Lesson 4
Data Sources

Overview

Where does one start digging publish-worthy data? Journalists can collect or "mine" qualitative and quantitative data from a wide range of sources. "Data mining" refers to the process of searching massive datasets to identify patterns. Datasets are available in government databases, civil society websites, businesses' servers, and open data portals. Where there is absence of data, an individual or independent groups can also create their own datasets.

In this part of the course, we will learn where to get data.

Duration: 1 hour

Introduction

Google it! It's easy to say this when tasked to find a piece of information. When looking for data, Google can also be a friend, given its index running 100 petabytes of data (1 petabyte is equal to an astounding one million gigabytes). But, Google can be overwhelming with its nearly 50 billion pages of search results. While the World Wide Web can be a handy tool in digging for data, a best practice in data journalism is to keep tabs on specific sources of accurate, credible, and reliable datasets. Depending on the development story one is chasing, a journalist may look at the national government's records, a global organization's report, or somebody's smartphone.

To set expectations, data is never clean and ready for use in reporting and analysis. Worse, much of available data are in PDFs and not in spreadsheets. Usually, these datasets are limited or anonymized or scrubbed. You will encounter data with different granularities, level of cleanliness, availability of documentation, or varied structures (can be structured like spreadsheets, semi-structured like emails, or unstructured like documents). Whatever form that might be, the first step is finding it.

Objectives

At the end of the lesson, you should be able to:

1. Enumerate good sources of data that can be used for stories;
2. Describe the types of data from these sources;
3. Discuss how to mine data from various sources

Lesson proper

Discussion flow:

1. Government research data and public records
2. Non-government sources (civil society, academe, businesses)
3. Open data portals
4. Mining one's own data

GOVERNMENT RESEARCH DATA AND PUBLIC RECORDS

Government

National governments allocate resources for research to better inform policymakers about public issues. In a functioning state, data is crucial in the decision-making processes on how to improve the lives of its citizens. Government units regularly publish sets of data through their official statistics portal. In the United States, data about its people and its economy are found in the US Census Bureau; in the United Kingdom, the British government has its Office for National Statistics. In the Philippines, we have the Philippine Statistics Authority or PSA. Every quarter, business and financial journalists are on their toes for the release of closely watched statistics from PSA. Like other national statistics authorities, PSA announces a calendar of "release dates" for data such as consumer price index, trade statistics, and performance of agriculture.

At the landing page of PSA (psa.gov.ph), major data entries are visible:

- population (births, marriages, deaths, fertility rate, and total population count)
- economic growth rate (gross domestic product and gross national income)
- prices/inflation rate
- trade (exports and imports)
- incidence of poverty
- average family income
- employment rate
- proportion of seats held by women in parliament

Another important source of data, especially for development journalists, is the National Economic and Development Authority (NEDA). According to its website, NEDA is

> the government's socioeconomic planning body, highly regarded in macroeconomic forecasting, policy research, and analysis; an acknowledged institution in providing high-level policy advice, developing consensus, and setting the agenda for inclusive development.

NEDA publishes weekly economic updates based on data from the PSA and the Banko Sentral ng Pilipinas (BSP), but it also publishes annual reports on development projects in the Philippines and other research-heavy publications. Central banks such as the BSP are also an authority in data, especially when it comes to interest rates, which experts think greatly affect consumer spending in a country.

Since every government department is responsible for a specific sector (e.g., transportation, education, health), each of these releases public data.

Public records

Governments have been keeping a census of their population since time immemorial. A census is "an official count or survey of a population, typically recording various details of individuals." It is said to have originated in the early 17th century and denoted a "poll tax" as it was applied to register citizens and property in ancient Rome, usually for taxation purposes.

In modern times, census data includes data on population such as age and sex, employment, housing, education, health, income and poverty, families and living arrangement, business and economy, and population estimates.

In the Philippines, government departments release public data. Usually these are journalists' first stop when gathering information. In instances when specific data is not available, they request for it —as we learned in a previous module.

