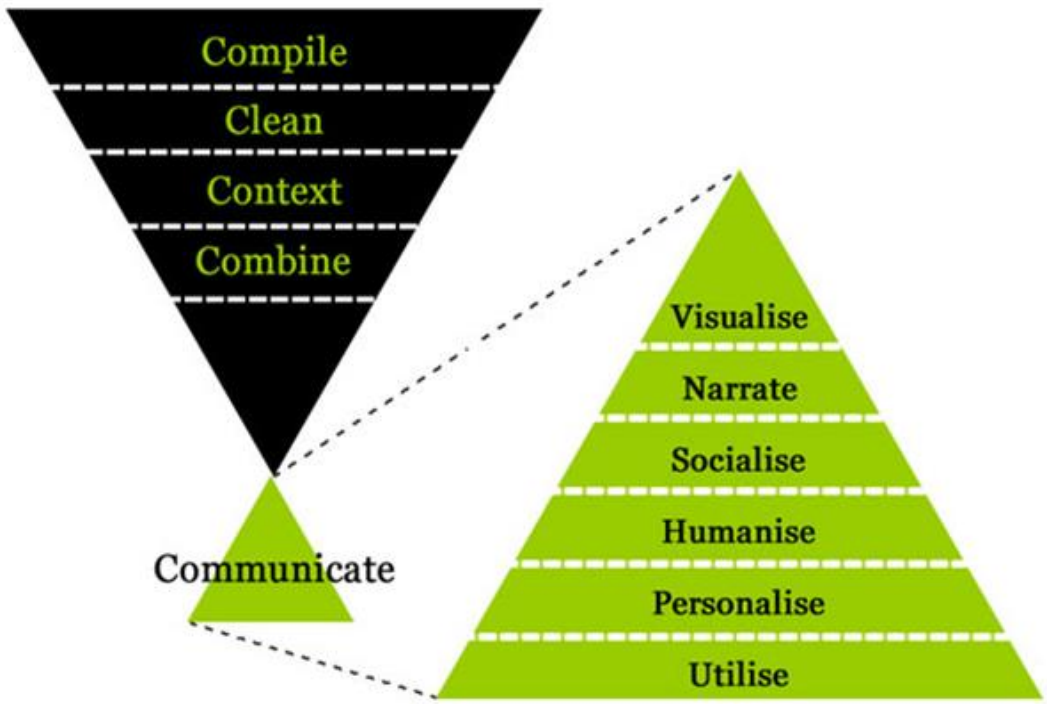
Module 2: Finding Data for Stories

STUDENT WORKBOOK

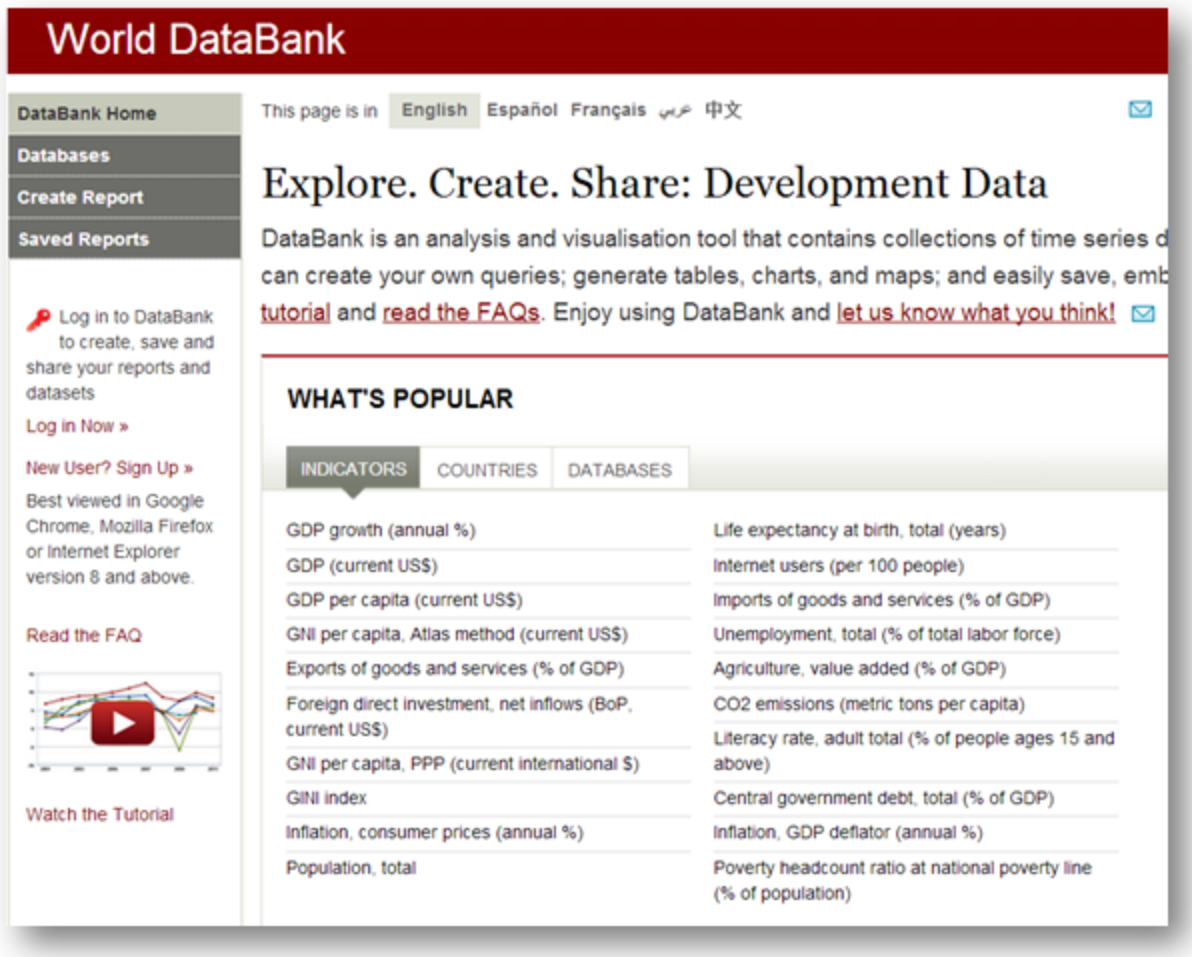
This module instills basic knowledge of data formats, the skills to find data online and the concepts to transform data into stories. Starting with a review of data formats, the unit moves on to Google search techniques to find different data types, introduces the formal process for designing a data project, evaluates the hypothesis of stories that use data and provide a chance for students to transform basic data fact sheets into stories. After completing this module, you will be able to:

- Identify basic data formats
- Search for data in different formats
- Automate data searches
- Transform fact sheets into simple stories
- Develop a hypothesis and questions for a story
- Practice evaluating the hypotheses of other stories

Contents



Getting Started



Public databases like the World Bank's DataBank have a wealth of data about more subjects than we can count or imagine: http://databank.worldbank.org/

But as you can see on this site, a lot of the jargon around data can be confusing. Looking at the first few categories, what do GDP growth (annual %); GDP (current US$); GDP per capita (current US$); GNI per capita, Atlas method (current US$; and Exports of goods and services (% of GDP) mean? We don't expect everyone who uses economic data in their reporting to be an economist.

Rather, this unit will help you start navigating your way around complex data by not only introducing some basic data formats and some questions that we should always ask of datasets, but also the resources we need to find out what is beyond our knowledge to help us understand what the data is measuring and how.

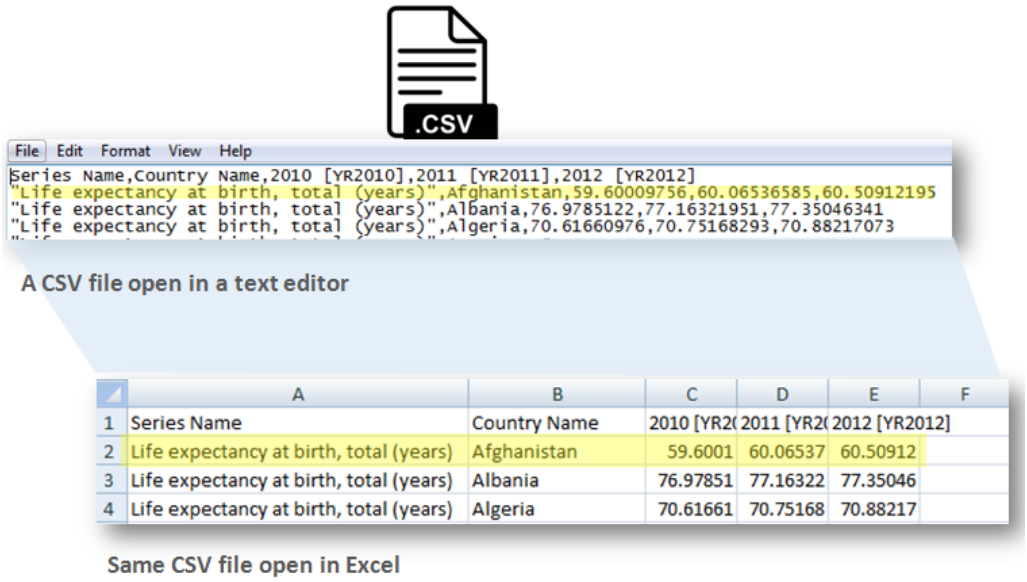
Lesson 1: Common Data Formats

Modern data analysis relies on software to do the heavy lifting involved in data analysis for us. We cannot work with data until we convert data into a format that the computer understands so that it can organize data into rows, columns, and cells.

