Actualisation des bases de données de l'OMC avec R: apercu pratique et enjeux

Présenté par:

Carolle Elisabeth KEMPA NANGUE

- Doctorante au département d'économique de l'Université Laval
- PJP à l'OMC en 2018, Division de la Recherche Économique et des Statistiques, Section des Statistiques du Commerce International

- "Les propos contenus dans ce document n'engagent que l'auteur et en aucune manière l'OMC"

UNIVERSITÉ LAVAL

Plan de la présentation

- Brève présentation de l'OMC
- Le rôle de la Division de la Recherche Économique et des Statistiques
- Actualisation des bases de données (BD) avec R: les principaux packages
- Illustration avec des flux commerciaux mensuels du département de commerce américain (US Department of Commerce)
- Défis

Une brève présentation de l'OMC

- L'Organisation Mondiale du Commerce (OMC) a été créée en janvier 1995 à la suite de l'Uruguay Round. Son siège est à Genève en Suisse

- Elle gère les accords internationaux du **commerce de marchandises et de services**, et de **propriété intellectuelle**, ainsi qu'un mécanisme de **règlement de litiges commerciaux**

- Elle analyse les politiques commerciales des 164 pays membres et œuvre à renforcer leurs capacités commerciales. Son but principal est **de favoriser la bonne marche, la prévisibilité et la liberté des échanges entre pays**

(www.wto.org/french/thewto_f/whatis_f/what_we_do_f.htm)#

Rôle de la Section des Statistiques du Commerce Internationale (SSCI)

- Le Secrétariat de l'OMC joue un rôle important de coordination des états membres. Il est structuré en 20 divisions, dont la Division de la Recherche Économique et des Statistiques qui fournit des travaux d'analyse et de recherche économique

- La SSCI fournit les données quantitatives liées aux sujets de politique économique et commerciale

- Une de ses principales missions est la **mise à jour** des bases de données sur le commerce international

- Un de mes mandats à l'OMC était d'automatiser cette mise à jour des données avec R. Logiciel très populaire auprès du personnel scientifique de l'OMC

Nécessité d'automatiser la mise à jour

- 72 pays sont concernés par cette mise à jour. Les données pouvant être mensuelles, trimestrielles et annuelles
- Principaux enjeux:
  - les dates de mise à jour varient en fonction des pays
  - nombreux formats de fichiers à traiter : *txt*, *csv*, *zip*, Excel, Word, *pdf, ODS*, *XML*, *php* et *html*
  - le format des pages web officielles diffère d'un pays à l'autre (listes déroulantes, liens URL, etc.)

Utilisation de R pour la mise à jour des BD: les principaux packages

- Le code R rédigé collecte les données sur la valeur des flux de commerce de marchandises entre pays

- Le code utilise les liens URL des fichiers provenant des sites gouvernementaux

- Plusieurs ***packages*** sont exploités suivant le format/provenance des fichiers:
  - ***Base*** et ***utils*** sont deux packages fondamentaux préinstallés dans R
  - **Stringr** est très utile pour manipuler les chaines de caractères
  - ***CANSIM2R*** et ***eurostat*** sont proposés par, respectivement, Statistique Canada et Eurostat pour accéder aux données statistiques avec R
  - ***Antiword*** et ***readxl*** lisent les fichiers *Microsoft Word* et *Microsoft Excel*
  - ***readODS*** permet l'accès aux fichiers *ODS* (OpenDoc Spreadsheet*)*
  - ***Pdftools*** donne accès au format *pdf,* ***xml*** et ***rvest*** au format *XML*
  - ***SelectorGadjet*** *est* une application JAVA pour la collecte sélective de données sur des pages web *(*Web Scraping ou Harvesting) avec R

Utilisation de R pour la mise à jour des BD: *base, utils, stringr, antiword* et *pdftools*

- Quelques fonctions utiles des différents packages:
  - ***base*** et ***utils***: read.txt(), read.csv(), strsplit(), subset(), trimws() et unzip()
  - ***stringr***: str_sub()
  - ***antiword****:* antiword()

