Module 1 - Session 2 - Data Collection

Working effectively with data

CivicDataLab

2021/07/24 (updated: 2021-07-26)

Session 1 - Recap

What we discussed

Session 1 - Recap

What we discussed

- Data Collection - How to (Primary vs Secondary)

Session 1 - Recap

What we discussed

- Data Collection - How to (Primary vs Secondary)
- Collecting data from secondary sources

Session 1 - Recap

What we discussed

- Data Collection - How to (Primary vs Secondary)
- Collecting data from secondary sources
- The data biography (A must when using data from secondary data sources)

Session 1 - Recap

What we discussed

- Data Collection - How to (Primary vs Secondary)
- Collecting data from secondary sources
- The data biography (A must when using data from secondary data sources)
- Structured vs Unstructured data sources (Limited but easy to analyse vs Vast but hard to work with)

Session 1 - Recap

- Collecting (structured) data from secondary sources

Session 1 - Recap

- Collecting (structured) data from secondary sources
- Working with CSV files (Quite prone to certain data collection errors)

Session 1 - Recap

- Collecting (structured) data from secondary sources
- Working with CSV files (Quite prone to certain data collection errors)
- Extracting data from PDF files (Working with Tabula)

Session 1 - Recap

- Collecting (structured) data from secondary sources
- Working with CSV files (Quite prone to certain data collection errors)
- Extracting data from PDF files (Working with Tabula)
- Collecting data from the web (Ethical web scraping)

Session 1 - Recap

- Collecting (structured) data from secondary sources
- Working with CSV files (Quite prone to certain data collection errors)
- Extracting data from PDF files (Working with Tabula)
- Collecting data from the web (Ethical web scraping)
- Using google sheets to scrape data from web pages

> Resources from all sessions can be accessed at this link

Session 2 - Introduction

- Common issues with real-world datasets and how to resolve them
- Data Collection - Supreme Court Judges
- Data scraping alternatives
- Data Standards
  - What
  - Why
  - What can be standardised
  - Examples
- Exploring data standards
  - Popolo
  - Akoma Ntoso
  - ICCS
  - GTFS
- Data standards - Other resources
- Data exploration - State wise mortality data from Devdatalab
- The Tyranny of Spreadsheets

Public data is messy

Here is a guide that will help you in dealing with a few common problems that we associate with datasets.

> Most of these problems can be solved. Some of them can't be solved and that means you should not use the data. Others can't be solved, but with precautions you can continue using the data. In order to allow for these ambiguities, this guide is organized by who is best equipped to solve the problem: you, your source, an expert, etc. In the description of each problem you may also find suggestions for what to do if that person can't help you. [1]

[1] The Quartz guide to bad data

Data collection use-cases

Supreme Court Judge Profiles

Website - https://main.sci.gov.in/chief-justice-judges

Requirement - Get judge profile details in a structured format

Challenges:

1. Direct download not available
2. Judge profile hidden behind a link
3. Current and Former judges are on different tabs but on the same URL

Data scraping Alternatives

1. Crowd sourced data collection *More work for humans*

Data scraping Alternatives

1. Crowd sourced data collection *More work for humans*
2. Training bots using AI *Challenge with implementation and accuracy of information*

Data scraping Alternatives

1. Crowd sourced data collection *More work for humans*
2. Training bots using AI *Challenge with implementation and accuracy of information*
3. Creating data standards *More work for data publishers*

Data scraping Alternatives

1. Crowd sourced data collection *More work for humans*
2. Training bots using AI *Challenge with implementation and accuracy of information*
3. Creating data standards *More work for data publishers*

> If your goal is to build something that's going to be collecting data for a long time, like a government monitoring tool, think about how to identify, rally around, and target common standards. There's some momentum for building shared code on top of shared data, and we're starting to see real success stories. And so my closing message is to think long-term about how you collect the data; don't plan on scraping forever. [1]

