

Local Government Training Institute (LGTI)

Course Data Curriculum

Data Fundamentals Course

Module 3: Defining, Finding and Getting Data

Disclaimer: This work is developed by School of Data with funding from The World Bank Tanzania Data programme. For more information visit <https://www.schoolofdata.org>.

Module 3: Defining, Finding and Getting Data

STUDENT WORKBOOK

In this module, we will begin to dive deeper into the concepts, techniques and tools used in the first 3 stages of the Data Pipeline. Learners will go through how to properly define the relevant scope of the question(s) they want to gain insights into and learn about where they can find that data. Once this data has been identified, it can be a simple process of just clicking to acquire the data or a tedious process of having to convert the data into the proper structured format.

And the end of this module, a learner should be able to:

- Define the target, audience, stakeholders and scope (geographic, time frame or sector) of an expected insights.
- Use different techniques and tools to find data from different sources
- Obtain identified data from different sources in machine-readable formats.
- Use several steps to ensure that the source, values, units, methodology and assumptions of data a clear and verified.

Contents

- Lesson 1: Defining Your Interest
- Lesson 2: Finding Data
- Lesson 3: Getting Data

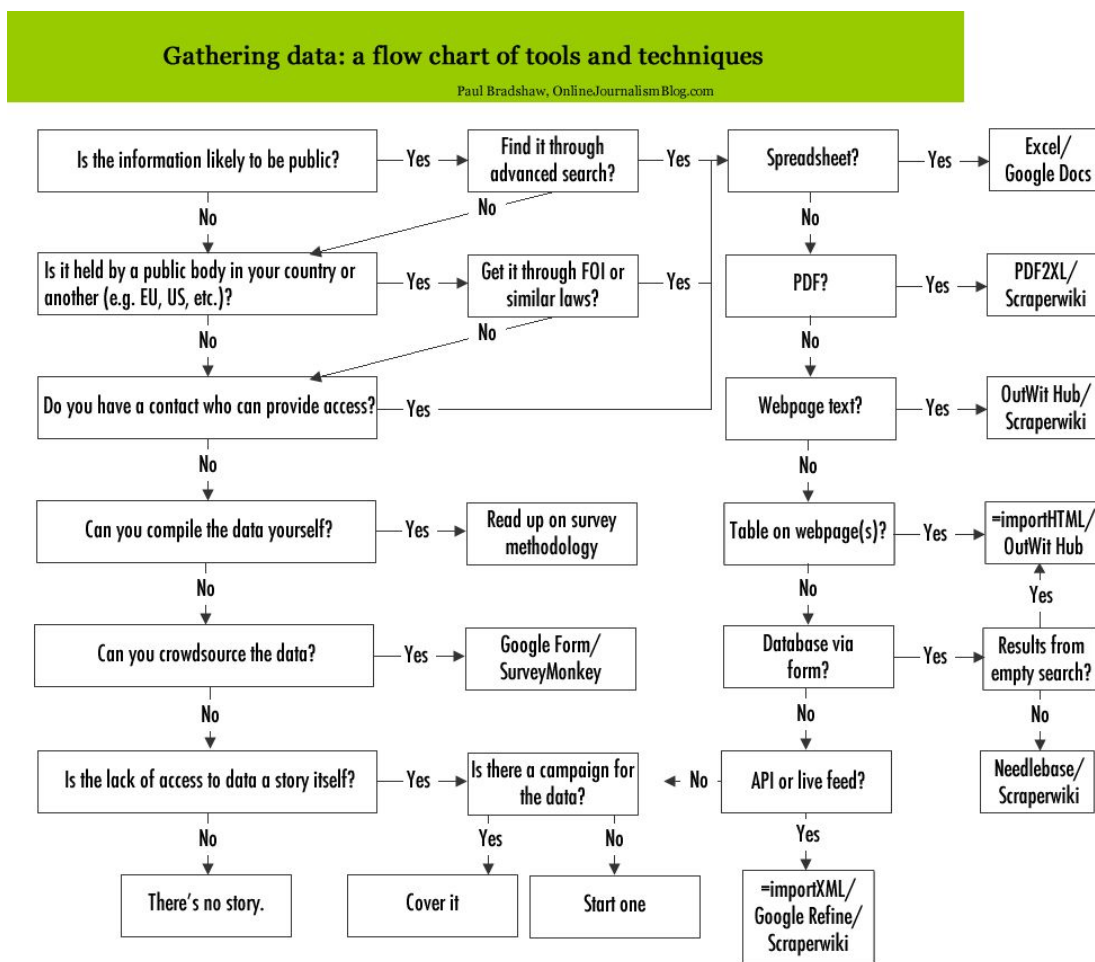
Lesson 1: Defining The Problem

Data-driven projects always have a “define the problem you’re trying to solve” component. It’s in this stage you start asking questions and come around the issues that will matter in the end. Defining your problem means going from a theme (e.g air pollution) to one or multiple specific questions (has bikesharing reduced air pollution?). Being specific forces you to formulate your question in a way that hints at what kind of data will be needed. Which in turns helps you scope your project: is the data needed easily available? Or does it sound like some key datasets will probably be hard to get?

Activity : Population Question or Problem

- Let's say someone comes to you to help them with a data story on population. Is this question specific enough for you to start your work?
- How can you make it more specific?
- Propose a more specific question and see if you can find the data to answer it.

Lesson 2: Finding Data



Gathering data: a flow chart of tools and techniques¹

In the digital age, more data is more available than ever before. In fact, sometimes it feels like we are drowning in data and it is difficult to find the data we are actually looking for. In this lesson, we will explore ways to find data online both through portals and by searching for it. We will also look at options

¹ The flowchart created by Paul Bradshaw flowchart shows common ways journalists try to access data and what they do when they face road blocks along the way. Despite being created for journalists, this can be a reference chart for you anyone searching for data online, has hit a wall, and doesn't know what the next step is:

<http://onlinejournalismblog.com/2011/09/06/gathering-data-a-flow-chart-for-data-journalists-2/>

for when the data we want isn't available and we need to collect data ourselves through 'crowdsourcing' or sensors.

Using Advanced Search

The screenshot shows the Google Advanced Search page. At the top left is the Google logo. Below it, the text "Advanced Search" is displayed in red. The page is divided into two main sections: "Find pages with..." and "Then narrow your results by...".

Find pages with...

Find pages with...	To do this in the search box
all these words:	Type the important words: tricolor rat terrier
this exact word or phrase:	Put exact words in quotes: "rat terrier"
any of these words:	Type OR between all the words you want: miniature OR standard
none of these words:	Put a minus sign just before words you don't want: -rodent, -"Jack Russell"
numbers ranging from:	Put 2 periods between the numbers and add a unit of measure: 10..35 lb, \$300..\$500, 2010..2011

Then narrow your results by...

language:	any language	Find pages in the language you select.
region:	any region	Find pages published in a particular region.
last update:	anytime	Find pages updated within the time you specify.
site or domain:		Search one site (like wikipedia.org) or limit your results to a domain like .edu, .org or .gov
terms appearing:	anywhere in the page	Search for terms in the whole page, page title, or web address, or links to the page you're looking for.
SafeSearch:	Show most relevant results	Tell SafeSearch whether to filter sexually explicit content.
file type:	any format	Find pages in the format you prefer.
usage rights:	not filtered by license	Find pages you are free to use yourself.

At the bottom right of the form is a blue button labeled "Advanced Search".

Google Advanced Search

There are many sources on data on the internet. A useful technique of finding data online is to use Google's advanced search.

Open http://www.google.com/advanced_search

A screen with several search fields appears. The following table explains various search options within Google advanced search. It also provides alternative shortcuts to perform the same search using the regular Google search that you may be familiar with.

