Data Analysis and Visualization

"Using World Bank data in R"

Search, utilize, download, graph, and model data from the World Bank database using R

Ricardo Lima

World Bank Data Base use in R

‘WDI package’

Search, utilize, download, graph, and model data from the World Bank database using R

Ricardo Lima

2023

1. Introduction

The World Bank (WB) is a prominent international financial institution with the primary objective of extending financial support, in the form of loans and grants, to governments of low- and middle-income nations for the execution of capital-intensive ventures. Moreover, the World Bank actively engages in offering valuable policy counsel, conducting research initiatives, and providing technical assistance to aid countries in the effective implementation of their developmental projects and policies. In order to accomplish these endeavors, the World Bank diligently assembles and disseminates an extensive array of data encompassing diverse facets of global development. These data encompass various domains such as economic progress, poverty and inequality, education, healthcare, infrastructure, climate change, among other pertinent topics. These invaluable datasets are conveniently accessible through the World Bank's transparent and comprehensive open data platform, commonly referred to as World Bank Data.

In addition, the World Bank has developed Statistical Performance Indicators (SPI) to monitor countries' statistical performance. The SPI focuses on five main dimensions of a country's statistical performance: (i) data use, (ii) data services, (iii) data products, (iv) data sources, and (v) data infrastructure, including more than 50 indicators and containing data from 174 countries. This set of countries covers about 99.2% of the world's population. Data cover 2016-2019, with some indicators going back to 2004.[1]

To access and analyze data from the World Bank's World Development Indicators (WDI), R software WDI package allows data retrieval directly from the WB data base. The WDI function offers convenient accessibility to a comprehensive range of over 40 databases that are hosted by the World Bank. These databases encompass various crucial datasets such as the World Development Indicators (WDI), International Debt Statistics, Doing Business, Human Capital Index, and Sub-national poverty indicators. To expedite the search process, the WDI package includes a local repository of accessible data series. This local repository can be refreshed to incorporate the most recent version by

[1] https://datacatalog.worldbank.org/search/dataset/0037996/Statistical-Performance-Indicators

employing the WDI cache function. The Package author is Vincent Arel-Bundock and detailed information can be found in the CRAN Package 'WDI'.[2]

2. The WDI Package

It is important to notice that you just have to install WDI package once in your computer. To install WDI, type the following command into the R console or script editor:

```
install.packages("WDI")
```

After the first time you use WDI package, you just have to load the package with the command:

```
library(WDI)
```

Prior to selecting a specific country, indicator, and time period for data analysis, it is advisable to refer to the World Bank Data web page available at: https://data.worldbank.org/.

This platform provides valuable resources that can assist in retrieving relevant data for your research or analysis. It is important to note that the chosen country and indicator will be utilized within the code framework to facilitate data retrieval and analysis procedures. The code for Canada, for instance, is "CAN", and the code for GDP per capita é "NY.GDP.PCAP.CD". A complete list of all country codes (ISO3) can be found in the appendix of this publication and at the URL (webpage location) below:

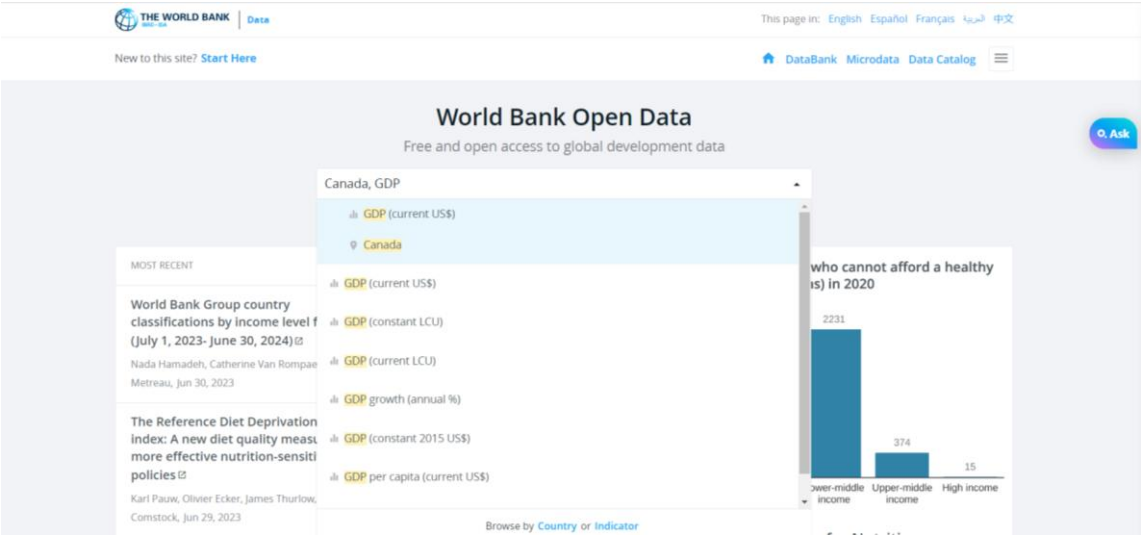
https://wits.worldbank.org/countryprofile/metadata/en/country/all

