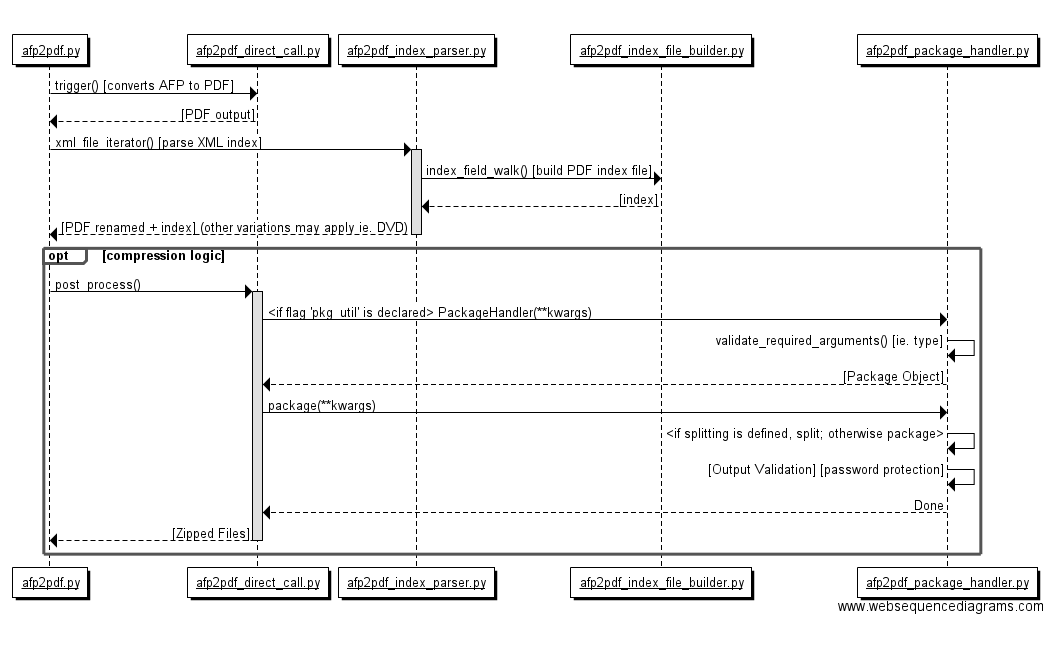
**Project Update – AFP2WEB Packaging Module**

**Business Requirement:**

* The creation of a Python packaging class that will integrate into the AFP2WEB services and perform packaging regardless of the PDF archival variation.
* The program should be capable of splitting a process into multiple packaged files based on either file size and file count.
* Initial support will be built for zip and 7z but should be easily scalable to accommodate future packaging type requirements (ie. tar should be easy to add).
* Allow command line processing for all of the above to act upon a folder of PDFs by giving calling the program with the correct arguments.

**Sequence Flow:**

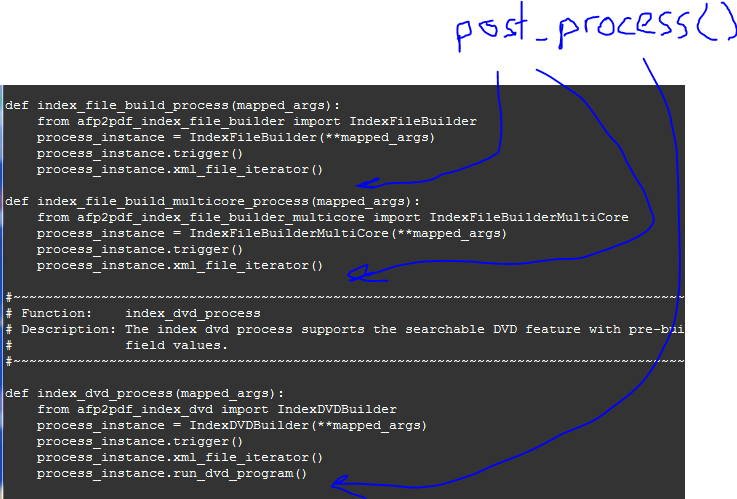


**Add a new “Type” argument to afp2pdf.py:**

This argument will be called “package” which will only apply packaging without AFP2PDF transformation. DP or IS may use this option to manually perform zipping of PDF files for jobs that are still manual.

**Class Design Direction:**

1. For each of the process functions in the entry point script afp2pdf.py, add a post process method call: process\_instance.post\_process(). This will trigger the post process step that you will introduce to handle the packaging class.