  Exemple de données mensuelle commerciales en Word (Kazakhstan)

  http://stat.gov.kz/getImg?id=ESTAT254665

Сыртқы сауда айналымы
Внешнеторговый оборот

млн. АҚШ доллары — млн. долларов С

	қаңтар	ақпан	наурыз	сәуір	мамыр	маусым	шілде	тамыз	қыркүйек	қазан	қараша	желтоқсан	
	январь	февраль	март	апрель	май	июнь	июль	август	сентябрь	октябрь	ноябрь	декабрь	
Сыртқы сауда айналымы													Внешнеторговый оборот
2015	7 503,1	5 984,4	5 829,8	6 839,3	6 921,0	7 216,7	6 659,6	6 349,6	6 109,4	5 762,6	5 137,0	6 211,0	2015
2016	4 351,8	4 540,6	4 619,2	4 823,1	4 811,5	4 989,7	5 070,4	5 353,6	5 543,7	5 332,7	6 090,1	6 587,2	2016
2017	5 255,4	5 538,4	6 206,7	6 278,1	6 875,2	6 822,3	6 275,7	5 976,5	6 719,1	6 700,8	7 329,8	8 124,9	2017
2018	6 325,1	7 011,9	7 809,1	7 547,1	7 394,3	8 205,4	7 976,8	7 979,1	8 217,1	8 653,1	7 582,9	8 787,8	2018
соның ішінде: экспорт													в том числе: экспорт
2015	4 803,4	3 763,0	3 483,0	4 003,6	4 169,1	4 357,6	3 938,0	3 659,8	3 744,0	3 301,6	2 841,5	3 891,2	2015
2016	2 757,3	2 872,8	2 653,6	2 737,1	2 735,2	2 994,2	3 161,8	3 024,8	3 284,0	2 826,8	3 686,5	4 002,8	2016
2017	3 267,4	3 630,0	3 976,5	3 976,6	4 159,9	4 156,9	3 827,7	3 483,3	4 159,6	4 034,6	4 591,0	5 239,8	2017
2018	4 147,5	4 747,2	5 010,8	4 828,3	4 714,0	5 495,8	5 178,6	5 095,1	5 375,8	5 551,0	4 685,9	6 126,2	2018
импорт													импорт
2015	2 699,7	2 221,4	2 346,8	2 835,7	2 751,9	2 859,1	2 721,6	2 689,8	2 365,4	2 461,0	2 295,5	2 319,8	2015
2016	1 594,5	1 667,8	1 965,6	2 086,0	2 076,3	1 995,5	1 908,6	2 328,8	2 259,7	2 505,9	2 403,6	2 584,4	2016
2017	1 988,0	1 908,4	2 230,2	2 301,5	2 715,3	2 665,4	2 448,0	2 493,2	2 559,5	2 666,2	2 738,8	2 885,1	2017
2018	2 177,6	2 264,7	2 798,3	2 718,8	2 680,3	2 709,6	2 798,2	2 884,0	2 841,3	3 102,1	2 897,0	2 661,6	2018

Utilisation de R pour la mise à jour des BD: *base, utils, stringr, antiword* et *pdftools*

- Quelques fonctions utiles des différents packages:

  - ***pdftools****:* pdf_txt()

    Exemple de données mensuelles commerciales en *pdf* (Vietnam)
    https://www.customs.gov.vn/Lists/EnglishStatisticsCalendars/Attachments/821/2018-T05T-2N(EN-RV).pdf