To obtain the indicator code, you should follow these steps:

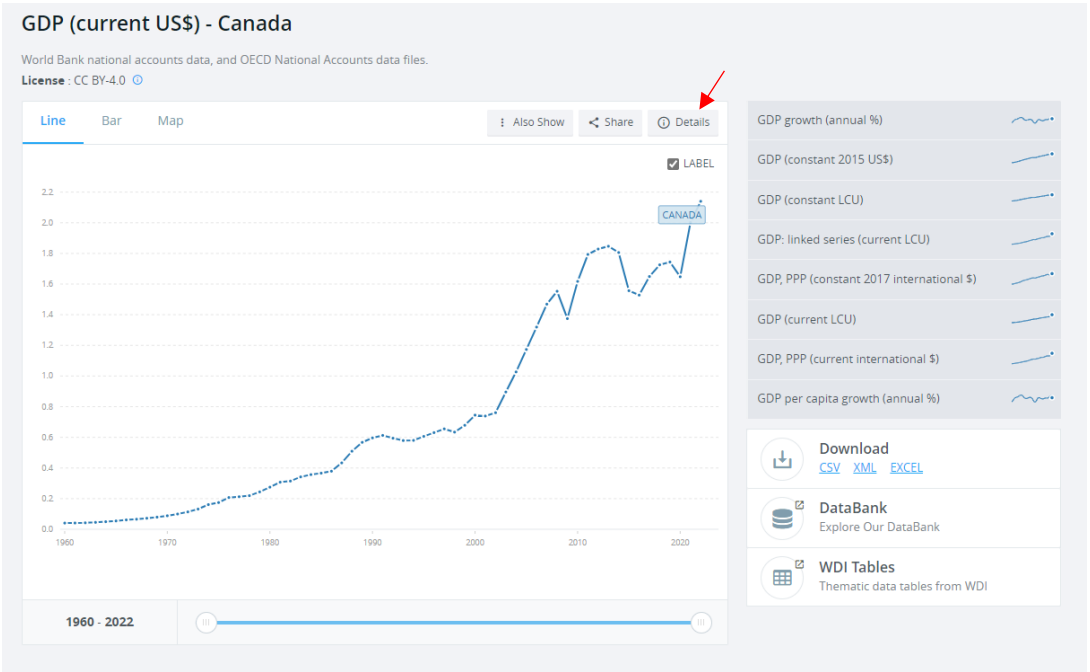
1. Access the WB Data webpage.
2. Select the desired country.
3. Choose the specific indicator you are interested in.
4. On the upper right part of the graphic panel, click on the "Details" button. This action will open a new window displaying the indicator details.
5. Locate the indicator code within the "ID" option.

[2] https://cran.r-project.org/web/packages/WDI/WDI.pdf

Please be aware that when searching on the World Bank website, you will need to use the country name and indicator name. However, when using R programming codes, you must provide both the country code and the indicator code. The most straightforward approach to obtaining the country code and indicator code is by searching on the World Bank Data page, as illustrated below:



When you chose country and the indicator, it is possible to see a graph of the variable in the screen. To get the code and a full description of the indicator, press the details button on the upright part of the graph.



