Local Government Training Institute (LGTI) Short Course Data Curriculum

Data Fundamentals Course

Module 2: Introduction to 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 2: Introduction to Data

STUDENT WORKBOOK

This module introduces data by providing a common understanding of what constitutes data and why it is of such value to local government professionals. It then provides some examples of data from simple, intermediate to advanced types including the opportunities and limitations that data presents. The following lessons explore different types of data access with a focus on defining open data and examples. An exploration of why open data is be a crucial component of the work of local government officials in Tanzania is further developed with an understanding of the common terms and concepts learners should be familiar with when working with open data. Finally, this module ends with an overview of the School of Data Pipeline, a framework that makes it easy to work through various stages to generate useful insights or products from data.

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

- Define and provide sufficient examples of data and open data.
- Explain different data related terms and concepts used in the data ecosystem.
- Explain and provide examples of the value of data in a given context or sector.
- Identify the various stages of the Data Pipeline and how it can be used in data work.

Contents

Introduction

When people usually think or talk of data, it can be in reference to different ideas in a particular context. Some people use data to refer to the amount of bandwidth they have on their smartphones or laptop in order to access internet services while others may be referring statistics of some form. In this lesson, we will develop a working foundational definition of data that encompasses the origins of the term and the value it offers to people who are driven to generate insights for their work.

Class Exercise 1: Defining Data

janneke staaks (CC) by Research Data Management on Flickr

This activity aims to develop a definition of data which will serve as a foundation for the rest of this course. It is always crucial that anyone working with data has a valid definition and understanding of what constitutes data and what does not. This provides a common foundation to explore what is necessary to properly work with data and also provide a common platform for collaboration.

Below is a group exercise that will allow you to explore what you understand by data and help you work with other individuals and the entire class to develop an acceptable understanding of data for the course. Read the instruction below and see your instructor for any resources you need to complete this exercise.

1. Each student should get a red, yellow and green index card/post-It from the instructor for this activity.
2. Without using an additional resources, give a definition and example of data in two sentences on the red index card/post-It. [3 mins]
3. Using the internet or other resources (except other students or instructors), refine your definition and example of data on the yellow post-It. Repeat the same definition and example even if this is the same as the first definition. [5 mins]
4. Find three other students and discuss your definition and example from the yellow post-It. [10 mins]
5. Write down your final definition and example of data after the peer discussion on the green post-It. [2 mins]
6. Regroup to discuss with the entire class and instructor. [10 mins]

What is Data?

Data is all around us. But what exactly is it? Data is a value assigned to a thing. It can also be thought of as a snapshot of an entity or activity. Take for example the child in the picture below. His birth is being registered by an official.

The data that is being recorded could be about his or her name: first name, middle name or last name. It could be about gender, date of birth, place of birth, name of parents. Data about his physiology such as eye colour, hair colour, weight, height and so on could be recorded. All these things are data and they have value.

Another important way to think about data is as a snapshot of a thing which means it captures some aspect (but not all) of a thing. This means that data does not always give the complete picture about a thing but based on what knowledge one is interested in, the captured data may be enough or not. In the case of the child above, we do not have data on the child's blood type or any genetic disorders, which in the case of a medical emergency doctors cannot take any action without that data.

Common Data Types

In the example above, we can already see that there are different types of data. The two major categories are qualitative and quantitative data.

Qualitative data is everything that refers to the quality of something: A description of hair and eye colours , names, and additional notes made by the birth official are all qualitative data of the child above.

Quantitative data is data that refers to a number. In the case of a baby, the age, height, weight etc are all examples of this.

However there are also other categories that you will most likely encounter:

Categorical data puts the item you are describing into a category: In our example the gender “male” or “female” would be categorical, as will also the blood type (A, B, AB or O).

Discrete data is numerical data that has gaps in it: e.g. the age of the child. There can only be whole numbers of age (there is no such thing as 0.5 years)..

Continuous data is numerical data with a continuous range: such as the weight of the child which can be any value (4.5 kg, 4.57 kg or 4.579 kg). In continuous data, all values are possible with no gaps in between.

Task: Take the example of our newly-born(less than a year) child: can you find data of all different categories?

What is the value of data?

Data, when collected and structured suddenly becomes a lot more useful. Let’s do this in the table below.

