BEC Data Extraction

1. Introduction:

BEC or Better Energy Communities (BEC) is SEAI’s national retrofit initiative with grant support. They support new approaches to achieving energy efficiency in Irish communities. They aim to deliver energy savings to homeowners, communities, and private sector organisations.

1. Goal:

The purpose of the project is to extract data and review Non Domestic elements of BEC for Brian O’ Mahony (Energy Efficiency Finance) to evaluate the cost credits and measures undertaken by BEC over the last number of years. He will use extracted data for informing decisions on the direction of future users of investment and understanding where the market in relation to different upgrade measures at present.

1. Brief Analysis:

The application is developed to get data of all BEC files since 2015 from than 150 excel files. Tabs in each file need to be extracted are “Project Summary” and many “Non Domestic” tabs.

In Project Summary tab, the data need extracting is under table “Better Energy Communities Programme – Non Domestic Costs” and other four columns alongside with the table: Total Project Cost, Eligible VAT, SEAI Funding, Beneficiary

In Non Domestic tab, a table represents for site reference and table represents for data measures are all extracted

Note: The data is changed slightly every year and there may have no exact pattern for every year, every file but they do have similar format.

1. Resources used:

* Programming Language: Python 3.7
* Software:
* Python 3.7
* Excel
* Co-operators:
* Database: Liam Costello
* Customer: Brian O’ Mahony
* BEC manager: David Boyle