1. Write the post\_process() method in the base class afp2pdf\_direct\_call.py. This method will check the self.data dictionary for a key named ‘pkg\_util’:

**[Example: if ‘pkg\_util’ in self.data.keys(): ….code continues… ]**

The dictionary self.data contains all of the data that will come from the configuration file ${cid}${jid}\_pdf.control from the processing script. This means that this will check the values in the processing script for the packaging key.

1. If ‘pkg\_util’ exists (section #2 is true) import your new class PackageHandler

**[Example: from afp2pdf\_package\_handler.py import PackageHandler ]**

This is called “lazy loading”. You won’t load the necessary modules unless a job will require it.

1. Create an instance of PackageHandler and pass it the full self.data dictionary. We will have the PackageHandler constructor [ \_\_init\_\_() ] determine which fields from the self.data dictionary is needed assign the value configs into the kwargs when instantiating:

**[ Example: pkg\_handler = PackageHandler(configs = self.data) ]**

Your Package Handler will take the arguments as \*\*kwargs and parse out configs. **This will allow us to dynamically grow PackageHandler without modifying afp2pdf\_direct\_call.py.**

1. For each configuration that begins with “pkg” assign it to the object attribute (ie. self.pkg\_output\_name = ). Any configurations that can contain multiple values (ie. \_#), create a dictionary using the number as the key and the value as the nested dictionary under the dictionary **“pkg\_values”**. Each nested dictionary will have **“pkg\_src\_path”**, **“pkg\_output\_name”** and/or **“pkg\_src\_file\_pattern**” as the keys. If **“pkg\_single\_output”** is declared use the same **“pkg\_output\_name”** for all \_#s.

Example:

‘pkg\_util: ‘zip,

‘pkg\_single\_output’ : ‘y’,

‘pkg\_values’ : {

‘1’ : {

‘pkg\_src\_path’ : ‘<< d\_dir >>/pdf\_afp/entc/entcms1/pdf/<< cycle >>\_disk01’

‘pkg\_output\_name’ : ‘<< d\_dir >>/pdf\_afp/entc/entcms1/zip/ ENTC\_COMBINED\_<< cycle >>\_<< segment >>\_<< sequence >>.zip’

‘pkg\_src\_file\_pattern’ : ‘entc\_archival\_\*\_\*.pdf | entc\_idx\_file.txt’

‘2’ : {

‘pkg\_src\_path’ : ‘<< d\_dir >>/pdf\_afp/entc/entcms1/pdf/<< cycle >>\_disk01’

‘pkg\_output\_name’ : ‘<< d\_dir >>/pdf\_afp/entc/entcms1/zip/ ENTC\_COMBINED\_<< cycle >>\_<< segment >>\_<< sequence >>.zip’

}

1. The PackageHandler constructor should call a validation method (similar to afp2pdf\_direct\_call’s check\_files() ) but call it validate\_arguments(). This method will perform validations on required fields for packaging. Currently there is only 3 required fields: **‘pkg\_util’**, **‘pkg\_src\_path’** (for all “pkg\_values”) and **‘pkg\_output\_name’** (for all “pkg\_values”)
2. The constructor should also call the translation function translate\_arguments() which will resolve all fields that are nested within << >> with their corresponding field values. If a field value or field is declared in the configurations but not passed in, the program should exit out.
3. After \_\_init\_\_() runs, the object is returned back to afp2pdf\_direct\_call.py. At this point the PackageHandler instances should have all of the arguments it will need. Do NOT carry over the entire configs dictionary from self.data in afp2pdf\_direct\_call.
4. afp2pdf\_direct\_call will receive the instantiated object and call the package() function.
5. The package() function will perform the actual packaging. It should check self.pkg\_util to find out if it is using a 7z or a normal zip packaging. It will iterate through each self.pkg\_values’ \_# field. For each field that has a “self.pkg\_value.#.pkg\_src\_file\_pattern”, it should separate the fields by pipe to search for each value of “pkg\_src\_file\_pattern” recursively in each subdirectory starting from the pkg\_src\_path to generate the pkg\_output\_name file.
6. The package() function will perform translations on the variables enclosed in << >>. If a variable is not defined error out stating that variable needs to be passed in when calling afp2pdf.py.
7. The package() function should also handle the split by count # or size as well as the password. These functionality requirements should preferably come from the Python library or an open source Python module. The package() function should also validate that the ‘pkg\_util’ has the capabilities to support the split by or password features. If it does not, it should error out and tell the user split by count/size or defining a password is not available for the selected ‘pkg\_util’.
8. At the end there should be a validation for the total count of files in the packaged file and the files in the system. This is to make sure that no files are left out (there has been cases were zip totals don’t match and we didn’t know until it got to the client).
9. If there is a selection, use the highest degree of compression by default.
10. If time allows, we may look into implementing the package() function in a multi-threading fashion such that multiple threads are packaging individual folders.