First Name	Biruk
Last Name	Asante
Date of Birth	04/10/2016
Place of Birth	Amhara. Ethiopia
Gender	Male
Weight	4.5 kg
Height	67.8cm
Blood Type	AB

But each of the data values is still rather meaningless by itself. To create information out of data, we need to interpret that data.

Let's take the weight for instance: a weight of 4.5kg doesn't tell us much. It is only meaningful when we compare it to other males babies of a similar age. In natal health, a weight between 2.5kg and 5.5 kg is expected. Good, our baby has a healthy weight. This is information. But it still is not knowledge. Knowledge is created when the information is learned, applied and understood.

Given that data is just a snapshot of a thing or activity, it is also important to pay attention to the gaps that data can create in our knowledge. Collecting more data on a thing or activity can help reduce this gap but the key takeaway is that there will always be a limitation since we are unable to collect all data for a given entity in every scenario. For instance, we are not able to collect the temperature of our baby for every second of his life. However this only becomes a limitation if we need that data during medical treatment.

Unstructured vs. Structured data

Data for Humans

A plain sentence – "the baby is tremendously healthy with dark black eyes, black hair, expected weight of 4.5kg and height of 67.8cm for his age" – might be easy to understand for a human, but for a computer this is hard to understand. The above sentence is what we call unstructured data. Unstructured has no fixed underlying structure – the sentence could easily be changed and it's not clear which word refers to what exactly. Likewise, PDFs and images may contain information which is pleasing to the human-eye as it is laid-out nicely, but they are not ***machine-readable***[1].

Data for Computers

Computers are inherently different from humans. It can be exceptionally hard to make computers extract information from certain sources. Some tasks that humans find easy are still difficult to automate with computers. For example, interpreting text that is presented as an image is still a challenge for a computer. If you want your computer to process and analyse your data, it has to be able to read and process the data. This means it needs to be structured and in a *machine-readable* form.

One of the most commonly used formats for exchanging data is CSV. CSV stands for comma separated values. The same thing expressed as CSV can look something like:

"first name", "last name", gender,"place of birth", "date of birth", "weight (kg)", height (cm)", "blood type"
Biruk, Asante, male, Amhara,, "4/10/2016", 4.5, 67.8, AB

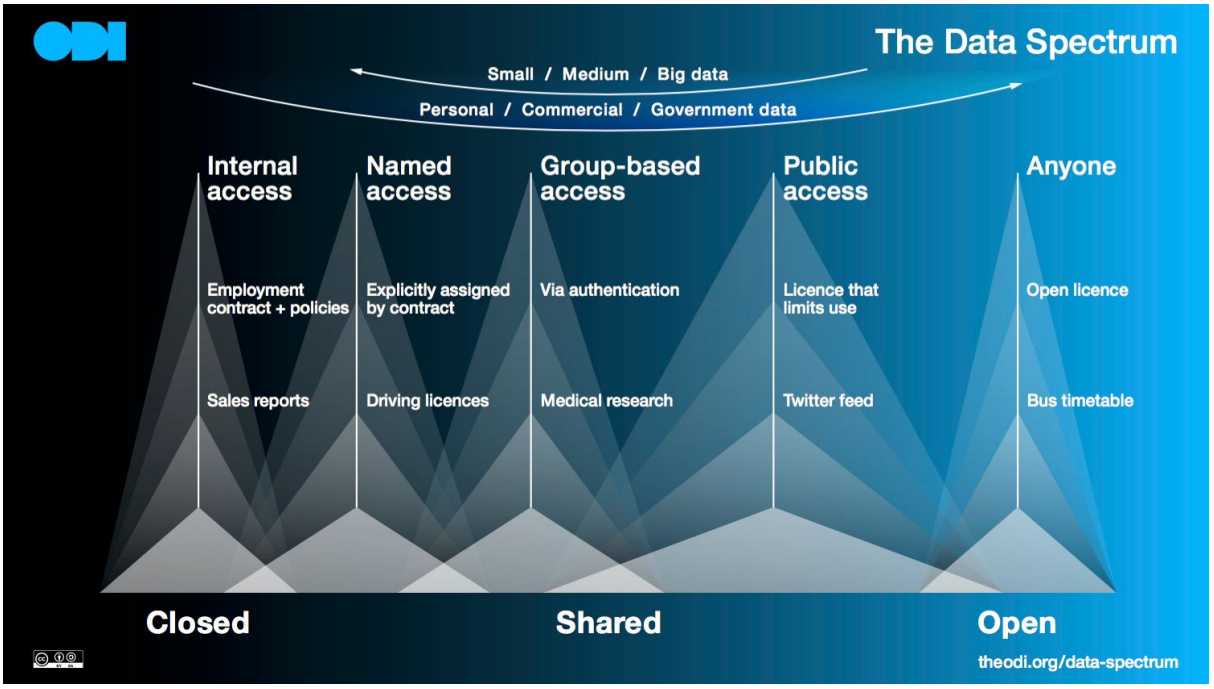
This is way simpler for your computer to understand and can be read directly by spreadsheet and other data software. Note that words have quotes around them:

