

**数科文件转换迁移系统**

**V3.0**

**技术白皮书**

**北京数科网维技术有限责任公司**

**2020年4月**

# 产品概述

数科文件转换迁移系统部署在服务器端，提供长期驻留的格式转换和版式加工服务，系统内置多种转换引擎，支持集群配置，通过并行计算、自动负载等技术，可有效应对高精度、大数据量、高速度、高可靠性要求的文档转换需求，提供实时转换、批量转换和套版转换功能，支持元数据处理、数字签名处理、批量盖章、二维码处理等版式文档二次功能。通过JAVA API、C++、COM、WebService、HTTP和虚拟打印等技术形态的功能服务调用接口实现与电子公文、电子档案、电子证照、电子凭据等业务系统的无缝集成。

文件转换迁移系统具有转换速度快、转换精度高、服务运行稳定、海量转换支持、可扩展性强等显著优势，适合于部署在政府机关及企事业单位内部提供版式转换、加工服务。



**图 文件转换迁移系统示意图**

# 技术路线

数科文件转换迁移系统采用国家或软件行业已有标准，强调技术的实用性、安全性、可靠性、先进性，从标准和技术上保障系统的可扩充性、易维护性、开放性和统一性。

**1）基于跨平台技术构建**

系统采用跨平台技术开发，实现对所有的硬件和操作系统平台的支持，特别是国产软硬件环境的支持，使用户在操作系统和硬件的选择上有更大的自由度，具有跨平台性，支持多样的部署方式，满足一次编写、到处运行等技术要求。

**2）采用基于开放标准的技术路线**

系统在设计时充分考虑了“标准和开放”原则，支持各种相应的软硬件接口，使之具有灵活性和扩展性，具备与多种系统互连互通的特性，在结构上实现真正开放，具备良好的可移植性、互操作性和集成性。

**3）采用面向服务架构（SOA）架构**

系统的设计和实现方案完全采用SOA面向服务架构，应用功能模块化，高内聚、低耦合，系统构架层次化，降低信息数据层、业务应用层、用户交互层资源间的耦合度，实现基于标准、松散耦合、共享服务、粗粒度和联合控制。

**4）国家标准支持**

系统全面符合国家自主版式OFD标准（GB/T 33190-2016），同时还符合电子公文标准规范体系中的电子公文结构（GB/T 33476.1-2016）、显现（GB/T 33476.2-2016）、实施指南（GB/T 33476.3-2016）中关于版式部分的要求；符合电子公文元数据（GB/T 33480-2016）和电子公文标识（GB/T 33477-2016）的规范要求；符合应用接口（GB/T 33478-2016）中关于版式插件和印章接入规范（GB/T 33481-2016）的要求。

# 产品功能

## 多格式转换

支持将Office及其高版本的办公文档、WPS系列的文档、PDF、XPS、永中Office系列（如UOF）、BMP、JPG、PNG、TIFF、HTM/HTML/SHTML网页、TXT、RTF、PS、CEB、SEP、CAD等格式文档转换为OFD文件；同时支持将OFD文件转换为PDF文件；支持将OFD文件转换为图片文件，且图片的宽度和分辨率可自定义；支持流式文件带修订痕迹的格式转换；可以通过URL地址实现远程源文件的格式转换。图片转换为OFD文件时，可设置页面的宽度、高度以及图片分辨率。

可根据需要提供定制化的格式转换服务。

## 高速转换

系统支持集群部署，内置多线程、多点并行计算、自动负载均衡等机制，非常容易扩充，其中的工作节点数量可以进行“热插拔式”扩展（即在不影响服务对外调用的前提下增加或减少转换节点）。通过合理配置转换工作节点的数量，可以应对TB级存量文档的高速转换需求。

## 多引擎转换

系统内置虚拟打印和其他转换引擎，当进行文档转换时，优先使用该文档的创建软件作为排版引擎，很大程度上避免了不同版本引擎之间的跑版现象引起的转换失真，转换后的版面效果更为精确。另外，电子文件使用其创建软件解析和排版的成功率也显著提高，这使得转换节点工作更为稳定。