Once you have obtained the country and indicator code, you are now ready to initiate your search.

Assuming that the R WDI package is already installed in your computer, you can load the library (package) and begin your search. The WDI command options are as follows:

WDI(country = "all", indicator = "NY.GDP.MKTP.CD ", start = 1960, end = 2022, extra = FALSE, cache = NULL, latest = NULL, language = "en")

And the arguments are:

country: The country code can be represented using either two letters (ISO2) or three letters (ISO3). In the case of multiple countries, it is necessary to use vector notation. For example, you can use the format c("US", "BR") to represent both Brazil and the United States. To select all countries, you can use the keyword "all";

indicator: indicator code. The indicator code can be represented using vector notation when working with more than one indicator. For example, if you have multiple indicators, you can use the format c("NY.GDP.MKTP.CD", "SP.POP.TOTL") to represent both indicators. Additionally, you have the option to change the name of an indicator for convenience. For instance, if you want to refer to the indicator "NY.GDP.MKTP.CD" as "GDP", you can assign it a new name using the format 'GDP' = 'NY.GDP.MKTP.CD'. This way, you can use the shorthand "GDP" in your code instead of the full indicator code;

start, end: You can set the starting and ending periods for your indicators by specifying the desired years. By default, the entire series is considered, ranging from 1960 to 2022. To customize the time period, you can use the "start" and "end" parameters. For example, if you want to set the starting period to 1990 and the ending period to 2020, you can specify start = 1990 and end = 2020. This way, the data will be limited to the specified time range for your chosen indicators;

extra: When the 'extra = TRUE' parameter is used, additional variables are included in the data set. These variables provide supplementary information such as the observation status (e.g., whether the observation is a forecast), region, name of the capital city, latitude and longitude coordinates, income categories of the World Bank, and lending information;

latest: An integer that indicates the number of the latest available values to read (e.g., if "latest = 5", it reads the five most recent observations);

language: A two-letter code (ISO2) in lowercase indicating the language in which the characters should be provided (e.g., if "language = 'en'", it stands for English language). For a list of supported languages and their codes, you can use WDI::languages_supported(). The default language is English.

The following R commands demonstrate how to search for Canada's GDP in current US dollars and save it in a variable named "CGDP" (you can choose any name for the variable):

```
# Load the WDI library
library(WDI)
# Read Canada's GDP in current US dollars
CGDP <- WDI(country = "CA", indicator = "NY.GDP.MKTP.CD")
CGDP
```

When you run the line CGDP, it will display the annual data of Canadian GDP in current US dollars from 1960 to 2022, resulting in 63 observations. Each observation represents a specific year's GDP value. To display only the first six and last six observations of the Canadian GDP data, you can use the 'head()' and 'tail()' functions as follows:

```
# Display the first six observations
head(CGDP)
# Display the last six observations
tail(CGDP)
```

The result will be as follows:

```
  country iso2c iso3c year NY.GDP.MKTP.CD
1  Canada    CA   CAN 2022   2.139840e+12
2  Canada    CA   CAN 2021   2.001487e+12
3  Canada    CA   CAN 2020   1.647598e+12
4  Canada    CA   CAN 2019   1.743725e+12
5  Canada    CA   CAN 2018   1.725298e+12
6  Canada    CA   CAN 2017   1.649266e+12

   country iso2c iso3c year NY.GDP.MKTP.CD
58  Canada    CA   CAN 1965    54515115736
59  Canada    CA   CAN 1964    49377963149
60  Canada    CA   CAN 1963    45029724490
61  Canada    CA   CAN 1962    42227357845
62  Canada    CA   CAN 1961    40935133543
63  Canada    CA   CAN 1960    40462398502
```

The data shows five columns with the following information:

1) country name,
2) country code (iso2),

3) country ode (iso3),
4) year
5) and GDP value (in current US dollars).

It is important to thoroughly examine the entire time series to identify any missing data or instances where information is unavailable. This approach will provide a comprehensive understanding of the data's completeness. To view the complete time series of Canadian GDP data, you can easily print the CGDP variable. If you are working with two or more time series that are available for different periods, you can use the 'start=' and 'end=' options to establish an equal time period for all variables.