MINISTRY OF FINANCE OF VIETNAM
GENERAL DEPARTMENT OF CUSTOMS

Table: 2N/TCHQ
Customs IT & Statistics Department
Revised Jun 21, 2018

STATISTICS OF MAIN IMPORTS BY MONTH
May, 2018

No.	Main imports	Units	Reporting month		Compared with previous month (%)		Year to date		Compared with previous year (%)	
			Volume	Value (USD)	Volume	Value	Volume	Value (USD)	Volume	Value
	TOTAL VALUE	USD		**20,765,404,595**		**20.7**		**91,517,291,414**		**10.4**
	In which: Import of Foreign Direct Investment (FDI) companies	USD		***11,755,629,423***		***19.8***		***53,810,329,337***		***8.6***
1	Fishery products	USD		163,352,736		*29.4*		697,509,517		*33.3*
2	Milk and milk products	USD		86,120,178		*2.2*		421,643,902		*10.4*
3	Fruits and vegetables	USD		145,552,473		*30.2*		601,521,550		*20.4*
4	Cashew nut	Ton	119,512	247,306,511	*118.1*	*120.1*	385,883	849,230,062	*-9.9*	*1.7*
5	Wheats	Ton	735,866	173,898,854	*37.9*	*45.2*	2,462,173	578,779,890	*14.3*	*30.8*
6	Maize (corn)	Ton	1,048,556	215,463,652	*-6.0*	*-4.6*	4,095,150	811,890,963	*32.3*	*29.9*
7	Soya beans	Ton	158,086	70,697,826	*-7.6*	*-6.3*	720,005	312,959,453	*-1.5*	*-2.7*
8	Animal, vegetable fats and oils	USD		53,414,015		*3.6*		286,406,425		*0.4*
9	Pastrycooks, sweets and cereal products	USD		27,389,911		*33.4*		121,326,102		*26.1*
10	Other edible food preparations	USD		66,343,887		*15.4*		273,668,858		*6.8*

Utilisation de R pour la mise à jour des BD: *readODS, xml, rvest* et *Web Scraping*

- ***readODS**:* read_ods()

Exemple de données mensuelles commerciales en *ods* (Brésil)

http://www.mdic.gov.br/images/balanca-semanal/5_Semana_04_Mes_BCB_Semanal.ods

- xml et **rvest**: read_html(), html_nodes() et html_text()

Exemple de données mensuelles commerciales en *html (*Pérou)

https://estadisticas.bcrp.gob.pe/estadisticas/series/mensuales/resultados/PN01448BM/html

Utilisation de R pour la mise à jour des BD: *CANSIM2R*

- ***CANSIM2R****:* getCANSIM(12100001,…)

https://www.statcan.gc.ca/fra/developpeurs/concordance

https://www150.statcan.gc.ca/t1/tbl1/fr/tv.action?pid=1210000101

Archivé - Commerce international de marchandises par classification des produits, inactif (x 1 000 000)[1 2 3]

Fréquence : Mensuelle

Tableau : 12-10-0001-01 (anciennement CANSIM 228-0059)

Géographie : Canada

Remplacé par:

- Commerce international de marchandises par classification des produits, mensuel

Système de classification des produits …

Total de toutes les marchandises | Appliquer

Enregistrer mes personnalisations | Aide

Ajouter ou enlever une période de référence | Ajouter ou enlever des données | Options de téléchargement

			Total de toutes les marchandises				
			Canada (carte)				
Commerce	**Base**	**Désaisonnalisation**	**mai 2018**	**juin 2018**	**juillet 2018**	**août 2018**	**septembre 2018**
			Dollars				
Importations	Douanière	Non désaisonnalisées	53 054,9	52 689,4	49 651,5	51 982,4	48 848,8
		Désaisonnalisées	50 426,0	50 453,3	50 456,4	49 905,9	49 722,9
	Balance des	Non désaisonnalisées	54 144,9	53 762,9	50 527,3	53 104,5	49 784,6

Utilisation de R pour la mise à jour des BD: *eurostat*

- ***eurostat**:* get_eurostat ('ext_st_28msbec')…
  https://ec.europa.eu/eurostat/fr/data/database

Base de données par thème> commerce international des biens>… données agrégées> …indicateurs à court terme> Commerce des États membres (UE28)… (ext_st_28msbec)