[1] https://schoolofdata.org/handbook/courses/appendix/glossary/#term-machine-readable

This distinguishes them as text (string values in computer speak) – whereas numbers do not have quotes. It is worth mentioning that there are many more formats out there that are structured and machine readable which we will discuss later.

Task: Think of the last book you read. What data is connected to it and how would you make it structured data?

Data Access: Closed, Shared & Open Data



The Open Data Institute Data Spectrum (theodi.org/data-spectrum)

As we have seen, having access to key data lays a foundation to obtain knowledge that may be relevant to a given context. For a health official, collecting data on a child allows her to determine if the child has the characteristics of a healthy child or not. This knowledge helps her achieve her determine whether to take any extra measure to ensure the health of the child, the mother and the community.

This value of data has also led to various level of access which can be categorised as: *closed, shared or open data.* At one end of the data access spectrum is ***closed data*** which only provides access to authorised individuals or entities. The opposite end of the spectrum is **open data** which removes any form of authorisation and makes the data accessible to anyone in the public. **Shared data** sits between these two extremes which a combination of reduced permissions or limited access to a subset of data or specific group of data users based on a defined context.

We will delve deeper into the different types of data access particularly open data in the next section.

Lesson 2: Open Data

Class Exercise 2: Defining Open Data

This activity aims to provide a definition of open data and its value in public decision-making. The aim is to give learners a working definition of open data and examples to reference.

1. Each student should get a red, yellow and green post-It from the instructor for this activity.

2. Without using an additional resources, give a definition and example of open in two sentences on the red index card/post-It. [2 mins]
3. Using the internet or other resources (except other students or instructors), refine your definition and example of open data on the yellow post-It. Repeat the same definition and example even if this is the same as the first definition. [3 mins]
4. Find two or three other students and discuss your definition and example from the yellow post-It. [5 mins]
5. Write down your final definition and example of open data after the peer discussion on the green post-It. [2 mins]
6. Regroup to discuss with the entire class and instructor. [8 mins]

What is Open Data?

One way to define open data (by the Open Definition) is *data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike.*

Let us explore this loaded definition more in-depth:

- **Availability and Access:** the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.
- **Re-use and Redistribution:** the data must be provided under terms that permit re-use and redistribution including the intermixing with other datasets.
- **Universal Participation:** everyone must be able to use, re-use and redistribute - there should be no discrimination against fields of endeavour or against persons or groups. For example, ‘non-commercial’

> restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.

If you're wondering why it is so important to be clear about what open means and why this definition is used, there's a simple answer: ***interoperability***.

It is common to hear people talk about the advent of the information age, where many of our activities and interactions are being recorded by and through devices and sensors. To be clear, data is not something that just got invented in the last couple of decades. Any activity that has ever occurred in life has produced data. The only difference today is that we have the ability to collect these data more efficiently and at a cheaper cost. For example, the city of Dar es Salaam can now collect data on vehicles that drive through the Julius Nyerere International Airport on a daily basis. This data can now be used to generate various insights that are important to the city of Dar es Salaam.

