

档案软件 - PDF 格式原生导入功能【云丰档案软件功能介绍】

档案管理软件 · 功能介绍

PDF 已是电子文件交换的事实标准，红头文件、报表、证照大量以 PDF 流转。PDF 格式原生导入在不重新排版、不破坏原版式的前提下直接接收 PDF 文件，保留原始版面与印章，并自动判断其为文字版还是扫描版，分别走可检索或 OCR 流程。

一、为什么需要PDF 格式原生导入

很多归档文件本身就是 PDF，但系统对待 PDF 的方式往往决定了它能否被真正利用：

版式被重排：为了入库把 PDF 转成 Word 再转存，字体错位、印章位置偏移，文件“变了样”。

扫描 PDF 不可检索：由图片拼成的 PDF 里文字选不中、搜不到，只能一张张翻着找，形同哑档案。

版本混乱：同一份文件几经转换，系统里留存的是哪个版本难以确认，引用时存在风险。

附件与表单丢失：复杂 PDF 里的表单域、内嵌附件、书签在导入后丢失，信息不完整。

原生导入的核心是“原样接收、智能增值”：版式不动、版本可溯，再按 PDF 的真实内容类型自动决定是否需要 OCR。

二、怎么用

把 PDF 文件直接拖入归档入口，系统先做格式与内容类型检测，再分别处理。主要环节如下：

表1 PDF 格式原生导入流程

环节	系统处理说明
上传接收	单份或批量拖入 PDF，保留原始文件不重新排版
类型判断	自动识别是文字版 PDF 还是扫描影像版 PDF
文字版处理	直接抽取文字层建立索引，原版式与印章完整保留
扫描版处理	自动 OCR 生成双层可检索 PDF，原件一并保存

入库归档	著录字段自动填充，病毒与完整性校验后归入目标案卷
------	--------------------------

导入全程不改写 PDF 内部结构，内嵌字体、签章、书签原样保留。文字版直接可检索，扫描版自动补齐文字层，两类文件入库后都能被全文检索命中。对于同时含有文字层和扫描页的混合型 PDF，系统按页分别处理：有文字的页直接索引，纯图像页触发 OCR，整份文件最终都能被全文检索命中，不必人工区分哪几页需要处理。

三、核心能力

表2 核心能力一览

能力	一句话价值	关键细节
原版接收	文件不变样	不重排不转存，字体、印章、表格版面与原文件一致，保证凭证效力。
类型识别	智能分流	自动判断文字版与扫描版，前者直接索引，后者触发 OCR，处理不重复。
文字层建立	扫描件也能搜	对影像 PDF 自动 OCR 生成双层 PDF，文字可选可搜，告别哑档案。
结构保留	附件书签不丢	内嵌附件、表单域、书签目录在导入后完整保留，信息不残缺。
版本留痕	入库可追溯	原始文件留存并记录导入时间与操作人，版本变更全程留痕可查。
混合处理	整份可检索	同一份 PDF 内文字页与扫描页按页分别处理，最终全文均可检索。

四、典型使用场景

- 红头文件归档：机关下发的红头 PDF 文件原版入库，印章与版式不变，检索时可直接定位到具体条款。
- 报表与证照存档：年度报表、营业执照等 PDF 批量接收，扫描件自动 OCR 后即可按关键字检索。
- 外部来文管理：往来单位以 PDF 报送的材料原样归档，版本清晰，调阅时与收到的文件完全一致。

五、价值与产品定位

PDF 原生导入让占比最高的一类电子文件“进得来、留得住、搜得到”：版式不动保证凭证严肃，自动类型判断让扫描件也具备检索能力，避免了反复转

换带来的失真与混乱。无论文件来自哪个业务系统、以何种 PDF 子格式流转，云丰档案都能原样接收、智能增值，让占比最高的电子文件类型在档案库里既保真又好用。

它是云丰档案多格式采集与 OCR 能力在 PDF 场景下的落点，与全文检索、密级管控、在线预览贯通。多全宗、三员分离与信创适配的整体架构，使大量 PDF 文件在统一平台内安全汇集并被高效利用。