Much of what stops citizens from using data, either intentionally or unintentionally, is that data is provided in formats that can't be immediately used or read by a computer. This lesson explains such data formats and the processes to transform them.

Data analysis, storytelling and visualization all depend on a computer program being able to read our data. Unfortunately, often data comes in formats that computers do not understand.

Data formats: Machine-readable, Computer Generated, Structured



In these data formats, computer software recognizes an explicit structure to the data - most commonly in a table - with columns and rows that organize and describe discrete data points. Excel and CSV are common examples.

- Excel file (XLS): data is saved as a table readable by Microsoft Excel
- Comma separated values (CSV): Plain text file with each data entry separated by a comma

These formats are typically the best suited for analysis, and you can easily work with them in a spreadsheet program - like Excel. When searching for data, if you can find Excel or CSV formats, this is a good sign that you won't have to spend a lot of time cleaning and formatting.

Note that CSV (comma-separated values) and TSV (tab-separated values) formats are formats for "encoding" tabular data. In simple words, CSV and TSV files are plain text files in which:

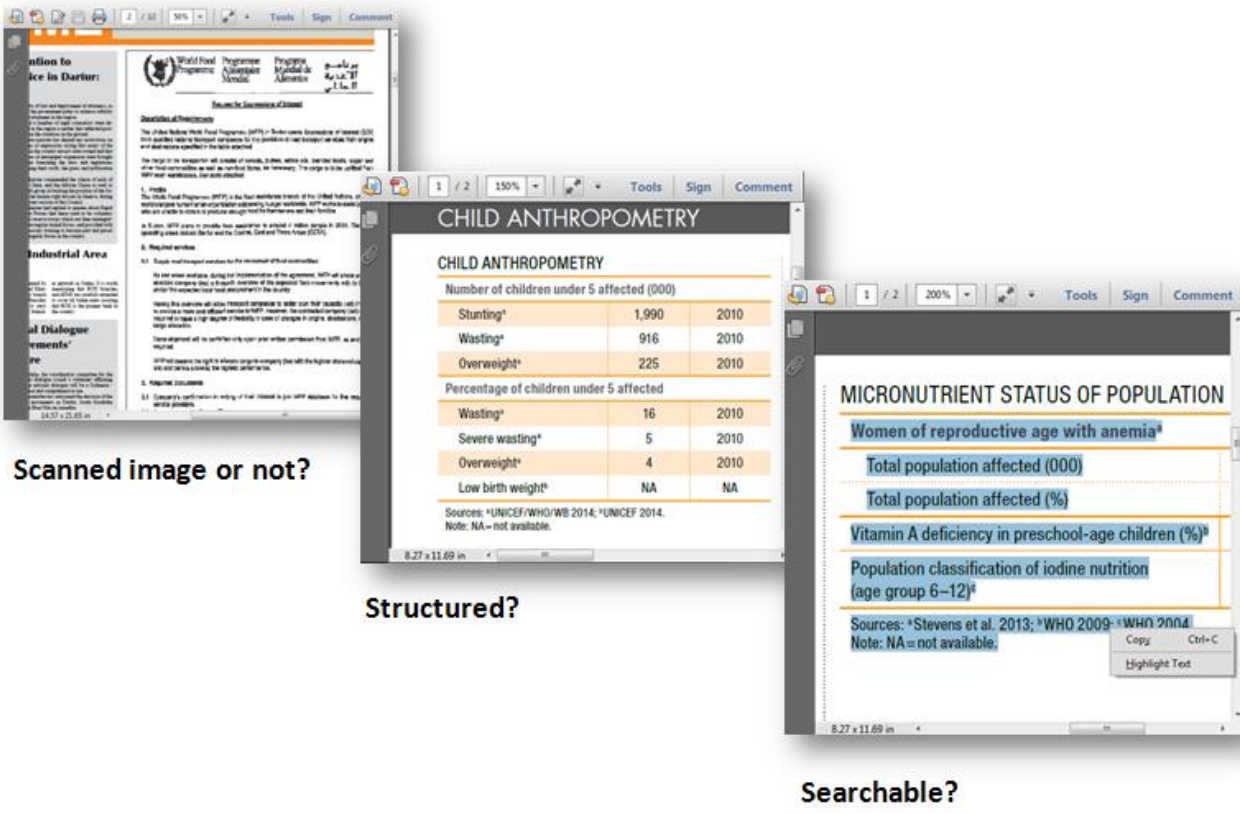
- Each line represents a row and
- Within each line, a comma (for CSV) or a tab character (for TSV) separates columns

Excel files also uses on a similar structure, but relies on Microsoft software.

Tools:

Google spreadsheets, Microsoft Excel are commonly available tools that help you work these formats.

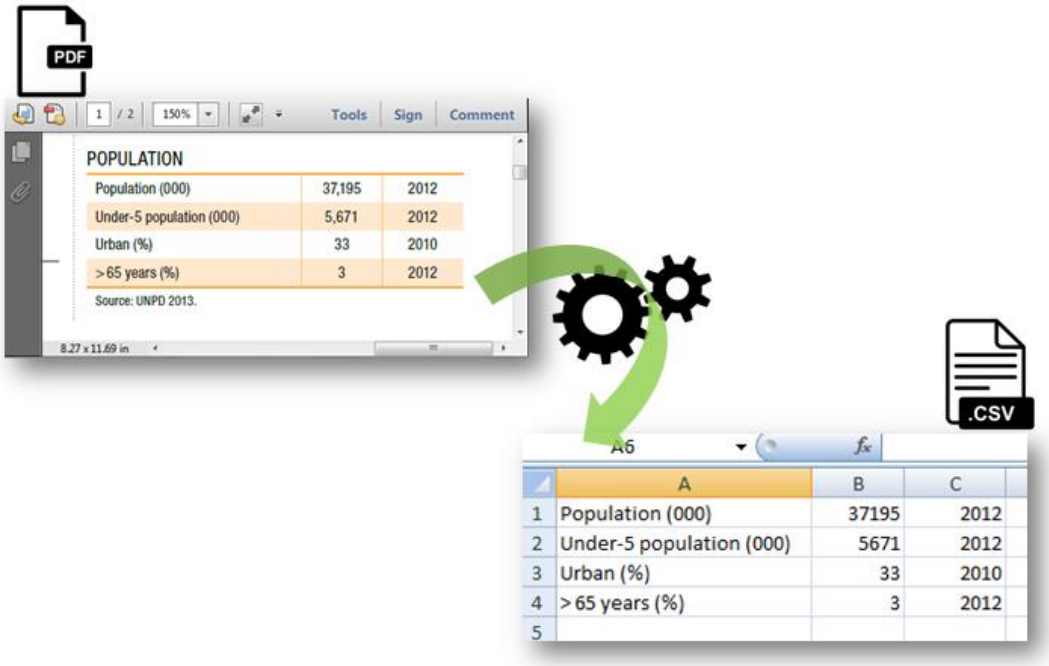
Portable Document Formats (PDF)



PDF files come in a few different varieties.

- The first question to ask is if they are computer-generated or not? That is, if a file was saved in a PDF format or if it was actually printed and scanned back in as an image not generated by software.
- The next question is if the data within the PDF is structured, as in, it's available in columns and rows published in a table.
- Finally if it is searchable - which has to do with whether it was generated by a computer. Basically searchable means that you can highlight the text and the computer recognizes the letters and numbers as characters.

From Document Formats to Machine Readable Data Formats



Typically, the best suited data formats for analysis are structured and machine readable – like CSV or Excel. When you find data in other formats, say a PDF, it's useful to convert it into a structured and machine readable format.

Data in PDFs

PDFs often contain structured, computer generated tables but a PDF is not a data format. The table has to be converted into a format that can be opened by a spreadsheet program. So these data tables require extraction into a data format through special software. You will practice extracting data in the Scraping lab.