The Value of Open Data

It is this value of collected data that leads to the access types that was mentioned before. In the past, the default access type for collected data such as the airport vehicles example mentioned above will be stored in a format that only give access to authorised airport officials. These formats could encompass writing data in hard copy notebooks to be stored on shelves in a locked room to storing data in a proprietary, internal database system requiring an administrative password. A more relaxed access level will be making it easier for other city agencies to get copies of the airport vehicle data for use. Closed data has its place in the data ecosystem including ensure privacy of individually-identifiable data (i.i.d) such as individuals medical records, data deemed as critical to national security or proprietary data. This leaves a subset of public data that has potential value but typically remained closed to a limited group of authorised user.

The push for open data arose from the drive to leverage of these datasets that do not fall under the privacy definitions listed above to be made publicly accessible in order to be used in developing valuable insights and products that will be of public benefit. It is in the light that formal definitions of what encompasses open data arose.

With a global push for open data many governments, international organizations are creating their own open data portals with the data they have aggregated and opened. These portals are a source of rich information for civic engagement and generating public interest stories. Although, these portals are by no means the only sources of data since much data lives on various government ministry websites. Some important datasets that are (or could be) open come from people's personal data that is
aggregated and anonymized. Much statistical information ultimately comes from surveys of individuals, but the end results are heavily aggregated so that individuals cannot be identified. The open data community should remain cautious about the need to protect privacy and individual well-being in the data's march towards public good and transparency.

Open data, especially open government data, is a tremendous resource that is as yet largely untapped. Many individuals and organisations collect a broad range of different types of data in order to perform their tasks. Government is particularly significant in this respect, both because of the quantity and centrality of the data it collects, but also because most of that government data is public data by law, and therefore could be made open and made available for others to use. Why is that of interest?

There are many areas where we can expect open data to be of value, and where examples of how it has been used already exist. There are also many different groups of people and organisations who can benefit from the availability of open data, including government itself. At the same time it is impossible to predict

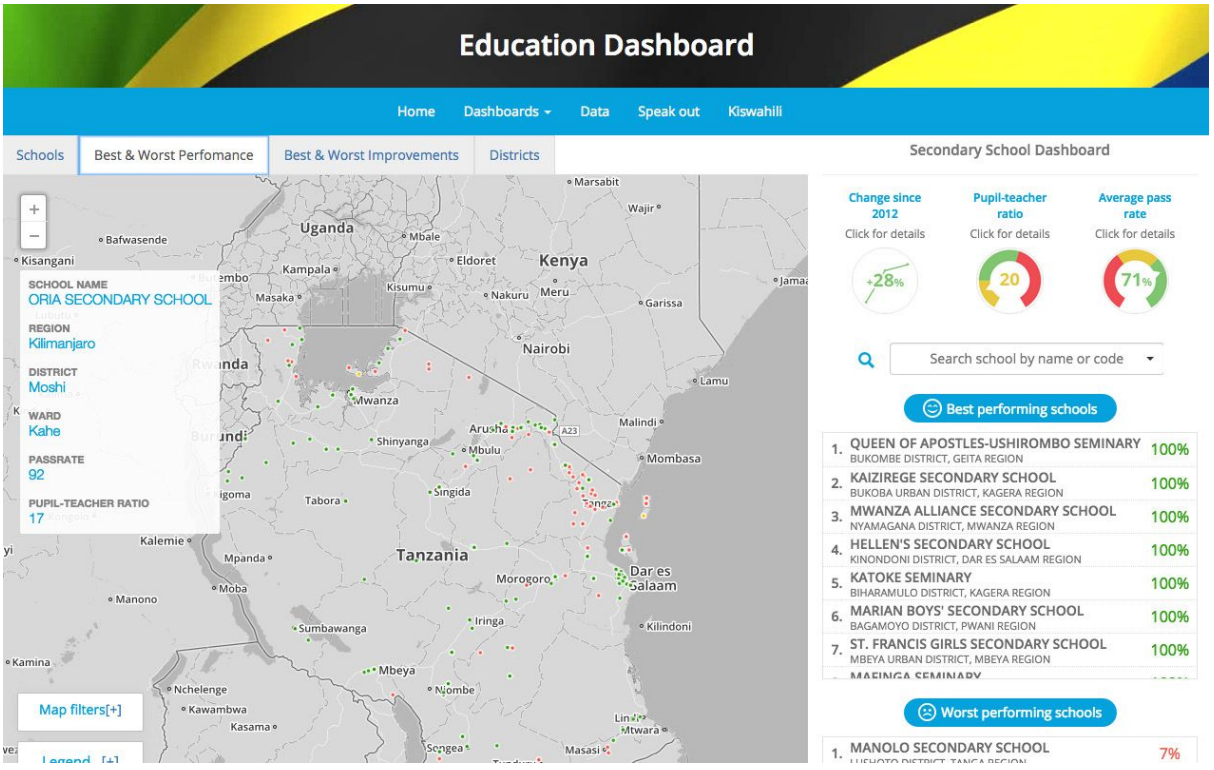
precisely how and where value will be created in the future. The nature of innovation is that developments often comes from unlikely places.

