DATA INSIGHTS

WORKSHOP ON OPEN DATA & ODE

- Frictionless Data
- Open Data Formats
- Navigate the ODE Interface
- Prepare Data for Analysis

SUPPORTED BY

Open Knowledge Foundation

ORGANISED BY

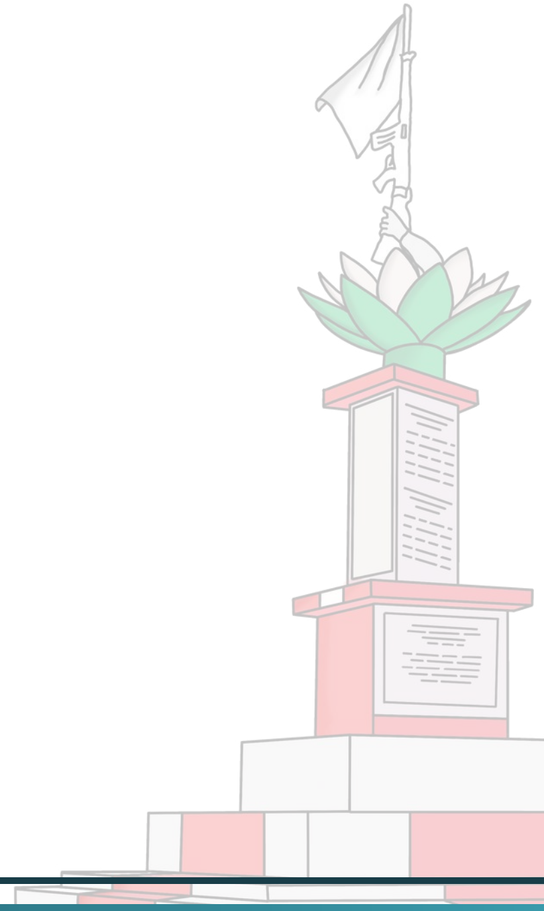
BAUCC

CO-ORGANISED BY

KORIKATH KNOWLEDGE

DAY ONE

Open Source and Open Data



AFTER THIS SESSION

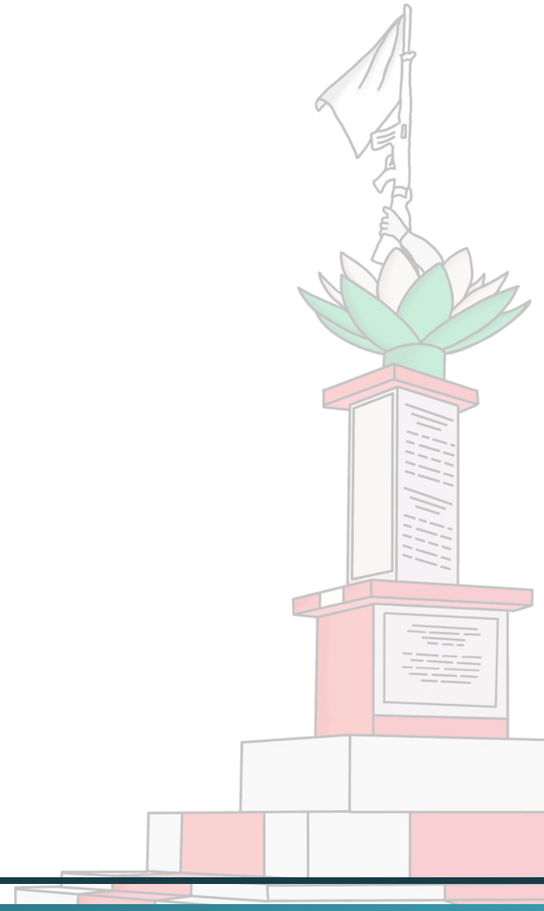
- We will understand the rich history of the open source movement.
- We will describe how open data affects our society.
- We will have a basic idea of the standards used in the open data field.
- We will be able to do a rough evaluation of a dataset's quality.

TIME BREAKDOWN

Intro	20 minutes
History of Open Source	20 minutes
Break 1	10 minutes
Open Licenses	20 minutes
Open Data	40 minutes
The world of Data	40 minutes
Break 2	20 minutes
Open Data Tools	20 minutes
Q&A	20 minutes

INTRO

Who on the earth are we?

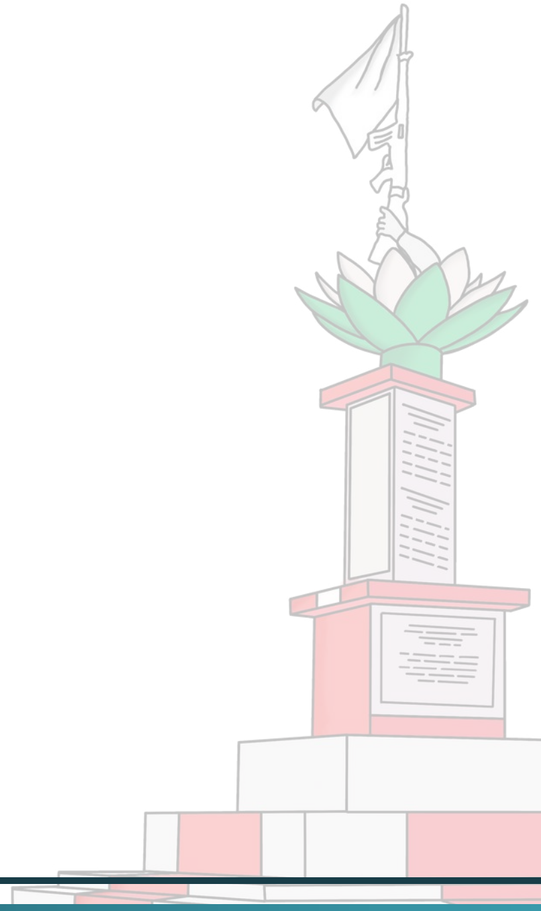


SELF BRAGGING

- Google “Mrb Rafi” 😜
- Or get connected over LinkedIn

WHO ARE "WE"