Tools: Tabula, CometDoc, PDFtoExcel, Zamzar

Data in Scanned Images

These are primarily image files that are read as one giant block instead of discrete parts. These require Optical Character Recognition Software to recognize the text in the file. Usually, these used to be computer generated, but then someone printed the document and scanned it back into the computer, turning it into giant image file.

Examples: Some PDF and all bitmap images (GIF, JPEG, PNG, BMP)

Tools: Google Docs OCR, Document Cloud

Data in Unstructured Formats

Some data has been generated by a computer but does not have a structure recognized by machines. Examples of this include data that has been entered into a text document in paragraph format and some data on websites. Basically, in this case, a developer has to teach the computer what the pattern is in the data and then extract it into a data format.

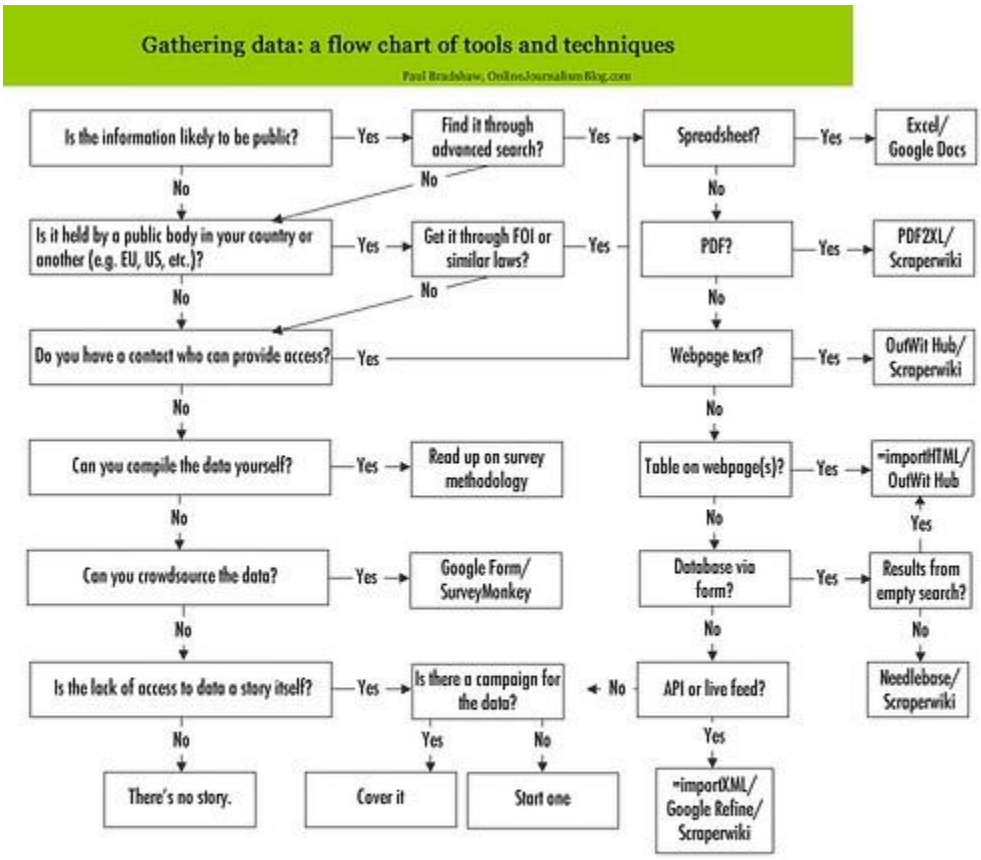
Tools: Python or Ruby programming languages to scrape data using https://morph.io/

Less Common Data formats

Some data, especially large databases, are saved in packages designed to be coded into websites or read by statistical software like Stata or R. These require conversion to CSV or Excel for use with spreadsheets software.

Examples: JSON (JavaScript Object Notation) or XML (extensible Markup Language) for programming and .SAV or .R. Try using https://konklone.io/json/ to convert JSON files to CSV.

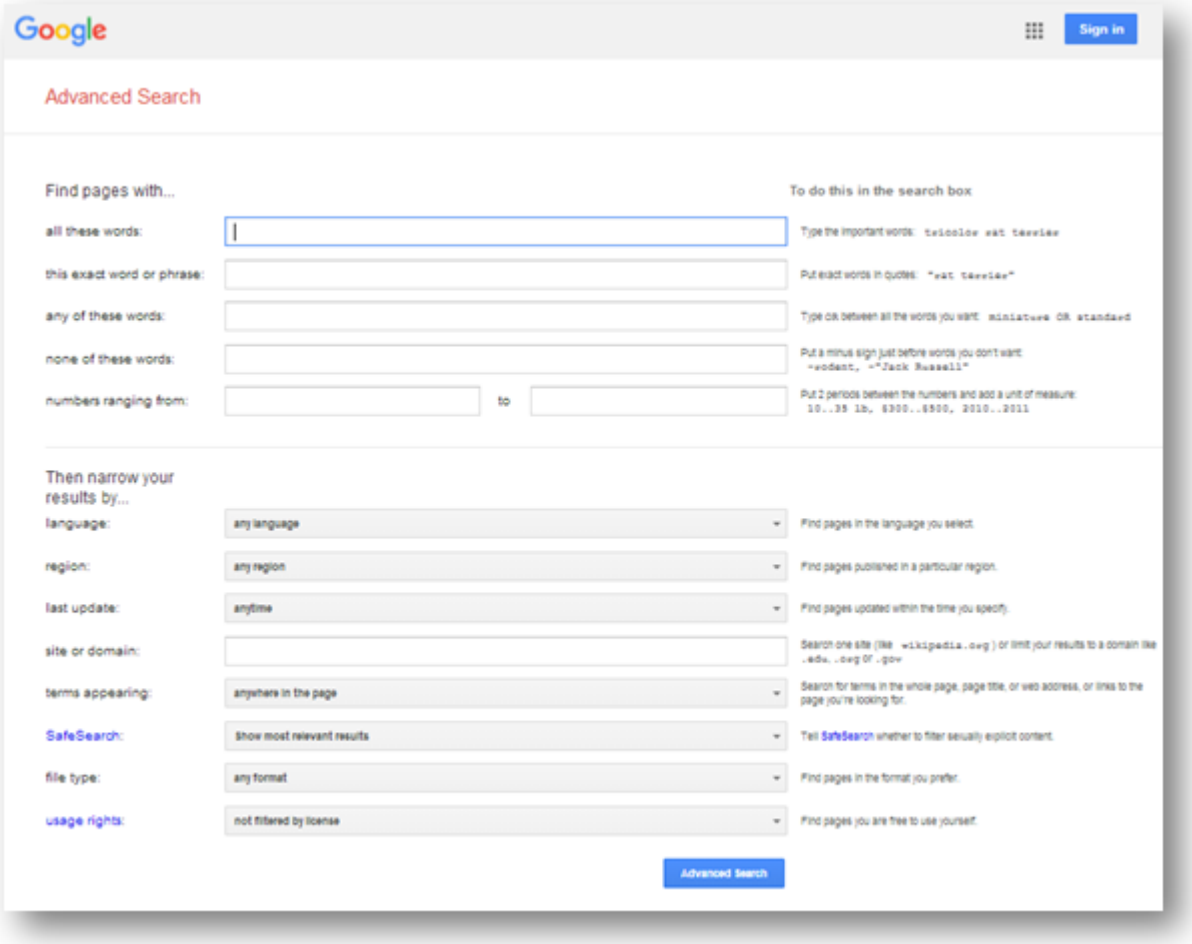
Lesson 2: Finding Data Online[1]



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

[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. This should be a reference chart for you when you start your own data search, hit a wall, and don't know what the next step is: http://onlinejournalismblog.com/2011/09/06/gathering-data-a-flow-chart-for-data-journalists-2/

Using 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.

