。

2. 实践效益

档案质检工作模式从“人工密集”到“自动批量”，有效降低了人工成本，释放档案人力资源。档案验收环节从“集中质检”转为“实时质检”，缩减了档案进馆周期。通过构建应用检测与人工复核的双重校验机制，确保档案质量达到验收标准，推进档案利用效能。