- BAUCC
- Korikath Knowledge
- Open Knowledge Foundation



HISTORY OF OPEN SOURCE

HACKERS!

- Richard Stallman
- MIT AI lab
- Hacker ethoes
- The GNU Project
- GNU GPL

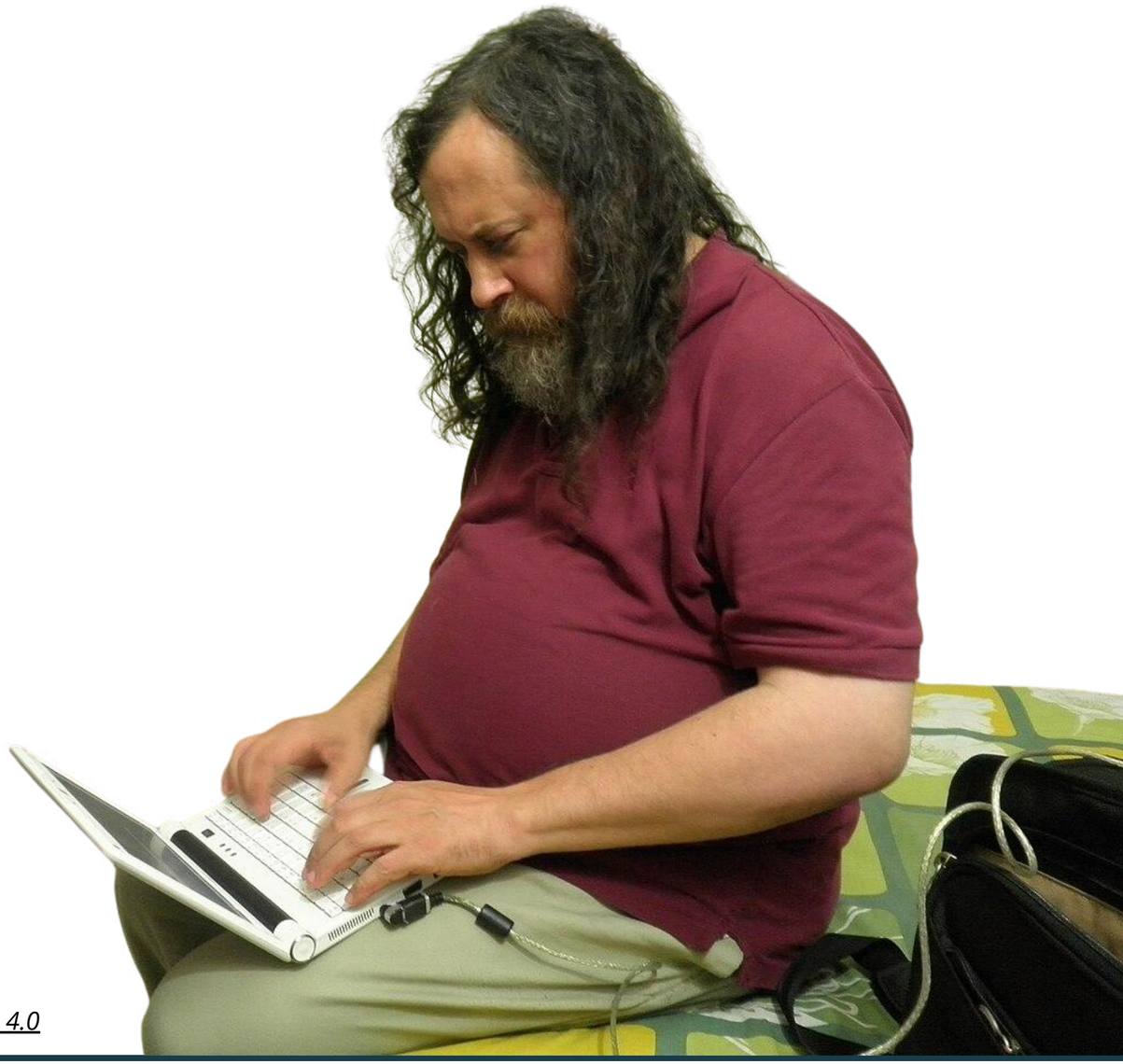


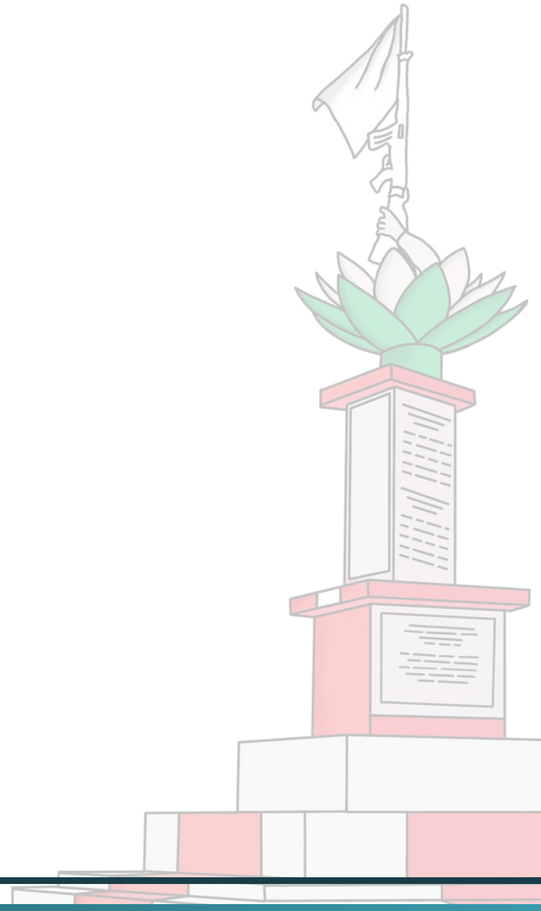
Image: Outofindia, CC BY SA 4.0

HAVE A READ!

- উইকিমিডিয়া: যেভাবে উন্মুক্ত লাইসেন্সের আওতায় বিশ্বজুড়ে জ্ঞান হলো উন্মুক্ত - বিজ্ঞানচিন্তা, October 30, 2024

BREAK 1

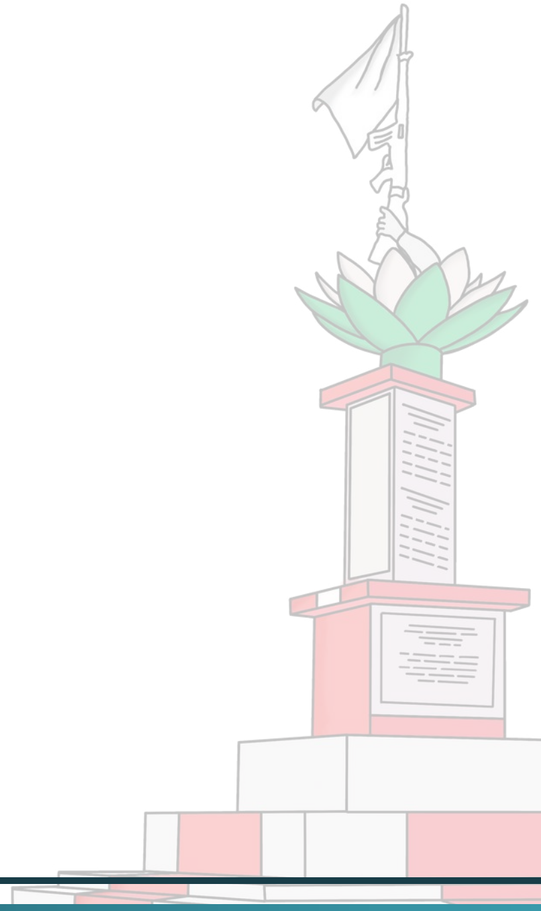
Only 10 minutes!



OPEN LICENSES

DIFFERENT OPEN LICENSES

- Creative Commons and its types
- GNU licenses
- The MIT license
- Open Data Commons - EU Sui Generis Rights