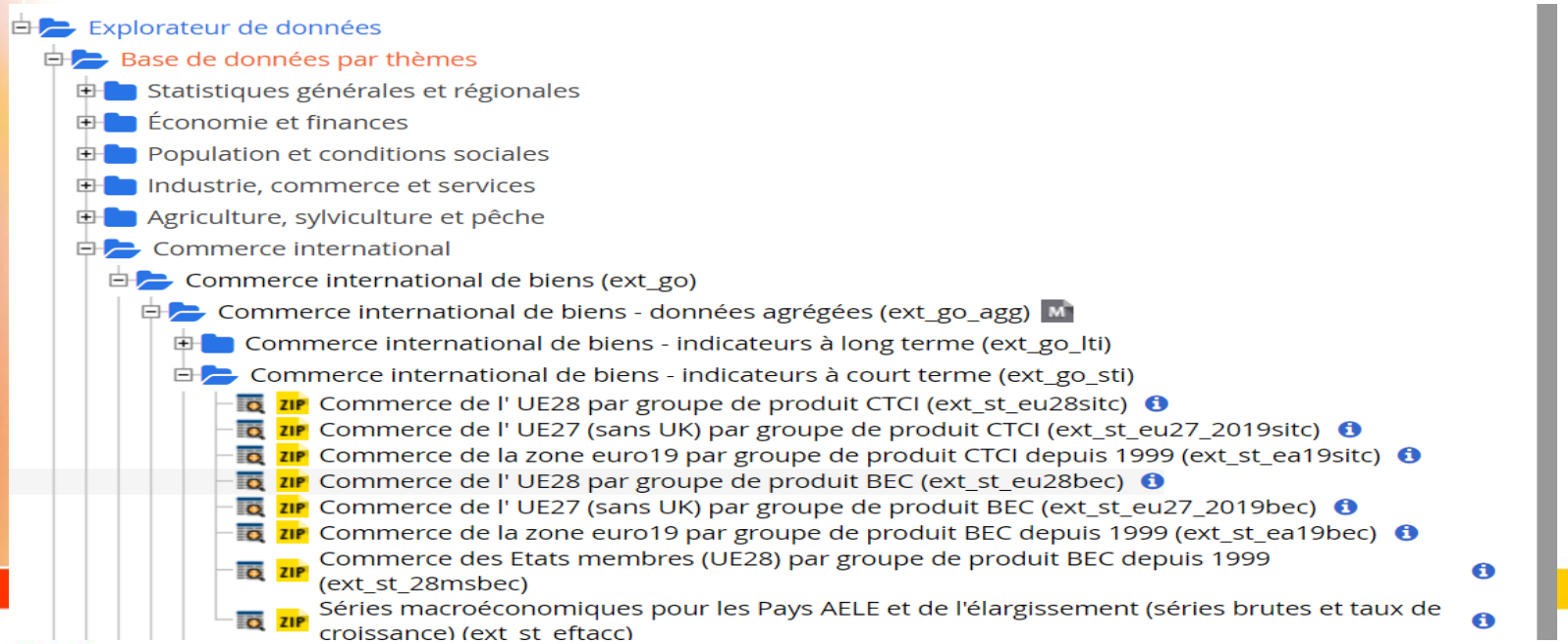


Illustration: site du US Department of commerce

Lien URL du fichier Excel:

Foreign Trade>data>Press releases>FT900>March 2019> Supplement Exhibits> Exhibit 3s-Exports,…

- https://www.census.gov/foreign-trade/Press-Release/current_press_release/exh3s.xls

Illustration: données du US Department of commerce

Exhibit 3. Exports, Imports, and Trade Balance of Goods

In millions of dollars. Details may not equal totals due to seasonal adjustment and rounding.(R) - Revised.

Period	Balance		Exports F.A.S. Value	Imports	
	Customs	C.I.F.		Customs value	C.I.F. Value
	Seasonally Adjusted				
2018					
Jan.- Dec.	-878 700,9	-950 239,1	1 663 982,3	2 542 683,1	2 614 221,4
Jan.- Mar.	-220 589,8	-237 922,4	407 037,6	627 627,4	644 960,0
January	-74 449,1	-80 157,7	132 240,8	206 689,8	212 398,5
February	-76 797,6	-82 643,7	135 206,3	212 003,9	217 849,9
March	-69 343,1	-75 121,1	139 590,6	208 933,7	214 711,6
April	-68 217,2	-73 998,0	139 822,5	208 039,8	213 820,5
May	-65 511,2	-71 419,1	143 463,8	208 975,0	214 882,9
June	-68 726,0	-74 483,4	141 760,2	210 486,3	216 243,6
July	-72 864,9	-78 731,1	139 465,5	212 330,4	218 196,6
August	-76 180,1	-82 031,9	137 773,8	213 954,0	219 805,7
September	-77 100,5	-83 223,7	140 538,1	217 638,6	223 761,7
October	-77 719,2	-84 037,8	140 155,6	217 874,8	224 193,3
November	-71 415,5	-77 518,4	138 907,7	210 323,2	216 426,1
December	-80 376,4	-86 873,3	135 057,4	215 433,8	221 930,8
2019					
Jan.- Mar.	-214 219,1	-232 340,4	416 984,1	631 203,2	649 324,6
January	-72 069,2	-78 281,8	136 907,4	208 976,6	215 189,2
February (R)	-70 822,1	-76 833,1	138 981,9	209 804,0	215 815,0
March	-71 327,9	-77 225,5	141 094,8	212 422,6	218 320,3
April					