[1] Monitoring Legislatures: The Long Game

Open data standards - What

An open data standard is a set of specifications (or requirements) for how some sets of data should be made publicly available. [1]

[1] Open data standards directory

Open data standards - Why

- Scraping is unsustainable

Open data standards - Why

- Scraping is unsustainable
- Machine-readable data is not enough

Open data standards - Why

- Scraping is unsustainable
- Machine-readable data is not enough
- To standardise data collection across time and space - ***A shared language*** that allows error free data exchange

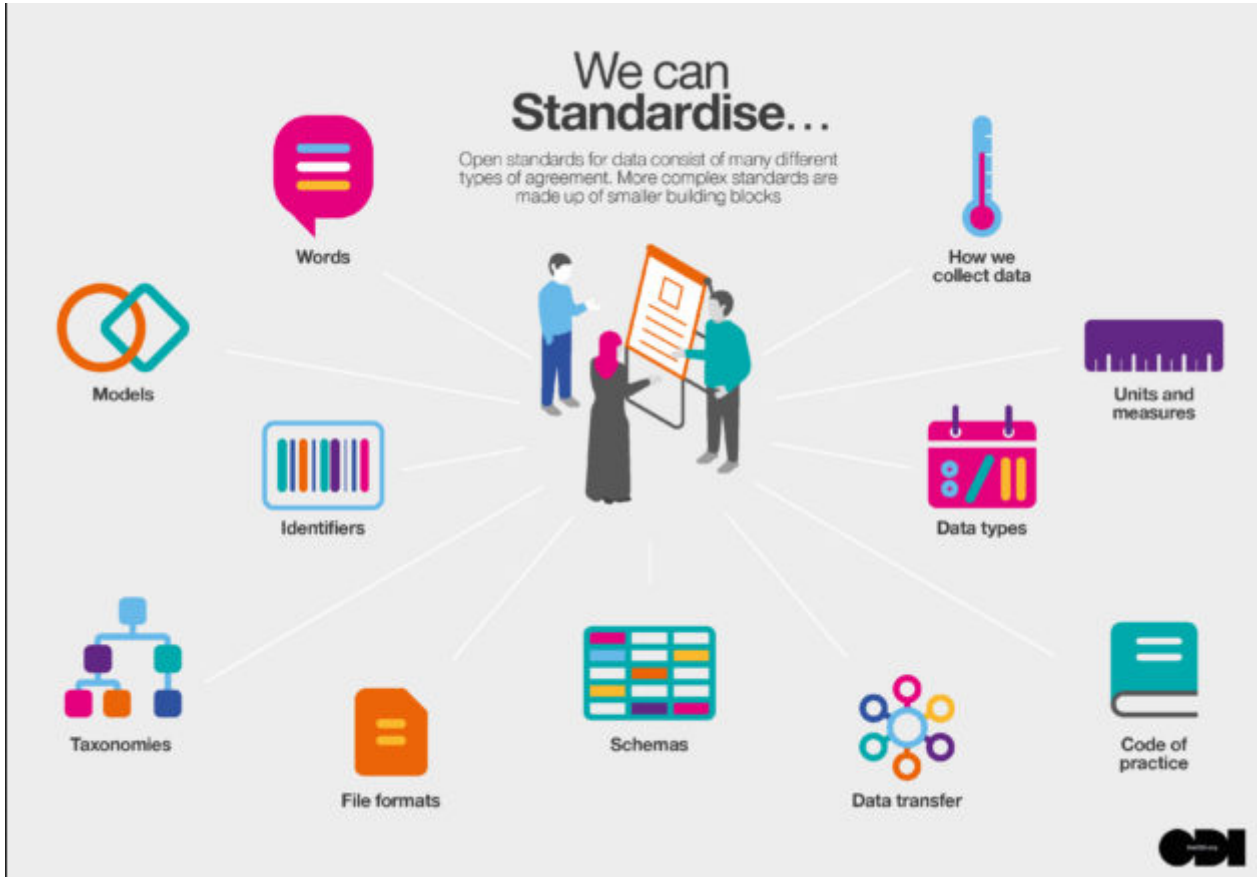
Open data standards - Why

- Scraping is unsustainable
- Machine-readable data is not enough
- To standardise data collection across time and space - ***A shared language*** that allows error free data exchange
- For generating more interoperable datasets