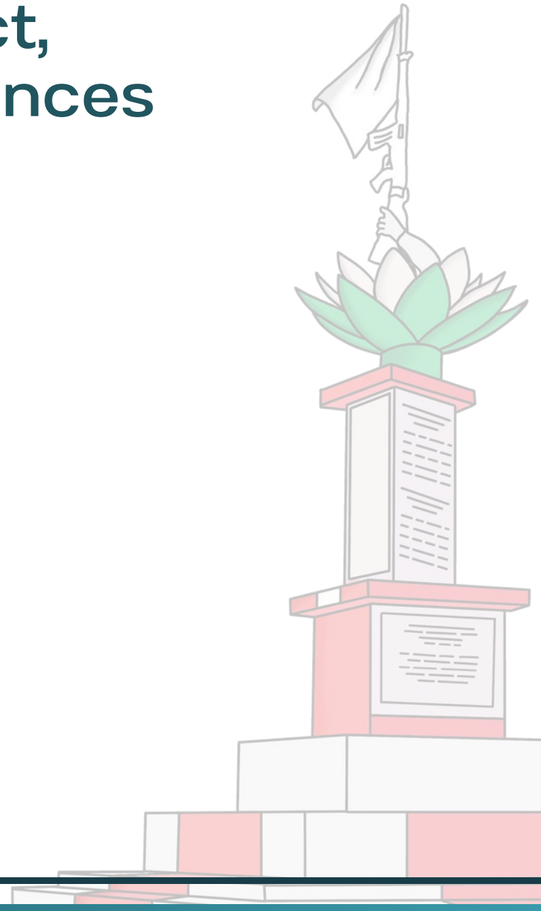


OPEN DATA

DATA

Data are a collection of discrete or continuous values that convey information, describing the quantity, quality, fact, statistics, other basic units of meaning, or simply sequences of symbols that may be further interpreted formally.

-Data (Wikipedia)



OPEN DATA

Open data are data that are openly accessible, exploitable, editable and shareable by anyone for any purpose. Open data are generally licensed under an open license.

-Open Data, Wikipedia

THE OPEN DEFINITION

"Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)."