List of Government Websites

Department of Agrarian Reform	http://dar.gov.ph/
Department of Agriculture	http://www.da.gov.ph/
Department of Budget and Management	http://www.dbm.gov.ph/
Department of Education	http://www.deped.gov.ph/
Department of Energy	https://www.doe.gov.ph/
Department of Environment and Natural Resources	http://denr.gov.ph/
Department of Finance	http://www.dof.gov.ph/
Department of Foreign Affairs	http://www.dfa.gov.ph/
Department of Health	http://www.doh.gov.ph/
Department of the Interior and Local Government	http://www.dilg.gov.ph/
Department of Justice	http://www.doj.gov.ph/
Department of Labor and Employment	http://www.dole.gov.ph/
Department of National Defense	http://www.dnd.gov.ph/
Department of Public Works and Highways	http://www.dpwh.gov.ph/
Department of Science and Technology	http://www.dost.gov.ph/
Department of Social Welfare and Development	http://www.dswd.gov.ph/
Department of Tourism	http://www.tourism.gov.ph/pages/default.aspx
Department of Trade and Industry	http://www.dti.gov.ph/dti/index.php
Department of Transportation and Communications	http://www.dotc.gov.ph/

DATA FROM CIVIL SOCIETY, ACADEME, BUSINESSES

A professor emeritus of Decision Sciences said in his book *Analytics Stories* that "part of human nature is the desire to boil down a complex concept like inequality to a single number" (Winston, 2021). And so, numerous organizations, scholars, and enterprises come up with data through indexes to measure human aspirations such as inequality and development. Some examples are the Gini index, the Palm index, Facebook Social Connectedness Index, World Happiness Index, and the Human Development Index.

Clearly, aside from government databases, data from civil society, academic institutions, and business groups can also be a pool of sources for the development communicator keen to improve the lives of people.

Civil Society

Because government agencies' data may not be exhaustive and because some journalistic stories are critical of the government, journalists and communicators also turn to civil society.

"Civil society" refers to non-state, non-profit, voluntary entities formed by people in a society, such as The Asia Foundation, Greenpeace, Save the Children, and the Organization for Economic Co-operation and Development or OECD. Civil society organizations can be community groups, non-governmental organizations, labor unions, indigenous groups, charity groups, faith-based organizations, professional associations, and foundations. The World Economic Forum (WEF) said in its website that the term, which became popular in political and economic discussions in the 1980s, was identified with non-state movements that defied authoritarian regimes, especially in Eastern Europe and Latin America.

The trillion-dollar sector sometimes dubbed "Volunteerland" hold institutions to account and raise awareness of societal issues while implementing actions, too, to help the marginalized:

> When mobilized, civil society - sometimes called the "third sector" (after government and commerce) - has the power to influence the actions of elected policy-makers and businesses. But the nature of civil society - what it is and what it does - is evolving, in response to both technological developments and more nuanced changes within societies. (WEF)

Aside from OECD and other organizations mentioned above, some of the popular groups include the World Wildlife Fund and Amnesty International. In the Philippines, the Asian Development Bank has listed several civil society groups and their types (e.g., farmers' associations, civic clubs, athletic associations, social welfare organizations, cooperatives, rural workers' associations, and umbrella groups like Bagong Alyansang Makabayan).

Around the world, intergovernmental organizations like the United Nations and World Trade Organization actively participate in civil society circles, and regularly publish resources such as statistics, economic research, and knowledge products. In other words, their databases and publications provide access to data. The UN's numerous agencies (WHO, FAO, ILO, IMF, WFP, UNDP, UNICEF, etc.) are a rich source of data.

Academe

In 2016, a team of data scientists at the Computational Story Lab at the University of Vermont in the US analyzed with its computers more than 1,300 digitized works of fiction from the Gutenberg Project, a library of over 60,000 digitized and archived cultural works from around the world. Their research discovered that a thousand classic literary stories have, in general, six main emotional arcs: rags to riches (steady rise); man in a hole (fall-rise); Cinderella (rise-fall-rise); tragedy (steady fall); Icarus (rise-fall); and Oedipus (fall-rise-fall). Based on digitized texts, data has confirmed the emotional arcs in stories.

This is one example of brilliant research that the academe produces by generating datasets and analyzing them. Research universities are a good source of data, especially if they excel in certain fields; University of Amsterdam, for example, is known for its excellent research track in communication compared with its peers around the world; Oxford University's research has been consistently ranked number 1 in the THE World University Rankings; in Asia, scholars from various academic institutions have been contributing to economic and social development of their home countries. In the Philippines, UPLB and the International Rice Research Institute are on top of research in agriculture and other sciences.