Open data standards - Why

- Scraping is unsustainable
- Machine-readable data is not enough
- To standardise data collection across time and space - ***A shared language*** that allows error free data exchange
- For generating more interoperable datasets
- Opportunity for a dialogue between consumers and producers of datasets

What can be standardised



Source: The Open Data Institute

Open data standards - Examples

civic data lab

Legislature | Crime | Transit | Finance

Title	Description
Popolo	The standard was designed as a simple and transferable vocabulary for sharing international government open data
Akoma Ntoso	Akoma Ntoso introduces common structures and ontologies for parliamentary, legislative and judiciary documents. These include parliamentary debates, committee briefs, and the entire life-cycle of a piece of legislation

Open data standards - Examples

civic data lab

Legislature | Crime | Transit | Finance

Title	Description
International Classification of Crime for Statistical Purposes (ICCS)	The ICCS provides a comprehensive framework for producing statistics on crime and criminal justice.

Open data standards - Examples

civic data lab

Legislature | Crime | Transit | Finance

Title	Description
General Transit Feed Specification (GTFS)	GTFS allows public transportation agencies to provide application developers with real-time updates about the locations, estimated arrival times and other important information regarding transit vehicles.

Open data standards - Examples

civic data lab

Legislature | Crime | Transit | Finance

Title	Description
Open Contracting Data Standard (OCDS)	For disclosing public procurement data in open formats about contracting processes from planning to implementation stage.

Exploring Data Standards

Popolo

International open government data specifications

Specification - https://www.popoloproject.com/

- For storing data related to elected officials
- Platforms like EveryPolitician that have data for elected officials from 233 countries use the format for storing info
- OpenAustralia developed TheyVoteforyou, a legislative vote tracking tool, to import Popolo data

Link to explore data for the members of the 16th Lok Sabha.