It is already possible to point to a large number of areas where open government data is creating value. Some of these areas include:

- Transparency and democratic control
- Participation
- Self-empowerment
- Improved or new private products and services
- Innovation
- Improved efficiency of government services
- Improved effectiveness of government services
- Impact measurement of policies
- New knowledge from combined data sources and patterns in large data volumes

Examples exist for most of these areas.

Tanzania's Open Education Dashboards: Improving education with open data



Mapping of school performance on the educationdashboard.org

Two recently established portals in Tanzania tried to improve low national examination pass rates, providing the public with more data on education and Tanzania's schools. The first, the Education Open Data Dashboard (educationdashboard.org), is a project established by the Tanzania Open Data Initiative, a government program supported by the World Bank and the United Kingdom Department for International Development (DFID) to support open data publication, accessibility and use. The second, Shule (shule.info), was spearheaded by a lone programmer, entrepreneur, and open data enthusiast who has developed a number of technologies and businesses focused on catalyzing social change in Tanzania. Although these projects initially encouraged citizens to demand greater accountability from their school system and public officials, both are in a state of near abandonment resulting from the lack of a clear sustainability and long-term management strategy.

Elections in Burkina Faso

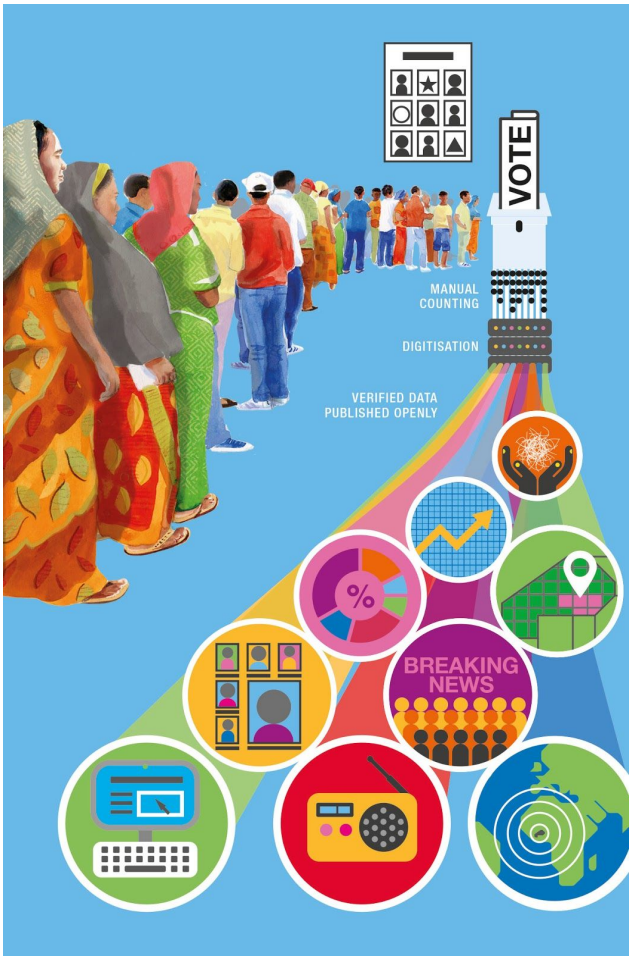


Image from ODI's article on [Open elections: lessons for open data from Burkina Faso and Zambia by ODI]

To ensure elections in Burkina Faso were conducted fairly, poll results were made available in real time via an official election website, which tracked candidates leading in each of the provinces. This project, run by the Burkina Open Data Initiative (BODI, http://data.gov.bf/about) with the support of the ODI, sought to promote democracy and trust between Burkina Faso's citizens and elected officials. For a country in transition like Burkina Faso, opening electoral data was seen as an important first step toward establishing longer term political stability and citizen trust in the electoral process, though the

number of citizens or organizations who actually accessed and acted upon the data is unclear[2].