Google Advanced Search Feature	Alternative Option in Regular Google Search
All these words is like a regular Google search	Type in all the words you want to find in the regular search bar
Exact word or phrase helps find results in which the words appear in the exact order you mention	Use quotes to search – for example “Local Government Training Institute”
Any of these words helps find results where any of the mentioned words appear	Use OR between words in a search – for example, agriculture OR farming OR crops
None of these words will filter out search results with words that you specify	Type the minus sign before the word you want to omit in a search – for example, Tanzania -Zanzibar
Language: specify the language of the results	-
Region: limit results to only websites from a geographical region	-
Last Update: limit results to recent content	-
Site or domain: Narrow search to specific website	Use this format to search - site:url For example:

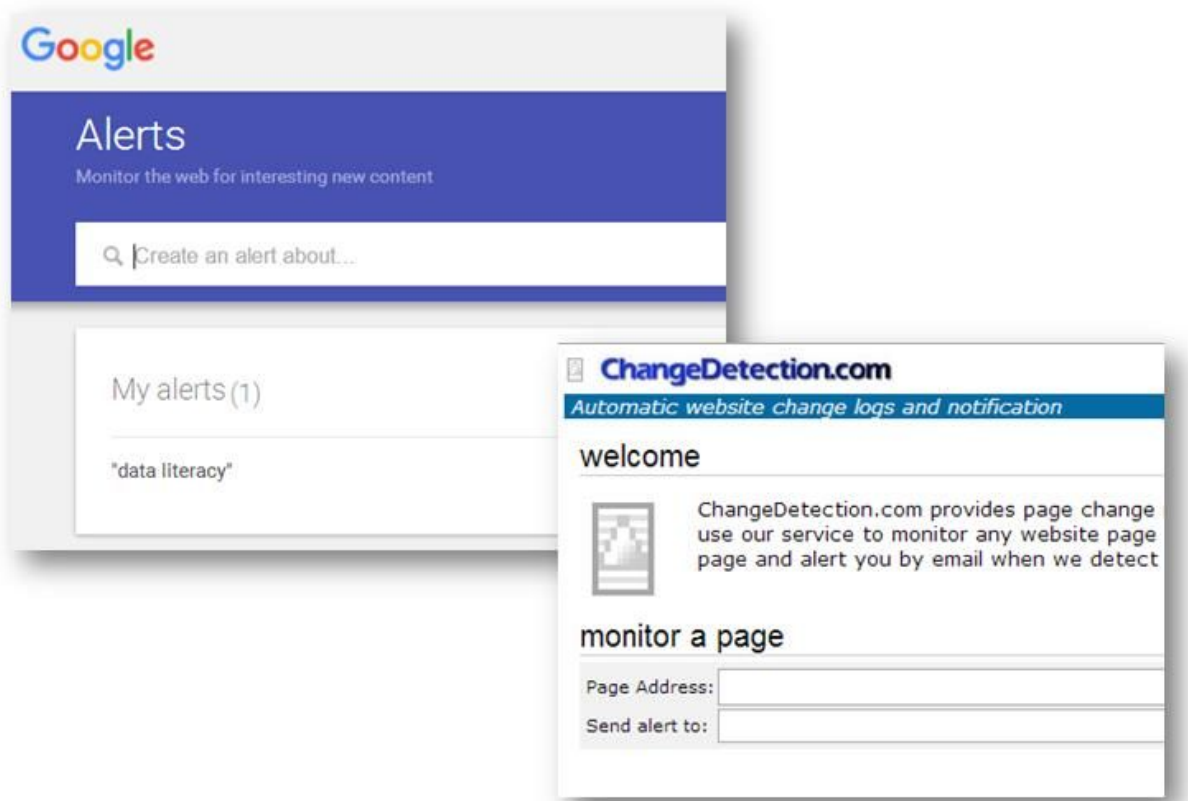
	<p>site:https://tanzania.go.tz/</p> <p>Note that the website address has to be EXACT.</p> <p>CORRECT site: https://tanzania.go.tz/</p> <p>WRONG site:tanzania.go.tz</p>
<p>Filetype: Search only for files with a specific extension (for instance: xls, pdf, doc, ppt etc)</p>	<p>Use this search format - Filetype:[extension]</p> <p>For example, here is a search term to look for XLS files:</p> <p>CORRECT filetype:xls</p> <p>WRONG filetype:Excel</p>

Now let's try using Google advanced search:

- Use 'any of these words' to find content about agriculture, farming or crops in your Tanzania.
- Use 'none of these words' to find information about agriculture in Tanzania and not Zanzibar.
- Find content about Groundnuts only in Tanzania.
- Find content about Groundnuts only from Tanzanian websites
- Find content about Groundnuts published in the last week.
- Search the Tanzanian Ministry of Agriculture website for Excel files

- Search for PDFs about pesticide use in your country.

Setting Up Alerts



If you interested in a particular topic, you can also use the following techniques to receive alerts or updates when something new appears online.

Google Alerts to follow topics

- Step 1: Sign into your Gmail

- Step 2: Go to <https://www.google.com/alerts>
 - Alternatively, you can use <http://www.talkwalker.com/alerts>
- Step 3: Create alert. Be specific. Put in the topic and region or person of interest.
- Step 4: Select how often, source, language, region and how many.
- Step 5: Turn alerts on and off as you follow stories.

Change Detection to track new content uploaded on websites

- Step 1: Open www.changedetection.com/
 - Alternatively you can use Update Scanner:
<http://updatescanner.mozdev.org/en/>
- Step 2: Open a website that regularly (but not too frequently) uploads new data or reports
- Step 3: Copy the URL of that website into the search window of the change detection software
- Step 4: Receive alerts when new content is uploaded to the site

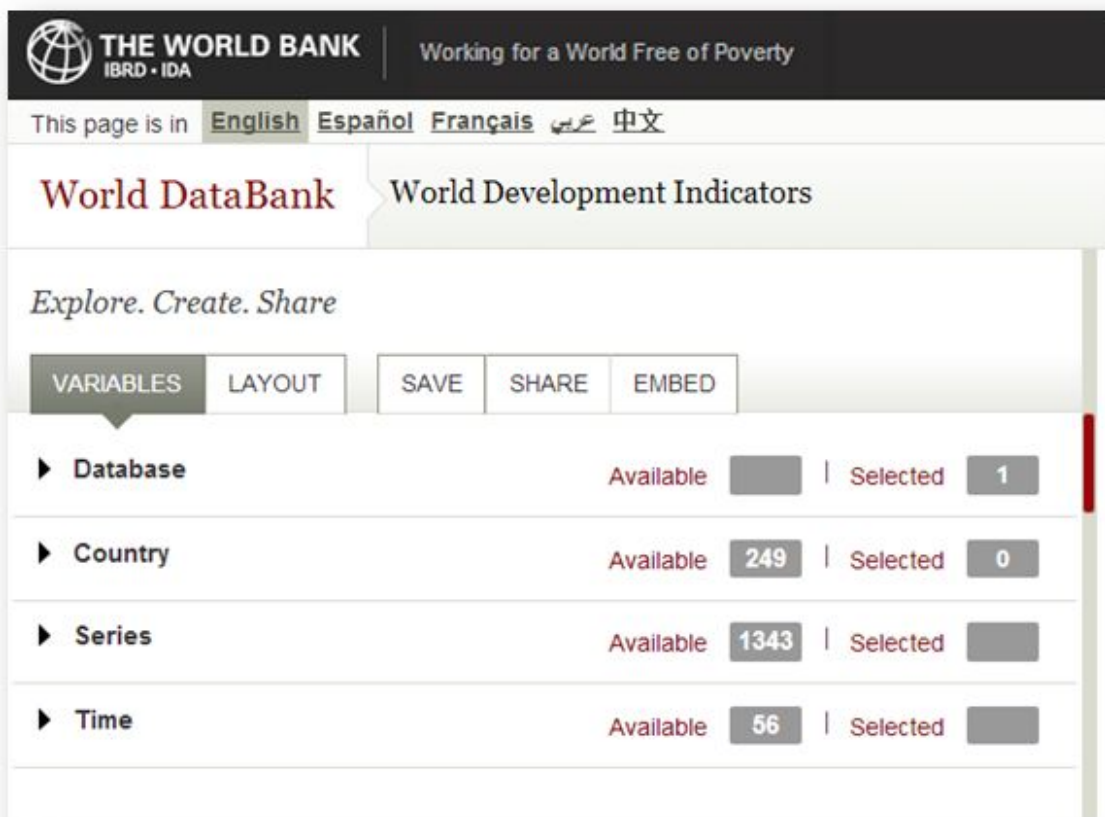
Advanced Google Searches: Scavenger Hunt!