Your search may include more than one country, multiple indicators, and you can also define a specific time period for the search using the options 'start=' and 'end='. Here are a few examples to illustrate this:

1) All countries and two indicators, starting in 1990 and ending in 2000:
   WDI(country="all",indicator=c("AG.AGR.TRAC.NO","TM.TAX.TCOM.BC.ZS"), start=1990, end=2000)
2) Renaming the indicator:
   WDI(country = 'CAN', indicator = c('women_private_sector' = 'BI.PWK.PRVS.FE.ZS', 'women_public_sector' = 'BI.PWK.PUBS.FE.ZS'))
3) Five last available observation for two countries and one indicator:
   WDI(country=c("US","BR"), indicator="NY.GNS.ICTR.GN.ZS", latest = 5).
4) Search names and descriptions of available WDI series:
   WDIsearch(string = "gdp", field = "name", short = TRUE, cache = NULL)

   - String: Character string. Use grep with ignore.case=TRUE to search for this string.
   - Field: Character string. Search within this field. Admissible fields include 'indicator', 'name', 'description', 'sourceDatabase', and 'sourceOrganization'.
   - Short: TRUE - Returns only the indicator's code and name. FALSE - Returns the indicator's code, name, description, and source.
   - Cache: Data list generated by the WDIcache function. If not provided, WDIsearch will search within a local list of series.

3. Organizing the Data

It should be noted that R reads World Bank data starting with the latest observations, as in the example of the Canadian GDP below.

```
country iso2c iso3c year NY.GDP.MKTP.CD
1  Canada    CA   CAN 2022   2.139840e+12
2  Canada    CA   CAN 2021   2.001487e+12
3  Canada    CA   CAN 2020   1.647598e+12
4  Canada    CA   CAN 2019   1.743725e+12
5  Canada    CA   CAN 2018   1.725298e+12
6  Canada    CA   CAN 2017   1.649266e+12
...
```

The result is that R plots a line that begins with the more recent data and progresses towards older information, the codes for the graph are presented below: as illustrated in Figure 1 presented below.:

```
library(WDI)
### GDP  (constant 2015 US$) - trilhoes
Y=WDI(country="KOR", indicator = "NY.GDP.MKTP.KD")
y=ts(Y[,5], start=1960, frequency=1)/10**12
plot.ts(y,ylab="Canadian GDP in US$ trillions")
```

Where,

- library(WDI): Activates the World Bank data reading program (WDI). If you are using WDI for the first time, you need to install it on your computer using the function install.packages(WDI);
- =WDI(): Function to specify the country code (country="KOR") and the indicator code (indicator="NY.GDP.MKTP.KD"). In this case, the country is South Korea (KOR) and the indicator is GDP (NY.GDP.MKTP.KD);
- Y=: Saves the complete indicator matrix in the variable Y. This matrix, however, has 5 columns containing information such as country name, country code, years of data, etc. Only column 5 contains the data relevant to our study;
- y=Y[,5]/(10**12): Creates a new variable y, using only column 5, and divides it by (10^{12}) to obtain the result in billions.

The graph is plotted below:

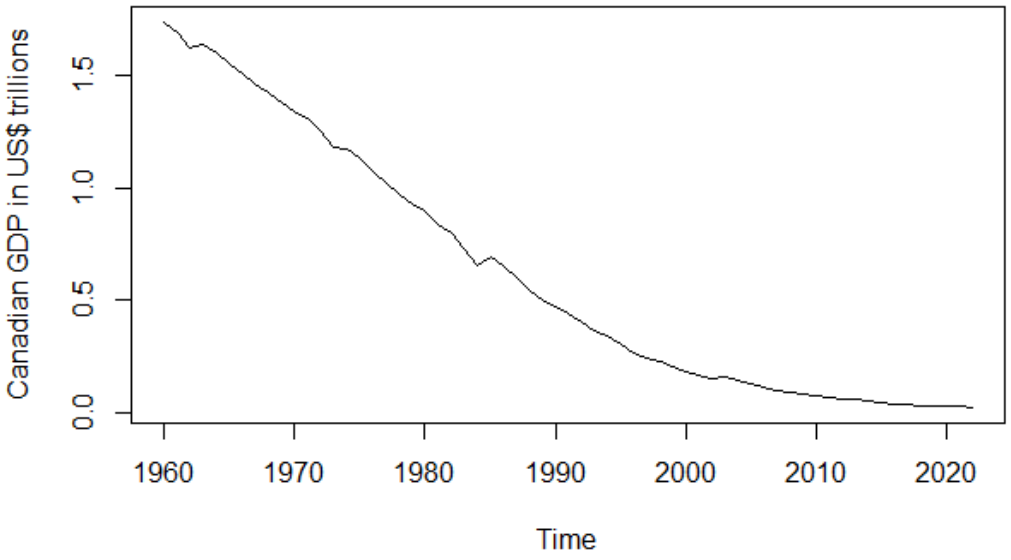


Figure 01 – Canadian GDP in US$ trillions: 2022 to 1960

To address this issue, the command **rev**() can be used to reverse the order of a variable. The updated code utilizing **rev**() is provided below:

```
library(WDI)
### GDP  (constant 2015 US$) - trillions
Y=WDI(country="KOR", indicator = "NY.GDP.MKTP.KD")
y=ts(Y[,5], start=1960, frequency=1)/10**12
plot.ts(rev(y),ylab="Canadian GDP in US$ trillions")
```

And the resulting graph is displayed below:

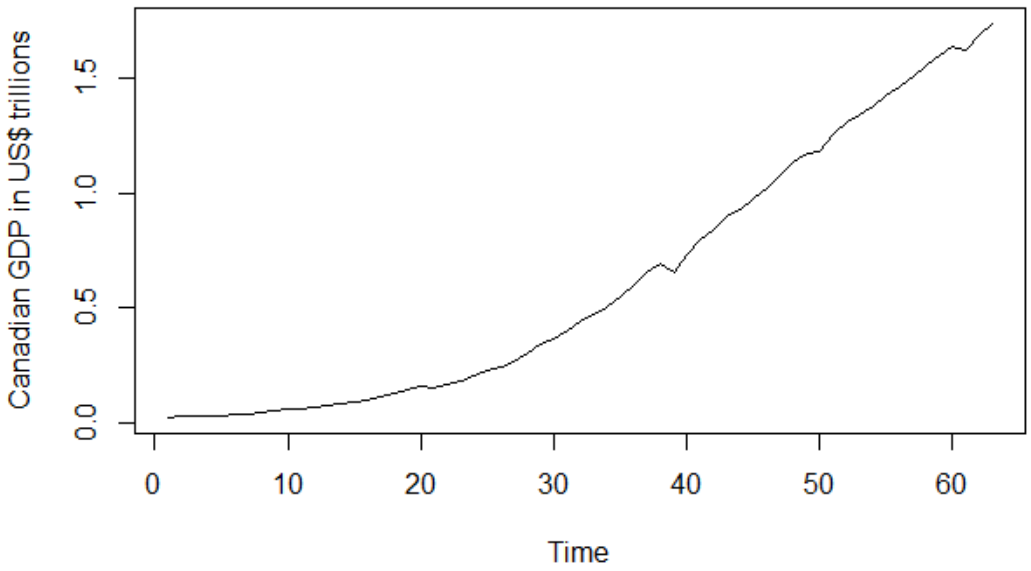


Figure 02 – Canadian GDP in US$ trillions: 1960-2022

4. An Exercise with the World Bank Data

The Long-Term Growth Model (LTGM), published on the World Bank (WB) website, relates a country's economic growth to its capital stock and labor force. In simplified terms, the LTGM can be presented as follows:

$$y_t = AK_t^{\alpha} L_t^{\beta} \quad (2.4)$$

Where:

$\boldsymbol{y_t}$: Gross Domestic Product (GDP) at time t;

A: A constant known as Total Factor Productivity (TFP);

$\boldsymbol{K_t}$: Capital stock at time t;

$\boldsymbol{L_t}$: Labor force at time t;

α and β: Elasticities of capital (K) and labor (L), respectively, where α takes values from 0 to 1 and β = (1- α).

