

---

# XML Scraper Documentation

Last updated on Sunday, 14.04.19.

Written by Stephanie Chua, modified by Selina Chua.

---

## Quick Overview

The XML Scraper scrapes all policies from the XML files found in the download bundle here: <https://data.gov.au/dataset/ds-dga-8ab10b1f-6eac-423c-abc5-bbffc31b216c/details?q=%20private> It scrapes all the required information from the XML files and populates an excel sheet for each file found.

## Program Installation & Setup

### Requirements

- Python 3
- Google Chrome
- OpenPyXL
- bs4
- Download bundle from above link

### Installing Python 3

Video Instructions:

<https://www.google.com/search?q=installing+python3+for+windows+10&oq=installing+python3+for+windows+10&aqs=chrome..69i57j0l5.5284j0j7&sourceid=chrome&ie=UTF-8#kpvalbx=1>

1. Download Python 3.7 installer package from the [official Python website](#) for your machine (Windows 32 or 64-bit).
  2. Run the installer. (You can follow the steps below or watch the video linked above.)
  3. Click **Customize Installation**. Optional Features should all be ticked. Click Next.
  4. In **Advanced Options**, click **Browse** on the bottom right under Customize install location. Change it to C:\Python37
  5. Click **Install**.
  6. Right click on the **Windows Start menu** and click **System**.
  7. Click on Advanced System Settings > Environment Variables.
-

8. In System Variables > Path > Edit > New  
**Variable Name:** PYTHON3\_HOME  
**Variable Value:** C:\Python37  
Click OK.
9. Under **System Variables** find **Path** and click on it. Click **Edit**.
10. Click New > copy %PYTHON3\_HOME% into the box > Enter  
Click New > copy %PYTHON3\_HOME%\Scripts into the box > Enter  
Click OK.
11. Go to the C:\Python37 folder. Copy and paste the file called *python.exe* and paste it into the **same folder**. Rename this file python3.exe.
12. Go to Start Menu again, right click Command Prompt (Windows) and type "**python3**" and press Enter. Type "**print("hello world")**" and see if Python works.

## Installing Python Modules

The following modules are used in the XML Scraper and must be installed for it to work.


1. On the Command Prompt type "**cd C:\Python37\Scripts**".
2. Type "**pip3 install bs4**". This will download and install the beautiful soup module.
3. Type "**pip3 install openpyxl**". This will download and install the [OpenPyXL](#) module.

## Downloading the XML Scraper

1. Download the zip file including the required XML files from the government link:  
<https://data.gov.au/dataset/ds-dga-8ab10b1f-6eac-423c-abc5-bbffc31b216c/details?q=%20private>
2. Download the most recent XML Scraper from [here](#) and unzip the folder. It should contain:  
**.py files:** constants.py, xml\_scraper.py, general.py, hosp.py, parse\_funds.py, policy.py  
**folders:** results folder, privatehealth folder
3. Copy the unzipped file containing the XML files into the xml\_scraper folder.

## Running the XML Scraper

1. Open the Command Prompt.
2. Type "cd <location of the *xml\_scraper* folder>" and press Enter.
3. Type "**python3 xml\_scraper.py**" and press Enter.



Note: The program will take a while to run. Leave it while it's running. You can use the computer as you would usually do, although it might be slowed down.

## Important User Information

- **Folder downloaded from government link is currently assumed to be named "privatehealth-04-apr-2019".** If it changes, the program will break. This can be fixed by going to **constants.py** and changing the value for XML\_FILES\_DIR.  
*Change into:*  
XML\_FILES\_DIR = "new\_folder\_name"
- The same error as above may persist if the name of the XML file for funds changes. This can be solved by updating the **constants.py** file.  
*Change into:*  
FUND\_FILE\_NAME = "new funds file name"
- XML Scraper produces Excel sheets that may have duplicates with the same links but with different criteria
- The XML Scraper creates an Excel sheet for **each** XML file that is found inside the government-downloaded folder.
- Program stores sheets in a folder called **results** in the folder. This can also be changed in **constants.py**.

## Technical Guide

1. Program scans all the fund information from **Fund 04-Apr-2019.xml**. (Name may change)
2. Program opens each file in the privatehealth folder and scrapes them for information.
3. After scraping the entire file, the program outputs all the information into an Excel spreadsheet.
4. 2 and 3 are repeated for each file in the downloaded folder.

### xml\_scraper.py

**main()** : Main function running the whole system.

- Parses fund xml file for information.
- Opens each file in the privatehealth folder and scrapes them for information on each product.
- Stores this information and then outputs them to an excel spreadsheet.

**schema3(...)**: Function collects all information for the newly formatted PDFs.

**schema2(...)**: Function collects all information for the old formatted PDFs.

**get\_hosp\_details(...)**: Takes in a product tag and gathers all hospital cover information.

**get\_general\_services(...)**: Takes in a product tag and gathers all general services for the product and stores them in a dictionary.

### constants.py

- **MAIN\_URL** = < Main URL for privatehealth.gov.au >
- **XML\_FILES\_DIR** = < Name of government downloaded folder >
- **FUND\_FILE\_NAME** = < Name of fund xml file in government downloaded folder >
- **RESULTS\_FOLDER** = < Name of folder results will be stored in >

→ This folder has to exist (even empty) inside the xml\_scraper folder for program to work.

### parse\_funds.py

**parse\_funds\_file()**: Function that parses through the government's funds xml file and stores it into a dictionary with Fund objects as values.

**Fund(class)**: Class declaration containing all fields required to store relevant information.

## policy.py, general.py, hosp.py

Python scripts containing only class declarations.

**policy.py:** OldPolicy is used for old formatted PDFs. NewPolicy is used for newly formatted PDFs.

**general.py:** Contains class declaration for a general service.

**Hosp.py:** Contains class declaration for hospital details for a policy.

## Other Files

**results :** Folder containing all the spreadsheet results of previous scrapes.

## Contact Information

Selina Chua ([selina.a.chua@gmail.com](mailto:selina.a.chua@gmail.com))