通过转换引擎，支持流式文件中的文本、表格、自选图形、图像、OLE、艺术字、书签、链接等各种文档元素的转换，书签、链接转换后保持其活性；将文档结构图直接转为OFD大纲；支持公文标识、公文元数据和公文语义；支持音视频附件和对应的链接转换；支持将嵌入OLE对象转换为OFD的附件和相关链接，转换后的版面效果精确完整。通过接口，可将公文体数据导出供其他应用使用。

## 批量转换

可采用后处理方式实现大批量文档的异步转换，主要用于历史文件批量转换为国标OFD格式和多业务系统的批量转换需求。

## 套版转换

支持无格式XML文件和结构化数据的套版转换，应用系统将数据组织为XML文件，指定引用模板文件，转换服务依据模板定义将数据内容转换为特定显示格式和版面的OFD文件，满足应用系统中各种套打功能和报表的版式化需求。生成OFD文件时可同时生成缩略图。

## 元数据处理

可在生成OFD文件的同时添加符合《党政机关电子公文元数据规范》（GB/T 33480-2016）的元数据，元数据包括公文标识、文种、份号、密级和保密期限、紧急程度、发文机关标志、发文字号、签发人、标题、主送机关、附件说明、发文机关署名、成文日期、附注、抄送机关、印发机关、印发日期、发布层次。

同时支持用户自定义元数据，提供元数据的添加、修改、删除功能。通过接口，可提取文件中的元数据信息，以便实现元数据的深度利用。

## 附件处理

可将指定附件添加到OFD文档中，可以指定附件标题、附件的类型等信息。可以从OFD文档中移除附件。

## 文档权限处理

可在OFD文档中增加、删除文档权限，文档权限包括：打印、导出（保存与另存）、复制、下载、签章、注释等，一旦定义了文档处理权限，则当打开OFD文件时将只有允许的操作权限。

## 文件标识处理

系统支持添加文件标识；系统支持密级标识的添加和删除，系统支持在文件转换时或在转换核验之后，在版式文件中加入密级标识符号并将其与文档绑定。支持依据密级自动形成对应的文档操作限制，添加到文档权限声明中；支持依据密级信息生成密级标识图片作为注释的可见部分；支持针对元数据和全文内容计算消息摘要，用给定密钥对中的私钥进行签名，并嵌入验证用的数字证书；支持将签名值与密级标识图片嵌入绑定。

## 文档页面处理

可为版式文档添加封面、尾页、在指定位置插入页面或删除指定的页面，支持指定页面的旋转及自定义旋转角度，以满足不同页面不同内容的显示需求。同时，系统支持OFD版式文档页面位置交换功能，可指定交换页码。

## 二维码处理

系统可以在OFD文档的指定位置嵌入二维码，支持宽高、DPI等参数的自定义。

## 嵌入字体

系统可以在OFD文档中嵌入字体文件，从而保证各种环境下的呈现效果一致。

## 文档组合拆分

系统支持将多个文档按要求组合成一个OFD文件，支持将一个OFD文件按要求拆分为多个OFD文件或图片文件。

可组合的文档格式包括Office文件、WPS文件、PDF、XPS、图片、网页等，对图片文件可指定组合后的宽度、分辨率。

## 文本内容提取

支持检索服务，可快速提取OFD文件的文本内容，并提交搜索引擎以生成文件内容索引库，从而实现基于文本的版式文件内容检索。

## 水印/背景处理

系统支持在转化后文档的指定页面中插入水印和背景，水印和背景支持图形、图像和文字的组合描述。

添加的图片水印可指定位置、是否可以打印、是否视图显示等属性，文字水印可指定位置、文本内容、字体、字体颜色、旋转角度等属性。

## 数字签名

系统可在OFD文件中添加符合OFD格式规范和密码技术应用规范要求的数字签名。

系统提供两种方式实现数字签名的添加：接口方式（须提供签名信息和关联类名）和证书方式（须提供封装的CA证书）。系统同时提供数字签名的删除。

系统提供数字签名的验证接口供应用系统调用。

## 批量盖章

配合电子签章服务，可实现后台批量盖章，可指定盖章的位置和大小。支持电子印章的删除（可保留印章图元信息）；提供验章接口，可通过接口在系统中进行电子印章的验证。

对于已有印章的历史文件，在转换时可删除数字签名而保留印章图元，然后再新加数字签名。或者，将旧的印章图元和数字签名全部删除，再加盖新的电子印章。

## 加密封装

系统支持将转换后的版式文件封装为指定类型的加密信封，该信封可被具备条件的阅读终端打开，而未授权的用户和软件无法打开。系统支持使用用户指定的加解密算法进行文件的加解密，支持在OFD文档中添加阅读有效期控制信息。