GOOGLE 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 on 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. Pakistan health, for example.
Exact word or phrase helps find results in	Use quotes to search – for example "Ministry of National Health Services, Regulations and Coordination"

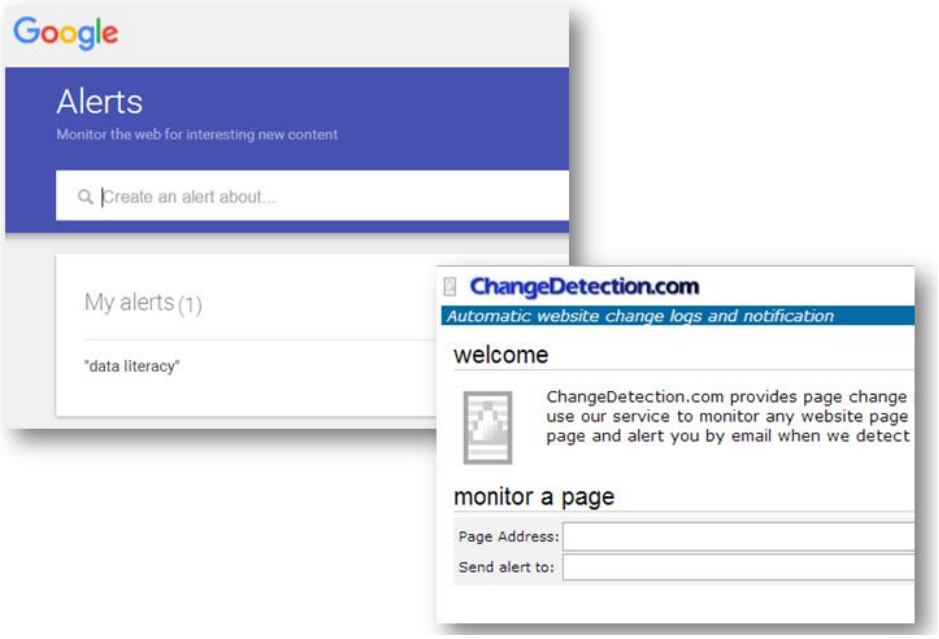
which the words appear in the exact order you mention	
Any of these words helps find results where any of the mentioned words appear	Use OR between words in a search – for example, *agriculture OR farming OR crops*
None of these words will filter out search results with words that you specify	Type the minus sign before the word you want to emit in a search – for example, Pakistan FATA -terrorism
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: *site:http://www.who.int/* Note that the website address has to be EXACT. • CORRECT *site: https://www.unodc.org/* • WRONG *site: WorldHealthOrganization.org*
Filetype: Search only for files with a specific extension (for instance: xls, pdf, csv, doc)	Use this search format - *Filetype:[extension]* For example, here is a search term to look for XLS files: • CORRECT filetype:xls • WRONG filetype:Excel

Now let's try using Google advanced search:

- Use '**any of these words'** to find content about malnutrition, hunger or starvation in your region.
- Use '**none of these words'** to find information about malnutrition not about children.
- Find content about polio only in French.
- Find content about polio only from Pakistani websites
- Find content about polio published in the last week.
- Search the Ministry of Health website for Excel files
- Search for PDFs about maternal health in your region.

.

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

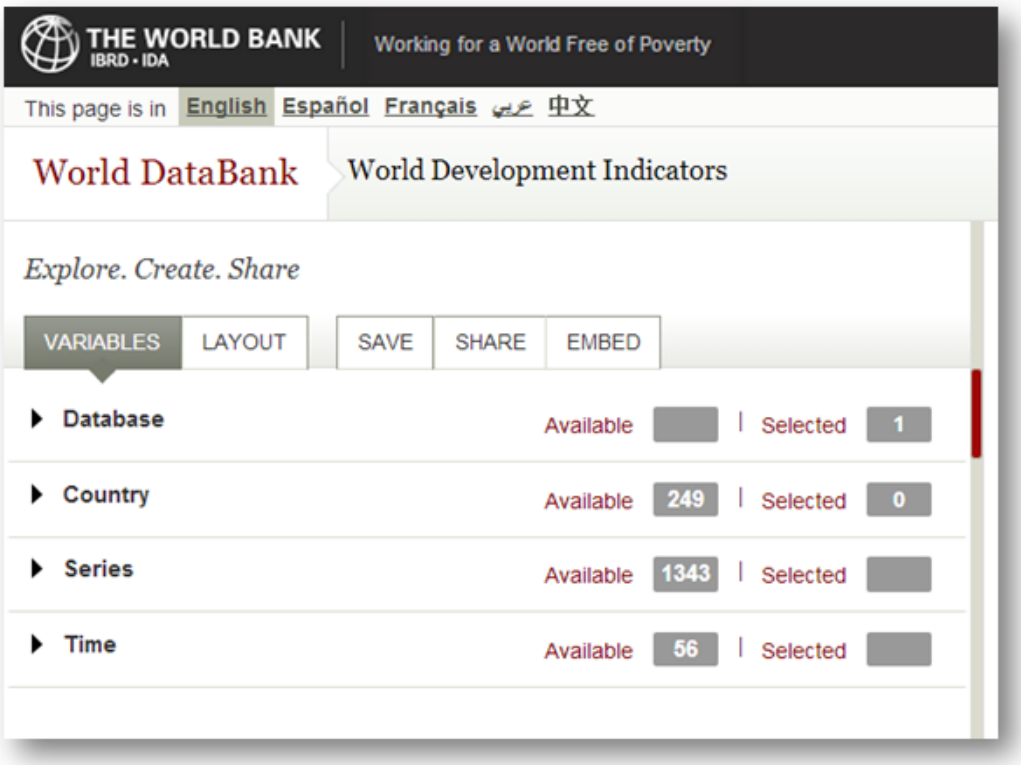
- **Step 1:** Open www.changedetection.com/
    - Alternatively you can use Update Scanner: http://updatescanner.mozdev.org/en/
- **Step 2:** Open a website that regularly (but not too frequently) uploads new data or reports
- **Step 3:** Copy the URL of that website into the search window of the change 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 of that includes the Number of patients treated (Indoor/Outdoor) in Government Hospitals and Dispensaries in KPK in 2014.
- The 2016 budget from your region
- Estimates of Foreign Assistance 2016-17 for Pakistan
- An estimated population projection from the national statistics website
- The GDP to tax revenue rate for the last five years
- Pakistan Social And Living Standards Measurement Survey (2014-15) in Excel format
- All the Pakistani people and companies listed in the Panama Papers
- The crop production in Pakistan by district
- Which Punjab district received the most funding for literacy on http://odi.itu.edu.pk/
- News about HIV in Pakistan from the last month

Using Data Portals



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 compares

	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

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/

Navigating National Databases

Here's an example of how you can use a national database to download data. In this example, you will download data to find out which region we are investigating have the highest rate of honour killings:

1. Open the HRCP database in your web browser http://hrcp-web.org/hrcpweb/campaigns/
2. Select **Honour Crimes (Women/Men).** The *dashboard* screen opens.
3. From the '**Incident:**' drop-down menu, select **Honour Killings**
4. Next, select **'Generate Report'** This will produce a PDF of summary statistics of the database.
5. Continue to the next the **Generate Report** option.
6. A PDF of data should open automatically.
7. What do you see that's interesting or surprising in the data?
8. Repeat the search for FATA and KPK. Choose more specific options once you see the general report.

Exercise: Navigating National Databases

1. Choose another category of human rights violations and answer the following questions:
    - In which year were the most cases recorded by the media?
    - What were the demographics of the victim?
    - What is a good news angle for a story about the issue?
    - How would you explain the source of the data?
2. What other national databases do you know about?

International Databases

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

- World Health Organization
- United Nations
- Population Reference Bureau
- UNICEF Data
- The Guardian's world government data portal
- Google's public data directory
- The data hub
- DBPedia Datasets
- Factual
- Free GIS data
- List of open data resources
- Energy data repositories
- World Research Institute
- Data wrangling
- Quora thread: "Where can I find large datasets open to the public?"
- Directory of APIs
- Infochimps
- Datamarket
- Offshore Leaks
- Investigative Dashboard
- Open Corporates
- Natural Earth data
- UNEP Data
- Transparency International Corruption Index

Land Ownership Database

Gapminder World

Global Data Lab

Navigating International Databases

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

1. Open http://databank.worldbank.org/
2. Under 'EXPLORE DATABASES', select **World Development Indicators.** The search window opens.
3. Under 'Country', select:
    - Afghanistan
    - India
    - Pakistan
    - Bangladesh
    - Sri Lanka
    - Nepal
    - Bhutan
    - Maldives
4. Next, click **APPLY CHANGES**.
5. Scroll down, and click **Series**. The available indicators are listed. Filter to "Health Indicators" then "Health Services"
6. Select the indicators of your interest. You can click on the information icon to read the long definition of the indicator. In this case we want:
    - Health expenditure, public (% of total health expenditure)
    - External resources for health (% of total expenditure on health)
    - Out-of-pocket health expenditure (% of total expenditure on health)

    and click **APPLY CHANGES**.
7. Now, click **Years**. The available years are listed.
8. Select the last 15 years, and click **APPLY CHANGES**.

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 or CSV file.

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 journalists 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 Pakistan visit Article 19's country report on Pakistan.

These are many of the excuses that you will get for denying access to information requests. Remember, they are just excuses! 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."