Use Google Search to find:

- A PDF report on education in Tanzania.
- A PDF report on the UNICEF website about immunization in the Tanzania
- The 2016 national budget for Tanzania

- Tanzania's annual exports from www.tradingeconomics.com/
- An Excel file with data about migrants from Tanzania
- An estimated population projection from the national statistics website
- The inflation rate in Tanzania for the last 20 years
- News about tourism in Tanzania from the last month

Using Data Portals



The screenshot shows the World Bank World DataBank interface. At the top, it features the World Bank logo and the tagline "Working for a World Free of Poverty". Below this, there are language options: English, Español, Français, عربي, and 中文. The main heading is "World DataBank" with a sub-heading "World Development Indicators". The interface includes a navigation bar with "VARIABLES", "LAYOUT", "SAVE", "SHARE", and "EMBED" buttons. Below the navigation bar, there are four filter categories: Database, Country, Series, and Time. Each category has an "Available" count and a "Selected" count.

Category	Available	Selected
Database	Available	1
Country	249	0
Series	1343	
Time	56	

With a global push for open data many governments, international organizations are creating their own open data portals. These portals are a source of rich data and it's important to understand how to use a variety of interfaces to access and download desired data.

International, government, civil society and university databases are all fantastic sources of data. However, they all have their own interface that is a little bit different and require some exploration to understand how to navigate.

This is a general guide for how to navigate databases:

Select a database In many cases, a website will have many databases and the first step is to select which database you want to search. For example, on the World Bank Data portal, you can select to search only for health data, only for education data or development indicators, among many more options.

Select a geographical region There are many ways to compare how your geographical area compare to others. You can compare neighboring cities, states, or countries, regions with a similar level of economic development or population.

Select indicators Often databases will allow you to check boxes to identify which indicators you want to compare. It is best to select a wide range, look for interesting trends, and narrow down your focus later.

Select a time period There is a higher probability of finding enough data points to identify trends over a large span of time. In many cases, data will be collected in different countries in different years so it is best to start with a wide search and then narrow down the time period once you know what years have data points.

Select a format Often databases will allow you to see a table, map or visualization of the data. These can be useful overview

tools. What we are most interested in is downloading the data either in CSV or Excel format. Visualizations can be useful to identify patterns but generally we want to work with the raw dataset ourselves

National Databases

The screenshot shows the homepage of the Tanzania Open Data Portal. At the top, there is a blue navigation bar with the text "The United Republic of Tanzania - Government Open Data Portal" on the left and "Log In | Newsletter | FAQ | Kiswahili" on the right. Below this is a white header with the portal's logo "opendata.go.tz" and a navigation menu with links for "Home", "Datasets", "Sectors", "Organizations", "Visualization Gallery", "Suggestions", and "About". A search icon is also present. The main content area features a large graphic with a circular flow of icons representing various actions: SUGGEST, FEEDBACK, CONNECT, SHARE, DOWNLOAD, and VISUALISE. To the right of this graphic, the text reads "OPEN DATA PORTAL" and "There are different ways that the portal can be used: Download data, Visualise data, Suggest data, Offer feedback, Connect, Share, etc...". Below this, there are four categories of data: Education (759 Resources), Health (167 Datasets), National Statistics (11 Organizations), and Water (4 Sectors). Each category includes an icon and a brief description of the data source.

Tanzania Open Data Portal: <http://opendata.go.tz/>

The screenshot shows the Tanzania National Bureau of Statistics (NBS) Central Data Catalog. The header includes the NBS logo, the text 'TANZANIA NATIONAL BUREAU OF STATISTICS' and 'STATISTICS FOR DEVELOPMENT', and a 'LOGIN' button. A navigation menu contains 'NBS Home', 'Data Catalog', 'Policies and Procedures', 'Acknowledgements', 'Citations', and 'Contact'. Below the header, there are social media icons for WhatsApp, Facebook, and Twitter. The main content area is titled 'Central Data Catalog' and has tabs for 'COLLECTIONS', 'DATASETS', and 'CITATIONS'. It displays 'Found 37 studies out of 37'. On the left, there are filter sections: 'SEARCH BY KEYWORD' with input fields for 'in study description' and 'in variable description'; 'FILTER BY YEAR' with a range from 1988 to 2016; 'FILTER BY DATA ACCESS' with options for 'All', 'Public use data files', and 'Data not available'; 'FILTER BY COLLECTION' with an 'All' option; and 'FILTER BY TOPIC' with an 'All' option. The search results list four studies, each with a green circular icon, title, year, author information, collection name, and creation/modification dates.

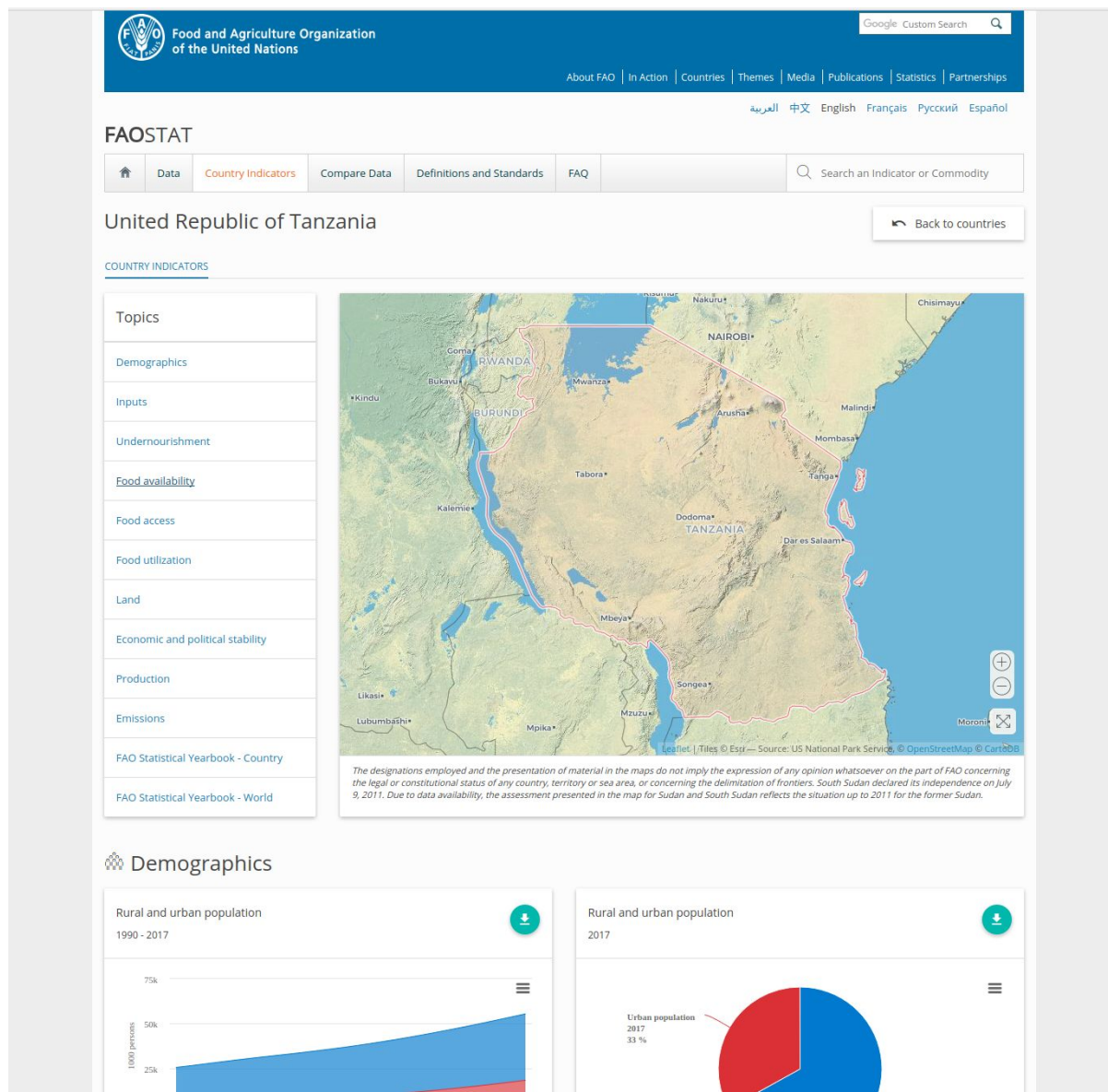
Tanzania National Bureau of Statistics Data Portal:

<http://nbs.go.tz/catalog/index.php/catalog>

There are several places to access national data portals:

- http://unstats.un.org/unsd/methods/inter-natlinks/sd_natstat.asp
- <https://www.opendatasoft.com/a-comprehensive-list-of-all-open-data-portals-around-the-world/>
- <https://investigativedashboard.org/>

International Databases



The screenshot displays the FAO STAT website interface for the United Republic of Tanzania. At the top, the FAO logo and name are visible, along with a search bar and navigation links. The main navigation bar includes 'Data', 'Country Indicators', 'Compare Data', 'Definitions and Standards', and 'FAQ'. A search bar is also present with the text 'Search an Indicator or Commodity'. The page title is 'United Republic of Tanzania', and there is a 'Back to countries' button. The 'COUNTRY INDICATORS' section is active, showing a list of topics on the left sidebar, including Demographics, Inputs, Undernourishment, Food availability, Food access, Food utilization, Land, Economic and political stability, Production, Emissions, and FAO Statistical Yearbook - Country/World. The main content area features a map of Tanzania with various cities labeled. Below the map is a disclaimer. The 'Demographics' section contains two charts: a stacked area chart for 'Rural and urban population 1990 - 2017' and a pie chart for 'Rural and urban population 2017' showing that 33% of the population is urban.

United Republic of Tanzania FAO Stats Country Page

Other similar examples of global data portals include:

In addition to national databases, there are many international data sources:

- World Health Organization: <http://www.who.int/gho/en/>
- United Nations: <http://data.un.org/>

- Population Reference Bureau: <http://www.prb.org/>
- UNICEF Data: <https://data.unicef.org/>
- Google's public data directory:
<https://www.google.com/publicdata/directory>
- The Datahub: <http://datahub.io/>
- DBPedia Datasets: <http://wiki.dbpedia.org/Datasets>
- Factual: <http://www.factual.com/>
- Free GIS data: <http://freegisdata.rtwilson.com/>
- List of open data repositories:
http://oad.simmons.edu/oadwiki/Data_repositories
- Open Energy Information: https://openei.org/wiki/Main_Page
- Extractive Industry Transparency Initiative (EITI):
<https://eiti.org/explore-data-portal>
- World Research Institute: <http://www.wri.org/>
- Quora thread: "Where can I find large datasets open to the public?":
<https://www.quora.com/Data/Where-can-I-find-large-datasets-open-to-the-public>
- Directory of APIs: <https://www.programmableweb.com/apis/directory>
- Infochimps: <http://www.infochimps.com/>
- Offshore Leaks: <https://offshoreleaks.icij.org/>
- Open Corporates: <https://opencorporates.com/>
- Transparency International Corruption Index:
<https://www.transparency.org/research/cpi/overview>
- Land Ownership Database: <http://www.landmatrix.org/en/>

- Gapminder World: <https://www.gapminder.org/data/>
- Global Data Lab: <https://globaldatalab.org/>

Navigating International Databases

Try this example to download data about **Tanzania** and its neighboring countries from an international database:

1. Open <http://databank.worldbank.org/>
2. Under 'EXPLORE DATABASES', select **Health Nutrition and Population Statistics**. The Health Nutrition and Population Statistics screen opens.
3. Under **Country**, select:
 - Burundi
 - Congo, Dem. Rep.
 - Kenya
 - Malawi
 - Mozambique
 - Rwanda
 - Tanzania
 - Uganda
 - Zambia
4. Scroll down, and click **Series**. The available indicators are listed.
5. Click on the **Filter** icon and select **Medical Resources and Usage**.
6. Check the boxes for **Nurses and midwives (per 1,000 people)** and **Physicians (per 1,000 people)**.

7. Now, click **Years**. The available years are listed.
8. Select the years of your interest, say the last 15 years. Click on **Apply Changes** on the right side of the page.
9. Then click on **Table** on the top right corner when your selection is ready.
You can always click on the menu on the right to change selections
10. Click the **Download Options** button, and download your data as an Excel file.
11. Open your data file in Excel.

Making Requests for Government Data

Article 19 of the Universal Declaration of Human Rights states that everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.

Many countries that have access to information laws lack rigorous regulations and procedures to respond to data and information requests. With the law still very new in many countries, it's essential that data users including local government professionals actively submit requests to ensure that these procedures and regulations are developed and pave the way for data sharing systems between government and citizens.

To find out more information about freedom of information in general visit

<http://foiadvocates.net/>.

For more information about Tanzania's freedom of information law check:

<http://www.freedominfo.org/regions/africa/tanzania>

These are many of the excuses that you will get for denying access to information requests. Remember, you have to be specific in your requests and persistent in order to get the data you need.

- “We don't have that data on a computer.”
- High fees
- Delay tactics
- “Your request was unclear.”
- Sending the wrong data
- “Our database is too complicated to give you access.”
- “Our database software is proprietary.”
- “That information is protected by privacy law.”

Lesson 3: Getting Data

In the ideal scenario, once you have been able to find the data you are interested in, all it should take is a click of a button to get that data to continue your data

work. However this is not the reality. There are several instances whether data is available on some platform or through some medium but still not in a format that can be obtained seamlessly. For instance, the Tanzania Ministry of Agriculture produces very useful reports on different aspects of the agricultural industry. However, many of these documents are in PDF format which is difficult if not impossible to analysing or process with common software. On other occasions, a website may have useful data in an HTML table that will be cumbersome or impossible to copy and paste into the relevant tool. This lesson will explore a few techniques to help you get this data.

Extracting data from PDFs using Tabula

Do you want a document that is secured, difficult to edit, easily compressed and portable? If so, please stick with PDFs. But at a time when the world is moving toward collaborative practices with data at the core of this movement, how can we keep it locked in a Portable Document Format, difficult to reach? If you have faced the challenge of having to extract data from a pdf, this tutorial is for you. If you haven't, read on anyway because if you want to keep up with the growing trend of data, extracting is something you should know.