**Configuration Rules:**

The below is a representation of the configurations using ENTC as a placeholder CID.

* The fields **“pkg\_src\_file\_pattern”**, **“pkg\_src\_path”** and **“pkg\_output\_name”** should support multiple folders/outputs. This means that the \_# must to appended to each of the 3 fields listed.
* Packaging field **“pkg\_src\_file\_pattern”** is pipe delimited for all valid patterns. The program will need to iteratively look for all files for each field In all possible subdirectories starting from the **“pkg\_src\_path”.**
* If **“pkg\_src\_file\_pattern”** is not declared in the configs, the program should package the entire tree structure from **“pkg\_src\_path”** as the **“pkg\_output\_name”**
* Packaging fields **“pkg\_src\_path”**, **“pkg\_output\_name”** and **“pkg\_src\_file\_pattern”** will be in a 1:1:1 relationship unless **“pkg\_single\_output”** is declared as “y”. If “pkg\_single\_output” is declared as “y”, combine all **“pkg\_src\_path\_#”** into one **“pkg\_output\_name”**.
* The other exception to the breakage of the 1:1:1 relationship is if the **“pkg\_src\_file\_pattern”** is not supplied. This means that the \_# without a **“pkg\_src\_file\_pattern”** will package all files within the folder path.
* Packaging fields **“pkg\_src\_path”** and **“pkg\_output\_name”** have the ability to use variable fields surrounded by << >>. These fields will be passed in when calling afp2pdf.py as a key value pair similar to the format used for **“InputFilename”** and **“OutputFilePath”**. These will be translated when executing the package() method call.
* **pkg\_src\_file\_pattern\_[#]:** If this field is not declared, the program will zip all files located in the pkg\_src\_path[#]
* If a packaging utility does not support a specified feature such as splitting, password protection, etc. exit out and issue a fatal error with the message “ERROR: Feature $feature not supported with $pkg\_util” .

**Configuration Example:**

pkg\_util = zip

pkg\_src\_file\_pattern\_1 = entc\_archival\_\*\_\*.pdf | entc\_idx\_file.txt # Put \* for any variable arguments.

pkg\_src\_path\_1 = << d\_dir >>/pdf\_afp/entc/entcms1/pdf/<< cycle >>\_disk01/

pkg\_src\_path\_2 = << d\_dir >>/pdf\_afp/entc/entcms1/pdf/<< cycle >>\_disk02/

pkg\_output\_name\_1 = << d\_dir >>/pdf\_afp/entc/entcms1/zip/ ENTC\_ZIP\_DISK1\_<< cycle >>\_<< segment >>\_<< sequence >>.zip

pkg\_output\_name\_2 = << d\_dir >>/pdf\_afp/entc/entcms1/zip/ENTC\_ZIP\_DISK2\_<< cycle >>\_<< segment >>\_<< sequence >>.zip

pkg\_output\_ignore\_subdir = y # If yes, junk the path past “pkg\_src\_path”.

pkg\_output\_split\_size = count|10000 # Can accept size in similar format

pkg\_output\_password = 12345678 # Password value

**Supporting Updates:**

A supporting update must be made to the core program **afp2pdf\_index\_parser.py** and **afp2pdf\_index\_file\_builder.py** to support output generation into subdirectories.

* If output needs to be redirected into a subfolder, the GROUPINDEX field **OUTPUT\_FILE\_PATH** must be declared.
* If a variable cannot be resolved on the DFA level (ie. cycle or segment) the GROUPINDEX field **OUTPUT\_FILE\_PATH** can contain values enclosed in << >> to be resolved by the processing script. This should match with that is contained in the **“pkg\_src\_path”**.
* Do a translation for all fields nested in << >> before moving the PDF file. If a field does not have a value, error out.
* When performing the PDF renaming step under **‘rename\_idx\_pdf\_after\_parse’**, rename the file under the path specified with **OUTPUT\_FILE\_PATH** if OUTPUT\_FILE\_PATH is declared. Otherwise, use the existing logic.
* If the path does not exist, create the path before renaming.