Exercise

Under each of these sectors, please find and download a recent, relevant dataset (either Excel, CSV or a table in a PDF) for your region using either google searches or international or national databases. Provide the URL for where you found it.

1. Governance and Security
2. Economic Indicators and Budgets
3. Education
4. Health
5. Human Rights
6. Environment

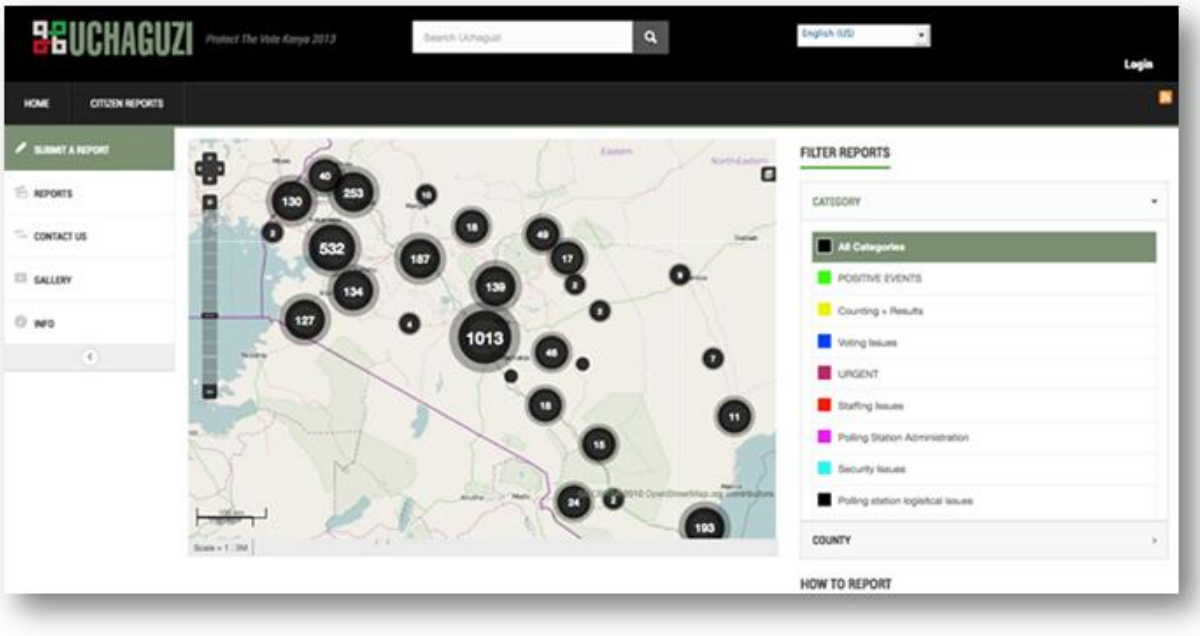
Lesson 3: Alternative Data Sources

Often, when official data source are not available, organizations may use “crowd-sourcing” to solicit data from citizens or from a trained network of volunteers. They may also use sensors, citizen reports, media reports or leaked data as alternative data sources.

For example, take a look at the Afghan Election Violence Map: http://www.tfp.nai.org.af/map/main

In this map, election monitors in Afghanistan sent SMS alerts documenting violence and irregularities on election day. The monitors were trained in how to use the system. The data was used by journalists to report on the status of polling stations across the country.

Crowd-sourced data

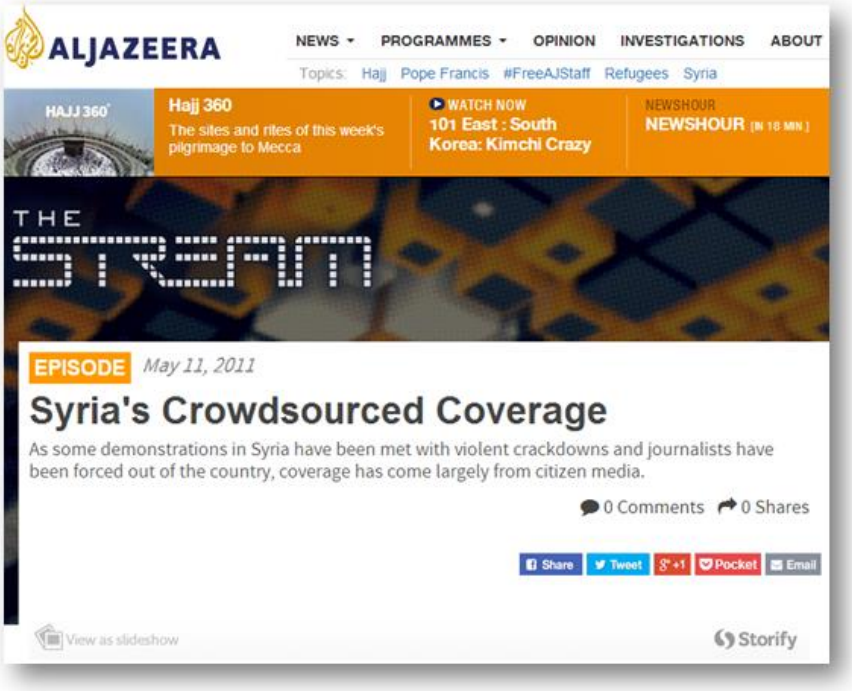


One of the most well-known platforms for collecting 'crowd-sourced' data is Ushahidi. This platform has been deployed to map natural disasters, political crises, and other events where live data collection can inform a response. This system enables an open or closed network to submit reports of incidents (such as violence, ballot box tampering, police harassment), which is sent to a centralized system for verification, addition to the database, and mapping.

For example, a website called Uchaguzi[2] was set up using Ushahidi for Kenyan Election Monitoring. In this example, media houses tapped into a centralized network of election monitors who were tracking and categorizing election incidents.

[2]https://www.facebook.com/ushahidi/photos/a.193585313994844.42244.116038145082895/543807175639321/?type=1&theater

Citizen Media Contributions

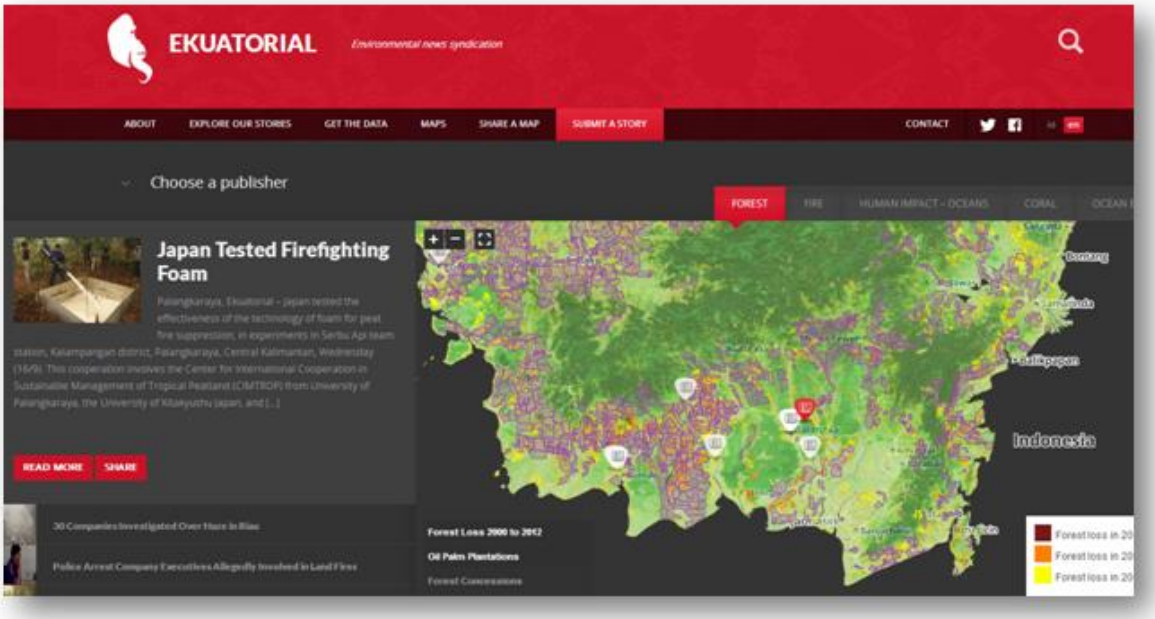


In many cases, media houses engage their audience to send in their own raw material or eye witness reports via SMS, video, or photos – referred to as citizen media contributions.

For instance, Al Jazeera[3] is one of many media who have resorted to social media and citizen media accounts to report on the situation within Syria. Most media houses have verification policies in place but they are not immune to publishing false content.