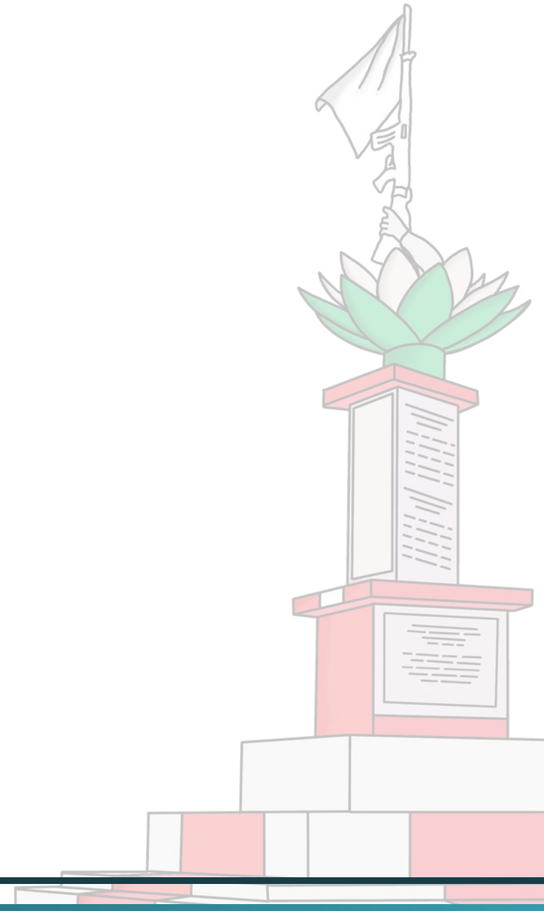
A project by the Open Knowledge Foundation to define and clarify the meaning of "open" in different knowledge context

FOUR FREE SOFTWARE FREEDOM

- Access
- Use
- Modify
- Share

MACHINE READABLE LICENSE

```
{
  "NAME": "MY-DATASET",
  "TITLE": "MY SAMPLE DATASET",
  "DESCRIPTION": "THIS DATASET CONTAINS SAMPLE DATA FOR DEMONSTRATION PURPOSES.",
  "LICENSES": [
   {
   "NAME": "CC-BY-4.0",
   "TITLE": "CREATIVE COMMONS ATTRIBUTION 4.0",
   "PATH": "HTTPS://CREATIVECOMMONS.ORG/LICENSES/BY/4.0/"
   }
   ],
  "RESOURCES": [
   {
   "NAME": "DATA",
   "PATH": "DATA.CSV",
   "FORMAT": "CSV",
   "MEDIATYPE": "TEXT/CSV",
   "SCHEMA": {
   "FIELDS": [
   { "NAME": "ID", "TYPE": "INTEGER" },
   { "NAME": "NAME", "TYPE": "STRING" }
   ]
   }
   }
   ]
  }
```



FIVE STARS OF OPEN DATA

Tim Berners-Lee, the inventor of the Web and initiator of the Linked Data project, suggested a 5 star deployment scheme for Linked Data. The 5 Star Linked Data system is cumulative. Each additional star presumes the data meets the criteria of the previous step(s).
-5 Star Linked Data, W3C

FIVE STARS OF OPEN DATA

★: Make your data available on the web in any format under an open license.

★★: Provide the data as structured data (an Excel instead of a scanned image of a table).

★★★: Use a non-proprietary open format (CSV instead of Excel).

★★★★: Use URIs to identify things, enabling others to link directly to your data.

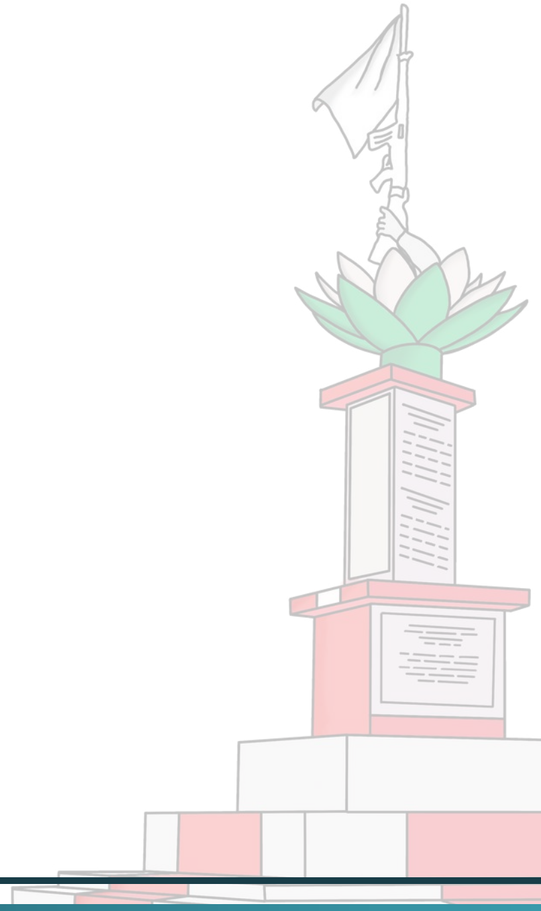
★★★★★: Link your data to other datasets to provide context.