1. Approach:

Here is a flow chart representing the approach of the application

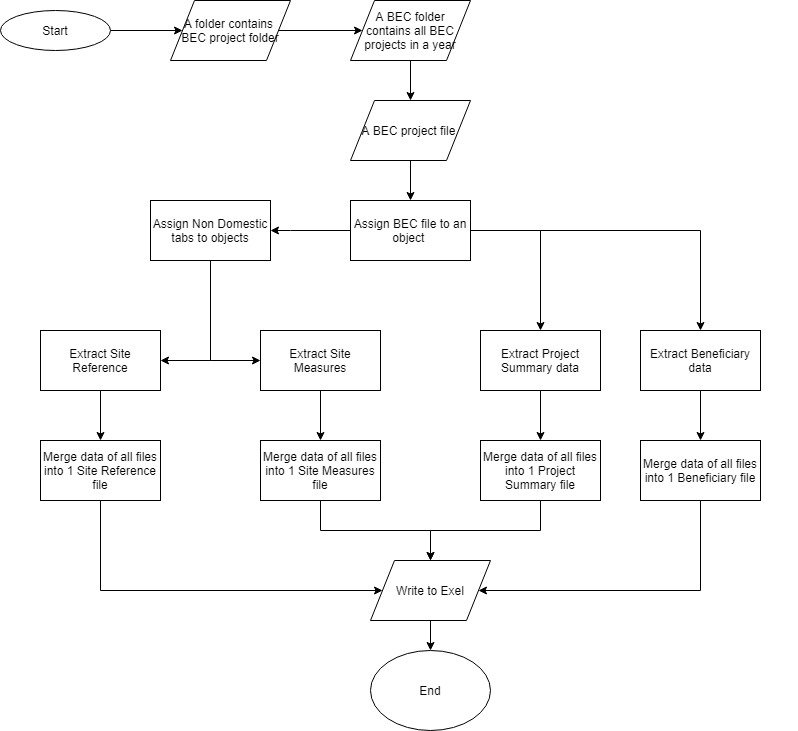
[](https://www.draw.io/?lightbox=1&highlight=0000ff&edit=_blank&layers=1&nav=1&title=Untitled%20Diagram.xml#R7Vtbk9ogFP41PraTm1Ef92Lb6XTbnTqd3e0ba9hIm4SUYNX%2B%2BkIgV%2BJuXK2g0xeFAwQ45%2FsOB0gG7lW8fk9AurjBAYwGjhWsB%2B71wHEmns9%2BuWAjBMOxLQQhQYEQ1QQz9AdKoSWlSxTArFGRYhxRlDaFc5wkcE4bMkAIXjWrPeGo2WsKQqgIZnMQqdI7FNCFlNr%2BpCr4AFG4kF2PnZEoiEFRWc4kW4AAr2oidzpwrwjGVKTi9RWMuO4KvYh277aUlgMjMKF9Gvz6llzDN2g1yW4%2Bvr%2F7%2BeXu8%2Ffpm0LNv0G0lDOWo6WbQgUwYBqRWUzoAoc4AdG0kl4SvEwCyPuxWG5B44glbZb8ASndSKOCJcVMVD3hE8aprCf65B1tnZwUZXhJ5vC5GUmQABJC%2Bkw9vzQBgy7EMaRkw9oRGAGKfjfHASSIwrJepWeWkKreRe2K1mdswFRVfRQxoHMVrxaIwlkK8rmvGNWaipbPg4TC9fMqVKcsG3gSCpKoBWxXddRL2aIG%2BKLZ4XWkKunEoen3hOZYJzT9Dq37ERvuZZaChKVDnr6QvhQS4XspQEnGCkHMgZk8ZmmuOutyesWVRfAP7pyrNuKJbID1hyrmZS4z5UmmbBBFMMIhATGrmEKC2GzZg1plt1XBERhTLlKb5ppVp4zTQRn%2Fn1HGOTfKjHtSZqKTMuPelBF06CCOBfJuGnThYpTwMg4xCM6FNu7ENNp450abSU%2Fa2K5O3kx25E1tGUFM5cpqc57s0E%2BPrpDgpOkhYN%2BHH8M9%2BZE3vSAEbGoVUowS7t3LJ99yQQ0CbisU96yWEcUTK5OWQ3u9lctt6kGMCteI3tfSD7zK26HMXa9lizyzKTIJm8F9VZFnH%2BplVbM8V7QzGkAjnQ62OIv4b9ND2tTZ1yns541dxRtfZBkKkyq8zBdHiymNdc%2FF%2BFEsm%2B31sgWNpuGPv9LpDwSd8b%2FgS8WRh1rJC3yxG2ypyGMQX4Z9%2BbIFBkfygRNjbGqdkU0drT5wuN0Hfsb895r1lFE0z%2BfymBXuUHjCTLvrs41zfe75Bfmjnlh29WJ5pCh%2BuqYE5Pvc23LHO1vGMciHFQAKtAN46DQB7HraAXx2hzjlfeeLALa6bXWkBVa9p6kAPGNAZH9f4RNkCmDT1Q3csXG4VdeyU8et0xe3egND52Xc3kCQLQnUHzC4lnG4VdetU8dt3%2BDX1Xpq7qgOo8It20LAJzRHBoUKvnEu1%2Fu%2FJezPiuIs%2BkVWePu%2BUfOqs3K79dqKbx3hrLzQSY2CN5Cpp6SchZ8GxZ0uP44TV7n5FtTujIrkoZ1ZwZE31M3UYgAmMNX4A2y3735hb6buFzio%2B4XXkKcKzczgTjtCM4A8tqKUI5NHBwn2PczpXma8UdO6w1HLaoJ0slXLcIdYcdTNyk6k6TpIMoI27YMkA2hz0H3MmV8CuX0vTfVEhyMt0aF6Q7sTV5t7OCNY2t7D6Wep1%2FXS5UkfP3h9Xw%2F3tL5T4qkXRXdEhGQ5gqdr%2FgFQyxKGv3Pn2VtWd33w7rgVSgIV4cf8PqSlJN%2FvqaTx7kpi2erDKOGXq6%2FL3Olf)

Figure 1: A flow chart shows the idea of BEX data extraction application

The application will scan a folder to find BEC folders. After that, the application will iterate through each folder and look at BEC excels. There will be an object, which represents for each excel file. This object then decides which tab contains Non Domestic data and which one contain Project Summary and Beneficiary data. For Non Domestic tab, it has its own object while other tabs do not. For each tab, the data will be extracted and stored inside BEC object. This BEC object has a responsibility to collect all outputs and write them respectively to different excel files.

