

档案软件 - OCR 文字识别功能【云丰档案软件功能介绍】

档案管理软件 · 功能介绍

纸质档案扫描成图片后只是看得见，OCR 让它变得可检索、可复制。OCR 文字识别自动识别扫描件图片中的文字，转为可编辑、可检索的文本，让纸质档案数字化后真正“活”起来，是检索与利用的基础。

一、为什么需要 OCR 文字识别

把纸质档案扫成图片容易，但图片里的字不能被检索、不能被复制，只能一页页翻，数字化价值大打折扣。

图片不可搜：扫描件是死图片，输入关键词根本找不到对应页；

人工录入慢：要引用其中文字，只能照着敲，慢且易错；

利用低效：查一卷档案仍要逐页翻，与翻纸本差别不大；

著录靠手敲：题名、日期要人工对着扫描件录入，工作量大。

OCR 识别把图片里的文字读出来：可搜、可复制、可自动著录，纸质档案数字化才真正产生利用价值。

二、怎么用：上传即识别

扫描件上传后，系统自动运行 OCR，识别其中文字并与图像叠加，生成可检索文本层。

表1 OCR 文字识别流程

步骤	操作要点
上传扫描件	纸质档案扫描后上传系统；
自动识别	OCR 引擎识别图片中的文字；
生成文本层	文字与原图像叠加，形成双层 PDF；
校对修正	对识别不准处人工校对；
入库检索	识别文本参与全文检索与著录。

识别要准，更要可校。OCR 对印刷体识别率高，但对手写、模糊、印章可能出错。系统把识别结果置于可校对状态，错处人工修正后再入库，保证检索质量。

识别结果置于可校对状态，错处人工修正后再入库。印刷体公文识别率高，手写与印章处需重点复核。

OCR 支持常见公文版式，自动识别红头、表格与栏线结构，不误把表格框线当文字。识别结果与原图像层叠保存，校对时图文对照。

OCR 是纸质档案数字化释放价值的钥匙。系统支持常见公文版式，自动识别红头、表格与栏线结构，不把框线误当文字，识别结果置于可校对状态，错处人工修正后再入库。更实用的是，OCR 可框选著录、自动填充题名与日期等字段，著录人员只需复核，录入负担大减。识别文本与原始图像层叠保存，校对时图文对照。经过 OCR，扫描件从只能翻看变为可全文检索，几十年前的老档案也能按关键词一秒定位，档案利用的深度由此发生质变。

三、核心能力

表2 核心能力一览

能力	一句话价值	关键细节
印刷体识别	图片变文字	自动识别扫描件中的印刷文字，转为可编辑文本。
文字叠加图像	双层可检索	识别文字与原图叠加，形成图文对照的双层 PDF。
自动著录	录字省人力	识别结果可自动填充题名、日期等著录字段。
支持校对	识别可修正	错字漏字可人工校对，保证文本质量。
中文优化	中文档案好用	针对公文、表格等常见档案版式优化识别。

OCR 框选著录能自动填充题名、日期等字段，著录人员只需复核，大幅减轻录入负担。

四、典型使用场景

存量数字化：批量老档案扫描后 OCR，立即可全文检索。

快速著录：新扫描件识别后自动带出题名著录，人工复核即可。

内容引用：需要引用档案文字时，直接复制识别文本。

五、价值与产品定位

OCR 文字识别的价值在于让扫描件从“只能看”变成“可检索”：文字可读、可搜、可复制，纸质档案数字化的利用价值由此释放。让扫描件从只能看变成可检索，纸质档案数字化价值由此释放。

它是云丰档案数字化加工的核心引擎，与双层 PDF、全文检索、智能著录衔接，是存量档案电子化利用的基石。