In terms of transparency, websites such as the Odekro (odekro.org) in Ghana track activity in parliament and the law making processes, so you can see what exactly is happening, and which parliamentarians are involved. Open data can also help you to make better decisions in your own life, or enable you to be more active in society. In Kampala, Uganda Outbox a civic-tech hub through its PiMaa project has been developing local environmental sensors to record air and noise data that can be used to improve the conditions in the city[3].

Fighting Ebola in Sierra Leone

In the parts of West Africa affected by the Ebola epidemic, roads, village names, and villages were missing on many online maps. OpenStreetMap (OSM), a free, crowdsourced mapping tool provided critical mapping information to Sierra Leone's National Ebola Response Centre (NERC), the United Nation's Humanitarian Data Exchange, and to the Ebola GeoNode to assist them in coordinating public health strategies in response to the epidemic. The OSM data was then often mashed up with open data from affected governments and international organizations. Although the direct impact of open data in the Ebola response was difficult to empirically measure, those working on the ground during the response made clear that providing missing data in open formats played an important role in fighting a complex epidemic and coordinating relief efforts of those working in a chaotic, fast-developing context.

[2]https://theodi.org/case-study-burkina-fasos-open-elections
[3]https://medium.com/outbox-research/creating-open-data-using-the-internet-of-things-to-build-open-source-environment-monitoring-a0c0510b9065

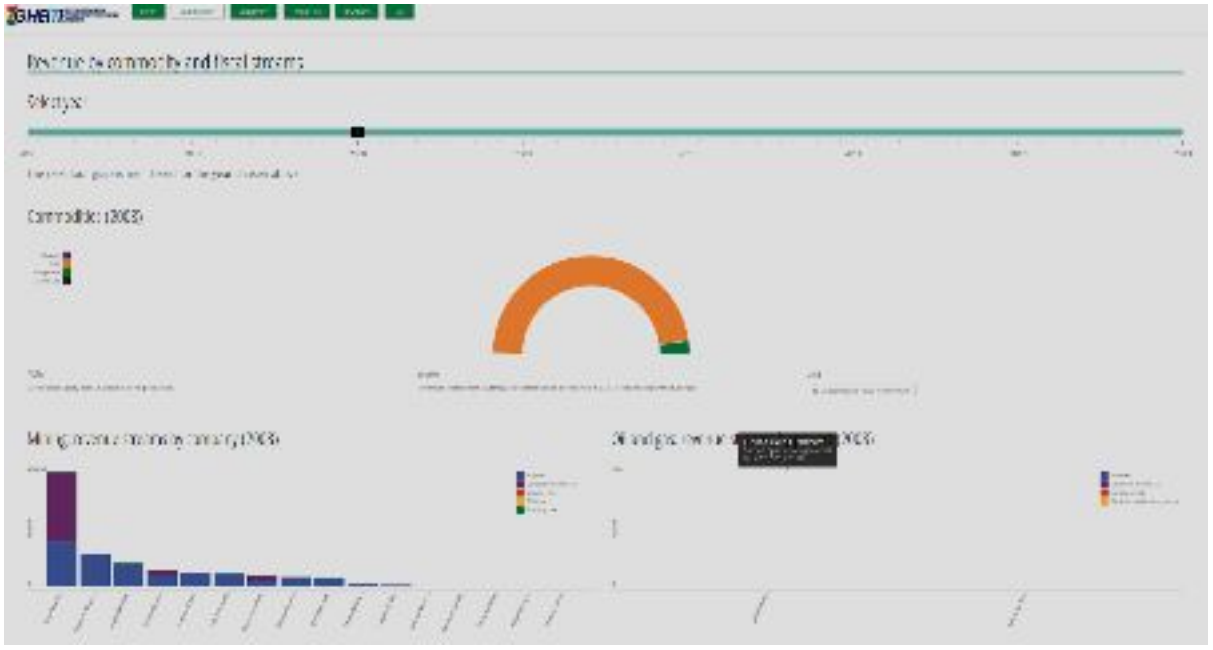
GotToVote! Kenya: Improving voter turnout with open data

