User Guide

# About

The purpose of this app is to collect data from an input file (`in\_file`) and add the information into a master Excel sheet (`master\_file `/`master.xlsx`), using column mappings defined by a configuration JSON file (`col\_file `/`col\_config.json`), and save it as a new file.

# Requirements

The required library as of November 2023 is pandas 2.0, pytesseract, and pdf2image. Tesseract is also required.

# Definitions

`in\_file `: The input file that you want to take information from. The accepted file types are Excel sheets and PDF files, using the extensions `.xlsx` and `.pdf`. The Excel file must be comprised of a single sheet of data, and the PDF file must contain the data on a single page.

`master.xlsx`: The Excel workbook to update with the input file’s information. This file’s data must be on a single sheet.

`col\_config.json`: The file that contains the file metadata and column mappings from input to master. For more information, refer to the section [Column Configuration JSON](#_Column_Configuration_JSON).

# Usage Instructions

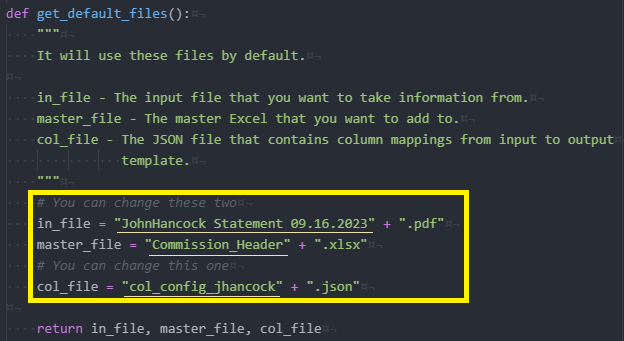
## Quick Overview

### Method: Visual Studio Code

Open `mv\_comm.py` in Visual Studio Code and run the program.

## Modifying File Names

To change the file names or file paths, modify the strings in the function `get\_default\_files()` in `mv\_comm.py`, which starts on line 13.



For example, if you wanted to change the input file, you might change `"JohnHancock Statement 09.16.2023"` to `"Allianz Statement 10.04.2023"`.

Regardless of whether there are whitespaces or not in your file/directory path, you must enclose them in quotation marks. Using single quotes is also acceptable, but you must keep it the same for each variable. Furthermore, backslashes may need to be doubled up to indicate that a backslash is not going to be an escape character itself, for example `"C:\\Documents\\Destination\\filename.xlsx"`.

## Column Configuration JSON

Example, from November 2023 request:

|  |
| --- |
| {  "src\_meta": {  "name": "Allianz",  "offset": 10  },  "col\_map": {  "Policy Number": "Policy #",  "Insured/Annuitant Name": "Applicant",  "Product Type": "Plan",  "Line of Business": "Product Type",  "Bill Mode Description": "Mode",  "Comm Base": "Premium",  "Commission Amount": {  "Description": {  "First Year": "First Year",  "Renewal Premium": "Renewal"  }  },  "Agent Name": "Employee Name",  "Transaction Date": "Effective Date"  }  } |

If you want to understand JSON’s more, please look it up somewhere else. Basically, the structure used here involves key-value pairs.

The file must contain two keys at the first level: “src\_meta” and “col\_map” *exactly*. The “src\_meta” value must be a dictionary with at least two more key-value pairs, the “name” and the “offset”. The “name” corresponds to the commission source’s name, and it must be one that has been requested and completed in the past.

For Excel sheet inputs, the “offset” indicates the row number of the input Excel sheet where the *header information* is contained, so in the above case, it would be the row where “Policy Number”, “Insured/Annuitant Name”, and so on are stored.

For PDF sheet inputs, the “offset” indicates the page number of the input PDF file where *all the relevant information* is contained. It would be where the header and the table of data is.

Example template:

|  |
| --- |
| {  "src\_meta": {  "name": "Company",  "offset": 3  },  "col\_map": {  "Input Column1": "Column1 in Master Sheet",  "Input Column2": "Column2 in Master Sheet"  }  } |

In “col\_map”, the string to the left of the colon should exist as a header in `input.xlsx`. Ideally, the string to the right of the colon should exist in `master.xlsx`, but this is not verified in this app. Each string should map to another string, but there can be exceptions to this rule. Exceptions must be defined by you and subsequently implemented by the coder. Feel free to submit a request.

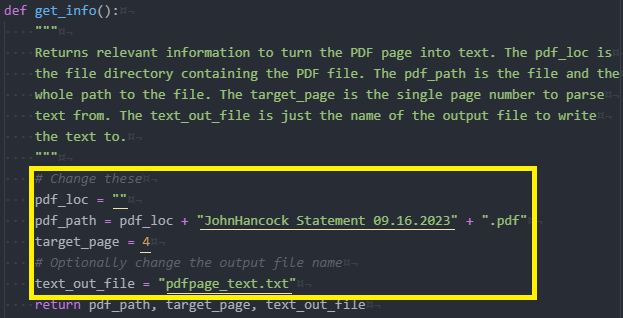
### Valid Names and How to Configure Them