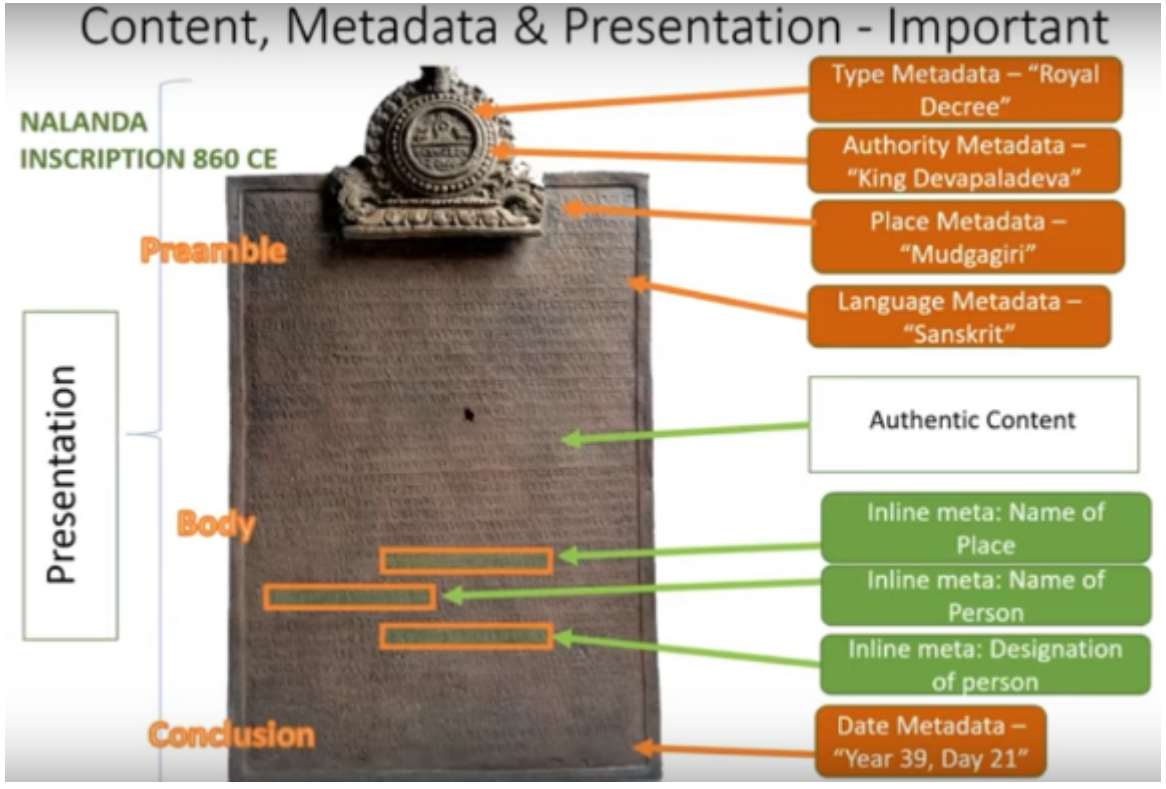
Akoma Ntoso

Architecture for **K**nowledge-**O**riented **M**anagement of **A**frican **N**ormative **T**exts using **O**pen **S**tandards and **O**ntologies - *A data standard for parliamentary, legislative and judiciary documents*