Extracting data from PDF

- **PDF to Word/Excel converters** which allow you to copy the information you need. But the result is often messy if there are tables in the pdf. Some free tools include **Excel Online**².

² <http://www.pdfstoexcelonline.com/>

- **OCR (Optical Character Recognition)** which “reads” the PDF and then copy its content in a different format, usually simple text. Quality varies between the OCR engines, and often the licences are not free. You could always go with the free and open source **Tesseract OCR**³, but it requires some programming know-how.
- **Programming**, with some libraries existing for Python (PDFMiner), Java (Tika, PDFBoc), and the command line (pdftotext, pdftohtml).
- **Crowdsourcing**, which is not specifically for PDF, but can be used when you have many documents to transcript.
- **Tabula** is specifically designed to get data out of PDF tables, which is often where the data you’re looking for lives.

The Tabula way

What is Tabula and how does it work?

Tabula is an offline software, available under MIT open-source license for Windows, Mac and Linux operating systems, that allows you upload a PDF file and extract a selection of rows and columns from any table it may contain.

Getting Tabula

Tabula is available for the 3 major operating systems. Download it for Windows⁴, Mac⁵ and Linux⁶. It works in a Java environment so you will have to download **Java runtime environment (JRE)** if you don’t already have it.

³ <https://github.com/tesseract-ocr>

⁴ <https://github.com/tabulapdf/tabula/releases/download/v1.0.1/tabula-win-1.0.1.zip>

⁵ <https://github.com/tabulapdf/tabula/releases/download/v1.0.1/tabula-mac-1.0.1.zip>

⁶ <https://github.com/tabulapdf/tabula/releases/download/v1.0.1/tabula-jar-1.0.1.zip>

Note: Tabula for Mac OS X comes with Java

Tips for installing

- Once the program is downloaded, you are halfway toward your first table extraction. Follow these steps to get Tabula set up and ready to go.
- Your downloaded file would be a zip file, so extract the folder within
- Go into the extracted folder and run the Tabula program in it
- It should automatically open in your browser (chrome, firefox, safari are all confirmed browsers that work)
- If it does not launch on you browser, use this URL –
`http://localhost:8080`
- You should now see the user interface of Tabula.

Extracting your table

Tabula is a pretty easy application to use once installed. This steps should see through the process:

1. **Upload your PDF file:** Run the application file in your extracted folder. Tabula should launch and show the interface in figure 1 below. click on the Browse button as highlighted on the image to select among your documents the PDF you want to extract from. Here is an [example pdf](#) that you could use. The uploaded file should show on the right hand side as shown in Figure 1.

Fork 168 Star 1,619



Tabula

Liberate data tables trapped inside PDF files

Upload a PDF

Auto-Detect Tables

Table auto-detection can be time-consuming, especially for large PDFs.

Browse...

Submit

Uploaded files

[sample data for scraping.pdf](#) (2015-09-03 10:31)

[sample data for scraping.pdf](#) (2015-09-01 14:55)

[Excel worksheet 2.pdf](#) (2015-09-01 12:41)

[Excel worksheet 2.pdf](#) (2015-09-01 12:40)

[Excel worksheet 2.pdf](#) (2015-09-01 12:39)

[NDDC projects as at April, 2013.pdf](#) (2015-08-11 10:18)

[Information as at 2013.pdf](#) (2015-08-11 10:12)

Tabula User Interface

1. **Viewing the PDF document for Extraction:** From the same screen seen in Figure 1, click on your uploaded file and you should get a view like Figure 2 below. Select the section of the table you want to extract, or select all if you are extracting the full table. Note: you can always adjust your selection.

Tabula is experimental software [Home](#) [About](#)

How to use: make a rectangular selection over a table on the PDF pages. That's it!

Hint: table headers are (still) problematic. Try to exclude it from your selection.

[Download All Data](#)

[Clear All Selections](#)

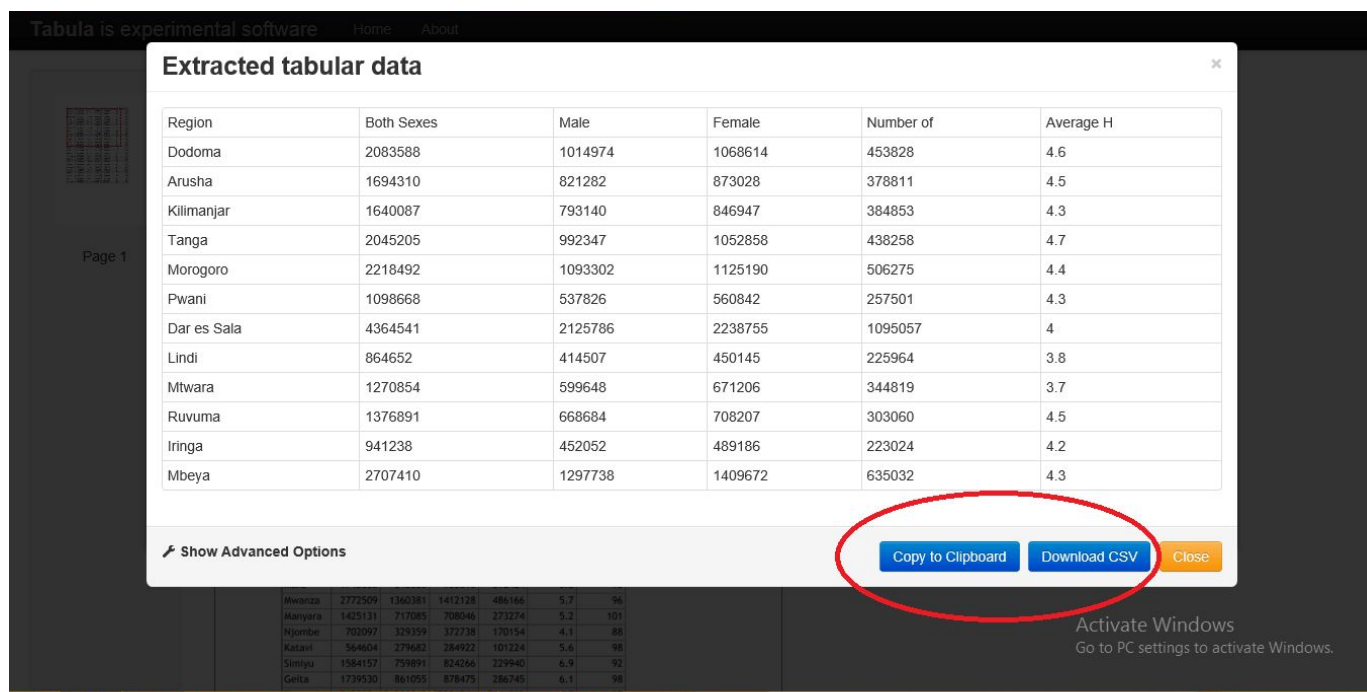
Preview Data Automatically? [Help](#)

Page 1

Region	Both Sexes	Male	Female	Number of	Average H	Sex Ratio
Dodoma	2083588	1014974	1068614	453828	4.6	95
Arusha	1694310	821282	873028	378811	4.5	94
Kilimanjar	1640087	793140	846947	384853	4.3	94
Tanga	2045205	992347	1052858	438258	4.7	94
Morogoro	2218492	1093302	1125190	506275	4.4	97
Pwani	1098668	537826	560842	257501	4.3	96
Dar es Sal	4364541	2125786	2238755	1095057	4	95
Lindi	864652	414507	450145	225964	3.8	92
Mtwara	1270854	599648	671206	344819	3.7	89
Ruvuma	1376891	668684	708207	303060	4.5	94
Iringa	941238	452052	489186	223024	4.2	92
Mbeya	2707410	1297738	1409672	635032	4.3	92
Singida	1370637	677995	692642	260441	5.3	98
Tabora	2291623	1129730	1161893	383419	6	97
Pwani	1004539	487311	517228	199764	5	94
Kigoma	2127930	1028994	1098936	374479	5.7	94
Shinyanga	1534808	750841	783967	261720	5.9	96
Kagera	2458023	1205683	1252340	524770	4.7	96
Mara	1743830	840020	903810	312424	5.6	93
Mwanza	2772509	1360381	1412128	486166	5.7	96
Manyara	1425131	717085	708046	273274	5.2	101
Njombe	702097	329359	372738	170154	4.1	88
Katavi	564604	279682	284922	101224	5.6	98
Simiyu	1584157	759891	824266	229940	6.9	92
Geita	1739530	861055	878475	286745	6.1	98

Selecting the table to extract

1. **Exporting the data:** Immediately after making your selection, your data should immediately show in a similar screen like Figure 3 below. You have an option to copy to clipboard and paste wherever you like or download your CSV file which can be opened in any spreadsheet application (Microsoft Excel, LibreOffice Calc, Google Spreadsheet...).. Simple and easy!