[3] http://www.stream.aljazeera.com/story/201105112039-003652

Sensors



Data collected by sensor data has been used for reporting on environmental issues. For instance, a media house may distribute small, inexpensive sensors to a trained community of volunteers to gather data such as air quality, water temperature, or earthquake activity.

For example, Ekuatorial[4] is an attempt to collect data about the Indonesia rainforest, an area that is remote and difficult to monitor. Sensors have been left with members of communities of these remote areas to collect ground level data on environmental conditions which are combined with national and satellite data to track environmental degradation.

[4] http://ekuatorial.com/en

Drones

Increasingly, media are using unmanned aircraft to estimate the size of protests, measure garbage dumps, calculate the rate of deforestation, and gather other useful information from an aerial view.

For example the Wall Street Journal shared videos taken by drones to demonstrate the scale of protests in Hong Kong: http://www.wsj.com/video/aerial-drone-captures-scale-of-hong-kong-protests/76AA792E-7AB9-4D2B-88BB-E9B5F9D707EC.html

Though the technology is available, privacy issues and creating the effect that protesters are being spied on by drones that could be from government can cause suspicion.

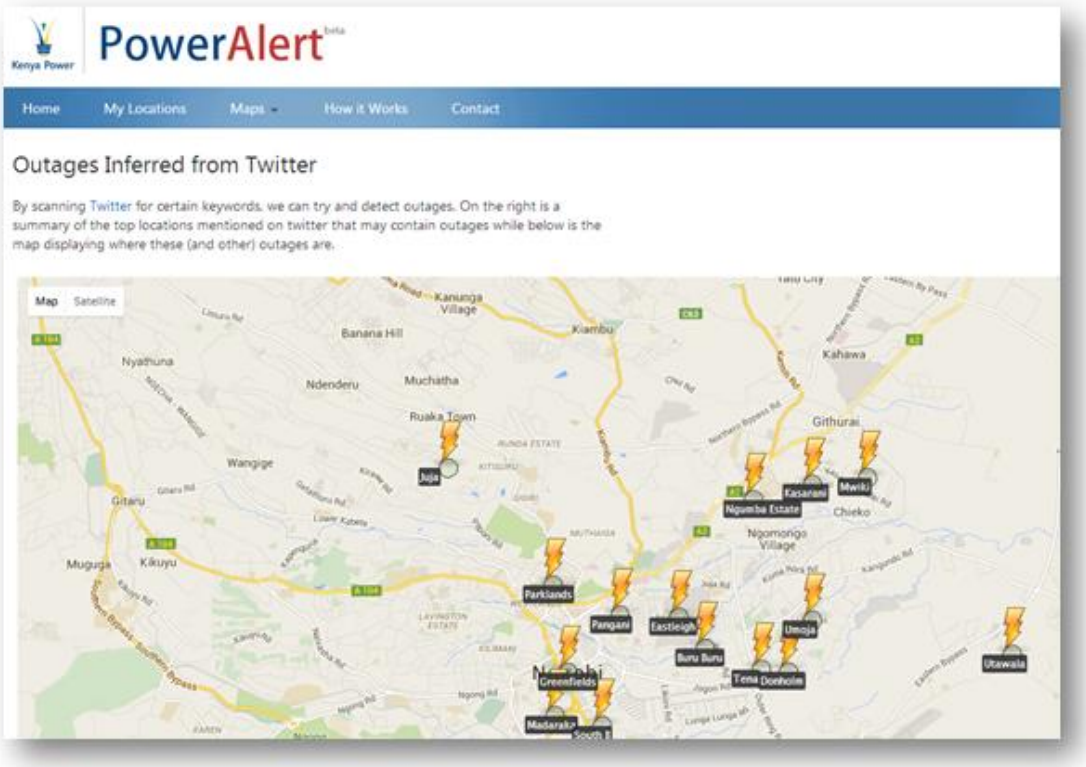
Mining Newsroom Data

For subjects with poor official documentation, crowd-sourcing media reports on the topic can often yield a rich store of data. This strategy has been used to collect data on violence against women, people killed in police shootings, and Chinese aid to Africa.

For example 'The Migrant Files'[5] sources reports of migrant deaths in the Mediterranean from global media reports. When combined, these reports paint a much more complete picture of migrants' deaths in their journey to Europe.

[5] http://www.themigrantsfiles.com/

Risks



There is an inherent problem with using crowd-sourcing to measure public service delivery for several reasons: those with poor access to public services usually also have poor access to the telecommunications needed to report on the issue.

- **Selection bias:** only people with time, resources and motivation are likely to contribute
- **Verification process:** people could flood the system with erroneous reports and it is difficult to find out which are real
- **Context:** from crowd sourced data, we only know what the crowd tells us so much of the contextual information that we would usually include to explain data, is lost.
- **Privacy:** sometimes personal data about contributers can be accessed by agencies that might want to target critics

For example, in this power outage reporting system in Nairobi[6], slum areas - with poor access to electricity in the first place - are under-reporting power outages.

[6] http://poweralerts.kenyapower.co.ke/tweetmap

Exercise

Identify one newsworthy issue that would benefit from alternative data sources and identify what data should be collected.

Lesson 4: Planning a Data Story

The prevalence of data in the modern world has changed the way humans receive information. Now, more insight and solutions can be developed by enriching traditional information channels with data. This lesson will review how journalists organize data projects to ensure a successful story.

When the Sun Sets in Turkana

Take a look at this news story: https://www.youtube.com/watch?v=Ga8CEYVALo4

http://www.internewskenya.org/summaries/internews52e7747b74fff.pdf

In this example, the journalist investigates a news story: the drought in Turkana and widespread starvation, with a data lens:

- She looks at **climate data** to determine that droughts are increasing in severity and frequency.

From Evidence to Stories: Thinking Like a Data Journalist

- She uses **health data** to determine the health impact of malnutrition on children.
- She uses **international aid data** to determine if there is a long-term solution to the problem: investing in food security instead of humanitarian aid.

The key to success for any data journalist is organization. Unlike in many other kinds of journalism, how you decide to organize your information and narrate your story can make or break your story. The process we will follow for organizing a data journalism story consists of six steps:

1. Background
2. Hypothesis and Questions
3. Analysis
4. Interviews
5. Story Organization
6. Visualization

As we proceed through this course, we will cover the tasks and skills required for each step. In this lesson, we will cover the two planning stages: background and hypothesis and questions.

Background

When first identifying a topic for a data story, the first step is to search for other data stories produced by other journalists on the same topic. This serves several purposes. It familiarizes you with how other journalists have approached the issue, where and what kind of data he or she used and what storytelling strategy was effective.

Cases of violence against women: Is mediation the best option?

In this story, a team of journalists decided to cover the issues of domestic violence services in Afghanistan.

Case studies

Using advanced Google search techniques, they identified three similar data stories in the media:

India is a Nation of Violent, Stressed Men, IndiaSpend, India

Till Death Do Us Part, Post and Courier, USA

Most Dangerous Transport System for Women, Global Post, Global

From these examples the journalists noted:

- The government's ability to provide services is key in determining whether or not victims of domestic violence survive their experience.

- Rates of domestic violence and reporting vary widely across geographical regions
- Visualizations can be effective in showing the scope of the problem

Reports/Data

In the next stage, journalists identify reports and data related the subject of the investigation. These reports can be found through searches, through data requests to the government and CSOs or through the creation of data for the investigation.

For the domestic violence story, journalists identified the following key reports:

Justice through the Eyes of Afghan Women: Cases of Violence against Women Addressed through Mediation and Court Adjudication UNAM?OCHA
An Update on Implementation of the Law on Elimination of Violence against Women in Afghanistan
AFGHANISTAN Ending Child Marriage and Domestic Violence
USIP Women's Access to Justice in Afghanistan
World Bank Gender Data Portal

Methodology

To evaluate the source of the data, journalists answer the following questions (here with sample answers from the first report)

Who gathered the data?	United Nations Assistance Mission in Afghanistan
When was the data gathered?	Detailed information from 18 of Afghanistan's 34 provinces for the one-year period October 2012 to September 2013 with technical review by the UN Office of the High Commissioner for Human Rights
What time period does the data cover?	Two years (2012-2013)
How was the data gathered?	Field monitoring and analysis of police and court records

Important findings

Finally, journalists read the executive summary of the report and write down 3-5 interesting findings.

1. The number of reports of domestic violence is rising
2. The economic and social vulnerability of women remains constant
3. Most cases of domestic violence are registered through the police

4. Most cases of domestic violence are resolved through mediation

Hypothesis and Questions

The most important stage of the data story process is the formulation of the hypothesis and questions. In scientific research, a researcher develops a hypothesis with a suspected set of ideas and then builds an experiment to support the hypothesis. The same process is true for data journalism. The journalist, using his or her news nose, develops a hypothesis that can be proved or disproved with data. News articles have a habit of a bringing up more questions than they answer, and following up with a hypothesis and investigation can uncovered details that lead to a data journalism story.