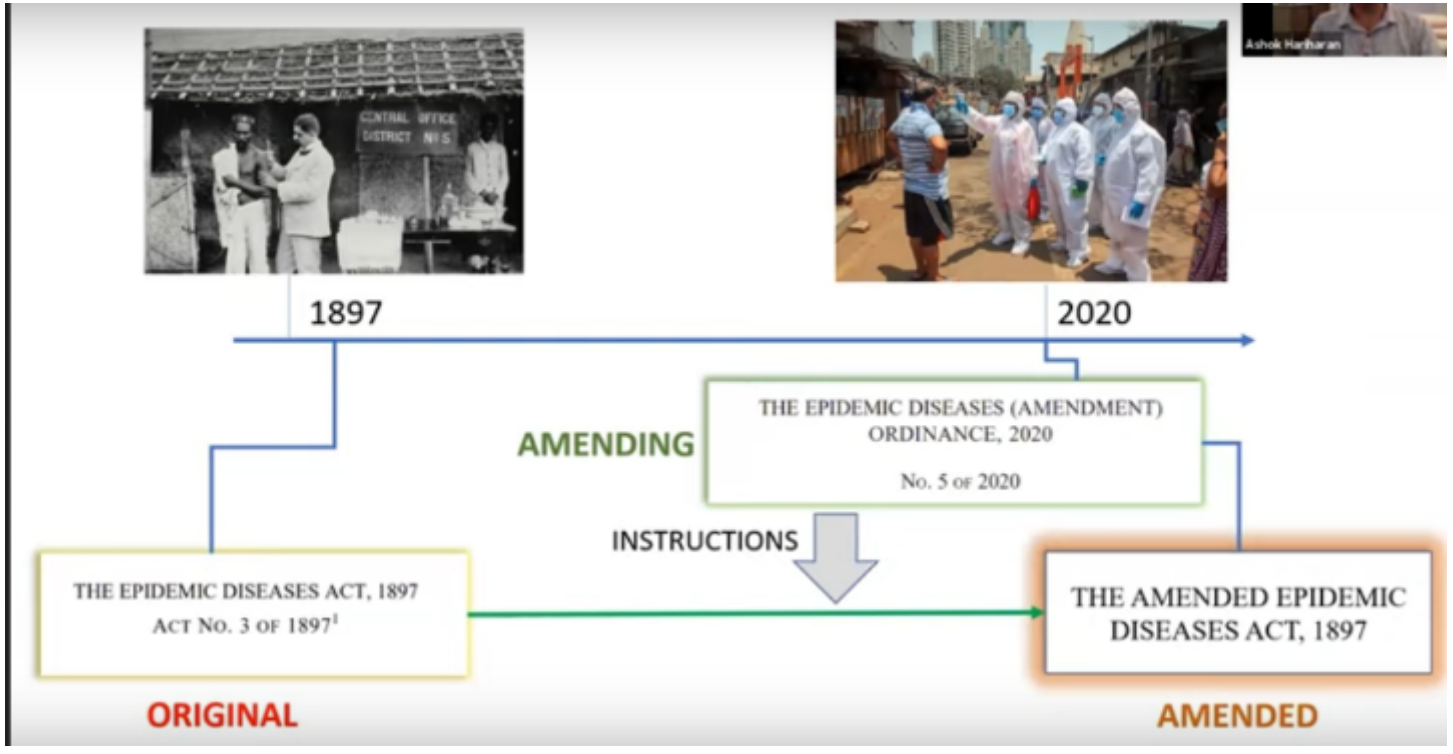
Specification - Akoma Ntoso

- Akoma Ntoso is an initiative of Africa i-Parliament Action Plan, a programme of UN/DESA.
- Schema explorer - Link