HANDY TERMS

OL: Open License
RE: machine REadable
OF: Open Format
URI: Uniform Resource Identifiers
LD: Linked Data
CC: Creative Commons
ODbL: Open Data Commons Open Database License

MICRODATA

In the study of survey and census data, microdata is information at the level of individual respondents.
-Microdata (statistics), Wikipedia



EXCEL EMBARASSMENT

Source: James Vincent, The Verge, Oct 5, 2020

The Verge / Tech / Reviews / Science / Entertainment / AI /

SCIENCE / HEALTH / CORONAVIRUS

Excel spreadsheet error blamed for UK's 16,000 missing coronavirus cases

/ The case went missing after the spreadsheet hit its filesize limit

Illustration by Alex Castro / The Verge

by James Vincent
Oct 5, 2020, 7:41 PM GMT+5

Comments

James Vincent is a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

The UK failed to add nearly 16,000 confirmed cases of coronavirus to its national track and trace system due to an Excel error. A number of reports, including from The Guardian, Sky News, and The Daily Mail, say the mistake was caused when an Excel spreadsheet used to track confirmed cases of the virus reached its maximum file size and failed to update.

The government agency that helps oversee the UK's pandemic response, Public Health England (PHE), said some 15,841 cases had been left out of national totals because of the error but did not specify what caused the so-called glitch.

> "The issue was caused by the fact that some files containing positive test results exceeded their maximum file size"

"Initial findings indicate that the issue was caused by the fact that some files containing positive test results exceeded their maximum file size, and they then failed to load into the central system because they couldn't get through," a spokesperson for PHE told BBC News journalist Leo Kelion. The spokesperson did not confirm or deny that Excel was involved.

The revelation caps a weekend of soaring coronavirus cases for the UK, which many have speculated could lead to a second national lockdown for the country: 12,782 new cases were announced on Saturday followed by 22,961 on Sunday.

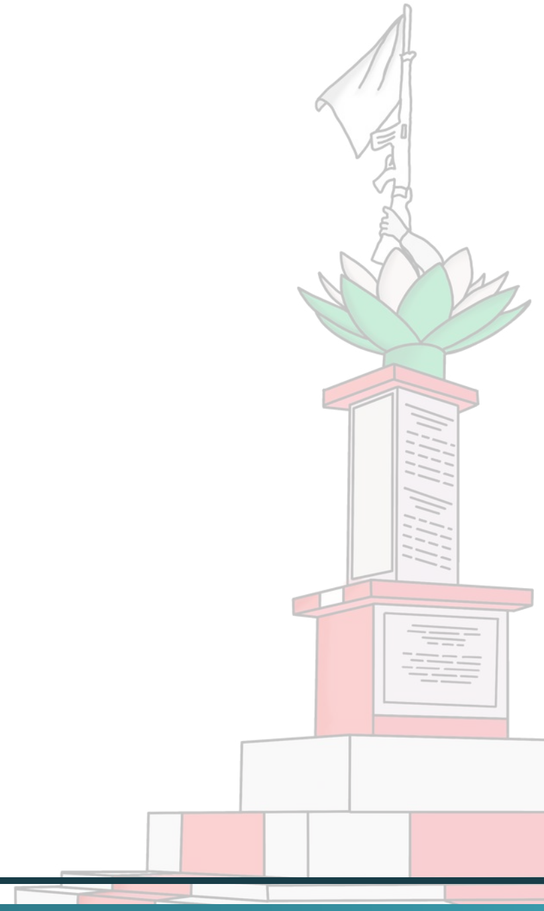
The 15,841 "missing" cases made public today were originally recorded between September 25th and October 2nd. All those who tested positive for COVID-19 were notified by the UK's health authorities, but the failure to upload these cases to the national database meant anyone who came into contact with these individuals was ... error that may have helped ...

MOST POPULAR

1. The Pro 3 is 8BitDo's first controller with swappable buttons
2. Sony's pocket-sized RX1R camera returns with its first update in 10 years
3. Elon Musk's AI bot adds a ridiculous anime companion with 'NSFW' mode
4. Superman is a box office hit, but the hard part comes next
5. Google's Pixel 8 Pro is $599, and includes a free Pixel Watch 2

