CF8 PDF Manipulation: Pulling Text Out

Posted At : December 3, 2007 3:48 PM | Posted By : Cutter
Related Categories: PDF, ColdFusion 8, Development, ColdFusion

So, this morning a friend called me up with a problem. They had received some PDF files from their insurance company, and they needed the data in Word or Excel for manipulation. Now, they could cut and paste the information, but this was time consuming. She went to the Adobe site, trying to find info, and saw 'ColdFusion' on the homepage. This sparked her brain, because she immediately went, "Hey, Cutter does something with ColdFusion! Maybe he can help me!"

Lucky for her, we now have ColdFusion 8, with it's built-in PDF support through the use of the CFPDF tag. I had to do a tiny bit of research on this, because Adobe's CF LiveDocs weren't overly clear, but I eventually found out that I could extract text with some very simple DDX processing directives.

Ray did a series of posts recently about working with PDF documents. Although none of them answered my question directly, he had written one about **using the DDX processing directives.** This sent me searching the Adobe site for more information, which is where I came upon the **Understanding DDX** developer documentation. Basically, by rewriting Ray's simple example, I was able to extract all of the *DocumentText* from the PDF and dump it into an XML file. First I need the DDX, which is just some simple XML:

```
<cfsavecontent variable="myddx">
<?xml version="1.0" encoding="UTF-8"?>
<DDX xmlns="http://ns.adobe.com/DDX/1.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://ns.adobe.com/DDX/1.0/ coldfusion_ddx.xsd">
 <DocumentText result="OutXML">
  <PDF source="Title"/>
 </DocumentText>
</DDX>
</cfsavecontent>
<cfset myddx = trim(myddx)>
```

Then, I verify the validity:

```
<cfif isDDX(myddx)>
yes, its ddx
<cfelse>
no its not
</cfif>
```

Now, a little explanation. Looking at the DDX, you'll notice I've defined a **result** and a **source**. I had tried to define my file names here directly, but ColdFusion didn't like that when I hit the CFPDF tag. Apparently, when using the *processddx* action of the tag, you are required to define your *inputfiles* and *outputfiles*. Further study of the LiveDocs shows that ColdFusion is expecting structures for these defininitions. So, the DDX references certain structure keys (OutXML and Title) which you must define prior to processing your pdf.

```
<cfset inputStruct = StructNew() />
<cfset inputStruct.Title = "rptLauncher2.pdf" />

<cfset outputStruct = StructNew() />
<cfset outputStruct.OutXML = "words2.xml" />
```

You now have all of the necessary pieces. All that's required is your call to process your DDX directives.

```
<cfpdf action="processddx" ddxfile="#myddx#" name="VARIABLES.doc" inputfiles="#inputStruct#" outputfiles="#outputStruct#" />
```

I CFDump the VARIABLES.doc to see my success or failure, which comes out just fine. I now have a file, words2.xml, sitting in my server's folder, which contains all of the content of the PDF file. Simple and sweet.