Equation (2.4) is known as the "Cobb-Douglas Production Function" and is usually linearized by taking the logarithm of both sides of the equation. However, for the first exercise, we will use a linear function without using logarithms. The data used will be from the Republic of Korea (South Korea) for the period 1990 to 2021, accessed directly from the World Bank's Application Programming Interface (API) through the R program. Therefore, the linear model to be estimated is presented below:

$$y_t = \beta_0 + \beta_1 k_t + \beta_2 l_t + \varepsilon_t$$

Where:

y_t: GDP of South Korea in constant 2015 dollars;

K_t:: Gross capital formation in constant 2015 dollars;

L_t: Labor force of 15 years and older;

β_j: Represents the parameters of the model (j = 0, 1, 2);

ε_t: Represents the random error.

The access to World Bank data in R can be done using programming codes, such as:

```
library(WDI)
### GDP (constant 2015 US$ trillions)
Y=WDI(country="KOR", indicator = "NY.GDP.MKTP.KD");Y
y=Y[2:33,5]/(10**12); y; plot.ts(rev(y))
### Labor: people or more (millions)
L=WDI(country="KOR", indicator = "SL.TLF.TOTL.IN"); L
l=L[2:33,5]/(10**6); l; plot.ts(rev(l))
### Gross capital formation (Constant 2015 US$ billions)
K=WDI(country="KOR", indicator = "NE.GDI.TOTL.KD");K
k=K[2:33,5]/(10**12); plot.ts(rev(k)); k
```

Where,

- y=Y[2:33,5]/(1012): Creates a new variable y, with information from rows 2 to 33 (1990 to 2021), only for column 5, and divides it by (10^{12}) to express the result in billions;
- plot.ts(rev(y)): Plots the series y. The command rev() reverses the order of the series since the World Bank data starts with the most recent and goes back in time;
- L, l, K, and k: Accessing variables L and K on the World Bank website and constructing variables l and k follow the same logic as variables Y and y. Once the variables are defined, the estimation in the R program is performed as follows:

The following codes are used to run the regression:

```
reg=lm(y~l+k)
summary(reg)
```

The regression results are shown below:

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.075908 -0.028730 -0.006255  0.033098  0.082386

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.89347    0.22865  -8.281 3.96e-09 ***
l            0.11047    0.01459   7.570 2.41e-08 ***
k            0.72442    0.39282   1.844   0.0754 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 0.04174 on 29 degrees of freedom
Multiple R-squared:  0.9901,    Adjusted R-squared:  0.9894
F-statistic:  1443 on 2 and 29 DF,  p-value: < 2.2e-16
```

The estimated model shows that the intercept (β_0) and the coefficient of the labor variable (β_1) are statistically significant at levels lower than 0.1%, while the coefficient of the capital stock variable (β_2) is statistically significant at levels lower than 10%. Given that the labor variable is in millions of workers, the value of the coefficient ($\beta_1 = 0.11$) means that a variation of 1 million people in the workforce causes a variation of 0.11 trillion dollars (or 110 billion) in the country's GDP, in the same direction. The estimated coefficient of the capital stock variable ($\beta_2 = 0.72$) indicates that a variation of 1 billion in the capital stock causes an impact of 0.72 trillion dollars (or 720 billion) on the country's GDP. The negative value of the intercept ($\beta_0 = -1.89$) does not make sense since there is no negative output; however, it can be understood that there is no autonomous GDP, meaning there is no GDP without capital and labor. The adjusted R-squared was equal to 99%, which means that 99% of the variation in GDP is due to variations in labor (l) and capital stock (k). The p-value of the F-test shows that the model as a whole is statistically significant.

The elasticities of the variables can be calculated in R using the following formulas:

```
#Elasticidades
Eb1 = coefficients(reg)[2]*(mean(l)/mean(y));Eb1
Eb2 = coefficients(reg)[3]*(mean(k)/mean(y));Eb2
```

The results are shown below:

```
> #Elasticidades
> Eb1 = coefficients(reg)[2]*(mean(l)/mean(y));Eb1
        l
2.574826
> Eb2 = coefficients(reg)[3]*(mean(k)/mean(y));Eb2
        k
0.2309348
```

The elasticity of the workforce (l) is 2.57, meaning that a 1% change in the workforce results in a 2.57% change in GDP, with all other variables held constant. In the case of capital stock, a 1% change in k leads to a 0.23% change in GDP, all else constant. Therefore, the relative importance of each explanatory variable in the explained variable is better analyzed through elasticities.

.

APPENDIX

Appendix A1 – World Bank Country Codes

Country	Code	Country	Code	Country	Code	Country	Code	Country	Code
Afghanistan	AFG	Congo, Dem. Rep.	ZAR	Heard/McD. Isl.	HMD	Namibia	NAM	Somalia	SOM
Albania	ALB	Congo, Rep.	COG	Holy See	VAT	Nauru	NRU	South Africa	ZAF
Algeria	DZA	Cook Islands	COK	Honduras	HND	Nepal	NPL	South Asia	SAS
American Samoa	ASM	Costa Rica	CRI	Hong Kong, China	HKG	Netherlands	NLD	S Georgia/ S Sa	SGS
Andorra	AND	Cote d'Ivoire	CIV	Hungary	HUN	Netherlands Ant.	ANT	South Sudan	SSD
Angola	AGO	Croatia	HRV	Iceland	ISL	Neutral Zone	NZE	Soviet Union	SVU
Anguila	AIA	Cuba	CUB	India	IND	New Caledonia	NCL	Spain	ESP
Antarctica	ATA	Curaçao	CUW	Indonesia	IDN	New Zealand	NZL	Special Categ.	SPE
Antigua and Barbuda	ATG	Cyprus	CYP	Iran, Islamic Rep.	IRN	Nicaragua	NIC	Sri Lanka	LKA
Argentina	ARG	Czech Republic	CZE	Iraq	IRQ	Niger	NER	St. Kitts/Nevis	KNA
Armenia	ARM	Czechoslovakia	CSK	Ireland	IRL	Nigeria	NGA	St. Lucia	LCA
Aruba	ABW	Denmark	DNK	Israel	ISR	Niue	NIU	St. Vinc./Grenad.	VCT
Australia	AUS	Djibouti	DJI	Italy	ITA	Norfolk Island	NFK	Sub-Sah Africa	SSF
Austria	AUT	Dominica	DMA	Jamaica	JAM	North America	NAC	Sudan	SUD
Azerbaijan	AZE	Dominican Republic	DOM	Japan	JPN	North Macedonia	MKD	Suriname	SUR
Bahamas, The	BHS	East Asia & Pacific	EAS	Jordan	JOR	Northern Mariana Is	MNP	Sweden	SWE
Bahrain	BHR	East Timor	TMP	Kazakhstan	KAZ	Norway	NOR	Switzerland	CHE
Bangladesh	BGD	Ecuador	ECU	Kenya	KEN	Occ.Pal.Terr	PSE	Syrian Arab Rep.	SYR
Barbados	BRB	Egypt, Arab Rep.	EGY	Kiribati	KIR	Oman	OMN	Tajikistan	TJK
Belarus	BLR	El Salvador	SLV	Korea, Dem. Rep.	PRK	Other Asia, nes	OAS	Tanzania	TZA
Belgium	BEL	Equatorial Guinea	GNQ	Korea, Rep.	KOR	Pacific Islands	PCE	Thailand	THA
Belgium-Luxembourg	BLX	Eritrea	ERI	Kuwait	KWT	Pakistan	PAK	Togo	TGO
Belize	BLZ	Estonia	EST	Kyrgyz Republic	KGZ	Palau	PLW	Tokelau	TKL
Benin	BEN	Eswatini	SWZ	Lao PDR	LAO	Panama	PAN	Tonga	TON
Bermuda	BMU	Ethiopia	ETH	Latin Am. & Carib.	LCN	Papua N Guinea	PNG	Trinidad/Tobago	TTO