Most Popular

THE WORLD OF DATA



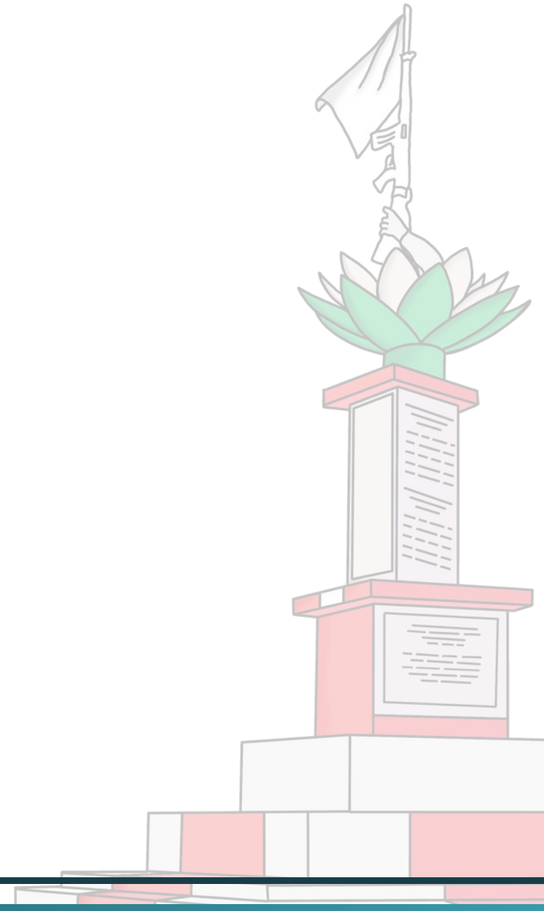
STRUCTURED DATA

Data that conforms to a fixed schema

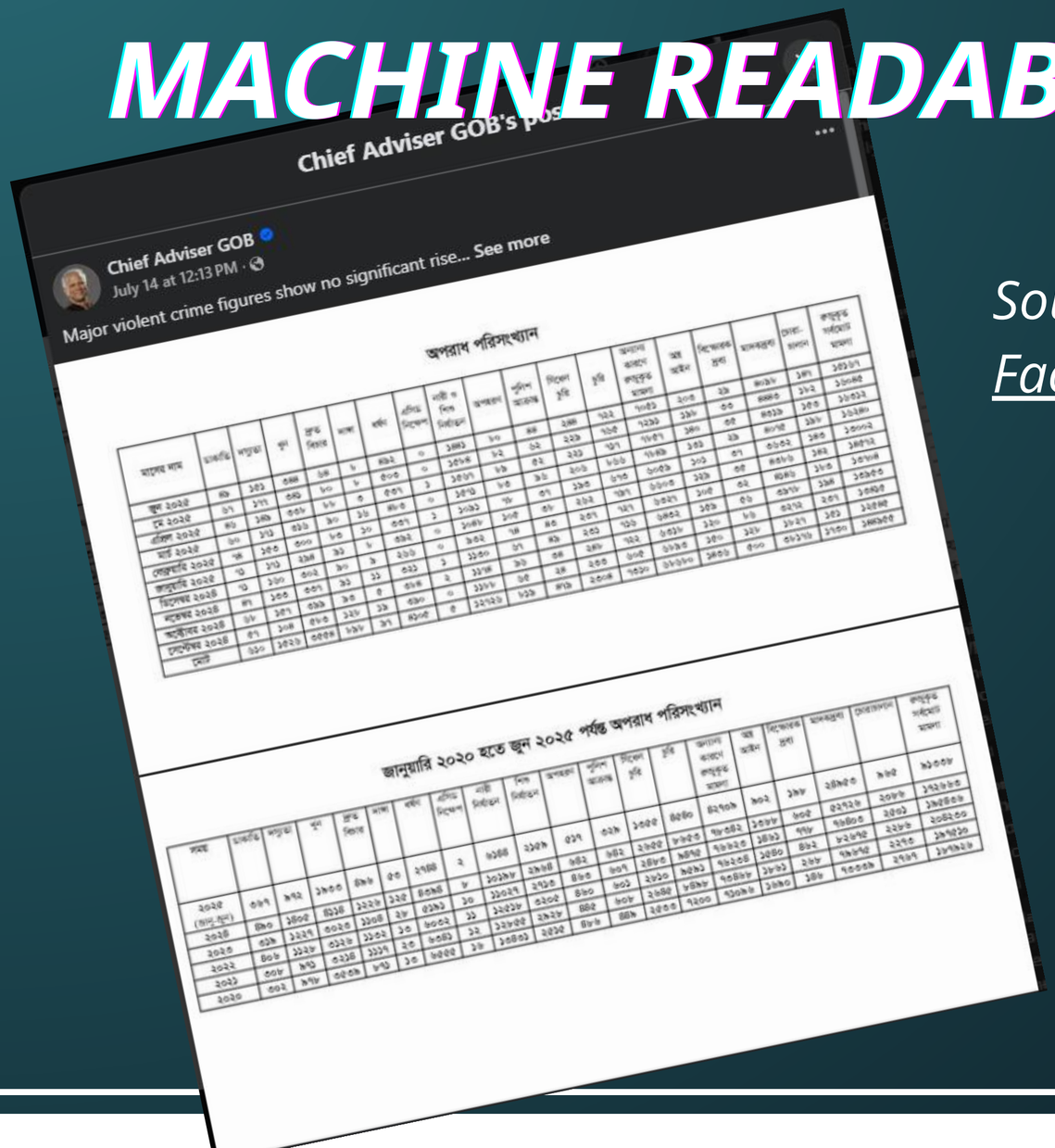
MACHINE READABLE DATA

As the name implies - data that can be read by the machines.

Example: csv, json, xlsx, shp



MACHINE READABLE OR NOT



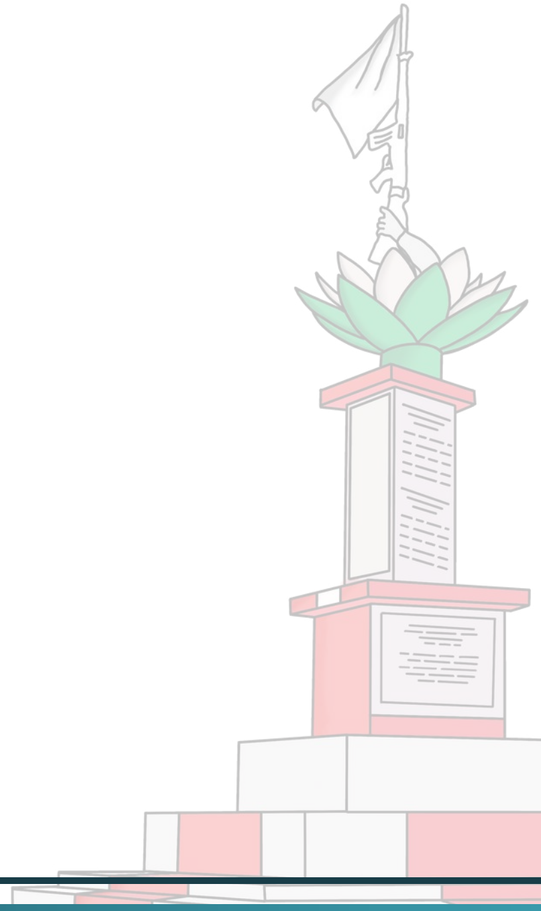
Source: Chief Advisor GOB, July 14, 2025, Facebook