Building a Hypothesis for Data and Investigative Stories[7]

1. **A hypothesis gives you something to verify, instead of trying to uncover a secret.** People do not give up their secrets without a very good reason. They are much more likely to offer confirmation of information that is already in your possession, simply because most people hate to lie. A hypothesis enables you to ask them to confirm something, rather than to advance information. It also puts you in the position of someone who is open to discovering that there is more to the story than he or she thought at first, because you are willing to accept that there are facts beyond what you suspected at the start.

2. **A hypothesis increases your chances of discovering secrets.** A lot of what we call "secrets" are simply facts that no one ever asked about. A hypothesis has the psychological effect of making you more sensitive to the material, so you can ask those questions. As the French investigator Edwy Plenel said, "If you want to find something, you have to be looking for it." We would add that if you're really looking for something, you'll find more than you were looking for.

3. **Hypotheses makes it easier to manage your project.** Having defined what you're looking for, and where to start looking for it, you can estimate how much time the initial steps of the investigation will require. This is the first step to treating an investigation as a project that you can manage. We'll return to this point at the end of this chapter.

4. **Hypotheses are a tool that you can use again and again**. When you can work in a methodical way, your career will change. More important, you will change. You will no longer need someone to tell you what to do. You will see what needs to be done to combat some of the chaos and suffering in this world, and you will be able to do it. Isn't that why you became a journalist in the first place?

[7] http://unesdoc.unesco.org/images/0019/001930/193078e.pdf

5. **A hypothesis virtually guarantees that you will deliver a story, not just a mass of data.** Editors want to know that at the end of a specific period of time – a specific investment of resources – there will be a story to publish. A hypothesis hugely increases the likelihood of that outcome. It enables you to predict a minimum and maximum positive result for your work, as well as a worst case.

    - The worst case is that verification of the hypothesis will quickly show there is no story, and the project can be ended without wasting significant resources.
    - The minimum positive outcome is that the initial hypothesis is true, and can be quickly verified
    - The maximum is that if this hypothesis is true, others must logically follow, and either a series of related stories or one big story will result.

Tips for a strong hypothesis:

- Posits a theory that can either be proven or disproven with data
- Is specific about what is being measured
- The data is available
- The topic is important to the public

Below is an example of how to transform a weak hypothesis into a strong hypothesis:

1. Children in this country are dying of malnutrition.
2. Most children in this country who die below the age of five die of malnutrition.
3. Most of the children in this country who die below the age of five die of malnutrition and live in the poorest provinces.
4. Most of the children in this country who die below the age of five die of malnutrition and live in the poorest provinces despite a donor funded feeding program that was supposed to cut malnutrition rates in half over five years.

Hypothesis

After completing the background section for their investigation the Afghan journalists developed the following hypothesis:

Government programs cannot keep up with the increased demand for domestic violence services.

As you can see this hypothesis posits two theories that can be proven or disproven with data:

- Domestic violence reporting rates are increasing'
- Government services to respond to these reports are inadequate.

Please evaluate the following hypotheses, writing **S** (Strong) or **W** (Weak). If weak, please rewrite in the space provided.

1. The Ministry of Health should spend more on public health ___
   _____
   _____
   _____
2. Average primary school test scores are rising but fewer rural and low income families are able to send their children to school.____
   _____
   _____
   _____
3. The decline in government spending for public health over the last five years has contributed to a lack of progress in achieving Millennium Development Goals. ___
   _____
   _____
   _____
4. Lack of hospitals are responsible for high disease rates. ____
   _____
   _____
   _____
5. The rate of under-five child deaths is going down. ____
   _____
   _____
   _____
6. The number of honour killings are not declining because of uneven enforcement of the law.
   _____
   _____
   _____

Questions

Once you have a strong hypothesis, you should develop at least five questions that can be answered with data to prove or disprove your hypothesis. All the questions should be able to be answered with a number. Other types of questions, such as interview questions, will be developed after analysis.

The questions you write should:

- Measure the trend.
- Compare different groups.
- Measure the causes.
- Measure the impact.

Questions

Hypothesis: Most of the children in this country who die below the age of five die of malnutrition and live in the poorest provinces despite a donor funded feeding program that was supposed to cut malnutrition rates in half over five years.

1. What are the rates of under-five child death in Pakistan over the last 10 years?
2. What percentage of those deaths each year are attributed to malnutrition?
3. What are the rates of malnutrition by province?
4. What are the rates of poverty by province?
5. What were malnutrition rates when the donor funded feeding program started?
6. What were the rates after five years?
7. How many children were served by the program?
8. How many would have to have been served to meet the goal?
9. What was the cost per child?
10. How do malnutrition under five death rates compare to those of neighboring countries?

Remember the hypothesis for the domestic violence story:

Government programs cannot keep up with the increased demand for domestic violence services.

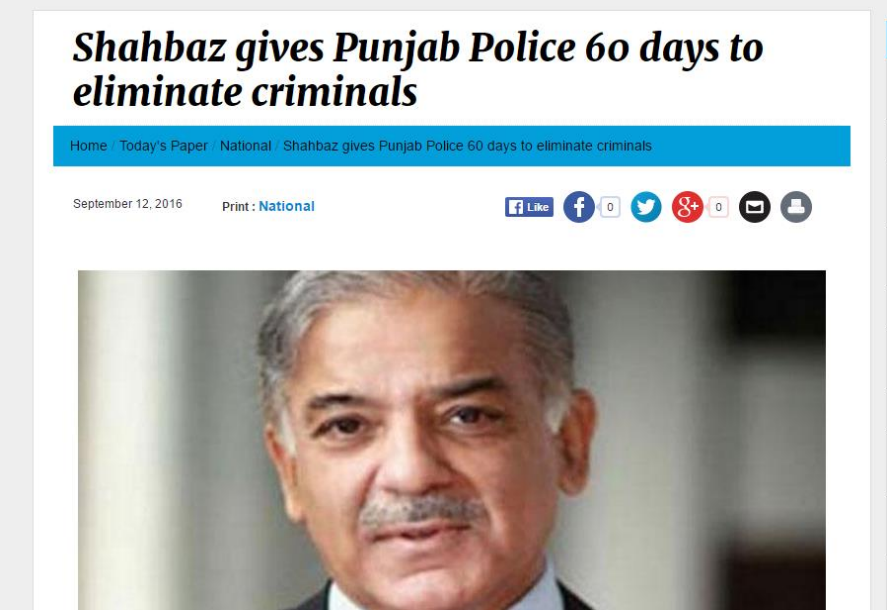
These are the questions the journalists wrote:

1. Are the number of cases being reported each year going up or down?
2. What is the age breakdown of women registering domestic violence cases under the new law? Are younger women registering cases?
3. What percentage of women have experienced domestic violence in Afghanistan
4. How bad is the situation for women in AF compared to the rest of the world?
5. What percentage of cases that are registered go to court? What happens to the rest of the cases?
6. What kinds of violence are being prosecuted under the new law?
7. What percentage of cases that make it to court result in a conviction?

Lesson 5: Enriching Stories With Data

Many stories that include data and statistics are not truly data journalism stories because they do not use data to explain the underlying issue. One of the most important skills as a data journalism is to recognize opportunities to transform an ordinary story into a data story. For each of the stories below, read the existing story and transform it into an idea for the data story. In some cases, the journalist has a hypothesis but fails to prove it with data. In other cases, there is data and statistics but it unclear what the journalist is trying to prove. Let's critically evaluate the claims in these news stories from a data perspective – and further suggest what data would be needed to support the claim.