Illustration (code R): cas des É-U

- Code R (Code principal, Code spécifique a chaque pays)
- Code principale (Répertoire de travail et package)
- Répertoire de travail

```
##--#--##--#--##--#--##--#--##--#--##--#
# general code (valid for all country) #
##--#--##--#--##--#--##--#--##--#--##--#

rm(list = ls());
if(class(dev.list()) == "integer") dev.off(dev.list()["RStudioGD"]);   # if some plots are open

# Set working directory
directory <- 'C:/DOC_CAROLLE/WTO_INTERNSHIP2018/R queries'
setwd(directory) # definition du nouveau lieu de sauvegarde
path_web_first <- 'C:/DOC_CAROLLE/WTO_INTERNSHIP2018/R queries/Data from web'

# date today
Today <- Sys.Date() #returns the current day in the current time zone. or we can also used Today

# Create Folder for data of Today
if(!file.exists(paste0(path_web_first,'/',Today))) dir.create(file.path(path_web_first,Today))#
path_web <- paste0(path_web_first,'/',Today) #set the name of the file I'm looking for
# paste0 function is to concatenate the elements in the link
```

Illustration (code R): cas des É-U

- Code R (Code principal, Code spécifique a chaque pays)
- Code principale (Répertoire de travail et package)
- Packages (fonction de vérification des packages)

```
#' loading.package
#' This function installs packages that are not installed
#' and load packages not loaded
loading.package <- function(list.of.packages = names){} # end function
# clear console
clear <- function() cat("/014")

# install the packages uninstalled
loading.package(c(
  'stringr',
  'eurostat',      # contains function to pull out data directly from Eurostat
  'CANSIM2R',      # contains function to pull out data directly from Statisti
  'dplyr',         # A fast, consistent tool for working with data frame like
  'magrittr',      # allows other way of witing code, including the pipes '%>%
  'lubridate',
  'tidyr',         # An evolution of 'reshape2' for data management (see e.g.
  'checkmate',
  'readxl',
  'xlsx',          # to read and write excel files - requires R-32 bit
  'openxlsx',      # allows to get cells (and much more) info in a Excel Sprea
```

Illustration (résultats): cas des É-U

- Code R (Code principal, Code spécifique a chaque pays)
- Exécution du code spécifique aux É-U (Résultats sur R)

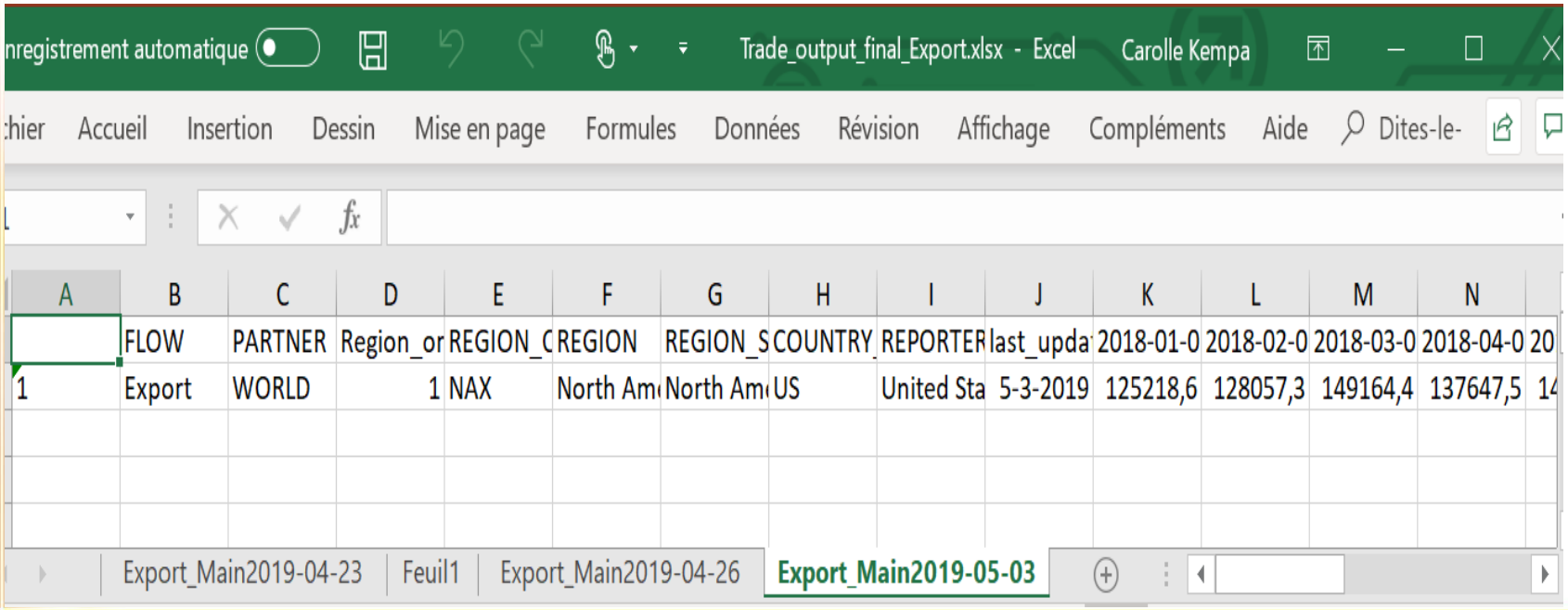
```
> Trade_final_US
# A tibble: 48 x 11
```