MACHINE READABLE OR NOT [CONTINUED]

More such examples

- Crime rate stats by BD police
- Population and Housing census data, BBS

I tried automated converters btw, but badluck :(



TABULAR DATA

Data that conforms to a fixed schema

Crime Statistics in June 2025

Names of Unit	Dacoity	Robbery	Murder	Speedy Trial	Riot	Woman & Child Repression	Kidnapping	Police Assault	Burglary	Theft	Other Cases	Recovery Cases				
												Arms Act	Explosive Act	Narcotics	Smuggling	Total
DMP	4	32	49	8	4	168	20	6	55	90	456	45	4	505	10	564
CMP	0	7	7	9	0	34	3	1	15	36	100	5	0	133	0	138
KMP	0	4	3	1	0	14	0	1	8	14	35	7	1	51	2	61
RMP	0	7	0	0	0	19	3	0	3	14	63	0	0	57	1	58
BMP	0	2	2	0	1	15	1	1	2	5	48	0	0	38	0	38
SMP	0	2	2	5	0	12	0	0	4	5	50	1	1	25	4	31
RPMP	0	1	2	0	2	11	2	0	0	7	28	4	0	20	0	24
GMP	1	5	8	0	0	24	6	3	7	12	76	1	1	57	0	59
Dhaka Range	22	37	82	9	0	318	17	8	50	104	1233	36	3	711	5	755
Mymensingh Range	1	6	19	5	0	133	2	1	8	39	536	6	0	276	36	318
Chittagong Range	8	13	53	11	1	246	10	9	43	98	1118	41	2	718	31	792
Sylhet Range	2	4	20	3	0	92	2	5	5	52	371	2	0	114	26	142
Khulna Range	5	9	28	9	0	156	6	4	19	61	651	19	4	393	16	432
Barishal Range	1	3	7	4	0	169	2	2	12	36	552	2	1	134	0	137
Rajshahi Range	3	11	38	0	0	230	4	3	8	56	890	28	12	441	12	493

ENCODING

- Text is made up of words and sentences, which are created from **characters** (letters, symbols etc.)
- **Character encoding** is a system that maps characters to bytes in computer memory
- This mapping is essential for computers to **process**, **store**, and **display text** correctly
- Think of it as **a key to unlock the code**

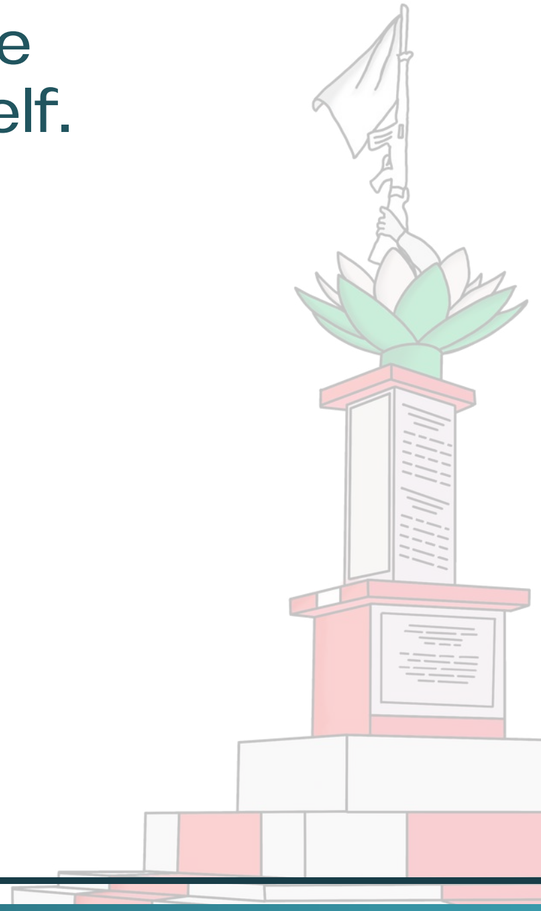
ENCODING

- ASCII: A basic encoding standard for English text
- UTF-8: A Unicode encoding that can handle any character and is widely used on the web
- ISO-8859: A family of encodings for different language scripts

METADATA

Metadata (or metainformation) is "data that provides information about other data",[1] but not the content of the data itself, such as the text of a message or the image itself.
-Metadata, Wikipedia

