

PDF文档转换

—— 韩丙得

本文讨论的文档类型转换狭义上指 `xlsx->pdf`、`docx->pdf`、`pptx->pdf`，其他类型转换使用本文所提及方法可能也能实现，本文未做验证

介绍

在工程在线2.0模板和文件处理相关模块中，文档类型的转换是一个核心功能，尤其是Office文件转PDF。

目前版本使用 `JodConverter` + `LibreOffice/OpenOffice` 实现文档转换功能，并作为子模块内嵌入文件模块提供转换服务。在使用过程中由于兼容性问题会造成转换生成的部分文档发生局部格式错乱问题。

例：

- 1) 测量弹线定位：接闪带安装前按照图纸设计的位置和数量，以结构轴线为基础进行弹线定位，控制好水平线确定支架的安装位置。
- 2) 支架安装：a. 用电锤在所定的支架位置钻孔，支架采用圆钢支架或植入式固定支架，支架应平直，距建筑物表面高度一致；b. 固定支架高度不低于100mm，c. 所有支架必须安装牢固。
- 3) 接闪带敷设：避雷带采用镀锌圆钢直径10mm，首先将所需圆钢用手锤（或钢筋扳子）进行调查，在圆钢一端用气焊熨制一个来回弯，使其错开一个圆钢直径，长度为10倍的圆钢直径，这样可使与其搭接的圆钢成一条直线，将调直的圆钢固定到支架上时应顺直，各段应受力均匀，不得有翘曲、扭转。
- 4) 接闪带刷银粉：接闪带安装后应进行局部调直，全长宜刷银粉漆，银粉

导致格式错乱的原因有两个：

1. 主要问题：`Office` 转换支持问题。
2. 次要问题：用户文档编写不规范问题。如：未正确使用分页符，而是使用空格换行；未正确设置段落格式。

几种技术路线

一、纯Java实现

这种方式主要是针对Open XML格式的Office文档，需要精通ECMA-376规范。

ECMA-376: <https://www.ecma-international.org/publications-and-standards/standards/ecma-376/>

1. Aspose

商业Jar包，需要付费licence授权。转换效果与MS Office原生基本一致，差别很小。

依赖

- Aspose Jar

优点

- 继承简单，只有jar包依赖
- 文档齐全
- 转换效果与原生几乎一致
- 跨平台

缺点

- 付费

二、利用现有Office软件

Office相关软件中基本都有自己文档转换实现，但是转换质量不一。可以通过某种方式调用已有的office转换，来间接实现文档转换。

1. LibreOffice/OpenOffice + JodConverter

实现逻辑：在模块启动时，由 **JodConverter** 启动并维护一个或多个LibreOffice或OpenOffice的 **soffice** 转换进程，通过 **cmd** 或 **shell** 命令利用该进程实现文档转换。

依赖

- 安装LibreOffice/OpenOffice
- JodConverter Jar

优点

- 跨平台(win + linux)
- 实现成熟，可维护多个转换进程



缺点

- 转换失真（excel转换偶现格式问题，word转换bug稍多）

2. MSOffice + document4j

documents4j: <https://github.com/documents4j/documents4j>

实现逻辑：模块启动时，由 **document4j** 启动并维护一个MS Office后台进程(**WINWORD.EXE** 或 **EXCEL.EXE**)。

 WINWORD.EXE	396	正在运行	hanbd	00	11,232 K	不允许
 EXCEL.EXE	5928	正在运行	hanbd	00	11,764 K	不允许