Business

Financial and business journalists are probably more exposed to datasets than any other types of journalists. This is because on a daily basis they report changes in stock exchange

or economic indicators and compare data per year, quarter, month, week, day, and even hour. In response to the volume of data that these journalists have to deal with every minute during a trading hour, some newsrooms have employed artificial intelligence tools.

Typically, the go-to data portals for finance and business journalists are Google Finance, Yahoo Finance, NASDAQ, NYSE market data, national statistics sites, and governments' securities exchange commission portals.

For macroeconomic trends, the World Bank has an open data catalog (a listing of its datasets), data bank (an analysis and visualization tool), and microdata library (data collected through sample surveys of households, business establishments or other facilities). However, some original dataset may not be shared or distributed outside of the World Bank Groups, and some analyses may be shared, but with citation. One example of a dataset with limited access is a collaboration by Facebook, OECD and World Bank, "Future of Business Survey - December 2019." The joint survey provides monthly data on the perceptions, challenges, and outlook of online Small and Medium Enterprises.

OPEN DATA PORTALS AND MINING OWN DATA

Open data

Open data are searchable, machine-readable, license-free sources of data. If data journalists are to keep their resources organized, they should have a database of databases (DoD). Examples of these DoD or metadata are news databases such as Factiva (Dow Jones) and Wisers News (Hong Kong); tabulation services or large and high-dimensional datasets aggregated into smaller tables (e.g., statcompiler.com, gapminder.org/tools, datausa.io); and statistics or charts search engine, large collection of charts (e.g. theatlas.com, statista.com, ourworldindata.org, data.worldbank.org).

An amazing DoD is this "Awesome Public Datasets" which includes the following:

Academic torrents of data sharing from UMB (83TB of research data)
Harvard dataverse (a network of scientific data)
OpenDataMonitor (catalogues and datasets of European open data)
OpenDataNetwork (US-focused search tool within open data ecosystem)
Universities Worldwide (database of ~7,700 universities in 200+ countries)

World Inequality Database (comparison of inequality between countries)

Own data

While an overwhelming volume of data is available online, journalists and communicators may still opt to produce their own data depending on their topic and their community's needs. They can get numbers by interviewing experts, retrieve statistics from local research reports and surveys, and mine raw data from massive user generated content. ABS-CBN, for instance, has its own data analytics team that gathers and analyzes its own data, just like Bloomberg, which is a global media outlet that collects, catalogues, visualizes, and reports its own data across decades.

For media companies, collecting their own data will require more effort and programming skills. One can collect data from public streams (like Twitter), extract from documents or files, or set up a platform to crowdsource data.

REFERENCES

Duarte, N. (2019). *Data Story: Explain Data and Inspire Action Through Story.* Canada: Ideapress.

World Economic Forum. Who and what is 'civil society'?" Retrieved 24 July 2021 from https://www.weforum.org/agenda/2018/04/what-is-civil-society/

Winston, W. (2021). *Analytics Stories: Using Data to Make Good Things Happen.* Indianapolis: Wiley.

LECTURE OUTPUT: THE RECORDS REQUEST PROJECT

Acquire a dataset from the government that is not available on their website or portal. Document every step of your request, every answer of the agency, transfer of units, until the final dataset you get.

Submit your data and a short deck sharing your experience.

This is 70% of your final lecture requirements. This may take you a few days or the whole semester.

Deadline: End of the semester, final exams week (January 4)

Rubric

Criteria	9-10	6-8	4-5	1-3
1.Quality of dataset (multiplier: 5)	Dataset has great potential for data stories	Dataset has good potential for data stories	Dataset can be used for data stories	Dataset is unusable for data stories stories
2. Attention to detail (multiplier: 3)	Process is well documented and reflection on experience is insightful; raised important issues around acquisition of data and policy	Process is l documented and reflection on experience is insightful; did not raise important issues on data policies on acquisition	Process documentation has a few lapses; lack of insights on data acquisition experience	Process is not documented
3. Clarity of presentation (multiplier: 2)	Nature and content of dataset is explained in a clear and comprehensive manner; Process and experience are interwoven into a clear narrative	Nature and content of dataset is adequately explained; Process and experience are discussed in the presentation	Nature and content of dataset is not clearly explained; lapses in discussing process and experience	Poor presentation of nature and content of dataset; lack of transparency in sharing process

Exercise 1: Identification of sectors/topics, preparation of data sources, and account set-up

Part A: Setting up of accounts

Overview

Like in other forms of journalism, the journalist needs to scope out topics, research, gather data, book interviews, and draft a story, but with focus on data gathering and data analysis. All these can begin with simple steps, and one of these is by setting up accounts.