1. Summary of outcome:

Have successfully extracted data from more than 150 BEC files over 4 years since 2015 into four big excel files, which are Beneficiar.xlsx, Project Summary.xlsx, Site Meaures.xlsx, Site References.xlsx. Each file contains exactly number of projects of each year and the data is in right column, extracted from right section. However, there may be some data points missing just because they contain empty data such as empty string, “0” (not supposed to be), etc.

1. Technical notes:
2. Notes:

Manual changes:

* Some files are removed:
* BEC 00355\_Drombane\_LoO workbook\_2015.xlsx in BEC 2015: it does not have the same format
* BEC00434\_Technical Workbook\_LOO\_2015.xls in BEC 2015: it is an empty file and cannot readable
* Changing tab names:
* BEC 00540\_ TechnicalWorkbook \_v5\_LoO IU Rev A 2016.xlsm in BEC 2016: change tab names to tab numbers from tab Non Domestic 1 to Non Domestic 12
* BEC00816\_TechnicalWorkbook\_2018\_Scope .xlsm in BEC 2018: change tab names to tab numbers from tab Non Domestic 1 to Non Domestic 18
* Remove some column in some tab in some file in some project (cannot recall what that file and that column is)
* Adding ‘0000’ for BEC EcoMerit 2014 Technical Workbook Rev 28Jul14.xlsm as it doesn’t have the right format.
* Decrypt manually BEC00193 Technical KCC BEC2014 Application\_rv\_JD.xls because the script cannot bypass the protection

Automatic changes:

* Remove tabs in Non Domestic tabs having a watermark as “Removed” by listing all of them manually
* Do not record “Cost per Primary kWh (excl VAT/Other Costs”) which appears only in files of BEC 2016 and 2015 in Non Domestic tabs.
* Remove empty lines and columns or some default texts presents for empty values.
* Remove Milestone Claim, Amount, Milestone, Claim, Additional information, Description of Minimum Data Required for Existing Specification, Description of minimum Data Required for Proposed Specification in Non Domestic tabs.

1. Assumptions:

* All files in a year have the same format. However, there are some exceptions.
* All tabs in a year have the same format. However, there are some exceptions.
* There will be slight changes between years in the way the name of headers, the index changes. However, the order of columns should not be changed.
* If there is any different between headers of each tab, then that problem must be dealt manually as the script doesn’t have function for that
* If there is any different between headers of each file, then it will be dealt automatically. Headers will follow the latest changes except Beneficiary file. Its header is fixed as “Beneficiary Name”
* Fuzzywuzzy library is used to identify how different headers are. In addition, the standard correctness is 92%. If lower than that, then headers will be labelled as “missing”, otherwise “different”.
* Two tables’ headers are considered the same when they have exactly the same order and texts.
* The extension name of all excel files must be “.xlsm”,”.xlsx”,”.xls”.
* The folder name of BEC in year must be “BEC <year>”
* The outputs will be written into “.xlsx” files
* The output data will contain identification such as ID, Project Year, Project Code, etc

1. How to install Python 3.7:

* Download link: <https://www.python.org/downloads/release/python-372/>
* Choose [Windows x86-64 executable installer](https://www.python.org/ftp/python/3.7.2/python-3.7.2-amd64.exe) if you are using windows
* Instruction: <https://realpython.com/installing-python/#step-1-download-the-python-3-installer> .Under Window section and start with step 1 then step 2.
* After installing., launch CMD -> type python

1. Install Pycharm (or any editor):

* Download link: <https://www.jetbrains.com/pycharm/download/#section=windows>
* Choose Community version
* Instruction: <https://beginnersbook.com/2018/01/python-install-pycharm-windows-mac-linux/>

1. How to run the application:

* Before running, please make sure that all modules are imported (Go to File -> Settings -> Project: [name of project] -> Project Interpreter -> “+” symbol -> search for missing modules which are underlined as red)
* Change the path that contains the folders BEC [year]. For example:

path = os.path.join(**'C:/Users/pphuc/Desktop/Docs/Current Using Docs/'**)

* Then Press Run on the top right of editor