在执行文档转换时由 **document4j** 使用 **cmd** 命令调用相应的 **.vbs** 脚本来执行转换任务。

```
word_assert.vbs
word_convert.vbs
word_shutdown.vbs
word_start.vbs
```

```
' Transforms a file using MS Word into the given format.
Function ConvertFile( inputFile, outputFile, formatEnumeration )

    Dim fileSystemObject
    Dim wordApplication
    Dim wordDocument

    ' Get the running instance of MS Word. If Word is not running, exit the conversion.
    On Error Resume Next
    Set wordApplication = GetObject(, "Word.Application")
    If Err <> 0 Then
        WScript.Quit -6
    End If
    On Error GoTo 0

    ' Find the source file on the file system.
    Set fileSystemObject = CreateObject("Scripting.FileSystemObject")
    inputFile = fileSystemObject.GetAbsolutePathName(inputFile)

    ' Convert the source file only if it exists.
    If fileSystemObject.FileExists(inputFile) Then

        ' Attempt to open the source document.
        On Error Resume Next
```

依赖

- winserver/windows + MS Office 2010以上版本
- document4j Jar

优点

- 原汁原味，不失真
- 实现成熟，能完整维护转换进程的生命周期，可以池化

缺点

1. 只支持Windows平台，不支持Linux
2. document4j 作者暂未就 PPT 的转换做出支持，目前只支持 WORD 和 EXCEL

3. MSOffice + jacob

jacob: JACOB一个Java-COM中间件.通过这个组件你可以在Java应用程序中调用COM组件和Win32程序库


<https://sourceforge.net/projects/jacob-project/>

<https://github.com/freemansoft/jacob-project>

ms Office vba Api:

确切来说，**Jacob** 只是一个 **Java-COM** 中间件，与文档转换无关，但是我们可以使用 **Jacob** 的相关 **JNI** 支持来调用MS Office。

实现逻辑：与 **MSOffice + document4j** 基本类似，不同的是 **document4j** 是通过cmd命令调用vbs脚本来实现转换，本方式是通过 **JNI** 直接操作 **MS Office**。

 WINWORD.EXE	396	正在运行	hanbd	00	11,232 K	不允许
---	-----	------	-------	----	----------	-----

Word Demo：

```
private static final ActiveXComponent jacobWordApp;

static {
    jacobWordApp = new ActiveXComponent("Word.Application");
    jacobWordApp.setProperty("Visible", false);
}

private void jacobWord2Pdf(String sourceFile, String destFile) {

    try {
        // Documents对象
        Dispatch docs = jacobWordApp.getProperty("Documents").toDispatch();
        // Document对象
        Dispatch doc = Dispatch.call(docs, "Open", new Object[]{sourceFile, false,
true}).toDispatch();
        Dispatch.call(doc, "ExportAsFixedFormat", new Object[]{destFile, 17});
        Dispatch.call(doc, "Close", new Object[]{false});
    } catch (Exception e) {
        log.error("[jacob] 转换出错: {}", e.getMessage());
    } finally {
        jacobWordApp.invoke("Quit");
    }
}
```

依赖

1. winserver/windows + MS Office 2010以上版本
2. jacob Jar
3. **jacob.dll**

优点

1. 原汁原味，无格式问题
2. 支持 **WORD**、**EXCEL**、**PPT**

缺点

1. 只支持Windows平台，不支持Linux
2. 实现不成熟，需要自己实现转换进程的生命周期管理，开发量较大

对比与优化

对比

	Aspose	JodConverter + LibreOffice	Documents4j + MsOffice	Jacob + MsOffice
windows平台	✓	✓	✓	✓
linux平台	✓	✓		
word -> pdf	✓	✓	✓	✓
excel -> pdf	✓	✓	✓	✓
ppt -> pdf	✓	✓		✓
转换效果(5分制)	好 (4.5)	一般 (3)	非常好 (5)	非常好 (5)
转换速度(5分制)	非常快 (5)	一般 (3)	一般 (3)	一般 (3)
是否免费		✓	✓	✓
集成难度	简单	简单	简单	困难

优化方案

首先排除付费方案 `Aspose` 。根据实际生产环境混合搭配使用其余三种，可以配置为转换链。

winserver

配置转换链，， 优先级为 `Documents4j + MsOffice` > `Jacob + MsOffice` > `JodConverter + LibreOffice` 。 `JodConverter + LibreOffice` 作为最终兜底，保证文档转换能够执行。

linux

保持现有方案不变， `JodConverter + LibreOffice`