Exporting your data

The limits of Tabula

As great as Tabula is, it has some shortcomings.

- It does not work on Multi-lines rows or merged cells.
- Tabula cannot detect a scanned PDF document. it only works on text-based PDF
- Quickly pick one of those PDF files and see how the extraction goes.

For more information, see the references below.

Tabula and command line

If you are at ease with the command line, and would like to use Tabula on a batch of similar documents, then you could use the `tabula-extractor` library directly. All information about this can be found online here⁷.

Making data on the web useful: scraping

Many times data is not easily accessible – although it does exist. As much as we wish everything was available in CSV or the format of our choice – most data is published in different forms on the web. What if you want to use the data to combine it with other datasets and explore it independently?

Scraping to the rescue!

Scraping describes the method to extract data hidden in documents – such as Web Pages and PDFs and make it useable for further processing. It is among the most useful skills if you set out to investigate data – and most of the time it's not especially challenging. For the most simple ways of scraping you don't even need to know how to write code.

This example relies heavily on Google Chrome for the first part. Some things work well with other browsers, however we will be using one specific browser extension only available on Chrome. If you can't install Chrome, don't worry the principles remain similar.

⁷ <https://github.com/tabulapdf/tabula-extractor/wiki/Using-the-command-line-tabula-extractor-tool>

Code-free Scraping in 5 minutes using Google Spreadsheets & Google Chrome

Knowing the structure of a website is the first step towards extracting and using the data. Let's get our data into a spreadsheet – so we can use it further. An easy way to do this is provided by a special formula in Google Spreadsheets.

Save yourselves hours of time in copy-paste agony with the ImportHTML command in Google Spreadsheets. It really is magic!

Liberating HTML Data Tables

It's not uncommon to see small data sets published on the web using an HTML table element. If you have a quick click around Wikipedia, you're likely to find a wide variety of examples. Some sites will use Javascript libraries to enhance the presentation or usability of a table, for example, by making columns sortable; but most of the time, we are faced with a flat HTML table, and the data locked in it.

In this section, we look at some quick tricks for liberating data from HTML tables on public webpages and turning them into something more useful.

Screenscraping HTML Tables Using Google Spreadsheets

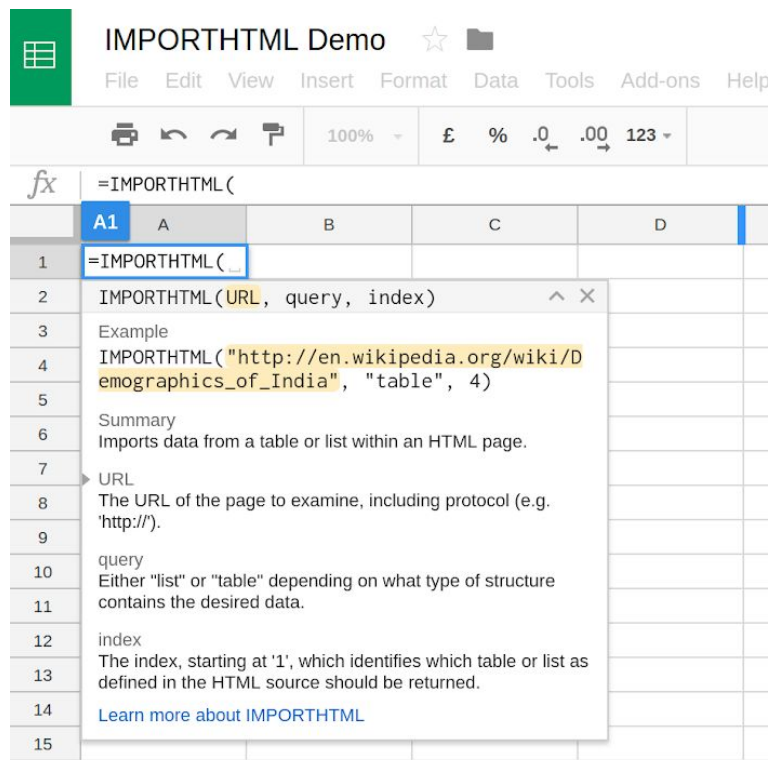
In order to follow this section of the lesson, you need to be familiar with Google Spreadsheets. If you are not, check out the Google Drive section in Lesson 2 of Module.

The Google Sheet formula:

`=IMPORTHTML(url,"table",index)`

will scrape a table from an HTML web page into a Google spreadsheet. The **url** of the target web page, and the table element both **need to be in double quotes**.

The index (which is a number) identifies the N'th table in the page (counting starts at 1) as the target table for data scraping.



So for example, have a look at the following Wikipedia page – [Regions of Tanzania](#) (found using a search on Wikipedia for regions of Tanzania):

Regions of Tanzania

From Wikipedia, the free encyclopedia

Tanzania is divided into thirty-one regions or *mkoa*.^[1]

Contents [hide]

- 1 History
- 2 List of regions
- 3 See also
- 4 Notes
- 5 References

History [edit]

In 1975 there were twenty-five regions in Tanzania. In 2002 one region changed its name, from Ziwa Magharibi Region (West Lake Region) to *Kagera Region*. In 2003 *Manyara Region* was created out of part of Arusha Region.^[1] In 2012 four more regions were created: Geita, Katavi, Njombe and Simiyu.^[2] In 2016 *Songwe Region* was created from the western part of *Mbeya Region*.^{[3][4]}



List of regions [edit]

Region	Capital	Districts	Area (km²)	Population (2012)	Postcode	Zone	Map
Arusha Region	Arusha	7	37,576	1,694,310	23xxx	Northern	
Dar es Salaam Region	Dar es Salaam	5	1,393	4,364,541	11xxx	Coastal	
Dodoma Region	Dodoma	7	41,311	2,083,588	41xxx	Central	

Grab the URL, fire up a new Google spreadsheet, and start to enter the formula

`=importHTML` into one of the cells:

The screenshot shows the Microsoft Excel interface. The title bar reads "IMPORTHTML Demo". The menu bar includes "File", "Edit", "View", "Insert", "Format", "Data", "Tools", "Add-ons", and "Help". The ribbon shows icons for printing, undo, redo, and paste, along with settings for zoom (100%), currency (£), percentage (%), decimal places (.0, .00), and font (Arial). The formula bar contains "=import". A dropdown menu is open, listing the following functions: IMPORTXML, Imports data from structured data online., IMPORTDATA, IMPORTFEED, IMPORTHTML, and IMPORTRANGE. The spreadsheet grid shows column headers A, B, C, D and row numbers 1 through 9. Cell A1 contains "=import".