Image from https://www.iebc.or.ke/

Seeking to Improve Voter Turnout in Kenya with Open Data: Kenya's national Independent Electoral and Boundaries Commission (IEBC) released information about polling center locations on its website in the lead up to Kenya's 2013 general election. The information, however, was difficult to access and reuse. Seizing on the gap between opening government data and citizens' actual ability to use that data, two Code 4 Kenya fellows conducted an experiment in unlocking government data to make it useful to the public. The fellows scraped the released IEBC data and built a simple website where it could be more easily accessed. The result was the initial version of GotToVote! a site that provided citizens with voter registration center information, and also helped them

navigate the sometimes complex world of registration procedures. This first version was developed in just 24 hours at minimum cost, garnered over 6,000 site visits in just its first week of existence, and has since been replicated across sub-Saharan Africa.

Extractives Dataset in Ghana



Revenue by commodity and fiscal streams (source: http://data.gheiti.gov.gh/#commodities)

Lesson 3: Data Formats, Frameworks and Terms

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

Portable Document Format (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)

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

Data Process Frameworks

The Data Pipeline

The Data Pipeline is School of Data's approach to working with data from beginning to end. Once you understand your action cycle and the stakeholders, it will be time to work with the data and we have broken down this process in steps. The steps are:

Define: 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?

Find: While the problem definition phase hints at what data is needed, finding the data is another step, of varying difficulty. There are a lot of tools and techniques to do that, ranging from a simple question on your social network, to using the tools provided by a search engine (such as Google search operators), open data portals or a Freedom of Information request querying about what data is available in that branch of government. This phase can make or break your project, as you can't do much if you can't find the data! But this is also where creativity can make a difference: using proxy indicators, searching in non-obvious locations... don't give up too soon!

Get: To get the data from its initial location to your computer can be short and easy or long and painful. Luckily, there's plenty of ways of doing that. You can crowdsource using online forms, you can perform offline data collection, you can use some crazy web scraping skills, or you could simply download the datasets from government sites, using their data portals or through a Freedom of Information request.

Verify: We got our hands in the data, but that doesn't mean it's the data we need. We have to check out if details are valid, such as the meta-data, the methodology of collection, if we know who organised the dataset and it's a credible source. We've heard a joke once, but it's only funny because it's true: all data is bad, we just need to find out how bad it is!

Clean: It's often the case the data we get and validate is messy. Duplicated rows, column names that don't match the records, values that contain characters which will make it difficult for a computer to process and so on. In this step, we need skills and tools that will help us get the data into a machine-readable format, so that we can analyse it. We're talking about tools like OpenRefine or LibreOffice Calc and concepts like relational databases.

Analyse: This is it! It's here where we get insights about the problem we defined in the beginning. We're gonna use our mad mathematical and statistical skills to interview a dataset like any good journalist. But we won't be using a recorder and a notebook. We can analyse datasets using many, many skills and tools. We can use visualisations to get insights of different variables, we can use programming languages packages, such as Pandas (Python) or simply R, we can use spreadsheet processors, such as LibreOffice Calc or even statistical suites like PSPP.

Present: And, of course, you will need to present your data. Presenting it is all about thinking of your audience, the questions you set out to answer and the medium you select to convey your message or start your conversation. You don't have to do all by yourself, it's good practice to get support from professional designers and storytellers, who are experts at understanding what are the best ways to present data visually and with words.

Glossary

Anonymisation

The process of treating data such that it cannot be used for the identification of individuals.

Application Programming Interface (API)

A way computer programmes talk to one another. Can be understood in terms of how a programmer sends instructions between programmes.

Attribution Licence

A licence that requires attributing the original source of the licensed material.

BitTorrent

BitTorrent is a protocol for distributing the bandwith for transferring very large files between the computers which are participating in the transfer. Rather than downloading a file from a specific source, BitTorrent allows peers to download from each other.

Boolean logic

A form of algebra in which all values are reduced to either TRUE or FALSE.

Categorical Data

Data that helps put things into categories. E.g.: Country names, Groups, Conditions, Tags

Choropleth Map

A choropleth map is a map where value are encoded onto regions using colormapping. The whole region is colored using the underlying value.