In this exercise, your laboratory instructor will introduce you to different Data Journalism platforms and tools. Then, you will sign up for free accounts and start exploring these platforms.

Objective

Set up accounts in Google Sheet, Data Studio, Flourish, and Datawrapper.

Mechanics

1. Form a group of four (notify your lab instructor if you prefer to work by yourself). Then, using your UP email accounts, create a shared Google Sheet (sheets.google.com). Get the link through the Share button at the upper right corner, and send this to your instructor. You and mates will work on your data using this shared sheet.
2. Create individual accounts in https://flourish.studio and https://www.datawrapper.de and start exploring these platforms for data visualization, too. Your lab instructor will also help you explore Flourish and Datawrapper, and introduce you to other data viz platforms such as Tableau in later modules.

Part B: Choosing a topic

Overview

Humanity's most intractable headaches are usually the challenges of development: illiteracy, malnutrition, inequality, poverty, corruption, and other systemic barriers to getting a government or a community to work right. For your data story, your team will choose a development-oriented topic under a pre-identified sector.

Objective

Identify a development-oriented topic and objectives for exploration.

Mechanics

1. Suggest three topics and seek your professor's approval.
2. Identify the affected sectors and the development issues.
3. Enumerate your strategies and methods and explain how you will go about data collection.

Part C: Pooling data sources

Overview

In the Lesson 4 of your course pack, a detailed discussion covers different sources of data. With your approved topic, you will search for data sources and download a data set or two that can be useful for your topic. Data sources for your project can be government research data, public records, or other open data portals.

Objectives

At the end of the exercise, you should be able to:

1. Enumerate data sources i.e. websites to get the data from; and
2. Produce a data set for your chosen topic from the listed data sources

Mechanics and Deadlines

With the array of data available around, a journalist may seem unsure how to start a data journalism project. Approaches vary, you can start by gathering a particular data set available or of your interest and start analyzing what story comes up or you can have a general topic in mind and start your search for data sets to build on stories.

The web is a source of data but it can get confusing. And, if you don't know how to look, you might get lost. A first good step can be the good old search engine (Google, DuckDuckGo, Bing, etc.)

Here are some tips in using search engines:

1. Use keywords or terms and the format or source that you want the data to be in (ex. Filetype: pdf or Filetype:XLS; include inurl:downloads; indicate domain name (ex. site:agency.gov)
2. Browse data sites and services (official data portals) or look for metadata sites or a portal that lists data portals.
3. Write an FOI request to a government agency (if info is not available online or could not be retrieved online); Study the FOI provision in the Philippines; submit specific requests; send multiple requests (different agencies, etc); keep a record; ask for raw data.

Expected Outputs

1. Proposed topic and list of data sources
2. Data set in spreadsheet format, in your shared Google Sheet

Assessment

Criterion	Performance Level		
	8-10 points	**4-7 points**	**1-3 points**
Quality of Data set/s (Multiplier: 5)	Relevant, up-to-date/latest available, credible, complete; Format wise: available in xls or csv, and with clear source of dataset	Somewhat relevant, recent but not the latest available, lacking information or entries; Format wise: in Excel file but no clear source	Somewhat irrelevant data, outdated, incomplete; Format wise: not in xls or csv format (e.g., jpeg, pdf, or flat image)
Development orientation of the proposed topic (Multiplier: 5)	The output clearly established the development orientation of the topic, localized the topic/issue, and provided sufficient details to establish the context of the topic.	The output fairly established the development orientation of the topic, localized the topic/issue, and provided sufficient details to establish the context of the topic.	The output was unable to establish the development orientation of the topic and failed to localize the topic/issue. The output did not provide enough details to establish the context of the topic.
Total			

Perfect score=100 points