Autocompletion works a treat, so finish off the expression and add in the URL and table number:

```
=importHTML("https://en.wikipedia.org/wiki/Regions_of_Tanzania", "table", 1)
```

The table numbers are not always obvious – start with 1 and increment the table number until you get the correct one.

The screenshot shows the Microsoft Excel interface with the formula bar containing the completed formula: "=IMPORTHTML("https://en.wikipedia.org/wiki/Regions_of_Tanzania", "table", 1)". The spreadsheet grid shows columns A through F and rows 1 and 2. Cell A1 contains the completed formula.

As if by magic, a data table appears in the spreadsheet, pulled in directly from the Wikipedia page:

	A	B	C	D	E	F	G	H	I
1	Region	Capital	Districts	Area (km2)	Population (2012)	Postcode	Zone	Map	
2	Arusha Region	Arusha		7	37,576	1,694,310	23xxx	Northern	
3	Dar es Salaam Region	Dar es Salaam		5	1,393	4,364,541	11xxx	Coastal	
4	Dodoma Region	Dodoma		7	41,311	2,083,588	41xxx	Central	
5	Geita Region	Geita		5	20,054	1,739,530	30xxx	Lake	
6	Iringa Region	Iringa		5	35,503	941,238	51xxx	Southern Highlands	
7	Kagera Region	Bukoba		8	25,265	2,458,023	35xxx	Lake	
8	Katavi Region	Mpanda		3	45,843	564,604	50xxx	Western	
9	Kigoma Region	Kigoma		8	37,040	2,127,930	47xxx	Western	
10	Kilimanjaro Region	Moshi		7	13,250	1,640,087	25xxx	Northern	
11	Lindi Region	Lindi		6	66,040	864,652	65xxx	Coastal	
12	Manyara Region	Babati		6	44,522	1,425,131	27xxx	Northern	
13	Mara Region	Musoma		7	21,760	1,743,830	31xxx	Lake	
14	Mbeya Region	Mbeya		7	35,954	2,707,410[a]	53xxx	Southern Highlands	
15	Morogoro Region	Morogoro		7	70,624	2,218,492	67xxx	Coastal	
16	Mtwara Region	Mtwara		7	16,710	1,270,854	63xxx	Coastal	
17	Mwanza Region	Mwanza		7	9,467	2,772,509	33xxx	Lake	
18	Njombe Region	Njombe		6	21,347	702,097	59xxx	Southern Highlands	
19	Pemba North Region	Wete		2	574	211,732	75xxx	Zanzibar	
20	Pemba South Region	Chake Chake		2	332	195,116	74xxx	Zanzibar	
21	Pwani Region	Kibaha		7	32,547	1,098,668	61xxx	Coastal	
22	Rukwa Region	Sumbawanga		4	22,792	1,004,539	55xxx	Southern Highlands	
23	Ruvuma Region	Songea		6	63,669	1,376,891	57xxx	Southern Highlands	

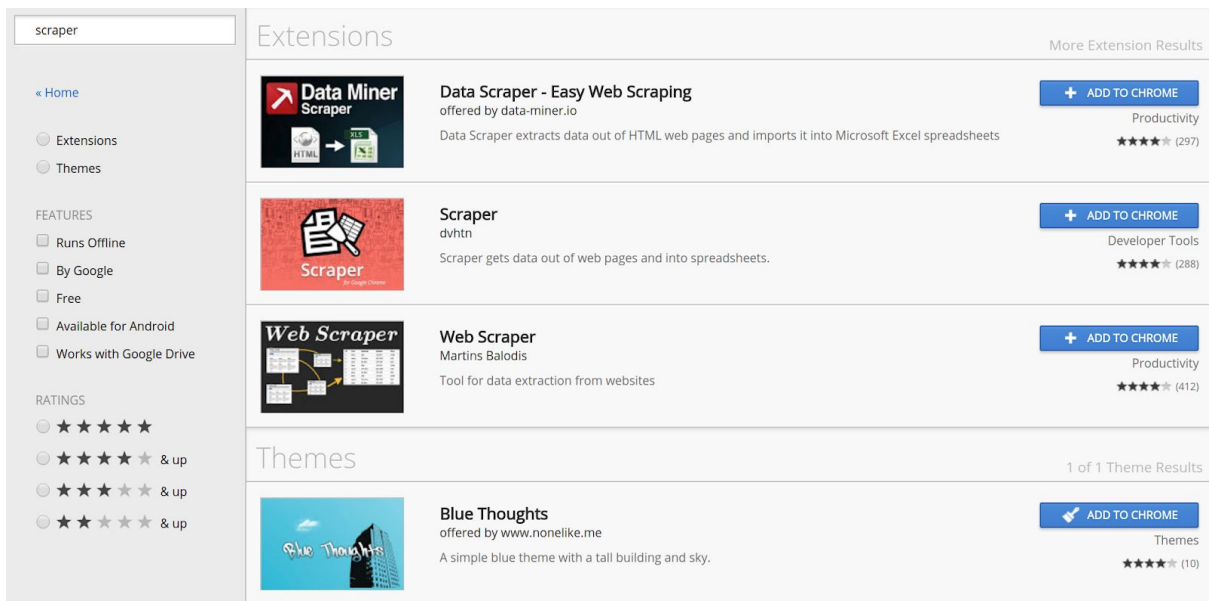
If the data in the HTML table is updated, the data in the spreadsheet will also be updated when you refresh or call the spreadsheet page.

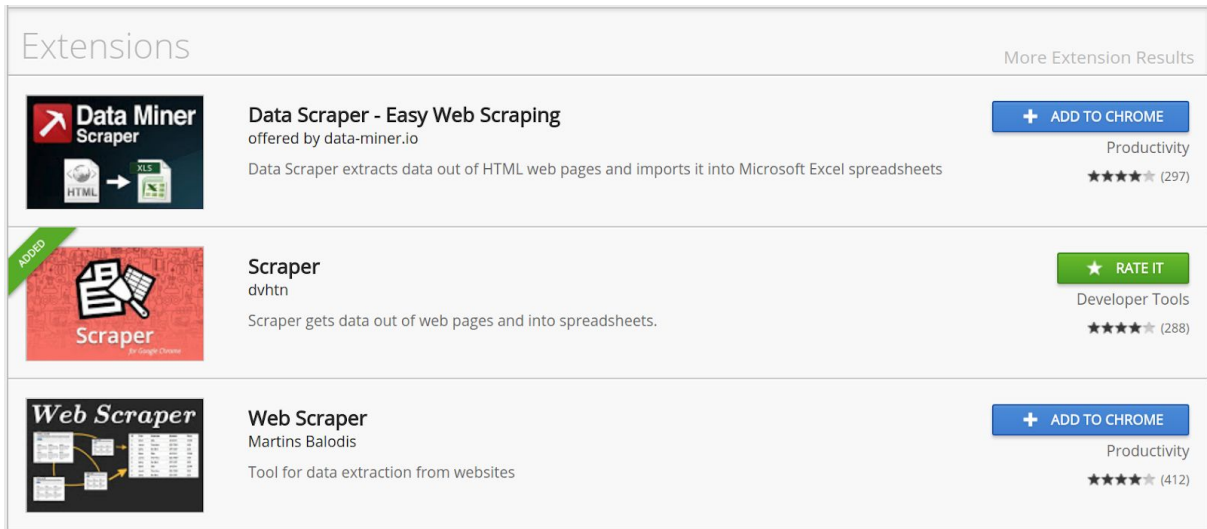
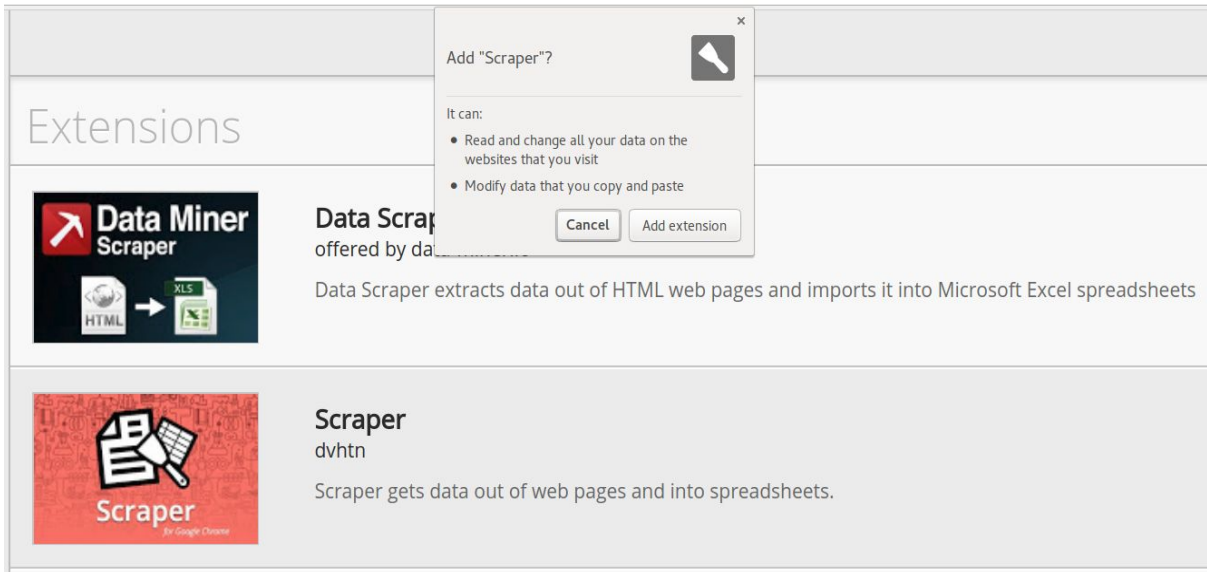
Scraping websites using the Scraper extension for Chrome