Continuous Data

Numerical data that, if you plot all possible values, has no gaps. E.g. Sizes (you can be 155.55 or 155.56cm tall etc.) Compare to *Discrete Data*

Crowdsourcing

Mashup of crowd and outsourcing: Having a lot of people do simple tasks to complete the whole work.

Comma Separated Values (CSV)

Comma Separated Values. A very simple, open format for tabular data which can be exported and imported by all spreadsheet applications and is easily manipulable with command line tools.

curl

http://curl.haxx.se/ – a command line tool for transferring data to and from online systems over standard internet protocols including FTP and HTTP. Very powerful and great for working with *Web API* s from the command line.

Data Access Protocol (DAP)

A system that allows outsiders to be granted access to databases without overloading either system.

Discrete Data

Numerical Data that, if you plot all possible values, has gaps in it. E.g. the count of things (there are no 1.5 children). Compare to *Continuous Data*

etherpad

A piece of software for collaborative real-time editing of text. See http://etherpad.org/.

GDP

Gross domestic product (GDP) is the market value of all officially recognized goods and services produced within a country in a given period of time. GDP per capita is often considered an indicator of a country's standard of living. (Source: Wikipedia.)

"**Geocoding**

From Geographical Coding. Describes the practice of attaching geographical coordinates to items.

GeoJSON

GeoJSON is a format for encoding a variety of geographic data structures. It is based on the JSON specification. More documentation can be found on http://www.geojson.org

Intellectual property rights

Monopolies granted to individuals for intellectual creations.

IP rights

See *Intellectual property rights*.

JSON

JavaScript Object Notation. A common format to exchange data. Although it is derived from Javascript, libraries to parse JSON data exist for many programming languages. Its compact style and ease of use has made it widespread. To make viewing JSON in a

browser easier you can install a plugin such as JSONView in Chrome and JSONView in Firefox.

Machine-readable

Formats that are machine readable are ones which are able to have their data extracted by computer programs easily. PDF documents are not machine readable. Computers can display the text nicely, but have great difficulty understanding the context that surrounds the text. Common machine-readable file formats are CSV and Excel Files.

Mean

The arithmetic mean of a set of values. Calculated by summing up all values and then dividing by the number of values.

Median

The median is defined as the value where 50% of values in a range will be below, 50% of values above the value.

Normal Distribution

The normal (or Gaussian) distribution is a continuous probability distribution with a bell shaped curve.

Open Data

Open data is data that can be used, reused and redistributed freely by anyone for any purpose. More details can be found at at opendefinition.org.

Open standards

Generally understood as technical standards which are free from licencing restrictions. Can also be interpreted to mean standards which are developed in a vendor-neutral manner.

Percentiles

Percentiles are a value where n% of values are below in a given range. e.g. the 5th percentile: 5 percent of values are lower than this value.

Public domain

No copyright exists over the work. Does not exist in all jurisdictions.

Qualitative Data

Qualitative data is data telling you something about qualities: e.g. description, colors etc. Interviews count as qualitative data

Quantitative Data

Quantitative data tells you something about a measure or quantification. Such as the quantity of things you have, the size (if measured) etc.

Quartiles

Quartiles are the values where 25, 50 and 75% of values in a range are below the given value.

Readme

A file (usually named README or README.txt) that explains new users what the current directory or set of files is about. This is very commonly found in open source software projects and is considered good practice to be included with various publications (including datasets). The file usually contains a short description of what to expect.

Scraping

The process of extracting data in *machine-readable* formats of non-pure data sources e.g.: webpages or PDF documents. Often prefixed with the source (web-scraping PDF-scraping).

Share-alike Licence

A licence that requires users of a work to provide the content under the same or similar conditions as the original.

Tab-separated values

Tab-separated values (TSV) are a very common form of text file format for sharing tabular data. The format is extremely simple and highly *machine-readable*.

Taxonomy

Classification. Taxonomy refers to hierarchical classification of things. One of the best known is the Linnean classification of species – still used today to classify all living beings.

Web API

An API that is designed to work over the Internet.

Extra Reading

1. Open Data Handbook: What is Open Data?
2. Defining Open Data
3. Open Data in Developing Economies: Toward Building an Evidence Base on What Works and How
4. The Data Pipeline: School of Data's Methodology