	Region_order	REGION	REGION_S	REGION_CODE	COUNTRY_CODE	TIME	COUNTRY	last_update	PARTNER	FLOW	VALUE
	<dbl>	<chr>	<chr>	<chr>	<chr>	<date>	<chr>	<date>	<chr>	<chr>	<dbl>
1	1.	North A~	North Am~	NAX	US	2018-01-01	United S~	2019-05-03	WORLD	Expo~	1.25e5
2	1.	North A~	North Am~	NAX	US	2018-02-01	United S~	2019-05-03	WORLD	Expo~	1.28e5
3	1.	North A~	North Am~	NAX	US	2018-03-01	United S~	2019-05-03	WORLD	Expo~	1.49e5
4	1.	North A~	North Am~	NAX	US	2018-04-01	United S~	2019-05-03	WORLD	Expo~	1.38e5
5	1.	North A~	North Am~	NAX	US	2018-05-01	United S~	2019-05-03	WORLD	Expo~	1.45e5
6	1.	North A~	North Am~	NAX	US	2018-06-01	United S~	2019-05-03	WORLD	Expo~	1.45e5
7	1.	North A~	North Am~	NAX	US	2018-07-01	United S~	2019-05-03	WORLD	Expo~	1.33e5
8	1.	North A~	North Am~	NAX	US	2018-08-01	United S~	2019-05-03	WORLD	Expo~	1.40e5
9	1.	North A~	North Am~	NAX	US	2018-09-01	United S~	2019-05-03	WORLD	Expo~	1.39e5
10	1.	North A~	North Am~	NAX	US	2018-10-01	United S~	2019-05-03	WORLD	Expo~	1.47e5

```
#     with 38 more rows
```

Illustration (résultats): cas des É-U

- Code R (code principal, code spécifique à chaque pays)
- Exécution du code spécifique aux É-U (Output export)



Défis et tâches supplémentaires

- De manière plus pratique pour certains pays:
  - on fait parfois face à des extensions URL inconnues dans R:
    https://www.bps.go.id/tabel/tblblnthn-xls.php?adodb_next_page=&jenis1=&thn1=2018&sumber1=2&coport=&hs_2=&qtr1=&hscd1=&nmhs1 (Indonesia)

    http://px.hagstofa.is/pxen/en/Efnahagur/ Efnahagur__utanrikisverslun__1_voruvidskipti__01_voruskipti/UTA06002.px/table/tableViewLayout1/?rxid=eac2d01c-60bf-47f4-a262-987af99c6df0&downloadfile=FileTypeExcelX (Iceland)
  - Le lien URL peux changer (nécessité de mise à jour du code)
  - on aimerait
    - automatiser le processus de mise à jour à des dates précises
    - programmer la détection de valeurs actualisées aberrantes
    - gérer la quantité des informations

Merci de votre aimable attention