对于保密要求比较高的电子文件，一般在格式转换时或标识、盖章后，通过调用加密封装服务或接口包装为加密信封文件，包装后的加密文件需要持有对应的密钥才可以阅读。

加密封装可支持不同的加密强度和授权阅读方式，可支持选择使用以下方式：

（1）阅读时需要提供密码才可阅读。封装时依据此密码对数据进行对称加密。

（2）指定对应的设备可以阅读。封装时以设备的对应信息作为加密密钥。

（3）指定持有特定私钥或Key的人可以阅读。使用256位或指定长度的随机密码对文件数据进行对称加密，再用对应的公钥对密码进行非对称加密。

## 后台管理

系统提供完善的后台管理功能，具体如下：

1. **监控管理**

系统采用图形化方式显示各转换节点的转换状态，包括成功数、失败数等信息：



**图 监控管理界面**

1. **在线预览**

系统提供在线预览，可直接上传本地文件进行转换，转换成功后将OFD文件下载到指定位置。通过此功能可快速验证系统部署是否成功，以及原始文档是否能够正确转换。



**图 在线预览界面图**

1. **转换日志**

系统以列表方式显示各业务系统的转换情况，可查看转换明细，系统提供失败任务重置功能。转换日志包括任务标识、来源系统、主机信息、任务原始名称、转换时间、转换状态、任务详情等信息。系统提供日志查询、导出等功能。



**图 转换日志列表界面**

1. **多系统注册**

支持多业务的转换请求调用，各业务系统在注册后，系统将采用按优先级排队转换的方式处理各业务系统的转换请求，并分别将转换结果返回各系统。



**图 多系统注册界面图**

1. **转换统计**

格式转换统计功能是展示转换成功率数据，通过自定义日期选择天、周、月、年维度的数据信息，可选择不同数据系统查看转换分析数据。统计样式支持柱形图和饼状图，也可以将统计结果导出为Excel文档。



**图 格式转换统计分析界面**

# 产品技术特色

* 转换速度快

通过并行计算和自动负载均衡设计，可以获得数倍甚至数十倍于传统文档转换方式的吞吐能力。

* 转换精度高

内置多种排版引擎，综合利用各个引擎的优点，有效规避其转换缺点，转出文档与源文件相比版面精确度更高。

* 服务运行稳定

转换服务的结构设计特点可实现比一般文件转换方式更为稳定，这依托两种因素：首先，转换服务不会因某一节点的失败而失败，系统提供持续服务的能力更强；其次，用文档对应的创建工具去排版和转换该文档将显著降低失败概率。

* 可转无格式文档

独创的“XML文件+模板”的方式支持无格式文件的转换，该模式可用于将数据库中的单条记录创建为可见的版式文件。

* 海量转换支持

系统支持集群部署，支持转换节点“热插拔”，可实现海量历史文件的转换。

* 可扩展性强

转换服务中的密码运算模块（对称和非对称加解密）、源文件获取模块和转换结果处理等均已抽象为可二次开发的接口，加以简单重写即可扩展为新的实现。

# 产品性能

* 支持文件150万份/24小时以上的高速转换；
* 系统采用多点并行计算、自动负载均衡等机制进行格式转换，转换节点可扩充，支持任意多节点的热插拔接入和退出，系统采用单服务器部署时流式文件的转换速度大于50页/秒，证照等单证类文件生成效率大于500张/秒。

# 部署环境

系统采用JAVA跨平台技术开发，在信创环境和成熟环境下均可稳定高效运行，部署环境如下表：

|  |
| --- |
| **部署运行环境** |
| CPU | Intel、AMD、龙芯、飞腾、申威、兆芯、鲲鹏等 |
| 操作系统 | Windows系列、中标麒麟、银河麒麟、中科方德、UOS等 |
| 数据库 | MySQL、达梦、金仓、神通、通用等 |
| 中间件 | Tomcat、东方通、金蝶、中创等 |
| 电子签章 | 中安网脉、方正、信安中心、数据所、卫士通、金格等 |