- Data Dictionary

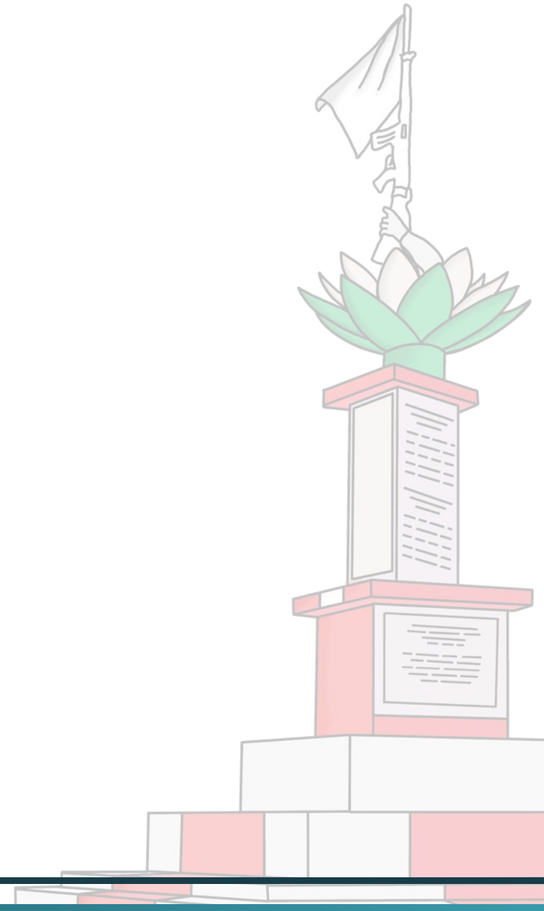


DATA GOVERNANCE

- A system of principles, practices and tools used to manage data assets across their entire lifespan
- It is more managing data, it is about making data accessible, usable and secure
- Data governance aligns data-related needs with business strategy
- It enables better data management, quality, visibility, security and compliance

BREAK 2

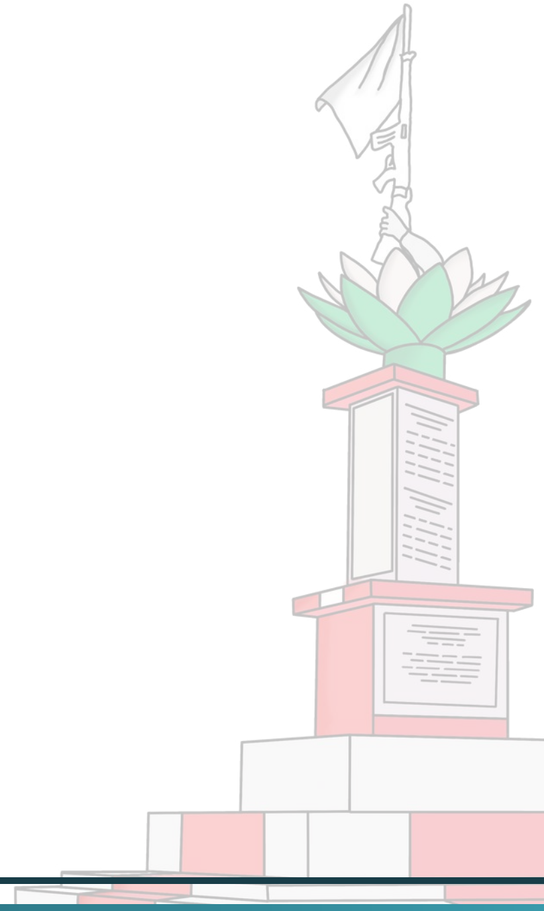
Only 20 minutes!



OPEN DATA TOOLS

OUR “OPEN” FRIENDS IN OPEN DATA

- Data Curator
- Openrefine
- Open Data Editor



BEFORE THE NEXT SESSION

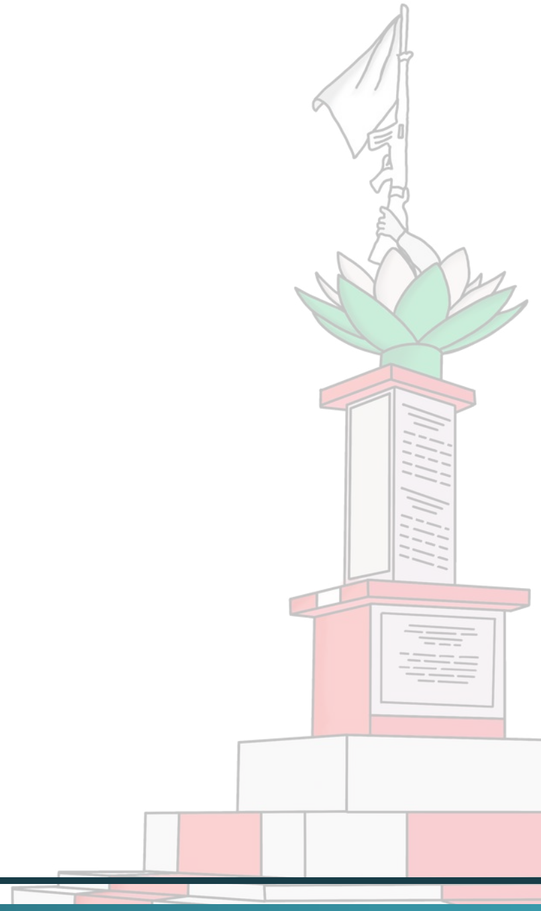
Please download the Open Data Editor on your PC using the following links.

- Windows
- Linux
- Mac

And feel free to ask questions at rafi.2109016@bau.edu.bd

Q&A

20 minutes



THE END [FOR TODAY]

Thank God!