If you are using Google Chrome there is a browser extension for scraping web pages. It's called "Scraper" and it is easy to use. It will help you scrape a website's content and upload the results to google docs.

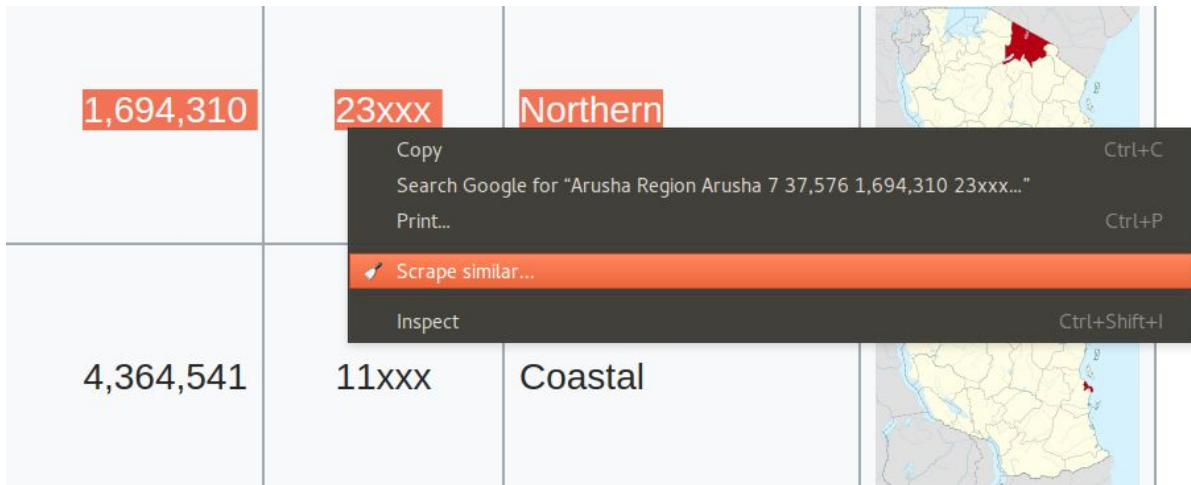
Walkthrough: Scraping a website with the Scraper extension

1. Open Google Chrome and go to the Chrome Web Store via the direct link (<https://chrome.google.com/webstore>) or by entering “Chrome Web Store” in your search bar to get the direct link.
2. Search for “Scraper” in extensions
3. The first search result is the “Scraper” extension
4. Click the Add to Chrome button.

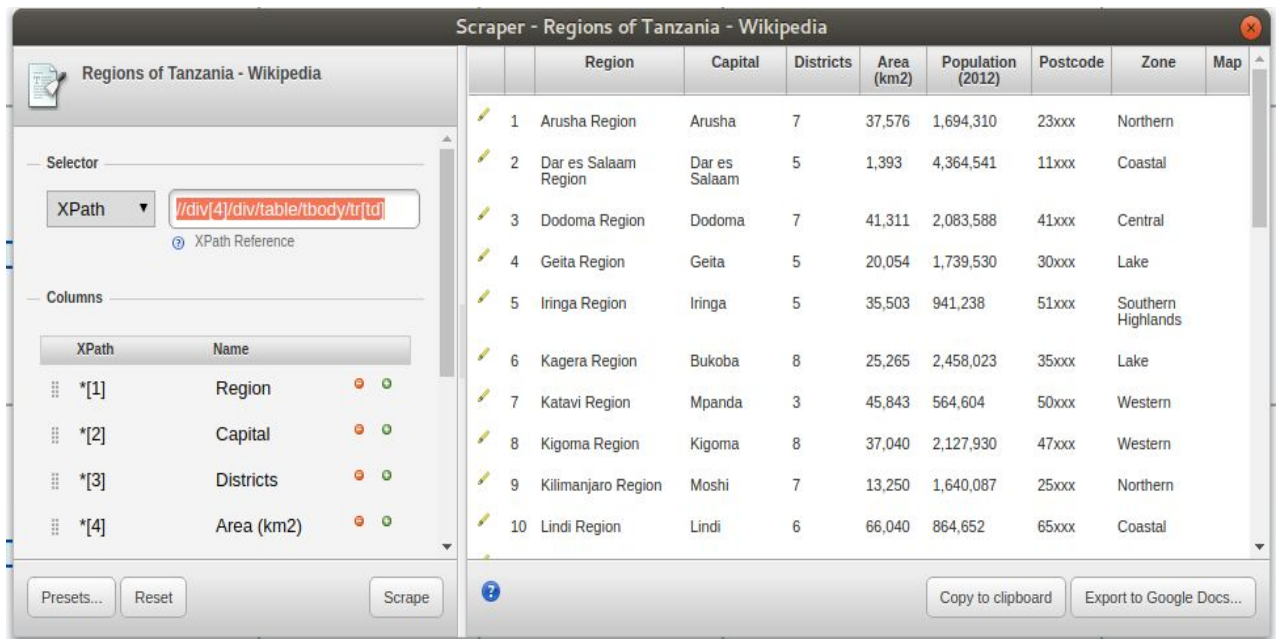




5. Now let's go back to the Regions of Tanzania Wikipedia page.
6. Open https://en.wikipedia.org/wiki/Regions_of_Tanzania
7. Now mark the entry or row for one region
8. Right click and select "scrape similar..."



9. A new window will appear – the scraper console



10. In the scraper console you will see the scraped content. If you do not see the table in the Scraper console. Refresh the Wikipedia page and repeat step 7.

11. Click on “Save to Google Docs...” to save the scraped content as a Google Spreadsheet.
12. You can also click on “Copy to clipboard” to paste the table in any other spreadsheet software.