As of November 16, 2023, valid names are as follows:

|  |
| --- |
| For Allianz |
| "src\_meta": {  "name": "Allianz",  "offset": <int: row number for the header>  } |
| For John Hancock |
| "src\_meta": {  "name": "JHancock",  "offset": <int: page number containing the relevant information>,  "text\_name\_line": <int: line number for the line containing "NAME:">,  "text\_table\_start": <int: line number for the first row of the table data>,  "text\_table\_end": <int: line number for the final row of the table data>,  "column\_names": [<list of strings that are the column names>]  } |

Further information on John Hancock (PDF) files…

* If the line numbers are not very clear, you may need to examine Tesseract’s parsed text to know what to set as the line numbers for the text\_name\_line, text\_table\_start, and text\_table\_end. To get a .txt output of the text extracted from the PDF page, use `pdfpage\_to\_text.py`. Like this program and similar requested scripts, you just need to set the file path and file name, but in addition, you must set the target\_page, which is the page of the PDF file that you wish to parse. You may change the name of the text\_out\_file if desired. These variables can be changed at the top of the script.



## Unique Map Functionality

As of November 16, 2023, there are two **unique mappings**, which is controlled by the src\_meta’s name in the col\_config file.

|  |  |
| --- | --- |
| name | Unique Mapping |
| Allianz | Rule 1 of 1:  If “Description” is “First Year”, then “Commission Amount” is saved as “First Year”.  Otherwise, if “Description” is “Renewal Premium”, then “Commission Amount” is saved as “Renewal”. |
| JHancock | Rule 1 of 1:  If “Earn Type\*” is “REN” or “RN-EX”, then “Amount” is “Renewal”.  Otherwise, “Amount” is “First Year”. |

# Troubleshooting

If you receive `JSONDecodeError`, check that your `col\_config.json` file is correctly formatted. For example, you can use this website (<https://jsonlint.com/>) to check your JSON input, or you can search for a different validator, using a search query such as “JSON validator”.

If you receive `UnicodeDecodeError`, check that you passed the correct JSON file, and that the file is correctly formatted.

If you receive `FileNotFoundError`, check the names of all your files.

Other types of errors may occur, and they usually will have some additional notes that will help you figure out what was found to be incorrect.

ISSUE: (2023/11/16) More error messages have been added and some have been changed, but the following table has not yet been updated.

|  |  |
| --- | --- |
| Message | Issue |
| Please verify the top level of "<col\_file>" contains two dictionaries, "src\_meta" and "col\_map". | Exactly two top-level key are required in the col\_config file. |
| Please verify the top level of "<col\_file>" contains the keys "src\_meta" and "col\_map". | Exactly the top-level keys, “src\_meta” and “col\_map”, are required in the col\_config file. |
| Please verify that src\_meta contains at least two keys, "name" and "hrow\_start". | The src\_meta in the col\_config file must contain a dictionary that contains at least the two keys. |
| Please verify that src\_meta contains at least the keys, "name" and "hrow\_start". | The src\_meta in the col\_config file must contain a dictionary that contains at least the keys, “name” and “hrow\_start”. |
| <Name> is not a valid name. | The name in the src\_meta in the col\_config file must be a valid name. Find a list of valid names in the section [Valid Names](#_Valid_Names). |
| src\_meta's name key must map to a string. | The name in src\_meta in the col\_config file must be a string. |
| src\_meta's hrow\_start is mapped to a non-numeric or negative string. | The hrow\_start in src\_meta in the col\_config file must be a positive number or a non-negative string. |
| src\_meta's hrow\_start key must map to a positive number or number string. | The hrow\_start in src\_meta in the col\_config file must be a positive number or number string. |
| src\_meta's hrow\_start must be greater than 0. | The hrow\_start in src\_meta in the col\_config file must be a positive number or number string. |
| Please verify that the columns in the column configuration file exist in the input Excel sheet and that the header row is correct. | The input column names in col\_config must exist in the input Excel. |
| This file name already exists in this directory. The new file was not generated. | An updated master spreadsheet cannot be created if the name is the same. Try running the app again or clearing unnecessary files from the directory. |
| Input Excel file and master file cannot share the same name. | The input Excel file and master Excel file should have different names. |

# Warnings, Issues, and Considerations

Warnings

* Strict Updating
  + If the column headers were not specified exactly, it may fail to merge the desired new data into the master spreadsheet. For example, if the `col\_config.json` had changes it to “Policy #”, and the master spreadsheet `master.xlsx` had a typo where there were two spaces between the words as such “Policy #”, then the new input data will not be added to the column because they are not exactly the same.
* Special Characters
  + Do not use fancy character encodings or very special characters. It might be possible to use special characters for the input spreadsheet’s data because these are just being copied, but the `col\_config.json` file and column headers rely on exact string matches. I would suggest avoiding special characters completely because they may encode differently.
* PDF Tesseract Image Misinterpretations
  + The PDF is converted to an image, and its text is parsed through OCR. This means that there may be errors. For example, “SI” might be interpreted as “Sl” or “S1”.
* File Safety
  + Be careful about overwriting files.

Issues or Further Considerations

* Be careful about using large file sizes.