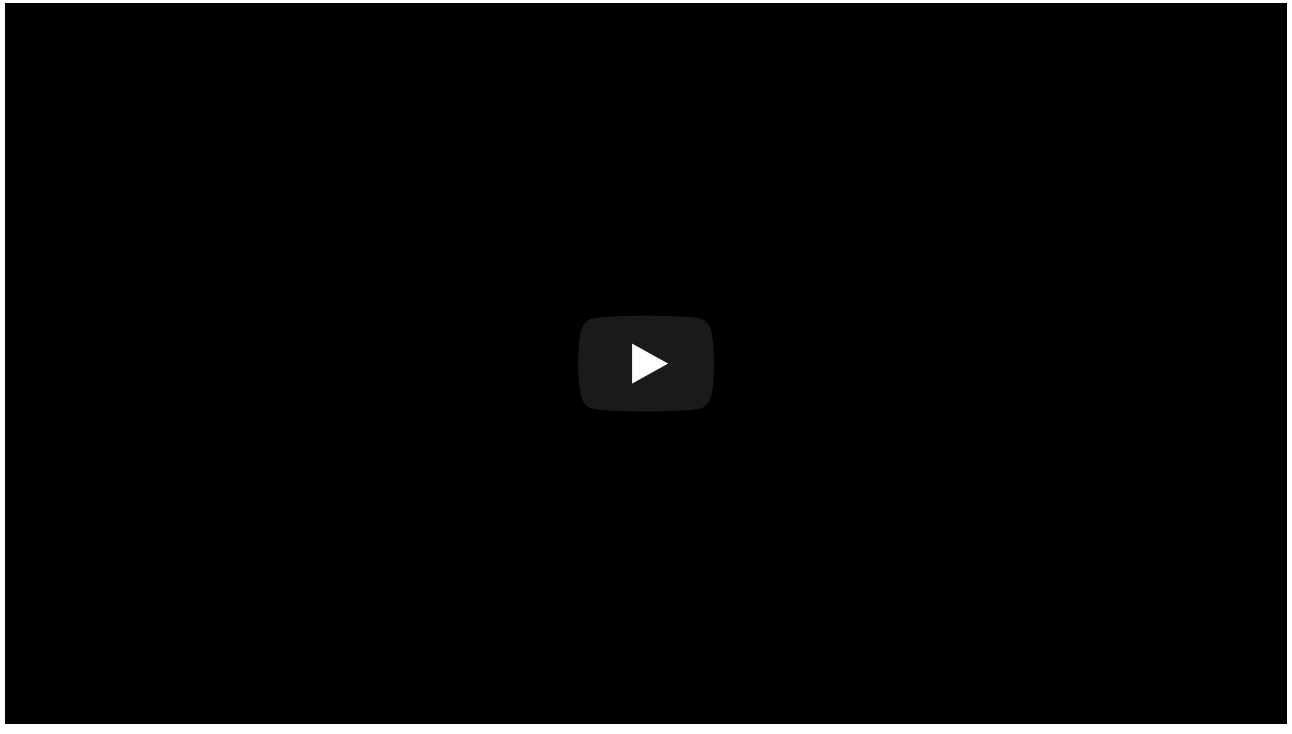
Semantics


Content, Metadata & Presentation - Important
NALANDA
INSCRIPTION 860 CE
Preamble
Presentation
Body
Conclusion
Type Metadata – "Royal Decree"
Authority Metadata – "King Devapaladeva"
Place Metadata – "Mudgagiri"
Language Metadata – "Sanskrit"
Authentic Content
Inline meta: Name of Place
Inline meta: Name of Person
Inline meta: Designation of person
Date Metadata – "Year 39, Day 21"

Use-Case


CENTRAL OFFICE
DISTRICT NHS
1897
2020
AMENDING
THE EPIDEMIC DISEASES (AMENDMENT) ORDINANCE, 2020
No. 5 OF 2020
INSTRUCTIONS
THE EPIDEMIC DISEASES ACT, 1897
ACT NO. 3 OF 1897[1]
THE AMENDED EPIDEMIC DISEASES ACT, 1897
ORIGINAL
AMENDED

Video - The Akoma Ntoso Open Standard



HasgeekTV - Open standards for law courts in Africa

Akoma Ntoso - Reading List

1. An Open Platform for Laws: How adoption of open standards and tools can strengthen the rule of law
2. Are Indian laws really ‘open’?
3. How The Law Factory turns the French parliamentary process into 300 version-controlled Open Data visualizations
4. Laws of India in the Akoma Ntoso format
5. Indigo platform deployed by IndianKanoon to manage the XML laws generated by the Nyaaya team
6. AkomaNtoso.io - *A resource on learning and using the Akoma Ntoso schema*
7. A presentation on why XML is important in the context of parliamentary documents

ICCS

International Classification of Crime for Statistical Purposes (ICCS) - *To enhance the consistency and international comparability of crime statistics, and improve analytical capabilities at both the national and international levels*

Specification - ICCS Version 1.0 PDF

An attempt by the UN to adopt a common data standard to respond to emerging data needs at national and international level, including data needs deriving from the Sustainable Development Goals (SDGs) in the areas of crime, violence, justice and the rule of law under UNODC mandate.

