Potential Data Entry
Harry Lo
Applying technologies to overcome real-world issues

109A Suva Street, Upper Riccarton, Christchurch 8041. | 02102824346 | harrykhlo@gmail.com

Brief Proposal for Auto Data Processing

To

Betasbi Limited

Table of Contents

Table of Figures

1 Introduction

This is a brief proposal to suggest potential methods for automatically extracting and manipulating required data from various sources to destinations. Further investigations into the details of implementation will be required.

The possible sources of data would be:

- Excel file
- PDF from documents (i.e. converted from word documents)
- Website

The possible destinations would be:

- Office Applications (e.g Words, Excel, Outlook and so on)
- PDF
- Images

Two different possible ways of extractions for particular sources and its pro and con are depicted in the following sections. A simple example of data extraction from a website (i.e. web scraping) for the purpose of personal learning only is given in the Appendix at the end of this proposal. Extracting data from websites is called web scraping.

This proposal would like to suggest to divide the development into two phases. First, the simple and low-cost method (i.e. Section 2 below) is adopted as a test or trial run at the beginning of the development for quickly getting more detailed requirements of what the actual needs in the existing system. Second, the comprehensive method (i.e. Section 3 below) is built according to the requirements obtained from the first-phase test run.

2 Power Query + VBA from Microsoft Excel or Office Script

The Power Query can be adopted for extracting data from various sources. The VBA or Office Script can be used to control the power query and manipulate the extracted data. Further details are given below. A simple example of the power query of Microsoft Excel is given in the Appendix at the end of this proposal for the purpose of personal learning and experiencing the data extraction from the table of a website.

2.1 Power Query

The Power Query, which is a series of functions in Microsoft Excel, can extract some structured data from various sources such as pdf (for Beta Channel of Office 365 only), websites, text, excel and some others.

2.2 VBA or Office Script

The VBA and Office Script, which are programming languages, can be written to carry out specific tasks in Microsoft Office. It can control the operations of Power Query and manipulate the extracted data.

2.3 Advantages

- Quick
- Easy
- Low cost

2.4 Disadvantages

- Low accuracy
- Lack of control

3 Python and / or R Coding

There are many libraries in Python which can be used for extracting data from pdf files, office applications and websites. Example libraries for data extractions are below:

- OpenPyXL, Pandas and etc. for extracting data from Excel
- PDFMiner, PyPDF2 and etc. for extracting data from pdf
- Beautiful Soup, Ixml, Request and etc. for extracting data from website

R language is a powerful tool for data mining and analysis. It can also be used for extracting data from websites too.

For the manipulation of extracted data, Python has a powerful capability. Further details can be given on requires in order to keep this proposal concise.

3.1 Basic coding concept of web scraping

The web scraping generally is to get data from a website according to target HTML (i.e. HyperText Markup Language) elements which are determined by analysing the contents and HTML structures of the website.

Note that web scraping for commercial purpose may be illegal depending on the policies, natures and usages of the data. Obtaining an approval from the owner of the target website is suggested for commercial purposes. The legality of the web scraping should be further investigated. Further information can be found at the link below:
https://www.tutorialspoint.com/python_web_scraping/legality_of_python_web_scraping.htm

3.2 Advantages

- Quick, once coding or system is developed
- Accurate
- Capable of achieving tailor-made requirements with proper control

3.3 Disadvantages

- Take time to develop the coding or system
- Relatively high initial cost
- Relatively high maintenance cost

4 Appendix

An example of power query of Microsoft Excel to extract a table from a website for the personal learning purpose only is given below:

1) Step 1 for extracting a table from a website (e.g. https://kiwisteel.co.nz/stock/galvanised-steel/)

Select data -> New Query -> From Other Sources -> From Web as below:

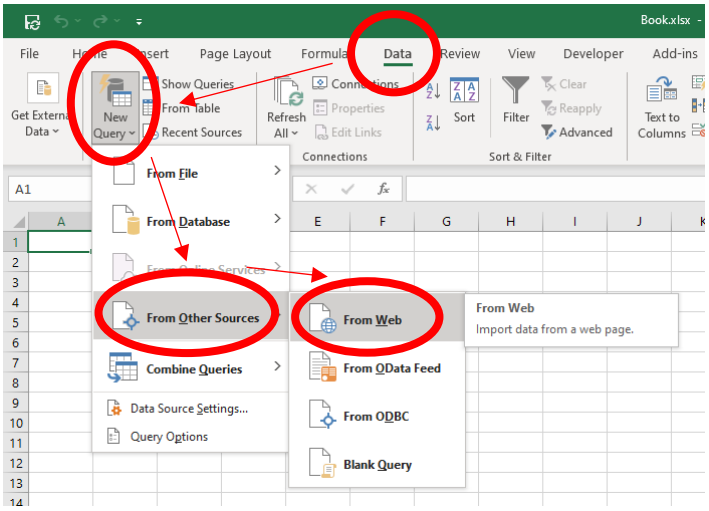

Figure 1: Select extracting data from web (power query example, Step 1)

2) Step 2

Put down the website as below:

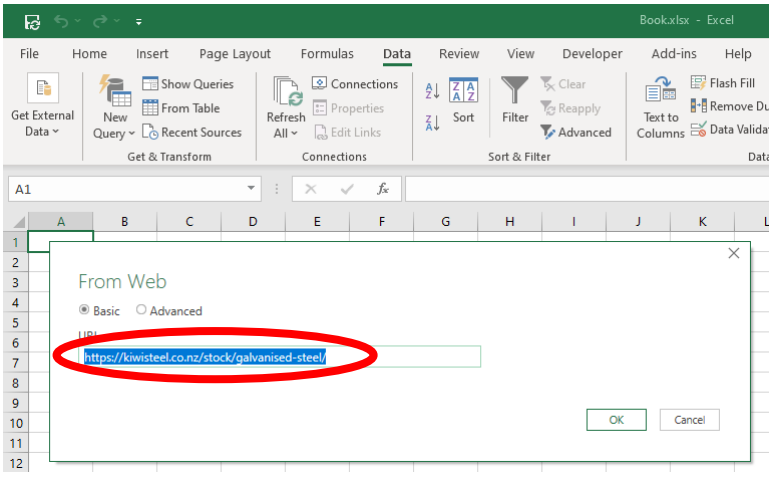

Figure 2: Enter website (power query example, Step 2)

3) Step 3

Select Table 0

Click Load, then the table of the website is loaded into Excel

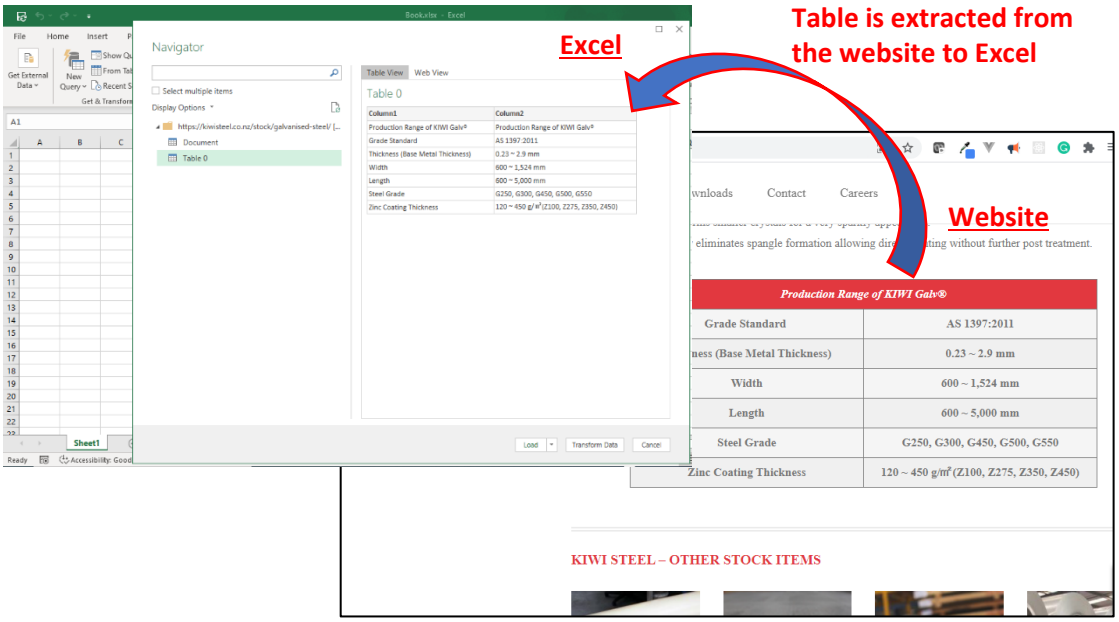

Figure 3: Table is extracted from web (power query example, Step 3)

4) Step 4

The table is loaded into the Excel file as below:

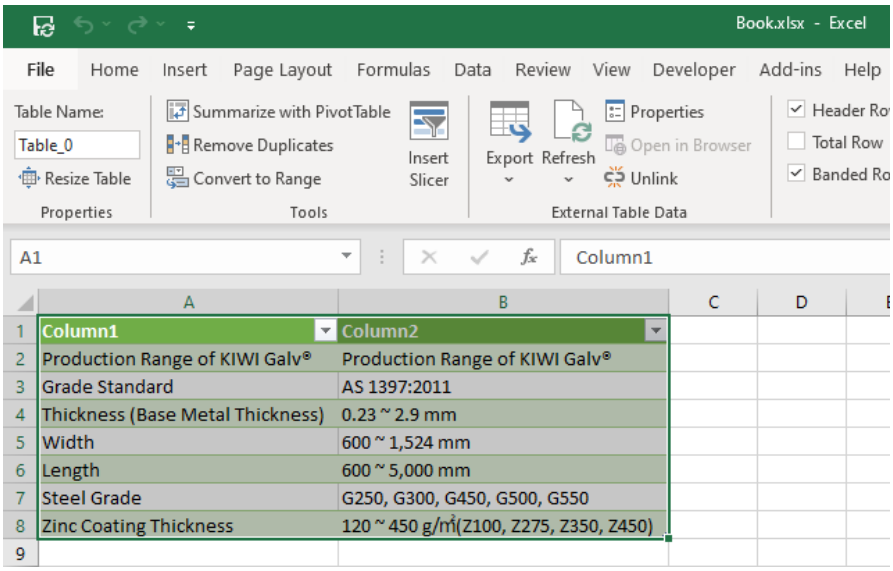

Figure 4: Table is loaded into Excel (power query example, Step 4)