Enriching Stories with Data: Governance and Security



Shahbaz gives Punjab Police 60 days to eliminate criminals

Home / Today's Paper / National / Shahbaz gives Punjab Police 60 days to eliminate criminals

September 12, 2016 Print : National

Here is a news article about crime reduction efforts that is surprisingly short on data

https://www.thenews.com.pk/print/149698-Shahbaz-gives-Punjab-Police-60-days-to-eliminate-criminals

Hypothesis:

_____
_____
_____

Questions:

_____
_____
_____
_____
_____
_____

Enriching Stories with Data: Economics and Budgets

Home» Pakistan » Banking & Finance

RS2389MN OUT OF TOTAL FATA BUDGET UTILIZED

Tuesday, 19 April 2016 15:18 221 Posted by Parvez Jabri

ISLAMABAD: Senate Standing Committee on States and Frontier Regions (SAFRON) was informed on Tuesday that Rs 2389.367 millions out of

Here is a news article that attempts to explain underspending of the allocated budget in FATA: http://www.brecorder.com/pakistan/banking-a-finance/291204-rs2389mn-out-of-total-fata-budget-utilized.html

Hypothesis:

_____

_____

_____

Questions:

_____

_____

_____

_____

_____

_____

Enriching Stories with Data: Education

This is an article that asserts that education in KPK is improving: http://dailytimes.com.pk/khyber-pakhtunkhwa/21-Aug-16/kp-education-reforms-produce-outstanding-results-report

Hypothesis:

_____
_____
_____

Questions:

_____
_____
_____
_____
_____
_____

Enriching News Stories with Data: Health

DAWN

EST POPULAR PAKISTAN TODAY'S PAPER OPINION WORLD SPORT BUSINESS MAGAZINE CULTURE BL

Polio monitoring board enhances targets for Pakistan

IKRAM JUNAIDI — UPDATED AUG 26, 2016 10:14AM

Like 10 Share Tweet 0 COMMENTS EMAIL PRINT

ISLAMABAD: The Independent Monitoring Board (IMB) for Polio has increased a number of targets for Pakistan to eliminate the crippling disease by the end of the year.

These included an enhancement in the quality of polio campaigns, reduction in the number of missed children, tracing children affected with polio-like diseases and the stability in the leadership of the polio programme.

The board also called for a peak performance from all stakeholders to

This is an article seeking to explain whether Pakistan is winning the fight against polio: http://www.dawn.com/news/1280074/polio-monitoring-board-enhances-targets-for-pakistan

Hypothesis:

_____
_____
_____

Questions:

_____
_____
_____
_____
_____
_____

Enriching News Stories with Data: Health

Upcoming anti-polio drive facing multiple challenges

Islamabad 23 HOURS AGO BY SHAH NAWAZ MOHAL

Shah Nawaz Mohal
The writer is a law graduate and member of staff, Islamabad Bureau.

All civic authorities asked to contribute personnel and machinery to make polio eradication campaign a success

The locals are concerned about the success of polio eradication drive, starting from September 26.

There are concerns pertaining to the availability of human resource in many departments and issue of reaching the targeted audience and access to populace that is in transition and most importantly, the willingness of people to let volunteers administer polio drops.

This is an article seeking to explain whether Pakistan is winning the fight against polio but from a negative perspetive: http://www.pakistantoday.com.pk/2016/09/17/city/islamabad/upcoming-anti-polio-drive-facing-multiple-challenges/

Hypothesis:

_____
_____
_____

Questions:

_____
_____
_____
_____
_____
_____

Enriching News Stories with Data: Human Rights

THE EXPRESS TRIBUNE BUSINESS

Implementing 27 UN conventions: 'Road map' developed to reduce gender wage gap

By Peer Muhammad Published: September 3, 2016

3 SHARES SHARE TWEET EMAIL

Treaty Implementation Cell discusses course of action. PHOTO: FILE

This is a typical summary report on an international human rights study:

http://tribune.com.pk/story/1175102/implementing-27-un-conventions-road-map-developed-reduce-gender-wage-gap/

Hypothesis:

_____
_____
_____

Questions:

_____
_____
_____
_____
_____
_____

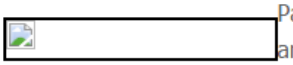
Enriching News Stories with Data: Environment and Agriculture

Installation Failures

By Hussain Ahmad Siddiqui

INSIGHT

The requirement to import two hundred bulldozers (track-type/crawler tractors) by the Balochistan government has, brought to limelight the myopic vision and lack of political will on the part of successive governments that failed to install the planned manufacturing facilities in the last few decades. Budget allocation of Rs3 billion was made by the provincial government in year 2014-15 for import of these bulldozers, but only 71 units have been procured so far.



Pakistan has a total land area of 79.61 million hectares (MH or Mha) and total physical area 57.99MH, including 4.55MH forest areas. At present, an area of 22.10MH is under agricultural use, whereas 23.01MH is not available for cultivation being under farm homesteads, farm roads etc. More than 8.25MH remains cultivable waste or uncultivated farm area, mostly in Balochistan. To improve the agriculture growth rate, there is a need to turn quickly the waste land into cultivable area. But nothing concrete has been done in this direction. In fact, currently, there is negative growth in the agriculture sector; due to a variety of factors though.

Read an analysis to a stalled potential solution to food security challenges:
https://www.thenews.com.pk/magazine/money-matters/149556-Installation-failures

Hypothesis:

_____
_____
_____

Questions:

_____
_____
_____
_____
_____
_____

Practice: Find Data Angles for Public Interest Topics

In this exercise, we will brainstorm public interest topics that would benefit from statistical analysis to put the issue into context for the reader.

Scenario

You have been asked to map out different scenarios in which a trigger event is an opportunity for a deeper data driven analysis of a phenomenon. For each of the following events, list a hypothesis that could explain the issue:

1. A teachers' strike in which teachers are demanding a pay raise and reduced class size

   Hypothesis:

   _____
   _____
   _____
   _____

2. A food shortage in two provinces in Pakistan

   Hypothesis:

   _____
   _____
   _____
   _____

3. An outbreak of measles among children in rural areas

   Hypothesis:

   _____
   _____
   _____
   _____

4. A program to install solar panels in villages without electricity

   Hypothesis:

   _____
   _____
   _____
   _____

5. A government health pledge: http://www.pakistantoday.com.pk/2016/09/08/city/islamabad/who-lauds-pakistans-vision-2025-for-improving-health-indicators/

   Hypothesis:

   _____
   _____
   _____
   _____

Lesson 6: Analyzing Fact Sheets

Overview

In this exercise we will look at data that has been analyzed and visualized to provide an overview of health in Pakistan. The goal is to interpret the data, begin to explore where the data came from, identify what is most interesting about the data, write a hypothesis and put together a story based on the findings.

To begin, open one of the followin:

Global burden of diseases, injuries, and risk factors profile:

http://www.healthdata.org/pakistan

WHO country profile:

- http://www.who.int/gho/countries/en/ (Choose a country, select Country Profiles and select General health statistical profile)
- **UNICEF country profile**
  http://www.unicef.org/infobycountry/ (Select a country and select statistics)
- WHO child and mother profile
  http://www.who.int/maternal_child_adolescent/epidemiology/profiles/neonatal_child/pak.pdf

Background

For each of the data sources, answer the following questions:

1. Who gathered the data?

   _____
   _____
   _____

2. When was the data gathered?

   _____
   _____
   _____

3. What time period does the data cover?

_____
_____
_____

4. How was the data gathered?

_____
_____
_____

Understanding the indicators

- What do the indicators mean?
- Can I look up definitions of indicators I don't understand?
- What are the differences between the categories?
- What indicators are not included in this data that would provide more context?

Underline 3-5 data points can answer these questions

- What is interesting or surprising about this data?
- What in this data could help citizens make better decisions about their health?
- What in this data could help policymakers make better decisions about health spending?
- What in this data could explain big picture health trends in the country?

Hypothesis

Write down a hypothesis that you can prove with this data:

_____
_____
_____

Prepare the story

Put your data points in order

- Start with the data that is most important to answer your hypothesis
- Add data that goes into further detail or provides context about the trend

Write the story

Evaluate the stories

1. Did the story answer the hypothesis?
2. Did each data point in the story support the hypothesis?
3. Did the data points come in a logical order in the story?

Sample Stories

To Grow a Healthy Adult Population, Evidence Says to Care for Children

Sudanese citizens are becoming sick throughout their lives due to preventable conditions during childhood. Weighing too little as children and poor breastfeeding are among the top three risk factors that make Sudanese citizens sicker throughout their lives.

2010 data collected by a global international health organization, the Institute for Health Metrics and Evaluation, explores the reasons that so many Sudanese are getting sick. They found that the two top two diseases that affect people in Sudan: diarrhea, nutritional deficiencies, are worse because children don't get enough food or time breastfeeding as children.

Lower respiratory infection, including severe coughs, diarrhea and malaria rob Sudanese of more years of their life than any other disease, meaning they are the biggest problem for life expectancy. For all these diseases, fewer years were lost than in 2010 than in 1990. But over 20 years, the percentage of years lost to those diseases on average has dropped less than 40%. In many countries, these most people survive these diseases, but vulnerability that starts with childhood risk factors, means that many Sudanese don't stand a fighting chance.

There has been some progress in growing healthier children. Malnutrition caused by not consuming enough protein dropped as a cause of years of lives lost by 44% over two years, meaning not as many people are dying from lack of protein, but from other diseases associated with not eating a balanced died and getting enough food. In 1990 not getting enough protein was the third cause of premature death and it has fallen to eighth place, after most deaths associated with the danger of being born and the more recent menace of HIV.