Bhutan	BTN	Ethiopia/Eritrea	ETF	Latvia	LVA	Paraguay	PRY	Tunisia	TUN
Bolivia	BOL	Europe & Central Asia	ECS	Lebanon	LBN	Peru	PER	Turkey	TUR
Bonaire	BES	Faeroe Islands	FRO	Lesotho	LSO	Philippines	PHL	Turkmenistan	TKM
Bosnia/Herzegovina	BIH	Falkland Island	FLK	Liberia	LBR	Pitcairn	PCN	Turks/ Caicos Isl.	TCA
Botswana	BWA	Fiji	FJI	Libya	LBY	Poland	POL	Tuvalu	TUV
Bouvet Island	BVT	Finland	FIN	Lithuania	LTU	Portugal	PRT	Uganda	UGA
Br. Antr. Terr	BAT	Fm Sudan	SDN	Luxembourg	LUX	Qatar	QAT	Ukraine	UKR
Brazil	BRA	Fr. So. Ant. Tr	ATF	Macao	MAC	Reunion	REU	U Arab Emirates	ARE
British Ind. Ocean Ter.	IOT	France	FRA	Madagascar	MDG	Romania	ROM	United Kingdom	GBR
British Virgin Islands	VGB	Free Zones	FRE	Malawi	MWI	Russian Fed.	RUS	United States	USA
Brunei	BRN	French Guiana	GUF	Malaysia	MYS	Rwanda	RWA	US Minor Outlying	UMI
Bulgaria	BGR	French Polynesia	PYF	Maldives	MDV	Saint Barthélemy	BLM	Unspecified	UNS
Bunkers	BUN	Gabon	GAB	Mali	MLI	Saint Helena	SHN	Uruguay	URY
Burkina Faso	BFA	Gambia, The	GMB	Malta	MLT	Saint M/Dutch	SXM	Us Msc.Pac.I	USP
Burundi	BDI	Georgia	GEO	Marshall Islands	MHL	S. Pierre and Miq.	SPM	Uzbekistan	UZB
Cambodia	KHM	German Dem. Rep.	DDR	Martinique	MTQ	Samoa	WSM	Vanuatu	VUT
Cameroon	CMR	Germany	DEU	Mauritania	MRT	San Marino	SMR	Venezuela	VEN
Canada	CAN	Ghana	GHA	Mauritius	MUS	S Tome and Princ.	STP	Vietnam	VNM
Cape Verde	CPV	Gibraltar	GIB	Mayotte	MYT	Saudi Arabia	SAU	Wallis and F Isl.	WLF
Cayman Islands	CYM	Greece	GRC	Mexico	MEX	Senegal	SEN	Western Sahara	ESH
Central African Rep.	CAF	Greenland	GRL	Micronesia, Fed. Sts.	FSM	Serbia (Serb./Mont.)	SER	World	WLD
Chad	TCD	Grenada	GRD	Middle East/N. Afr.	MEA	Seychelles	SYC	Yemen	YEM
Chile	CHL	Guadeloupe	GLP	Moldova	MDA	Sierra Leone	SLE	Yemen Democ.	YDR
China	CHN	Guam	GUM	Monaco	MCO	Singapore	SGP	Yug/Serb/Mont.	YUG
Christmas Island	CXR	Guatemala	GTM	Mongolia	MNG	Slovak Republic	SVK	Zambia	ZMB
Cocos Islands	CCK	Guinea	GIN	Montenegro	MNT	Slovenia	SVN	Zimbabwe	ZWE
Colombia	COL	Guyana	GUY	Montserrat	MSR	Solomon Islands	SLB	Mozambique	MOZ
Comoros	COM	Haiti	HTI	Morocco	MAR	Guinea-Bissau	GNB	Myanmar	MMR

Appendix A2 – **List of supported languages**

$fully
[1] "en (English)", "es (Spanish)", "fr (French)", "ar (Arabic)", "zh (Chinese)"

$locally
[1] "bg (Bulgarian)", "de (German)", "hi (Hindi)", "id (Indonesian)"
[5] "ja (Japanese)" , "km (Khmer)", "ko (Korean)", "mk (Macedonian)"
[9] "mn (Mongolian)", "pl (Polish)", "pt (Portuguese)", "ro (Romanian)"
[13] "ru (Russian)", "sq (Albanian)", "th (Thai)", "tr (Turkish)"
[17] "uk (Ukrainian)", "vi (Vietnamese)"