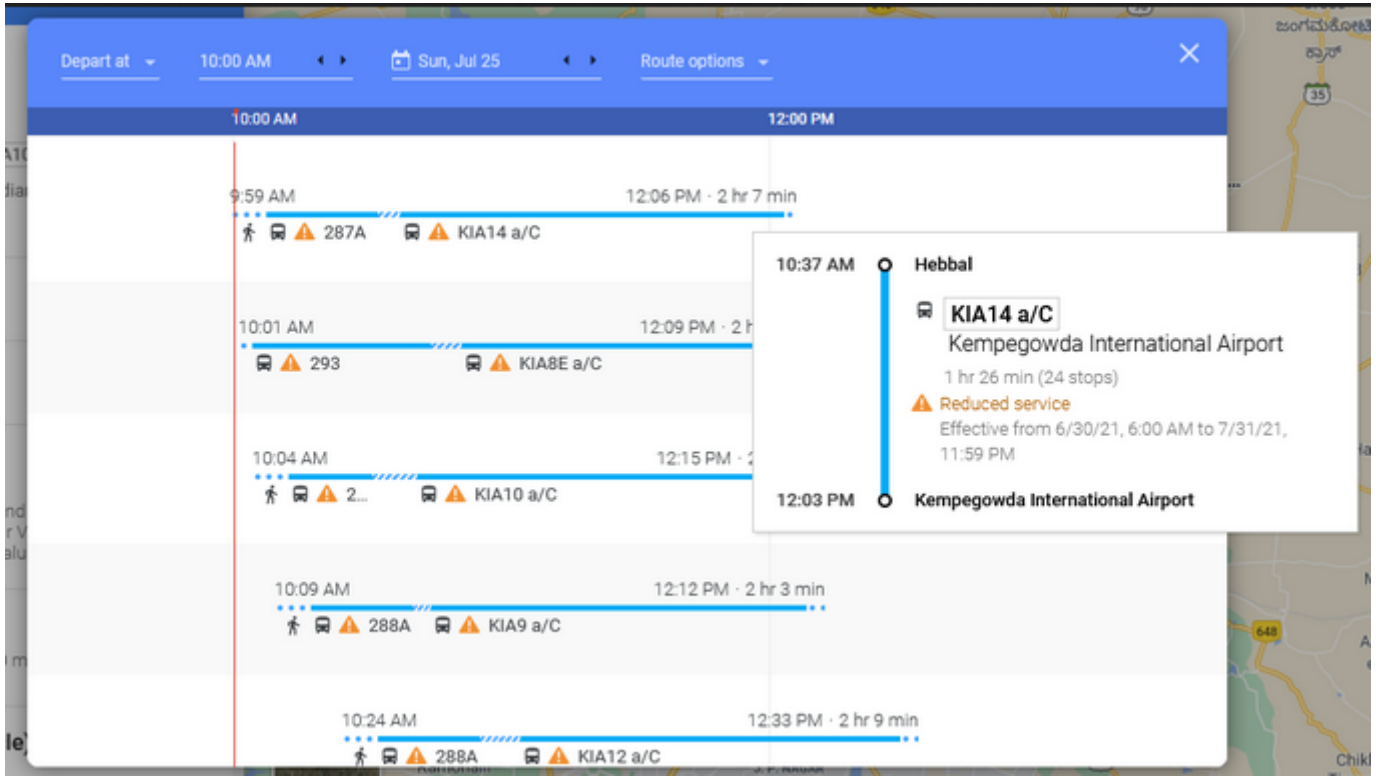
GTFS

General Transit Feed Specification (GTFS) - *A common format for public transportation schedules and associated geographic information*

Specification - GTFS Resource Center

- One of the most commonly used and adopted data standard
- If the transit information is on Google Maps, then it means that it is stored as GTFS
- Made it easier for transport agencies across the globe to share their datasets - schedules, rotues, etc
- Made transit data more accessible to the citizens

GTFS on google maps



GTFS Links

- Read more of the GTFS standard) about how the GTFS standard evolved with time.
- GTFS Datasets from selected cities can be accessed here
- OpenMobilityData gives access to GTFS feeds from around the world

Data Standards - Other Resources

- Guidebook by the ODI
  - Podcast
- Five critical questions for constructing data standards

State wise mortality data from Devdatalab

Link to dataset - Dropbox

Objectives:

1. Explore the dataset on google sheets

1. Compare state wise mortality figures in the last three years using a line chart

The Tyranny of Spreadsheets

Nearly 16,000 positive Covid cases had disappeared completely from the UK's contact tracing system. Why had the cases disappeared? Apparently, **Microsoft Excel had run out of numbers.**

> It was an astonishing story that would, in time, lead me to delve into the history of accountancy, epidemiology and vaccination, discuss file formatting with Microsoft's founder, Bill Gates, and even trace the aftershocks of the collapse of Enron. But above all, it was a story that would teach me about the way we take numbers for granted.

Read the story, by Tim Harford, here